Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, España.
Tel. y Fax +34.96.364.3560
E-mail: jose.m.bernardo@uv.es, Web: http://www.uv.es/˜bernardo/

Compuesto el 28 de Febrero de 2005
Texto de las conferencias dictadas en las Universidades de La Laguna (Tenerife)
y de Las Palmas Gran Canaria, 8 y 9 de Marzo de 2005.

Probabilidad y Estadística en los Procesos Electorales

JOSÉ-MIGUEL BERNARDO

Universitat de València, España

RESUMEN

En un régimen parlamentario, la ley electoral debe especificar la forma de distribuir los escaños disponibles entre los partidos que concurren a las elecciones, de manera que su representación política responda al apoyo que han recibido de los electores. Las distintas leyes electorales españolas hacen uso para ello la ley d'Hondt; se trata, sin embargo, de un algoritmo demostrablemente mejorable. En este artículo se describe y se justifica una solución más apropiada.

Palabras Clave: LEYES ELECTORALES; LEY D'HONDT; DIVERGENCIAS EN $\Re^K$; DIVERGENCIA ENTRE DISTRIBUCIONES DE PROBABILIDAD; DISCREPANCIA INTRÍNSECA.

1. INTRODUCCIÓN

Los resultados de últimas elecciones autonómicas catalanas, en las que una formación política (CiU) obtuvo el mayor número de escaños (46 con el 30.93% de los votos) a pesar de ser superada en votos por otra formación política (el PSC, 42 escaños con el 31.17% de los votos) pusieron de manifiesto, una vez más, la falta de idoneidad de nuestras leyes electorales.

En los regímenes parlamentarios, una ley electoral viene definida cuatro elementos bien diferenciados: (i) el número total de escaños del parlamento, (ii) su posible distribución por circunscripciones, (iii) el porcentaje mínimo de votos que debe tener un partido para poder optar a algún escaño, y (iv) el algoritmo utilizado para distribuir los escaños entre los partidos que superan ese umbral. Por ejemplo, en el caso catalán, la ley electoral vigente (aprobada como *provisional* para las primeras elecciones tras la dictadura franquista, pero nunca modificada) ordena distribuir los 135 escaños de su Parlamento en cuatro circunscripciones (Barcelona, Girona, Lleida y Tarragona con 85, 17, 15 y 18 escaños cada una, respectivamente), exige al menos un 3% de los votos válidos en toda Cataluña para poder optar a representación parlamentaria, y utiliza la ley d'Hondt para, en cada una de las circunscripciones, distribuir los escaños que le corresponden entre los partidos que han superado el umbral del 3% (PSC, CiU, ERC, PP e ICV en las elecciones del 16 de Noviembre de 2003).

José Miguel Bernardo es Catedrático de Estadística en la Universidad de Valencia. Investigación financiada con el Proyecto BNF2001-2889 de la DGICYT, Madrid.

De los cuatro elementos que definen la ley electoral, los tres primeros deben ser el resultado de una negociación política en la que es necesario valorar argumentos de índole muy diversa. Un parlamento muy numeroso permite un reflejo más preciso del apoyo obtenido por las distintas fuerzas políticas, pero es más costoso y puede resultar menos operativo. La partición del territorio en circunscripciones permite garantizar una representación mínima para cada circunscripción, pero limita seriamente la proporcionalidad del resultado final: cuanto menores sean las circunscripciones electorales, mayor será la ventaja relativa de los partidos grandes frente a los pequeños, *cualquiera* que sea el mecanismo con el que se atribuyan los escaños (Bernardo, 1999). La existencia de un nivel umbral simplifica las posibles negociaciones entre los partidos, pero puede distorsionar la pluralidad política expresada por los resultados electorales. Sin embargo, el último elemento, el algoritmo utilizado para la asignación de escaños es la solución a un problema técnico y debería ser discutido en términos técnicos. Es matemáticamente verificable que la Ley d'Hondt *no* es la solución más adecuada.

Una vez especificado el número de escaños atribuidos a cada circunscripción, todas las leyes electorales *pretenden* distribuirlos entre los partidos que han alcanzado el umbral requerido de forma que su representación política responda al apoyo que han recibido de los electores. Idealmente, el porcentaje de escaños atribuidos a cada partido en una circunscripción debería ser *proporcional* al número de votos que han obtenido es esa circunscripción. En este sentido, el artículo 68.3 de la Constitución española especifica que la atribución de diputados en cada circunscripción se realizará "atendiendo a criterios de representación proporcional". Consecuentemente, si fuese posible, el porcentaje de escaños obtenido por cada partido debería *coincidir* con el porcentaje de votos que han obtenido entre los conseguidos por todos los partidos que han superado el umbral requerido (y que tienen, por lo tanto, derecho a entrar en el reparto de escaños). Naturalmente, la coincidencia *exacta* no es posible en general, debido a que los escaños atribuidos deben ser números *enteros* no negativos. La Ley d'Hondt proporciona una posible aproximación, pero se trata de una aproximación manifiestamente mejorable. En este artículo se describe un algoritmo que permite obtener una solución al problema planteado que puede ser defendida en la práctica como la *única* solución apropiada desde el punto de vista matemático (para una descripción no técnica del problema, véase Bernardo, 2004). En general, la solución matemáticamente correcta *no* coincide con la proporcionada por la Ley d'Hondt, que debería ser eliminada de nuestras leyes electorales "por imperativo constitucional".

La asignación *óptima* de escaños, esto es la distribución de escaños más *parecida* a la distribución de votos, en el sentido (matemáticamente preciso) de minimizar la divergencia entre las distribuciones porcentuales a las que dan lugar, puede ser determinada mediante un sencillo algoritmo, que llamaremos de *mínima discrepancia*. En general, el resultado puede depender de la forma en que decida medirse la divergencia entre dos distribuciones de probabilidad. En la Sección 2 se describen las medidas de divergencia más usuales entre distribuciones de probabilidad, y se argumenta la idoneidad de la *discrepancia intrínseca*, basada en la teoría de la información. En la Sección 3 se ilustra mediante un ejmeplo real como, en la práctica, la solución óptima es esencialmente independiente de la definición de divergencia que se utilice, se describe un algoritmo que permite determinarla con facilidad, y se analizan críticamente los resultados obtenidos. En la Sección 4 se mencionan otros problemas matemáticos asociados a los procesos electorales, y se ofrecen referencias adicionales.

2.DIVERGENCIA ENTRE DISTRIBUCIONES DE PROBABILIDAD

Tanto en teoría de la probabilidad y como en estadística matemática resulta frecuentemente necesario *medir*, de forma precisa, el grado de disparidad (divergencia) entre dos distribuciones de probabilidad de un mismo vector aleatorio, $\boldsymbol{x} \in \mathcal{X}$.

Definición 1. Una función real $\ell\{\boldsymbol{p}, \boldsymbol{q}\}$ es una medida de la *divergencia* entre dos distribuciones de un vector aleatorio $\boldsymbol{x} \in \mathcal{X}$ con funciones probabilidad (o de densidad de probabilidad) $\boldsymbol{p}(\boldsymbol{x})$ y $\boldsymbol{q}(\boldsymbol{x})$ si, y sólamente si,

(i) es simétrica: $\ell\{\boldsymbol{p},\, \boldsymbol{q}\,\} = \ell\{\boldsymbol{q},\, \boldsymbol{p}\,\}$

(ii) es no-negativa: $\ell\{\boldsymbol{p},\, \boldsymbol{q}\,\} \geq 0$

(iii) $\ell\{\boldsymbol{p},\, \boldsymbol{q}\,\} = 0$ si, y sólamente si, $\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{q}(\boldsymbol{x})$ casi por todas partes.

Existen muchas formas de medir la divergencia entre dos distribuciones de probabilidad. Limitando la atención al caso discreto finito, que es el único relevante para el problema estudiado en este trabajo, una medida de divergencia entre dos distribuciones de probabilidad $\boldsymbol{p} = \{p_1, \ldots, p_k\}$ y $\boldsymbol{q} = \{q_1, \ldots, q_k\}$, con $0 \leq p_j \leq 1$ y $\sum_{j=1}^{k} p_j = 1$, $0 \leq q_j \leq 1$ y $\sum_{j=1}^{k} q_j = 1$, es cualquier función real $\ell\{\boldsymbol{p},\, \boldsymbol{q}\,\}$ *simétrica* y *no-negativa*, tal que $\ell\{\boldsymbol{p},\, \boldsymbol{q}\,\} = 0$ si, y sólamente si, $p_j = q_j$ para todo j.

En principio, cualquier medida de divergencia entre vectores de $\Re^k$ (sea o no sea una distancia métrica) podría ser utilizada. Entre las medidas de divergencia más conocidas, están la *distancia euclídea*

$$\ell_e\{\boldsymbol{p},\, \boldsymbol{q}\,\} = \Big(\sum\nolimits_{j=1}^{k} (p_j - q_j)^2 \Big)^{1/2}, \tag{1}$$

la *distancia de Hellinger*

$$\ell_h\{\boldsymbol{p},\, \boldsymbol{q}\,\} = \tfrac{1}{2} \sum\nolimits_{j=1}^{k} (\sqrt{p_j} - \sqrt{q_j})^2 \Big), \tag{2}$$

y la norma L_∞

$$\ell_\infty\{\boldsymbol{p},\, \boldsymbol{q}\,\} = \max_j |p_j - q_j|. \tag{3}$$

Sin embargo, parece más razonable utilizar una medida de divergencia que tenga en cuenta el hecho de que $\boldsymbol{p}$ y $\boldsymbol{q}$ no son vectores arbitrarios de $\Re^k$, sino que se trata, especifícamente, de *distribuciones de probabilidad*. Existen importantes argumentos axiomáticos, basados en la teoría de la información (ver Bernardo, 2005 y referencias allí citadas) para afirmar que la medida de divergencia entre distribuciones de probabilidad más apropiada es la *discrepancia intrínseca* (Bernardo y Rueda, 2002):

Definición 2. La *discrepancia intrínseca* $\delta\{\boldsymbol{p},\, \boldsymbol{q}\,\}$ entre dos distribuciones de probabilidad discretas, $\boldsymbol{p} = \{p_j,\, j \in J\}$ y $\boldsymbol{q} = \{q_j,\, j \in J\}$, es la función simétrica y no-negativa

$$\delta\{\boldsymbol{p},\, \boldsymbol{q}\,\} = \min \Big\{ k\{\boldsymbol{p} \,|\, \boldsymbol{q}\,\},\, k\{\boldsymbol{q} \,|\, \boldsymbol{p}\,\} \Big\}, \tag{4}$$

donde

$$k\{\boldsymbol{q} \,|\, \boldsymbol{p}\,\} = \sum\nolimits_{j \in J} p_j \log \frac{p_j}{q_j}\,. \tag{5}$$

Como resulta inmediato de su definición, la discrepancia intrínseca es el *mínimo valor medio del logaritmo del cociente de probabilidades* de las dos distribuciones comparadas. Puesto que para cualquier $\epsilon > 0$ suficientemente pequeño, $\log(1 + \epsilon) \approx \epsilon$, una pequeña discrepancia de valor ϵ indica un mínimo cociente esperado de probabilidades del orden de $1 + \epsilon$, esto es un error relativo medio de al menos $100\epsilon\%$.

La definición de discrepancia intrínseca se generaliza sin dificultad al caso de vectores aleatorios continuos, y puede utilizarse para definir un nuevo tipo de convergencia para distribuciones de probabilidad que goza de propiedades muy interesantes (Bernardo (2005).

La función $k\{\boldsymbol{q} \,|\, \boldsymbol{p}\}$ que aparece en la Definición 2 es la *divergencia logarítmica* de $\boldsymbol{q}$ con respecto de $\boldsymbol{p}$ (Kullback-Leibler, 1951), o *entropía cruzada*. En particular, la discrepancia intrínseca $\delta\{\boldsymbol{p}, \boldsymbol{p}_0\}$ entre una distribución discreta finita $\boldsymbol{p} = \{p_1, \ldots, p_k\}$ y la distribución uniforme $\boldsymbol{p}_0 = \{1/k, \ldots, 1/k\}$ es

$$\delta\{\boldsymbol{p}, \boldsymbol{p}_0\} = k\{\boldsymbol{p}_0 \,|\, \boldsymbol{p}\} = \log k - H(\boldsymbol{p}),$$

donde $H(\boldsymbol{p}) = -\sum_{j=1}^{k} p_j \log p_j$ es la *entropía* de la distribución $\boldsymbol{p}$, de forma que $\delta\{\boldsymbol{p}, \boldsymbol{p}_0\}$ es precisamente la cantidad de información contenida en $\boldsymbol{p}$, esto es la diferencia entre la máxima entropía posible (en el caso discreto finito), $H(\boldsymbol{p}_0) = \log k$ correspondiente a la distribución uniforme, y la entropía $H(\boldsymbol{p})$ de la distribución $\boldsymbol{p}$, (Shannon, 1948; Kullback, 1959). En general, la discrepancia intrínseca $\delta\{\boldsymbol{p}, \boldsymbol{q}\}$ es la *mínima cantidad de información necesaria*, en unidades naturales de información o *nits* (en *bits* si se utilizan logaritmos en base 2), para *discriminar* entre $\boldsymbol{p}$ y $\boldsymbol{q}$.

Es importante subrayar que la discrepancia intrínseca está bien definida incluso cuando el soporte de una de las distribuciones está estrictamente contenido en el soporte de la otra, lo que permite utilizarla para medir la bondad de muchos tipos de aproximaciones entre distribuciones de probabilidad.

Ejemplo 1. *Aproximación Poisson a una distribución Binomial.*
La discrepancia intrínseca entre una distribución Binomial $\text{Bi}(r \,|\, n, \theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$, $0 < \theta < 1$, y su aproximación Poisson, $\text{Pn}(r \,|\, n\theta) = e^{-n\theta}(n\theta)^r/r!$ viene dada por

$$\delta\{\text{Bi}(\cdot \,|\, n, \theta), \text{Pn}(\cdot \,|\, n\theta)\} = \delta\{n, \theta\} = \sum_{r=0}^{n} \text{Bi}(r \,|\, n, \theta) \log \frac{\text{Bi}(r \,|\, n, \theta)}{\text{Pn}(r \,|\, n\theta)} \,,$$

puesto que la otra suma diverge, debido a que el soporte de la distribución de Binomial, $\{0, 1, \ldots, n\}$, está estrictamente contenido en el soporte de la distribución de Poisson, $\{0, 1, \ldots\}$.

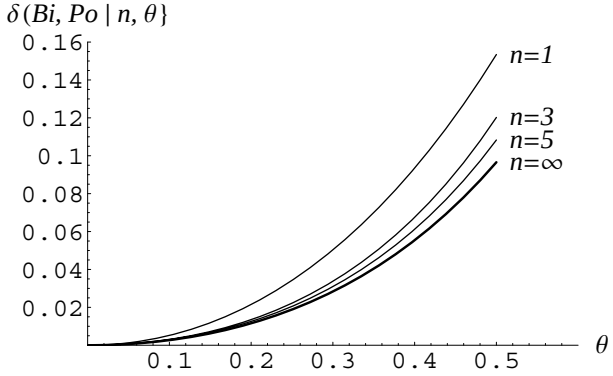


Figura 1. *Aproximación Poisson a una distribución Binomial*

En la Figura 1 se representa el valor de $\delta\{n, \theta\}$ como función de θ para distintos valores de n. Es inmediato observar que, contra lo que muchos parecen creer, la *única* condición esencial para que la aproximación sea buena es que el valor θ sea pequeño: el valor de n es prácticamente irrelevante. De hecho, cuando n crece, la discrepancia intrínseca converge rápidamente a $\delta\{\infty, \theta\} = \frac{1}{2}[-\theta - \log(1-\theta)]$, de forma que para θ fijo, el error de la aproximación no puede ser menor que ese límite por grande que sea el valor de n. Por ejemplo, para $\theta = 0.05$, $\delta\{3, \theta\} \approx 0.00074$, y $\delta\{\infty, \theta\} \approx 0.00065$, de forma que, para todo $n \geq 3$, el error relativo medio de aproximar $\text{Bi}(r \,|\, n, 0.05)$ por $\text{Pn}(r \,|\, n0.05)$ es del orden del 0.07%.

El concepto de discrepancia intrínseca permite proponer una definición general del grado de *asociación* entre dos vectores aleatorios cualesquiera.

Definición 3. La *asociación intrínseca* $\alpha_{\boldsymbol{xy}} = \alpha\{p(\boldsymbol{x}, \boldsymbol{y})\}$ entre dos vectores aleatorios discretos $\boldsymbol{x}$, $\boldsymbol{y}$ con función de probabilidad conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ es la discrepancia intrínseca $\alpha_{\boldsymbol{xy}} = \delta\{p_{\boldsymbol{xy}}, p_{\boldsymbol{x}} p_{\boldsymbol{y}}\}$ entre su función de probabilidad conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ y el producto de sus funciones de probabilidad marginales, $p(\boldsymbol{x})p(\boldsymbol{y})$.

Como en el caso de la discrepancia intrínseca, la medida de asociación intrínseca es inmediatamente generalizable al caso de vectores aleatorios continuos.

Ejemplo 2. *Medida de asociación en una tabla de contingencia.*
Sea $P = \{\pi_{ij} = \Pr[x_i, y_j]\}$, con $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$, $0 < \pi_{ij} < 1$, y $\sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} = 1$, la matriz de probabilidades asociada a una tabla de contingencia de tamaño $n \times m$, y sean $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ las correspondientes distribuciones marginales, de forma que $\boldsymbol{\alpha} = \{\alpha_i = \Pr[x_i] = \sum_{j=1}^{m} \pi_{ij}\}$, y $\boldsymbol{\beta} = \{\beta_j = \Pr[y_j] = \sum_{i=1}^{n} \pi_{ij}\}$. La medida de *asociación intrínseca* entre las variables aleatorias x e y que definen la tabla es

$$\delta\{P\} = \delta\Big\{\{\pi_{ij}\}, \{\alpha_i \beta_j\}\Big\} = \min\Big\{k\{P\},\, k_0\{P\}\Big\}$$

con $k\{P\} = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \log[\pi_{ij}/(\alpha_i \beta_j)]$, y $k_0\{P\} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \log[(\alpha_i \beta_j)/\pi_{ij}]$. El valor $\delta\{P\} = 0$ se obtiene si, y sólamente si, las variables aleatorias x e y son independientes.

Tabla 1. *Asociación intrínseca en tablas de contingencia* 2×2.

$P = \{\pi_{ij}\}$	$k\{P\}$	$k_0\{P\}$	$\delta\{P\}$
$\begin{pmatrix} 0.980 & 0.005 \\ 0.010 & 0.005 \end{pmatrix}$	0.015	0.007	0.007
$\begin{pmatrix} 0.3 & 0.2 \\ 0.1 & 0.4 \end{pmatrix}$	0.086	0.093	0.086
$\begin{pmatrix} \alpha\beta & \alpha(1-\beta) \\ (1-\alpha)\beta & (1-\alpha)(1-\beta) \end{pmatrix}$	0	0	0
$\lim_{\epsilon \to 0} \begin{pmatrix} 1/2-\epsilon & \epsilon \\ \epsilon & 1/2-\epsilon \end{pmatrix}$	$\log 2$	∞	$\log 2$

Obsérvese que el mínimo puede ser alcanzado mediante cualquiera de las dos sumas, $k\{P\}$ o $k_0\{P\}$. Por ejemplo, con $m = n = 2$, el mínimo se alcanza mediante $k_0\{P\}$ con la primera matriz de probabilidades de la Tabla 1, pero se alcanza mediante $k\{P\}$ con la segunda. En el tercer ejemplo las variables aleatorias son independientes y, consecuentemente, $\delta\{P\} = 0$. Cuando $m = n = 2$, $\alpha\{P\} < \log 2$; la asociación intrínseca correspondiente a la matriz del cuarto ejemplo tiende a $\log 2$ cuanto ϵ tiende a 0 y, consecuentemente, las variables aleatorias correspondientes tienden a una situación de máxima dependencia.

Tanto desde el punto de vista axiomático como desde el punto de vista de su comportamiento práctico (ilustrado en los ejemplos anteriores de la aproximación binomial y de la medida de asociación en tablas de contingencia), es posible afirmar que la discrepancia intrínseca es la forma más apropiada de medir la divergencia entre distribuciones de probabilidad.

En la próxima sección, se analizan las consecuencias de utilizar las distintas medidas de divergencia entre distribuciones consideradas en esta sección para determinar la forma óptima de distribuir los escaños de forma aproximadamente proporcional a los resultados electorales.

3. LA SOLUCIÓN ÓPTIMA

La asignación *óptima* es escaños es, por definición, aquella que proporciona una distribución de escaños más *parecida* a la distribución de votos en el sentido de minimizar la divergencia entre las distribuciones de probabilidad (de votos y de escaños) a que dan lugar. El resultado, en general, depende de la medida de divergencia utilizada.

Considérese primero el caso más sencillo no trivial, en el que hay que asignar *dos* escaños y en el que solamente concurren *dos* partidos A y B que han obtenido, respectivamente, una proporción p y $1-p$ de los votos. Sin pérdida de generalidad, supónganse que $0.5 < p < 1$, de forma que A es el partido más votado. Se trata de decidir a partir de que valor p_0 deberían asignarse al partido A los dos escaños en disputa. Es fácil comprobar que la ley d'Hondt asigna los dos escaños al partido mayoritario si (y solamente si) $p \geq 2/3$.

La distribución de votos entre los partidos A y B es $(p, 1-p)$. Si se asigna uno de los dos escaños al partido A, la distribución de escaños será $(1/2, 1/2)$, mientras que si se le asignasen los dos escaños al partido A la distribución de escaños sería $(1, 0)$. Consecuentemente, se trata de comparar la divergencias $\ell_1(p) = \ell\{(p, 1-p), (1/2, 1/2)\}$ y $\ell_2(p) = \ell\{(p, 1-p), (1, 0)\}$ y tomar, para cada valor de p, la menor de ellas; el punto de corte será el valor p_0 tal que $\ell_1(p_0) = \ell_2(p_0)$. En la Tabla 2, se recogen los puntos de corte correspondientes a las distintas medidas de divergencia consideradas (el valor exacto del punto corte para la distancia de Hellinger es $(2+\sqrt{2})/4 \approx 0.853$; el punto de corte correspondiente a la discrepancia intrínseca es la solución de la ecuación transcendente $\log(2p) = H(p)$, donde $H(p) = -p \log p - (1-p)\log(1-p)$ es la entropía de la distribución $(p, 1-p)$; el valor de esa solución es, aproximadamente, $p_0 = 0.811$.

Tabla 2. *Puntos de corte para la asignación de dos escaños con distintas medidas de divergencia.*

d'Hondt	Intrínseca	Euclídea	Hellinger	L_∞
2/3	0.811	3/4	0.853	3/4

Como puede observarse, la Ley d'Hondt favorece claramente al partido mayoritario, asignándole los dos escaños en disputa a partir de los 2/3 de los votos, cuando todas funciones matemáticas de divergencia propuestas lo hacen solamente a partir de los 3/4 (y la divergencia intrínseca, axiomáticamente justificable, a partir del 81%).

Considérese ahora la situación general, en la que un total de t escaños deben ser repartidos entre k partidos cuya distribución relativa de votos ha sido $\boldsymbol{p} = \{p_1, \ldots, p_k\}$, con $0 < p_i < 1$, y $\sum_{j=1}^{k} p_j = 1$. La distribución óptima de los t escaños es aquella solución *posible*, esto es de la forma $\boldsymbol{e} = \{e_1, \ldots, e_k\}$, con todos los e_j's enteros no negativos, y con $\sum_{j=1}^{k} e_j = t$, que da lugar a la distribución relativa de escaños $\boldsymbol{q} = \{q_1, \ldots, q_k\}$, con $q_j = e_j/t$, más próxima a $\boldsymbol{p}$. La solución óptima, por lo tanto, es aquella que minimiza, en el conjunto de todas las soluciones posibles, la divergencia $\ell\{\boldsymbol{p}, \boldsymbol{q}\}$ entre $\boldsymbol{p}$ y $\boldsymbol{q}$. Como se ha ilustrado en el caso particular de $t = 2$ escaños a repartir, la solución, en general, depende de la medida de divergencia entre distribuciones que se decida utilizar.

La solución *ideal* es la que distribuiría los escaños de forma exactamente proporcional a los votos obtenidos; en general, la solución ideal no suele ser una solución *posible* porque, en general, no da lugar a números enteros enteros. Sin embargo, utilizando las propiedades matemáticas de las medidas de discrepancia, es posible demostrar que, *cualquiera que sea el criterio de divergencia utilizado*, la solución óptima entre las soluciones posibles debe pertenecer al *entorno entero* de la solución ideal, constituido por todas las combinaciones

de sus aproximaciones enteras no-negativas, por defecto y por exceso, cuya suma sea igual al número t de escaños a repartir. Consecuentemente, la determinación de la solución óptima sólo requiere calcular las divergencias correspondientes a unas pocas soluciones posibles.

Como podría esperarse, las diferencias entre los resultados obtenidos para distintas medidas de divergencia tienden a desaparecer cuando aumenta el número de escaños a repartir y, en la práctica, proporcionan una misma solución (*la* solución óptima para cualquier medida de divergencia) en casi todos los casos reales, que puede ser determinada mediante un sencillo algoritmo. Este algoritmo, que llamaremos de *mínima discrepancia*, se reduce a determinar para cada partido, las diferencias absolutas entre la solución ideal y sus dos aproximaciones enteras, escogiendo sucesivamente los escaños atribuidos a cada partido por orden creciente de tales diferencias, y determinándose por diferencia los escaños que deben ser atribuidos al último partido que resulte en este proceso.

Tabla 3. *Algoritmo de asignación de escaños. Lleida 2003.*

Lleida (15 escaños)	PSC	CiU	ERC	PP	ICV	Total
Votos	**45214**	**83636**	**40131**	**19446**	**8750**	197177
Porcentaje de votos	22.93	42.42	20.35	9.96	4.44	100.00
Solución ideal	3.44	6.36	3.05	1.48	0.67	15
Límites inferiores	3	6	3	1	0	13
Límites superiores	4	7	4	2	1	18
Diferencias absolutas inferiores	0.44	0.36	0.05	0.48	0.67	
Diferencias absolutas superiores	0.56	0.64	0.95	0.52	0.33	
Solución óptima	**3**	**6**	**3**	**2**	**1**	15
Porcentaje de escaños	20.00	40.00	20.00	13.33	6.67	100.00
Solución d'Hondt	4	7	3	1	0	15
Porcentaje de escaños	26.67	46.67	20.00	6.67	0.00	100.00

Para ilustrar el algoritmo descrito se utilizan a continuación los resultados en la provincia de Lleida de las elecciones autonómicas catalanas de 2003 (ver Tabla 3). En ese caso, los votos finalmente obtenidos por los cinco partidos que podían optar a representación parlamentaria {PSC, CiU, ERC, PP, ICV} fueron, en ese orden, $\{45214, 83636, 40131, 19446, 8750\}$, es decir $\{22.93, 42.42, 20.35, 9.86, 4.44\}$ si los resultados se expresan en porcentaje de los votos obtenidos en Lleida por el conjunto de esos cinco partidos. La ley electoral vigente atribuye a Lleida 15 de los 135 escaños del parlamento catalán; para que su distribución fuese exactamente proporcional {PSC, CiU, ERC, PP, ICV} deberían recibir $\{3.44, 6.36, 3.05, 1.48, 0.67\}$ escaños respectivamente; esta sería la solución *ideal*. Se trata de *aproximar* estos valores por *números enteros*, para convertir esta solución *ideal* en una solución *posible*, y hacerlo de forma que el resultado represente una distribución porcentual de escaños cercana a la distribución porcentual de votos. El entorno entero de la solución ideal está constituido por las 10 únicas formas de asignar los 15 escaños de manera que el PSC tenga 3 o 4, CiU 6 o 7, ERC 3 o 4, PP 1 o 2 e ICV 0 o 1. La menor de las diez diferencias absolutas es 0.05, que corresponde a asignar 3 escaños a ERC; la menor de las ocho diferencias correspondientes a los cuatro partidos restantes es 0.33, que corresponde a asignar 1 escaño a ICV; la menor de las seis restantes es 0.36, que corresponde a asignar 6 escaños a CiU; la menor de las cuatro restantes es 0.44 que corresponde a asignar 3 escaños al PSC; finalmente, los 2 escaños restantes deben ser atribuidos al único partido cuyos escaños no han sido identificados todavía, el PP. La solución encontrada es

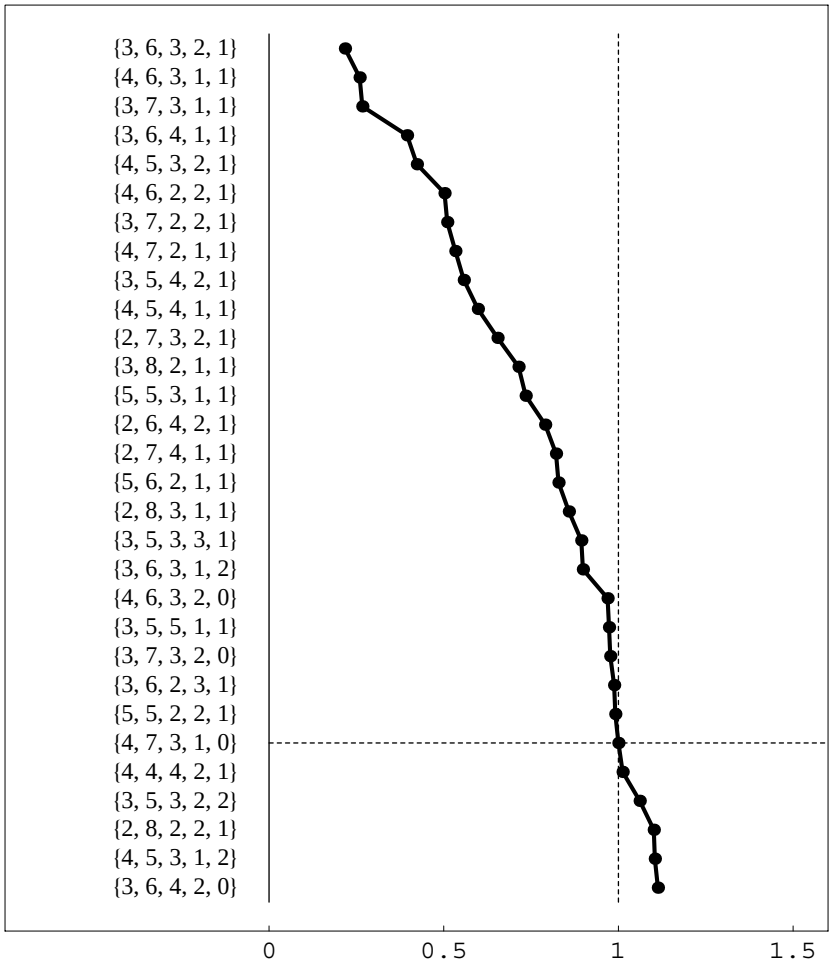


Figura 2. *Lleida 2003. Discrepancia relativa de distintas soluciones posibles para la distribución de sus* 15 *escaños con respecto a la solución d'Hondt.*

atribuir $\{3, 6, 3, 2, 1\}$ escaños a {PSC, CiU, ERC, PP, ICV}respectivamente, lo que representa el $\{20.00, 40.00, 20.00, 13.33, 6.67\}$ por ciento de los escaños.

La ley d'Hondt produce una asignación de $\{4, 7, 3, 1, 0\}$ escaños, lo que representa el $\{26.67, 46.67, 20.00, 6.67, 0.00\}$ por ciento de los escaños. Puede comprobarse que, *cualquiera que sea el criterio utilizado,* la distribución porcentual de escaños correspondiente a la solución propuesta $\{20.00, 40.00, 20.00, 13.33, 6.67\}$ está más próxima a la distribución porcentual de votos, $\{22.93, 42.42, 20.35, 9.86, 4.44\}$ que la correspondiente a la ley d'Hondt. De hecho, hemos comprobado que, entre las 3876 soluciones posibles, existen 24 asignaciones *mejores* que la proporcionada por la Ley d'Hondt, en el sentido de que dan lugar a una distribución porcentual de escaños más próxima a la distribución porcentual de votos. En La Figura 2 se listan las 30 mejores distribuciones de escaños para Lleida 2003, donde puede observarse que la solución d'Hondt ocupa el lugar 25; en la derecha de la figura se representa la discrepancia intrínseca respecto a la solución ideal de cada una de estas soluciones, utilizándose como unidad la discrepancia intrínseca de la solución d'Hondt. En particular, la solución óptima, $\{3, 6, 3, 2, 1\}$, está a 0.0120 *nits* (unidades naturales de información) de la solución ideal, el 21.7% de los 0.0552 *nits* a que se sitúa la solución d'Hondt, $\{4, 7, 3, 1, 0\}$. Puede comporbarse (ver Tabla 4) que con las demás medidas de divergencia estudiadas en la Seccíon 2 se obtienen

Tabla 4. *Divergencias con la solución ideal. Lleida 2003.*

Solución	PSC	CiU	ERC	PP	ICV	*Hellinger*	*Intrínseca*	*Euclídea*	L_∞
Ideal	3.44	6.36	3.05	1.48	0.67	0	0	0	0
Óptima	3	6	3	2	1	0.0031	0.0120	0.0562	0.52
d'Hondt	4	7	3	1	0	0.0250	0.0552	0.0788	0.67

resultados cualitativamente similares, poniendo expresamente de manifiesto la inferioridad de la solución d'Hondt.

Resulta interesante analizar la composición del parlamento catalán que se hubiese obtenido si los escaños hubiesen sido asignados de forma óptima en lugar de utilizar la Ley d'Hondt. Los partidos mayoritarios PSC y CiU hubieran perdido un escaño cada uno en favor de los dos minoritarios, PP e ICV; en particular, ICV hubiera conseguido representación en toda Cataluña. El resultado final hubiese sido $\{40, 43, 22, 16, 14\}$ en lugar de $\{41, 44, 22, 15, 13\}$.

Como podría esperarse, las diferencias entre la solución óptima y la ley d'Hondt tienden a desaparecer cuando aumenta el número de escaños a repartir. Por ejemplo, la solución d'Hondt para la distribución de los 85 escaños de la provincia de Barcelona en esas mismas elecciones, coincide con la solución óptima. Recíprocamente, las diferencias tienden a aumentar cuando en número de escaños a repartir disminuye.

El algoritmo descrito en esta sección proporciona siempre la solución óptima cuando se utiliza la distancia euclídea como medida de divergencia pero, como se ha ilustrado en el caso de Lleida, esta solución es generalmente también la solución óptima con respecto a cualquier otra medida de divergencia cuando el número de escaños a distribuir (como típicamente sucede en la práctica en España) no es extremadamente pequeño.

Finalmente, debe señalarse una ventaja política importante del algoritmo de discrepancia mínima: su extraordinaria sencillez. En marcado contraste con la Ley d'Hondt (que muy pocos ciudadanos saben utilizar, y que tan sólo los especialistas pueden pretender justificar), el algoritmo de discrepancia mínima es inmediatamente aplicable por cualquier ciudadano, y le permite apreciar con facilidad que se trata la mejor aproximación posible a la solución ideal. La substitución de la Ley d'Hondt por el algoritmo de mínima discrepancia contribuiría pues de dos formas distintas a perfeccionar nuestro sistema electoral; por una parte, haría las leyes electorales más próximas a la comprensión del ciudadano; por otra las haría más cercanas al mandato constitucional de proporcionalidad.

4. OTROS PROBLEMAS

La teoría de la probabilidad y la estadística matemática (y, muy especialmente, los métodos bayesianos objetivos) permiten ofrecer soluciones a muchos más problemas relacionados con los procesos electorales; este trabajo concluye mencionando dos de los más importantes, y proporcionando algunas referencias para su estudio.

1. Tanto los partidos políticos como los medios de comunicación conceden una notable importancia a poder disponer de predicciones muy fiables de los resultados de unas elecciones al poco tiempo de cerrar las urnas. Tales predicciones son posibles analizando, mediante métodos estadísticos *bayesianos* objetivos, los resultados de un muestreo de los primeros resultados escrutados en un conjunto de mesas electorales apropiadamente elegidas. La selección de mesas utiliza un algoritmo, basado en el uso de la discrepancia intrínseca, que procesa resultados

electorales anteriores. Las predicciones, en forma de una distribución de probabilidad sobre las posibles configuraciones del Parlamento, son obtenidas mediante el análisis bayesiano de modelos jerárquicos, implementados mediante métodos numéricos de Monte Carlo. El lector interesado en los detalles matemáticos puede consultar Bernardo (1984, 1990, 1994a), Bernardo y Girón (1992), y Bernardo (1997).

2. Una vez concluidas las elecciones, son frecuentes en los medios de comunicación las polémicas sobre las transiciones de votos que han dado lugar al nuevo mapa electoral. Tales polémicas son típicamente estériles, porque se trata de un *problema estadístico con una solución precisa*. Aunque, obviamente, existen infinitas matrices de transición de voto compatibles con los resultados *globales* de dos elecciones consecutivas, el hecho de disponer de los resultados electorales para cada una de las mesas electorales del territorio permite estimar, con un error prácticamente despreciable, la verdadera matriz de transición de voto que ha dado lugar a los nuevos resultados. Bernardo (1994b) describe uno de los algorítmos que permiten determinarla.

REFERENCIAS

Bernardo, J. M. (1984). Monitoring the 1982 Spanish socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510–515.

Bernardo, J. M. (1990). Bayesian Election Forecasting. *The New Zealand Statistician* **25**, 66–73.

Bernardo, J. M. (1994a). Optimal prediction with hierarchical models: Bayesian clustering. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 67–76.

Bernardo, J. M. (1994b). Bayesian estimation of political transition matrices. *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.). Berlin: Springer, 135–140.

Bernardo, J. M. (1997) Probing public opinion: the State of Valencia experience. *Case Studies in Bayesian Statistics 3* (C. Gatsonis, J. S. Hodges, R. E. Kass, R. McCulloch and N. D. Singpurwalla, eds.). Berlin: Springer, 3–35, (con discusión).

Bernardo, J. M. (1999). Ley d'Hondt y elecciones catalanas. *El País*, 2 de Noviembre de 1999. Madrid: Prisa.

Bernardo, J. M. (2004). Una alternativa a la Ley d'Hondt. *El País*, 2 de marzo de 2004. Madrid: Prisa.

Bernardo, J. M. (2005) Reference analysis. *Handbook of Statistics* **25**, (D. Dipak, ed.) Amsterdam: North-Holland(en prensa)

Bernardo, J. M. and Girón F. J. (1992). Robust sequential prediction from random samples: the election night forecasting case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 651–660, (con discusión).

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley. Second edition in 1968, New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656. Reprinted in *The Mathematical Theory of Communication* (Shannon, C. E. and Weaver, W., 1949). Urbana, IL.: Univ. Illinois Press.

Probabilidad y Estadística en los Procesos Electorales

José-Miguel Bernardo

Universitat de València
<jose.m.bernardo@uv.es>
www.uv.es/˜bernardo

Universidad de La Laguna, 8 Marzo 2005
Universidad de Las Palmas de Gran Canaria, 9 Marzo 2005

Contenido

1. *El problema de la asignación de escaños*
   - Elementos de una ley electoral
   - Características de una solución general
2. *Divergencia entre distribuciones de probabilidad*
   - Medidas de divergencia
   - Discrepancia intrínseca
   - Asociación intrínseca
3. *Distribución óptima de escaños*
   - El caso de dos escaños para dos partidos
   - El algoritmo de mínima discrepancia
   - Ejemplo: Lleida, autonómicas de 2003
4. *Otros problemas electorales*
   - Predicciones en la noche electoral
   - Selección de mesas electorales representativas
   - Matriz de transición de voto

1. El problema de la asignación de escaños

- En las elecciones autonómicas catalanas de Noviembre de 2003
  CiU obtuvo 46 escaños con el 30.93% de los votos.
  PSC obtuvo 42 escaños con el 31.17% de los votos.
- El artículo 68.3 de la Constitución española afirma que la asignación de escaños en cada circunscripción se realizará " atendiendo a criterios de representación porporcional"

- *Elementos de una ley electoral*
  - Número total de escaños en el Parlamento
  - Posible distribución por cirscumscripciones (*e.g.*, provincias)
  - Porcentaje umbral mínimo (*e.g.*, 3%)
  - Algoritmo utilizado para la asignación de escaños en cada circunscripción (*e.g.*, ley d'Hondt)
    Este un problema matemático. La solución proporcionada por la ley d'Hondt es incorrecta "atendiendo a criterios de representación porporcional"

- *Características de una solución general*
  - Considerese una circunscripción a la que corresponden t escaños.
  - Sean k los partidos que han superado el umbral requerido, y sean $\boldsymbol{v} = \{v_1, \ldots, v_k\}$ los votos válidos obtenidos en ella por cada uno de los partidos, lo que produce una distrubución del voto $\boldsymbol{p} = \{p_1, \ldots, p_k\}$, con $p_j = v_j/(\sum_{j=1}^{k} v_j)$, de forma que $0 < p_j < 1$, $\sum_{j=1}^{k} p_j = 1$ y $\boldsymbol{p}$ is una distribución (discreta finita) de probabilidad (la distribución del voto).
  - Sea $\boldsymbol{e} = \{e_1, \ldots, e_k\}$, una posible asignación de los t escaños: los e_j's son enteros no negativos tales que $\sum_{j=1}^{k} e_j = t$, y sea $\boldsymbol{q} = \{q_1, \ldots, q_k\}$, con $q_j = e_j/t$, $(0 \leq q_j \leq 1, \sum_{j=1}^{k} q_j = 1)$ la correspondiente distribución de los escaños.
  - El problema es elegir una asignación $\boldsymbol{e}$ de los t escaños de forma que $\boldsymbol{p}$ y $\boldsymbol{q}$ sean distribuciones tan parecidas como sea posible.

2. Divergencia entre distribuciones de probabilidad

- *Medidas de divergencia*

  Definición 1. La función real $\ell\{\boldsymbol{p}, \boldsymbol{q}\}$ es una *medida de divergencia* entre dos distribuciones de un vector aleatorio $\boldsymbol{x} \in \mathcal{X}$ con funciones de probabilidad (o de densidad de probabilidad) $\boldsymbol{p}(\boldsymbol{x})$ y $\boldsymbol{q}(\boldsymbol{x})$ si, y sólamente si,
  (i) es simétrica: $\ell\{\boldsymbol{p}, \boldsymbol{q}\} = \ell\{\boldsymbol{q}, \boldsymbol{p}\}$
  (ii) es no-negativa: $\ell\{\boldsymbol{p}, \boldsymbol{q}\} \geq 0$
  (iii) $\ell\{\boldsymbol{p}, \boldsymbol{q}\} = 0$ sii $\boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{q}(\boldsymbol{x})$ casi por todas partes.

  - Ejemplos

$$\ell_e\{\boldsymbol{p}, \boldsymbol{q}\} = (\sum_{j=1}^{k}(p_j - q_j)^2)^{1/2}, \quad \text{(Euclídea)}$$

$$\ell_h\{\boldsymbol{p}, \boldsymbol{q}\} = \tfrac{1}{2}\sum_{j=1}^{k}(\sqrt{p_j} - \sqrt{q_j})^2), \quad \text{(Hellinger)}$$

$$\ell_\infty\{\boldsymbol{p}, \boldsymbol{q}\} = \max_j |p_j - q_j|, \quad \text{(Norma } L_\infty\text{)}$$

- *Discrepancia intrínseca*

  Definición 2. La *discrepancia intrínseca* $\delta\{\boldsymbol{p}, \boldsymbol{q}\}$ entre dos distribuciones de probabilidad $\boldsymbol{p}$ y $\boldsymbol{q}$, es la función simétrica y no-negativa

$$\delta\{\boldsymbol{p}, \boldsymbol{q}\} = \min\{ k\{\boldsymbol{p} \,|\, \boldsymbol{q}\}, k\{\boldsymbol{q} \,|\, \boldsymbol{p}\}\},$$

$$k\{\boldsymbol{q} \,|\, \boldsymbol{p}\} = \sum_{j \in J} p_j \log \frac{p_j}{q_j}, \quad \text{(caso discreto)}$$

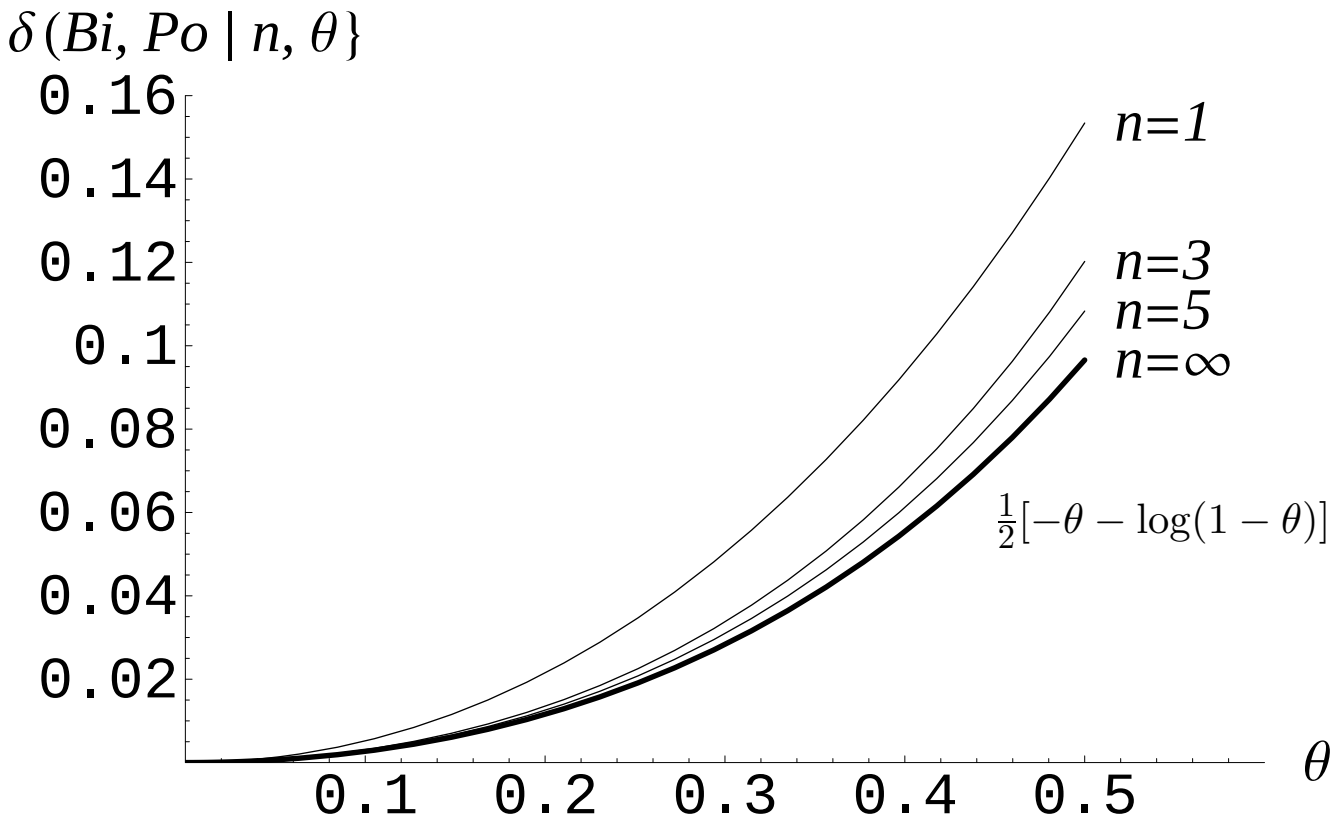
$$k\{\boldsymbol{q} \,|\, \boldsymbol{p}\} = \int_{\mathcal{X}} \boldsymbol{p}(\boldsymbol{x}) \log \frac{\boldsymbol{p}(\boldsymbol{x})}{\boldsymbol{p}(\boldsymbol{x})} d\boldsymbol{x}, \quad \text{(caso continuo)}$$

  - $\delta\{\boldsymbol{p}, \boldsymbol{q}\}$ es el *mínimo valor medio del logaritmo del cociente de probabilidades* de las dos distribuciones comparadas.
  - Puesto que para cualquier $\forall \epsilon > 0$ pequeño, $\log(1+\epsilon) \approx \epsilon$ una pequeña discrepancia ϵ indica un mínimo cociente esperado de probabilidades del orden de $1+\epsilon$, *i.e.*, un error relativo medio de al menos $100\epsilon\%$.

Ejemplo 1. *Aproximación Poisson a una distribución Binomial*.

$$\delta\{\mathrm{Bi}(\cdot\,|\,n,\theta),\mathrm{Pn}(\cdot\,|\,n\theta)\} = \delta\{n,\theta\}$$

$$= \sum_{r=0}^{n} \mathrm{Bi}(r\,|\,n,\theta)\log\frac{\mathrm{Bi}(r\,|\,n,\theta)}{\mathrm{Pn}(r\,|\,n\theta)}\ ,$$



- *Asociación intrínseca*

  Definición 3. La *asociación intrínseca* $\alpha_{\boldsymbol{xy}} = \alpha\{p(\boldsymbol{x}, \boldsymbol{y})\}$ entre dos vectores aleatorios $\boldsymbol{x}$, $\boldsymbol{y}$ con función de probabilidad (densidad de probabilidad) conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ es la discrepancia intrínseca $\alpha_{\boldsymbol{xy}} = \delta\{p_{\boldsymbol{xy}}, p_{\boldsymbol{x}} p_{\boldsymbol{y}}\}$ entre su distribución conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ y el producto $p(\boldsymbol{x})p(\boldsymbol{y})$ de sus distribuciones marginales.

  Ejemplo 2. *Medida de asociación en una tabla de contingencia*. Sea $P = \{\pi_{ij} = \Pr[x_i, y_j]\}$, la matriz de probabilidades de una tabla de contingencia de tamaño $n \times m$, y sean $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}$ sus distribuciones marginales, $\boldsymbol{\alpha} = \{\alpha_i = \Pr[x_i] = \sum_{j=1}^{m} \pi_{ij}\}$, y $\boldsymbol{\beta} = \{\beta_j = \Pr[y_j] = \sum_{i=1}^{n} \pi_{ij}\}$. La *asociación intrínseca* entre las variables aleatorias x e y que definen la tabla es $\delta\{P\} = \delta\{\{\pi_{ij}\}, \{\alpha_i \beta_j\}\} = \min\{k\{P\}, k_0\{P\}\}$, con $k\{P\} = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \log[\pi_{ij}/(\alpha_i \beta_j)]$, y $k_0\{P\} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \log[(\alpha_i \beta_j)/\pi_{ij}]$.

3. Distribución óptima de escaños

- Dada una circunscripción con t escaños a repartir entre k partidos con distrubución del voto $\boldsymbol{p} = \{p_1, \ldots, p_k\}$, se trata elegir una asignación $\boldsymbol{e}$ de los t escaños de forma que las distribuciones de votos y de escaños sean tan parecidas como sea posible.
- La distribución optima de escaños $\boldsymbol{e}^*$ se define como aquella asignación posible $\boldsymbol{e} = \{e_1, \ldots, e_k\}$ (e_j's enteros no negativos que suman t) que minimiza la discrepancia $\ell\{\boldsymbol{p}, \boldsymbol{q}\}$ entre la distribución del voto $\boldsymbol{p} = \{p_1, \ldots, p_k\}$ y la distribución de los escaños $\boldsymbol{q}$ (con $q_j = e_j/t$).
- La solución óptima e^* puede depender de la medida ℓ de discrepancia que se utilice, especialmente si el número de escaños a distribuir t es muy pequeño.
- En las elecciones generales españolas (con $t \geq 3$), la solución óptima es frecuentemente independiente de la medida de discrepancia elegida, especialmente en las provincias muy pobladas.

- *El caso de dos escaños para dos partidos*
  - Con $t = 2$, y distribucion del voto $\boldsymbol{p} = \{p, 1-p\}$, $p \geq 1/2$.
    el partido mayoritario recibe los dos escaños sii
    $\ell\{\{p, 1-p\}, \{1, 0\}\} \leq \ell\{\{p, 1-p\}, \{1/2, 1/2\}\}$.
    El punto de corte es la solución p_0 de la ecuación
    $\ell\{\{p_0, 1-p_0\}, \{1, 0\}\} = \ell\{\{p_0, 1-p_0\}, \{1/2, 1/2\}\}$

	d'Hondt	Intrínseca	Euclídea	Hellinger	L_∞
p_0	2/3	0.811	3/4	0.853	3/4

  - La ley d'Hondt favorece injustificadamente al partido mayoritario, otorgándole los dos escaños a paertir de los 2/3 de los votos, cuando todas las medidas de divergenmcia exigen al menos los 3/4 de los votos.
  - La discrepancia intrínseca, con una base axiomática, requiere al menos el 81.1% de los votos para asignar los dos escaños.

- *El algoritmo de mínima discrepancia*
  - La solución *ideal* es la que distribuiría los escaños de forma exactamente proporcional a los votos obtenidos; en general, no es solución *posible*.
  - La solución óptima debe pertenecer al *entorno entero* de la solución ideal, constituido por todas las combinaciones de sus aproximaciones enteras no-negativas, por defecto y por exceso, cuya suma sea igual al número t de escaños a repartir.
  - Algoritmo de *mínima discrepancia*: (solución euclídea)

    (i) determinar para cada partido, las diferencias absolutas entre la solución ideal y sus dos aproximaciones enteras,

    (ii) escoger sucesivamente los escaños atribuidos a $k-1$ partidos por orden creciente de esas diferencias,

    (iii)determinar por diferencia los escaños que correspondiente al partido restante.

- *Ejemplo: Lleida, autonómicas de 2003*

15 escaños	PSC	CiU	ERC	PP	ICV	Total
Votos	**45214**	**83636**	**40131**	**19446**	**8750**	197177
%votos	22.93	42.42	20.35	9.96	4.44	100.00
Ideal	3.44	6.36	3.05	1.48	0.67	15
Lím inf	3	6	3	1	0	13
Lím sup	4	7	4	2	1	18
Dif inf	0.44	0.36	0.05	0.48	0.67	
Dif sup	0.56	0.64	0.95	0.52	0.33	
Óptima	**3**	**6**	**3**	**2**	**1**	15
% escaños	20.00	40.00	20.00	13.33	6.67	100.00
d'Hondt	4	7	3	1	0	15
% escaños	26.67	46.67	20.00	6.67	0.00	100.00

(i) ERC$\rightarrow 3$, (ii) ICV$\rightarrow 1$, (iii) CiU$\rightarrow 6$, (iv) PSC$\rightarrow 3$,

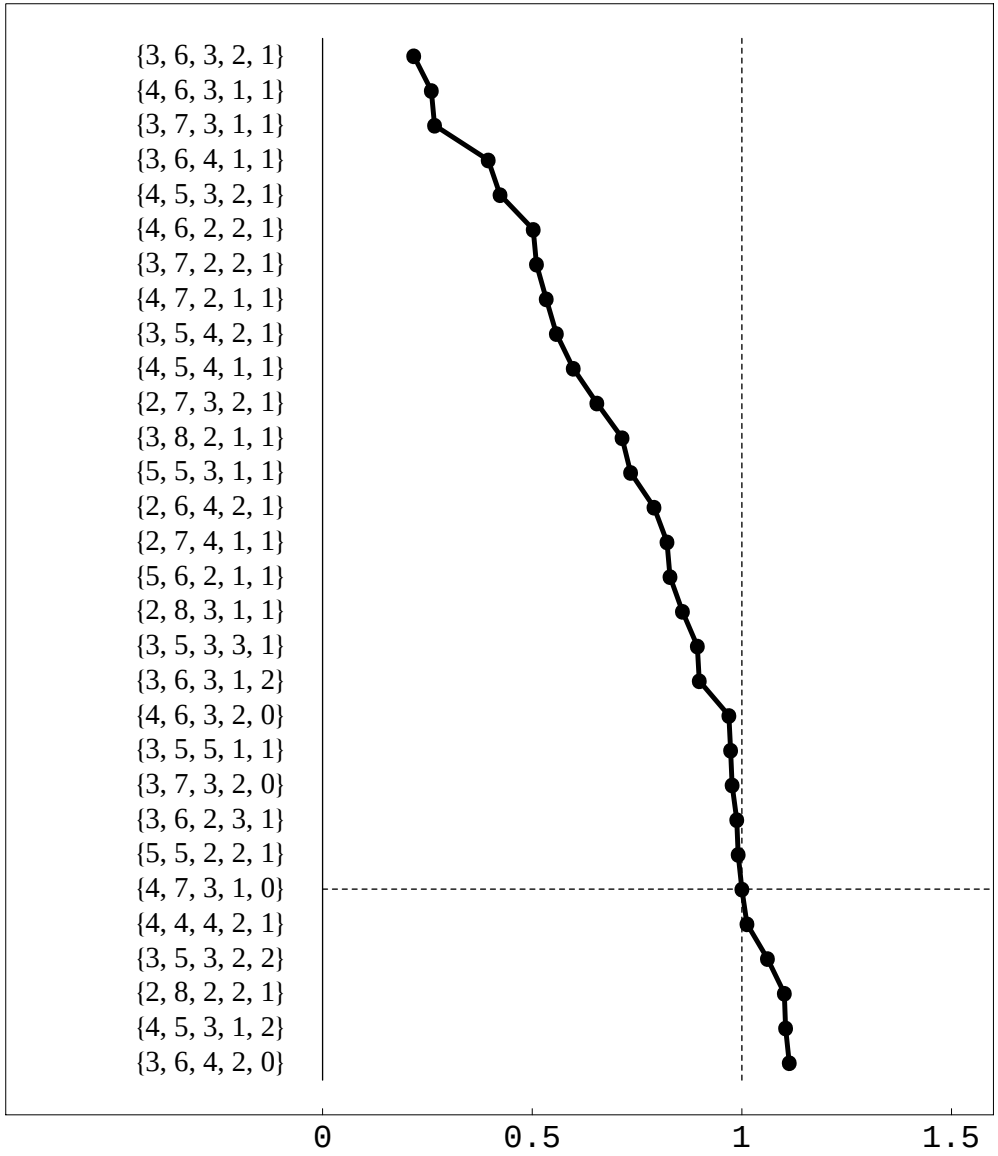
(v) PP$\rightarrow 2$, $(15-3-1-6-3=2)$

- En este caso había 24 soluciones mejores que d'Hondt:

Optima

$$\frac{\delta\{\text{óptima, ideal}\}}{\delta\{\text{d'Hondt, ideal}\}} = 0.217$$

d'Hondt



- Divergencias con respecto a la solución ideal:

Solución	PSC	CiU	ERC	PP	ICV	*Hell*	*Intr*	*Eucl*	L_∞
Ideal	3.44	6.36	3.05	1.48	0.67	0	0	0	0
Óptima	3	6	3	2	1	0.003	0.012	0.056	0.52
d'Hondt	4	7	3	1	0	0.025	0.056	0.079	0.67

- La solución $\{3, 6, 3, 2, 1\}$ es óptima con respecto a las cuatro medidas de divergencia y, en todos los casos, apreciablemente mejor que la solución d'Hondt. La solución propuesta es, bajo cualquier criterio, mucho más cercana al ideal constitucional de proporcionalidad.
- En marcado contraste con la ley d'Hondt, el algoritmo de mínima discrepancia es muy sencillo. De hecho, es fácilmente aplicable por el ciudadano medio, y le permite apreciar que se trata una buena aproximación a la solución ideal.
- La Ley d'Hondt debería desaparecer de nuestras leyes electorales.

4. Otros problemas electorales

- *Predicciones en la noche electoral*

  Predicciones precisas sobre la composición del Parlamento poco después de cerrar las urnas analizando, mediante métodos estadísticos *bayesianos* objetivos, los primeros resultados escrutados en un conjunto de mesas electorales apropiadamente elegidas.

- *Selección de mesas electorales representativas*

  El conjunto de mesas representativas minimiza su *discrepancia intrínseca media* con el resultado electoral global en una sucesión de elecciones anteriores.

- *Matriz de transición de voto*

  Aunque existen infinitas matrices de transición de voto compatibles con los resultados *globales* de dos elecciones consecutivas, los resultados electorales parciales permiten *estimar*, con un error despreciable, la matriz de transición de voto que ha dado lugar a los nuevos resultados.

O'Bayes 5

Fifth International Workshop on Objective Bayes Methodology

Branson, Missouri, USA
June 5th-8th, 2005

www.stat.missouri.edu/~bayes/Obayes5

Local Organizer and Chair:

Dongchu Sun *University of Missouri, USA.*
<dsun@stat.missouri.edu>

Organizing Committee:

Susie J. Bayarri *University of València, Spain*; James O. Berger *Duke University, USA*; José M. Bernardo *University of València, Spain*; Brunero Liseo *University of Rome, Italy*; Peter Müller*U.T. M.D. Anderson Cancer Center, USA*; Christian P. Robert *University Paris-Dauphine, France*; Paul L. Speckman *University of Missouri, USA*

Valencia Mailing List

- The Valencia Mailing List contains about 1,800 entries of people interested in Bayesian Statistics. It sends information about the Valencia Meetings and other material of interest to the Bayesian community.

8th Valencia International Meeting on Bayesian Statistics

Benidorm (Alicante), June 1st – 7th 2006

- If you do not belong to the list and want to be included, please send your data to **<valenciameeting@uv.es>**

Family name, Given name
Department, Institution
Country
Preferred e-mail address
Institutional web-site
Personal web-site
Areas of interest within Bayesian Statistics.

BAYESIAN STATISTICS 4, pp. 61±77
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, (Eds.)

Robust Sequential Prediction from Non-random Samples: the Election Night Forecasting Case*

JOSÉ M. BERNARDO and F. JAVIER GIRÓN
Generalitat Valenciana, Spain and *Universidad de Málaga, Spain*

SUMMARY

On Election Night, returns from polling stations occur in a highly non-random manner, thus posing special difÆculties in forecasting the Ænal result. Using a data base which contains the results of past elections for all polling stations, a robust hierarchical multivariate regression model is set up which uses the available returns as a training sample and the outcome of the campaign surveys as a prior. This model produces accurate predictions of the Ænal results, even with only a fraction of the returns, and it is extremely robust against data transmission errors.

Keywords: HIERARCHICAL BAYESIAN REGRESSION; PREDICTIVE POSTERIOR DISTRIBUTIONS; ROBUST BAYESIAN METHODS.

1. THE PROBLEM

Consider a situation where, on election night, one is requested to produce a sequence of forecasts of the Ænal result, based on incoming returns. Unfortunately, one cannot treat the available results at a given time as a random sample from all polling stations; indeed, returns from small rural communities typically come in early, with a vote distribution which is far removed from the overall vote distribution.

Naturally, one expects a certain geographical consistency among elections in the sense that areas with, say, a proportionally high socialist vote in the last election will still have a proportionally high socialist vote in the present election. Since the results of the past election are available for each polling station, each incoming result may be compared with the corresponding result in the past election in order to learn about the direction and magnitude of the swing for each party. Combining the results already known with a prediction of those yet to come, based on an estimation of the swings, one may hope to produce accurate forecasts of the Ænal results.

Since the whole process is done in real time, with very limited checking possibilities, it is of paramount importance that the forecast procedure (i) should deal appropriately with missing data, since reports from some polling stations may be very delayed, and (ii) should be fairly robust against the inØuence of potentially misleading data, such as clerical mistakes in the actual typing of the incoming data, or in the identiÆcation of the corresponding polling station.

* This paper has been prepared with partial Ænancial help from project number PB87-0607-C02-01/02 of the *Programa Sectorial de Promoción General del Conocimiento* granted by the *Ministerio de Educación y Ciencia*, Spain. Professor José M. Bernardo is on leave of absence from the *Departamento de Estadística e I.O., Universidad de Valencia*, Spain.

In this paper, we offer a possible answer to the problem described. Section 2 describes a solution in terms of a hierarchical linear model with heavy tailed error distributions. In Section 3, we develop the required theory as an extension of the normal hierarchical model; in Section 4, this theory is applied to the proposed model. Section 5 provides an example of the behaviour of the solution, using data from the last (1989) Spanish general election, where intentional ™errors" have been planted in order to test the robustness of the procedure. Finally, Section 6 includes additional discussion and identifies areas for future research.

2. THE MODEL

In the Spanish electoral system, a certain number of parliamentary seats are assigned to each province, roughly proportional to its population, and those seats are allocated to the competing parties using a corrected proportional system known as the Jefferson-d'Hondt algorithm (see e.g., Bernardo, 1984, for details). Moreover, because of important regional differences deeply rooted in history, electoral data in a given region are only mildly relevant to a different region. Thus, a sensible strategy for the analysis of Spanish electoral data is to proceed province by province, leaving for a final step the combination of the different provincial predictions into a final overall forecast.

Let r_{ijkl} be the proportion of the valid vote which was obtained in the last election by party i in polling station j, of electoral district k, in county l of a given province. Here, $i = 1, \ldots, p$, where p is the number of studied parties, $j = 1, \ldots, n_{kl}$, where n_{kl} is the number of polling stations in district k of county l; $k = 1, \ldots, n_l$, where n_l is the number of electoral districts in county l, and $l = 1, \ldots, m$, where m is the number of counties (*municipios*) in the province. Thus, we will be dealing with a total of

$$N = \sum_{l=1}^{m} \sum_{k=1}^{n_l} n_{kl}$$

polling stations in the province, distributed over m counties. For convenience, let $\boldsymbol{r}$ generically denote the p-dimensional vector which contains the past results of a given polling station.

Similarly, let y_{ijkl} be the proportion of the valid vote which party i obtains in the present election in polling station j, of electoral district k, in county l of the province under study. As before, let $\boldsymbol{y}$ generically denote the p-dimensional vector which contains the incoming results of a given polling station.

At any given moment, only some of the $\boldsymbol{y}$'s, say $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $0 \leq n \leq N$, will be known. An estimate of the final distribution of the vote $\boldsymbol{z} = \{z_1, \ldots, z_p\}$ will be given by

$$\hat{\boldsymbol{z}} = \sum_{i=1}^{n} \omega_i \boldsymbol{y}_i + \sum_{i=n+1}^{N} \omega_i \hat{\boldsymbol{y}}_i, \qquad \sum_{i=1}^{N} \omega_i = 1,$$

where the ω's are the relative weights of the polling stations, in terms of number of voters, and the $\hat{\boldsymbol{y}}_j$'s are estimates of the $N - n$ unobserved $\boldsymbol{y}$'s, to be obtained from the n observed results.

Within each electoral district, one may expect similar political behaviour, so that it seems plausible to assume that the observed swings should be exchangeable, i.e.,

$$\boldsymbol{y}_{jkl} - \boldsymbol{r}_{jkl} = \boldsymbol{\alpha}_{kl} + \boldsymbol{e}_{jkl}, \qquad j = 1, \ldots, n_{kl};$$

where the $\boldsymbol{\alpha}$'s describe the average swings within each electoral district and where, for robustness, the $\boldsymbol{e}$'s should be assumed to be from a heavy tailed error distribution.

Moreover, electoral districts may safely be assumed to be exchangeable within each county, so that

$$\boldsymbol{\alpha}_{kl} = \boldsymbol{\beta}_l + \boldsymbol{u}_{kl}, \qquad k = 1, \ldots, n_l,$$

where the $\boldsymbol{\beta}$'s describe the average swings within each county and where, again for robustness, the $\boldsymbol{u}$'s should be assumed to be from a heavy tailed error distribution.

Finally, county swings may be assumed to be exchangeable within the province, and thus

$$\boldsymbol{\beta}_l = \boldsymbol{\gamma} + \boldsymbol{v}_l, \qquad l = 1, \ldots, m;$$

where $\boldsymbol{\gamma}$ describes the average expected swing within the province, which will be assumed to be known from the last campaign survey. Again, for robustness, the distribution of the $\boldsymbol{v}$'s should have heavy tails.

In Section 4, we shall make the speciÆcalculations assuming that $\boldsymbol{e}$, $\boldsymbol{u}$ and $\boldsymbol{v}$ have p-variate Cauchy distributions, centered at the origin and with known precision matrices $\boldsymbol{P}_\alpha$, $\boldsymbol{P}_\beta$ and $\boldsymbol{P}_\gamma$ which, in practice, are estimated from the swings recorded between the last two elections held. The model may however be easily extended to the far more general class of elliptical symmetric distributions.

From these assumptions, one may obtain the joint posterior distribution of the average swings of the electoral districts, i.e.,

$$p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{n_m} \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$$

and thus, one may compute the posterior predictive distribution

$$p(\boldsymbol{z} \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$$

of the Ænaldistribution of the vote,

$$\boldsymbol{z} = \sum_{i=1}^{n} \omega_i \boldsymbol{y}_i + \sum_{i=n+1}^{N} \omega_i (\boldsymbol{\alpha}_i + \boldsymbol{r}_i), \qquad \sum_{i=1}^{N} \omega_i = 1,$$

where, for each i, $\boldsymbol{\alpha}_i$ is the swing which corresponds to the electoral district to which the polling station i belongs.

A Ænaltransformation, using the d'Hondt algorithm, $\boldsymbol{s} = \text{Hondt}[\boldsymbol{z}]$, which associates a partition

$$\boldsymbol{s} = \{s_1, \ldots, s_p\}, \qquad s_1 + \cdots + s_p = S$$

among the p parties of the S seats allocated to the province as a function of the vote distribution $\boldsymbol{z}$, may then be used to obtain a predictive posterior distribution

$$p(\boldsymbol{s} \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) \tag{2.1}$$

over the possible distributions among the p parties of the S disputed seats.

The predictive distributions thus obtained from each province may Ænallybe combined to obtain the desired Ænalresult, i.e., a predictive distribution over the possible Parliamentary seat conÆgurations.

3. ROBUST HIERARCHICAL LINEAR MODELS

One of the most useful models in Bayesian practice is the Normal Hierarchical Linear Model (NHLM) developed by Lindley and Smith (1972) and Smith (1973). In their model the assumption of normality was essential for the derivation of the exact posterior distributions of the parameters of every hierarchy and the corresponding predictive likelihoods. Within this setup, all the distributions involved were normal and, accordingly, the computation of all parameters in these distributions was straightforward. However, the usefulness of the model was limited, to a great extent, by the assumption of independent normal errors in every stage of the hierarchy. In this section,

(i) We first generalize the NHLM model to a multivariate setting, to be denoted NMHLM, in a form which may be extended to more general error structures.
(ii) We then generalize that model to a Multivariate Hierarchical Linear Model (MHLM) with rather general error structures, in a form which retains the main features of the NMHLN.
(iii) Next, we show that the MHLM is weakly robust, in a sense to be made precise later, which, loosely speaking, means that the usual NMHLM estimates of the parameters in every stage are distribution independent for a large class of error structures.
(iv) We then develop the theory, and give exact distributional results, for error structures which may be written as scale mixtures of matrix-normal distributions.
(v) Finally, we give more precise results for the subclass of Student's matrix-variate t distributions.

These results generalize the standard multivariate linear model and also extend some previous work by Zellner (1976) for the usual linear regression model.

A k-stage general multivariate normal hierarchical linear model MNHLM, which generalizes the usual univariate model, is given by the following equations, each representing the conditional distribution of one hyperparameter given the next in the hierarchy. It is supposed that the last stage hyperparameter, Θ_k, is known.

$$\begin{aligned} \boldsymbol{Y} \mid \Theta_1 &\sim N(\boldsymbol{A}_1\Theta_1, \boldsymbol{C}_1 \otimes \Sigma) \\ \Theta_i \mid \Theta_{i+1} &\sim N(\boldsymbol{A}_{i+1}\Theta_{i+1}, \boldsymbol{C}_{i+1} \otimes \Sigma); \qquad i = 1, \ldots k-1. \end{aligned} \tag{3.1}$$

In these equations $\boldsymbol{Y}$ is an $n \times p$ matrix which represents the observed data, the Θ_i's are the i-th stage hyperparameter matrices of dimensions $n_i \times p$ and the $\boldsymbol{A}_i$'s are design matrices of dimensions $n_{i-1} \times n_i$ (assuming that $n_0 = n$). The $\boldsymbol{C}_i$'s are positive definite matrices of dimensions $n_{i-1} \times n_{i-1}$ and, finally, Σ is a $p \times p$ positive definite matrix. The matrix of means for the conditional matrix-normal distribution at stage i is $\boldsymbol{A}_i\Theta_i$ and the corresponding covariance matrix is $\boldsymbol{C}_i \otimes \Sigma$, where $\otimes$ denotes the Kronecker product of matrices.

From this model, using standard properties of the matrix-normal distributions, one may derive the marginal distribution of the hyperparameter Θ_i, which is given by

$$\Theta_i \sim N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Sigma), \qquad i = 1, \ldots k-1,$$

where

$$\begin{aligned} \boldsymbol{B}_{ij} &= \boldsymbol{A}_{i+1} \cdots \boldsymbol{A}_j, \qquad i < j; \\ \boldsymbol{P}_i &= \boldsymbol{C}_{i+1} + \sum_{j=i+1}^{k-1} \boldsymbol{B}_{ij}\boldsymbol{C}_{j+1}B'_{ij}. \end{aligned}$$

The predictive distribution of $\boldsymbol{Y}$ given Θ_i is

$$\boldsymbol{Y} \mid \Theta_i \sim N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Sigma),$$

where

$$\boldsymbol{A}_i^* = \boldsymbol{A}_0\boldsymbol{A}_1\cdots\boldsymbol{A}_i \quad \text{with} \quad \boldsymbol{A}_0 = \boldsymbol{I};$$
$$\boldsymbol{Q}_i = \sum_{j=0}^{i-1} \boldsymbol{A}_j^*\boldsymbol{C}_{j+1}\boldsymbol{A}_i^{*\prime}.$$

From this, the posterior distribution of Θ_i given the data $\boldsymbol{Y}$, $\{\boldsymbol{A}_i\}$ and $\{\boldsymbol{C}_i\}$ is

$$\Theta_i \,|\, \boldsymbol{Y} \sim N(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i \otimes \Sigma),$$

with

$$\boldsymbol{D}_i^{-1} = \boldsymbol{A}_i^{*\prime}\boldsymbol{Q}_i^{-1}\boldsymbol{A}_i^* + \boldsymbol{P}_i^{-1};$$
$$\boldsymbol{d}_i = \boldsymbol{A}_i^{*\prime}\boldsymbol{Q}_i^{-1}\boldsymbol{Y} + \boldsymbol{P}_i^{-1}\boldsymbol{B}_{ik}\Theta_k.$$

In order to prove the basic result of this section, the MNHLM (3.1) can be more usefully written in the form

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{A}_1\Theta_1 + \boldsymbol{U}_1 \\ \Theta_i &= \boldsymbol{A}_{i+1}\Theta_{i+1} + \boldsymbol{U}_{i+1}; \qquad i = 1, \ldots k-1, \end{aligned} \tag{3.2}$$

where the matrix of error terms $\boldsymbol{U}_i$ are assumed independent $N(\boldsymbol{O}, \boldsymbol{C}_i \otimes \Sigma)$ or, equivalently, that the matrix $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_k)$ is distributed as

$$\begin{pmatrix} \boldsymbol{U}_1 \\ \vdots \\ \boldsymbol{U}_k \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{O} \\ \vdots \\ \boldsymbol{O} \end{pmatrix}; \begin{pmatrix} \boldsymbol{C}_1 & \ldots & \boldsymbol{O} \\ \vdots & \ddots & \vdots \\ \boldsymbol{O} & \ldots & \boldsymbol{C}_k \end{pmatrix} \otimes \Sigma \right]. \tag{3.3}$$

Predictive distributions for future data $\boldsymbol{Z}$ following the linear model

$$\boldsymbol{Z} = \boldsymbol{W}_1\Theta_1 + \boldsymbol{U}_W, \qquad \boldsymbol{U}_W \sim N(\boldsymbol{O}, \boldsymbol{C}_W \otimes \Sigma), \tag{3.4}$$

where $\boldsymbol{Z}$ is a $m \times p$ matrix and $\boldsymbol{U}_W$ is independent of the matrix $\boldsymbol{U}$, can now be easily derived. Indeed, from properties of the matrix-normal distributions it follows that

$$\boldsymbol{Z} \,|\, \boldsymbol{Y} \sim N(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, (\boldsymbol{W}\boldsymbol{D}_i\boldsymbol{W}' + \boldsymbol{C}_W) \otimes \Sigma). \tag{3.5}$$

Suppose now that the error vector $\boldsymbol{U}$ is distributed according to the scale mixture

$$\boldsymbol{U} \sim \int N(0, \boldsymbol{C} \otimes \Lambda)\, dF(\Lambda), \tag{3.6}$$

where $\boldsymbol{C}$ represents the matrix whose diagonal elements are the matrices $\boldsymbol{C}_i$ and the remaining elements are zero matrices of the appropriate dimensions, i.e., the diagonal covariance matrix of equation (3.3), and $F(\Lambda)$ is any matrix-distribution with support in the class of positive deÆnite $p \times p$ matrices. Clearly, the usual MNHLM (3.2) can be viewed as choosing a degenerate distribution at $\Lambda = \Sigma$ for F, while, for example, the hypothesis of $\boldsymbol{U}$ being distributed as a matrix-variate Student t distribution is equivalent to F being distributed as an inverted-Wishart distribution with appropriate parameters.

With this notation we can state the following theorem

Theorem 3.1 . *If the random matrix $\boldsymbol{U}$ is distributed according to (3.6), then*

i) the marginal distribution of Θ_i is

$$\Theta_i \sim \int N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda)\, dF(\Lambda) \qquad i = 1, \ldots k-1;$$

ii) the predictive distribution of $\boldsymbol{Y}$ given Θ_i is

$$\boldsymbol{Y} \,|\, \Theta_i \sim \int N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Lambda)\, dF(\Lambda \,|\, \Theta_i), \qquad i = 1, \ldots k-1;$$

where the posterior distribution of Λ given Θ_i, $F(\Lambda \,|\, \Theta_i)$, is given by

$$dF(\Lambda \,|\, \Theta_i) \propto |\Lambda|^{-n_i/2} \exp\Big\{ -\frac{1}{2} tr\Lambda^{-1}(\Theta_i - \boldsymbol{B}_{ik}\Theta_k)'\boldsymbol{P}_i^{-1}(\Theta_i - \boldsymbol{B}_{ik}\Theta_k)\Big\} dF(\Lambda);$$

iii) the posterior distribution of Θ_i given the data $\boldsymbol{Y}$ is

$$\Theta_i \,|\, \boldsymbol{Y} \sim \int N(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i \otimes \Lambda)\, dF(\Lambda \,|\, \boldsymbol{Y}), \qquad i = 1, \ldots k-1;$$

where the posterior distribution of Λ given $\boldsymbol{Y}$, $F(\Lambda \,|\, \boldsymbol{Y})$, is given by

$$dF(\Lambda \,|\, \boldsymbol{Y}) \propto |\Lambda|^{-n/2} \exp\Big\{ -\frac{1}{2} tr\Lambda^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)'\boldsymbol{Q}_k^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)\Big\} dF(\Lambda).$$

Proof. The main idea is, simply, to work conditionally on the scale hyperparameter Λ and, then, apply the results of the MNHLM stated above.

Conditionally on Λ, the error matrices $\boldsymbol{U}_i$ are independent and normally distributed as $\boldsymbol{U}_i \sim N(\boldsymbol{O}, \boldsymbol{C}_i \otimes \Lambda)$; therefore, with the same notation as above, we have

$$\begin{aligned} \Theta_i \,|\, \Lambda &\sim N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda), \\ \boldsymbol{Y} \,|\, \Theta_i, \Lambda &\sim N(\boldsymbol{A}_i^*\Theta_i, \boldsymbol{Q}_i \otimes \Lambda), \end{aligned}$$

and

$$\Theta_i \,|\, \boldsymbol{Y}, \Lambda \sim N(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i \otimes \Lambda); \qquad i = 1, \ldots, k.$$

Now, by Bayes theorem,

$$\frac{dF(\Lambda \,|\, \Theta_i)}{dF(\Lambda)} \propto g(\Theta_i \,|\, \Lambda), \quad \frac{dF(\Lambda \,|\, \boldsymbol{Y})}{dF(\Lambda)} \propto h(\boldsymbol{Y} \,|\, \Lambda),$$

where $g(\Theta_i \,|\, \Lambda)$ and $h(\boldsymbol{Y} \,|\, \Lambda)$ represent the conditional densities of Θ_i given Λ and $\boldsymbol{Y}$ given Λ, which are $N(\boldsymbol{B}_{ik}\Theta_k, \boldsymbol{P}_i \otimes \Lambda)$ and $N(\boldsymbol{A}_k^*\Theta_k, \boldsymbol{Q}_k \otimes \Lambda)$, respectively.

From this, by integrating out the scale hyperparameter Λ with respect to the corresponding distribution, we obtain the stated results. ◁

The theorem shows that all distributions involved are also scale mixtures of matrix-normal distributions. In particular, the most interesting distributions are the posteriors of the hyperparameters at every stage given the data, i.e., $\Theta_i \,|\, \boldsymbol{Y}$. These distributions turn out to be just a scale mixture of matrix-normals. This implies that the usual modal estimator of the Θ_i's, i.e., the mode of the posterior distribution, which is also the matrix of means for those F's with finite first moments, is $\boldsymbol{D}_i\boldsymbol{d}_i$, whatever the prior distribution F of Λ. In this sense,

these estimates are robust, that is, they do not depend on F. However, other parameters and characteristics of these distributions such as the H.P.D. regions for the hyperparameters in the hierarchy depend on the distribution F of Λ.

Note that from this theorem and formula (3.5) we can also compute the predictive distribution of future data $\boldsymbol{Z}$ generated by the model (3.4), which is also a scale mixture.

$$\boldsymbol{Z} \mid \boldsymbol{Y} \sim \int N(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, (\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W) \otimes \Lambda)\, dF(\Lambda \mid \boldsymbol{Y}). \tag{3.7}$$

More precise results can be derived for the special case in which the $\boldsymbol{U}$ matrix is distributed as a matrix-variate Student t. For the definition of the matrix-variate Student t, we follow the same notation as in Box and Tiao (1973, Chapter 8).

Theorem 3.2. *If $\boldsymbol{U} \sim t(\boldsymbol{O}, \boldsymbol{C}, \boldsymbol{S}; \nu)$ with dispersion matrix $\boldsymbol{C} \otimes \boldsymbol{S}$ and ν degrees of freedom, then*

(i) the posterior distribution of Θ_i given $\boldsymbol{Y}$ is

$$\Theta_i \mid \boldsymbol{Y} \sim t_{n_i p}(\boldsymbol{D}_i\boldsymbol{d}_i, \boldsymbol{D}_i, (\boldsymbol{S} + \boldsymbol{T}); \nu + n),$$

where the matrix $\boldsymbol{T} = (\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)'\boldsymbol{Q}_k^{-1}(\boldsymbol{Y} - \boldsymbol{A}_k^*\Theta_k)$;

(ii) the posterior distribution of Λ is an inverted-Wishart,

$$\Lambda \mid \boldsymbol{Y} \sim InW(\boldsymbol{S} + \boldsymbol{T}, \nu + n).$$

(iii) the predictive distribution of $\boldsymbol{Z} = \boldsymbol{W}_1\Theta_1 + \boldsymbol{U}_W$ is

$$\boldsymbol{Z} \mid \boldsymbol{Y} \sim t_{mp}(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, (\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W), \boldsymbol{S} + \boldsymbol{T}; \nu + n).$$

Proof. The first result is a simple consequence of the fact that a matrix-variate Student t distribution is a scale mixture of matrix-variate normals. More precisely, if $\boldsymbol{U} \sim t(\boldsymbol{O}, \boldsymbol{C}, \boldsymbol{S}; \nu)$, then $\boldsymbol{U}$ is the mixture given by (3.6), with $F \sim InW(\boldsymbol{S}, \nu)$.

From this representation and Theorem 3.1. iii), we obtain that the inverted-Wishart family for Λ is a conjugate one. In fact,

$$\begin{aligned}\frac{dF(\Lambda \mid \boldsymbol{Y})}{d\Lambda} &\propto |\Lambda|^{-n/2} \exp\Big\{ -\frac{1}{2}\mathrm{tr}\Lambda^{-1}\boldsymbol{T}\Big\} \cdot |\Lambda|^{-(\nu/2+p)} \exp\Big\{ -\frac{1}{2}\mathrm{tr}\Lambda^{-1}\boldsymbol{S}\Big\} \\ &\propto |\Lambda|^{-((\nu+n)/2+p)} \exp\Big\{ -\frac{1}{2}\mathrm{tr}\Lambda^{-1}(\boldsymbol{T} + \boldsymbol{S})\Big\};\end{aligned}$$

and (ii) follows. Finally, substitution of (ii) into (3.7) establishes (iii). ◁

4. PREDICTIVE POSTERIOR DISTRIBUTIONS OF INTEREST

In this section we specialize the results just established to the particular case of the model described in Section 2. In order to derive the predictive distribution of the random quantity $\boldsymbol{z}$ let us introduce some useful notation. Let $\boldsymbol{Y}$ denote the full $N \times p$ matrix whose rows are the vectors $\boldsymbol{y}_i$ of observed and potentially observed results, as defined in Section 2. Partition this matrix into the already observed part $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, i.e., the $n \times p$ matrix $\boldsymbol{Y}_1$ and the unobserved part, the $(N - n) \times p$ matrix $\boldsymbol{Y}_2$ formed with the remaining $N - n$ rows of $\boldsymbol{Y}$. Let $\boldsymbol{R}$ denote the $N \times p$ matrix whose rows are the vectors $\boldsymbol{r}_i$ of past results and $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ the corresponding partitions. By $\boldsymbol{X}$ we denote the matrix of swings, i.e., $\boldsymbol{X} = \boldsymbol{Y} - \boldsymbol{R}$ with $\boldsymbol{X}_1$,

$\boldsymbol{X}_2$ representing the corresponding partitions. Finally, let $\boldsymbol{\omega}$ be the row vector of weights $(\omega_1, \ldots, \omega_N)$ and $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ the corresponding partition.

With this notation the model presented in Section 2, which in a sense is similar to a random effect model with missing data, can be written as a hierarchical model in three stages as follows

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{A}_1\Theta_1 + \boldsymbol{U}_1, \\ \Theta_1 &= \boldsymbol{A}_2\Theta_2 + \boldsymbol{U}_2, \\ \Theta_2 &= \boldsymbol{A}_3\Theta_3 + \boldsymbol{U}_3; \end{aligned} \tag{4.1}$$

where $\boldsymbol{X}_1$ is a $n \times p$ matrix of known data, whose rows are of the form $\boldsymbol{y}_{jkl} - \boldsymbol{r}_{jkl}$ for those indexes corresponding to the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, Θ_1 is an $N \times p$ matrix whose rows are the p-dimensional vectors $\boldsymbol{\alpha}_{kl}$, Θ_2 is an $m \times p$ matrix whose rows are the p-dimensional vectors $\boldsymbol{\beta}_l$ and, Ænally, Θ_3 is the p-dimensional row vector $\boldsymbol{\gamma}$. The matrices $\boldsymbol{A}_i$ for $i = 1, 2, 3$ have special forms; in fact $\boldsymbol{A}_1$ is an $n \times N$ matrix whose rows are N-dimensional unit vectors, with the one in the place that matches the polling station in district k of county l from which the data arose. $\boldsymbol{A}_2$ is an $N \times m$ matrix whose rows are m-dimensional units vectors, as follows: the Ærst n_1 rows are equal to the unit vector $\boldsymbol{e}_1$, the next n_2 rows are equal to the unit vector $\boldsymbol{e}_2$, and so on, so that the last n_m rows are equal to the unit vector $\boldsymbol{e}_m$. Finally, the $m \times 1$ matrix $\boldsymbol{A}_3$ is the m-dimensional column vector $(1, \ldots, 1)$.

The main objective is to obtain the predictive distribution of $\boldsymbol{z}$ given the observed data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and the results from the last election $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$. From this, using the d'Hondt algorithm, it is easy to obtain the predictive distribution of the seats among the p parties.

The Ærst step is to derive the posterior of the $\boldsymbol{\alpha}$'s or, equivalently, the posterior of Θ_1 given $\boldsymbol{Y}$ or, equivalently, $\boldsymbol{X}_1$.

From Theorem 3.2, for $k = 3$ we have

$$\begin{aligned} \boldsymbol{D}_1^{-1} &= \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{A}_1 + (\boldsymbol{C}_2 + \boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2')^{-1} \\ \boldsymbol{d}_1 &= \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + (\boldsymbol{C}_2 + \boldsymbol{A}_2\boldsymbol{C}_3\boldsymbol{A}_2')^{-1}\boldsymbol{A}_2\boldsymbol{A}_3\boldsymbol{\gamma}. \end{aligned}$$

The computation of $\boldsymbol{D}^{-1}$ involves the inversion of an $N \times N$ matrix. Using standard matrix identities, $\boldsymbol{D}^{-1}$ can also be written in the form

$$\boldsymbol{D}_1^{-1} = \boldsymbol{A}_1'\boldsymbol{C}_1^{-1}\boldsymbol{A}_1 + \boldsymbol{C}_2^{-1} - \boldsymbol{C}_2^{-1}\boldsymbol{A}_2(\boldsymbol{A}_2'\boldsymbol{C}_2^{-1}\boldsymbol{A}_2 + \boldsymbol{C}_3^{-1})^{-1}\boldsymbol{A}_2'\boldsymbol{C}_2^{-1}$$

which may be computationally more efÆcient when the matrix $\boldsymbol{C}_2$ is diagonal and m, as in our case, is much smaller than N.

Further simpliÆcations in the formulae and subsequent computations result from the hypothesis of exchangeability of the swings formulated in Section 2. This implies that the matrices $\boldsymbol{C}_i$ are of the form $k_i\boldsymbol{I}$, where k_i are positive constants and $\boldsymbol{I}$ are identity matrices of the appropiate dimensions.

Now, the predictive model for future observations is

$$\boldsymbol{X}_2 = \boldsymbol{Y}_2 - \boldsymbol{R}_2 = \boldsymbol{W}\Theta_1 + \boldsymbol{U}_W, \qquad \boldsymbol{U}_W \sim N(\boldsymbol{O}, \boldsymbol{C}_W \otimes \boldsymbol{S});$$

where $\boldsymbol{W}$ is the $(N - n) \times N$ matrix whose rows are N-dimensional unit vectors that have exactly the same meaning as those of matrix $\boldsymbol{A}_1$.

Then, using the results of the preceding section, the predictive ditribution of $\boldsymbol{Y}_2$ given the data $\boldsymbol{Y}_1$ and $\boldsymbol{R}$ is

$$\boldsymbol{Y}_2 \sim t_{(N-n)p}(\boldsymbol{R}_2 + \boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, \boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W, \boldsymbol{S} + (\boldsymbol{Y}_1 - 1\boldsymbol{\gamma})'\boldsymbol{Q}_3^{-1}(\boldsymbol{Y}_1 - 1\boldsymbol{\gamma}); \nu + n)$$

due to the fact that the matrix $\boldsymbol{A}_3^* = 1$, where 1 is an n column vector with all entries equal to 1.

From this distribution, using properties of the matrix-variate Student t, the posterior of $\boldsymbol{z}$ which is a linear combination of $\boldsymbol{Y}_2$ is

$$\boldsymbol{z} \,|\, \boldsymbol{Y}_1, \boldsymbol{R} \sim t_{1p}(\boldsymbol{\omega}_1\boldsymbol{Y}_1 + \boldsymbol{\omega}_2\boldsymbol{R}_2 + \boldsymbol{\omega}_2\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1, \\ \boldsymbol{\omega}_2(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W)\boldsymbol{\omega}_2', \boldsymbol{S} + (\boldsymbol{Y}_1 - 1\boldsymbol{\gamma})'\boldsymbol{Q}_3^{-1}(\boldsymbol{Y}_1 - 1\boldsymbol{\gamma}); \nu + n).$$

This matrix-variate t is, in fact, a multivariate Student t distribution, so that, in the notation of Section 2,

$$p(\boldsymbol{z} \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \mathrm{St}_p(\boldsymbol{z} \,|\, \boldsymbol{m}_z, \boldsymbol{S}_z, \nu + n) \tag{4.2}$$

i.e., a p-dimensional Student t, with mean

$$\boldsymbol{m}_z = \boldsymbol{\omega}_1\boldsymbol{Y}_1 + \boldsymbol{\omega}_2\boldsymbol{R}_2 + \boldsymbol{\omega}_2\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{d}_1,$$

dispersion matrix,

$$\frac{\boldsymbol{\omega}_2(\boldsymbol{W}\boldsymbol{D}_1\boldsymbol{W}' + \boldsymbol{C}_W)\boldsymbol{\omega}_2'}{\nu + n}(\boldsymbol{S} + (\boldsymbol{Y}_1 - 1\boldsymbol{\gamma})'\boldsymbol{Q}_3^{-1}(\boldsymbol{Y}_1 - 1\boldsymbol{\gamma});$$

and $\nu + n$ degrees of freedom.

5. A CASE STUDY: THE 1989 SPANISH GENERAL ELECTION

The methodology described in Section 4 has been tested using the results, for the Province of Valencia, of the last two elections which have been held in Spain, namely the European Parliamentary Elections of June 1989, and the Spanish General Elections of October 1989.

The Province of Valencia has N = 1566 polling stations, distributed among m = 264 counties. The number n_l of electoral districts whithin each county varies between 1 and 19, and the number n_{kl} of polling stations within each electoral district varies between 1 and 57.

The outcome of the October General Election for the p = 5 parties with parliamentary representation in Valencia has been predicted, pretending that their returns are partially unknown, and using the June European Elections as the database. The parties considered were PSOE (socialist), PP (conservative), CDS (liberal), UV (conservative regionalist) and IU (communist).

	5%			20%			90%			**Final**
	Mean	Dev.	Error	Mean	Dev.	Error	Mean	Dev.	Error	
PSOE	40.08	0.46	±0.43	40.39	0.40	±0.13	40.50	0.16	±0.02	40.52
PP	23.72	0.49	±0.40	24.19	0.45	0.07	24.19	0.18	0.07	24.12
CDS	6.28	0.36	±0.20	6.33	0.33	±0.15	6.49	0.13	0.01	6.49
UV	11.88	0.50	0.44	11.62	0.46	0.17	11.42	0.17	±0.02	11.45
IU	10.05	0.40	0.03	9.93	0.37	±0.09	10.01	0.14	±0.02	10.02

10. *Table 1.*

Evolution of the percentages of valid votes.

For several proportions of known returns (5%, 20% and 90% of the total number of votes), Table 1 shows the means and standard deviations of the marginal posterior distributions of

the percentages of valid votes obtained by each of the Æve parties. The absolute error of the means with respect to the Ænal result actually obtained are also quoted.

It is fairly impressive to observe that, with only 5% of the returns, the absolute errors of the posterior modes are all smaller than 0.5%, and that those errors drop to about 0.15% with just 20% of the returns, a proportion of the vote which is usually available about two hours after the polling stations close. With 90% of the returns, we are able to quote a ™practically Ænal" result without having to wait for the small proportion of returns which typically get delayed for one reason or another; indeed, the errors all drop below 0.1% and, on election night, vote percentages are never quoted to more than one decimal place.

In Table 2, we show the evolution, as the proportion of the returns grows, of the posterior probability distribution over the possible allocation of the S=16 disputed seats.

PSOE	PP	CDS	UV	IU	5%	20%	90%	**Final**
8	4	1	2	1	0.476	0.665	0.799	1.000
7	4	1	2	2	0.521	0.324	0.201	0.000
7	5	1	2	1	0.003	0.010	0.000	0.000

10. *Table 2.*

Evolution of the probability distribution over seat partitions.

Interestingly, two seat distributions, namely $\{8, 4, 1, 2, 1\}$ and $\{7, 4, 1, 2, 2\}$, have a relatively large probability from the very beginning. This gives advance warning of the fact that, because of the intrinsically discontinuous features of the d'Hondt algorithm, the last seat is going to be allocated by a few number of votes, to either the socialists or the communists. In fact, the socialists won that seat, but, had the communists obtained 1,667 more votes (they obtained 118,567) they would have won that seat.

Tables 1 and 2 are the product of a very realistic simulation. The numbers appear to be very stable even if the sampling mechanism in the simulation is heavily biased, as when the returns are introduced by city size. The next Valencia State Elections will be held on May 26th, 1991; that night, will be the *première* of this model in *real* time.

6. DISCUSSION

The multivariate normal model NMHLM developed in Section 3 is a natural extension of the usual NHLM; indeed, this is just the particular case which obtains when $p = 1$ and the matrix $\boldsymbol{S}$ is an scalar equal to 1. As deÆned in (3.1), our multivariate model imposes some restrictions on the structure of the global covariance matrix but, this is what makes possible the derivation of simple formulae for the posterior distributions of the parameters and for the predictive distributions of future observations, all of which are matrix-variate-normal. Moreover, within this setting it is also possible, as we have demonstrated, to extend the model to error structures generated by scale mixtures of matrix-variate-normals. Actually, this may be futher extended to the class of elliptically symmetric distributions, which contains the class of scale mixtures of matrix-variate-normals as a particular case; this will be reported elsewhere. Without the restrictions we have imposed on the covariance structure, further progress on the general model seems difÆcult.

One additional characteristic of this hierarchical model, that we have not developed in this paper but merits careful attention, is the possibility of sequential updating of the hyperparameters, in a Kalman-like fashion, when the observational errors are assumed to be conditionally independent given the scale matrix hyperparameter. The possibility of combining

the Øexibility of modelling the data according to a hierarchical model, with the computational advantages of the sequential characteristics of the Kalman Ælter deserves, we believe, some attention and further research.

As shown in our motivating example, the use of sophisticated Bayesian modelling in forecasting may provide qualitatively different answers, to the point of modifying the possible uses of the forecast.

REFERENCES

Bernardo, J. M. (1984). Monitoring the 1982 Spanish Socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510±515.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B* **34**, 1±41, (with discussion).

Smith, A. F. M. (1973). A general Bayesian linear model. *J. Roy. Statist. Soc. B* **35**, 67±75.

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. *J. Amer. Statist. Assoc.* **71**, 400±405.

APPENDIX

Tables 3 and 4 below describe, with the notation used in Tables 1 and 2, what actually happened in the Province of Valencia on election night, May 26th, 1991, when $S = 37$ State Parliament seats were being contested.

	5%			20%			90%			**Final**
	Mean	Dev.	Error	Mean	Dev.	Error	Mean	Dev.	Error	
PSOE	41.5	3.6	±1.0	41.6	2.6	±0.9	42.4	2.2	±0.1	42.5
PP	23.5	3.1	0.0	23.4	2.8	±0.1	23.5	1.9	0.0	23.5
CDS	4.4	1.4	1.9	4.8	0.5	2.3	2.9	0.5	0.4	2.5
UV	14.4	2.3	±2.0	13.6	1.3	±2.8	16.0	2.0	±0.4	16.4
IU	9.2	2.0	0.9	9.4	2.2	1.1	8.6	1.9	0.3	8.3

10. *Table 3.*
Evolution of the percentages of valid votes.

PSOE	PP	CDS	UV	IU	5%	20%	90%	**Final**
18	10	0	6	3	0.06	0.02	0.82	1.00
18	9	0	7	3	0.03	0.02	0.04	0.00
17	10	2	5	3	0.03	0.47	0.01	0.00
17	9	2	5	4	0.03	0.17	0.01	0.00
17	10	1	6	3	0.36	0.02	0.01	0.00
18	9	1	6	3	0.11	0.02	0.01	0.00

10. *Table 4.*
Evolution of the probability distribution over seat partitions.

It is easily appreciated by comparison that both the standard deviations of the marginal posteriors, and the actual estimation errors, were far larger in real life than in the example. A general explanation lies in the fact that state elections have a far larger local component

than national elections, so that variances within strata were far larger, specially with the regionalists (UV). Moreover, the liberals (CDS) performed very badly in this election (motivating the resignation from their leadership of former prime minister Adolfo Suarez); this poor performance was very inhomogeneous, however, thus adding to the inflated variances. Nevertheless, essentially accurate final predictions were made with 60% of the returns, and this was done over two hours before any other forecaster was able to produce a decent approximation to the final results.

DISCUSSION

L. R. PERICCHI (*Universidad Simón Bolívar, Venezuela*)

This paper addresses a problem that has captured statisticians' attention in the past. It is one of these public problems where the case for sophisticated statistical techniques, and moreover the case for the Bayesian approach, is put to the test: quick and accurate forecasts are demanded.

The proposal described here has some characteristics in common with previous approaches and some novel improvements. In general this article raises issues of modelling and robustness.

The problem is one on which there is substantial prior information from different sources, like past elections, surveys, etc. Also, exchangeability relationships in a hierarchy are natural. Furthermore, the objective is one of prediction in the form of a probability distribution of the possible configurations of the parlament. Thus, not surprisingly, this paper, as previous articles on the same subject, Brown and Payne (1975, 1984) and Bernardo (1984), have obtained shrinkage estimators, "borrowing strength", setting the problem as a Bayesian Hierarchical Linear model. Bernardo and Girón in the present article get closer to the Brown and Payne modelling than that of Bernardo (1984), since they resort to modelling directly the "swings" rather than modelling the log-odds of the multinomial probabilities. All this, coupled with the great amount of prior information, offers the possibility of very accurate predictions from the very begining of the exercise.

A limitation of the model, as has been pointed out by the authors, is the lack of sequential updating. The incoming data is highly structured –there is certainly a bias of order of declaration– producing a trend rather than a random ordering. This prompts the need for sequential updating in a dynamic model that may be in place just before the election, as the authors confirmed in their verbal reply to the discussion.

The second limitation, is in our opinion of even greater importance and that is the lack of "strong" robustness (see below), protecting against unbounded influence of wrong information of counts and/or wrong classification of polling stations; i.e. gross errors or atypical data should not influence unduly the general prediction of the swigns. The usual hierarchical normal model has been found extremely sensitive to gross errors, possibly producing large shrinkages in the wrong direction.

At this point a short general discussion is in order. The term 'Bayesian Robustness' covers a wide field within which it can have quite different meanings. The first meaning begins with the recognition of the inevitability of imprecision of probability specifications. Even this first approach admits two different interpretations (that have similarities but also important differences). One is the "*sensitivy analysis*" interpretation (Berger, 1990), which is widely known. The second is the *upper and lower probability* interpretation. The latter is a more radical departure from precise analysis, which rejects the usual axiomatic foundations and derives directly the lower probability from its own axioms for rational behaviour, (Walley, 1990). The second meaning of robustness is closer to the Huber-Hampel notion of

assuming models (likelihoods and/or priors) that avoid unbounded inØuence of assumptions, but still work with a single probability model. The present paper uses this second meaning of robustness.

The authors address the need for robustness by replacing the normal errors throughout, by scale mixtures of normal errors. Scale mixtures of normal errors as outlier prone distributions have a long history in Bayesian analyses. They were, perhaps, Ærst proposed as a Bayesian way of dealing with outliers by de Finetti (1961) and have been sucessfully used in static and dynamic linear regression, West (1981, 1984).

Let us note in passing that the class of scale mixture of normals has been considered as a class (in the Ærst meaning of robustness mentioned above) by Moreno and Pericchi (1990). They consider an ε-contaminated model but the base prior π_0 is a scale mixture and the mixing distribution is only assumed to belong to a class H, i.e.

$$\Gamma_{\varepsilon,\pi_0}(H,Q) = \left\{ \pi(\theta) = (1-\varepsilon) \int \pi_0(\theta|r) h(dr) + \varepsilon q(\theta), q \in Q, h \in H \right\}$$

Examples of different classes of mixing distributions considered are

$$H_1 = \left\{ h(d_r) : \int_0^{r_i} h(d_r) = h_i, i = 1 \ldots n \right\}$$

$$H_2 = \left\{ h(d_r) : h(r) \text{ unimodal at } r_0 \text{ and } \int_0^{r_0} h(d_r) = h_0 \right\}$$

When π_0 is normal and $\varepsilon = 0$ then $\Gamma(H)$ is the class of scale mixtures of normal distributions with mixing distributions in H. The authors report sensible posterior ranges for probabilities of sets using H_1 and H_2.

Going back to the particular scale mixture of normals considered by Bernardo and Girón, they Ærst conveniently write the usual Multivariate Normal Hierarchical model and by restricting to a common scale matrix (Σ in (3.3) or Λ in (3.6)), they are able to obtain an elegant expression of the posterior distributions (Theorem 3.1.). Furthermore in Theorem 3.2, by specializing to a particular combination of Student-t distributions, they are able to get closed form results. This would be surprising, were it not for Zellner's (1976) conjecture: ™similar results (as those for regression) will be found with errors following a matrix Student-t". However, as with Zellner's results the authors get ™weak" rather than ™strong" robustness, in the sense that the posterior mean turns out to be linear in the observations (and therefore non-robust), although other characteristics of the distributions will be robust. However, ™*strong*" robustness is what is required, and some *ad hoc* ways to protect against outlying data (like screening) may be required. Also, approximations on combination of models that yield ™strong" robustness may be more useful than exact results. Having said that, we should bear in mind that compromises due to time pressure on election night, may have to be made given the insufÆcient development of the theory of scale mixtures of normals.

Finally, we remark that the elegant (even if too restricted) development of this paper opens wide possibilities for modelling. We should strive for more theoretical insight in the scale mixture of normals, to guide the assessment. For example O'Hagan's ™Credence" theory is still quite incomplete. Moreover, scale mixture of normals offers a much wider choice than just the Student-t, that should be explored. So far Bernardo and Girón have shown us encouraging simulations. Let us wish them well on the actual election night.

A. P. DAWID (*University College London, UK*)

It seems worth emphasising that the ™robustness" considered in this paper refers to the invariance of the results (formulae for means) in the face of varying Σ in (3.3) or (what is equivalent) the distribution F of (3.6). This distribution can be thought of either as part of the prior (Σ being a parameter) or, on using (3.6) in (3.2), as part of the model – although note that, in this latter case, the important independence (Markov) properties of the system (3.2) are lost. Relevant theory and formulae for both the general ™left-spherical" case and the particular Student-t case may be found in Dawid (1977) – see also Dawid (1981, 1988).

At the presentation of this paper at the meeting, I understood the authors to suggest that the methods also exhibit robustness in the more common sense of insensitivity to extreme data values. One Bayesian approach to this involves modelling with heavy tailed prior and error distributions, as in Dawid (1973), O'Hagan (1979, 1988) –in particular, Student-t forms are often suitable. And indeed, as pointed out at the meeting, the model does allow the possibility of obtaining such distributions for all relevant quantitities. In order to avoid any ambiguity, therefore, it must be clearly realized that, even with this choice, this model does *not* possess robustness against outliers. The Bayesian outlier-robustness theory does not apply because, as mentioned above, after using (3.6) with $F \sim InW(S, \nu)$ the (U_i) are no longer independent. Independence is vital for the heavy-tails theory to work – zero correlation is simply not an acceptable alternative. In fact, since the predictive means under the model turn out to be linear in the data, it is obvious that the methods developed in this paper can *not* be outlier-robust.

S. E. FIENBERG (*York University, Canada*)

As Bernardo and Girón are aware, others have used hierarchical Bayesian models for election night predictions. As far as I am aware the earliest such prediction system was set up in the United States.

In the 1960s a group of statisticians working for the NBC televion network developed a computer-based statistical model for predicting the winner in the U.S. national elections for President (by state) and for individual state elections for Senator anf Governor. In a presidential-election year, close to 100 predictions are made, otherwise only half that number are required. The statistical model used can be viewed as a primitive version of a Bayesian hierarchical linear model (with a fair bit of what I. J. Good would call ad hockery) and it predates the work of Lindley and Smith by several years. Primary contributors to the election prediction model development included D. Brillinger, J. Tukey, and D. Wallace. Since the actual model is still proprietary, the following description is somewhat general, and is based on my memory of the system as it operated in the 1970s.

In the 1960s an organization called the News Election Service (NES) was formed through a cooperative effort of the three national television networks and two wire services. NES collects data by precint, from individual precincts and the 3000 county reporting centers and forwards them to the networks and wire services by county (for more details, see Link, 1989). All networks get the same data at the same time from NES.

For each state, at any point in time, there are data from four sources: (i) a prior estimate of the outcome, (ii) key precints (chosen by their previous correlation with the actual outcome), (iii) county data, (iv) whole-state data (which are the numbers the networks ™oÆciallyrèport). The NBC model works with estimates of the swings of the differences between % Republican vote and % Democratic vote (a more elaborate version is used for multiple candidates) *relative* to the difference from some previous election. In addition there is a related model for turnout ratios.

The four sources of data are combined to produce an estimate of $[\%R - \%D]/2$ with

an estimated mean square error based on the sampling variance, historical information, and various bias parameters which can be varied depending on circumstances. A somewhat more elaborate structure is used to accomodate elections involving three or more major candidates. For each race the NBC model requires special settings for 78 different sets of parameters, for biases and variances, turnout adjustment factors, stratification of the state, etc. The model usually involves a geographic stratification of the state into four "substates" based on urban/suburban/rural structure and produces estimates by strata, which are then weighted by turnout to produce statewide estimates.

Even with such a computer-based model about a dozen statisticians are required to monitor the flow of data and the model performance. Special attention to the robustness of predictions relative to different historical bases for swings is an important factor, as is collateral information about where the early data are from (e.g., the city of Chicago vs. the Chicago suburbs vs. downstate Illinois).

Getting accurate early predictions is the name of the game in election night forecasting because NBC competes with the other networks on making forecasts. Borrowing strength in the Bayesian-model sense originally gave NBC an advantage over the raw data-based models employed by the other networks. For example, in 1976, NBC called 94 out of 95 races correctly (only the Presidential race in Oregon remained too close to determine) and made several calls of outcomes when the overall percentages favored the eventual loser. In the Texas Presidential race, another network called the Republican candidate as the winner early in the evening at a time when the NBC model was showing the Democratic candidate ahead (but with a large mean square error). Later this call was retracted and NBC was the first to call the Democrat the winner.

The 1980s brought a new phenomenon to U.S. election night predictions: the exit survey of voters (see Link, 1989). As a consequence, the television networks have been able to call most races long before the election polls have closed and before the precinct totals are available. All of the fancy bells and whistles of the kind of Bayesian prediction system designed by Bernardo and Girón or the earlier system designed by NBC have little use in such circumstances, unless the election race is extremely close.

REPLY TO THE DISCUSSION

We are grateful to Professor Pericchi for his valuable comments and for his wish that all go worked well on election night. As described in the Appendix above, his wish was reasonably well achieved.

He also refers to the possibility of sequential updating, also mentioned in our final discussion. Assuming, as we do in sections 2 and 4, the hypothesis of exchangeability in the swings –which implies that the $\boldsymbol{C}_i$ matrices in the model are of the form $k_i\boldsymbol{I}$– the derivation of recursive updating equations for the parameters of the posterior of Θ_1 given the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$, for $t = 1, \ldots, n$, is straightforward. However, no *simple* recursive updating formulae seem to exist for the parameters of the predictive distribution (4.2), due to the complexity of the model (4.1) and to the fact that the order in which data from the polling stations arrive is unknown a priori and, hence, the matrix $\boldsymbol{W}$ used for prediction varies with n in a form which depends on the identity of the new data.

We agree with Pericchi that weak robustness, while being an interesting theoretical extension to the usual hierarchical normal model, may not be enough for detecting gross errors. As we prove in the paper, weak robustness of the posterior mean –which is linear in the observations– is obtained under the error specification given by (3.6), independently of $F(\Lambda)$.

To obtain strong robustness of the estimators, exchangeabilty should be abandoned in favour of independence. Thus, the first equation in model (4.1), should be replaced by

$$\boldsymbol{x}_i = \boldsymbol{a}_i' \Theta_1 + \boldsymbol{u}_i, \qquad i = 1, \ldots, n,$$

where the $\boldsymbol{a}_i'$'s are the rows of matrix $\boldsymbol{A}_1$, and the error matrix $\boldsymbol{U}_i' = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_n')$ is such that the error vectors $\boldsymbol{u}_i$ are independent and identically distributed as scale mixtures of multivariate normals, i.e., $\boldsymbol{u}_i \sim \int N(0, k_1 \Lambda)\, dF(\Lambda)$.

Unfortunately, under these conditions, no closed form for the posterior is possible, except for the trivial case where $F(\cdot)$ is degenerate at some matrix, say, Σ. In fact, the posterior distribution of Θ_1 given the data is a very complex infinite mixture of matrix-normal distributions. Thus, in order to derive useful robust estimators, we have to resort to approximate methods. One possibility, which has been explored by Rojano (1991) in the context of dynamic linear models, is to update the parameters of the MHLM sequentially, considering one observation at a time, as pointed out above, thus obtaining a simple infinite mixture of matrix-normals, and then to approximate this mixture by a matrix-normal distribution, and proceed sequentially.

Professor Dawid refers again to the fact that the method described is not outlier-robust. Pragmatically, we protected ourselves from extreme outliers by screening out from the forecasting mechanism any values which were more than three standard deviations off under the appropriate predictive distribution, conditional on the information currently variable. Actually, we are developing a sequential robust updating procedure based on an approximate Kalman filter scheme adapted to the hierarchical model, that both detects and accomodates outliers on line.

We are grateful to Professor Fienberg for his detailed description of previous work on election forecasting. We should like however to make a couple of points on his final remarks.

(i) Predicting the winner in a two party race is *far* easier that predicting a parliamentary *seat distribution* among *several* parties.

(ii) In our experience, exit surveys show too much uncontrolled bias to be useful, at least if you have to forecast a seat distribution.

REFERENCES IN THE DISCUSSION

Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Planning and Inference* **25**, 303±328.

Brown, P. J. and Payne, C. (1975). Election night forecasting. *J. Roy. Statist. Soc. A* **138**, 463±498.

Brown, P. J. and Payne, C. (1984). Forecasting the 1983 British General Election. *The Statistician* **33**, 217±228.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664±667.

Dawid, A. P. (1977). Spherical matrix distributions and a multivariate model. *J. Roy. Statist. Soc. B* **39**, 254±261.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68**, 265±274.

Dawid, A. P. (1988). The infinite regress and its conjugate analysis. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Oxford: University Press, 95±110, (with discussion).

de Finetti, B. (1961). The Bayesian approach to the rejection of outliers. *Proceedings 4th Berkeley Symp. Math. Prob. Statist.* **1**, Berkeley, CA: University Press, 199±210.

Link, R. F. (1989). Election night on television. *Statistics: A Guide to the Unknown* (J. M. Tanur *et al.* eds.), Pacific Grove, CA: Wadsworth & Brooks, 104±112.

Moreno, E. and Pericchi, L. R. (1990). An ε-contaminated hierarchical model. *Tech. Rep.* Universidad de Granada, Spain.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference, *J. Roy. Statist. Soc. B* **41**, 358±367.

O'Hagan, A. (1988). Modelling with heavy tails. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Oxford: University Press, 345±359, (with discussion).

O'Hagan, A. (1990). Outliers and credence for location parameter inference. *J. Amer. Statist. Assoc.* **85**, 172±176.

Rojano, J. C. (1991). *Métodos Bayesianos Aproximados para Mixturas de Distribuciones*. Ph.D. Thesis, University of Málaga, Spain.

Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

West, M. (1981). Robust sequential approximate Bayesian estimation. *J. Roy. Statist. Soc. B* **43**, 157±166.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regressions. *J. Roy. Statist. Soc. B* **46**, 431±439.

Technical Report **06/93**, (August 30, 1993).
Presidencia de la Generalidad. Caballeros 9, 46001 - Valencia, Spain.
Tel. (34)(6) 386.6138, Fax (34)(6) 386.3626, e-mail: bernardo@mac.uv.es

Optimizing Prediction with Hierarchical Models: Bayesian Clustering

JOSÉ M. BERNARDO
Universidad de Valencia, Spain
Presidencia de la Generalidad Valenciana, Spain

SUMMARY

A frequent statistical problem is that of predicting a set of quantities given the values of some covariates, and the information provided by a training sample. These prediction problems are often structured with hierarchical models that make use of the similarities existing within classes of the population. Hierarchical models are typically based on a 'natural' definition of the clustering which defines the hierarchy, which is context dependent. However, there is no assurance that this 'natural' clustering is optimal in any sense for the stated prediction purposes. In this paper we explore the this issue by treating the choice of the clustering which defines the hierarchy as a formal decision problem. Actually, the methodology described may be seen as describing a large class of new clustering algorithms. The application which motivated this research is briefly described. The argument lies entirely within the Bayesian framework.

Keywords: BAYESIAN PREDICTION; HIERARCHICAL MODELLING; ELECTION FORECASTING; LOGARITHMIC DIVERGENCE; PROPER SCORING RULES; SIMULATED ANNEALING.

1. INTRODUCTION

Dennis Lindley taught me that interesting problems often come from interesting applications. Furthermore, he has always championed the use of Bayesian analysis in practice, specially when this has social implications. Thus, when I was asked to prepare a paper for a book in his honour, I thought it would be specially appropriate to describe some research which originated on a socially interesting area, –politics–, and may be used to broaden the applications of one of the methodologies he pioneered, –hierarchical linear models–.

2. THE PREDICTION PROBLEM

Let Ω be a set of N elements, let $\boldsymbol{y}$ be a, possibly multivariate, *quantity of interest* which is defined for each of those elements, and suppose that we are interested in some, possibly multivariate, function

$$\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$$

José M. Bernardo is Professor of Statistics at the University of Valencia, and Adviser for Decision Analysis to the President of the State of Valencia. This paper will appear in *Aspects of Uncertainty, a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.) New York: Wiley, 1994.

of the values of these vectors over Ω. Suppose, furthermore, that a vector $\boldsymbol{x}$ of covariates is also defined, that its values $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are known for all the elements is Ω, and that a random *training sample*

$$\boldsymbol{z}_n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n\},$$

which consists of n pairs of vectors $(\boldsymbol{x}, \boldsymbol{y})$, has been obtained. From a Bayesian viewpoint, we are interested in the predictive distribution

$$p(\boldsymbol{t} \mid \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N).$$

If the set Ω could be partitioned into a class $\boldsymbol{C} = \{C_i, i \in I\}$ of disjoint sets such that within each C_i the relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ could easily be modelled, it would be natural to use a hierarchical model of the general form

$$\begin{gathered} p(\boldsymbol{y}_j \mid \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}), \quad \forall j \in C_i \\ p(\boldsymbol{\theta} \mid \boldsymbol{\varphi}) \\ p(\boldsymbol{\varphi}) \end{gathered} \tag{1}$$

where $i[j]$ idenfifies the class C_i to which the j-th element belongs, $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_i)$ is a conditional probability density, totally specified by $\boldsymbol{\theta}_i$, which models the stochastic relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$ within C_i, $p(\boldsymbol{\theta} \mid \boldsymbol{\varphi})$ describes the possible interrelation among the behaviour of the different classes, and $p(\boldsymbol{\varphi})$ specifies the prior information which is available about such interrelation.

Given a specific partition $\boldsymbol{C}$, the desired predictive density $p(\boldsymbol{t} \mid \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N)$ may be computed by:

(i) deriving the posterior distribution of the $\boldsymbol{\theta}_i$'s,

$$p(\boldsymbol{\theta} \mid \boldsymbol{z}_n, \boldsymbol{C}) \propto \int \prod_{j=1}^{n} p(\boldsymbol{y}_j \mid \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}) p(\boldsymbol{\theta} \mid \boldsymbol{\varphi})\, p(\boldsymbol{\varphi})\, d\boldsymbol{\varphi}; \tag{2}$$

(ii) using this to obtain the conditional predictive distribution of the unknown $\boldsymbol{y}$'s,

$$p(\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_N \mid \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_n, \boldsymbol{C}) = \int \prod_{j=n+1}^{N} p(\boldsymbol{y}_j \mid \boldsymbol{x}_j, \boldsymbol{\theta}_{i[j]}) p(\boldsymbol{\theta} \mid \boldsymbol{z}_n, \boldsymbol{C})\, d\boldsymbol{\theta}; \tag{3}$$

(iii) computing the desired predictive density

$$p(\boldsymbol{t} \mid \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C}) = f[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n,\, p(\boldsymbol{y}_{n+1}, \ldots, \boldsymbol{y}_N \mid \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{z}_n)] \tag{4}$$

of the function of interest $\boldsymbol{t}$ as a well-defined probability transformation f of the joint predictive distribution of the unknown $\boldsymbol{y}$'s, given the appropriate covariate values $\{\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\}$ and the known $\boldsymbol{y}$ values $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$.

This solution is obviously dependent on the particular choice of the partition $\boldsymbol{C}$. In this paper, we consider the choice of $\boldsymbol{C}$ as a formal decision problem, propose a solution, which actually provides a new class of (Bayesian) clustering algorithms, and succinctly describe the case study, –Mexican State elections–, which actually motivated this research.

3. THE DECISION PROBLEM

The choice of the partition $\boldsymbol{C}$ may be seen as a decision problem where the decision space is the class of the 2^N parts of Ω , and the relevant uncertain elements are the unknown value of the quantity of interest $\boldsymbol{t}$, and the actual values of the training sample $\boldsymbol{z}_n$. Hence, to complete the specification of the decision problem, we have to define a utility function $u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]$ which measures, for each pair $(\boldsymbol{t}, \boldsymbol{z}_n)$, the desirability of the particular partition $\boldsymbol{C}$ used to build a hierarchical model designed to provide inferences about the value of $\boldsymbol{t}$, given the information provided by $\boldsymbol{z}_n$.

Since, by assumption, we are only interested in predicting $\boldsymbol{t}$ given $\boldsymbol{z}_n$, the utility function should only depend on the *reported* predictive distribution for $\boldsymbol{t}$, say $q_t(. \mid \boldsymbol{z}_n, \boldsymbol{C})$, and the actual value of $\boldsymbol{t}$, i.e., should be of the form

$$u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)] = s[q_t(. \mid \boldsymbol{z}_n, \boldsymbol{C}), \boldsymbol{t}]. \tag{5}$$

The function s is known is the literature as a *score function*, and it is natural to assume that it should be *proper*, i.e., such that its expected value should be maximized if, and only if, the reported prediction *is* the predictive distribution $p_t(. \mid \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C})$. Furthermore, in a pure inferential situation, one may want the utility of the prediction to depend only on the probability density it attaches to the true value of $\boldsymbol{t}$. In this case (Bernardo, 1979), the score function must be of the form

$$s[q_t(. \mid \boldsymbol{z}_n, \boldsymbol{C}), \boldsymbol{t}] = A \log[p(\boldsymbol{t} \mid \boldsymbol{z}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N, \boldsymbol{C})] + B, \quad A > 0. \tag{6}$$

Although, in our applications, we have always worked with this particular utility function, the algorithms we are about to describe may naturally be used with *any* utility function $u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]$.

For a given utility function u and sample size n the optimal choice of $\boldsymbol{C}$ is obviously that which maximizes the expected utility

$$u^*[\boldsymbol{C} \mid n] = \int \int u[\boldsymbol{C}, (\boldsymbol{t}, \boldsymbol{z}_n)]\, p(\boldsymbol{t}, \boldsymbol{z}_n)\, d\boldsymbol{t}\, d\boldsymbol{z}_n. \tag{7}$$

An analytic expression for $u^*[\boldsymbol{C} \mid n]$ is hardly ever attainable. However, it is not difficult to obtain a numerical approximation. Indeed, using Monte Carlo to approximate the outer integral, the value of $u^*[\boldsymbol{C} \mid m]$, for $m < n$ may be expressed as

$$u^*[\boldsymbol{C} \mid m] \approx \frac{1}{k} \sum_{l=1}^{k} \int u[\boldsymbol{C}, \boldsymbol{z}_{m(l)}, \boldsymbol{t})]\, p(\boldsymbol{t} \mid \boldsymbol{z}_{m(l)})\, d\boldsymbol{t}, \tag{8}$$

where $\boldsymbol{z}_{m(l)}$ is one of k random subselections of size $m < n$ from $\boldsymbol{z}_n$. This, in turn, may be approximated by

$$u^*[\boldsymbol{C} \mid m] \approx \frac{1}{k} \sum_{l=1}^{k} \frac{1}{n_s} \sum_{j=1}^{n_s} u[\boldsymbol{C}, \boldsymbol{z}_{m(l)}, \boldsymbol{t}_j)], \tag{9}$$

where t_j is one of n_j simulations obtained, possibly by Gibbs sampling, from $p(\boldsymbol{t} \mid \boldsymbol{z}_{m(l)})$.

Equation (9) may be used to obtain an approximation to the expected utility of any given partition C. By construction, the optimal partition will agglomerate the elements of Ω in a form which is most efficient if one is to predict $\boldsymbol{t}$ given $\boldsymbol{z}_n$. However, the practical determination of the optimal $\boldsymbol{C}$ is far from trivial.

4. THE CLUSTERING ALGORITHM

In practical situations, where N may be several thousands, an exhaustive search among all partitions $\boldsymbol{C}$ is obviously not feasible. However, the use of an agglomerative procedure to obtain a sensible initial solution, followed by an application of a simulated annealing search procedure, leads to practical solutions in a reasonable computing time.

In the aglomerative initial step, we start from the partition which consists of all the N elements as classes with a single element, and proceed to a systematic agglomeration until the expected utility is not increased by the process. The following, is a pseudocode for this procedure.

$\boldsymbol{C} := \{$all elements in$\Omega\}$
repeat
  for i:=1 **to** N
    for j:=$i+1$ **to** N
      begin
        $C^* := C \ominus (i,j), \quad \{C_i \rightarrow C_i \cup C_j)\}$
        if $u^*[\boldsymbol{C}^*] > u^*[\boldsymbol{C}]$ **then** $\boldsymbol{C} := \boldsymbol{C}^*$
      end
until No_Change

The result of this algorithm may then be used as an initial solution for a simulated annealing procedure. Simulated annealing is an algorithm of random optimization which uses as a heuristic base the process of obtaining pure crystals (annealing), where the material is slowly cooled, giving time at each step for the atomic structure of the crystal to reach its lowest energy level at the current temperature. The method was described by Kirkpatrick, Gelatt and Vechhi (1983) and has seen some statistical applications, such as Lundy (1985) and Haines (1987). The algorithm is special in that, at each iteration, one may move with positive probability to solutions with lower values of the function to maximize, rather than directly jumping to the point with the highest value within the neighborhood, thus drastically reducing the chances of getting trapped in local maxima. The following, is a pseudocode for this procedure.

get Initial_Solution $\boldsymbol{C}_0$, Initial_Temperature t_0, Initial_Distance d_0;
$\boldsymbol{C} := \boldsymbol{C}_0; t := t_0; d := d_0;$
repeat
while (**not** d-Finished) **do**
  begin
    while (**not** t-Optimized) **do**
      begin
        Choose_Random$(\boldsymbol{C}_i \mid d)$
        $\delta := u^*[\boldsymbol{C}_i] - u^*[\boldsymbol{C}_0]$
        if $(\delta \geq 0)$ **then** $\boldsymbol{C} := \boldsymbol{C}_i$
        else if $(\exp\{-\delta/t\} \leq$ Random$)$ **then** $\boldsymbol{C} := \boldsymbol{C}_i$
      end;
    $t := t/2$
  end;
Reduce_Distance(d)
until $d < \varepsilon$

In the annealing procedure, the distance among two partitions is defined as the number of different classes it contains.

5. AN APPLICATION TO ELECTION FORECASTING

Consider a situation where, on election night, one is requested to produce a sequence of forecasts of the final result, based on incoming returns. Since the results of the past election are available for each polling station, each incoming result may be compared with the corresponding result in the past election in order to learn about the direction and magnitude of the swing for each party. Combining the results already known with a prediction of those yet to come, based on an estimation of the swings, in each of a set of appropriately chosen strata, one may hope to produce accurate forecasts of the final results.

In Bernardo and Girón (1992), a hierarchical prediction model for this problem was developed, using electoral districts within counties as a 'natural' partition for a three stage hierarchy, and the results were successfully applied some weeks later to the Valencia State Elections. One may wonder, however, whether the geographical clustering used in the definition of the hierarchical model is optimal for the stated prediction purposes.

With the notation of this paper, a two-stage hierarchical model for this problem is defined by the set of equations

$$\begin{gathered}\boldsymbol{y}_{j[i]} = \boldsymbol{x}_{j[i]} + \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_{0j[i]}, \quad j \in C_i, \qquad p(\boldsymbol{\varepsilon}_{0j[i]} \,|\, \boldsymbol{\alpha}_0), \quad E[\boldsymbol{\varepsilon}_{0j[i]}] = 0 \\ \boldsymbol{\theta}_i = \boldsymbol{\varphi} + \boldsymbol{\varepsilon}_{1i}, \quad i \in I, \qquad p(\boldsymbol{\varepsilon}_{1i} \,|\, \boldsymbol{\alpha}_1), \quad E[\boldsymbol{\varepsilon}_{1i}] = 0 \\ \pi(\boldsymbol{\varphi}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)\end{gathered} \tag{10}$$

where $\boldsymbol{y}_{j[i]}$ is the vector which describes the results on the new election in polling station j which belongs to class C_i, $\boldsymbol{x}_{j[i]}$ contains the corresponding results in the past election, the error distributions of $\boldsymbol{\varepsilon}_0 = (\boldsymbol{\varepsilon}_{01[1]}, \ldots,)$ and $\boldsymbol{\varepsilon}_1 = (\boldsymbol{\varepsilon}_{11}, \ldots,)$, $p(\boldsymbol{\varepsilon}_0 \,|\, \boldsymbol{\alpha}_0)$ and $p(\boldsymbol{\varepsilon}_1 \,|\, \boldsymbol{\alpha}_1)$, have zero mean and are fully specified by the hiperparameters $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$, and $\pi(\boldsymbol{\varphi}, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1)$ is the reference distribution (Berger and Bernardo, 1992) which corresponds to this model.

The function of interest is the probability vector which describes the final results of the new election, i.e.,

$$\boldsymbol{t} = \sum_{i \in I} \sum_{j \in C_i} \beta_{j[i]} \boldsymbol{y}_{j[i]} \tag{11}$$

where $\beta_{j[i]}$ is the (known) proportion of the population which lives in the poling station j of class C_i. The posterior distribution of $\boldsymbol{t}$ may be derived using the methods described above.

In this particular application, however, interest is essentially centered on a good estimate of $\boldsymbol{t}$. Given some results from the new election, i.e., the training sample $\boldsymbol{z}_n$, the value of $\boldsymbol{t}$ may be decomposed into its known and unknown parts, so that the expected value of the posterior distribution of $\boldsymbol{t}$ may be written as

$$E[\boldsymbol{t} \,|\, \boldsymbol{z}_n] = \sum_{i \in I} \sum_{j \in \text{ Obs}} \beta_{j[i]} \boldsymbol{y}_{j[i]} + \sum_{i \in I} \sum_{j \in \text{ NoObs}} \beta_{j[i]} E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n], \tag{12}$$

where

$$E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n] = \boldsymbol{x}_{j[i]} + \int \int E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1] \, p(\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1 \,|\, \boldsymbol{z}_n) \, d\boldsymbol{\alpha}_0 \, d\boldsymbol{\alpha}_1. \tag{13}$$

The conditional expectation within the double integral may be analytically found under different sets of conditions. In their seminal paper on hierarchical models, Lindley and Smith (1972) already provided the relevant expressions under normality, when $\boldsymbol{y}$ is univariate. Bernardo and Girón (1992) generalize this to multivariate models with error distributions which may be expressed as scales mixtures of normals; this includes heavy tailed error distributions such

as the matrix-variate Student t's. If an analytical expression for the conditional expectation $E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, \boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1]$ may be found, then an approximation to $E[\boldsymbol{y}_{j[i]} \,|\, \boldsymbol{z}_n]$ may be obtained by using Gibbs sampling to approximate the expectation integral.

In particular, when the error structure may be assumed to have the simple form

$$D^2[\boldsymbol{\varepsilon}_0 \,|\, h_0, \Sigma] = \frac{1}{h_0}(\boldsymbol{I} \otimes \Sigma), \qquad D^2[\boldsymbol{\varepsilon}_1 \,|\, h_1, \Sigma]] = \frac{1}{h_1}(\boldsymbol{I} \otimes \Sigma), \tag{14}$$

where the $\boldsymbol{I}$'s are identity matrices of appropriate dimensions and $\otimes$ denotes the Kronecker product of matrices, and when the error distribution is expressable as a scale mixture of normals, then the conditional reference reference distribution $\pi(\boldsymbol{\varphi}, \,|\, h_0, h_1, \Sigma)$ is uniform and the first moments of the conditional posterior distribution of the $\boldsymbol{\theta}_i$'s are given by

$$E[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, h_0, h_1, \Sigma] = \frac{n_i h_0 \boldsymbol{r}_{.i} + h_1 \boldsymbol{r}_{..}}{n_i h_0 + h_1} \tag{15}$$

$$D^2[\boldsymbol{\theta}_i \,|\, \boldsymbol{z}_n, h_0, h_1, \Sigma] = \frac{1}{n_i h_0 + h_1}\Sigma, \tag{16}$$

where n_i is the number of polling stations the sample which belong to class $\boldsymbol{C}_i$,

$$\boldsymbol{r}_{.i} = \frac{1}{n_i} \sum_{j \in \boldsymbol{C}_i} \left(\boldsymbol{y}_{j[i]} - \boldsymbol{x}_{j[i]} \right), \quad i \in I \tag{17}$$

are the average sample swings within class $\boldsymbol{C}_i$, and

$$\boldsymbol{r}_{..} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{y}_j - \boldsymbol{x}_j = \overline{\boldsymbol{r}}_{.i} \tag{18}$$

is the overall average swing.

Since (14) are the rather natural assumptions of exchangeability within classes, and exchangeability among classes, and (15) remains valid for rather general error distributions, (12), (13), and Gibbs integration over (15) provide together a practical mechanism to implement the model described.

6. A CASE STUDY: STATE ELECTIONS IN MEXICO

On February 1993, I was invited by the Mexican authorities to observe their Hidalgo State elections, in order to report on the feasibility of implementing in Mexico the methods developed in Valencia. Although I was not supposed to do any specific analysis of this election, I could not resist the temptation of trying out some methods.

I had taken with me the code of the algorithm I use to select a set of constituencies which, when viewed as a whole, have historically produced, for each election, a result close to the global result. The procedure, which is another application of simulated annealing, is described in Bernardo (1992).

Using the results of the 1989 election in Hidalgo (which were the only available ones), I used that algorithm to select a set of 70 polling stations whose joint behaviour had been similar to that of the State as a whole, and suggested that the local authorities should send agents to those polling stations to report on the phone the corresponding returns as soon as they were counted. A number of practical problems reduced to 58 the total number of results which were available about two hours after the polling stations closed.

In the mean time, I was busy setting up a very simple forecasting model –with no hierarchies included–, programmed in Pascal in a hurry on a resident Macintosh, to forecast the final results based on those early returns. This was in fact the particular case which corresponds to the model described in Section 4, if the partition C is taken to have a single class, namely the whole Ω.

About 24 hours later, just before the farewell dinner, the provisional official results came in. Table 1, Line 1, contains the official results, in percentage of valid votes of PAN (right wing), PRI (government party), PRD (left wing) and other parties. As it is apparent from Table 1, Line 2, my forecasts were not very good; the mean absolute error (displayed as the loss column in the table, was 3.28. Naturally, as soon as I was back in Valencia, I adapted the hierarchical software which I have been using here. The results (Table 1, Line 3) were certainly far better, but did not quite met the standards I was used to in Spain.

State of Hidalgo, February 21st, 1993					
	PAN	PRI	PRD	Others	Loss
Oficial Results	8.30	80.56	5.56	5.56	
No hierarchies	5.5	76.8	9.3	8.4	3.28
Districts as clusters	6.4	80.6	7.7	5.3	1.09
Optimal clustering	8.23	80.32	6.18	5.27	0.31

Table 1. *Comparative methodological analysis.*

On closer inspection, I discovered that the variances within the districts used as clusters in the hierarchical model were far higher than the corresponding variances in Spain. This prompted an investigation on the possible ways to reduce such variances and, naturally, this lead to the general procedures described in this paper.

We used repeated random subselection of size 58 from the last election results in Hidalgo in order to obtain, –using the algorithms described in Section 3–, the 1989 optimal partition of the polling stations. In practice, we made the exangeability assumptions described by (14), assumed Cauchy error distributions, and chose a logarithmic scoring rule. We then used this partition to predict the 1993 election, using the two-stage hirearchical model described in Section 4 and the 58 available polling station results. The results are shown in Table 1, Line 4; it is obvious from them that the research effort did indeed have a practical effect in the Hidalgo data set.

7. DISCUSSION

Prediction with hierarchical models is a very wide field. Although very often, the clustering which defines the hierarchy has a natural definition, this is not necessarily optimal from a prediction point of view. If the main object of the model is prediction, it may be worth to explore alternative hierarchies, and the preceding methods provide a promising way to do this.

Moreover, there are other situations where the appropriate clustering is less than obvious. For instance, a model similar to that described here may be used to estimate the total personal income of a country, based on the covariates provided by the census and a training sample which consists of the personal incomes of a random sample of the population and their associated census covariates. The clustering which would be provided by the methods described here may have indeed an intrinsic sociological interest, which goes beyond the stated prediction problem.

Finally, the whole system may be seen as a proposal of a large class of well-defined clustering algorithms, where –as one would expect in any Bayesian solution–, the objectives of the problem are precisely defined. These could be compared with the rather *ad hoc* standard clustering algorithms as explorative data analysis methods used to improve our understanding of complex multivariate data sets.

REFERENCES

Berger, J. O. and Bernardo, J. M. (1992). On the development of the reference prior method. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 35–60 (with discussion).

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1992). Simulated annealing in Bayesian decision theory. *Computational Statistics* **1** (Y. Dodge and J. Whittaker, eds.) Heidelberg: Physica-Verlag, pp.547–552.

Bernardo, J. M. and Girón, F. J. (1992). Robust sequential prediction form non-random samples: the election night forecasting case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), Oxford: University Press, 61–77, (with discussion).

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. *Technometrics* **29**, 439–447.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. B* **34**, 1-41, (with discussion).

Lundy, M. (1985). Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika* **72**, 191–198.

Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, Spain.
Tel. 34.6.363.6048, Fax 34.6.363.6048 (direct), 34.6.386.4735 (office)
Internet: bernardo@uv.es, Web: http://www.uv.es/~bernardo/

Printed on December 9, 1996
Invited paper presented at the *Third Workshop on Bayesian Statistics in Science and Technology: Case Studies*, held at Carnegie Mellon University, Pittsburgh, U. S. A., October 5–7, 1995

Probing Public Opinion: the State of Valencia Experience

JOSÉ M. BERNARDO

Universitat de València, Spain

SUMMARY

This paper summarizes the procedures which have been set up during the last years at the Government of the State of Valencia, Spain, to systematically probe its public opinion as an important input into its decision processes.

After a brief description of the electoral setup, we (i) outline the use of a simulated annealing algorithm, designed to find a good design for sample surveys, which is based on the identification of representative electoral sections, (ii) describe the methods used to analyze the data obtained from sample surveys on politically relevant topics, (iii) outline the proceedings of election day —detailing the special problems posed by the analysis of exit poll, representative sections, and early returns data— and (iv) describe a solution to the problem of estimating the political transition matrices which identify the reallocation of the vote of each individual party between two political elections.

Throughout the paper, special attention is given to the illustration of the methods with real data. The arguments fall entirely within the Bayesian framework.

Keywords: BAYESIAN PREDICTION; HIERARCHICAL MODELLING; ELECTION FORECASTING; LOGARITHMIC DIVERGENCE; SAMPLE SURVEYS; SIMULATED ANNEALING.

1. INTRODUCTION

The elections held in the State of Valencia on May 28, 1995 gave the power to the Conservatives after sixteen years of Socialist government. During most of the socialist period, the author acted as a scientific advisor to the State President, introducing Bayesian inference and decision analysis to systematically probe the State's public opinion, with the stated aim of improving the democratic system, by closely taking into account the peoples' beliefs and preferences. This paper summarizes the methods used —always within the Bayesian framework— and illustrates their behaviour with real data.

Section 2 briefly describes the electoral setup, which allows a very detailed knowledge of the electoral results —at the level of polling stations,— and which uses Jefferson-d'Hondt

José M. Bernardo is Professor of Statistics at the University of Valencia. Research was partially funded with grant PB93-1204 of the DGICYT, Madrid, Spain.

algorithm for seat allocation. Section 3 focuses on data selection, describing the use of a simulated annealing algorithm in order to find a good design for sample surveys, which is based on the identification of a small subset of electoral sections that closely duplicates the State political behaviour.

Section 4 describes the methods which we have mostly used to analyze the data obtained from sample surveys, while Section 5 specializes on election day, describing the methods used to obtain election forecasts from exit poll, representative sections, and early returns data. Special attention is given to the actual performance of the methods described in the May 95 State election.

Section 6 describes a solution to the problem of estimating the political transition matrices which identify the reallocation of the vote of each individual party between two political elections. Finally, Section 7 contains some concluding remarks and suggests areas of additional research.

2. THE ELECTORAL SYSTEM

The State of Valencia is divided into three main electoral units, or provinces, Alicante, Castellón and Valencia, each of which elects a number or seats which is roughly proportional to its population. Thus, the State Parliament consists of a single House with 89 seats, 30 of which are elected by Alicante, 22 by Castellón and 37 by Valencia. The leader of the party or coalition that has a plurality of the seats is appointed by the King to be President of the State.

The seats in each province are divided among the parties that obtain at least 5% of the vote in the State according to a corrected proportional system, usually known as the d'Hondt rule —invented by Thomas Jefferson nearly a century before Victor d'Hondt rediscovered and popularized the system— and used, with variations, in most parliamentary democracies with proportional representation systems.

Table 1. *d'Hondt table for the results of province of Valencia in 1995 State elections*

	PP	PSOE	EU	UV
1	532524	429840	166676	137277
2	266262	214920	83338	68639
3	177508	143280	55559	45759
4	133131	107460	41669	**34319**
5	106505	85968	**33335**	27455
6	88754	71640	27779	—
7	76075	61406	—	—
8	66566	53730	—	—
9	59169	47760	—	—
10	53252	42984	—	—
11	48411	39076	—	—
12	44377	**35820**	—	—
13	40963	33065	—	—
14	38037	—	—	—
15	35502	—	—	—
16	**33283**	—	—	—
17	31325	—	—	—
18	—	—	—	—

According to d'Hondt rule, to distribute n_s seats among the, say, k parties that have overcome the 5% barrier, one (i) computes the $n_s \times k$ matrix of quotients with general element

$$z_{ij} = n_j/i, \quad i = 1, \ldots, n_s, \quad j = 1, \ldots, k,$$

where n_j is the number of valid votes obtained by the jth party, (ii) selects the largest n_s elements and (iii) allocates to party j a number of seats equal to the number of these n_s largest elements found in its corresponding column. Clearly, to apply d'Hondt rule, one may equivalently use the proportion of valid votes obtained by each party, rather than the absolute number of votes.

Thus if, for example, the 37 seats that corresponds to the province of Valencia are to be distributed among the four parties PP (conservatives), PSOE (socialists), EU (communists) and UV (conservative nationalists) who have respectively obtained (May 1995 results) 532524, 429840, 166676 and 137277 votes in the province of Valencia *and* over 5% of the State votes —the remaining 46094 counted votes being distributed among parties which did not make the overall 5% barrier— one forms the matrix in Table 1 and, according to the algorithm described, associates 16 seats to PP, 12 to PSOE, 5 to EU and 4 to UV.

It may be verified that the d'Hondt rule provides a corrected proportional system that enhances the representation of the big parties to the detriment of the smaller ones, but the correction becomes smaller as the number of seats increases, so that a pragmatically perfect proportional representation may be achieved with d'Hondt rule if the number of seats is sufficiently large. Indeed, if only one seat is allocated, d'Hondt rule obviously reduces to majority rule but, as the number of seats increases, d'Hondt rule rapidly converges to proportional rule: with the results described above, a proportional representation would yield 15.56, 12.56, 4.87 and 4.01, not far from the 16, 12, 5 and 4 integer partition provided by d'Hondt rule. Note that the last, 37th seat, was allocated to the conservative PP rather than to the socialist PSOE by only an small proportion, $(33283 - 33065) * 13 = 2836$ or 0.22/%, of the 1312411 votes counted

Since seats —and hence political power— are allocated by province results, and since there are some very noticeable differences in the political behaviour of the provinces —for instance the conservative nationalists UV are barely present outside the province of Valencia— most political analysis of the State are better done at province level, aggregating the provincial forecast in a final step.

Each province is divided into a variable number of electoral sections, each containing between 500 and 2000 electors living in a tiny, often socially homogeneous, geographical area. The State of Valencia is divided into 4484 electoral sections, 1483, 588 and 2410 of which respectively correspond to the provinces of Alicante, Castellón and Valencia. Votes are counted in public at each electoral section, just after the vote is closed at 8 pm. This means that at about 9 pm someone attending the counting may telephone to the analysis center the final results from that section; these data may be used to make early predictions of the results. Since the definition of the electoral sections has remained fairly stable since democracy was restored in Spain in 1976, this also means that a huge electoral data base, which contains the results of all elections (referendums, local, state, national and european elections) at electoral section level, is publicly available. In the next section we will describe how this is used at the design stage.

3. SURVEY DESIGN

In sample surveys, one typically has to obtain a representative sample from a human population, in order to determine the proportion $\boldsymbol{\psi} \equiv \{\psi_1, \ldots, \psi_k\}$, $(\psi_j > 0, \Sigma\psi_j = 1)$ of people who favor one of a set of, say k, possible alternative answers to a question. Naturally, most surveys contain more than one question, but we may safely ignore this fact in this discussion. Typically,

the questionnaire also includes information on possible relevant *covariates*, such as sex, age, education, or political preferences.

Within the Bayesian framework, the analysis of the survey results essentially consists on the derivation of the posterior distribution of $\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_k\}$. A particular case of frequent interest is that of *election forecasting*, where the ψ_j's, $j = 1, \ldots, k$ describe the proportion of the valid vote which each of the, say, k parties will eventually obtain.

The selection of the sample has traditionally been made by the use of the so-called "random" routes, which, regrettably, are often far from random. The problem lies in the fact that there is no way to guarantee that the attitudes of the population with respect to the question posed are homogeneous relative to the design of the "random" route. Indeed, this has produced a number of historical blunders.

An obvious alternative would be to use a real random sample, *i.e.*, to obtain a random sample from the population census —which is publicly available and contains name, address, age, sex and level of education of all citizens with the right to vote— and to interview the resulting people. The problem with this approach is that it produces highly scattered samples, what typically implies a very high cost. A practical alternative would be to determine a set of geographically small units who could jointly be considered to behave like the population as a whole, and to obtain the sample by simple random sampling within those units. Since the political spectrum of a democratic society is supposed to describe its diversity, and since the results of political elections are known for the small units defined by the electoral sections, a practical implementation of this idea would be to find a small set of electoral sections whose *joint* political behaviour has historically been as similar as possible to that of the whole population, and to use those as the basis for the selection of the samples. We now describe how did we formalize this idea.

To design a survey on a province with, say, n_p electoral sections —which on election day become n_p polling stations— may be seen as a *decision problem* where the action space is the class of the 2^{n_p} possible subsets of electoral sections, and where the loss function which describes the consequences of choosing the subset $\boldsymbol{s}$ should be a measure of the *discrepancy* $l(\hat{\boldsymbol{\psi}}_s, \boldsymbol{\psi})$ between the actual proportions $\boldsymbol{\psi} \equiv \{\psi_1, \ldots, \psi_k\}$ of people which favor each of the k alternatives considered, and the estimated proportions $\hat{\boldsymbol{\psi}}_s \equiv \{\hat{\psi}_{s1}, \ldots, \hat{\psi}_{sk}\}$ which would be obtained from a survey based of random sampling from the subset $\boldsymbol{s}$. The optimal choice would be that minimizing the expected loss

$$E[l(s) \mid D] = \int_{\Psi} l(\hat{\boldsymbol{\psi}}_s, \boldsymbol{\psi})\, p(\boldsymbol{\psi} \mid D)\, d\boldsymbol{\psi}, \tag{1}$$

where D is the database of relevant available information.

Since preferences within socially important questions may safely be assumed to be closely related with political preferences, the results of previous elections may be taken a as proxy for a random sample of questions, in order to approximate by Monte Carlo the integral above.

To be more specific, we have to introduce some notation. Let $\boldsymbol{\theta}_e = \{\theta_{e1}, \ldots, \theta_{ek(e)}\}$, for $e = 1, \ldots, n_e$, be the province results on n_e elections; thus, θ_{ej} is the proportion of the valid vote obtained by party j in election e, which was disputed among $k(e)$ parties. Similarly, let $\boldsymbol{w}_{el} = \{w_{el1}, \ldots, w_{elk(e)}\}$, $e = 1, \ldots, n_e$ and, $l = 1, \ldots, n_p$ be the results of the n_e elections in each of the n_p electoral sections in which the province is divided.

Each of the 2^{n_p} possible subsets may be represented by a sequence of 0's and 1's of length n_p, so that $\boldsymbol{s} \equiv \{s_1, \ldots, s_{n_p}\}$ is the subset of electoral sections for which $s_l = 1$. Taken as a

whole, those electoral sections would produce an estimate of the provincial result for election e, which is simply given by the arithmetic average of the results obtained in them, *i.e.*,

$$\hat{\boldsymbol{\theta}}_{es} = \frac{1}{\sum_{l=1}^{n_p} s_l} \sum_{l=1}^{n_p} s_l \boldsymbol{w}_{el}. \tag{2}$$

Thus, if election preferences may be considered representative of the type of questions posed, a Monte Carlo approximation to the integral (1) is given by

$$E[l(s) \,|\, D] \simeq \frac{1}{n_e} \sum_{e=1}^{n_e} l(\hat{\boldsymbol{\theta}}_{es}, \boldsymbol{\theta}_e) \tag{3}$$

A large number of axiomatically based arguments (see *e.g.*, Good, 1952, and Bernardo, 1979) suggest that the most appropriate measure of discrepancy between probability distributions is the logarithmic divergence

$$\delta\{\hat{\boldsymbol{\theta}}_{es}, \boldsymbol{\theta}_e\} = \sum_{j=1}^{k(e)} \theta_{ej} \log \frac{\theta_{ej}}{\hat{\theta}_{sej}} \tag{4}$$

so that we have to minimize

$$\sum_{e=1}^{n_e} \sum_{j=1}^{k(e)} \theta_{ej} \log \frac{\theta_{ej}}{\hat{\theta}_{sej}} \; . \tag{5}$$

However, this is a really huge minimization problem. For instance, for the province of Alicante, the action space thus has 2^{1483} points, what absolutely forbids to compute them all. To obtain a solution, we decided to use a random optimization algorithm, known as *simulated annealing*.

Simulated annealing is an algorithm of random optimization which uses as a heuristic base the process of obtaining pure crystals (annealing), where the material is slowly cooled, giving time at each step for the atomic structure of the crystal to reach its lowest energy level at the current temperature. The method was described by Kirkpatrick, Gelatt and Vecchi (1983) and has seen some statistical applications, such as Lundy (1985) and Haines (1987).

Consider a function $f(\boldsymbol{x})$ to be *minimized* for $\boldsymbol{x} \in \boldsymbol{X}$. Starting from an origin $\boldsymbol{x}_0$ with value $f(\boldsymbol{x}_0)$ —maybe a possible first guess on where the minimum may lie—, the idea consists of computing the value $f(\boldsymbol{x}_{i+1})$ of the objective function f at a *random* point $\boldsymbol{x}_{i+1}$ at distance d of $\boldsymbol{x}_i$; one then moves to $\boldsymbol{x}_{i+1}$ with probability one if $f(\boldsymbol{x}_{i+1}) < f(\boldsymbol{x}_i)$, and with probability $\exp\{-\delta/t\}$ otherwise, where $\delta = f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_i)$, and where t is a parameter —initially set at a large value— which mimics the temperature in the physical process of crystallization.

Thus, at high temperature, *i.e.*, at the beginning of the process, it is not unlikely to move to points where the function actually increases, thus limiting the chances of getting trapped in local minima. This process is repeated until a temporary equilibrium situation is reached, where the objective value does not change for a while.

Once in temporary equilibrium, the value of t is reduced, and a new temporary equilibrium is obtained. The sequence is repeated until, for small t values, the algorithm reduces to a rapidly convergent non-random search. The method is applied to progressively smaller distances, until an acceptable precision in reached.

The optimization cycle is typically ended when the objective value does not change for a fixed number of consecutive tries. The iteration is finished when the final non-random search is concluded. The algorithm is terminated when the final search distance is smaller than the desired precision.

In order implement the simulated annealing algorithm is further necessary to define what it is understood by "distance" among sets of electoral sections. It is natural to define the class of sets which are at distance d of $\boldsymbol{s}_j$ as those which differ from $\boldsymbol{s}_j$ in precisely d electoral sections. Thus

$$d\{\boldsymbol{s}_i, \boldsymbol{s}_j\} = \sum_{l=1}^{np} ||s_{il} - s_{jl}|| \tag{6}$$

All which is left is to adjust the sequence of "temperatures" t —what we do interactively— and to choose a starting set $\boldsymbol{s}_0$ which may reasonably chosen to be that of the, say, n_0, polling stations which are closest in average to the global value, *i.e.*, those which minimize

$$\frac{1}{n_e} \sum_{i=1}^{ne} \delta\{\boldsymbol{\omega}_{el}, \boldsymbol{\theta}_e\}. \tag{7}$$

To offer an idea of the practical power of the method, we conclude this section by describing the results actually obtained in the province of Alicante.

The province has 1483 electoral sections, so we have $2^{1483} \approx 10^{446}$ possible subsets. For these these 1483 sections we used the results obtained by the four major parties —PP, PSOE, EU and UV, grouping as "others" a large group of small, nearly testimonial parties— in four consecutive elections, local (1991), State (1991), national (1993) and european (1994). Thus, with the notation above we had $n_e = 4$, $n_p = 1483$ and $k(e) = 5$. For a mixture of economical and political considerations, we wanted to use at least 20 and no more than 40 electoral sections. Thus, starting with the set $\boldsymbol{s}_0$ of the 20 sections which, averaging over these four elections, where closest to the provincial result in a logarithmic divergence set, we run the annealing algorithm with imposed boundaries at sizes 20 and 40. The final solution —which took 7 hours on a Macintosh— was a set of 25 sections whose behaviour is described in Table 2.

For each of the four elections whose data were used, the table provides the actual results in the province of Alicante —in percentages of valid votes—, the estimators obtained as the arithmetic means of the results obtained in the 25 selected sections, and their absolute differences. It may be seen that those absolute differences are all between 0.01% 0.36%. The final block in Table 2 provides the corresponding data for the May 95 State elections, which were *not* used to find the design. The corresponding absolute errors —around 0.4, with corresponding *relative* errors of about 3%— are *much smaller* than the sampling errors which correspond to the sample sizes (about 400 in each province) which were typically used. Very similar results were obtained for the other provinces.

Our sample surveys have always been implemented with *home interviews* on citizens randomly selected from the representative sections using the electoral census. Thus, we could provide the interviewers with list of the people to be interviewed which included their name, address, and the covariates sex, age and level of education. The lists contained possible replacements with people of identical covariates, thus avoiding the danger of over representing the profiles which corresponded to people who are more often home.

Table 2. *Performance of the design algorithm for the province of Alicante in the 1995 State elections*

		PP	PSOE	EU	UV
Local 91	Results	31.50	43.17	7.23	1.22
	Estimators	31.30	43.32	7.24	1.32
	Abs. Dif.	0.20	0.15	0.01	0.09
State 91	Results	33.55	45.05	7.37	1.75
	Estimators	33.36	45.05	7.33	1.74
	Abs. Dif.	0.19	0.01	0.05	0.01
National 93	Results	43.87	39.94	10.32	0.57
	Estimators	43.64	39.75	10.62	0.49
	Abs. Dif.	0.22	0.19	0.30	0.07
European 94	Results	47.62	32.38	13.53	1.43
	Estimators	47.69	32.02	13.51	1.46
	Abs. Dif.	0.07	0.36	0.02	0.03
State 95	Results	47.24	36.30	11.06	2.11
	Estimators	48.26	36.33	10.50	1.79
	Abs. Dif.	1.02	0.03	0.56	0.32

4. SURVEY DATA ANALYSIS

The structure of the questionnaires we mostly used typically consisted of a sequence of *closed* questions –where a set of possible answers is given for each question, always leaving an "other options" possibility for those who do no agree with any of the stated alternatives, and a "non-response" option for those who refuse to answer a particular question. This was followed by a number of questions on the covariates which identify the social profile of the person interviewed; these typically include items such as age, sex, level of education, mother language or area of origin.

Let us consider one of the questions included in a survey and suppose that it is posed as a set of, say, k alternatives $\{\delta_1, \ldots, \delta_k\}$ (including the "other options" possibility) among which the person interviewed has to choose one and only one. The objective is to know the proportions of people which favor each of the alternatives, both globally, and in socially or politically interesting subsets of the population —that we shall call *classes*— as defined by either geographical or social characteristics. When the possible answers are *not* incompatible and the subject is allowed to mark more than one of them, we treated the multiple answer as a *uniform distribution* of the person's preferences over the marked answers and randomly choose one of them, thus reducing the situation to one with incompatible answers.

Thus, if $\boldsymbol{x} = \{x_1, \ldots, x_v\}$ denotes the set of, say, v covariates used to define the population classes we may be interested in, the data D relevant to a particular question included in a survey may described as a matrix which contains in each row the value of the covariates and the answer to that question provided by each of the persons interviewed. Naturally, a certain proportion of the people interviewed —typically between 20% and 40%— refuse to answer some the questions; thus, if, say, n persons have actually answered and m have refused to answer a

particular question its associated $(n+m)\times(v+1)$ matrix is defined to be

$$D=\left(\frac{D_1}{D_2}\right)=\left(\begin{array}{ccc|c} x_{1,1} & \cdots & x_{1,v} & \delta_{(1)} \\ \vdots & \cdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,v} & \delta_{(n)} \\ \hline x_{n+1,1} & \cdots & x_{n+1,v} & - \\ \vdots & \cdots & \vdots & \vdots \\ x_{n+m,1} & \cdots & x_{n+m,v} & - \end{array}\right) \tag{8}$$

where x_{ij} is the value of jth covariate for the ith subject, and $\delta_{(i)}$ denotes his or her preferences among the proposed alternatives.

Our main objective is the set of posterior probabilities

$$E[\boldsymbol{\psi}\,|\,D,c]=p(\boldsymbol{\delta}\,|\,D,c)=\{p(\delta_1\,|\,D,c),\ldots,p(\delta_k\,|\,D,c)\},\quad c\in C, \tag{9}$$

which describe the probabilities that a person in class c prefers each of the k possible alternatives, for each of the classes $c\in C$ being considered. The particular class which contains all the citizens naturally provides the global results.

To compute these, we used the total probability theorem to 'extend the conversation' to include the covariates $\boldsymbol{x}=\{x_1,\ldots,x_k\}$, so that

$$p(\boldsymbol{\delta}\,|\,D,c)=\int_{\boldsymbol{X}} p(\boldsymbol{\delta}\,|\,\boldsymbol{x},D,c)\,p(\boldsymbol{x}\,|\,D,c)\,d\boldsymbol{x} \tag{10}$$

where $p(\boldsymbol{x}\,|\,D,c)$ is the predictive distribution of the covariates vector.

Usually, the joint predictive $p(\boldsymbol{x}\,|\,D,c)$ is too complicated to handle, so we introduce a *relevant function* $\boldsymbol{t}=\boldsymbol{t}(\boldsymbol{x})$ which could be thought to be *approximately sufficient* in the sense that, for all classes,

$$p(\boldsymbol{\delta}\,|\,\boldsymbol{x},D,c)\approx p(\boldsymbol{\delta}\,|\,\boldsymbol{t},D,c),\quad \boldsymbol{x}\in\boldsymbol{X} \tag{11}$$

and, thus, (10) may be rewritten as

$$p(\boldsymbol{\delta}\,|\,D,c)\approx\int_{\boldsymbol{T}} p(\boldsymbol{\delta}\,|\,\boldsymbol{t},D,c)\,p(\boldsymbol{t}\,|\,D,c)\,d\boldsymbol{t}. \tag{12}$$

We pragmatically distinguished two different situations:

1. *Known marginal predictive.* In many situations, $\boldsymbol{t}$ has only a finite number of possible values with known distribution. For instance, we have often used as values for the relevant function $\boldsymbol{t}$ the cartesian product of sex, age group (less than 35, 35–65 and over 65) and level of education (no formal education, primary, high school and university); this produces a relevant function with 2*3*4=24 possible values, whose probability distribution within the more obvious classes, the politically relevant geographical areas, is precisely known from the electoral census. In this case,

$$p(\boldsymbol{\delta}\,|\,D,c)=\sum\nolimits_j p(\boldsymbol{\delta}\,|\,\boldsymbol{t}_j,D,c)\,w_{jc},\quad \sum\nolimits_j w_{jc}=1, \tag{13}$$

where w_{jc} denotes the weight within population class c of the subset of people with relevant function value $\boldsymbol{t}_j$.

2. *Unknown marginal predictive.* If the predictive distribution of $\boldsymbol{t}$ is unknown, or too difficult to handle, we used the $n+m$ random values of $\boldsymbol{t}$ included in the data matrix to approximate by Monte Carlo the integral (12), so that

$$p(\boldsymbol{\delta} \,|\, D, c) = \frac{1}{n+m} \sum_{j=1}^{n+m} p(\boldsymbol{\delta} \,|\, \boldsymbol{t}_j, D, c). \tag{14}$$

It is important to note that, in both cases, the 'extension of the conversation' to include the covariates automatically solved the otherwise complex problem of the *non-response*. Indeed, by expressing the required posterior distributions as weighted averages of posterior distributions *conditional* to the value of the relevant function, a *correct weight* was given to the different political sectors of the population —as described by their relevant $\boldsymbol{t}$ values— whether or not this distribution is the same within the non-response group and the rest of the population.

When the marginal predictive is known, those weights were directly input in (13), and only the data contained in D_1, *i.e.*, those which correspond to the people who answered the question, are relevant. When the marginal predictive is unknown, the weighting was done through (14) and the whole data matrix D become relevant.

The unknown predictive case is an interesting example of *probabilistic classification*. Indeed, it is *as if*, for each person with relevant function $\boldsymbol{t}$ who refuses to 'vote' for one of the alternatives $\{\delta_1, \ldots, \delta_k\}$, one would distribute his or her vote as

$$\{p(\delta_1 \,|\, \boldsymbol{t}, D, c), \ldots, p(\delta_k \,|\, \boldsymbol{t}, D, c)\}, \quad \sum_{i=1}^{k} p(\delta_i \,|\, \boldsymbol{t}, D, c) = 1, \tag{15}$$

i.e., proportionally to the chance that a person, in the same class and with the same $\boldsymbol{t}$ value, would prefer each of the alternatives.

Equations (13) and (14) reformulate the original problem in terms of estimating the conditional posterior probabilities (15). But, by Bayes' theorem,

$$p(\delta_i \,|\, \boldsymbol{t}, D, c) \propto p(\boldsymbol{t} \,|\, \delta_i, D, c)\, p(\delta_i \,|\, D, c), \quad i = 1, \ldots, k. \tag{16}$$

Computable expressions for the two factors in (16) are now derived.

If, as one would expect, the $\boldsymbol{t}$'s may be considered exchangeable within each group of citizens who share the same class and the same preferences, the representation theorems (see *e.g.*, Bernardo and Smith, 1994, Chapter 4, and references therein) imply that, for each class c and preference δ_i, *there exists* a sampling model $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}_{ic})$, indexed by a parameter $\boldsymbol{\theta}_{ic}$ which is some limiting form of the observable $\boldsymbol{t}$'s, *and* a prior distribution $p(\boldsymbol{\theta}_{ic})$ such that

$$p(\boldsymbol{t} \,|\, \delta_i, D, c) = \int_{\Theta_{ic}} p(\boldsymbol{t} \,|\, \boldsymbol{\theta}_{ic})\, p(\boldsymbol{\theta}_{ic} \,|\, D)\, d\boldsymbol{\theta}_{ic} \tag{17}$$

$$p(\boldsymbol{\theta}_{ic} \,|\, D) \propto \prod_{j=1}^{n_{ic}} p(\boldsymbol{t}_j \,|\, \boldsymbol{\theta}_{ic})\, p(\boldsymbol{\theta}_{ic}), \tag{18}$$

where, n_{ic} is the number of citizens in the survey which belong to class c and prefer option δ_i.

In practice, we have mostly worked with a *finite* number of $\boldsymbol{t}$ values. In this case, for each preference δ_i and class c, one typically has

$$p(\boldsymbol{t}_j \,|\, \boldsymbol{\theta}_{ic}) = \theta_{jic} \quad \sum\nolimits_j \theta_{jic} = 1, \quad i = 1, \ldots, k, \quad c \in C \tag{19}$$

where θ_{jic} is the chance that a person in class c who prefers the ith alternative would have relevant value $\boldsymbol{t}_j$, *i.e.*, a multinomial model for each pair $\{\delta_i, c\}$.

We were always requested to produce answers which would only depend on the survey results, without using any personal information that the politicians might have, or any prior knowledge which we could elicitate from previous work, so we systematically produced reference analyses. Using the multinomial reference prior, (Berger and Bernardo, 1992)

$$\pi(\boldsymbol{\theta}_{ic}) \propto \prod_j \left\{ \theta_{jic}^{-1/2} \left(1 - \sum\nolimits_{l=1}^{j} \theta_{lic} \right)^{-1/2} \right\}, \tag{20}$$

we find

$$\pi(\boldsymbol{\theta}_{ic} \,|\, D) \propto \prod_j \left\{ \theta_{jic}^{n_{jic}} \right\} \pi(\boldsymbol{\theta}_{ic}), \tag{21}$$

$$\begin{aligned} p(\boldsymbol{t}_j \,|\, \delta_i, D, c) &= \int_{\Theta_{ic}} \theta_{jic}\, \pi(\boldsymbol{\theta}_{ic} \,|\, D)\, d\boldsymbol{\theta}_{ic} \\ &= E[\theta_{jic} \,|\, D] = \frac{n_{jic} + 0.5}{n_{ic} + 1} \end{aligned} \tag{22}$$

where n_{jic} is the number of citizens in the survey which share the relevant value $\boldsymbol{t}_j$ among those which belong to class c and prefer option δ_i. Note that the reference analysis produces a result which is *independent of the actual number of different* $\boldsymbol{t}$ *values*, an important consequence of the use of the reference prior.

The second factor in (16) is the unconditional posterior probability that a person in class c would prefer option δ_i. With no other source of information, a similar reference multinomial analysis yields

$$p(\delta_i \,|\, D, c) = \frac{n_{ic} + 0.5}{n_c + 1}, \quad i = 1, \ldots, k, \tag{23}$$

where, again, n_{ic} is the number of citizens in the survey which belong to class c and prefer option δ_i, and n_c is the number of people in the survey that belong to class c and have answered the question. Note again that the reference prior produces a result which is *independent of the number of alternatives*, k.

Substituting (22) and (23) into (16) one finally has

$$p(\delta_i \,|\, \boldsymbol{t}_j, D, c) \propto \frac{n_{jic} + 0.5}{n_{ic} + 1} \, \frac{n_{ic} + 0.5}{n_c + 1}, \tag{24}$$

which is then used in either (13) or (14) to produce the final results.

Occasionally, we have used a more sophisticated hierarchical model, by assuming that for each preference δ_i, the $\{\theta_{1ic}, \theta_{2ic}, \ldots, \}$'s, $c \in C$, *i.e.*, the parameters which correspond to the classes actually used, are a random sample from some population of classes. In practice,

Prioridades de la Generalitat

De entre los diferentes servicios públicos que gestiona la ***Generalitat Valenciana*** ¿puede decirme los que en estos momentos deberían considerarse prioritarios?

1. Sanidad (ambulatorios, hospitales, control de alimentos, . . .).
2. Seguridad Ciudadana.
3. Vivienda (oferta y precios).
4. Educación (pública o subvencionada).
5. Medio Ambiente (humos, ruidos, basuras, . . .).
6. Tiempo Libre (instalaciones deportivas, espectáculos, exposiciones, . . .).
7. Infraestructuras viarias (autobuses, ferrocarriles, . . .).
8. Transporte público (autobuses, ferrocarriles,. . .)
9. Otras

		1	2	3	4	5	Otr	Totales
Comunidad Valenciana		34.9	19.1	13.6	14.2	11.4	6.8	1545
Provincia de Alicante		34.3	21.0	14.9	15.5	9.0	5.2	380
Provincia de Castellón		36.7	17.8	10.6	14.6	12.6	7.7	386
Provincia de Valencia		34.9	18.2	13.6	13.4	12.5	7.4	779
Ciudad de Valencia		34.1	17.6	15.6	14.3	10.5	8.0	389
Resto de Valencia		35.3	18.5	12.4	12.9	13.6	7.2	390
Intención voto	Abs	33.0	21.2	18.4	13.6	8.5	5.3	255
	PP	37.8	19.1	13.7	12.7	8.6	8.0	445
	PSOE	36.4	22.9	10.6	11.0	11.6	7.6	340
	EU	33.0	14.8	12.0	18.4	17.4	4.5	164
	UV	39.4	21.2	5.3	10.0	16.2	8.0	68

Figure 1. *Partial output of the analysis of one survey question*

however we have found few instances where a hierarchical structure of this type may safely be assured.

The methods described above were written in Pascal with the output formatted as a TEX file, with all the necessary code built in. This meant that we were able to produce reports of *presentation quality* only some *minutes* after the data were introduced, with the added important advantage of eliminating the possibility of clerical errors in the preparation of the reports.

Figure 1 is part of the actual output of such a file. It describes a fraction of the analysis of what the citizens of the State of Valencia thought the main *priorities of the State Government* should be at the time when the 1995 budget was being prepared. The first row of the table gives the mean of the posterior distribution of the proportions of the people over 18 in the State who favors each of the listed alternatives, and also includes the total number of responses over which the analysis is based. The other rows contain similar information relative to some conditional distributions (area of residence and favoured political party). The software combines together in 'Others' (Otr) all options which do not reach 5%. It may be seen from the table that it is estimateat that about 34.9% of the population believes the highest priority should be given to the health services, while 19.1% believes it should be given to law and order, and 14.2% believes

it should be given to education; these estimates are based on the answers of the 1545 people who completed this question. The proportion of people who believe education should be the highest priority becomes 15.5% among the citizens of the province of Alicante, 13.6% among those who have no intention to vote, 11.0% among the socialist voters and 18.4% among the communist voters. The estimates provided were actually the means of the appropriate posterior distributions; the corresponding standard deviations were also computed, but not included in the reports in order to make those complex tables as readable as possble to politicians under stress.

Occasionally, we posed questions on a numerical scale, often the [0–10] scale used at Spanish schools. These included requests for an evaluation of the performance of a political leader, and questions on the level of agreement (0=total disagreement, 10=total agreement) with a sequence of statements designed to identify the people's values. The answers to these numerical questions were treated with the methods described above to produce probability distributions over the eleven $\{0, 1, \ldots, 10\}$ possible values. These distributions were graphically reported as histograms, together with their expected values. For instance, within the city of Valencia in late 1994, the statement "My children will have a better life than I" got an average level of agreement of 7.0, while "Sex is one of the more important things in life" got 5.0, "Spain should have never joined the European union" 3.2, and "Man should not enter the kitchen or look after the kids" only 2.0.

5. ELECTION NIGHT FORECASTING

On election days, we systematically produced several hours of evolving information. In this section we summarize the methods we used, and illustrate them with the results obtained at the May 28th, 1995 State election; the procedures used in other elections have been very similar.

Some weeks before any election we used the methodology described in Section 3 to obtain a set of representative electoral sections for each of the areas we wanted to produce specific results. In the May 95 election, a total of 100 sections were selected, in four groups of 25, respectively reproducing the political behaviour of the provinces of Alicante and Castellón, the city of Valencia, and the rest of the province of Valencia; these are the representative sections we will be referring to.

5.1. *The exit poll*

An exit poll was conducted from the moment the polling stations opened at 9 am. People were approached in their way out from the 100 representative polling stations. Interviewers handed simple forms to as many people as possible, where they were asked to mark by themselves their vote and a few covariates (sex, age, level of education, and vote in the previous election), and to introduce the completed forms in portable urns held by the interviewers.

Mobile supervisors collected the completed forms, each cycling through a few stations, and phoned their contents to the analysis center. Those answers (seven digits per person including the code to identify the polling station) where typed in, and a dedicated programme automatically updated the relevant sufficient statistics every few minutes.

The analysis was an extension of that described in Section 4. Each electoral section s was considered a class, and an estimation of the proportion of votes,

$$\{p(\delta_1 \mid D, s), \ldots, p(\delta_1 \mid D, s)\}, \quad s \in S, \tag{25}$$

that each of the parties $\delta_1, \ldots, \delta_k$ could expect in that section, given the relevant data D, was obtained by extending the conversation to include sex and age group, and using (13) rather than

(14), since the proportions of people within each sex and age group combination was known from the electoral census for all sections.

We had repeatdly observed that the logit transformations or the proportions are better behaved than the proportions themselves. A normal hierarchical model on the logit transformations of the section estimates was then used to integrate the results from all the sections in each province. Specifically, the logit transformations of the collection of k-variate vectors (25) where treated as a random sample from some k-variate normal distribution with an unknown mean vector $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_k\}$ —which identify the logit transformation of the global results in the province— and were used to obtain the corresponding reference posterior distribution for $\boldsymbol{\mu}$, *i.e.*, the usual k-variate Student t (see *e.g.*, Bernardo and Smith, 1994, p. 441).

Monte Carlo integration was then used to obtain the corresponding probability distribution over the seat allocation in the province. This was done by simulating 2,000 observations from the posterior distribution of $\boldsymbol{\mu}$, using d'Hondt rule to obtain for each of those the corresponding seat allocation, and counting the results to obtain a probability distribution over the possible seat allocations and the corresponding marginal distributions on the number of seats which each party may expect to obtain in the province. The simulations from the three provinces were finally integrated to produce a forecast at State level.

The performance achieved by this type of forecast in practice is summarized in the first block of Table 3.

5.2. *The representative sections forecast*

By the time the polls closed (8 pm) the results of the exit poll could be made public. The interviewers located at the selected representative stations were then instructed to attend the scrutiny and to phone twice to the analysis center. They first transmitted the result of the first 200 counted votes, and then the final result.

The analysis of these data is much simpler than that of those from the exit poll. Indeed, we do not have here any covariates, nor any need for them, for these data do not have any non-response problems.

The results from each representative section were treated as a random sample from a multinomial model with a parameter vector describing the vote distribution within that section. Again, a hierarchical argument was invoked to treat the logit transformation of those parameters as a normal random sample centered in the logit transformation of a parameter vector describing the vote distribution in the province.

Numerical integration was then used to produce the reference posterior distribution of the province vote distribution and the implied reference posterior distribution on the seat allocation within that province. The simulations from the three provinces were then combined to produce a global forecast.

In the many elections we have tried, the technique just described produced very accurate forecasts of the final results about one hour after the stations closed. Figure 2 is a reproduction of the actual forecast made at 22h52 of May 28th, 1995, which was based on the 94 representative stations (from a total of 100) that had been received before we switched to the model which used the final returns.

Elecciones Autonómicas 1995
Comunidad Valenciana

Datos históricos relevantes

Autonómicas 1991	PP	PSOE	EU	UV	UPV	Otr
% votos	28.1	43.2	7.6	10.4	3.7	7.1
Escaños (89)	31	45	6	7	0	0

Datos procedentes del escrutinio de 94 mesas escogidas
Proyección a las 22 horas 52 min

	PP	PSOE	EU	UV	UPV	Otr
% votos válidos	**43.0**	**33.4**	**12.4**	**7.2**	2.8	1.1
Desviaciones	0.8	0.8	0.9	0.4	0.8	0.3
Escaños (89)	**42**	**32**	**10**	**5**	0	0
0.20	42	32	10	5	0	0
0.13	42	31	11	5	0	0
0.11	41	32	11	5	0	0
0.09	41	33	10	5	0	0
0.08	43	31	10	5	0	0
0.08	42	33	9	5	0	0
0.07	43	32	9	5	0	0
0.03	41	31	12	5	0	0
0.03	40	33	11	5	0	0
0.02	41	34	9	5	0	0

Distribución de diputados por partidos

PP	40	**41**	**42**	**43**	44
	0.05	0.28	0.46	0.20	0.02
PSOE	30	**31**	**32**	**33**	34
	0.03	0.26	0.42	0.24	0.04
EU	8	**9**	**10**	**11**	12
	0.03	0.18	0.42	0.30	0.06
UV	4	**5**	6		
	0.06	0.94	0.01		

Figure 2. *Actual forecast on election night, 1995*

5.3. *The early returns forecast*

By 11 pm, the return from the electoral sections which have been more efficient at the scrutiny started to come in through a modem line connected to the main computer where the official results were being accumulated. Unfortunately, one could not treat the available results as

a random sample from all electoral sections; indeed, returns from small rural communities typically come in early, with a vote distribution which is far removed from the overall vote distribution.

Naturally, we expected a certain geographical consistency among elections in the sense that areas with, say, a proportionally high socialist vote in the last election will still have a proportionally high socialist vote in the present election. Since the results of the past election were available for each electoral section, each incoming result could be compared with the corresponding result in the past election in order to learn about the direction and magnitude of the swing for each party. Combining the results already known with a prediction of those yet to come, based on an estimation of the swings, we could hope to produce accurate forecasts of the final results.

Let r_{ij} be the proportion of the valid vote which was obtained in the last election by party i in electoral section j of a given province. Here, $i = 1, \ldots, k$, where k is the number of parties considered in the analysis, and $j = 1, \ldots, N$, where N is the number of electoral sections in the province. For convenience, let $\boldsymbol{r}$ generically denote the k-dimensional vector which contains the past results of a given electoral section. Similarly, let y_{ij} be the proportion of the valid vote which party i obtains in the present election in electoral section j of the province under study. As before, let $\boldsymbol{y}$ generically denote the k-dimensional vector which contains the incoming results of a given electoral section.

At any given moment, only some of the $\boldsymbol{y}$'s, say $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, $0 \leq n \leq N$, will be known. An estimate of the final distribution of the vote $\boldsymbol{z} = \{z_1, \ldots, z_k\}$ will be given by

$$\hat{\boldsymbol{z}} = \sum_{j=1}^{n} \omega_j \boldsymbol{y}_j + \sum_{j=n+1}^{N} \omega_j \hat{\boldsymbol{y}}_j, \qquad \sum_{j=1}^{N} \omega_j = 1, \tag{26}$$

where ω_j is the relative number of voters in the electoral section j, known from the census, and the $\hat{\boldsymbol{y}}_j$'s are estimates of the $N - n$ unobserved $\boldsymbol{y}$'s, to be obtained from the n observed results.

The analysis of previous election results showed that the logit transformations of the proportion of the votes in consecutive elections were roughly linearly related. Moreover, within the province, one may expect a related political behaviour, so that it seems plausible to assume that the corresponding residuals should be exchangeable. Thus, we assumed

$$\begin{aligned} \log\left\{\frac{y_{ij}}{1-y_{ij}}\right\} &= \alpha_i \log\left\{\frac{r_{ij}}{1-r_{ij}}\right\} + \beta_i + \varepsilon_{ij}, \quad j = i, \ldots, k, \quad j = 1, \ldots, n, \\ p(\varepsilon_{ij}) &= \mathrm{N}(\varepsilon_{ij} \,|\, 0, \sigma_i) \end{aligned} \tag{27}$$

and obtained the corresponding reference predictive distribution for the logit transformation of the $\boldsymbol{y}_{ij}$'s (Bernardo and Smith, 1994, p. 442) and hence, a reference predictive for $\boldsymbol{z}$.

Again, numerical integration was used to obtain the corresponding predictive distribution for the seat allocation in the province implied by the d'Hondt algorithm, and the simulations for the three provinces combined to obtain a forecast for the State Parliament.

The performance of this model in practice, summarized in the last two blocks of Table 3, is nearly as good as the considerably more complex model developed by Bernardo and Girón (1992), first tested in the 1991 State elections.

Table 3. *Vote distribution and seat allocation forecasts on election day 1995*

Parties	PP	PSOE	EU	UV	
Exit poll	44.0±1.3	30.9±1.2	12.6±0.7	6.1±1.1	
(14h29)	45	30	10	4	$p = 0.05$
Representative sections	43.0±0.8	33.4±0.8	12.4±0.9	7.2±0.4	
(22h52)	42	32	10	5	$p = 0.20$
First 77% scrutinized	43.80±0.40	34.21±0.20	11.74±0.04	6.77±0.04	
(23h58)	42	32	10	5	$p = 0.45$
First 91% scrutinized	43.47±0.32	34.28±0.17	11.69±0.02	6.96±0.03	
(00h53)	42	32	10	5	$p = 1.00$
Final	**43.3**	**34.2**	**11.6**	**7.0**	
	42	**32**	**10**	**5**	

5.4. *The communication of the results*

All the algorithms were programmed in Pascal with the output formatted as a TEX file which also included information on past relevant data to make easier its political analysis. A macro defined on a Macintosh chained the different programmes involved to capture the available data, perform the analysis, typeset the corresponding TEX file, print the output on a laser printer and fax a copy to the relevant authorities. The whole procedure needed about 12 minutes.

Table 3 summarizes the results obtained on May 95 election with the methods described. The timing was about one hour later than usual, because the counting for the local elections held on the same day was done before the counting for the State elections. Fore several forecasts, we reproduce the means and standard deviations of the posterior distribution of the percentages of valid vote at State level, and the mode and associated probability of the corresponding posterior distribution of the seat allocation. These include an exit poll forecast (at 14h29, with 5,683 answers introduced), a forecast based on the final results of the 94 representative sections received at 22h52 (when six of them were still missing), and two forecasts respectively based on the first 77% (reached at 23h58) and the first 91% (reached at 00h53) scrutinized stations. The final block of the table reproduces, for comparison, the official final results.

The analysis of Table 3 shows the progressive convergence of the forecasts to the final results. Pragmatically, the important qualitative outcome of the election, namely the conservative victory, was obvious from the very first forecast, in the early afternoon (when only about 60% of the people had actually voted!), but nothing precise could then be said about the actual seat distribution. The final seat allocation was already the mode of its posterior distribution with the forecast made with representative stations, but its probability was then only 0.20. That probability was 0.45 at midnight (with 77% of the results) and 1.00, to two decimal places, at 1 am (with 91%), about three hours before the scrutiny was actually finished (the scrutiny typically takes several hours to be actually completed because of bureaucratic problems always appearing at one place or another).

46250 Valencia

Partidos	1991		1995	
Censo	586 719	100.00	619 302	100.00
Abst.	218 112	37.17	169 588	27.38
psoe	134 943	36.60	116 034	25.80
pp	101 414	27.51	212 878	47.33
eu	30 761	8.34	64 469	14.33
uv	72 055	19.54	47 492	10.56
upv	8 568	2.32	4 610	1.02
cds	7 881	2.13	1 597	0.35
env	0	0.00	372	0.08
pra	0	0.00	457	0.10
pcpe	0	0.00	512	0.11
au	0	0.00	128	0.02
fe-jons	0	0.00	293	0.06
pie	0	0.00	331	0.07
lae	0	0.00	194	0.04
r v	0	0.00	0	0.00
ph	0	0.00	347	0.07

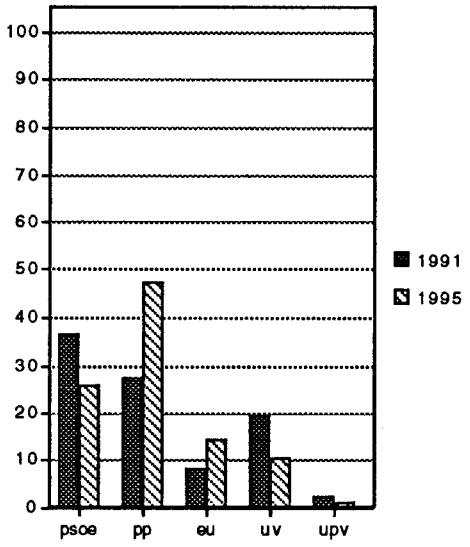


Figure 3. *Reproduction of the city of Valencia output from the 1995 election book*

By about 4 am, all the results were in, and have been automatically introduced into a relational data base (*4th Dimension* ™) which already contained the results from past elections. An script had been programmed to produce, format, and print, a graphical display of the elections results for each of the 442 counties in the State, including for comparison the results form the last, 1991, State election. Figure 3 reproduces the output which corresponds to the city of Valencia. Besides, the results where automatically aggregated to produce similar outputs for each of the 34 geographical regions of the State, for the 3 provinces, and for the State as a whole.

While this was being printed, a program in *Mathematica*™, using digital cartography of the State, produced colour maps where the geographical distribution of the vote was vividly described. Figure 4 is a black and white reproduction of a colour map of the province of Alicante, where each county is coded as a function the two parties coming first and second in the election. Meanwhile, the author prepared a short, introductory analysis to the election results.

Thus, at about 9 am, we had a camera-ready copy of a commercial quality, 397 pages book which, together with a diskette containing the detailed results, was printed, bounded and, *distributed* 24 hours later to the media and the relevant authorities, and immediately available to the public at bookshops.

6. THE DAY AFTER

After the elections have been held, both the media and the politicians' discussions often center on the *transition probabilities* $\Phi = \{\varphi_{ij}\}$ where

$$\varphi_{ij} = \Pr\{\text{a person has just voted for party } i \mid \text{ he (she) voted for party } j\}, \qquad (28)$$

which describe the reallocation of the vote of each individual party between the present and the past election.

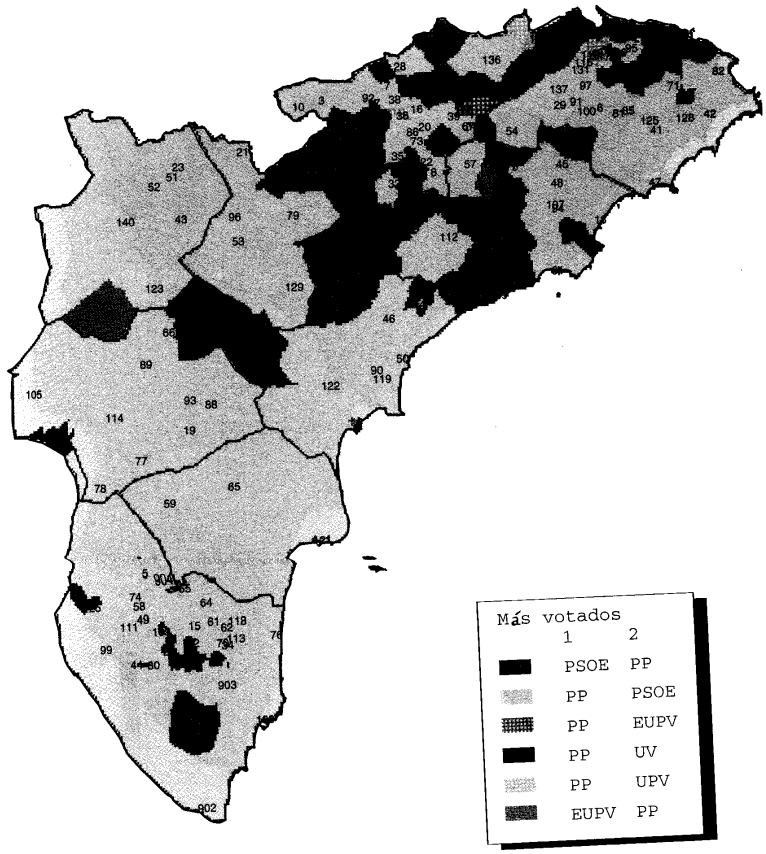


Figure 4. *Reproduction of a page on electoral geography from the 1995 election book*

Let N be the number of people included in either of the two electoral censuses involved. It is desired to analyze the aggregate behaviour of *all* those people, including those who never voted or only voted in one of the two elections. Let $\boldsymbol{p} = \{p_1, \ldots, p_k\}$ describe the distribution of the behaviour of the people in the present election; thus, p_j is the proportion of those N people who have just voted for party j, and p_k is the proportion of those N people who did not vote, either because they decide to abstain or because they could not vote for whatever reason (business trip, illness, or whatever), including those who died between the two elections. Similarly, let $\boldsymbol{q} = \{q_1, \ldots, q_m\}$ be the distribution of the people's behaviour in the previous election, including as specific categories not only what people voted for, if they did, but also whether they abstained in that election, or whether they were under 18 (and, hence, could not vote) at the time that election was held.

Obviously, by the total probability theorem, the transition matrix Φ has to satisfy

$$p_i = \sum_{j=1}^{m} \varphi_{ij}\, q_j, \quad i = 1, \ldots, k. \tag{29}$$

A "global" estimation $\hat{\Phi}$ of the transition matrix Φ is most useful if it successfully "explains" the transference of vote in *each* politically interesting area, *i.e.*, if for each of these areas l,

$$p_{il} \simeq \sum_{j=1}^{m} \hat{\varphi_{ij}}\, q_{jl}, \quad j = 1, \ldots, m. \tag{30}$$

The exit poll had provided us with a politically representative sample of the entire population of, say, size n, for which

$$\begin{aligned} \boldsymbol{x} &= \{\text{NewVote, PastVote, Class}\} \\ \text{Class} &= \{\text{Sex, AgeGroup, Education}\} \end{aligned} \tag{31}$$

had been recorded, where Class is a discrete variable whose distribution in the population, say $p(c)$, is precisely known from the census.

For each pair $\{\text{PastVote} = j, \text{Class} = c\}$, the $\boldsymbol{x}$'s provide a multinomial random sample with parameters $\{\varphi_{1jc}, \ldots, \varphi_{kjc}, \}$ where φ_{ijc} is the probability that a person in class c had just voted party i, if he (she) voted j in the past election. The corresponding reference prior is

$$\pi(\boldsymbol{\varphi}_{jc}) \propto \prod_{i=1}^{k} \left\{ \varphi_{ijc}^{-1/2} \left(1 - \textstyle\sum_{l=1}^{i} \varphi_{ljc} \right)^{-1/2} \right\}. \tag{32}$$

Hence, for each pair (j, c) we obtain the modified Dirichlet reference posterior distribution

$$\pi(\boldsymbol{\varphi}_{jc} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \propto \pi(\boldsymbol{\varphi}_{jc}) \prod_{i=1}^{k} \varphi_{ijc}^{n_{ijc}}, \tag{33}$$

where n_{ijc} is the number of citizens of type c in the exit poll survey who declared that have just voted i and that had voted j in the past election. The *global* posteriors for the transition probabilities $\{\pi(\varphi_{1j}, \ldots, \varphi_{kj}), j = 1, \ldots, m$ are then

$$\pi(\boldsymbol{\varphi}_j \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} = \sum_{c} \pi(\boldsymbol{\varphi}_{jc} \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) p(c), \tag{34}$$

where the $p(c)$'s are known from the census. The mean, standard deviation, and any other interesting functions of the transition probabilities φ_{ij}, may easily be obtained by simulation.

Equation (34) encapsulates the information about the transition probabilities provided by the exit poll data but, once the new results $p_1, \ldots, p_k$ are known, equation (29) has to be *exactly* satisfied. However, the (continuous) posterior distribution of the φ_{ij}'s cannot be updated using Bayes theorem, for this set of restrictions constitute an event of zero measure.

Deming and Stephan proposed in the forties an iterative adjustment of sampled frequency tables when expected marginal totals are known, which preserves the association structure and matches the marginal constraints; this is further analyzed in Bishop, Fienberg and Holland (1975). With a simulation technique, we may repeatedly use this algorithm to obtain a posterior sample of φ_{ij}'s which satisfy the conditions. Specifically, to obtain a simulated sample from each of the m conditional posterior distributions of the transition probabilities given the final results,

$$\pi(\boldsymbol{\varphi}_j \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{p}_j), \quad j = 1, \ldots, m, \tag{35}$$

	EU	PP	PSOE	...	Abs	Totales
EU	82327	11189	11796	...	41300	151242
	54.4	7.4	7.8	...	27.3	100.0
PP	2744	422648	8215	...	118082	558617
	0.5	75.7	1.5	...	21.1	100.0
PSOE	32735	85758	531739	...	192087	860429
	3.8	10.0	61.8	...	22.3	100.0
UV	7304	44056	6130	...	57728	208126
	3.5	21.2	2.9	...	27.7	100.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Menores	10073	27046	15089	...	18314	75174
	13.4	36.0	20.1	...	24.4	100.0
Totales	**271606**	**1010702**	**798537**	...	**762419**	**3093574**
	8.8	32.7	25.8	...	**24.6**	100.0
	11.7	**43.4**	**34.3**	...	100.0	—

Figure 5. *Part of the transition matrix between the 1991 and the 1995 Valencia State Elections*

we (i) simulated from the unrestricted conditional posteriors a set of φ_{ij}'s, (ii) derived the corresponding joint probabilities $t_{ij} = \varphi_{ij}\, q_j$; (iii) applied the iterative algorithm to obtain a estimate $\hat{t}_{ij}$ which agrees with the marginals $\boldsymbol{p}$ and $\boldsymbol{q}$ and (iv) retransformed into the conditional probabilities $\hat{\varphi}_{ij} = \hat{t}_{ij}/q_j$.

The posterior mean, standard deviation, and any other interesting functions of the transition probabilities φ_{ij}, given the final electoral results $\boldsymbol{p}$, were then easily obtained from this simulated sample. Finally, we used the final estimates of the transition probabilities to derive estimates of the absolute vote transfers, obviously given by $v_{ij} = N\, \hat{\varphi}_{ij}\, q_j$, where N is the total population of the area analyzed.

Figure 5 reproduces some of the means of the posterior distribution of the transition probabilities between the 1991 and the 1995 elections in the State of Valencia, which were obtained with the methods just described. For instance, we estimated that the socialist PSOE retained 61.8% of its 1991 vote, and lost 10.0% (85,758 votes) to the conservative PP, and 22.3% (192,087) votes in people who did not vote.

7. FINAL REMARKS

Due to space limitations and to the nature of this meeting, we have concentrated on the methods we have mostly used in *practice*. Those have continually evolved since our first efforts at the national elections of 1982, described in Bernardo (1984). A number of interesting research issues have appeared however in connection with this work. A recent example (Bernardo, 1994) is the investigation of the optimal hierarchical strategy which could be used to predict election results based on early returns; this naturally leads to *Bayesian clustering* algorithms where, as one would expect from any Bayesian analysis, clearly specified preferences define the algorithm, thus avoiding the 'adhockeries' which plague standard cluster analysis.

ACKNOWLEDGEMENTS

In many senses, the work described in this paper has been joint with a team of people working under the author's supervision at *Presidència de la Generalitat*, the office of the State President. Explicit mention is at least required to Rafael Bellver and Rosa López, who respectively supervised the field work and the hardware setup and, —most specially— to Javier Muñoz, who did important parts of the necessary programming.

REFERENCES

Berger, J. O. and Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**, 25–37.

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 636–690.

Bernardo, J. M. (1984). Monitoring the 1982 Spanish Socialist victory: a Bayesian analysis. *J. Amer. Statist. Assoc.* **79**, 510-515.

Bernardo, J. M. (1994). Optimizing prediction with hierarchical models: Bayesian clustering. *Aspects of Uncertainty: a Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.). Chichester: Wiley, 67–76.

Bernardo, J. M. and Girón, F. J. (1992). Robust sequential prediction form non-random samples: the election night forecasting case. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 61–77, (with discussion).

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107–114.

Haines, L. M. (1987). The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. *Technometrics* **29**, 439–447.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–680.

Lundy, M. (1985). Applications of the annealing algorithm to combinatorial problems in statistics. *Biometrika* **72**, 191–198.

TRIBUNA

Ley d"Hondt y elecciones catalanas

José Miguel Bernardo es catedrático de Estadística de la Universidad de Valencia.
EL PAÍS - C. Valenciana - 02-11-1999

Los apretados resultados de las recientes elecciones catalanas en las que una formación política (CiU) ha obtenido el mayor número de escaños (56 con el 37,68% de los votos) a pesar de ser superada en votos por otra formación política (PCS-CpC, 52 escaños con el 37,88% de los votos) ha despertado de nuevo la polémica sobre la utilización de la Ley d"Hondt en nuestras leyes electorales y, una vez más, se ha repetido la afirmación según la cual la Ley d"Hondt distorsiona la voluntad popular expresada por los porcentajes de votos obtenidos por cada formación política.En esta situación es necesario reiterar que tal afirmación es manifiestamente errónea. En un sistema de representación parlamentria es obviamente necesario asignar un número entero de escaños a cada formación política, y la Ley d"Hondt es el mejor algoritmo conocido para repartir el número total de escaños que forman el Parlamento de manera que cada formación política reciba un número entero de escaños aproximadamente proporcional al porcentaje de votos válidos que ha obtenido.

Como es sabido, el algoritmo de Jefferson-d"Hondt (propuesto por Thomas Jefferson casi un siglo antes de que Victor d"Hondt la redescubriese y popularizase), consigue esta aproximación entera mediante el uso de cocientes sucesivos. Específicamente, si se trata de un Parlamento con N escaños disputados por p formaciones políticas que han obtenido (n1, ...np) votos, se calcula la matriz de cocientes rij=ni/j, j=1, ... N, se seleccionan sus N mayores elementos, y se asigna a cada formación política un número de escaños igual al número de esos N elementos que corresponden a sus propios cocientes. En este algoritmo no hay mecanismo distorsionador alguno, más allá de la aproximación necesaria para poder encontrar una partición entera.

Nuestra ley electoral distorsiona efectivamente la voluntad popular, pero esto no es debido al uso del algoritmo de Jefferson-d"Hondt, sino al empleo de las provincias como circunscripciones electorales y, en menor medida, al requisito de un porcentaje mínimo de votos válidos para obtener representación parlamentaria. Cuanto menores sean las circunscripciones electorales, mayor será la ventaja relativa de los partidos grandes frente a los pequeños, cualquiera que sea el algoritmo de asignación empleado. En un extremo, si cada circunscripción electoral elige un único diputado (como actualmente sucede en el Reino Unido), se tiene un sistema de representación mayoritario. En el otro extremo, si se utiliza una circunscripción única (como se hace en las elecciones europeas, en las que toda España es una circunscripción electoral) se obtiene una representación parlamentaria lo más proxima posible a una representación proporcional perfecta.

Generalmente, las leyes electorales exigen un porcentaje mínimo de votos válidos para acceder a la representación parlamentaria (en España es el 3% para las elecciones generales y para la mayor parte de las autonómicas, pero sólo el 1% para las elecciones europeas). Naturalmente, este requisito constituye otro elemento distorsionador de la pluralidad política expresada por los resultados electorales, tanto mayor cuanto mayor sea el porcentaje exigido (en las elecciones autonómicas valencianas se sitúa en un injustificable 5%).

Unos sencillos ejercicios aritméticos con los resultados provisionales de las recientes elecciones catalanas permiten apreciar las consecuencias políticas de los efectos distorsionadores mencionados.

Las dos primeras columnas de la Tabla 1 describen (en número de votos y en porcentajes) los resultados globales de las elecciones en el conjunto de Cataluña. Con la ley electoral vigente (circunscripciones electorales provinciales y mínimo del 3%), la asignación de escaños da lugar a la columna I, en la que CiU, con 56 escaños, alcanza el mayor número de diputados. El uso de toda Cataluña como circunscripción única, manteniendo el requisito del 3%, da lugar a la columna II, en la que el empate técnico entre CiU y PSC-CpC se traduce en 55 escaños cada uno. El uso de toda Cataluña como circunscripción única, pero con requisito mínimo de sólo el 1% da lugar a la columna III, en la que la ligera ventaja en votos del PSC-CpC sobre CiU se traduce en un escaño más para la formación socialista. Este mismo resultado es el que se obtiene con estos datos si no se exige requisito mínimo alguno.

En la Tabla 2 se reproducen los porcentajes de escaños a que corresponden. Como podía esperarse, solamente la tercera opción representa una aproximación no distorsionada de los resultados electorales. En particular, CiU, con el 37,68% de los votos hubiera obtenido el 38,52% de los escaños (y no el 41,48% que le otorga la ley electoral vigente) mientras el PSC-CpC con un 37,88% de los votos hubiera obtenido el 39,26% de los escaños (y no el 38,52% que otorga la ley electoral vigente); EU, con el 1,43% de los votos hubiera obtenido el 1,48% de los escaños (en lugar de quedar sin representación parlamentaria). La lista más votada, el PSC-CpC habría tenido también la mayor representación parlamentaria y habría sido requerida para formar Gobierno.

To appear at the *International Statistical Review*
Typesetted on September 27, 2002, at 11h 5

Bayesian Hypothesis Testing: A Reference Approach

José M. Bernardo[1] **and Raúl Rueda**[2]

[1]*Dep. d'Estadística e IO, Universitat de València, 46100-Burjassot, Valencia, Spain.*
E-mail: jose.m.bernardo@uv.es

[2]*IIMAS, UNAM, Apartado Postal 20-726, 01000 Mexico DF, Mexico.*
E-mail: pinky@sigma.iimas.unam.mx

Summary

For any probability model $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ assumed to describe the probabilistic behaviour of data $x \in X$, it is argued that testing whether or not the available data are *compatible* with the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is best considered as a formal decision problem on whether to use (a_0), or not to use (a_1), the simpler probability model (or *null* model) $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, where the loss difference $L(a_0, \boldsymbol{\theta}, \boldsymbol{\omega}) - L(a_1, \boldsymbol{\theta}, \boldsymbol{\omega})$ is proportional to the amount of information $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ which would be lost if the simplified model M_0 were used as a proxy for the assumed model M. For any prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$, the appropriate normative solution is obtained by rejecting the null model M_0 whenever the corresponding posterior expectation $\int \int \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) \, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x}) \, d\boldsymbol{\theta} \, d\boldsymbol{\omega}$ is sufficiently large.

Specification of a subjective prior is always difficult, and often polemical, in scientific communication. Information theory may be used to specify a prior, the *reference* prior, which only depends on the assumed model M, and mathematically describes a situation where no prior information is available about the quantity of interest. The reference posterior expectation, $d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \int \delta \, \pi(\delta \mid \boldsymbol{x}) \, d\delta$, of the amount of information $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ which could be lost if the null model were used, provides an attractive non-negative test function, the *intrinsic statistic*, which is invariant under reparametrization.

The intrinsic statistic $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is measured in units of information, and it is easily calibrated (for any sample size and any dimensionality) in terms of some average log-likelihood ratios. The corresponding Bayes decision rule, the *Bayesian reference criterion (BRC)*, indicates that the null model M_0 should only be rejected if the posterior expected loss of information from using the simplified model M_0 is too large or, equivalently, if the associated expected average log-likelihood ratio is large enough.

The BRC criterion provides a general reference Bayesian solution to hypothesis testing which does not assume a probability mass concentrated on M_0 and, hence, it is immune to Lindley's paradox. The theory is illustrated within the context of multivariate normal data, where it is shown to avoid Rao's paradox on the inconsistency between univariate and multivariate frequentist hypothesis testing.

Keywords: Amount of Information; Decision Theory; Lindley's Paradox; Loss function; Model Criticism; Model Choice; Precise Hypothesis Testing; Rao's Paradox; Reference Analysis; Reference Prior.

1. Introduction

1.1. Model Choice and Hypothesis Testing

Hypothesis testing has been subject to polemic since its early formulation by Neyman and Pearson in the 1930s. This is mainly due to the fact that its standard formulation often constitutes a serious oversimplification of the problem intended to solve. Indeed, many of the problems which traditionally have been formulated in terms of hypothesis testing are really complex decision problems on *model choice*, whose appropriate solution naturally depends on the structure of the problem. Some of these important structural elements are the motivation to choose a particular model (*e.g.*, simplification or prediction), the class of models considered (say a finite set of alternatives or a class of nested models), and the available prior information (say a sharp prior concentrated on a particular model or a relatively diffuse prior).

In the vast literature of model choice, reference is often made to the "true" probability model. Assuming the existence of a "true" model would be appropriate whenever one knew for sure that the real world mechanism which has generated the available data was one of a specified class. This would indeed be the case if data had been generated by computer simulation, but beyond such controlled situations it is difficult to accept the existence of a "true" model in a literal sense. There are many situations however where one is prepared to proceed "as if" such a true model existed, and furthermore belonged to some specified class of models. Naturally, any further conclusions will then be conditional on this (often strong) assumption being reasonable in the situation considered.

The natural mathematical framework for a systematic treatment of model choice is decision theory. One has to specify the range of models which one is willing to consider, to decide whether or not it may be assumed that this range includes the true model, to specify probability distributions describing prior information on all unknown elements in the problem, and to specify a loss function measuring the eventual consequences of each model choice. The best alternative within the range of models considered is then that model which minimizes the corresponding expected posterior loss. Bernardo and Smith (1994, Ch. 6) provide a detailed description of many of these options. In this paper attention focuses on one of the simplest problems of model choice, namely *hypothesis testing*, where a (typically large) model M is tentatively accepted, and it is desired to test whether or not available data are *compatible* with a particular submodel M_0. Note that this formulation includes most of the problems traditionally considered under the heading of hypothesis testing in the frequentist statistical literature.

1.2. Notation

It is assumed that probability distributions may be described through their probability mass or probability density functions, and no distinction is generally made between a random quantity and the particular values that it may take. Roman fonts are used for *observable* random quantities (typically data) and for known constants, while Greek fonts are used for *unobservable* random quantities (typically parameters). Bold face is used to denote row vectors, and $\boldsymbol{x}'$ to denote the transpose of the vector $\boldsymbol{x}$. Lower case is used for variables and upper case for their domains. The standard mathematical convention of referring to *functions*, say f and g of $\boldsymbol{x} \in X$, respectively, by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, will often be used. In particular, $p(\boldsymbol{x} \,|\, C)$ and $p(\boldsymbol{y} \,|\, C)$ will respectively represent general *probability densities* of the *observable* random vectors $\boldsymbol{x} \in X$ and $\boldsymbol{y} \in Y$ under conditions C, without any suggestion that the random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ have the same distribution. Similarly, $\pi(\boldsymbol{\theta} \,|\, C)$ and $\pi(\boldsymbol{\omega} \,|\, C)$ will respectively represent general probability densities of the *unobservable* parameter vectors $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\omega} \in \Omega$ under conditions C. Thus, $p(\boldsymbol{x} \,|\, C) \geq 0$, $\int_X p(\boldsymbol{x} \,|\, C)\, d\boldsymbol{x} = 1$, and $\pi(\boldsymbol{\theta} \,|\, C) \geq 0$, $\int_\Theta \pi(\boldsymbol{\theta} \,|\, C)\, d\boldsymbol{\theta} = 1$. If the random

vectors are discrete, these functions are probability mass functions, and integrals over their values become sums. $\mathrm{E}[\boldsymbol{x} \mid C]$ and $\mathrm{E}[\boldsymbol{\theta} \mid C]$ are respectively used to denote the expected values of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ under conditions C. Finally, $\Pr(\boldsymbol{\theta} \in A \mid \boldsymbol{x}, C) = \int_A p(\boldsymbol{\theta} \mid \boldsymbol{x}, C)\, d\boldsymbol{\theta}$ denotes the probability that the parameter $\boldsymbol{\theta}$ belongs to A, given data $\boldsymbol{x}$ and conditions C.

Specific density functions are denoted by appropriate names. Thus, if x is a univariate random quantity having a Normal distribution with mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x \mid \mu, \sigma^2)$; if θ has a Beta distribution with parameters a and b, its density function will be denoted $\mathrm{Be}(\theta \mid a, b)$.

A *probability model* for some data $\boldsymbol{x} \in X$ is defined as a *family* of probability distributions for $\boldsymbol{x}$ indexed by some *parameter*. Whenever a model has to be fully specified, the notation $\{p(\boldsymbol{x} \mid \boldsymbol{\phi}),\, \boldsymbol{\phi} \in \Phi,\, \boldsymbol{x} \in X\}$ is used, and it is assumed that $p(\boldsymbol{x} \mid \boldsymbol{\phi})$ is a probability density function (or a probability mass function) so that $p(\boldsymbol{x} \mid \boldsymbol{\phi}) \geq 0$, and $\int_X p(\boldsymbol{x} \mid \boldsymbol{\phi})\, d\boldsymbol{x} = 1$ for all $\boldsymbol{\phi} \in \Phi$. The parameter $\boldsymbol{\phi}$ will generally be assumed to be a vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)$ of finite dimension $k \geq 1$, so that $\Phi \subset \Re^k$. Often, the parameter vector $\boldsymbol{\phi}$ will be written in the form $\boldsymbol{\phi} = \{\boldsymbol{\theta}, \boldsymbol{\omega}\}$, where $\boldsymbol{\theta}$ is considered to be the vector of interest and $\boldsymbol{\omega}$ a vector of nuisance parameters. The sets X and Φ will be referred to, respectively, as the *sample space* and the *parameter space*. Occasionally, if there is no danger of confusion, reference is made to 'model' $\{p(\boldsymbol{x} \mid \boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi\}$, or even to 'model' $p(\boldsymbol{x} \mid \boldsymbol{\phi})$, without recalling the sample and the parameter spaces. In non-regular problems the sample space X depends on the parameter value $\boldsymbol{\phi}$; this will explicitly be indicated by writing $X = X(\boldsymbol{\phi})$. Considered as a function of the parameter $\boldsymbol{\phi}$, the probability density (or probability mass) $p(\boldsymbol{x} \mid \boldsymbol{\phi})$ will be referred to as the *likelihood function* of $\boldsymbol{\phi}$ given $\boldsymbol{x}$. Whenever this exists, a maximum of the likelihood function (*maximum likelihood estimate* or *mle*) will be denoted by $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}(\boldsymbol{x})$.

The *complete* set of available data is represented by $\boldsymbol{x}$. In many examples this will be a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a model of the form $\{p(x \mid \boldsymbol{\phi}), x \in \Re, \boldsymbol{\phi} \in \Phi\}$ so that the likelihood function will be of the form $p(\boldsymbol{x} \mid \boldsymbol{\phi}) = \prod_{j=1}^{n} p(x_j \mid \boldsymbol{\phi})$ and the sample space will be $X \subset \Re^n$, but it will *not* be assumed that this has to be the case. The notation $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, $\boldsymbol{t} \in T$, is used to refer to a general function of the data; often, but not necessarily, this will be a sufficient statistic.

1.3. Simple Model Choice

The simplest example of a model choice problem (and one which centers most discussions on model choice and model comparison) is one where (i) the range of models considered is a finite class $\mathcal{M} = \{M_1, \ldots, M_m\}$, of m fully specified models

$$M_i \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\phi}_i),\ \boldsymbol{x} \in X\}, \quad i = 1, \ldots, m \tag{1}$$

(ii) it is assumed that that the 'true' model is a member $M_t \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\phi}_t), \boldsymbol{x} \in X\}$ from that class, and (iii) the loss function is the simple step function

$$\begin{cases} \ell(a_t, \boldsymbol{\phi}_t) = 0, \\ \ell(a_i, \boldsymbol{\phi}_t) = c > 0, \quad i \neq t, \end{cases} \tag{2}$$

where a_i denotes the *decision to act as if* the true model was M_i. In this simplistic situation, it is immediate to verify that the optimal model choice is that which maximizes the posterior probability, $\pi(\boldsymbol{\phi}_i \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)\pi(\boldsymbol{\phi}_i)$. Moreover, an intuitive measure of paired comparison of plausibility between any two of the models M_i and M_j is provided by the ratio of the posterior probabilities $\pi(\boldsymbol{\phi}_i \mid \boldsymbol{x})/\pi(\boldsymbol{\phi}_j \mid \boldsymbol{x})$. If, in particular, all m models are judged to be equally likely a priori, so that $\pi(\boldsymbol{\phi}_i) = 1/m$, for all i, then the optimal model is that which maximizes the likelihood, $p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)$, and the ratio of posterior probabilities reduces to the corresponding *Bayes*

factor $B_{ij} = p(\boldsymbol{x} \mid \boldsymbol{\phi}_i)/p(\boldsymbol{x} \mid \boldsymbol{\phi}_j)$ which, in this simple case (with no nuisance parameters), it is also the corresponding likelihood ratio.

The natural extension of this scenario to a continuous setting considers a non-countable class of models $\mathcal{M} = \{M_\phi, \boldsymbol{\phi} \in \Phi \subset \Re^k\}$,

$$M_\phi \equiv p(\boldsymbol{x} \mid \boldsymbol{\phi}); \quad \text{with} \quad p(\boldsymbol{x} \mid \boldsymbol{\phi}) > 0, \quad \int_X p(\boldsymbol{x} \mid \boldsymbol{\phi})\, d\boldsymbol{x} = 1, \tag{3}$$

an absolutely continuous and strictly positive prior, represented by its density $p(\boldsymbol{\phi}) > 0$, and a simple step loss function $\ell(a_\phi, \boldsymbol{\phi})$ such that

$$\begin{cases} \ell(a_\phi, \boldsymbol{\phi}_t) = 0, & \boldsymbol{\phi} \in B_\epsilon(\boldsymbol{\phi}_t) \\ \ell(a_\phi, \boldsymbol{\phi}_t) = c > 0, & \boldsymbol{\phi} \notin B_\epsilon(\boldsymbol{\phi}_t), \end{cases} \tag{4}$$

where a_ϕ denotes the decision to act as if the true model was M_ϕ, and $B_\epsilon(\boldsymbol{\phi}_t)$ is a radius ϵ neighbourhood of $\boldsymbol{\phi}_t$. In this case, it is easily shown that, as ϵ decreases, the optimal model choice converges to the model labelled by the mode of the corresponding posterior distribution $\pi(\boldsymbol{\phi} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\phi})\, \pi(\boldsymbol{\phi})$. Note that with this formulation, which strictly parallels the conventional formulation for model choice in the finite case, the problem of model choice is mathematically equivalent to the problem of point estimation with a zero-one loss function.

1.4. Hypothesis Testing

Within the context of an accepted, possibly very wide class of models, $\mathcal{M} = \{M_\phi, \boldsymbol{\phi} \in \Phi\}$, a subset $\mathcal{M}_0 = \{M_\phi,\ \boldsymbol{\phi} \in \Phi_0 \subset \Phi\}$ of the class $\mathcal{M}$, where Φ_0 may possibly consist of a single value $\boldsymbol{\phi}_0$, is sometimes suggested in the course of the investigation as deserving special attention. This may either be because restricting $\boldsymbol{\phi}$ to Φ_0 would greatly simplify the model, or because there are additional (context specific) arguments suggesting that $\boldsymbol{\phi} \in \Phi_0$. The conventional formulation of a *hypothesis testing* problem is stated within this framework. Thus, given data $\boldsymbol{x} \in X$ which are *assumed* to have been generated by $p(\boldsymbol{x} \mid \boldsymbol{\phi})$, for some $\boldsymbol{\phi} \in \Phi$, a procedure is required to advise on whether or not if may safely be assumed that $\boldsymbol{\phi} \in \Phi_0$. In conventional language, a procedure is desired to *test* the *null* hypothesis $H_0 \equiv \{\boldsymbol{\phi} \in \Phi_0\}$. The particular case where Φ_0 contains a *single* value $\boldsymbol{\phi}_0$, so that $\Phi_0 = \{\boldsymbol{\phi}_0\}$, is further referred to as a problem of *precise* hypothesis testing.

The standard frequentist approach to *precise* hypothesis testing requires to propose some one-dimensional test statistic $t = t(\boldsymbol{x}) \in T \subset \Re$, where large values of t cast doubt on H_0. The p-value (or observed significance level) associated to some observed data $\boldsymbol{x}_0 \in X$ is then the probability, conditional on the null hypothesis being true, of observing data as or more extreme than the data actually observed, that is,

$$p = \Pr[t \geq t(\boldsymbol{x}_0) \mid \boldsymbol{\phi} = \boldsymbol{\phi}_0] = \int_{\{\boldsymbol{x};\, t(\boldsymbol{x}) \geq t(\boldsymbol{x}_0)\}} p(\boldsymbol{x} \mid \boldsymbol{\phi}_0)\, d\boldsymbol{x} \tag{5}$$

Small values of the p-value are considered to be evidence against H_0, with the values 0.05 and 0.01 typically used as conventional cut-off points.

There are many well-known criticisms to this common procedure, some of which are briefly reviewed below. For further discussion see Jeffreys (1961), Edwards, Lindman and Savage (1963), Rao (1966), Lindley (1972), Good (1983), Berger and Delampady (1987), Berger and Sellke (1987), Matthews (2001), and references therein.

- *Arbitrary choice of the test statistic*. There is no generally accepted theory on the selection of the appropriate test statistic, and different choices may well lead to incompatible results.

- *Not a measure of evidence*. Observed significance levels are not direct measures of evidence. Although most users would like it to be true, in precise hypothesis testing there is no mathematical relation between the p-value and $\Pr[H_0 \,|\, \boldsymbol{x}_0]$, the probability that the null is true given the evidence.
- *Arbitrary cut-off points*. Conventional cut-off points for p-values (as the ubiquitous 0.05) are arbitrary, and ignore power. Moreover, despite frequent warnings in the literature, they are typically chosen with no regard for either the dimensionality of the problem or the sample size (possibly due to the fact that there is no accepted methodology to perform that adjustment).
- *Exaggerate significance*. Different arguments have been used to suggest that the conventional use of p-values exaggerate significance. Indeed, with common sample sizes, a 0.05 p-value is typically better seen as an indication that more data are needed than as firm evidence against the null.
- *Improper conditioning*. Observed significance levels are not based on the *observed* evidence, namely $t(\boldsymbol{x}) = t(\boldsymbol{x}_0)$, but on the (less than obviously relevant) event $\{t(\boldsymbol{x}) \geq t(\boldsymbol{x}_0)\}$ so that, to quote Jeffreys (1980, p. 453), the null hypothesis may be rejected by not predicting something that has not happened.
- *Contradictions*. Using fixed cut-off points for p-values easily leads to contradiction. For instance, in a multivariate setting, one may simultaneously reject all components $\phi_i = \phi_{i0}$ and yet accept $\boldsymbol{\phi} = \boldsymbol{\phi}_0$ (Rao's paradox).
- *No general procedure*. The procedure is not directly applicable to general hypothesis testing problems. Indeed, the p-value is a function of the sampling distribution of the test statistic under the null, and this is only well defined in the case of *precise* hypothesis testing. Extensions to the general case, $\mathcal{M}_0 = \{M_\phi,\ \boldsymbol{\phi} \in \Phi_0\}$, where Φ_0 contains more than one point, are less than obvious.

Hypothesis testing has been formulated as a decision problem. No wonder therefore that Bayesian approaches to hypothesis testing are best described within the unifying framework of decision theory. Those are reviewed below.

2. Hypothesis Testing as a Decision Problem

2.1. General Structure

Consider the probability model $M \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ which is currently assumed to provide an appropriate description of the probabilistic behaviour of observable data $\boldsymbol{x} \in X$ in terms of some *vector of interest* $\boldsymbol{\theta} \in \Theta$ and some *nuisance parameter vector* $\boldsymbol{\omega} \in \Omega$. From a Bayesian viewpoint, the complete final outcome of a problem of inference about any unknown quantity is the appropriate posterior distribution. Thus, given data $\boldsymbol{x}$ and a (joint) prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$, all that can be said about $\boldsymbol{\theta}$ is encapsulated in the corresponding posterior distribution

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \int_\Omega \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\omega}, \qquad \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega}). \tag{6}$$

In particular, the (marginal) posterior distribution of $\boldsymbol{\theta}$ immediately conveys information on those values of the vector of interest which (given the assumed model) may be taken to be *compatible* with the observed data $\boldsymbol{x}$, namely, those with a relatively high probability density. In some occasions, a particular value $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \in \Theta$ of the quantity of interest is suggested in the course of the investigation as deserving special consideration, either because assuming $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ would greatly simplify the model, or because there are additional (context specific) arguments

suggesting that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Intuitively, the *(null) hypothesis* $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ should be judged to be *compatible* with the observed data $\boldsymbol{x}$ if $\boldsymbol{\theta}_0$ has a relatively high posterior density; however, a more precise conclusion is often required, and this may be derived from a decision-oriented approach.

Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is defined as a *decision problem* where the action space has only two elements, namely to accept (a_0) or to reject (a_1) the use of the restricted model $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$ as a convenient proxy for the *assumed* model $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega\}$. To solve this decision problem, it is necessary to specify an appropriate loss function, $\{\ell[a_i, (\boldsymbol{\theta}, \boldsymbol{\omega})],\ i = 0, 1\}$, measuring the consequences of accepting or rejecting H_0 as a function of the actual values $(\boldsymbol{\theta}, \boldsymbol{\omega})$ of the parameters. Notice that this requires the statement of an *alternative* action a_1 to accepting H_0; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined.

Given data $\boldsymbol{x}$, the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_\Theta \int_\Omega \ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})]\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}$, is larger than the expected posterior loss of rejecting, $\int_\Theta \int_\Omega \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})]\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}$, *i.e.*, iff

$$\int_\Theta \int_\Omega \{\ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})] - \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})]\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > 0. \tag{7}$$

Therefore, *only* the *loss difference*

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \ell[a_0, (\boldsymbol{\theta}, \boldsymbol{\omega})] - \ell[a_1, (\boldsymbol{\theta}, \boldsymbol{\omega})], \tag{8}$$

which measures the *advantage* of rejecting H_0 as a function of $\{\boldsymbol{\theta}, \boldsymbol{\omega}\}$, has to be specified. Notice that no constraint has been imposed in the preceding formulation. It follows that *any* (generalized) Bayes solution to the decision problem posed (and hence any *admissible* solution, see *e.g.*, Berger, 1985, Ch. 8) *must* be of the form

$$\text{Reject } H_0 \quad \text{iff} \quad \int_\Theta \int_\Omega \Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > 0, \tag{9}$$

for some loss difference function $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, and some (possibly improper) prior $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$. Thus, as common sense dictates, the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive. In some examples, the loss difference function does not depend on the nuisance parameter vector $\boldsymbol{\omega}$; if this is the case, the decision criterion obviously simplifies to rejecting H_0 iff $\int_\Theta \Delta\ell(H_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \mid \boldsymbol{x})\, d\boldsymbol{\theta} > 0$.

A crucial element in the specification of the loss function is a description of what is precisely meant by rejecting H_0. By assumption, a_0 means to act *as if* model M_0 were true, *i.e.*, as if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, but there are at least two options for the alternative action a_1. This might mean the *negation* of H_0, that is to act as if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, or it might rather mean to reject the simplification to M_0 implied by $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, and to keep the unrestricted model M (with $\boldsymbol{\theta} \in \Theta$), which is acceptable by assumption. Both of these options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis where precise hypothesis testing procedures are typically used are better described by the second alternative. Indeed, this is the situation in two frequent scenarios: (i) an established model, identified by M_0, is *embedded* into a more general model M (so that $M_0 \subset M$), constructed to include possibly promising departures from M_0, and it is required to verify whether or not the extended model M provides a significant improvement in the description of the behaviour of the available data; or, (ii) a large model M is accepted, and it is required to verify whether or not the simpler model M_0 may be used as a sufficiently accurate approximation.

2.2. Bayes Factors

The *Bayes factor* approach to hypothesis testing is a particular case of the decision structure outlined above; it is obtained when the alternative action a_1 is taken to be to act as if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, and the difference loss function is taken to be a simplistic zero-one function. Indeed, if the *advantage* $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ of rejecting H_0 is of the form

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \Delta\ell(H_0, \boldsymbol{\theta}) = \begin{cases} -1 & \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ +1 & \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \end{cases} \tag{10}$$

then the corresponding decision criterion is

$$\text{Reject } H_0 \quad \text{iff} \quad \Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0 \,|\, \boldsymbol{x}) < \Pr(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \,|\, \boldsymbol{x}). \tag{11}$$

If the prior distribution is such that $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0) = \Pr(\boldsymbol{\theta} \neq \boldsymbol{\theta}_0) = 1/2$, and $\{\pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})\}$ respectively denote the conditional prior distributions of $\boldsymbol{\omega}$, when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then the criterion becomes

$$\text{Reject } H_0 \quad \text{iff} \quad B_{01}\{\boldsymbol{x}, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})\} = \frac{\int_\Omega p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}_0)\, d\boldsymbol{\omega}}{\int_\Theta \int_\Omega p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})\, d\boldsymbol{\theta} d\boldsymbol{\omega}} < 1 \tag{12}$$

where $B_{01}\{\boldsymbol{x}, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}_0), \pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})\}$ is the *Bayes factor* (or integrated likelihood ratio) in favour of H_0. Notice that the Bayes factor B_{01} crucially depends on the conditional priors $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta}_0)$ and $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{\theta})$, which must typically be proper for the Bayes factor to be well-defined.

It is important to realize that this formulation *requires* that $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0) > 0$, so that the hypothesis H_0 must have a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter, this *forces* the use of a *non-regular* (not absolutely continuous) 'sharp' prior concentrating a positive probability mass on $\boldsymbol{\theta}_0$. One unappealing consequence of this non-regular prior structure, noted by Lindley (1957) and generally known as *Lindley's paradox*, is that for any *fixed* value of the pertinent test statistic, the Bayes factor typically increases as $\sqrt{n}$ with the sample size; hence, with large samples, "evidence" in favor of H_0 *may* be overwhelming with data sets which are both extremely implausible under H_0 and quite likely under alternative $\boldsymbol{\theta}$ values, such as (say) the mle $\hat{\boldsymbol{\theta}}$. For further discussion of this polemical issue see Bernardo (1980), Shafer (1982), Berger and Delampady (1987), Casella and Berger (1987), Robert (1993), Bernardo (1999), and discussions therein.

The Bayes factor approach to hypothesis testing in a continuous parameter setting deals with situations of *concentrated* prior probability; it *assumes* important prior knowledge about the value of the vector of interest $\boldsymbol{\theta}$ (described by a prior sharply spiked on $\boldsymbol{\theta}_0$) and analyzes how such *very strong* prior beliefs about the value of $\boldsymbol{\theta}$ should be modified by the data. Hence, Bayes factors should *not* be used unless this strong prior formulation is an appropriate assumption. In particular, Bayes factors should *not* be used to test the *compatibility* of the data with H_0, for they inextricably combine what data have to say with (typically subjective) *strong* beliefs about the value of $\boldsymbol{\theta}$.

2.3. Continuous Loss Functions

It is often natural to assume that the loss difference $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, a conditional measure of the loss suffered if $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega})$ were used as a proxy for $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$, has to be some *continuous* function of the 'discrepancy' between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Moreover, one would expect $\Delta\ell(H_0, \boldsymbol{\theta}_0, \boldsymbol{\omega})$ to be negative, for there must be some positive advantage, say $\ell^* > 0$, in accepting the null when it is true. A simple example is the quadratic loss

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \Delta\ell(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^2 - \ell^*, \quad \ell^* > 0, \tag{13}$$

Notice that continuous difference loss functions do not require the use of non-regular priors. As a consequence, their use does not *force* the assumption of strong prior beliefs and, in particular,

they may be used with improper priors. However, (i) there are many possible choices for continuous difference loss functions; (ii) the resulting criteria are typically not invariant under one-to-one reparametrization of the quantity of interest; and (iii) their use requires some form of calibration, that is, an appropriate choice of the utility constant ℓ^*, which is often context dependent.

In the next section we justify the choice of a particular continuous invariant difference loss function, the *intrinsic discrepancy*. This is combined with reference analysis to propose an attractive Bayesian solution to the problem of hypothesis testing, defined as the problem of deciding whether or not available data are statistically compatible with the hypothesis that the parameters of the model belong to some subset of the parameter space. The proposed solution sharpens a procedure suggested by Bernardo (1999) to make it applicable to non-regular models, and extends previous results to multivariate probability models. For earlier, related references, see Bernardo (1982, 1985), Bernardo and Bayarri (1985), Ferrándiz (1985), Gutiérrez-Peña (1992), and Rueda (1992). The argument lies entirely within a Bayesian decision-theoretical framework (in that the proposed solution is obtained by minimizing a posterior expected loss), and it is *objective* (in the precise sense that it only uses an "objective" prior, a prior uniquely defined in terms of the assumed model and the quantity of interest).

3. The Bayesian Reference Criterion

Let model $M \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}),\, \boldsymbol{\theta} \in \Theta,\, \boldsymbol{\omega} \in \Omega\}$ be a currently accepted description of the probabilistic behaviour of data $\boldsymbol{x} \in X$, let a_0 be the decision to work under the restricted model $M_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, and let a_1 be the decision to keep the general, unrestricted model M. In this situation, the loss advantage $\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ of rejecting H_0 as a function of $(\boldsymbol{\theta}, \boldsymbol{\omega})$ may safely be assumed to have the form

$$\Delta\ell(H_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) - d^*, \qquad d^* > 0, \tag{14}$$

where

(i) the function $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is some non-negative measure of the *discrepancy* between the assumed model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ and its closest approximation within $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, such that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0, \boldsymbol{\omega}) = 0$, and

(ii) the constant $d^* > 0$ is a context dependent *utility value* which measures the (necessarily positive) advantage of being able to work with the simpler model when it is true.

Choices of both $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ and d^* which might be appropriate for general use will now be discussed.

3.1. The Intrinsic Discrepancy

Conventional loss functions typically focus on the "distance" between the true and the null values of the quantity of interest, rather than on the "distance" between the models they label and, typically, they are *not* invariant under reparametrization. Intrinsic losses however (see *e.g.*, Robert, 1996) directly focus on how different the true model is from the null model, and they typically produce invariant solutions. We now introduce a new, particularly attractive, intrinsic loss function, the *intrinsic discrepancy loss*.

The basic idea is to define the discrepancy between two probability densities $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ as $\min\{k(p_1 \,|\, p_2), k(p_2 \,|\, p_1)\}$, where

$$k(p_2 \,|\, p_1) = \int_X p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x} \tag{15}$$

is the *directed logarithmic divergence* (Kullback and Leibler, 1951; Kullback, 1959) of $p_2(\boldsymbol{x})$ from $p_1(\boldsymbol{x})$. The discrepancy from a point to a set is further defined as the discrepancy from the point to its closest element in the set. The introduction of the minimum makes it possible to define a symmetric discrepancy between probability densities which is *finite* with strictly nested supports, a crucial property if a general theory (applicable to non-regular models) is required.

Definition 1. Intrinsic Discrepancies. The intrinsic discrepancy $\delta(p_1, p_2)$ between two probability densities $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ for the random quantity $\boldsymbol{x} \in X$ is

$$\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = \min\Big\{ \int_X p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x},\ \int_X p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\, d\boldsymbol{x} \Big\}$$

The intrinsic discrepancy between two families of probability densities for the random quantity $\boldsymbol{x} \in X$, $M_1 \equiv \{p_1(\boldsymbol{x} \,|\, \boldsymbol{\phi}), \boldsymbol{\phi} \in \Phi\}$ and $M_2 \equiv \{p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi}), \boldsymbol{\psi} \in \Psi\}$, is given by

$$\delta(M_1, M_2) = \min_{\boldsymbol{\phi} \in \Phi,\, \boldsymbol{\psi} \in \Psi} \delta\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\phi}), p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi})\}$$

◁

It immediately follows for Definition 1 that $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ provides the minimum expected log-density ratio $\log[p_i(\boldsymbol{x})/p_j(\boldsymbol{x})]$ in favour of the true density that one would obtain if data $\boldsymbol{x} \in X$ were sampled from either $p_1(\boldsymbol{x})$ or $p_2(\boldsymbol{x})$. In particular, if $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ are fully specified alternative probability models for data $\boldsymbol{x} \in X$, and it is assumed that one of them is true, then $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ is the minimum expected log-likelihood ratio for the true model.

Intrinsic discrepancies have a number of attractive properties. Some are directly inherited from the directed logarithmic divergence. Indeed,

(i) The intrinsic discrepancy $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ is *non-negative* and vanishes iff $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ almost everywhere.

(ii) The intrinsic discrepancy $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\}$ is invariant under one-to-one transformations $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$ of the random quantity $\boldsymbol{x}$.

(iii) The intrinsic discrepancy is *additive* in the sense that if the available data $\boldsymbol{x}$ consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from either $p_1(x)$ or $p_2(x)$, then $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = n\, \delta\{p_1(x), p_2(x)\}$.

(iv) If the densities $p_1(\boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\phi}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x} \,|\, \boldsymbol{\phi}_2)$ are two members of a parametric family $p(\boldsymbol{x} \,|\, \boldsymbol{\phi})$, then $\delta\{p(\boldsymbol{x} \,|\, \boldsymbol{\phi}_1), p(\boldsymbol{x} \,|\, \boldsymbol{\phi}_2)\} = \delta\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$ is *invariant* under one-to-one transformations for the parameter, so that for any such transformation $\boldsymbol{\psi}_i = \boldsymbol{\psi}(\boldsymbol{\phi}_i)$, one has $\delta\{p(\boldsymbol{x} \,|\, \boldsymbol{\psi}_1), p(\boldsymbol{x} \,|\, \boldsymbol{\psi}_2)\} = \delta\{\boldsymbol{\psi}(\boldsymbol{\phi}_1), \boldsymbol{\psi}(\boldsymbol{\phi}_2)\} = \delta\{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2\}$.

(v) The intrinsic discrepancy between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ measures the minimum amount of information (in natural information units, *nits*) that one observation $\boldsymbol{x} \in X$ may be expected to provide in order to discriminate between $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ (Kullback, 1959).

Moreover, the intrinsic discrepancy has two further important properties which the directed logarithmic divergence does *not* have:

(vi) The intrinsic discrepancy is *symmetric* so that $\delta\{p_1(\boldsymbol{x}), p_2(\boldsymbol{x})\} = \delta\{p_2(\boldsymbol{x}), p_1(\boldsymbol{x})\}$.

(vii) If the two densities have strictly nested supports, so that $p_1(\boldsymbol{x}) > 0$ iff $\boldsymbol{x} \in X_1$, $p_2(\boldsymbol{x}) > 0$ iff $\boldsymbol{x} \in X_2$, and either $X_1 \subset X_2$ or $X_2 \subset X_1$, then the intrinsic discrepancy is still typically *finite*. More specifically, the intrinsic discrepancy then reduces to one of the directed logarithmic divergences while the other diverges, so that $\delta\{p_1, p_2\} = k(p_1 \,|\, p_2)$ when $X_2 \subset X_1$, and $\delta\{p_1, p_2\} = k(p_2 \,|\, p_1)$ when $X_1 \subset X_2$.

Example 1. Discrepancy between a Binomial distribution and its Poisson approximation. Let $p_1(x)$ be a binomial distribution $\text{Bi}(x \mid n, \theta)$, and let $p_2(x)$ be its Poisson approximation $\text{Pn}(x \mid n\theta)$. Since $X_1 \subset X_2$, $\delta(p_1, p_2) = k(p_2 \mid p_1)$; thus,

$$\delta\{p_1(x), p_2(x)\} = \delta(n, \theta) = \sum_{x=0}^{n} \text{Bi}(x \mid n, \theta) \log \frac{\text{Bi}(x \mid n, \theta)}{\text{Pn}(x \mid n\theta)} \ .$$

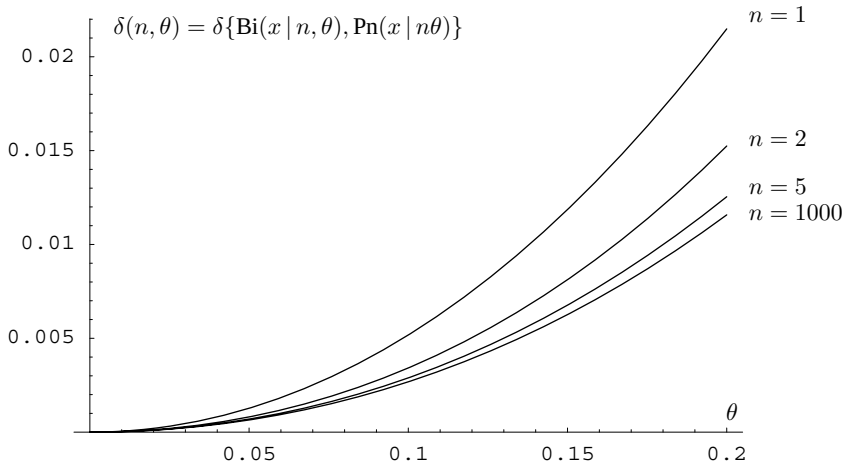


Figure 1. *Intrinsic discrepancy between a Binomial distribution Bi$(x \mid n, \theta)$ and a Poisson distribution Pn$(x \mid n\theta)$ as a function of θ, for $n = 1, 2, 5$ and 1000.*

The resulting discrepancy, $\delta(n, \theta)$ is plotted in Figure 1 as a function of θ for several values of n. As one might expect, the discrepancy converges to zero as θ decreases and as n increases, but it is apparent from the graph that the important condition for the approximation to work is that θ has to be small. ◁

The definition of the intrinsic divergence suggests an interesting new form of convergence for probability distributions:

Definition 2. Intrinsic Convergence. A sequence of probability distributions represented by their density functions $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ is said to converge *intrinsically* to a probability distribution with density $p(\boldsymbol{x})$ whenever $\lim_{i\to\infty} \delta(p_i, p) = 0$, that is, whenever the intrinsic discrepancy between $p_i(x)$ and $p(x)$ converges to zero. ◁

Example 2. Intrinsic convergence of Student densities to a Normal density. The intrinsic discrepancy between a standard Normal and a standard Student with α degrees of freedom is $\delta(\alpha) = \delta\{\text{St}(x \mid 0, 1, \alpha), \text{N}(x \mid 0, 1)\}$, *i.e.*,

$$\min\Big\{ \int_{-\infty}^{\infty} \text{St}(x \mid 0, 1, \alpha) \log \frac{\text{St}(x \mid 0, 1, \alpha)}{\text{N}(x \mid 0, 1)}\, dx,\ \int_{-\infty}^{\infty} \text{N}(x \mid 0, 1) \log \frac{\text{N}(x \mid 0, 1)}{\text{St}(x \mid 0, 1, \alpha)}\, dx \Big\};$$

The second integral may be shown to be always smaller than the first, and to yield an analytical result (in terms of the Hypergeometric and Beta functions) which, for large α values, may be approximated by Stirling to obtain

$$\delta(\alpha) = \int_{-\infty}^{\infty} \text{N}(x \mid 0, 1) \log \frac{\text{N}(x \mid 0, 1)}{\text{St}(x \mid 0, 1, \alpha)}\, dx = \frac{1}{(1+\alpha)^2} + o(\alpha^{-2}) \ ,$$

a function which rapidly converges to zero. Thus, a sequence of standard Student densities with increasing degrees of freedom intrinsically converges to a standard normal density. ◁

In this paper, intrinsic discrepancies are basically used to measure the "distance" between alternative model assumptions about data $\boldsymbol{x} \in X$. Thus, $\delta\{p_1(\boldsymbol{x} \mid \boldsymbol{\phi}), p_2(\boldsymbol{x} \mid \boldsymbol{\psi})\}$ is a symmetric measure (in natural information units, *nits*) of how different the probability densities $p_1(\boldsymbol{x} \mid \boldsymbol{\phi})$ and $p_2(\boldsymbol{x} \mid \boldsymbol{\psi})$ are from each other as a function of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. Since, for any given data $\boldsymbol{x} \in X$, $p_1(\boldsymbol{x} \mid \boldsymbol{\phi})$ and $p_2(\boldsymbol{x} \mid \boldsymbol{\psi})$ are the respective likelihood functions, it follows from Definition 1 that $\delta\{p_1(\boldsymbol{x} \mid \boldsymbol{\phi}), p_2(\boldsymbol{x} \mid \boldsymbol{\psi})\} = \delta(\boldsymbol{\phi}, \boldsymbol{\psi})$ may immediately be interpreted as the *minimum expected log-likelihood ratio in favour of the true model*, assuming that one of the two models is true. Indeed, if $p_1(\boldsymbol{x} \mid \boldsymbol{\phi}_0) = p_2(\boldsymbol{x} \mid \boldsymbol{\psi}_0)$ almost everywhere (and hence the models $p_1(\boldsymbol{x} \mid \boldsymbol{\phi}_0)$ and $p_2(\boldsymbol{x} \mid \boldsymbol{\psi}_0)$ are indistinguishable), then $\delta\{\boldsymbol{\phi}_0, \boldsymbol{\psi}_0)\} = 0$. In general, if either $p_1(\boldsymbol{x} \mid \boldsymbol{\phi}_0)$ or $p_2(\boldsymbol{x} \mid \boldsymbol{\psi}_0)$ is correct, then an intrinsic discrepancy $\delta(\boldsymbol{\phi}_0, \boldsymbol{\psi}_0) = d$ implies an average log-likelihood ratio for the true model of at least d, *i.e.*, minimum likelihood ratios for the true model of about e^d. If $\delta\{\boldsymbol{\phi}_0, \boldsymbol{\psi}_0)\} = 5$, $e^5 \approx 150$, so that data $\boldsymbol{x} \in X$ should then be expected to provide *strong evidence* to discriminate between $p_1(\boldsymbol{x} \mid \boldsymbol{\phi}_0)$ and $p_2(\boldsymbol{x} \mid \boldsymbol{\psi}_0)$. Similarly, if $\delta\{\boldsymbol{\phi}_0, \boldsymbol{\psi}_0)\} = 2.5$, $e^{2.5} \approx 12$, so that data $\boldsymbol{x} \in X$ should then only be expected to provide *mild evidence* to discriminate between $p_1(\boldsymbol{x} \mid \boldsymbol{\phi}_0)$ and $p_2(\boldsymbol{x} \mid \boldsymbol{\psi}_0)$.

Definition 3. Intrinsic Discrepancy Loss. The intrinsic discrepancy loss $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ from replacing the probability model $M = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega,\ \boldsymbol{x} \in X\}$ by its restriction with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $M_0 = \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega,\ \boldsymbol{x} \in X\}$ is the intrinsic discrepancy between the probability density $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and the family of probability densities $\{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$, that is

$$\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) = \min_{\boldsymbol{\omega}_0 \in \Omega} \delta\{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}),\ p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)\}$$

◁

The intrinsic discrepancy $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ between $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and M_0 is the intrinsic discrepancy between the assumed probability density $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ and its closest approximation with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Notice that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is invariant under reparametrization of either $\boldsymbol{\theta}$ or $\boldsymbol{\omega}$. Moreover, if $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is a sufficient statistic for model M, then

$$\int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{x} = \int_T p(\boldsymbol{t} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{t} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{t} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{t}; \tag{16}$$

thus, if convenient, $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ may be computed in terms of the sampling distribution of the sufficient statistic $p(\boldsymbol{t} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$, rather than in terms of the complete probability model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$. Moreover, although not explicitly shown in the notation, the intrinsic discrepancy function typically depends on the sample size. Indeed, if data $\boldsymbol{x} \in X \subset \Re^n$, consist of a *random sample* $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n from $p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})$, then

$$\int_X p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, d\boldsymbol{x} = n \int_{\Re} p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}) \log \frac{p(x \mid \boldsymbol{\theta}_i, \boldsymbol{\omega})}{p(x \mid \boldsymbol{\theta}_j, \boldsymbol{\omega}_j)}\, dx, \tag{17}$$

so that the intrinsic discrepancy associated with the full model $p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ is simply n times the intrinsic discrepancy associated to the model $p(x \mid \boldsymbol{\theta}, \boldsymbol{\omega})$ which corresponds to a single observation. Definition 3 may be used however in problems (say time series) where $\boldsymbol{x}$ does *not* consist of a random sample.

It immediately follows from (9) and (14) that, with an intrinsic discrepancy loss function, the hypothesis H_0 should be rejected if (and only if) the posterior expected advantage of rejecting $\boldsymbol{\theta}_0$, given model M and data $\boldsymbol{x}$, is sufficiently large, so that the decision criterion becomes

$$\text{Reject } H_0 \quad \text{iff} \quad d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \int_\Theta \int_\Omega \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\theta}, \boldsymbol{\omega} \mid \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega} > d^*, \tag{18}$$

for some $d^* > 0$. Since $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ is non-negative, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is nonnegative. Moreover, if $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, then $d(\boldsymbol{\phi}(\boldsymbol{\theta}_0), \boldsymbol{x}) = d(\boldsymbol{\theta}_0, \boldsymbol{x})$, so that the expected intrinsic loss of rejecting H_0 is invariant under reparametrization.

The function $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a continuous, non-negative measure of how inappropriate (in loss of information units) may be expected to be to simplify the model by accepting H_0. Indeed, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a precise measure of the (posterior) expected amount information (in *nits*) which would be necessary to recover the assumed probability density $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ from its closest approximation within $M_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$; it is a measure of the 'strength of evidence' against M_0 given $M \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}),\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\omega} \in \Omega\}$ (cf. Good, 1950). In traditional language, $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ is a (monotone) *test statistic* for H_0, and the null hypothesis should be rejected if the value of $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ exceeds some *critical value* d^*. Notice however that, in sharp contrast to conventional hypothesis testing, the critical value d^* is found to be a positive *utility constant*, which may precisely be described as the number of information units which the decision maker is prepared to lose in order to be able to work with the simpler model H_0, and which does *not* depend on the sampling properties of the test statistic. The procedure may be used with standard, continuous (possibly improper) regular priors when $\boldsymbol{\theta}$ is a continuous parameter (and hence $M_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ is a zero measure set).

Naturally, to implement the decision criterion, both the prior $\pi(\boldsymbol{\theta}, \boldsymbol{\omega})$ and the utility constant d^* must be chosen. These two important issues are now successively addressed, leading to a general decision criterion for hypothesis testing, the *Bayesian reference criterion*.

3.2. The Bayesian Reference Criterion (BRC)

Prior specification. An objective Bayesian procedure (objective in the sense that it depends exclusively on the the assumed model and the observed data), requires an objective "non-informative" prior which mathematically describes lack on relevant information about the quantity of interest, and which only depends on the assumed statistical model and on the quantity of interest. Recent literature contains a number of requirements which may be regarded as necessary properties of any algorithm proposed to derive these 'baseline' priors; those requirements include general applicability, invariance under reparametrization, consistent marginalization, and appropriate coverage properties. The *reference analysis* algorithm, introduced by Bernardo (1979) and further developed by Berger and Bernardo (1989, 1992) is, to the best of our knowledge, the only available method to derive objective priors which satisfy all these desiderata. For an introduction to reference analysis, see Bernardo and Ramón (1998); for a textbook level description see Bernardo and Smith (1994, Ch. 5); for a critical overview of the topic, see Bernardo (1997), references therein and ensuing discussion.

Within a given probability model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$, the joint prior $\pi_\phi(\boldsymbol{\theta}, \boldsymbol{\omega})$ required to obtain the (marginal) *reference* posterior $\pi(\phi \,|\, \boldsymbol{x})$ of some function of interest $\phi = \phi(\boldsymbol{\theta}, \boldsymbol{\omega})$ generally depends on the function of interest, and its derivation is not necessarily trivial. However, under regularity conditions (often met in practice) the required reference prior may easily be found. For instance, if the marginal posterior distribution of the function of interest $\pi(\phi \,|\, \boldsymbol{x})$ has an asymptotic approximation $\hat{\pi}(\phi \,|\, \boldsymbol{x}) = \hat{\pi}(\phi \,|\, \hat{\phi})$ which only depends on the data through a consistent estimator $\hat{\phi} = \hat{\phi}(\boldsymbol{x})$ of ϕ, then the ϕ-reference prior is simply obtained as

$$\pi(\phi) \propto \hat{\pi}(\phi \,|\, \hat{\phi}) \Big|_{\hat{\phi}=\phi}. \tag{19}$$

In particular, if the posterior distribution of ϕ is asymptotically normal $\mathrm{N}(\phi \,|\, \hat{\phi}, s(\hat{\phi})/\sqrt{n})$, then $\pi(\phi) \propto s(\phi)^{-1}$, so that the reference prior reduces to Jeffreys' prior in one-dimensional, asymptotically normal conditions. If, moreover, the sampling distribution of $\hat{\phi}$ only depends

on ϕ, so that $p(\hat{\phi} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}) = p(\hat{\phi} \,|\, \phi)$, then, by Bayes theorem, the corresponding reference posterior is

$$\pi(\phi \,|\, \boldsymbol{x}) \approx \pi(\phi \,|\, \hat{\phi}) \propto p(\hat{\phi} \,|\, \phi)\, \pi(\phi), \tag{20}$$

and the approximation is exact if, given the ϕ-reference prior $\pi_\phi(\boldsymbol{\theta}, \boldsymbol{\omega})$, $\hat{\phi}$ is marginally sufficient for ϕ (rather than just *asymptotically* marginally sufficient).

In our formulation of hypothesis testing, the function of interest (*i.e.*, the function of the parameters which drives the utility function) is the intrinsic discrepancy $\delta = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$. Thus, we propose to use the joint reference prior $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$ which corresponds to the function of interest $\delta = \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$. This implies rejecting the null if (and only if) the reference posterior expectation of the intrinsic discrepancy, which will be referred to as the *intrinsic statistic* $d(\boldsymbol{\theta}_0, \boldsymbol{x})$, is sufficiently large. The proposed test statistic is thus

$$d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \int_\Delta \delta\, \pi_\delta(\delta \,|\, \boldsymbol{x})\, d\delta = \int_\Theta \int_\Omega \delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} d\boldsymbol{\omega}, \tag{21}$$

where $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$ is the posterior distribution which corresponds to the δ-reference prior $\pi_\delta(\boldsymbol{\theta}, \boldsymbol{\omega})$.

Loss calibration. As described in Section 3.1, the intrinsic discrepancy between two fully specified probability models is simply the minimum expected log-likelihood ratio for the true model from data sampled from either of them. It follows that $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$ measures, as a function of $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$, the minimum expected log-likelihood ratio for $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$, against a model of the form $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$, for some $\boldsymbol{\omega}_0 \in \Omega$.

Consequently, given some data $\boldsymbol{x}$, the intrinsic statistic $d(\boldsymbol{\theta}_0, \boldsymbol{x})$, which is simply the reference posterior expectation of $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$, is an estimate (given the available data) of the expected log-likelihood ratio against the null model. This is a continuous measure of the evidence provided by the data against the (null) hypothesis that a model of the form $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\omega}_0)$, for some $\boldsymbol{\omega}_0 \in \Omega$, may safely be used as a proxy for the assumed model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$. In particular, values of $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ of about about 2.5 or 5.0 should respectively be regarded as mild and strong evidence against the (null) hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Example 3. Testing the value of a Normal mean, σ known. Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma^2)$, where σ is assumed to be known, and consider the canonical problem of testing whether these data are (or are not) compatible with some precise hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean. Given σ, the logarithmic divergence of $p(\boldsymbol{x} \,|\, \mu_0, \sigma)$ from $p(\boldsymbol{x} \,|\, \mu, \sigma)$ is the symmetric function

$$k(\mu_0 \,|\, \mu) = n \int_\Re N(x \,|\, \mu, \sigma^2) \log \frac{N(x \,|\, \mu, \sigma^2)}{N(x \,|\, \mu_0, \sigma^2)}\, dx = \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2. \tag{22}$$

Thus, the intrinsic discrepancy in this problem is simply

$$\delta(\mu_0, \mu) = \frac{n}{2} \left(\frac{\mu - \mu_0}{\sigma} \right)^2 = \frac{1}{2} \left(\frac{\mu - \mu_0}{\sigma / \sqrt{n}} \right)^2, \tag{23}$$

half the square of the standardized distance between μ and μ_0. For known σ, the intrinsic discrepancy $\delta(\mu_0, \mu)$ is a piecewise invertible transformation of μ and, hence, the δ-reference prior is simply $\pi_\delta(\mu) = \pi_\mu(\mu) = 1$. The corresponding reference posterior distribution of μ is $\pi_\delta(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma^2/n)$ and, therefore, the intrinsic statistic (the reference posterior expectation of the intrinsic discrepancy) is

$$d(\mu_0, \boldsymbol{x}) = \frac{n}{2} \int_\Re \left(\frac{\mu - \mu_0}{\sigma} \right)^2 \mathrm{N}\left(\mu \,\middle|\, \overline{x}, \frac{\sigma^2}{n} \right) d\mu = \tfrac{1}{2}(1 + z^2), \tag{24}$$

where $z = (\overline{x} - \mu_0)/(\sigma/\sqrt{n})$. Thus, $d(\mu_0, \boldsymbol{x})$ is a simple transformation of z, the number of standard deviations which μ_0 lies away from the data mean $\overline{x}$. The sampling distribution of z^2 is noncentral Chi squared with one degree of freedom and noncentrality parameter 2δ, and its expected value is $1 + 2\delta$, where $\delta = \delta(\mu_0, \mu)$ is the intrinsic discrepancy given by (23). It follows that, in this canonical problem, the expected value under repeated sampling of the reference statistic $d(\mu_0, \boldsymbol{x})$ is equal to one if $\mu = \mu_0$, and increases linearly with n if $\mu \neq \mu_0$.

Scientists have often expressed the view (see ***e.g.***, Jaynes, 1980, or Jeffreys, 1980) that, in this canonical situation, $|z| \approx 2$ should be considered as a mild indication of evidence against $\mu = \mu_0$, while $|z| > 3$ should be regarded as strong evidence against $\mu = \mu_0$. In terms of the intrinsic statistic $d(\mu_0, \boldsymbol{x}) = (1 + z^2)/2$ this precisely corresponds to issuing warning signals whenever $d(\mu_0, \boldsymbol{x})$ is about 2.5 nits, and to reject the null whenever $d(\mu_0, \boldsymbol{x})$ is larger than 5 nits, in perfect agreement with the log-likelihood ratio calibration mentioned above. ◁

Notice, however, that the information scale suggested is an *absolute* scale which is independent of the problem considered, so that rejecting the null whenever its (reference posterior) expected intrinsic discrepancy from the true model is larger than (say) $d^* = 5$ natural units of information is a *general* rule (and one which corresponds to the conventional '3σ' rule in the canonical normal case). Notice too that the use of the ubiquitous 5% confidence level in this problem would correspond to $z = 1.96$, or $d^* = 2.42$ nits, which only indicates mild evidence against the null; this is consistent with other arguments (see ***e.g.***, Berger and Delampady, 1987) suggesting that a p-value of about 0.05 does *not* generally provide sufficient evidence to definitely reject the null hypothesis.

The preceding discussion justifies the following formal definition of an (objective) Bayesian reference criterion for hypothesis testing:

Definition 3. Bayesian Reference Criterion (BRC). Let $\{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$, be a statistical model which is assumed to have been generated some data $\boldsymbol{x} \in X$, and consider a precise value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ among those which remain *possible* after $\boldsymbol{x}$ has ben observed. To decide whether or not the precise value $\boldsymbol{\theta}_0$ may be used as a proxy for the unknown value of $\boldsymbol{\theta}$,

(i) compute the intrinsic discrepancy $\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega})$;

(ii) derive the corresponding reference posterior expectation $d(\boldsymbol{\theta}_0, \boldsymbol{x}) = \mathrm{E}[\delta(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \boldsymbol{\omega}) \,|\, \boldsymbol{x}]$, and state this number as a measure of evidence against the (null) hypothesis $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$.

(iii) If a formal decision is required, reject the null if, and only if, $d(\boldsymbol{\theta}_0, \boldsymbol{x}) > d^*$, for some context dependent d^*. The values $d^* \approx 1.0$ (no evidence against the null), $d^* \approx 2.5$ (mild evidence against the null) and $d^* > 5$ (significant evidence against the null) may conveniently be used for scientific communication. ◁

The results derived in Example 3 may be used to analyze the large sample behaviour of the proposed criterion in one-parameter problems. Indeed, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a large random sample from a one-parameter regular model $\{p(x \,|\, \theta), \theta \in \Theta\}$, the relevant reference prior will be Jeffreys' prior $\pi(\theta) \propto i(\theta)^{1/2}$, where $i(\theta)$ is Fisher's information function, Hence, the reference prior of $\phi(\theta) = \int^{\theta} i(\theta)^{1/2} \, d\theta$ will be uniform, and the reference posterior of ϕ approximately normal $\mathrm{N}(\phi \,|\, \hat{\phi}, 1/\sqrt{n})$. Thus, using Example 3 and the fact that the intrinsic statistic is invariant under one-to-one parameter transformations, one gets the approximation $d(\theta_0, \boldsymbol{x}) = d(\phi_0, \boldsymbol{x}) \approx \frac{1}{2}(1 + z^2)$, where $z = \sqrt{n}(\hat{\phi} - \phi_0)$. Moreover, the sampling distribution of z will approximately be a non-central χ^2 with one degree of freedom and non centrality parameter $n(\phi - \phi_0)^2$. Hence, the expected value of $d(\phi_0, \boldsymbol{x})$ under repeated sampling from

$p(x \mid \theta)$ will approximately be one if $\theta = \theta_0$ and will linearly increase with $n(\theta - \theta_0)^2$ otherwise. More formally, we may state

Proposition 1. One-Dimensional Asymptotic Behaviour. If $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from a regular model $\{p(x \mid \theta),\ \theta \in \Theta \subset \Re,\ x \in X \subset \Re\}$ with one continuous parameter, and $\phi(\theta) = \int^{\theta} i(\theta)^{1/2}\, d\theta$, where $i(\theta) = -\mathrm{E}_{x|\theta}[\partial^2 \log p(x \mid \theta)/\partial\theta^2]$, then the intrinsic statistic $d(\theta_0, \boldsymbol{x})$ to test $\{\theta = \theta_0\}$ is

$$d(\theta_0, \boldsymbol{x}) = \tfrac{1}{2}[1 + z^2(\theta_0, \hat{\theta})] + o(n^{-1}), \quad z(\theta_0, \hat{\theta}) = \sqrt{n}[\phi(\hat{\theta}) - \phi(\theta)].$$

where $\hat{\theta} = \hat{\theta}(\boldsymbol{x}) = \arg\max p(\boldsymbol{x} \mid \theta)$. Moreover, the expected value of $d(\theta_0, \boldsymbol{x})$ under repeated sampling is

$$\mathrm{E}_{x \mid \theta}[d(\theta_0, \boldsymbol{x})] = 1 + n[\phi(\theta) - \phi(\theta_0)]^2 + o(n^{-1}),$$

so that $d(\theta_0, \boldsymbol{x})$ will concentrate around the value one if $\theta = \theta_0$, and will linearly increase with n otherwise. ◁

The arguments leading to Proposition 1 may be extended to multivariate situations, with or without nuisance parameters.

In the final section of this paper we illustrate the behaviour of the Bayesian reference criterion with three examples: (i) hypothesis testing on the value of a binomial parameter, which is used to illustrate the shape of an intrinsic discrepancy, (ii) a problem of precise hypothesis testing within a non-regular probability model, which is used to illustrate the exact behaviour of the BRC criterion under repeated sampling, and (iii) a multivariate normal problem which illustrates how the proposed procedure avoids Rao's paradox on incoherent multivariate frequentist testing.

4. Examples

4.1. Testing the Value of the Parameter of a Binomial Bistribution

Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n conditionally independent Bernoulli observations with parameter θ, so that $p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$, $0 < \theta < 1$, $x \in \{0, 1\}$, and consider testing whether or not the observed data $\boldsymbol{x}$ are compatible with the null hypothesis $\{\theta = \theta_0\}$. The directed logarithmic divergence of $p(x \mid \theta_j)$ from $p(x \mid \theta_i)$ is

$$k(\theta_j \mid \theta_i) = \theta_i \log \frac{\theta_i}{\theta_j} + (1 - \theta_i) \log \frac{(1 - \theta_i)}{(1 - \theta_j)}\,, \tag{25}$$

and it is easily verified that $k(\theta_j \mid \theta_i) < k(\theta_i \mid \theta_j)$ iff $\theta_i < \theta_j < 1 - \theta_i$; thus, the intrinsic discrepancy between $p(\boldsymbol{x} \mid \theta_0)$ and $p(\boldsymbol{x} \mid \theta)$, represented in Figure 2, is

$$\delta(\theta_0, \theta) = n \begin{cases} k(\theta \mid \theta_0) & \theta \in (\theta_0, 1 - \theta_0), \\ k(\theta_0 \mid \theta) & \text{otherwise} \end{cases} \tag{26}$$

Since $\delta(\theta_0, \theta)$ is a piecewise invertible function of θ, the δ-reference prior is just the θ-reference prior and, since Bernoulli is a regular model, this is Jeffreys' prior, $\pi(\theta) = \mathrm{Be}(\theta \mid 1/2, 1/2)$. The reference posterior is the Beta distribution $\pi(\theta \mid \boldsymbol{x}) = \pi(\theta \mid r, n) = \mathrm{Be}(\theta \mid r + 1/2, n - r + 1/2)$, with $r = \sum x_i$, and the intrinsic statistic $d(\theta_0, \boldsymbol{x})$ is the concave function

$$d(\theta_0, \boldsymbol{x}) = d(\theta_0, r, n) = \int_0^1 \delta(\theta_0, \theta)\ \pi(\theta \mid r, n)\, d\theta = \tfrac{1}{2}[1 + z(\theta_0, \hat{\theta})^2] + o(n^{-1}) \tag{27}$$

where $z(\theta_0, \hat{\theta}) = \sqrt{n}[\phi(\hat{\theta}) - \phi(\theta_0)]$, and $\phi(\theta) = 2\mathrm{ArcSin}(\sqrt{\theta})$. The exact value of the intrinsic statistic may easily be found by one-dimensional numerical integration, or may be expressed in

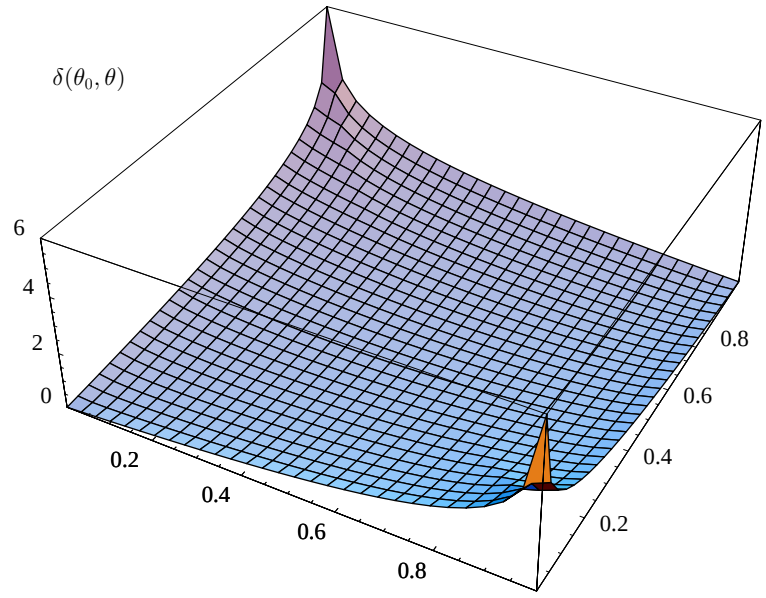


Figure 2. *Intrinsic discrepancy between two Bernoulli probability models.*

terms of Digamma and incomplete Beta functions, but the approximation given above, directly obtained from Proposition 1, is quite good, even for moderate samples.

The canonical particular case where $\theta_0 = 1/2$ deserves special attention. The exact value of the intrinsic statistic is then

$$d(1/2, r, n) = \psi(n+1) + \tilde{\theta}\,\psi(r+1/2) + (1-\tilde{\theta})\,\psi(n-r+1/2) - \log 2 \tag{28}$$

where $\tilde{\theta} = (r+1/2)/(n+1)$ is the reference posterior mean. As one would certainly expect, $d(1/2, 0, n) = d(1/2, n, n)$ increases with n; moreover, it is found that $d(1/2, 0, 6) = 2.92$ and that $d(1/2, 0, 10) = 5.41$. Thus, when $r = 0$ (all failures) or $r = n$ (all successes) the null value $\theta_0 = 1/2$ should be questioned ($d > 2.5$) for all $n > 5$ and definitely rejected ($d > 5$) for all $n > 9$.

4.2. Testing the Value of the Upper Limit of a Uniform Distribution

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_i \in X(\theta) = [0, \theta]$ be a random sample of n uniform observations in $[0, \theta]$, so that $p(x_i \,|\, \theta) = \theta^{-1}$, and consider testing the compatibility of data $\boldsymbol{x}$ with the precise value $\theta = \theta_0$. The logarithmic divergence of $p(\boldsymbol{x} \,|\, \theta_j)$ from $p(\boldsymbol{x} \,|\, \theta_i)$ is

$$k(\theta_j \,|\, \theta_i) = n \int_0^{\theta_i} p(x \,|\, \theta_i) \log \frac{p(x \,|\, \theta_i)}{p(x \,|\, \theta_j)}\, dx = \begin{cases} n\, \log(\theta_j/\theta_i) & \text{if } \theta_i < \theta_j \\ \infty & \text{otherwise} \end{cases} \tag{29}$$

and, therefore, the intrinsic discrepancy between $p(x \,|\, \theta)$ and $p(x \,|\, \theta_0)$ is

$$\delta(\theta_0, \theta) = \min\{k(\theta_0 \,|\, \theta), k(\theta \,|\, \theta_0)\} = \begin{cases} n\, \log(\theta_0/\theta) & \text{if } \theta_0 > \theta \\ n\, \log(\theta/\theta_0) & \text{if } \theta_0 \le \theta. \end{cases} \tag{30}$$

Let $x_{(n)} = \max\{x_1, \ldots, x_n\}$ be the largest observation in the sample. The likelihood function is $p(\boldsymbol{x} \,|\, \theta) = \theta^{-n}$, if $\theta > x_{(n)}$, and zero otherwise; hence, $x_{(n)}$ is a sufficient statistic, and a simple asymptotic approximation $\hat{\pi}(\theta \,|\, \boldsymbol{x})$ to the posterior distribution of θ is given by

$$\hat{\pi}(\theta \,|\, \boldsymbol{x}) = \hat{\pi}(\theta \,|\, x_{(n)}) = \frac{\theta^{-n}}{\int_{x_{(n)}}^{\infty} \theta^{-n}\, d\theta} = (n-1)\, x_{(n)}^{n-1}\, \theta^{-n}, \quad \theta > x_{(n)}. \tag{31}$$

It immediately follows from (31) that $x_{(n)}$ is a consistent estimator of θ; hence, using (19), the θ-reference prior is given by

$$\pi_\theta(\theta) \propto \hat{\pi}(\theta \,|\, x_{(n)})\Big|_{x_{(n)}=\theta} \propto \theta^{-1}. \tag{32}$$

Moreover, for any θ_0, $\delta = \delta(\theta_0, \theta)$ is a piecewise invertible function of θ and, hence, the δ-reference prior is also $\pi_\delta(\theta) = \theta^{-1}$. Using Bayes theorem, the corresponding reference posterior is

$$\pi_\delta(\theta \,|\, \boldsymbol{x}) = \pi_\delta(\theta \,|\, x_{(n)}) = n\, x_{(n)}^n\, \theta^{-(n+1)}, \quad \theta > x_{(n)}; \tag{33}$$

thus, the intrinsic statistic to test the compatibility of the data with any possible value θ_0, *i.e.*, such that $\theta_0 > x_{(n)}$, is given by

$$d(\theta_0, \boldsymbol{x}) = d(t) = \int_{x_{(n)}}^{\infty} \delta(\theta_0, \theta)\, \pi_\delta(\theta \,|\, x_{(n)})\, d\theta = 2t - \log t - 1, \quad t = (x_{(n)}/\theta_0)^n, \tag{34}$$

which only depends on $t = t(\theta_0, x_{(n)}, n) = (x_{(n)}/\theta_0)^n \in [0, 1]$. The intrinsic statistic $d(t)$ is the concave function represented in Figure 3, which has a unique minimum at $t = 1/2$. Hence, the value of $d(\theta_0, \boldsymbol{x})$ is minimized iff $(x_{(n)}/\theta_0)^n = 1/2$, *i.e.*, iff $\theta_0 = 2^{1/n} x_{(n)}$, which is the Bayes estimator for this loss function (and the median of the reference posterior distribution).

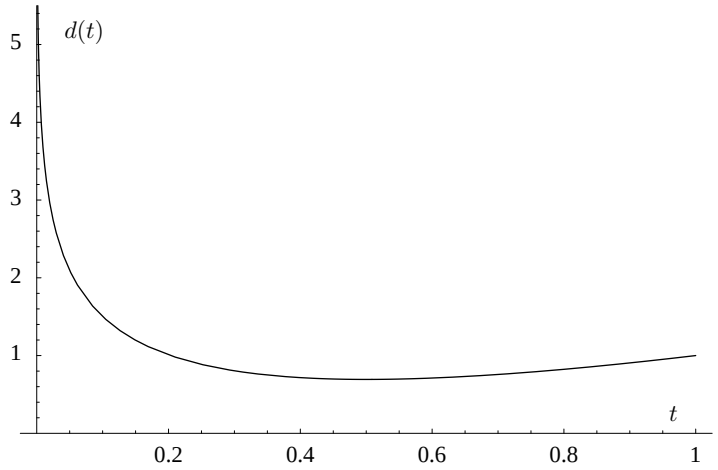


Figure 3. *The intrinsic statistic $d(\theta_0, \boldsymbol{x}) = d(t) = 2t - \log t - 1$ to test $\theta = \theta_0$ which corresponds to a random sample $\{x_1 \ldots, x_n\}$ from uniform distribution Un$(x \,|\, 0, \theta)$, as a function of $t = (x_{(n)}/\theta_0)^n$.*

It may easily be shown that the distribution of t under repeated sampling is uniform in $[0, (\theta/\theta_0)^n]$ and, hence, the expected value of $d(\theta_0, \boldsymbol{x}) = d(t)$ under repeated sampling is

$$\mathrm{E}[d(t) \,|\, \theta] = \int_0^{(\theta/\theta_0)^n} (2t - \log t - 1)\, dt = (\theta/\theta_0)^n - n \log(\theta/\theta_0), \tag{35}$$

which is precisely equal to one if $\theta = \theta_0$, and increases linearly with n otherwise. Thus, once again, one would expect $d(t)$ values to be about one under the null, and one would expect to always reject a false null for a large enough sample. It could have been argued that $t = (x_{(n)}/\theta_0)^n$ is indeed a 'natural' intuitive measure of the evidence provided by the data against the precise value θ_0, but this is not needed; the procedure outlined *automatically* provides an appropriate test function for *any* hypothesis testing problem.

The relationship between BRC and both frequentist testing and Bayesian tail area testing procedures is easily established in this example. Indeed,

(i) The sampling distribution of t under the null is uniform in $[0, 1]$, so that t is precisely the p-value which corresponds to a frequentist test based on any one-to-one function of t.

(ii) The posterior tail area, that is, the reference posterior probability that θ is larger than θ_0, is $\int_{\theta_0}^{\infty} \pi(\theta \mid x_{(n)})\, d\theta = (x_{(n)}/\theta_0)^n = t$, so that t is *also* the reference posterior tail area.

It is immediately verified that $d(0.035) = 2.42$, and that $d(0.0025) = 5$. It follows that, in this problem, the bounds $d^* = 2.42$ and $d^* = 5$, respectively correspond to the p-values 0.035 and 0.0025. Notice that these numbers are *not* equal to the the values 0.05 and 0.0027 obtained when testing a value $\mu = \mu_0$ for a univariate normal mean. This illustrates an important general point: for comparable strength of evidence in terms of information loss, the significance level should depend on the assumed statistical model (even in simple, one-dimensional problems).

4.3. Testing the Value of a Multivariate Normal Mean

Let $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample from $\mathrm{N}_k(\boldsymbol{x} \mid \boldsymbol{\mu}, \sigma^2\Sigma)$, a multivariate normal distribution of dimension k, where Σ is a known symmetric positive-definite matrix. In this final example, tests on the value of $\boldsymbol{\mu}$ are presented for the case where σ is known. Tests for the case where σ is unknown, tests on the value of some of the components of $\boldsymbol{\mu}$, and tests on the values of regression coefficients $\boldsymbol{\beta}$ in normal regression models of the form $\mathrm{N}_k(\boldsymbol{y} \mid X\boldsymbol{\beta}, \sigma^2\Sigma)$, may be obtained from appropriate extensions of the results described below, and will be presented elsewhere.

Intrinsic discrepancy. Without loss of generality, it may be assumed that $\sigma = 1$, for otherwise σ may be included in the matrix Σ; since Σ is known, the vector of means $\overline{\boldsymbol{x}}$ is a sufficient statistic. The sampling distribution of $\overline{\boldsymbol{x}}$ is $p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}) = \mathrm{N}_k(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}, n^{-1}\Sigma)$; thus, using (16), the logarithmic divergence of $p(\boldsymbol{x} \mid \boldsymbol{\mu}_j)$ from $p(\boldsymbol{x} \mid \boldsymbol{\mu}_i)$ is the symmetric function

$$k(\boldsymbol{\mu}_j \mid \boldsymbol{\mu}_i) = \int_{\Re^k} p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_i) \log \frac{p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_i)}{p(\overline{\boldsymbol{x}} \mid \boldsymbol{\mu}_j)}\, d\overline{\boldsymbol{x}} = \frac{n}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{36}$$

It follows that the intrinsic discrepancy between the null model $p(\boldsymbol{x} \mid \boldsymbol{\mu}_0)$ and the assumed model $p(\boldsymbol{x} \mid \boldsymbol{\mu})$ has the quadratic form

$$\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}) = \frac{n}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0). \tag{37}$$

The required test statistic, the intrinsic statistic, is the reference posterior expectation of $\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu})$, $d(\boldsymbol{\mu}_0, \boldsymbol{x}) = \int_{\Re^k} \delta(\boldsymbol{\mu}_0, \boldsymbol{\mu})\, \pi_\delta(\boldsymbol{\mu} \mid \boldsymbol{x})\, d\boldsymbol{\mu}$.

Marginal reference prior. We first make use of standard normal distribution theory to obtain the marginal reference prior distribution of $\lambda = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, and hence that of $\delta = n\,\lambda/2$. Reference priors only depend on the asymptotic behaviour of the model and, for any regular prior, the posterior distribution of $\boldsymbol{\mu}$ is asymptotically multivariate normal $\mathrm{N}_k(\boldsymbol{\mu} \mid \overline{\boldsymbol{x}}, n^{-1}\Sigma)$. Consider $\boldsymbol{\eta} = A(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, where $A'A = \Sigma^{-1}$, so that $\lambda = \boldsymbol{\eta}'\boldsymbol{\eta}$; the posterior distribution of $\boldsymbol{\eta}$ is asymptotically normal $\mathrm{N}_k(\boldsymbol{\eta} \mid A(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0), n^{-1}I_k)$. Hence (see *e.g.*, Rao, 1973, Ch. 3), the posterior distribution of $n\,\lambda = n\,\boldsymbol{\eta}'\boldsymbol{\eta} = n\,(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)$ is asymptotically a non-central Chi squared with k degrees of freedom and non-centrality parameter $n\,\hat{\lambda}$, with $\hat{\lambda} = (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$, and this distribution has mean $k + n\,\hat{\lambda}$ and variance $2(k + 2n\,\hat{\lambda})$. It follows that the marginal posterior distribution of λ is asymptotically normal; specifically,

$$p(\lambda \mid \boldsymbol{x}) \approx \mathrm{N}(\lambda \mid (k + n\,\hat{\lambda})/n,\, 2(k + 2n\,\hat{\lambda})/n^2) \approx \mathrm{N}(\lambda \mid \hat{\lambda},\, 4\hat{\lambda}/n). \tag{38}$$

Hence, the posterior distribution of λ has an asymptotic approximation $\hat{\pi}(\lambda \mid \hat{\lambda})$ which only depends on the data through $\hat{\lambda}$, a consistent estimator of λ. Therefore, using (19), the λ-reference prior is

$$\pi_\lambda(\lambda) \propto \hat{\pi}(\lambda \mid \hat{\lambda})\Big|_{\hat{\lambda}=\lambda} \propto \lambda^{-1/2}. \tag{39}$$

But the parameter of interest, $\delta = n\lambda/2$, is a linear transformation of λ and, therefore, the δ-reference prior is

$$\pi_\delta(\delta) \propto \pi_\lambda(\lambda)|\partial\lambda/\partial\delta| \propto \delta^{-1/2}. \tag{40}$$

Reference posterior and intrinsic statistic. Normal distribution theory may be used to derive the *exact* sampling distribution of the asymptotically sufficient estimator $\hat{\lambda} = (\overline{\boldsymbol{x}}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\overline{\boldsymbol{x}}-\boldsymbol{\mu}_0)$. Indeed, letting $\boldsymbol{y} = A(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$, with $A'A = \Sigma^{-1}$, the sampling distribution of $\boldsymbol{y}$ is normal $\mathrm{N}_k(\boldsymbol{y} \mid A(\boldsymbol{\mu} - \boldsymbol{\mu}_0), n^{-1}I_k)$; thus, the sampling distribution of $n\,\boldsymbol{y}'\boldsymbol{y} = n\,\hat{\lambda}$ is a non-central Chi squared with k degrees of freedom and non-centrality parameter $n\,(\boldsymbol{\mu}-\boldsymbol{\mu}_0)'\Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)$, which by equation (37) is precisely equal to 2δ. Thus, the asymptotic marginal posterior distribution of δ only depends on the data through the statistic,

$$z^2 = n\,\hat{\lambda} = n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0), \tag{41}$$

whose sampling distribution only depends on δ. Therefore, using (20), the reference posterior distribution of δ given z^2 is

$$\pi(\delta \mid z^2) \propto \pi(\delta)\, p(z^2 \mid \delta) = \delta^{-1/2}\chi^2(z^2 \mid k, 2\delta). \tag{42}$$

Transforming to polar coordinates it may be shown (Berger, Philippe, and Robert, 1998) that (42) is actually the reference posterior distribution of δ which corresponds to the ordered parametrization $\{\delta, \boldsymbol{\omega}\}$, where $\boldsymbol{\omega}$ is the vector of the angles, so that, using such a prior, $\pi(\delta \mid \boldsymbol{x}) = \pi(\delta \mid z^2)$, and z^2 encapsulates all available information about the value of δ.

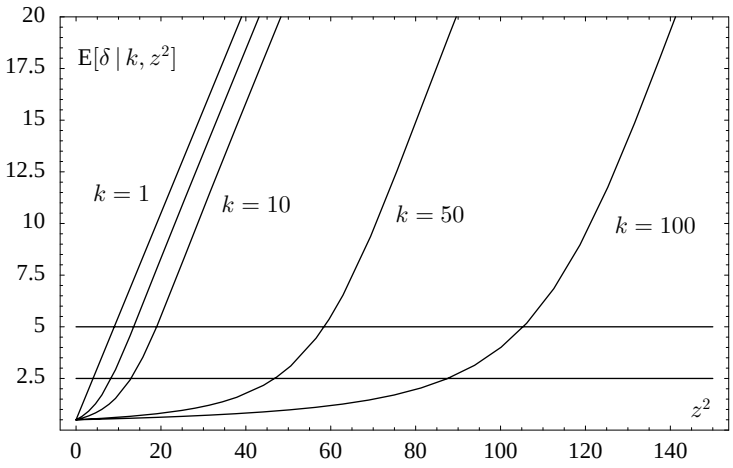


Figure 4. *Approximate behaviour of the intrinsic statistic* $d(\boldsymbol{\mu}_0, \boldsymbol{x}) \approx \mathrm{E}[\delta \mid k, z^2]$ *as a function of* $z^2 = n\,(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\Sigma^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_0)$*, for* $k = 1, 5, 10, 50$ *and* 100*.*

After some tedious algebra, both the missing proportionality constant, and the expected value of $\pi(\delta \mid z^2)$ may be obtained in terms of the ${}_1F_1$ confluent hypergeometric function, leading to

$$d(\boldsymbol{\mu}_0, z^2) = \mathrm{E}[\delta \mid k, z^2] = \frac{1}{2}\,\frac{{}_1F_1(3/2; k/2, z^2/2)}{{}_1F_1(1/2; k/2, z^2/2)}. \tag{43}$$

Moreover, the exact value for $\mathrm{E}[\delta \mid k, z^2]$ given by (43) has a simple linear approximation for large values of z^2, namely,

$$\mathrm{E}[\delta \mid k, z^2] \approx \frac{1}{2}\,(2 - k + z^2). \tag{44}$$

Notice that, in general, (44) is only appropriate for values of z^2 which are large relative to k (showing strong evidence against the null), but it is actually exact for $k = 1$, so that (43) provides a multivariate generalization of (24). Figure 4 shows the form of $\mathrm{E}[\delta \mid k, z^2]$ as a function of z^2 for different values of the dimension k.

Numerical Example: Rao's paradox. As an illustrative numerical example, consider one observation $\boldsymbol{x} = (2.06, 2.06)$ from a bivariate normal density with variances $\sigma_1^2 = \sigma_2^2 = 1$ and correlation coefficient $\rho = 0.5$; the problem is to test whether or not the data $\boldsymbol{x}$ are compatible with the null hypothesis $\boldsymbol{\mu} = (0, 0)$. These data were used by Rao (1966) (and reassessed by Healy, 1969), to illustrate the often neglected fact that using standard significance tests, it can happen that a test for $\mu_1 = 0$ can lead to rejection at the same time as one for $\mu_2 = 0$, whereas the test for $\boldsymbol{\mu} = (0, 0)$ can result in acceptance, a clear example of frequentist incoherence, often known as Rao's paradox. Indeed, with those data, both $\mu_1 = 0$ and $\mu_2 = 0$ are rejected at the 5% level (since $x_1^2 = x_2^2 = 2.06^2 = 4.244$, larger than 3.841, the 0.95 quantile of a χ_1^2), while the same (Hottelling's T^2) test leads to acceptance of $\boldsymbol{\mu} = (0, 0)$ at the same level (since $z^2 = \boldsymbol{x}'\Sigma^{-1}\boldsymbol{x} = 5.658$, smaller than 5.991, the 0.95 quantile of a χ_2^2). However, using (43), we find,

$$\begin{cases} E[\delta \mid 1, 2.06^2] = \frac{1}{2}(1 + 2.06^2) = 2.622, \\ E[\delta \mid 2, 5.658] = \frac{1}{2}\frac{{}_1F_1(3/2;\, 1,\, 5.658/2)}{{}_1F_1(1/2;\, 1,\, 5.658/2)} = 2.727. \end{cases} \tag{45}$$

Thus, the BRC criterion suggests tentative rejection in both cases (since both numbers are larger than 2.5, the '2σ' rule in the canonical normal case), with some extra evidence in the bivariate case, as intuition clearly suggests.

Acknowledgements

The authors thank Professor Dennis Lindley, the Journal Editor Professor Elja Arjas, and an anonymous referee, for helpful comments on an earlier version of the paper. J. M. Bernardo was funded with grants BFM2001-2889 of the DGICYT Madrid and GV01-7 of Generalitat Valenciana (Spain). R. Rueda was funded with grant CONACyT 32256-E (Mexico).

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.* **2**, 317–352 (with discussion).

Berger, J. O. Philippe, A. and Robert, C. P. (1998). Estimation of Quadratic Functions: Noninformative priors for non-centrality parameters. *Statistica Sinica* **8**, 359–376.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of significance levels and evidence. *J. Amer. Statist. Assoc.* **82**, 112–133 (with discussion).

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.), Brookfield, VT: Edward Elgar, (1995), 229–263.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 605–647 (with discussion).

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J. M. (1985). Análisis Bayesiano de los contrastes de hipótesis paramétricos. *Trab. Estadist.* **36**, 45–54.

Bernardo, J. M. (1997). Noninformative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. (1999). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 101–130 (with discussion).

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. *Model Choice* (J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, eds.). Brussels: Pub. Fac. Univ. Saint Louis, 43–59.

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 101–135.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–135, (with discussion).

Edwards, W. L., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* **70**, 193–242. Reprinted in *Robustness of Bayesian Analysis* (J. B. Kadane, ed.). Amsterdam: North-Holland, 1984, 1–62. Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.). Brookfield, VT: Edward Elgar, 1995, 140–189.

Ferrándiz, J. R. (1985). Bayesian inference on Mahalanobis distance: an alternative approach to Bayesian model testing. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 645–654.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London : Griffin; New York: Hafner Press.

Good, I. J. (1983). *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis: Univ. Minnesota Press.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 669–674.

Healy, J. R. (1969). Rao's paradox concerning multivariate tests of significance. *Biometrics* **25**, 411–413.

Jaynes, E. T. (1980). Discussion to the session on hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 618–629. Reprinted in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*. (R. D. Rosenkranz, ed.). Dordrecht: Kluwer (1983), 378–400.

Jeffreys, H. (1961). *Theory of Probability*. (3rd edition) Oxford: University Press.

Jeffreys, H. (1980). Some general points in probability theory. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 451–453.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley. Second edition in 1968, New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lindley, D. V. (1972). *Bayesian Statistics, a Review*. Philadelphia, PA: SIAM.

Matthews, R. A. J. (2001). Why *should* clinicians care about Bayesian methods? *J. Statist. Planning and Inference* **94**, 43–71 (with discussion).

Rao, C. R. (1966). Covariance adjustment and related problems in multivariate analysis. *Multivariate Analysis*. (P. E. Krishnaiah, ed.). New York: Academic Press, 87–103.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. New York: Wiley

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 603–608.

Robert, C.P. (1996). Intrinsic Losses. *Theory and Decision* **40**, 191–214.

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1**, 61-67.

Shafer, G. (1982). Lindley's paradox. *J. Amer. Statist. Assoc.* **77**, 325–351 (with discussion).

Résumé

Pour un modèle probabiliste $M \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\omega}), \boldsymbol{\theta} \in \Theta, \boldsymbol{\omega} \in \Omega\}$ censé décrire le comportement probabiliste de données $\boldsymbol{x} \in X$, nous soutenons que tester si les données sont compatibles avec une hypothèse $H_0 \equiv \{\boldsymbol{\theta} = \boldsymbol{\theta}_0\}$ doit être considéré comme un problème décisionnel concernant l'usage du modèle $M_0 \equiv \{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, avec une fonction de coût qui mesure la quantité d'information qui peut être perdue si le modèle simplifié M_0 est utilisé comme approximation du véritable modèle M. Le coût moyen, calculé par rapport à une loi a priori de référence idoine fournit une statistique de test pertinente, la statistique intrinsèque $d(\boldsymbol{\theta}_0, \boldsymbol{x})$, invariante par reparamétrisation. La statistique intrinsèque $d(\boldsymbol{\theta}_0, \boldsymbol{x})$ est mesurée en unités d'information, et sa calibrage, qui est independante de la taille de l'échantillon et de la dimension du paramétre, ne dépend pas de sa distribution à l'échantillonage. La règle de Bayes correspondante, le critère de Bayes de référence (BRC), indique que H_0 doit seulement être rejeté si le coût a posteriori moyen de la perte d'information à utiliser le modèle simplifié M_0 est trop grande. Le critère BRC fournit une solution bayésienne générale et objective pour les tests d'hypothèses précises qui ne réclame pas une masse de Dirac concentrée sur M_0. Par conséquent, elle échappe au paradoxe de Lindley. Cette théorie est illustrée dans le contexte de variables normales multivariées, et on montre qu'elle évite le paradoxe de Rao sur l'inconsistence existant entre tests univariés et multivariés.

SISTEMA ELECTORAL

Una alternativa a la ley d'Hondt

JOSÉ MIGUEL BERNARDO

Por mandato constitucional, las leyes electorales deben especificar la forma de distribuir los escaños disponibles entre los partidos que concurren a las elecciones atendiendo a criterios de representación proporcional. En España se utiliza para ello un procedimiento, conocido como ley d'Hondt, que pretende proporcionar una buena aproximación a la representación proporcional. Sin embargo, un estudio reciente ha demostrado que el problema de determinar la mejor aproximación posible a una asignación proporcional de escaños tiene, en la práctica, una única solución matemáticamente correcta, que no es la que actualmente se utiliza. Se describe un procedimiento que permite obtenerla con facilidad, y se ilustra con resultados de las últimas elecciones catalanas.

El artículo 68 de la Constitución española señala que la circunscripción electoral es la provincia, especifica que ley electoral distribuirá el número total de Diputados del Congreso entre las circunscripciones asignando una representación mínima inicial a cada una y distribuyendo los demás en proporción a la población, y ordena que la distribución de escaños entre los partidos en cada circunscripción se efectúe "atendiendo a criterios de representación proporcional". Como el número de escaños que se asigna a cada partido debe ser un número entero, no es posible una distribución de escaños exactamente proporcional a los votos obtenidos por cada partido. Para resolver este problema, la ley electoral vigente utiliza un procedimiento conocido como ley d'Hondt. Es fácil comprobar, sin embargo, que este procedimiento distorsiona la voluntad popular, distribuyendo los escaños de una forma que no respeta la representación proporcional. Un reciente estudio matemático realizado en la Universidad de Valencia, demuestra que el problema del reparto entero de escaños de forma aproximadamente proporcional tiene, en la práctica, una única solución óptima, independiente del concepto de aproximación utilizado, y describe un método sencillo para obtenerla. La solución propuesta permite mejorar también la proporcionalidad de la representación obtenida tanto en las elecciones autonómicas como en las elecciones municipales, puesto que, en ambos casos, las leyes electorales vigentes hacen uso de la ley d'Hondt para distribuir los escaños autonómicos correspondientes a cada provincia, o los concejales correspondientes a cada municipio.

Asignación de escaños autonómicos (Lleida, 2003)

15 ESCAÑOS	CiU	PSC	ERC	PP	ICV	TOTAL
Votos	83.636	45.214	40.131	19.446	8.750	197.177
Porcentaje de votos	42,42%	22,93%	20,35%	9,96%	4,44%	100,00%
Solución ideal	6,36	3,44	3,05	1,48	0,67	15,00
Solución correcta	6	3	3	2	1	15
Porcentaje de escaños	40,00%	20,00%	20,00%	13,33%	6,67%	100,00%
Solución d'Hondt	7	4	3	1	0	15
Porcentaje de escaños	46,67%	26,67%	20,00%	6,67%	0	100,00%

Algoritmo para una correcta asignación de escaños (Lleida, 2003)

15 ESCAÑOS	CiU	PSC	ERC	PP	ICV	TOTAL
Votos	83.636	45.214	40.131	19.446	8.750	197.177
Solución ideal	6,36	3,44	3,05	1,48	0,67	15
Límites inferiores	6	3	3	1	0	
Límites superiores	7	4	4	2	1	
Diferencias absolutas inferiores	0,36	0,44	0,05	0,48	0,67	
Diferencias absolutas superiores	0,64	0,56	0,95	0,52	0,33	
Solución correcta	6	3	3	2	1	15

EL PAÍS

Se propone una modificación a la ley electoral para aproximarla al mandato constitucional de representación proporcional

▶ El análisis matemático proporciona la solución correcta

Las diferencias reales entre la solución proporcionada por la ley d'Hondt y la solución correcta pueden ser ilustradas con los resultados correspondientes a la provincia de Lleida en las elecciones autonómicas catalanas del pasado 16 de Noviembre (ver primer gráfico). En ese caso, los votos finalmente obtenidos por los cinco partidos que, habiendo conseguido al menos un 3% de los votos válidos en toda Catalunya, podían optar a representación parlamentaria (CiU, PSC, ERC, PP e ICV) fueron, en ese orden, 83.636, 45.214, 40.131, 19.446 y 8.750, es decir 42,42%, 22,93%, 20,35%, 9,86% y 4,44% del total de votos obtenidos en Lleida por esos cinco partidos. La ley electoral vigente atribuye a Lleida 15 de los 135 escaños del parlamento catalán; para que su distribución fuese exactamente proporcional CiU, PSC, ERC, PP e ICV deberían haber recibido 6,36, 3,44, 3,05, 1,48 y 0,67 escaños respectivamente. Si, de acuerdo con la Constitución, se trata de conseguir una distribución proporcional, ésta sería la solución ideal. El problema técnico consiste en aproximar estos valores por números enteros, para convertir la solución ideal en una solución posible, y hacerlo de forma que el resultado represente una distribución de escaños tan cercana a la distribución de votos como sea posible. Puede demostrarse que la solución correcta es atribuir 6, 3, 3, 2 y 1 escaños a CiU, PSC, ERC, PP e ICV, respectivamente, lo que representa el 40%, 20%, 20%, 13,33% y 6,67% de los 15 escaños.

Puede comprobarse que (como resulta casi evidente a simple vista), cualquiera que sea el criterio de aproximación que se quiera utilizar, la distribución de escaños correspondiente a la solución correcta está notablemente más próxima a la distribución de votos que la distribución de escaños correspondiente a la ley d'Hondt.

▶ La ley d'Hondt no respeta la proporcionalidad y favorece a los partidos mayoritarios

En análisis matemático existen muchas formas diferentes de medir la discrepancia entre dos distribuciones.

Entre las más utilizadas están la distancia euclídea, la distancia de Hellinger y la discrepancia logarítmica. En el caso de Lleida, se ha comprobado que, entre las 3876 formas posibles de distribuir sus 15 escaños entre los cinco partidos con derecho a representación parlamentaria, existen 24 asignaciones mejores que la proporcionada por la Ley d'Hondt, en el sentido de que definen una distribución de escaños más próxima a la distribución de votos para cualquiera de esas medidas de discrepancia. En particular, la solución correcta está 8,1 veces más cerca de la solución ideal que la solución d'Hondt si se utiliza la distancia de Hellinger para medir la proximidad, 4,6 veces más cerca si se utiliza la discrepancia logarítmica, y 1,4 veces más cerca si se utiliza la distancia euclídea.

Las diferencias entre la solución correcta y la ley d'Hondt tienden a desaparecer cuando aumenta el número de escaños a repartir. Por ejemplo, la soluciónd'Hondt para la distribución de los 85 escaños de la provincia de Barcelona en esas mismas elecciones coincide con la distribución correcta. Recíprocamente, las diferencias aumentan cuando el número de escaños a repartir disminuye. Por ejemplo, si sólo se repartiesen 2 escaños entre 2 partidos, la ley d'Hondt asignaría los 2 escaños al partido mayoritario siempre que éste obtuviese al menos dos terceras partes de los votos, mientras que la solución correcta con la distancia euclídea es hacerlo únicamente a partir de las tres cuartas partes, y la solución con la de Hellinger a partir de las cuatro quintas partes. La tendencia de la Ley d'Hondt a distorsionar la voluntad popular en el sentido de favorecer a los partidos mayoritarios resulta evidente.

▶ Fácil determinación de la solución correcta

Para cualquier conjunto de resultados electorales, la solución que minimiza la distancia euclídea (una extensión de la distancia entre dos puntos del plano dada por el teorema de Pitágoras) puede ser encontrada mediante un procedimiento muy sencillo (mucho más fácil que el procedimiento necesario para determinar la solución propuesta por d'Hondt). Como se indica en la Tabla 2 (correspondiente a Lleida 2003), se parte del número de votos obtenidos por cada uno de los partidos con derecho a representación parlamentaria; se determina la solución ideal, repartiendo los escaños correspondientes a la provincia de forma proporcional a los votos obtenidos por cada uno de esos partidos; se especifican sus aproximaciones enteras, es decir los números enteros más cercanos (por defecto y por exceso) a la solución ideal, y se calculan los errores correspondientes a cada una de las aproximaciones enteras (es decir los valores absolutos de sus diferencias con la solución ideal). La solución correcta se obtiene entonces partiendo del más pequeño de los errores absolutos y procediendo por orden, de menor a mayor error, para asignar a cada partido la solución con mínimo error que sea compatible con el número total de escaños que deben ser distribuidos.

El el caso de Lleida (tabla 2), el menor de los errores absolutos es 0,05, que corresponde a asignar 3 escaños a ERC, lo que constituye el primer elemento de la solución. El menor de los errores absolutos correspondientes a los demás partidos es 0,33, que corresponde a asignar 1 escaño a ICV, el segundo elemento de la solución. El menor de los errores restantes es 0,36, que corresponde a asignar 6 escaños a CiU; le sigue 0,44, que corresponde a asignar 3 escaños al PSOE. Como el número total de escaños a asignar es 15, al PP se le deben asignar los 2 escaños restantes (única asignación compatible con las ya realizadas), con lo que se completa la la solución correcta para la distancia euclídea.

En casos extremos, cuando el número de escaños a repartir es muy pequeño, la solución óptima puede depender de la distancia elegida, pero en la práctica, con el número de escaños por circunscripción que se utilizan en España, la solución óptima es independiente de la medida de distancia elegida, y distinta de la que proporciona la ley d'Hondt.

▶ Por respeto a la Constitución, las leyes deberían ser modificadas

La ley electoral define el número total de escaños del Parlamento, su distribución por circunscripciones, el porcentaje mínimo de votos exigido, y el procedimiento utilizado para distribuir las escaños entre los partidos que superan ese umbral. Los tres primeros de estos aspectos deben ser el resultado de una negociación política en la que es necesario valorar argumentos muy diversos. Sin embargo, el último elemento, el procedimiento utilizado para la asignación de escaños, es la solución a un problema matemático, y debe ser discutido en términos matemáticos.

El mandato constitucional de distribuir los escaños de cada circunscripción "atendiendo a criterios de representación proporcional" tiene, para cada función de distancia, una única solución matemáticamente correcta. En la práctica, con el número de escaños que se distribuyen en España en cada circunscripción, la solución no depende del criterio de aproximación que quiera utilizarse. Esta solución óptima es muy fácil de implementar, y no es la que actualmente se utiliza. Por respeto a los ideales democráticos consagrados en la Constitución, nuestras leyes electorales deberían ser adecuadamente modificadas.

José Miguel Bernardo es catedrático de Estadística. Los detalles matemáticos pueden ser consultados en *Proportionality in parliamentary democracy: An alternative to Jefferson-d'Hondt rule.* J. M. Bernardo (2004). Universidad de Valencia.

Reference Analysis

José M. Bernardo [1]

Departamento de Estadística e I.0., Universitat de València, Spain

Abstract

This chapter describes reference analysis, a method to produce Bayesian inferential statements which only depend on the assumed model and the available data. Statistical information theory is used to define the *reference* prior function as a mathematical description of that situation where data would best dominate prior knowledge about the quantity of interest. Reference priors are not descriptions of personal beliefs; they are proposed as formal *consensus* prior functions to be used as standards for scientific communication. Reference posteriors are obtained by formal use of Bayes theorem with a reference prior. Reference prediction is achieved by integration with a reference posterior. Reference decisions are derived by minimizing a reference posterior expected loss. An information theory based loss function, the *intrinsic discrepancy*, may be used to derive reference procedures for conventional inference problems in scientific investigation, such as point estimation, region estimation and hypothesis testing.

Key words: Amount of information, Intrinsic discrepancy, Bayesian asymptotics, Fisher information, Objective priors, Noninformative priors, Jeffreys priors, Reference priors, Maximum entropy, Consensus priors, Intrinsic statistic, Point Estimation, Region Estimation, Hypothesis testing,

1 Introduction and notation

This chapter is mainly concerned with statistical inference problems such as occur in scientific investigation. Those problems are typically solved conditional on the assumption that a particular statistical model is an appropriate description of the probabilistic mechanism which has generated the data, and the choice of that model naturally involves an element of subjectivity. It has become standard practice, however, to describe as "objective" any statistical

Email address: jose.m.bernardo@uv.es (José M. Bernardo).
URL: www.uv.es/~bernardo (José M. Bernardo).

[1] Supported by grant BMF2001-2889 of the MCyT, Madrid, Spain

analysis which only depends on the model assumed and the data observed. In this precise sense (and only in this sense) reference analysis is a method to produce "objective" Bayesian inference.

Foundational arguments (Savage, 1954; de Finetti, 1970; Bernardo and Smith, 1994) dictate that scientists should elicit a unique (joint) prior distribution on all unknown elements of the problem on the basis of available information, and use Bayes theorem to combine this with the information provided by the data, encapsulated in the likelihood function. Unfortunately however, this elicitation is a formidable task, specially in realistic models with many nuisance parameters which rarely have a simple interpretation. Weakly informative priors have here a role to play as approximations to genuine proper prior distributions. In this context, the (unfortunately very frequent) naïve use of simple proper "flat" priors (often a limiting form of a conjugate family) as presumed "noninformative" priors often hides important unwarranted assumptions which may easily dominate, or even invalidate, the analysis: see *e.g.*, Hobert and Casella (1996, 1998), Casella (1996), Palmer and Pettit (1996), Hadjicostas and Berry (1999) or Berger (2000). The uncritical (ab)use of such "flat" priors should be strongly discouraged. An appropriate *reference* prior (see below) should instead be used. With numerical simulation techniques, where a proper prior is often needed, a proper approximation to the *reference* prior may be employed.

Prior elicitation would be even harder in the important case of scientific inference, where some sort of *consensus* on the elicited prior would obviously be required. A fairly natural candidate for such a consensus prior would be a "noninformative" prior, where prior knowledge could be argued to be dominated by the information provided by the data. Indeed, scientific investigation is seldom undertaken unless it is likely to substantially increase knowledge and, even if the scientist holds strong prior beliefs, the analysis would be most convincing to the scientific community if done with a consensus prior which is dominated by the data. Notice that the concept of a "noninformative" prior is *relative* to the information provided by the data.

As evidenced by the long list of references which concludes this chapter, there has been a considerable body of conceptual and theoretical literature devoted to identifying appropriate procedures for the formulation of "noninformative" priors. Beginning with the work of Bayes (1763) and Laplace (1825) under the name of inverse probability, the use of "noninformative" priors became central to the early statistical literature, which at that time was mainly objective Bayesian. The obvious limitations of the principle of insufficient reason used to justify the (by then) ubiquitous uniform priors, motivated the developments of Fisher and Neyman, which overshadowed Bayesian statistics during the first half of the 20th century. The work of Jeffreys (1946) prompted a strong revival of objective Bayesian statistics; the seminal books by Jeffreys (1961), Lindley (1965), Zellner (1971), Press (1972) and Box and Tiao (1973), demonstrated that the conventional textbook problems which frequentist statistics were able

to handle could better be solved from a unifying objective Bayesian perspective. Gradual realization of the fact that no *single* "noninformative" prior could possibly be always appropriate for all inference problems within a given multiparameter model (Dawid, Stone and Zidek, 1973; Efron, 1986) suggested that the long search for a *unique* "noninformative" prior representing "ignorance" within a given model was misguided. Instead, efforts concentrated in identifying, for each particular inference problem, a specific (joint) *reference prior* on all the unknown elements of the problem which would lead to a (marginal) *reference posterior* for the quantity of interest, a posterior which would always be dominated by the information provided by the data (Bernardo, 1979b). As will later be described in detail, statistical information theory was used to provide a precise meaning to this dominance requirement.

Notice that reference priors were *not* proposed as an approximation to the scientist's (unique) personal beliefs, but as a collection of formal *consensus* (not necessarily proper) prior functions which could conveniently be used as standards for scientific communication. As Box and Tiao (1973, p. 23) required, using a reference prior the scientist employs the jury principle; as the jury is carefully screened among people with no connection with the case, so that testimony may be assumed to dominate prior ideas of the members of the jury, the reference prior is carefully chosen to guarantee that the information provided by the data will not be overshadowed by the scientist's prior beliefs.

Reference posteriors are obtained by formal use of Bayes theorem with a reference prior function. If required, they may be used to provide point or region estimates, to test hypothesis, or to predict the value of future observations. This provides a unified set of objective Bayesian solutions to the conventional problems of scientific inference, objective in the precise sense that those solutions only depend on the assumed model and the observed data.

By restricting the class $\mathcal{P}$ of candidate priors, the reference algorithm makes it possible to incorporate into the analysis any genuine prior knowledge (over which scientific consensus will presumably exist). From this point of view, derivation of reference priors may be described as a new, powerful method for *prior elicitation*. Moreover, when subjective prior information is actually specified, the corresponding subjective posterior may be compared with the *reference* posterior—hence its name—to assess the relative importance of the initial opinions in the final inference.

In this chapter, it is assumed that probability distributions may be described through their probability density functions, and no notational distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for observable random vectors (typically data) and bold italic greek fonts for unobservable random vectors (typically parameters); lower case is used for variables and upper case calligraphic for their dominion sets. Moreover, the standard mathematical convention of referring to functions, say $f_{\boldsymbol{x}}$ and $g_{\boldsymbol{x}}$ of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ will be

used throughout. Thus, the conditional probability density of data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given $\boldsymbol{\theta}$ will be represented by either $p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}$ or $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, with $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \geq 0$ and $\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x}\,|\,\boldsymbol{\theta})\, d\boldsymbol{x} = 1$, and the posterior distribution of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given $\boldsymbol{x}$ will be represented by either $p_{\boldsymbol{\theta}\,|\,\boldsymbol{x}}$ or $p(\boldsymbol{\theta}\,|\,\boldsymbol{x})$, with $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \geq 0$ and $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums. Density functions of specific distributions are denoted by appropriate names. Thus, if x is an observable random variable with a normal distribution of mean μ and variance σ^2, its probability density function will be denoted $\mathrm{N}(x\,|\,\mu, \sigma)$. If the posterior distribution of μ is Student with location $\overline{x}$, scale s, and $n-1$ degrees of freedom, its probability density function will be denoted $\mathrm{St}(\mu\,|\,\overline{x},\, s,\, n-1)$.

The reference analysis argument is always defined in terms of some *parametric model* of the general form $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, which describes the conditions under which data have been generated. Thus, data $\boldsymbol{x}$ are assumed to consist of one observation of the random vector $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, with probability density $p(\boldsymbol{x}\,|\,\boldsymbol{\omega})$ for some $\boldsymbol{\omega} \in \boldsymbol{\Omega}$. Often, but not necessarily, data will consist of a random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of fixed size n from some distribution with, say, density $p(\boldsymbol{y}\,|\,\boldsymbol{\omega})$, $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$, in which case $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}) = \prod_{j=1}^{n} p(\boldsymbol{y}_j\,|\,\boldsymbol{\omega})$ and $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}}^n$. In this case, reference priors relative to model $\mathcal{M}$ turn out to be the same as those relative to the simpler model $\mathcal{M}_{\boldsymbol{y}} \equiv \{p(\boldsymbol{y}\,|\,\boldsymbol{\omega}),\, \boldsymbol{y} \in \boldsymbol{\mathcal{Y}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$.

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be some vector of interest; without loss of generality, the assumed model $\mathcal{M}$ may be reparametrized in the form

$$\mathcal{M} \equiv \{\, p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\,\}, \tag{1}$$

where $\boldsymbol{\lambda}$ is some vector of nuisance parameters; this is often simply referred to as "model" $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$. Conditional on the assumed model, all valid Bayesian inferential statements about the value of $\boldsymbol{\theta}$ are encapsulated in its posterior distribution $p(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto \int_{\boldsymbol{\Lambda}}\, p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\, p(\boldsymbol{\theta}, \boldsymbol{\lambda})\, d\boldsymbol{\lambda}$, which *combines* the information provided by the data $\boldsymbol{x}$ with any other information about $\boldsymbol{\theta}$ contained in the prior density $p(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Intuitively, the *reference prior function* for $\boldsymbol{\theta}$, given model $\mathcal{M}$ and a class of candidate priors $\mathcal{P}$, is that (joint) prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\mathcal{M}, \mathcal{P})$ which may be expected to have a minimal effect on the posterior inference about the quantity of interest $\boldsymbol{\theta}$ among the class of priors which belong to $\mathcal{P}$, *relative* to data which could be obtained from $\mathcal{M}$. The reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega}\,|\,\mathcal{M}, \mathcal{P})$ is specifically designed to be a reasonable *consensus* prior (within the class $\mathcal{P}$ of priors compatible with assumed prior knowledge) for inferences about a *particular quantity of interest* $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$, and it is always conditional to the *specific experimental design* $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ which is assumed to have generated the data.

By definition, the reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\mathcal{M}, \mathcal{P})$ is "objective", in the sense that it is a well-defined mathematical function of the vector of interest $\boldsymbol{\theta}$, the assumed model $\mathcal{M}$, and the class $\mathcal{P}$ of candidate priors, with no additional subjective elements. By formal use of Bayes theorem and appropriate integ-

ration (provided the integral is finite), the (joint) reference prior produces a (marginal) *reference posterior* for the vector of interest

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P}) \propto \int_{\boldsymbol{\Lambda}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathcal{M}, \mathcal{P})\, d\boldsymbol{\lambda}, \tag{2}$$

which could be described as a mathematical expression of the inferential content of data $\boldsymbol{x}$ with respect to the value of $\boldsymbol{\theta}$, with no additional knowledge beyond that contained in the assumed statistical model $\mathcal{M}$ and the class $\mathcal{P}$ of candidate priors (which may well consist of the class $\mathcal{P}_0$ of *all* suitably regular priors). To simplify the exposition, the dependence of the reference prior on both the model and the class of candidate priors is frequently dropped from the notation, so that $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ are written instead of $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \mathcal{M}, \mathcal{P})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P})$.

The reference prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ often turns out to be an *improper* prior, *i.e.*, a positive function such that $\int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$ diverges and, hence, cannot be renormalized into a proper density function. Notice that this is not a problem provided the resulting posterior distribution (2) is proper for all suitable data. Indeed the declared objective of reference analysis is to provide appropriate reference *posterior* distributions; reference prior *functions* are merely useful technical devices for a simple computation (via formal use of Bayes theorem) of reference posterior *distributions*. For discussions on the axiomatic foundations which justify the use of improper prior functions, see Hartigan (1983) and references therein.

In the long quest for objective posterior distributions, several requirements have emerged which may reasonably be requested as *necessary* properties of any proposed solution:

(1) *Generality*. The procedure should be completely general, *i.e.*, applicable to any properly defined inference problem, and should produce no untenable answers which could be used as counterexamples. In particular, an objective posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ must be a *proper* probability distribution for any data set $\boldsymbol{x}$ large enough to identify the unknown parameters.

(2) *Invariance*. Jeffreys (1946), Hartigan (1964), Jaynes (1968), Box and Tiao (1973, Sec. 1.3), Villegas (1977b, 1990), Dawid (1983), Yang (1995), Datta and J. K. Ghosh (1995b), Datta and M. Ghosh (1996). For any one-to-one function $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, the posterior $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x})$ obtained from the reparametrized model $p(\boldsymbol{x} \,|\, \boldsymbol{\phi}, \boldsymbol{\lambda})$ must be coherent with the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from the original model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ in the sense that, for any data set $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, $\pi(\boldsymbol{\phi} \,|\, \boldsymbol{x}) = \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, |\, d\boldsymbol{\theta} / d\boldsymbol{\phi}|$. Moreover, if the model has a sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, then the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from the full model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ must be the same as the posterior $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the equivalent model $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$.

(3) *Consistent marginalization.* Stone and Dawid (1972), Dawid, Stone and Zidek (1973), Dawid (1980). If, for all data $\boldsymbol{x}$, the posterior $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ obtained from model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ is of the form $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ for some statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ whose sampling distribution $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{t} \,|\, \boldsymbol{\theta})$ only depends on $\boldsymbol{\theta}$, then the posterior $\pi_2(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the marginal model $p(\boldsymbol{t} \,|\, \boldsymbol{\theta})$ must be the same as the posterior $\pi_1(\boldsymbol{\theta} \,|\, \boldsymbol{t})$ obtained from the original full model.

(4) *Consistent sampling properties.* Neyman and Scott (1948), Stein (1959), Dawid and Stone (1972, 1973), Cox and Hinkley (1974, Sec. 2.4.3), Stone (1976), Lane and Sudderth (1984). The properties under repeated sampling of the posterior distribution must be consistent with the model. In particular, the family of posterior distributions $\{\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}_j),\, \boldsymbol{x}_j \in \boldsymbol{\mathcal{X}}\}$ which could be obtained by repeated sampling from $p(\boldsymbol{x}_j \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})$ should concentrate on a region of $\boldsymbol{\Theta}$ which contains the true value of $\boldsymbol{\theta}$.

Reference analysis, introduced by Bernardo (1979b) and further developed by Berger and Bernardo (1989, 1992a,b,c), appears to be the only available method to derive objective posterior distributions which satisfy all these desiderata. This chapter describes the basic elements of reference analysis, states its main properties, and provides signposts to the huge related literature.

Section 2 summarizes some necessary concepts of discrepancy and convergence, which are based on information theory. Section 3 provides a formal definition of reference distributions, and describes their main properties. Section 4 describes an integrated approach to point estimation, region estimation, and hypothesis testing, which is derived from the joint use of reference analysis and an information-theory based loss function, the *intrinsic discrepancy*. Section 5 provides many additional references for further reading on reference analysis and related topics.

2 Intrinsic discrepancy and expected information

Intuitively, a reference prior for $\boldsymbol{\theta}$ is one which maximizes what it is *not known* about $\boldsymbol{\theta}$, *relative* to what *could* possibly be learnt from repeated observations from a particular model. More formally, a reference prior for $\boldsymbol{\theta}$ is defined to be one which maximizes—within some class of candidate priors—the *missing information* about the quantity of interest $\boldsymbol{\theta}$, defined as a limiting form of the amount of information about its value which repeated data from the assumed model could possibly provide. In this section, the notions of discrepancy, convergence, and expected information—which are required to make these ideas precise—are introduced and illustrated.

Probability theory makes frequent use of *divergence measures* between probability distributions. The total variation distance, Hellinger distance, Kullback-Leibler logarithmic divergence, and Jeffreys logarithmic divergence are fre-

quently cited; see, for example, Kullback (1968, 1983, 1987) for precise definitions and properties. Each of those divergence measures may be used to define a type of convergence. It has been found, however, that the behaviour of many important limiting processes, in both probability theory and statistical inference, is better described in terms of another information-theory related divergence measure, the *intrinsic discrepancy* (Bernardo and Rueda, 2002), which is now defined and illustrated.

Definition 1 (Intrinsic discrepancy) *The* intrinsic discrepancy $\delta\{p_1, p_2\}$ *between two probability distributions of a random vector* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*, specified by their density functions* $p_1(\boldsymbol{x})$*,* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1 \subset \boldsymbol{\mathcal{X}}$*, and* $p_2(\boldsymbol{x})$*,* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2 \subset \boldsymbol{\mathcal{X}}$*, with either identical or nested supports, is*

$$\delta\{p_1, p_2\} = \min\left\{ \int_{\boldsymbol{\mathcal{X}}_1} p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\, d\boldsymbol{x},\ \int_{\boldsymbol{\mathcal{X}}_2} p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\, d\boldsymbol{x} \right\}, \tag{3}$$

provided one of the integrals (or sums) is finite. The intrinsic discrepancy between two parametric models for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*,* $\mathcal{M}_1 \equiv \{p_1(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}_1\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *and* $\mathcal{M}_2 \equiv \{p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}_2\, \boldsymbol{\psi} \in \boldsymbol{\Psi}\}$*, is the minimum intrinsic discrepancy between their elements,*

$$\delta\{\mathcal{M}_1, \mathcal{M}_2\} = \inf_{\boldsymbol{\omega} \in \boldsymbol{\Omega},\, \boldsymbol{\psi} \in \boldsymbol{\Psi}} \delta\{p_1(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, p_2(\boldsymbol{x} \,|\, \boldsymbol{\psi})\}. \tag{4}$$

The *intrinsic discrepancy* is a new element of the class of *intrinsic loss functions* defined by Robert (1996); the concept is *not* related to the concepts of "intrinsic Bayes factors" and "intrinsic priors" introduced by Berger and Pericchi (1996), and reviewed in Pericchi (2005).

Notice that, as one would require, the intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ between two parametric families of distributions $\mathcal{M}_1$ and $\mathcal{M}_2$ does not depend on the particular parametrizations used to describe them. This will be crucial to guarantee the desired invariance properties of the statistical procedures described later.

It follows from Definition 1 that the intrinsic discrepancy between two probability distributions may be written in terms of their two possible Kullback-Leibler *directed divergences* as

$$\delta\{p_2, p_1\} = \min\left\{ k\{p_2 \,|\, p_1\},\ k\{p_1 \,|\, p_2\} \right\} \tag{5}$$

where (Kullback and Leibler, 1951) the $k\{p_j \,|\, p_i\}$'s are the non-negative invariant quantities defined by

$$k\{p_j \,|\, p_i\} = \int_{\boldsymbol{\mathcal{X}}_i} p_i(\boldsymbol{x}) \log \frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\, d\boldsymbol{x}, \quad \text{with} \quad \boldsymbol{\mathcal{X}}_i \subseteq \boldsymbol{\mathcal{X}}_j. \tag{6}$$

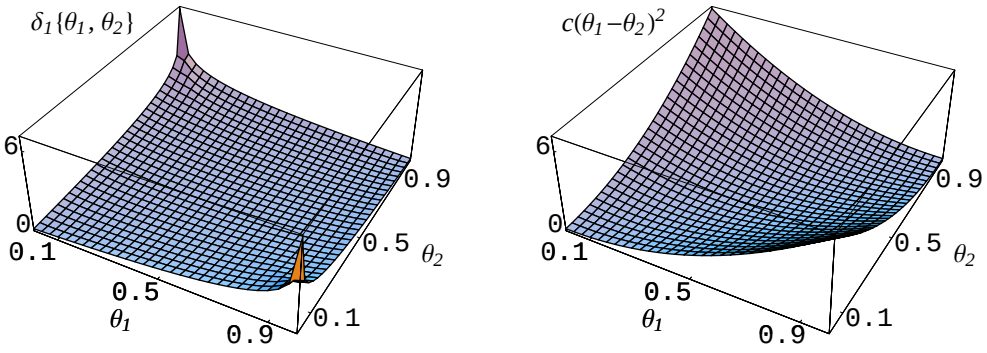
Since $k\{p_j \,|\, p_i\}$ is the expected value of the logarithm of the density (or probability) ratio for p_i against p_j, when p_i is true, it also follows from Definition 1 that, if $\mathcal{M}_1$ and $\mathcal{M}_2$ describe two alternative models, one of which is assumed to generate the data, their intrinsic discrepancy $\delta\{\mathcal{M}_1, \mathcal{M}_2\}$ is the *minimum expected log-likelihood ratio in favour of the model which generates the data* (the "true" model). This will be important in the interpretation of many of the results described in this chapter.

The intrinsic discrepancy is obviously *symmetric*. It is non-negative, vanishes if (and only if) $p_1(\boldsymbol{x}) = p_2(\boldsymbol{x})$ almost everywhere, and it is invariant under one-to-one transformations of $\boldsymbol{x}$. Moreover, if $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ have strictly nested supports, one of the two directed divergences will not be finite, but their intrinsic discrepancy is still defined, and reduces to the other directed divergence. Thus, if $\boldsymbol{\mathcal{X}}_i \subset \boldsymbol{\mathcal{X}}_j$, then $\delta\{p_i, p_j\} = \delta\{p_j, p_i\} = k\{p_j \,|\, p_i\}$.

The intrinsic discrepancy is *information additive*. Thus, if $\boldsymbol{x}$ consists of n independent observations, so that $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ and $p_i(\boldsymbol{x}) = \prod_{j=1}^{n} q_i(\boldsymbol{y}_j)$, then $\delta\{p_1, p_2\} = n\,\delta\{q_1, q_2\}$. This statistically important additive property is essentially unique to logarithmic discrepancies; it is basically a consequence of the fact that the joint density of independent random quantities is the product of their marginals, and the logarithm is the only analytic function which transforms products into sums.

Example 1 *Intrinsic discrepancy between binomial distributions.* The intrinsic discrepancy $\delta\{\theta_1, \theta_2 \,|\, n\}$ between the two binomial distributions

Figure 1 *Intrinsic discrepancy between Bernoulli variables.*



with common value for n, $p_1(r) = \mathrm{Bi}(r \,|\, n, \theta_1)$ and $p_2(r) = \mathrm{Bi}(r \,|\, n, \theta_2)$, is

$$\begin{aligned} \delta\{p_1, p_2\} &= \delta\{\theta_1, \theta_2 \,|\, n\} = n\,\delta_1\{\theta_1, \theta_2\}, \\ \delta_1\{\theta_1, \theta_2\} &= \min[\, k\{\theta_1 \,|\, \theta_2\},\, k\{\theta_2 \,|\, \theta_1\}\,] \\ k(\theta_i \,|\, \theta_j) &= \theta_j \log[\theta_j/\theta_i] + (1-\theta_j)\log[(1-\theta_j)/(1-\theta_i)], \end{aligned} \tag{7}$$

where $\delta_1\{\theta_1, \theta_2\}$ (represented in the left panel of Figure 1) is the intrinsic discrepancy $\delta\{q_1, q_2\}$ between the corresponding Bernoulli distributions,

$q_i(y) = \theta_i^y(1-\theta_i)^{1-y}$, $y \in \{0, 1\}$. It may be appreciated that, specially near the extremes, the behaviour of the intrinsic discrepancy is rather different from that of the conventional quadratic loss $c\,(\theta_1 - \theta_2)^2$ (represented in the right panel of Figure 1 with c chosen to preserve the vertical scale).

As a direct consequence of the information-theoretical interpretation of the Kullback-Leibler directed divergences (Kullback, 1968, Ch. 1), the intrinsic discrepancy $\delta\{p_1, p_2\}$ is a measure, in natural information units or *nits* (Boulton and Wallace, 1970), of the *minimum* amount of expected information, in Shannon (1948) sense, required to discriminate between p_1 and p_2. If base 2 logarithms were used instead of natural logarithms, the intrinsic discrepancy would be measured in binary units of information (*bits*).

The quadratic loss $\ell\{\theta_1, \theta_2\} = (\theta_1 - \theta_2)^2$, often (over)used in statistical inference as measure of the discrepancy between two distributions $p(\boldsymbol{x} \,|\, \theta_1)$ and $p(\boldsymbol{x} \,|\, \theta_2)$ of the same parametric family $\{p(\boldsymbol{x} \,|\, \theta), \theta \in \Theta\}$, heavily depends on the parametrization chosen. As a consequence, the corresponding point estimate, the posterior expectation is not coherent under one-to-one transformations of the parameter. For instance, under quadratic loss, the "best" estimate of the logarithm of some positive physical magnitude is *not* the logarithm of the "best" estimate of such magnitude, a situation hardly acceptable by the scientific community. In sharp contrast to conventional loss functions, the intrinsic discrepancy is invariant under one-to-one reparametrizations. Some important consequences of this fact are summarized below.

Let $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be a family of probability densities, with no nuisance parameters, and let $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ be a possible point estimate of the quantity of interest $\boldsymbol{\theta}$. The intrinsic discrepancy $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}},\, p_{\boldsymbol{x} \,|\, \boldsymbol{\theta}}\}$ between the estimated model and the true model measures, as a function of $\boldsymbol{\theta}$, the loss which would be suffered if model $p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}})$ were used as a proxy for model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$. Notice that this directly measures how different the two *models* are, as opposed to measuring how different their *labels* are, which is what conventional loss functions—like the quadratic loss—typically do. As a consequence, the resulting discrepancy measure is independent of the particular parametrization used; indeed, $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$ provides a natural, *invariant* loss function for estimation, the *intrinsic loss*. The *intrinsic estimate* is that value $\boldsymbol{\theta}^*$ which minimizes $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}\, p(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, the posterior expected intrinsic loss, among all $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$. Since $\delta\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$ is invariant under reparametrization, the intrinsic estimate of any one-to-one transformation of $\boldsymbol{\theta}$, $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, is simply $\boldsymbol{\phi}^* = \boldsymbol{\phi}(\boldsymbol{\theta}^*)$ (Bernardo and Juárez, 2003).

The posterior expected loss function $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ may further be used to define posterior *intrinsic p-credible regions* $R_p = \{\tilde{\boldsymbol{\theta}};\ d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) < k(p)\}$, where $k(p)$ is chosen such that $\Pr[\boldsymbol{\theta} \in R_p \,|\, \boldsymbol{x}] = p$. In contrast to conventional highest posterior density (HPD) credible regions, which do *not* remain HPD under one-to-one transformations of $\boldsymbol{\theta}$, these *lowest posterior loss* (LPL) credible regions *remain* LPL under those transformations.

Similarly, if $\boldsymbol{\theta}_0$ is a parameter value of special interest, the intrinsic discrepancy $\delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\} = \delta\{p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}_0},\, p_{\boldsymbol{x}\,|\,\boldsymbol{\theta}}\}$ provides, as a function of $\boldsymbol{\theta}$, a measure of how far the particular density $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)$ (often referred to as the *null* model) is from the assumed model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, suggesting a natural invariant loss function for precise hypothesis testing. The null model $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)$ will be rejected if the corresponding posterior expected loss (called the *intrinsic statistic*) $d(\boldsymbol{\theta}_0\,|\,\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \delta\{\boldsymbol{\theta}_0, \boldsymbol{\theta}\}\, p(\boldsymbol{\theta}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta}$, is too large. As one should surely require, for any one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, testing whether of not data are compatible with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ yields precisely the same result as testing $\boldsymbol{\phi} = \boldsymbol{\phi}_0 = \boldsymbol{\phi}(\boldsymbol{\theta}_0)$ (Bernardo and Rueda, 2002).

These ideas, extended to include the possible presence of nuisance parameters, will be further analyzed in Section 4.

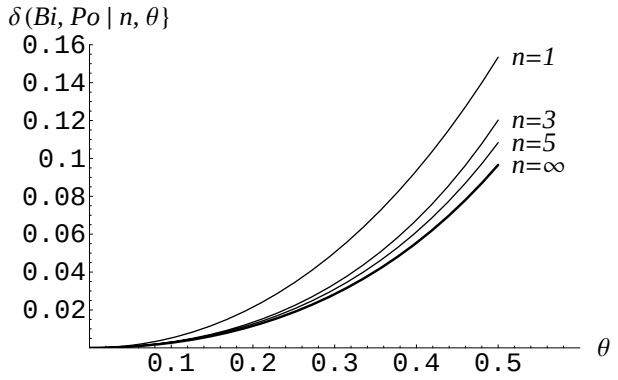
Definition 2 (Intrinsic convergence) *A sequence of probability distributions specified by their density functions* $\{p_i(\boldsymbol{x})\}_{i=1}^{\infty}$ *is said to* converge intrinsically *to a probability distribution with density* $p(\boldsymbol{x})$ *whenever the sequence of their intrinsic discrepancies* $\{\delta(p_i,\, p)\}_{i=1}^{\infty}$ *converges to zero.*

Example 2 *Poisson approximation to a Binomial distribution.* The intrinsic discrepancy between a Binomial distribution with probability function $\mathrm{Bi}(r\,|\,n, \theta)$ and its Poisson approximation $\mathrm{Po}(r\,|\,n\theta)$, is

$$\delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \theta\} = \sum_{r=0}^{n} \mathrm{Bi}(r\,|\,n,\, \theta) \log \frac{\mathrm{Bi}(r\,|\,n,\, \theta)}{\mathrm{Po}(r\,|\,n\theta)} \,,$$

since the second sum in Definition 1 diverges. It may easily be verified that $\lim_{n\to\infty} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \lambda/n\} = 0$ and $\lim_{\theta\to 0} \delta\{\mathrm{Bi}, \mathrm{Po}\,|\,\lambda/\theta,\, \theta\} = 0$; thus, as one would expect from standard probability theory, the sequences of Binomials $\mathrm{Bi}(r\,|\,n, \lambda/n)$ and $\mathrm{Bi}(r\,|\,\lambda/\theta_i,\, \theta_i)$ both intrinsically converge to a Poisson $\mathrm{Po}(r\,|\,\lambda)$ when $n \to \infty$ and $\theta_i \to 0$, respectively.

Figure 2 *Intrinsic discrepancy* $\delta\{\mathrm{Bi}, \mathrm{Po}\,|\,n, \theta\}$ *between a Binomial* $\mathrm{Bi}(r\,|\,n, \theta)$ *and a Poisson* $\mathrm{Po}(r\,|\,n\theta)$ *as a function of* θ*, for* $n = 1, 3, 5$ *and* ∞.



However, if one is interest in approximatiing a binomial $\mathrm{Bi}(r\,|\,n, \theta)$ by a

Poisson $\mathrm{Po}(r \,|\, n\theta)$ the rôles of n and θ are far from similar: the important condition for the Poisson approximation to the Binomial to work is that the value of θ must be small, while the value of n is largely irrelevant. Indeed, (see Figure 2), $\lim_{\theta \to 0} \delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} = 0$, for all $n > 0$, but $\lim_{n \to \infty} \delta\{\mathrm{Bi}, \mathrm{Po} \,|\, n, \theta\} = \frac{1}{2}[-\theta - \log(1-\theta)]$ for all $\theta > 0$. Thus, arbitrarily good approximations are possible with any n, provided θ is sufficiently small. However, for fixed θ, the quality of the approximation cannot improve over a certain limit, no matter how large n might be. For example, $\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, 3, 0.05\} = 0.00074$ and $\delta\{\mathrm{Bi}, \mathrm{Po} \,|\, 5000, 0.05\} = 0.00065$, both yielding an expected log-probability ratio of about 0.0007. Thus, for all $n \geq 3$ the Binomial distribution $\mathrm{Bi}(r \,|\, n,\, 0.05)$ is quite well approximated by the Poisson distribution $\mathrm{Po}(r \,|\, 0.05n)$, and the quality of the approximation is very much the same for any value n.

Many standard approximations in probability theory may benefit from an analysis similar to that of Example 2. For instance, the sequence of Student distributions $\{\mathrm{St}(x \,|\, \mu, \sigma, \nu)\}_{\nu=1}^{\infty}$ converges intrinsically to the normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$ with the same location and scale parameters, and the discrepancy $\delta(\nu) = \delta\{\mathrm{St}(x \,|\, \mu, \sigma, \nu), \mathrm{N}(x \,|\, \mu, \sigma)\}$ (which only depends on the degrees of freedom ν) is smaller than 0.001 when $\nu > 40$. Thus approximating a Student with more than 40 degrees of freedom by a normal yields an expected log-density ratio smaller than 0.001, suggesting quite a good approximation.

As mentioned before, a reference prior is often an improper prior function. Justification of its use as a *formal* prior in Bayes theorem to obtain a reference posterior necessitates proving that the reference posterior thus obtained is an appropriate limit of a sequence of posteriors obtained from proper priors.

Theorem 1 *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. *If* $\pi(\boldsymbol{\omega})$ *is a strictly positive improper prior,* $\{\boldsymbol{\Omega}_i\}_{i=1}^{\infty}$ *is an increasing sequence of subsets of the parameter space which converges to* $\boldsymbol{\Omega}$ *and such that* $\int_{\boldsymbol{\Omega}_i} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$, *and* $\pi_i(\boldsymbol{\omega})$ *is the renormalized proper density obtained by restricting* $\pi(\boldsymbol{\omega})$ *to* $\boldsymbol{\Omega}_i$, *then, for any data set* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, *the sequence of the corresponding posteriors* $\{\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ *converges intrinsically to the posterior* $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$ *obtained by formal use of Bayes theorem with the improper prior* $\pi(\boldsymbol{\omega})$.

However, to avoid possible pathologies, a stronger form of convergence is needed; for a sequence of proper priors $\{\pi_i\}_{i=1}^{\infty}$ to converge to a (possibly improper) prior function π, it will further be required that the *predicted* intrinsic discrepancy between the corresponding posteriors converges to zero. For a motivating example, see Berger and Bernardo (1992c, p. 43), where the model

$$\Big\{ p(x \,|\, \theta) = \tfrac{1}{3}, \quad x \in \{[\tfrac{\theta}{2}], 2\theta, 2\theta + 1\}, \quad \theta \in \{1, 2, \ldots\} \Big\},$$

where $[u]$ denotes the integer part of u (and $[\frac{1}{2}]$ is separately defined as 1), originally proposed by Fraser, Monette and Ng (1985), is reanalysed.

Definition 3 (Permissible prior function) *A positive function* $\pi(\boldsymbol{\omega})$ *is an* permissible *prior function for model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *if for all* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ *one has* $\int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$*, and for some increasing sequence* $\{\boldsymbol{\Omega}_i\}_{i=1}^{\infty}$ *of subsets of* $\boldsymbol{\Omega}$*, such that* $\lim_{i \to \infty} \boldsymbol{\Omega}_i = \boldsymbol{\Omega}$*, and* $\int_{\boldsymbol{\Omega}_i} \pi(\boldsymbol{\omega})\, d\boldsymbol{\omega} < \infty$*,*

$$\lim_{i \to \infty} \int_{\boldsymbol{\mathcal{X}}} p_i(\boldsymbol{x})\; \delta\{\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x}),\, \pi(\boldsymbol{\omega} \,|\, \boldsymbol{x})\}\, d\boldsymbol{x} = 0,$$

where $\pi_i(\boldsymbol{\omega})$ *is the renormalized restriction of* $\pi(\boldsymbol{\omega})$ *to* $\boldsymbol{\Omega}_i$*,* $\pi_i(\boldsymbol{\omega} \,|\, \boldsymbol{x})$ *is the corresponding posterior,* $p_i(\boldsymbol{x}) = \int_{\boldsymbol{\Omega}_i} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi_i(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ *is the corresponding predictive, and* $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \pi(\boldsymbol{\omega})$*.*

In words, $\pi(\boldsymbol{\omega})$ is a permissible prior function for model $\mathcal{M}$ if it always yields proper posteriors, and the sequence of the *predicted* intrinsic discrepancies between the corresponding posterior $\pi(\boldsymbol{\omega} \,|\, \boldsymbol{x})$ and its renormalized restrictions to $\boldsymbol{\Omega}_i$ converges to zero for some suitable approximating sequence of the parameter space. All proper priors are permissible in the sense of Definition 3, but improper priors may or may not be permissible, even if they seem to be arbitrarily close to proper priors.

Example 3 *Exponential model.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $p(x \,|\, \theta) = \theta e^{-\theta\, x}$, $\theta > 0$, so that $p(\boldsymbol{x} \,|\, \theta) = \theta^n e^{-\theta\, t}$, with sufficient statistic $t = \sum_{j=1}^{n} x_j$. Consider a positive function $\pi(\theta) \propto \theta^{-1}$, so that $\pi(\theta \,|\, t) \propto \theta^{n-1} e^{-\theta\, t}$, a gamma density $\mathrm{Ga}(\theta \,|\, n, t)$, which is a proper distribution for all possible data sets. Take now some sequence of pairs of positive real numbers $\{a_i, b_i\}$, with $a_i < b_i$, and let $\Theta_i = (a_i, b_i)$; the intrinsic discrepancy between $\pi(\theta \,|\, t)$ and its renormalized restriction to Θ_i, denoted $\pi_i(\theta \,|\, t)$, is $\delta_i(n, t) = k\{\pi(\theta \,|\, t) \,|\, \pi_i(\theta \,|\, t)\} = \log[c_i(n, t)]$, where $c_i(n, t) = \Gamma(n)/\{\Gamma(n, a_i\, t) - \Gamma(n, b_i\, t)\}$. The renormalized restriction of $\pi(\theta)$ to Θ_i is $\pi_i(\theta) = \theta^{-1}/\log[b_i/a_i]$, and the corresponding (prior) predictive of t is $p_i(t \,|\, n) = c_i^{-1}(n, t)\; t^{-1}/\log[b_i/a_i]$. It may be verified that, for all $n \geq 1$, the expected intrinsic discrepancy $\int_0^{\infty} p_i(t \,|\, n)\, \delta_i(n, t)\, dt$ converges to zero as $i \to \infty$. Hence, all positive functions of the form $\pi(\theta) \propto \theta^{-1}$ are permissible priors for the parameter of an exponential model.

Example 4 *Mixture model.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from $\mathcal{M} \equiv \{\frac{1}{2}\mathrm{N}(x \,|\, \theta, 1) + \frac{1}{2}\mathrm{N}(x \,|\, 0, 1),\, x \in \mathbb{R},\, \theta \in \mathbb{R}\}$. It is easily verified that the likelihood function $p(\boldsymbol{x} \,|\, \theta) = \prod_{j=1}^{n} p(x_j \,|\, \theta)$ is always bounded below by a strictly positive function of $\boldsymbol{x}$. Hence, $\int_{-\infty}^{\infty} p(\boldsymbol{x} \,|\, \theta)\, d\theta = \infty$ for all $\boldsymbol{x}$, and the "natural" objective uniform prior function $\pi(\theta) = 1$ is obviously *not* permissible, although it may be pointwise arbitrarily well approximated by a sequence of proper "flat" priors.

Definition 4 (Intrinsic association) *The* intrinsic association $\alpha_{\boldsymbol{x}\boldsymbol{y}}$ *between two random vectors* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ *and* $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ *with joint density* $p(\boldsymbol{x}, \boldsymbol{y})$ *and marginals* $p(\boldsymbol{x})$ *and* $p(\boldsymbol{y})$ *is the intrinsic discrepancy* $\alpha_{\boldsymbol{x}\boldsymbol{y}} = \delta\{p_{\boldsymbol{x}\,\boldsymbol{y}},\, p_{\boldsymbol{x}} p_{\boldsymbol{y}}\}$ *between their joint density and the product of their marginals.*

The intrinsic association is a non-negative invariant measure of association between two random vectors, which vanishes if they are independent, and tends to infinity as $\boldsymbol{y}$ and $\boldsymbol{x}$ approach a functional relationship. If their joint distribution is bivariate normal, it reduces to $-\frac{1}{2}\log(1-\rho^2)$, a simple function of their coefficient of correlation ρ.

The concept of intrinsic association extends that of *mutual information*; see *e.g.*, Cover and Thomas (1991), and references therein. Important differences arise in the context of contingency tables, where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are discrete random variables which may only take a finite number of different values.

Definition 5 (Expected intrinsic information) *The* expected intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ *from one observation of* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ *about the value of* $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ *when the prior density is* $p(\boldsymbol{\omega})$*, is the intrinsic association* $\alpha_{\boldsymbol{x}\boldsymbol{\omega}} = \delta\{p_{\boldsymbol{x}\,\boldsymbol{\omega}},\ p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\}$ *between* $\boldsymbol{x}$ *and* $\boldsymbol{\omega}$*, where* $p(\boldsymbol{x}, \boldsymbol{\omega}) = p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, p(\boldsymbol{\omega})$*, and* $p(\boldsymbol{x}) = \int_{\boldsymbol{\Omega}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$.

For a fixed model $\mathcal{M}$, the expected intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ is a concave, positive functional of the prior $p(\boldsymbol{\omega})$. Under appropriate regularity conditions, in particular when data consists of a *large* random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ from some model $\{p(\boldsymbol{y} \,|\, \boldsymbol{\omega}),\, \boldsymbol{y} \in \boldsymbol{\mathcal{Y}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, one has

$$\int \int_{\boldsymbol{\mathcal{X}} \times \boldsymbol{\Omega}} [p(\boldsymbol{x}) p(\boldsymbol{\omega}) + p(\boldsymbol{x}, \boldsymbol{\omega})]\, \log \frac{p(\boldsymbol{x})\, p(\boldsymbol{\omega})}{p(\boldsymbol{x}, \boldsymbol{\omega})}\, d\boldsymbol{x}\, d\boldsymbol{\omega} \geq 0 \tag{8}$$

so that $k\{p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}\,\boldsymbol{\omega}}\} \leq k\{p_{\boldsymbol{x}\,\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\}$. If this is the case,

$$\begin{aligned} I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\} &= \delta\{p_{\boldsymbol{x}\,\boldsymbol{\omega}},\ p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}}\} = k\{p_{\boldsymbol{x}}\, p_{\boldsymbol{\omega}} \,|\, p_{\boldsymbol{x}\,\boldsymbol{\omega}}\} \\ &= \int \int_{\boldsymbol{\mathcal{X}} \times \boldsymbol{\Omega}} p(\boldsymbol{x}, \boldsymbol{\omega})\, \log \frac{p(\boldsymbol{x}, \boldsymbol{\omega})}{p(\boldsymbol{x})\, p(\boldsymbol{\omega})}\, d\boldsymbol{x}\, d\boldsymbol{\omega} \qquad (9) \\ &= \int_{\boldsymbol{\Omega}} p(\boldsymbol{\omega}) \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, \log \frac{p(\boldsymbol{\omega} \,|\, \boldsymbol{x})}{p(\boldsymbol{\omega})}\, d\boldsymbol{x}\, d\boldsymbol{\omega} \qquad (10) \\ &= H[p_{\boldsymbol{\omega}}] - \int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x})\, H[p_{\boldsymbol{\omega} \,|\, \boldsymbol{x}}]\, d\boldsymbol{x}, \qquad (11) \end{aligned}$$

where $H[p_{\boldsymbol{\omega}}] = -\int_{\boldsymbol{\Omega}} p(\boldsymbol{\omega}) \log p(\boldsymbol{\omega})\, d\boldsymbol{\omega}$ is the *entropy* of $p_{\boldsymbol{\omega}}$, and the expected intrinsic information reduces to the Shannon's expected information (Shannon, 1948; Lindley, 1956; Stone, 1959; de Waal and Groenewald, 1989; Clarke and Barron, 1990).

For any fixed model $\mathcal{M}$, the intrinsic information $I\{p_{\boldsymbol{\omega}} \,|\, \mathcal{M}\}$ measures, as a functional of the prior $p_{\boldsymbol{\omega}}$, the amount of information about the value of $\boldsymbol{\omega}$ which one observation $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ may be expected to provide. The stronger the prior knowledge described by $p_{\boldsymbol{\omega}}$, the smaller the information the data may be expected to provide; conversely, weak initial knowledge about $\boldsymbol{\omega}$ will correspond to large expected information from the data. This is the intuitive basis for the definition of a reference prior.

3 Reference distributions

Let $\boldsymbol{x}$ be one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be some vector of interest, whose posterior distribution is required. Notice that $\boldsymbol{x}$ represents the *complete* available data; often, but not always, this will consist of a random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ of fixed size n from some simpler model. Let $\mathcal{P}$ be the *class of candidate priors* for $\boldsymbol{\omega}$, defined as those sufficiently regular priors which are compatible with whatever agreed "objective" initial information about the value of $\boldsymbol{\omega}$ one is willing to assume. A permissible prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ is desired which may be expected to have a minimal effect (in a sense to be made precise) among all priors in $\mathcal{P}$, on the posterior inferences about $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ which could be derived given data generated from $\mathcal{M}$. This will be named a *reference prior function* of $\boldsymbol{\omega}$ for the quantity of interest $\boldsymbol{\theta}$, relative to model $\mathcal{M}$ and class $\mathcal{P}$ of candidate priors, and will be denoted by $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$. The reference prior function $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ will then be used as a formal prior density to derive the required *reference posterior* distribution of the quantity of interest, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}, \mathcal{M}, \mathcal{P})$, via Bayes theorem and the required probability operations.

This section contains the definition and basic properties of reference distributions. The ideas are first formalized in one-parameter models, and then extended to multiparameter situations. Special attention is devoted to *restricted* reference distributions, where the class of candidate priors $\mathcal{P}$ consists of those which satisfy some set of assumed conditions. This provides a continuous collection of solutions, ranging from situations with no assumed prior information on the quantity of interest, when $\mathcal{P}$ is the class $\mathcal{P}_0$ of *all* sufficiently regular priors, to situations where accepted prior knowledge is sufficient to specify a unique prior $p_0(\boldsymbol{\omega})$, so that $\pi^{\boldsymbol{\theta}}(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P}) = p_0(\boldsymbol{\theta})$, the situation commonly assumed in Bayesian subjective analysis.

3.1 One parameter models

Let $\theta \in \Theta \subset \mathbb{R}$ be a real-valued quantity of interest, and let available data $\boldsymbol{x}$ consist of one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, so that there are no nuisance parameters. A permissible prior function $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ in a class $\mathcal{P}$ is desired with a minimal expected effect on the posteriors of θ which could be obtained after data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ generated from $\mathcal{M}$ have been observed.

Let $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ consist of k conditionally independent (given θ) observations from $\mathcal{M}$, so that $\boldsymbol{x}^{(k)}$ consists of one observation from the product model $\mathcal{M}^k = \{ \prod_{j=1}^k p(\boldsymbol{x}_j \,|\, \theta),\, \boldsymbol{x}_j \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta \}$. Let p_θ be a prior distribution for the quantity of interest, and consider the intrinsic information about θ, $I\{p_\theta \,|\, \mathcal{M}^k\}$, which could be expected from the vector $\boldsymbol{x}^{(k)} \in \boldsymbol{\mathcal{X}}^k$. For any sufficiently regular prior p_θ, the posterior distribution of θ would concentrate on its true value as k increases and therefore, as $k \to \infty$, the true value of θ

would get to be precisely known. Thus, as $k \to \infty$, the functional $I\{p_\theta \,|\, \mathcal{M}^k\}$ will approach a precise measure of the amount of *missing information* about θ which corresponds to the prior p_θ. It is natural to define the reference prior as that prior function $\pi^\theta = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ which *maximizes the missing information* about the value of θ within the class $\mathcal{P}$ of candidate priors.

Under regularity conditions, the expected intrinsic information $I\{p_\theta \,|\, \mathcal{M}^k\}$ becomes, for large k, Shannon's expected information and hence, using (11),

$$I\{p_\theta \,|\, \mathcal{M}^k\} = H[p_\theta] - \int_{\mathcal{X}^k} p(\boldsymbol{x}^{(k)})\, H[p_{\theta \,|\, \boldsymbol{x}^{(k)}}]\, d\boldsymbol{x}^{(k)}, \tag{12}$$

where $H[p_\theta] = -\int_\Theta p(\theta) \log p(\theta)\, d\theta$, is the *entropy* of p_θ. It follows that, when the parameter space $\Theta = \{\theta_1, \ldots, \theta_m\}$ is finite, the missing information which corresponds to any strictly positive prior p_θ is, *for any model* $\mathcal{M}$,

$$\lim_{k \to \infty} I\{p_\theta \,|\, \mathcal{M}^k\} = H[p_\theta] = -\sum\nolimits_{j=1}^{m} p(\theta_j) \log p(\theta_j), \tag{13}$$

since, as $k \to \infty$, the discrete posterior probability function $p(\theta \,|\, \boldsymbol{x}^{(k)})$ converges to a degenerate distribution with probability one on the true value of θ and zero on all others, and thus, the posterior entropy $H[p_{\theta \,|\, \boldsymbol{x}^{(k)}}]$ converges to zero. Hence, in finite parameter spaces, the reference prior for the parameter does not depend on the precise form of the model, and it is precisely that which *maximizes the entropy* within the class $\mathcal{P}$ of candidate priors. This was the solution proposed by Jaynes (1968), and it is often used in mathematical physics. In particular, if the class of candidate priors is the class $\mathcal{P}_0$ of *all* strictly positive probability distributions, the reference prior for θ is a uniform distribution over Θ, the "noninformative" prior suggested by the old insufficient reason argument (Laplace, 1812). For further information on the concept of *maximum entropy*, see Jaynes (1968, 1982, 1985, 1989), Akaike (1977), Csiszár (1985, 1991), Clarke and Barron (1994), Grünwald and Dawid (2004), and references therein.

In the continuous case, however, $I\{p_\theta \,|\, \mathcal{M}^k\}$ typically diverges as $k \to \infty$, since an infinite amount of information is required to know exactly the value of a real number. A general definition of the reference prior (which includes the finite case as a particular case), is nevertheless possible as an appropriate limit, when $k \to \infty$, of the sequence of priors maximizing $I\{p_\theta \,|\, \mathcal{M}^k\}$ within the class $\mathcal{P}$. Notice that this limiting procedure is *not* some kind of asymptotic approximation, but an essential element of the *concept* of a reference prior. Indeed, the reference prior is defined to maximize the *missing information* about the quantity of interest which *could* be obtained by repeated sampling from $\mathcal{M}$ (not just the information expected from a finite data set), and this is precisely achieved by maximizing the expected information from the arbitrarily large data set which could be obtained by unlimited repeated sampling from the assumed model.

Since $I\{p_\theta \,|\, \mathcal{M}^k\}$ is only defined for proper priors, and $I\{p_\theta \,|\, \mathcal{M}^k\}$ is not guaranteed to attain its maximum at a proper prior, the formal definition of a reference prior is stated as a limit, as $i \to \infty$, of the sequence of solutions obtained for restrictions $\{\Theta_i\}_{i=1}^{\infty}$ of the parameter space chosen to ensure that the maximum of $I\{p_\theta \,|\, \mathcal{M}^k\}$ is actually obtained at a proper prior. The definition below (Berger, Bernardo and Sun, 2005) generalizes those in Bernardo (1979b) and Berger and Bernardo (1992c), and addresses the problems described in Berger, Bernardo and Mendoza (1989).

Definition 6 (One-parameter reference priors) *Consider the one-parameter model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta \subset \mathbb{R}\}$*, and let* $\mathcal{P}$ *be a class of candidate priors for* θ*. The positive function* $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ *is a* reference prior *for model* $\mathcal{M}$ *given* $\mathcal{P}$ *if it is a permissible prior function such that, for some increasing sequence* $\{\Theta_i\}_{i=1}^{\infty}$ *with* $\lim_{i\to\infty} \Theta_i = \Theta$ *and* $\int_{\Theta_i} \pi(\theta)\, d\theta < \infty$*,*

$$\lim_{k\to\infty} \{I\{\pi_i \,|\, \mathcal{M}^k\} - I\{p_i \,|\, \mathcal{M}^k\}\} \geq 0, \quad \textit{for all } \Theta_i, \textit{ for all } p \in \mathcal{P},$$

where $\pi_i(\theta)$ *and* $p_i(\theta)$ *are the renormalized restrictions of* $\pi(\theta)$ *and* $p(\theta)$ *to* Θ_i*.*

Notice that Definition 6 involves two rather different limiting processes. The limiting process of the Θ_i's towards the whole parameter space Θ is only required to guarantee the existence of the expected informations; this may often (but not always) be avoided if the parameter space is (realistically) chosen to be some finite interval $[a, b]$. On the other hand, the limiting process as $k \to \infty$ is an *essential* part of the definition. Indeed, the reference prior is *defined* as that prior function which maximizes the *missing* information, which is the expected discrepancy between prior knowledge and *perfect* knowledge; but perfect knowledge is only approached *asymptotically*, as $k \to \infty$.

Definition 6 implies that reference priors *only* depend on the *asymptotic behaviour* of the assumed model, a feature which greatly simplifies their actual derivation; to obtain a reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ for the parameter θ of model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$, it is both necessary and sufficient to establish the asymptotic behaviour of its posterior distribution under (conceptual) repeated sampling from $\mathcal{M}$, that is the limiting form, as $k \to \infty$, of the posterior density (or probability function) $\pi(\theta \,|\, \boldsymbol{x}^{(k)}) = \pi(\theta \,|\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$.

As one would hope, Definition 6 yields the maximum entropy result in the case where the parameter space is finite and the quantity of interest is the actual value of the parameter:

Theorem 2 (Reference priors with finite parameter space) *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \theta \in \Theta\}$*, with a* finite *parameter space* $\Theta = \{\theta_1, \ldots, \theta_m\}$ *and such that, for all pairs* θ_i *and* θ_j*,* $\delta\{p_{x\,|\,\theta_i},\ p_{x\,|\,\theta_j}\} > 0$*, and let* $\mathcal{P}$ *be a class of probability distributions over* Θ*. Then the reference prior for the parameter* θ *is*

$$\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P}) = \arg\max_{p_\theta \in \mathcal{P}} H\{p_\theta\},$$

where $p_\theta = \{p(\theta_1), p(\theta_2), \ldots, p(\theta_m)\}$ *and* $H\{p_\theta\} = -\sum_{j=1}^m p(\theta_j) \log p(\theta_j)$ *is the* entropy *of* p_θ*. In particular, if the class of candidate priors for* θ *is the set* $\mathcal{P}_0$ *of all strictly positive probability distributions over* Θ*, then the reference prior is the uniform distribution* $\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \{1/m, \ldots, 1/m\}$.

Theorem 2 follows immediately from the fact that, if the intrinsic discrepancies $\delta\{p_{x\,|\,\theta_i},\ p_{x\,|\,\theta_j}\}$ are all positive (and hence the m models $p(x \,|\, \theta_i)$ are all distinguishable from each other), then the posterior distribution of θ asymptotically converges to a degenerate distribution with probability one on the true value of θ (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3) and references therein). Such asymptotic posterior has zero entropy and thus, by Equation 12, the missing information about θ when the prior is p_θ does not depend on $\mathcal{M}$, and is simply given by the prior entropy, $H\{p_\theta\}$. □

Consider now a model $\mathcal{M}$ indexed by a continuous parameter $\theta \in \Theta \subset \mathbb{R}$. If the family of candidate priors consist of the class $\mathcal{P}_0$ of *all* continuous priors with support Θ, then the reference prior, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ may be obtained as the result of an explicit limit. This provides a relatively simple procedure to obtain reference priors in models with one continuous parameter. Moreover, this analytical procedure may easily be converted into a programmable algorithm for numerical derivation of reference distributions. The results may conveniently be described in terms of any *asymptotically sufficient* statistic, *i.e.*, a function $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)})$ such that, for all θ and for all $\boldsymbol{x}^{(k)}$, $\lim_{k\to\infty}[p(\theta \,|\, \boldsymbol{x}^{(k)})/p(\theta \,|\, \boldsymbol{t}_k)] = 1$.

Obviously, the entire sample $\boldsymbol{x}^{(k)}$ is sufficient (and hence asymptotically sufficient), so there is no loss of generality in framing the results in terms of asymptotically sufficient statistics.

Theorem 3 (Explicit form of the reference prior) *Consider the model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta \subset \mathbb{R}\}$*, and let* $\mathcal{P}_0$ *be the class of all continuous priors with support* Θ*. Let* $\boldsymbol{x}^{(k)} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ *consist of* k *independent observations from* $\mathcal{M}$*, so that* $p(\boldsymbol{x}^{(k)} \,|\, \theta) = \prod_{j=1}^k p(\boldsymbol{x}_j \,|\, \theta)$*, and let* $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)}) \in \boldsymbol{\mathcal{T}}$ *be any asymptotically sufficient statistic. Let* $h(\theta)$ *be a continuous strictly positive function such that, for sufficiently large* k*,* $\int_\Theta p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)\, d\theta < \infty$*, and define*

$$f_k(\theta) = \exp\left\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\ \log\left(\frac{p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)}{\int_\Theta p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)\, d\theta} \right) d\boldsymbol{t}_k \right\}, \quad \textit{and} \tag{14}$$

$$f(\theta) = \lim_{k\to\infty} \frac{f_k(\theta)}{f_k(\theta_0)}\ , \tag{15}$$

where θ_0 *is any interior point of* Θ*. If* $f(\theta)$ *is a permissible prior function then, for any* $c > 0$*,* $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = c\, f(\theta)$ *is a reference prior.*

Intuitively, Theorem 3 states that the reference prior $\pi(\theta \,|\, \mathcal{M})$ relative to model $\mathcal{M}$ only depends on the asymptotic behaviour of the model and that, with no additional information to restrict the class of candidate priors, it has (from Equation 14), the form

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto \exp\Big\{ E_{\boldsymbol{t}_k \,|\, \theta} \Big[\log\, p(\theta \,|\, \boldsymbol{t}_k) \Big] \Big\}, \tag{16}$$

where $p(\theta \,|\, \boldsymbol{t}_k)$ is any asymptotic approximation to the posterior distribution of θ, and the expectation is taken with respect to the sampling distribution of the relevant asymptotically sufficient statistic $\boldsymbol{t}_k = \boldsymbol{t}_k(\boldsymbol{x}^{(k)})$. A heuristic derivation of Theorem 3 is provided below. For a precise statement of the regularity conditions and a formal proof, see Berger, Bernardo and Sun (2005).

Under fairly general regularity conditions, the intrinsic expected information reduces to Shannon's expected information when $k \to \infty$. Thus, starting from (10), the amount of information about θ to be expected from $\mathcal{M}^k$ when the prior is $p(\theta)$ may be rewritten as $I\{p_\theta \,|\, \mathcal{M}^k\} = \int_\Theta p(\theta)\, \log[h_k(\theta)/p(\theta)]\, d\theta$, where $h_k(\theta) = \exp\{\int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log p(\theta \,|\, \boldsymbol{t}_k)\, d\boldsymbol{t}_k\}$. If $c_k = \int_\Theta h_k(\theta)\, d\theta < \infty$, then $h_k(\theta)$ may be renormalized to get the proper density $h_k(\theta)/c_k$, and $I\{p_\theta \,|\, \mathcal{M}^k\}$ may be rewritten as

$$I\{p_\theta \,|\, \mathcal{M}^k\} = \log c_k - \int_\Theta p(\theta) \log \frac{p(\theta)}{h_k(\theta)/c_k}\, d\theta. \tag{17}$$

But the integral in (17) is the Kullback-Leibler directed divergence of $h_k(\theta)/c_k$ from $p(\theta)$, which is non-negative, and it is zero iff $p(\theta) = h_k(\theta)/c_k$ almost everywhere. Thus, $I\{p_\theta \,|\, \mathcal{M}^k\}$ would be maximized by a prior $\pi_k(\theta)$ which satisfies the functional equation

$$\pi_k(\theta) \propto h_k(\theta) = \exp\Big\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log \pi_k(\theta \,|\, \boldsymbol{t}_k)\, d\boldsymbol{t}_k \Big\}, \tag{18}$$

where $\pi_k(\theta \,|\, \boldsymbol{t}_k) \propto p(\boldsymbol{t}_k \,|\, \theta)\, \pi_k(\theta)$ and, therefore, the reference prior should be a limiting form, as $k \to \infty$ of the sequence of proper priors given by (18). This only provides an implicit solution, since the posterior density $\pi_k(\theta \,|\, \boldsymbol{t}_k)$ in the right hand side of (18) obviously depends on the prior $\pi_k(\theta)$; however, as $k \to \infty$, the posterior $\pi_k(\theta \,|\, \boldsymbol{t}_k)$ will approach its asymptotic form which, under the assumed conditions, is independent of the prior. Thus, the posterior density in (18) may be replaced by the posterior $\pi^0(\theta \,|\, \boldsymbol{t}_k) \propto p(\boldsymbol{t}_k \,|\, \theta)\, h(\theta)$ which corresponds to any fixed prior, say $\pi^0(\theta) = h(\theta)$, to obtain an explicit expression for a sequence of priors,

$$\pi_k(\theta) \propto f_k(\theta) = \exp\Big\{ \int_{\boldsymbol{\mathcal{T}}} p(\boldsymbol{t}_k \,|\, \theta)\, \log \pi^0(\theta \,|\, \boldsymbol{t}_k)\, d\boldsymbol{t}_k \Big\}, \tag{19}$$

whose limiting form will still maximize the missing information about θ. The preceding argument rests however on the assumption that (at least for suffi-

ciently large k) the integrals in Θ of $f_k(\theta)$ are finite, but those integrals may well diverge. The problem is solved by considering an increasing sequence $\{\Theta_i\}_{i=1}^{\infty}$ of subsets of Θ which converges to Θ and such that, for all i and sufficiently large k, $c_{ik} = \int_{\Theta_i} f_k(\theta)\, d\theta < \infty$, so that the required integrals are finite. An appropriate limiting form of the double sequence $\pi_{ik}(\theta) = f_k(\theta)/c_{ik}$, $\theta \in \Theta_i$ will then approach the required reference prior.

Such a limiting form is easily established; indeed, let $\pi_{ik}(\theta \,|\, \boldsymbol{x})$, $\theta \in \Theta_i$ be the posterior which corresponds to $\pi_{ik}(\theta)$ and, for some interior point θ_0 of all the Θ_i's, consider the limit

$$\lim_{k\to\infty} \frac{\pi_{ik}(\theta \,|\, \boldsymbol{x})}{\pi_{ik}(\theta_0 \,|\, \boldsymbol{x})} = \lim_{k\to\infty} \frac{p(\boldsymbol{x} \,|\, \theta)\, f_k(\theta)}{p(\boldsymbol{x} \,|\, \theta_0)\, f_k(\theta_0)} \propto p(\boldsymbol{x} \,|\, \theta)\, f(\theta), \tag{20}$$

where $f(\theta) = \lim_{k\to\infty} f_k(\theta)/f_k(\theta_0)$, which does not depend on the initial function $h(\theta)$ (and therefore $h(\theta)$ may be chosen by mathematical convenience). It follows from (20) that, for any data $\boldsymbol{x}$, the sequence of posteriors $\pi_{ik}(\theta \,|\, \boldsymbol{x})$ which maximize the missing information will approach the posterior $\pi(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, f(\theta)$ obtained by formal use of Bayes theorem, using $f(\theta)$ as the prior. This completes the heuristic justification of Theorem 3. □

3.2 Main properties

Reference priors enjoy many attractive properties, as stated below. For detailed proofs, see Bernardo and Smith (1994, Secs. 5.4 and 5.6).

In the frequently occurring situation where the available data consist of a random sample of fixed size n from some model $\mathcal{M}$ (so that the assumed model is $\mathcal{M}^n$), the reference prior relative to $\mathcal{M}^n$ is independent of n, and may simply be obtained as the reference prior relative to $\mathcal{M}$, assuming the later exists.

Theorem 4 (Independence of sample size) *If data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ consists of a random sample of size n from model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta\}$, with reference prior $\pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P})$ relative to the class of candidate priors $\mathcal{P}$, then, for any fixed sample size n, the reference prior for θ relative to $\mathcal{P}$ is $\pi^\theta(\theta \,|\, \mathcal{M}^n, \mathcal{P}) = \pi^\theta(\theta \,|\, \mathcal{M}, \mathcal{P})$.*

This follows from the additivity of the information measure. Indeed, for any sample size n and number of replicates k, $I\{p_\theta \,|\, \mathcal{M}^{nk}\} = n\, I\{p_\theta \,|\, \mathcal{M}^k\}$. □

Note, however, that Theorem 4 requires $\boldsymbol{x}$ to be a random sample from the assumed model. If the model entails dependence between the observations (as in time series, or in spatial models) the reference prior may well depend on the sample size; see, for example, Berger and Yang (1994), and Berger, de Oliveira and Sansó (2001).

The possible dependence of the reference prior on the sample size and, more generally, on the design of the experiment highlights the fact that a reference

prior is *not* a description of (personal) prior beliefs, but a possible *consensus* prior for a particular problem of scientific inference. Indeed, genuine prior beliefs about some quantity of interest should not depend on the design of the experiment performed to learn about its value (although they will typically influence the choice of the design), but a prior function to be used as a *consensus* prior to analyse the results of an experiment may be expected to depend on its design. Reference priors, which by definition maximize the missing information which repeated observations from a *particular* experiment could possibly provide, generally depend on the design of that experiment.

As one would hope, if the assumed model $\mathcal{M}$ has a sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$, the reference prior relative to $\mathcal{M}$ is the same as the reference prior relative to the equivalent model derived from the sampling distribution of $\boldsymbol{t}$:

Theorem 5 (Compatibility with sufficient statistics) *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta\}$ *with sufficient statistic* $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}) \in \boldsymbol{\mathcal{T}}$*, and let* $\mathcal{M}_{\boldsymbol{t}} \equiv \{p(\boldsymbol{t} \,|\, \theta), \boldsymbol{t} \in \boldsymbol{\mathcal{T}},\ \theta \in \Theta\}$ *be the corresponding model in terms of* $\boldsymbol{t}$*. Then, for any class of candidate priors* $\mathcal{P}$*, the reference prior for* θ *relative to model* $\mathcal{M}$ *is* $\pi^{\theta}(\theta \,|\, \mathcal{M}, \mathcal{P}) = \pi^{\theta}(\theta \,|\, \mathcal{M}_{\boldsymbol{t}}, \mathcal{P})$*.*

Theorem 5 follows from the fact that the expected information is invariant under such transformation, so that, for all k, $I\{p_\theta \,|\, \mathcal{M}^k\} = I\{p_\theta \,|\, \mathcal{M}^k_{\boldsymbol{t}}\}$. □

When data consist of a random sample of fixed size from some model, and there exists a sufficient statistic of fixed dimensionality, Theorems 3, 4 and 5 may be combined for an easy, direct derivation of the reference prior, as illustrated below.

Example 5 *Exponential model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample of size n from an exponential distribution. By Theorem 4, to obtain the corresponding reference prior it suffices to analyse the behaviour, as $k \to \infty$, of k replications of the model which corresponds to a single observation, $\mathcal{M} \equiv \{\theta\, e^{-\theta\, y},\ y > 0,\ \theta > 0\}$, as opposed to k replications of the actual model for data $\boldsymbol{x}$, $\mathcal{M}^n \equiv \{\prod_{j=1}^n \theta\, e^{-\theta\, x_j},\ x_j > 0,\ \theta > 0\}$.

Thus, consider $\boldsymbol{y}^{(k)} = \{y_1, \ldots, y_k\}$, a random sample of size k from the single observation model $\mathcal{M}$; clearly $t_k = \sum_{j=1}^k y_j$ is sufficient, and the sampling distribution of t_k has a gamma density $p(t_k \,|\, \theta) = \mathrm{Ga}(t_k \,|\, k, \theta)$. Using a constant for the arbitrary function $h(\theta)$ in Theorem 3, the corresponding posterior has a gamma density $\mathrm{Ga}(\theta \,|\, k+1, t_k)$ and, thus,

$$f_k(\theta) = \exp\left[\int_0^\infty \mathrm{Ga}(t_k \,|\, k, \theta) \log\left\{\mathrm{Ga}(\theta \,|\, k+1, t_k)\right\} dt_k\right] = c_k\, \theta^{-1},$$

where c_k is a constant which does not contain θ. Therefore, using (15), $f(\theta) = \theta_0/\theta$ and, since this is a permissible prior function (see Example 3), the unrestricted reference prior (for both the single observation model $\mathcal{M}$ and the actual model $\mathcal{M}^n$) is $\pi(\theta \,|\, \mathcal{M}^n, \mathcal{P}_0) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \theta^{-1}$.

Parametrizations are essentially arbitrary. As one would hope, reference priors are coherent under reparametrization in the sense that if $\phi = \phi(\theta)$ is a one-to-one mapping of Θ into $\Phi = \phi(\Theta)$ then, for all $\phi \in \Phi$,

(i) $\pi^{\phi}(\phi) = \pi^{\theta}\{\theta(\phi)\}$, if Θ is discrete;

(ii) $\pi^{\phi}(\phi) = \pi^{\theta}\{\theta(\phi)\} \, |\, \partial\theta(\phi)/\partial\phi \, |$, if Θ is continuous;

More generally, reference posteriors are coherent under piecewise invertible transformations $\phi = \phi(\theta)$ of the parameter θ in the sense that, for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, the reference posterior for ϕ derived from first principles, $\pi(\phi \,|\, \boldsymbol{x})$, is precisely the same as that which could be obtained from $\pi(\theta \,|\, \boldsymbol{x})$ by standard probability calculus:

Theorem 6 (Consistency under reparametrization) *Consider a model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$ *and let* $\phi(\theta)$ *be a piecewise invertible transformation of* θ*. For any data* $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$*, the reference posterior density of* ϕ*,* $\pi(\phi \,|\, \boldsymbol{x})$*, is that induced by the reference posterior density of* θ*,* $\pi(\theta \,|\, \boldsymbol{x})$*.*

If $\phi(\theta)$ is one-to-one, Theorem 6 follows immediately from the fact that the expected information is also invariant under such transformation, so that, for all k, $I\{p_\theta \,|\, \mathcal{M}_\theta^k\} = I\{p_\psi \,|\, \mathcal{M}_\psi^k\}$; this may also be directly verified using Theorems 2 and 3. Suppose now that $\phi(\theta) = \phi_j(\theta)$, $\theta \in \Theta_j$, where the Θ_j's form a partition of Θ, such that each of the $\phi_j(\theta)$'s is one-to-one in Θ_j. The reference prior for θ only depends on the asymptotic posterior of θ which, for sufficiently large samples, will concentrate on that subset Θ_j of the parameter space Θ to which the true value of θ belongs. Since $\phi(\theta)$ is one-to-one within Θ_j, and reference priors are coherent under one-to-one parametrizations, the general result follows. □

An important consequence of Theorem 6 is that the reference prior of any location parameter, and the reference prior of the logarithm of any scale parameter are both uniform:

Theorem 7 (Location models and scale models) *Consider a* location *model* $\mathcal{M}_1$*, so that for some function* f_1*,* $\mathcal{M}_1 \equiv \{f_1(x - \mu), x \in \mathbb{R}, \mu \in \mathbb{R}\}$*, and let* $\mathcal{P}_0$ *be the class of all continuous strictly positive priors on* $\mathbb{R}$*; then, if it exists, a reference prior for* μ *is of the form* $\pi(\mu \,|\, \mathcal{M}_1, \mathcal{P}_0) = c$*. Moreover, if* $\mathcal{M}_2$ *is a* scale *model,* $\mathcal{M}_2 \equiv \{\sigma^{-1} f_2(x/\sigma), x > 0, \sigma > 0\}$*, and* $\mathcal{P}_0$ *is the class of all continuous strictly positive priors on* $(0, \infty)$*, then a reference prior for* σ*, if it exists, is of the form* $\pi(\sigma \,|\, \mathcal{M}_2, \mathcal{P}_0) = c\, \sigma^{-1}$*.*

Let $\pi(\mu)$ be the reference prior which corresponds to model $\mathcal{M}_1$; the changes $y = x + \alpha$ and $\theta = \mu + \alpha$ produce $\{f_1(y - \theta), y \in \mathcal{Y}, \theta \in \mathbb{R}\}$, which is again model $\mathcal{M}_1$. Hence, using Theorem 6, $\pi(\mu) = \pi(\mu + \alpha)$ for all α and, therefore, $\pi(\mu)$ must be constant. Moreover, the obvious changes $y = \log x$ and $\phi = \log \sigma$ transform the scale model $\mathcal{M}_2$ into a location model; hence, $\pi(\phi) = c$ and, therefore, $\pi(\sigma) \propto \sigma^{-1}$. □

Example 6 *Cauchy data.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a Cauchy distribution with unknown location μ and known scale $\sigma = 1$, so that $p(x_j \,|\, \mu) \propto [1 + (x_j - \mu)^2]^{-1}$. Since this is a location model, the reference prior is uniform and, by Bayes theorem, the corresponding reference posterior is

$$\pi(\mu \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{n} \left[1 + (x_j - \mu)^2\right]^{-1}, \qquad \mu \in \mathbb{R}.$$

Using the change of variable theorem, the reference posterior of (say) the one-to-one transformation $\phi = e^{\mu}/(1+e^{\mu})$ mapping the original parameter space $\mathbb{R}$ into $(0, 1)$, is $\pi(\phi \,|\, \boldsymbol{x}) = \pi(\mu(\phi) \,|\, \boldsymbol{x})|\partial\mu/\partial\phi|$, $\phi \in (0, 1)$. Similarly, the reference posterior $\pi(\psi \,|\, \boldsymbol{x})$ of (say) $\psi = \mu^2$ may be derived from $\pi(\mu \,|\, \boldsymbol{x})$ using standard change of variable techniques, since $\psi = \mu^2$ is a piecewise invertible function of μ, and Theorem 6 may therefore be applied.

3.3 Approximate location parametrization

Another consequence of Theorem 6 is that, for any model with one continuous parameter $\theta \in \Theta$, there is a parametrization $\phi = \phi(\theta)$ (which is unique up to a largely irrelevant proportionality constant), for which the reference prior is uniform. By Theorem 6 this may be obtained from the reference prior $\pi(\theta)$ in the original parametrization as a function $\phi = \phi(\theta)$ which satisfies the differential equation $\pi(\theta)|\partial\phi(\theta)/\partial\theta|^{-1} = 1$, that is, any solution to the indefinite integral $\phi(\theta) = \int \pi(\theta)\, d\theta$. Intuitively, $\phi = \phi(\theta)$ may be expected to behave as an *approximate* location parameter; this links reference priors with the concept data translated likelihood inducing priors introduced by Box and Tiao (1973, Sec. 1.3). For many models, good simple approximations to the posterior distribution may be obtained in terms of this parametrization, which often yields an *exact* location model.

Definition 7 (Approximate location parametrization) *Consider the model* $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta \subset \mathbb{R}\}$. *An* approximate location parametrization $\phi = \phi(\theta)$ *for model* $\mathcal{M}$ *is one for which the reference prior is uniform. In continuous regular models, this is given by any solution to the indefinite integral* $\phi(\theta) = \int \pi(\theta)\, d\theta$, *where* $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ *is the (unrestricted) reference prior for the original parameter.*

Example 7 *Exponential model, continued.* Consider again the exponential model $\mathcal{M} \equiv \{\theta\, e^{-\theta\, x},\, x > 0,\, \theta > 0\}$. The reference prior for θ is (see Example 5) $\pi(\theta) = \theta^{-1}$; thus an approximate location parameter is $\phi = \phi(\theta) = \int \pi(\theta)\, d\theta = \log \theta$. Using $y = -\log x$, this yields

$$\mathcal{M}_y \equiv \left\{ \exp\left[-(y - \phi) + e^{-(y-\phi)} \right], \quad y \in \mathbb{R}, \quad \phi \in \mathbb{R} \right\},$$

where ϕ is an (actually exact) location parameter.

Example 8 *Uniform model on* $(0, \theta)$. Let $\boldsymbol{x} = \{x_1, \ldots, x_k\}$ be a random sample from the uniform model $\mathcal{M} \equiv \{p(x \,|\, \theta) = \theta^{-1},\ 0 < x < \theta,\ \theta > 0\}$, so that $t_k = \max_{j=1}^{k} x_j$ is sufficient, and the sampling distribution of t_k is the inverted Pareto $p(t_k \,|\, \theta) = \mathrm{IPa}(t_k \,|\, k, \theta^{-1}) = k\,\theta^{-k}\,t_k^{k-1}$, if $0 < t_k < \theta$, and zero otherwise. Using a uniform prior for the arbitrary function $h(\theta)$ in Theorem 3, the corresponding posterior distribution has the Pareto density $\mathrm{Pa}(\theta \,|\, k-1, t_k) = (k-1)\,t_k^{k-1}\,\theta^{-k}$, $\theta > t_k$, and (14) becomes

$$f_k(\theta) = \exp\left[\int_0^{\theta} \mathrm{IPa}(t_k \,|\, k, \theta^{-1}) \log \mathrm{Pa}(\theta \,|\, k-1, t_k)\, dt_k\right] = c_k\,\theta^{-1},$$

where c_k is a constant which does not contain θ. Therefore, using (15), $f(\theta) = \theta_0/\theta$, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) = \theta^{-1}$.

By Theorem 4, this is also the reference prior for samples of any size; hence, by Bayes theorem, the reference posterior density of θ after, say, a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of size n has been observed is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto \textstyle\prod_{j=1}^{n} p(x_j \,|\, \theta)\,\pi(\theta) = \theta^{-(n+1)}, \quad \theta > t_n,$$

where $t_n = \max\{x_1, \ldots, x_n\}$, which is a kernel of the Pareto density $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t_n) = \mathrm{Pa}(\theta \,|\, n, t_n) = n\,(t_n)^n\,\theta^{-(n+1)}$, $\theta > t_n$.

The approximate location parameter is $\phi(\theta) = \int \theta^{-1}\, d\theta = \log \theta$. The sampling distribution of the sufficient statistic $s_n = \log t_n$ in terms of the new parameter is the reversed exponential $p(s_n \,|\, n, \phi) = n\, e^{-n(\phi - s_n)}$, $s_n < \phi$, which explicitly shows ϕ as an (exact) location parameter. The reference prior of ϕ is indeed uniform, and the reference posterior after $\boldsymbol{x}$ has been observed is the shifted exponential $\pi(\phi \,|\, \boldsymbol{x}) = n\, e^{-n(\phi - s_n)}$, $\phi > s_n$, which may also be obtained by changing variables in $\pi(\theta \,|\, \boldsymbol{x})$.

3.4 Numerical reference priors

Analytical derivation of reference priors may be technically demanding in complex models. However, Theorem 3 may also be used to obtain a numerical approximation to the reference prior which corresponds to any one-parameter model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \theta \in \Theta\}$ from which random observations may be efficiently simulated.

The proposed algorithm requires a numerical evaluation of Equation (14). This is relatively straightforward, for simulation from the assumed model may be used to approximate by Monte Carlo the integral in (14), and the evaluation of its integrand for each simulated set of data only requires (cheap) one-dimensional numerical integration. Moderate values of k (to simulate the asymptotic posterior) are typically sufficient to obtain a good approximation to the reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ (up to an irrelevant proportionality constant). The appropriate pseudo code is quite simple:

(1) Starting values:
Choose a moderate value for k,
Choose an arbitrary positive function $h(\theta)$, say $h(\theta) = 1$.
Choose the number of m of samples to be simulated,

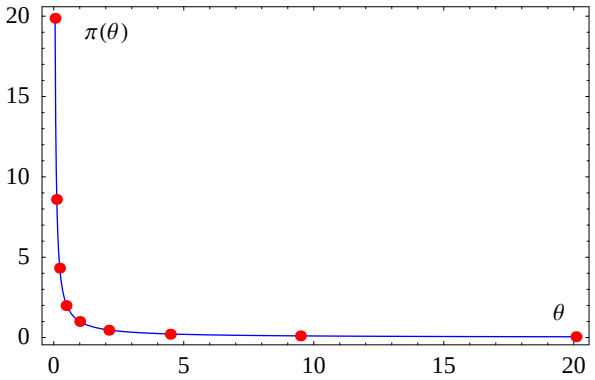
(2) For any given θ value, **repeat**, for $j = 1, \ldots, m$:
Simulate a random sample $\{\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}\}$ of size k from $p(\boldsymbol{x} \,|\, \theta)$.
Compute numerically the integral $c_j = \int_\Theta \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)\, d\theta$.
Evaluate $r_j(\theta) = \log[\, \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)/c_j \,]$.

(3) Compute $\pi(\theta) = \exp[\, m^{-1} \sum_{j=1}^m r_j(\theta) \,]$ and **store** the pair $\{\theta, \pi(\theta)\}$.

(4) **Repeat** routines (2) and (3) for all θ values for which the pair $\{\theta, \pi(\theta)\}$ is required.

Example 9 *Exponential data, continued.* Figure 3 represents the exact reference prior for the exponential model $\pi(\theta) = \theta^{-1}$ (continuous line) and the reference prior numerically calculated with the algorithm above for nine θ values, ranging from e^{-3} to e^3, uniformly log-spaced and rescaled to have $\pi(1) = 1$; $m = 500$ samples of $k = 25$ observations were used to compute each of the nine $\{\theta_i, \pi(\theta_i)\}$ points.

Figure 3 *Numerical reference prior for the exponential model*



If required, a continuous approximation to $\pi(\theta)$ may easily be obtained from the computed points using standard interpolation techniques.

An educated choice of the arbitrary function $h(\theta)$ often leads to an analytical form for the required posterior, $p(\theta \,|\, \boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}) \propto \prod_{i=1}^k p(\boldsymbol{x}_{ij} \,|\, \theta)\, h(\theta)$; for instance, this is the case in Example 9 if $h(\theta)$ is chosen to be of the form $h(\theta) = \theta^a$, for some $a \geq -1$. If the posterior may be analytically computed, then the values of the $r_j(\theta) = \log[\, p(\theta \,|\, \boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{kj}) \,]$ are immediately obtained, and the numerical algorithm reduces to only one Monte Carlo integration for each desired pair $\{\theta_i, \pi(\theta_i)\}$.

For an alternative, MCMC based, numerical computation method of reference priors, see Lafferty and Wasserman (2001).

3.5 Reference priors under regularity conditions

If data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of a model with one continuous parameter θ, it is often possible to find an *asymptotically sufficient* statistic $\tilde{\theta}_n = \tilde{\theta}_n(x_1, \ldots, x_n)$ which is also a *consistent* estimator of θ; for example, under regularity conditions, the maximum likelihood estimator (mle) $\hat{\theta}_n$ is consistent and asymptotically sufficient. In that case, the reference prior may easily be obtained in terms of either (i) an asymptotic approximation $\pi(\theta \,|\, \tilde{\theta}_n)$ to the posterior distribution of θ, or (ii) the sampling distribution $p(\tilde{\theta}_n \,|\, \theta)$ of the asymptotically sufficient consistent estimator $\tilde{\theta}_n$.

Theorem 8 (Reference priors under regularity conditions) *Let available data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ consist of a random sample of any size from a one-parameter model $\mathcal{M} \equiv \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$. Let $\boldsymbol{x}^{(k)} = \{x_1, \ldots, x_k\}$ be a random sample of size k from model $\mathcal{M}$, let $\tilde{\theta}_k = \tilde{\theta}_k(\boldsymbol{x}^{(k)}) \in \Theta$ be an asymptotically sufficient statistic which is a consistent estimator of θ, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. Let $\pi(\theta \,|\, \tilde{\theta}_k)$ be any asymptotic approximation (as $k \to \infty$) to the posterior distribution of θ, let $p(\tilde{\theta}_k \,|\, \theta)$ be the sampling distribution of $\tilde{\theta}_k$, and define*

$$f_k^a(\theta) = \pi(\theta \,|\, \tilde{\theta}_k)\Big|_{\tilde{\theta}_k = \theta}, \quad f^a(\theta) = \lim_{k \to \infty} \frac{f_k^a(\theta)}{f_k^a(\theta_0)} \tag{21}$$

$$f_k^b(\theta) = p(\tilde{\theta}_k \,|\, \theta)\Big|_{\tilde{\theta}_k = \theta}, \quad f^b(\theta) = \lim_{k \to \infty} \frac{f_k^b(\theta)}{f_k^b(\theta_0)}\,, \tag{22}$$

where θ_0 is any interior point of Θ. Then, under frequently occurring additional technical conditions, $f^a(\theta) = f^b(\theta) = f(\theta)$ and, if $f(\theta)$ is a permissible prior, any function of the form $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto f(\theta)$ is a reference prior for θ.

Since $\tilde{\theta}_k$ is asymptotically sufficient, Equation (14) in Theorem 3 becomes

$$f_k(\theta) = \exp\Big\{ \int_\Theta p(\tilde{\theta}_k \,|\, \theta) \log \pi_k(\theta \,|\, \tilde{\theta}_k)\, d\tilde{\theta}_k \Big\}.$$

Moreover, since $\tilde{\theta}_k$ is consistent, the sampling distribution of $\tilde{\theta}_k$ will concentrate on θ as $k \to \infty$, $f_k(\theta)$ will converge to $f_k^a(\theta)$, and Equation (21) will have the same limit as Equation (15). Moreover, for any formal prior function $h(\theta)$,

$$\pi(\theta \,|\, \tilde{\theta}_k) = \frac{p(\tilde{\theta}_k \,|\, \theta)\, h(\theta)}{\int_\Theta p(\tilde{\theta}_k \,|\, \theta)\, h(\theta)\, d\theta}\,.$$

As $k \to \infty$, the integral in the denominator converges to $h(\tilde{\theta}_k)$ and, therefore, $f_k^a(\theta) = \pi(\theta \,|\, \tilde{\theta}_k)\,|_{\tilde{\theta}_k = \theta}$ converges to $p(\tilde{\theta}_k \,|\, \theta)|_{\tilde{\theta}_k = \theta} = f_k^b(\theta)$. Thus, both limits in Equations (21) and (22) yield the same result, and their common value provides an explicit expression for the reference prior. For details, and precise technical conditions, see Berger, Bernardo and Sun (2005). □

Example 10 *Exponential model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample of n exponential observations from $\mathrm{Ex}(x \,|\, \theta)$. The mle is $\hat{\theta}_n(\boldsymbol{x}) = 1/\overline{x}$, a sufficient, consistent estimator of θ whose sampling distribution is the inverted gamma $p(\hat{\theta}_n \,|\, \theta) = \mathrm{IGa}(\hat{\theta}_n \,|\, n\theta, n)$. Therefore, $f_n^b(\theta) = p(\hat{\theta}_n \,|\, \theta)|_{\hat{\theta}_n = \theta} = c_n/\theta$, where $c_n = e^{-n} n^n / \Gamma(n)$ and, using Theorem 8, the reference prior is $\pi(\theta) = \theta^{-1}$.

Alternatively, the likelihood function is $\theta^n e^{-n\theta/\hat{\theta}_n}$; hence, for any positive function $h(\theta)$, $\pi^n(\theta \,|\, \hat{\theta}_n) \propto \theta^n e^{-n\theta/\hat{\theta}_n}\, h(\theta)$ is an asymptotic approximation to the posterior distribution of θ. Taking, for instance, $h(\theta) = 1$, this yields the gamma posterior $\pi^n(\theta \,|\, \hat{\theta}_n) = \mathrm{Ga}(\theta \,|\, n+1, n/\hat{\theta}_n)$. Consequently, $f_n^a(\theta) = \pi(\theta \,|\, \hat{\theta}_n)\,|_{\hat{\theta}_n = \theta} = c_n/\theta$, and $\pi(\theta) = \theta^{-1}$ as before.

Example 11 *Uniform model, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample of n uniform observations from $\mathrm{Un}(x \,|\, 0, \theta)$. The mle is $\hat{\theta}_n(\boldsymbol{x}) = \max\{x_1, \ldots, x_n\}$, a sufficient, consistent estimator of θ whose sampling distribution is the inverted Pareto $p(\hat{\theta}_n \,|\, \theta) = \mathrm{IPa}(\hat{\theta}_n \,|\, n, \theta^{-1})$. Therefore, $f_n^b(\theta) = p(\hat{\theta}_n \,|\, \theta)|_{\hat{\theta}_n = \theta} = n/\theta$ and, using Theorem 8, the reference prior is $\pi(\theta) = \theta^{-1}$.

Alternatively, the likelihood function is θ^{-n}, $\theta > \hat{\theta}_n$; hence, taking for instance a uniform prior, the Pareto $\pi^n(\theta \,|\, \hat{\theta}_n) = \mathrm{Pa}(\theta \,|\, n-1, \hat{\theta}_n)$ is found to be a particular asymptotic approximation of the posterior of θ; thus, $f_n^a(\theta) = \pi(\theta \,|\, \hat{\theta}_n)\,|_{\hat{\theta}_n = \theta} = (n-1)/\theta$, and $\pi(\theta) = \theta^{-1}$ as before.

The posterior distribution of the parameter is often asymptotically normal (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3), and references therein). In this case, the reference prior is easily derived. The result includes (univariate) Jeffreys (1946) and Perks (1947) rules as a particular cases:

Theorem 9 (Reference priors under asymptotic normality) *Let data consist of a random sample from model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta \subset \mathbb{R}\}$, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. If the posterior distribution of θ, $\pi(\theta \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, is asymptotically normal with standard deviation $s(\tilde{\theta}_n)/\sqrt{n}$, where $\tilde{\theta}_n$ is a consistent estimator of θ, and $s(\theta)^{-1}$ is a permissible prior function, then any function of the form*

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto s(\theta)^{-1} \tag{23}$$

is a reference prior. Under appropriate regularity conditions the posterior distribution of θ is asymptotically normal with variance $i(\hat{\theta}_n)^{-1}/n$, where $\hat{\theta}_n$ is the mle of θ and

$$i(\theta) = -\int_{\mathcal{Y}} p(\boldsymbol{y} \,|\, \theta)\, \frac{\partial^2}{\partial \theta^2} \log p(\boldsymbol{y} \,|\, \theta)\, d\boldsymbol{y} \tag{24}$$

is Fisher's information function. If this is the case, and $i(\theta)^{1/2}$ is a permissible prior function, the reference prior is Jeffreys prior, $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0) \propto i(\theta)^{1/2}$.

The result follows directly from Theorem 8 since, uder the assumed conditions, $f_n^a(\theta) = \pi(\theta \,|\, \hat{\theta}_n) \,|_{\hat{\theta}_n=\theta} = c_n s(\theta)^{-1}$. Jeffreys prior is the particular case which obtains when $s(\theta) = i(\theta)^{-1/2}$. □

Jeffreys (1946, 1961) prior, independently rediscovered by Perks (1947), was central in the early objective Bayesian reformulation of standard textbook problems of statistical inference (Lindley, 1965; Zellner, 1971; Press, 1972; Box and Tiao, 1973). By Theorem 9, this is also the reference prior in regular models with one continuous parameter, whose posterior distribution is asymptotically normal. By Theorem 6, reference priors are coherently transformed under one-to-one reparametrizations; hence, Theorem 9 may be typically applied with any mathematically convenient (re)parametrization. For conditions which preserve asymptotic normality under transformations see Mendoza (1994).

The posterior distribution of the exponential parameter in Example 10 is asymptotically normal; thus the corresponding reference prior may also be obtained using Theorem 9; the reference prior for the uniform parameter in Example 11 cannot be obtained however in this way, since the relevant posterior distribution is *not* asymptotically normal. Notice that, even under conditions which guarantee asymptotic normality, Jeffreys formula is not necessarily the easiest way to derive a reference prior; indeed, Theorem 8 often provides a simpler alternative.

3.6 Reference priors and the likelihood principle

By definition, reference priors are a function of the *entire* statistical model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, not of the *observed* likelihood. Indeed, the reference prior $\pi(\theta \,|\, \mathcal{M})$ is a mathematical description of lack of information about θ *relative* to the information about θ which could be obtained by repeated sampling from a particular experimental design $\mathcal{M}$. If the design is changed, the reference prior may be expected to change accordingly. This is now illustrated by comparing the reference priors which correspond to direct and inverse sampling of Bernoulli observations.

Example 12 *Binomial and negative binomial data.* Let available data $\boldsymbol{x} = \{r, m\}$ consist of m Bernoulli trials (with m fixed in advance) which contain r successes, so that the assumed model is binomial $\text{Bi}(r \,|\, m, \theta)$:

$$\mathcal{M}_1 \equiv \{p(r \,|\, m, \theta) = \binom{m}{r} \theta^r (1-\theta)^{m-r},\ r = 0, 1, \ldots, m, \quad 0 < \theta < 1\}$$

Using Theorem 9, with $n = 1$, m fixed, and $\boldsymbol{y} = r$, the reference prior for θ is the (proper) prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$; Bayes theorem yields the Beta reference posterior $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, r + 1/2,\, m - r + 1/2)$. Notice that $\pi(\theta \,|\, \boldsymbol{x})$ is proper, for all values of r; in particular, if $r = 0$, the reference posterior is $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, 1/2,\, m + 1/2)$, from which sensible

conclusions may be reached, even though there are no observed successes. This may be compared with the Haldane (1948) prior, also proposed by Jaynes (1968), $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, which produces an improper posterior until at least one success and one failure are observed.

Consider, however, that data $\boldsymbol{x} = \{r, m\}$ consist of the sequence of Bernoulli trials observed until r successes are obtained (with $r \geq 1$ fixed in advance), so that the assumed model is negative binomial:

$$\mathcal{M}_2 \equiv \{p(m \,|\, r, \theta) = \binom{m-1}{r-1} \theta^r (1-\theta)^{m-r},\ m = r, r+1, \ldots \quad 0 < \theta < 1\}$$

Using Theorem 9, with $n = 1$ and $\boldsymbol{y} = m$, the reference prior for θ is the (improper) prior $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$, and Bayes theorem yields the Beta reference posterior $\pi(\theta \,|\, \boldsymbol{x}) = \text{Be}(\theta \,|\, r,\, m - r + 1/2)$, which is proper whatever the number of observations m required to obtain r successes. Notice that $r = 0$ is *not* possible under this model: inverse binomial sampling implicitly assumes that $r \geq 1$ successes will occur for sure.

In reporting results, scientists are typically required to specify not only the observed data but also the conditions under which those where obtained, the *design* of the experiment, so that the data analyst has available the full specification of the model, $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. To carry out a reference analysis of the data, such a full specification (that is, including the experiment design) is indeed required. The reference prior $\pi(\boldsymbol{\omega} \,|\, \mathcal{M}, \mathcal{P})$ is proposed as a *consensus* prior to analyse data *associated to a particular design* $\mathcal{M}$ (and under any agreed assumptions about the value of $\boldsymbol{\omega}$ which might be encapsulated in the choice of $\mathcal{P}$).

The *likelihood principle* (Berger and Wolpert, 1988) says that all evidence about an unknown quantity $\boldsymbol{\omega}$, which is obtained from an experiment which has produced data $\boldsymbol{x}$, is contained in the likelihood function $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$ of $\boldsymbol{\omega}$ for the *observed* data $\boldsymbol{x}$. In particular, for any *specific* prior beliefs (described by a *fixed* prior), proportional likelihoods should produce the same posterior distribution.

As Example 12 demonstrates, it may be argued that formal use of reference priors is not compatible with the likelihood principle. However, the likelihood principle applies *after* data have been observed while reference priors are derived *before* the data are observed. Reference priors are a (limiting) form of rather specific beliefs, namely those which would maximize the missing information (about the quantity or interest) *associated to a particular design*, and thus depend on the particular design considered. There is no claim that these particular beliefs describe (or even approximate) those of any particular individual; instead, they are precisely defined as possible *consensus* prior functions, presumably useful as a *reference* for scientific communication. Notice that reference prior *functions* (often improper) should *not* be interpreted

as prior probability *distributions*: they are merely technical devices to facilitate the derivation of reference posteriors, and only reference posteriors support a probability interpretation.

Any statistical analysis should include an evaluation of the sensitivity of the results to accepted assumptions. In particular, any Bayesian analysis should include some discussion of the sensitivity of the results to the choice of the prior, and reference priors are better viewed as a useful tool for this important aspect of *sensitivity analysis*. The analyst is supposed to have a unique (often subjective) prior $p(\boldsymbol{\omega})$, independent of the design of the experiment, but the scientific community will presumably be interested in comparing the corresponding analyst's personal posterior with the *reference* (consensus) posterior associated to the published experimental design. To report reference posteriors (possibly for a range of alternative designs) should be seen as part of this sensitivity analysis. Indeed, reference analysis provides an answer to an important *conditional* question in scientific inference: the reference posterior encapsulates what *could* be said about the quantity of interest *if* prior information about its value were minimal *relative* to the information which repeated data from an specific experimental design $\mathcal{M}$ could possibly provide.

3.7 Restricted reference priors

The reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ is that which maximizes the missing information about θ relative to model $\mathcal{M}$ among the priors which belong to $\mathcal{P}$, the class of all sufficiently regular priors which are compatible with available knowledge (Definition 6). By restricting the class $\mathcal{P}$ of candidate priors to those which satisfy specific restrictions (derived from assumed knowledge) one may use the reference prior algorithm as an effective tool for *prior elicitation*: the corresponding reference prior will incorporate the accepted restrictions, but no other information.

Under regularity conditions, Theorems 3, 8 and 9, make it relatively simple to obtain the unrestricted reference prior $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ which corresponds to the case where the class of candidate priors is the class $\mathcal{P}_0$ of all continuous priors with support Θ. Hence, it is useful to be able to express a general reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ in terms of the corresponding unrestricted reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$, and the set of restrictions which define the class $\mathcal{P}$ of candidate priors.

If the unrestricted reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ is proper, then $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ is the closest prior in $\mathcal{P}$ to $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$, in the sense of minimizing the intrinsic discrepancy (see Definition 1) between them, so that

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) = \arg \inf_{p(\theta) \in \mathcal{P}} \delta\{\, p(\theta),\ \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)\,\}$$

If $\pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)$ is not proper it may be necessary to derive $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$ from its definition. However, in the rather large class of problems where the conditions

which define $\mathcal{P}$ may all be expressed in the general form $\int_\Theta g_i(\theta)\, p(\theta)\, \mathrm{d}\theta = \beta_i$, for appropriately chosen functions $g_i(\theta)$, (*i.e.*, as a collection of expected values which the prior $p(\theta)$ must satisfy), an explicit solution is available in terms of the unrestricted reference prior:

Theorem 10 (Explicit form of restricted reference priors) *Consider a model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta\}$, let $\mathcal{P}$ be the class of continuous proper priors with support Θ*

$$\mathcal{P} = \Big\{ p_\theta; \quad \int_\Theta p(\theta)\, d\theta = 1, \quad \int_\Theta g_i(\theta)\, p(\theta)\, d\theta = \beta_i, \quad i = 1, \ldots, m \Big\}$$

which satisfies the restrictions imposed by the expected values $\mathrm{E}[g_i(\theta)] = \beta_i$, and let $\mathcal{P}_0$ be the class of all continuous priors with support Θ. The reference prior $\pi(\theta \,|\, \mathcal{M}, \mathcal{P})$, if it exists, is then of the form

$$\pi(\theta \,|\, \mathcal{M}, \mathcal{P}) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P}_0)\ \exp\Big\{ \textstyle\sum_{i=1}^m \lambda_i\, g_i(\theta) \Big\}$$

where the λ_i's are constants determined by the conditions which define $\mathcal{P}$.

Theorem 10 may be proven using a standard calculus of variations argument. If $m = 0$, so that one only has the constraint that the prior is proper, then there typically is no restricted reference prior. For details, see Bernardo and Smith (1994, p. 316). □

Example 13 *Location models, continued.* Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a location model $\mathcal{M} \equiv \{f(x - \mu), x \in \mathcal{X}, \mu \in \mathbb{R}\}$, and suppose that the prior mean and variance of μ are restricted to be $\mathrm{E}[\mu] = \mu_0$, and $\mathrm{Var}[\mu] = \sigma_0^2$. By Theorem 7, the unrestricted reference prior $\pi(\mu \,|\, \mathcal{M}, \mathcal{P}_0)$ is uniform; hence, using Theorem 10, the (restricted) reference prior must be of the form

$$\pi(\mu \,|\, \mathcal{M}, \mathcal{P}) \propto \exp\{\lambda_1 \mu + \lambda_2 (\mu - \mu_0)^2\}$$

with $\int_{-\infty}^{\infty} \mu\, \pi(\mu \,|\, \mathcal{M}, \mathcal{P})\, d\mu = \mu_0$ and $\int_{-\infty}^{\infty} (\mu - \mu_0)^2\, \pi(\mu \,|\, \mathcal{M}, \mathcal{P})\, d\mu = \sigma_0^2$. It follows that $\lambda_1 = 0$ and $\lambda_2 = -1/(2\sigma_0^2)$ and, substituting above, the restricted reference prior is $\pi(\mu \,|\, \mathcal{M}, \mathcal{P}) \propto \exp\{-(\mu - \mu_0)^2/(2\sigma_0^2)\}$, which is the *normal* distribution $\mathrm{N}(\mu \,|\, \mu_0, \sigma_0)$ with the specified mean and variance. This provides a very powerful argument for the choice of a normal density to describe prior information in location models, when prior knowledge about the location parameter is *limited* to its first two moments.

3.8 One nuisance parameter

Consider now the case where the statistical model $\mathcal{M}$ contains one nuisance parameter, so that $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda\}$, the quantity of

interest is $\theta \in \Theta \subset \mathbb{R}$, and the nuisance parameter is $\lambda \in \Lambda \subset \mathbb{R}$. To obtain the required reference posterior for θ, $\pi(\theta \,|\, \boldsymbol{x})$, an appropriate *joint* reference prior $\pi^{\theta}(\theta, \lambda)$ is obviously needed: by Bayes theorem, the corresponding joint posterior is $\pi^{\theta}(\theta, \lambda \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)$ and, integrating out the nuisance parameter, the (marginal) reference posterior for the parameter of interest is

$$\pi(\theta \,|\, \boldsymbol{x}) = \int_{\Lambda} \pi^{\theta}(\theta, \lambda \,|\, \boldsymbol{x})\, d\lambda \propto \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)\, d\lambda.$$

The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the two parameter problem to a sequential application of the established procedure for the single parameter case. Thus, the reference algorithm proceeds by combining the results obtained in two successive applications of the one-parameter solution:

(1) Conditional on θ, $p(\boldsymbol{x} \,|\, \theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the *conditional* reference prior $\pi(\lambda \,|\, \theta) = \pi(\lambda \,|\, \theta, \mathcal{M}, \mathcal{P})$.

(2) If $\pi(\lambda \,|\, \theta)$ has a finite integral in Λ (so that, when normalized, yields a proper density with $\int_{\Lambda} \pi(\lambda \,|\, \theta)\, d\lambda = 1$), the conditional reference prior $\pi(\lambda \,|\, \theta)$ may be used to integrate out the nuisance parameter and derive the one-parameter integrated model,

$$p(\boldsymbol{x} \,|\, \theta) = \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda, \tag{25}$$

to which the one-parameter algorithm may be applied again to obtain the *marginal* reference prior $\pi(\theta) = \pi(\theta \,|\, \mathcal{M}, \mathcal{P})$.

(3) The desired θ-reference prior is then $\pi^{\theta}(\theta, \lambda) = \pi(\lambda \,|\, \theta)\, \pi(\theta)$, and the required reference posterior is

$$\pi(\theta \,|\, \boldsymbol{x}) \propto \int_{\Lambda} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^{\theta}(\theta, \lambda)\, d\lambda = p(\boldsymbol{x} \,|\, \theta)\, \pi(\theta). \tag{26}$$

Equation (25) suggests that conditional reference priors provides a general procedure to eliminate nuisance parameters, a major problem within the frequentist paradigm. For a review of this important topic, see Liseo (2005), in this volume.

If the conditional reference prior $\pi(\lambda \,|\, \theta)$ is *not* proper, Equation (25) does not define a valid statistical model and, as a consequence, a more subtle approach is needed to provide a general solution; this will be described later. Nevertheless, the simple algorithm described above may be used to obtain appropriate solutions to a number of interesting problems which serve to illustrate the crucial need to identify the quantity of interest, as is the following two examples.

Example 14 *Induction.* Consider a finite population of (known) size N, all of whose elements may or may not have a specified property. A random sample of size n is taken without replacement, and all the elements in the sample turn out to have that property. Scientific interest often centres in the probability that all the N elements in the population have the property under consideration (natural induction). It has often been argued that for relatively large n values, this should be close to one whatever might be the population size N (typically much larger than the sample size n). Thus, if all the $n = 225$ randomly chosen turtles in an isolated volcanic island are found to show a particular difference with respect to those in the mainland, zoologists would tend to believe that all the turtles in the island share that property. Formally, if r and R respectively denote the number of elements in the sample and in the population which have the property under study, the statistical model is

$$\mathcal{M} \equiv \Big\{ p(r \,|\, n, R, N), \quad r \in \{0, \ldots, n\}, \quad R \in \{0, \ldots, N\} \Big\},$$

where R is the unknown parameter, and $p(r \,|\, n, R, N) = \binom{R}{r}\binom{N-R}{n-r} / \binom{N}{n}$ is the relevant hypergeometric distribution. The required result,

$$p(R = N \,|\, r = n, N) = \frac{p(r = n \,|\, n, R, N)\, p(R = N)}{\sum_{R=0}^{N} p(r = n \,|\, n, R, N)\, p(R)} \,. \tag{27}$$

may immediately be obtained from Bayes theorem, once a prior $p(R)$ for the unknown number R of elements in the population which have the property has been established. If the parameter of interest were R itself, the reference prior would be uniform over its range (Theorem 2), so that $p(R) = (N + 1)^{-1}$; using (27) this would lead to the posterior probability $p(R = N \,|\, r = n, N) = (n + 1)/(N + 1)$ which will be small when (as it is usually the case) the sampling fraction n/N is small. However, the quantity of interest here is *not* the value of R but whether or not $R = N$, and a reference prior is desired which maximizes the missing information about this *specific* question. Rewriting the unknown parameter as $R = (\theta, \lambda)$, where $\theta = 1$ if $R = N$ and $\theta = 0$ otherwise, and $\lambda = 1$ if $R = N$ and $\lambda = R$ otherwise (so that the quantity of interest θ is explicitly shown), and using Theorem 2 and the argument above, one gets $\pi(\lambda \,|\, \theta = 1) = 1$, $\pi(\lambda \,|\, \theta = 0) = N^{-1}$, and $\pi(\theta = 0) = \pi(\theta = 1) = 1/2$, so that the θ-reference prior is $\pi^{\theta}(R) = 1/2$ if $R = N$ and $\pi^{\theta}(R) = 1/(2N)$ if $R \neq N$. Using (27), this leads to

$$p(R = N \,|\, r = n, N) = \left[1 + \frac{1}{n+1} \Big(1 - \frac{n}{N} \Big) \right]^{-1} \approx \frac{n+1}{n+2} \tag{28}$$

which, as expected, clearly displays the irrelevance of the sampling fraction, and the approach to unity for large n. In the turtles example (a real question posed to the author at the Galápagos Islands in the eighties), this

yields $p(R = N \,|\, r = n = 225,\, N) \approx 0.995$ for all large N. The *reference* result (28) does not necessarily represents any personal scientist's beliefs (although apparently it may approach actual scientists's beliefs in many situations), but the conclusions which should be reached from a situation where the missing information about the quantity of interest (whether or not $R = N$) is maximized, a situation mathematically characterized by the θ-reference prior described above. For further discussion of this problem (with important applications in philosophy of science, physical sciences and reliability), see Jeffreys (1961, pp. 128–132), Geisser (1984) and Bernardo (1985b).

Example 15 *Ratio of multinomial parameters.* Let data $\boldsymbol{x} = \{r_1, r_2, n\}$ consist of the result of n trinomial observations, with parameters α_1, α_2 and $\alpha_3 = 1 - \alpha_1 - \alpha_2$, so that, for $0 < \alpha_i < 1$, $\alpha_1 + \alpha_2 < 1$,

$$p(r_1, r_2 \,|\, n, \alpha_1, \alpha_2) = c(r_1, r_2, n)\, \alpha_1^{r_1} \alpha_2^{r_2} (1 - \alpha_1 - \alpha_2)^{n - r_1 - r_2},$$

where $c(r_1, r_2, n) = (n!)/(r_1!\, r_2!\, (n - r_1 - r_2)!)$, and suppose that the quantity of interest is the *ratio* $\theta = \alpha_1/\alpha_2$ of the first two original parameters. Reparametrization in terms of θ and (say) $\lambda = \alpha_2$ yields

$$p(r_1, r_2 \,|\, n, \theta, \lambda) = c(r_1, r_2, n)\, \theta^{r_1}\, \lambda^{r_1 + r_2}\, \{1 - \lambda(1 + \theta)\}^{n - r_1 - r_2},$$

for $\theta > 0$ and, given θ, $0 < \lambda < (1 + \theta)^{-1}$. Conditional on θ, this is a model with one continuous parameter λ, and the corresponding Fisher information function is $i(\lambda \,|\, \theta) = n(1 + \theta)/\{\lambda(1 - \lambda(1 + \theta))\}$; using Theorem 9 the conditional reference prior of the nuisance parameter is $\pi(\lambda \,|\, \theta) \propto i(\lambda \,|\, \theta)^{1/2}$ which is the *proper* beta-like prior $\pi(\lambda \,|\, \theta) \propto \lambda^{-1/2} \{1 - \lambda(1 + \theta)\}^{-1/2}$, with support on $\lambda \in [0, (1 + \theta)^{-1}]$ (which depends on θ). Integration of the full model $p(r_1, r_2 \,|\, n, \theta, \lambda)$ with the conditional reference prior $\pi(\lambda \,|\, \theta)$ yields $p(r_1, r_2 \,|\, n, \theta) = \int_0^{(1+\theta)^{-1}} p(r_1, r_2 \,|\, n, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda$, the *integrated* one-parameter model

$$p(r_1, r_2 \,|\, n, \theta) = \frac{\Gamma(r_1 + r_2 + \frac{1}{2})\, \Gamma(n - r_1 - r_2 + \frac{1}{2})}{r_1!\, r_2!\, (n - r_1 - r_2)!}\; \frac{\theta^{r_1}}{(1 + \theta)^{r_1 + r_2}}\,.$$

The corresponding Fisher information function is $i(\theta) = n/\{2\theta(1 + \theta)^2\}$; using again Theorem 9 the reference prior of the parameter of interest is $\pi(\theta) \propto i(\theta)^{1/2}$ which is the proper prior $\pi(\theta) \propto \theta^{-1/2}(1 + \theta)^{-1}$, $\theta > 0$. Hence, by Bayes theorem, the reference posterior of the quantity of interest is $\pi(\theta \,|\, r_1, r_2, n) \propto p(r_1, r_2 \,|\, n, \theta)\, \pi(\theta)$; this yields

$$\pi(\theta \,|\, r_1, r_2) = \frac{\Gamma(r_1 + r_2 + 1)}{\Gamma(r_1 + \frac{1}{2})\, \Gamma(r_2 + \frac{1}{2})}\; \frac{\theta^{r_1 - 1/2}}{(1 + \theta)^{r_1 + r_2 + 1}}\,, \quad \theta > 0.$$

Notice that $\pi(\theta \,|\, r_1, r_2)$ does *not* depend on n; to draw conclusions about the value of $\theta = \alpha_1/\alpha_2$ only the numbers r_1 and r_2 observed in the first

two classes matter: a result $\{55, 45, 100\}$ carries precisely the same information about the *ratio* α_1/α_2 than a result $\{55, 45, 10000\}$. For instance, if an electoral survey of size n yields r_1 voters for party A and r_2 voters for party B, the reference posterior distribution of the *ratio* θ of the proportion of A voters to B voters in the population only depends on their respective number of voters in the sample, r_1 and r_2, whatever the size and political intentions of the other $n - r_1 - r_2$ citizens in the sample. In particular, the reference posterior probability that party A gets better results than party B is $\Pr[\theta > 1 \,|\, r_1, r_2] = \int_1^\infty \pi(\theta \,|\, r_1, r_2)\, d\theta$. As one would expect, this is precisely equal to $1/2$ if, and only if, $r_1 = r_2$; one-dimensional numerical integration (or use of the incomplete beta function) is required to compute other values. For instance, whatever the total sample size n in each case, this yields $\Pr[\theta > 1 \,|\, r_1 = 55, r_2 = 45] = 0.841$ (with $r_1 + r_2 = 100$) and $\Pr[\theta > 1 \,|\, r_1 = 550, r_2 = 450] = 0.999$ (with the same ratio r_1/r_2, but $r_1 + r_2 = 1000$).

As illustrated by the preceding examples, in a multiparameter model, say $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ the required (joint) reference prior $\pi^{\theta}(\boldsymbol{\omega})$ may depend on the quantity of interest, $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ (although, as one would certainly expect, and will later be demonstrated, this will *not* be the case if the new quantity of interest $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\omega})$ say, is a one-to-one function of $\boldsymbol{\theta}$). Notice that this does *not* mean that the analyst's beliefs should depend on his or her interests; as stressed before, reference priors are not meant to describe the analyst's beliefs, but the mathematical formulation of a particular type of prior beliefs—those which would maximize the expected missing information about the quantity of interest—which could be adopted by consensus as a standard for scientific communication.

If the conditional reference prior $\pi(\lambda \,|\, \theta)$ is *not* proper, so that Equation (25) does not define a valid statistical model, then integration may be performed within each of the elements of an increasing sequence $\{\Lambda_i\}_{i=1}^{\infty}$ of subsets of Λ converging to Λ over which $\pi(\lambda \,|\, \theta)$ is integrable. Thus, Equation (25) is to be replaced by

$$p_i(\boldsymbol{x} \,|\, \theta) = \int_{\Lambda_i} p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi_i(\lambda \,|\, \theta)\, d\lambda, \tag{29}$$

where $\pi_i(\lambda \,|\, \theta)$ is the renormalized proper restriction of $\pi(\lambda \,|\, \theta)$ to Λ_i, from which the reference posterior $\pi_i(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, \mathcal{M}_i, \mathcal{P})$, which corresponds to model $\mathcal{M}_i \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda_i\}$ may be derived.

The use of the sequence $\{\Lambda_i\}_{i=1}^{\infty}$ makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta \,|\, \boldsymbol{x})\}_{i=1}^{\infty}$ for the quantity of interest θ which corresponds to the sequence of integrated models (29); the required reference posterior may then be found as the corresponding intrinsic limit $\pi(\theta \,|\, \boldsymbol{x}) = \lim_{i \to \infty} \pi_i(\theta \,|\, \boldsymbol{x})$. A θ-reference prior is then defined as any positive function $\pi^{\theta}(\theta, \lambda)$ which may formally be used in Bayes' theorem to directly

obtain the reference posterior, so that for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, the posterior density satisfies $\pi(\theta \,|\, \boldsymbol{x}) \propto \int_\Lambda p(\boldsymbol{x} \,|\, \theta, \lambda)\, \pi^\theta(\theta, \lambda)\, d\lambda$.

The approximating sequences should be *consistently* chosen within the same model: given a statistical model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ an appropriate approximating sequence $\{\boldsymbol{\Omega}_i\}$ should be chosen for the whole parameter space $\boldsymbol{\Omega}$. Thus, if the analysis is done in terms of $\psi = \{\psi_1, \psi_2\} \in \Psi(\boldsymbol{\Omega})$, the approximating sequence should be chosen such that $\Psi_i = \psi(\boldsymbol{\Omega}_i)$. A very natural approximating sequence in location-scale problems is $\{\mu, \log \sigma\} \in [-i, i]^2$; reparametrization to asymptotically independent parameters and approximate location reparametrizations (Definition 7) may be combined to choose appropriate approximating sequences in more complex situations. A formal definition of reference prior functions in multiparameter problems is possible along the lines of Definition 6.

As one would hope, the θ-reference prior does *not* depend on the choice of the nuisance parameter λ; thus, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ), the θ-reference prior in terms of (θ, ψ) is simply $\pi^\theta(\theta, \psi) = \pi^\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, the appropriate probability transformation of the θ-reference prior in terms of (θ, λ). Notice however that, as mentioned before, the reference prior *may* depend on the parameter of interest; thus, the θ-reference prior may differ from the ϕ-reference prior unless either ϕ is a one-to-one transformation of θ, or ϕ is asymptotically independent of θ. This is an expected consequence of the mathematical fact that the prior which maximizes the missing information about θ is not generally the same as the prior which maximizes the missing information about any function $\phi = \phi(\theta, \lambda)$.

The *non-existence* of a unique "noninformative" prior for all inference problems within a given model was established by Dawid, Stone and Zidek (1973) when they showed that this is incompatible with *consistent marginalization*. Indeed, given a two-parameter model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \theta, \lambda), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \theta \in \Theta, \lambda \in \Lambda\}$, if the reference posterior of the quantity of interest θ, $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, \boldsymbol{t})$, only depends on the data through a statistic $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ whose sampling distribution, $p(\boldsymbol{t} \,|\, \theta, \lambda) = p(\boldsymbol{t} \,|\, \theta)$, only depends on θ, one would expect the reference posterior to be of the form $\pi(\theta \,|\, \boldsymbol{t}) \propto p(\boldsymbol{t} \,|\, \theta)\, \pi(\theta)$ for some prior $\pi(\theta)$. However, examples were found where this *cannot* be the case if a *unique* joint "noninformative" prior were to be used for all possible quantities of interest within the same statistical model $\mathcal{M}$.

By definition, a reference prior must be a *permissible* prior function. In particular (Definition 3), it must yield *proper posteriors* for all data sets large enough to identify the parameters. For instance, if data $\boldsymbol{x}$ consist of a random sample of fixed size n from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ distribution, so that, $\mathcal{M} \equiv \{\prod_{j=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma),\ x_j \in \mathbb{R},\ \sigma > 0\}$, the function $\pi^\mu(\mu, \sigma) = \sigma^{-1}$ is only a permissible (joint) prior for μ if $n \geq 2$ (and, without restrictions in the class $\mathcal{P}$ of candidate priors, a reference prior function *does not exist* for $n = 1$).

Under posterior asymptotic normality, reference priors are easily obtained in terms of the relevant Fisher information matrix. The following result extends Theorem 9 to models with two continuous parameters:

Theorem 11 (Reference priors under asymptotic binormality) *Let data* $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ *consist of* n *conditionally independent (given* θ*) observations from a model* $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \theta, \lambda), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \theta \in \Theta, \lambda \in \Lambda\}$*, and let* $\mathcal{P}_0$ *be the class of all continuous (joint) priors with support* $\Theta \times \Lambda\}$*. If the posterior distribution of* $\{\theta, \lambda\}$ *is asymptotically normal with dispersion matrix* $V(\hat{\theta}_n, \hat{\lambda}_n)/n$*, where* $\{\hat{\theta}_n, \hat{\lambda}_n\}$ *is a consistent estimator of* $\{\theta, \lambda\}$*, define*

$$V(\theta, \lambda) = \begin{pmatrix} v_{\theta\theta}(\theta, \lambda) & v_{\theta\lambda}(\theta, \lambda) \\ v_{\theta\lambda}(\theta, \lambda) & v_{\lambda\lambda}(\theta, \lambda) \end{pmatrix}, \quad H(\theta, \lambda) = V^{-1}(\theta, \lambda), \quad \textit{and}$$

$$\pi(\lambda \,|\, \theta) \propto h^{1/2}_{\lambda\lambda}(\theta, \lambda), \quad \lambda \in \Lambda, \tag{30}$$

and, if $\pi(\lambda \,|\, \theta)$ *is proper,*

$$\pi(\theta) \propto \exp\Big\{ \int_\Lambda \pi(\lambda \,|\, \theta) \, \log[v^{-1/2}_{\theta\theta}(\theta, \lambda)] \, d\lambda \Big\}, \quad \theta \in \Theta. \tag{31}$$

Then, if $\pi(\lambda \,|\, \theta)\, \pi(\theta)$ *is a permissible prior function, the* θ*-reference prior is*

$$\pi(\theta \,|\, \mathcal{M}^n, \mathcal{P}_0) \propto \pi(\lambda \,|\, \theta)\, \pi(\theta).$$

If $\pi(\lambda \,|\, \theta)$ *is not proper, integration in* (31) *is performed on elements of an increasing sequence* $\{\Lambda_i\}_{i=1}^\infty$ *such that* $\int_{\Lambda_i} \pi(\lambda \,|\, \theta)\, d\lambda < \infty$*, to obtain the sequence* $\{\pi_i(\lambda \,|\, \theta)\, \pi_i(\theta)\}_{i=1}^\infty$*, where* $\pi_i(\lambda \,|\, \theta)$ *is the renormalization of* $\pi(\lambda \,|\, \theta)$ *to* Λ_i*, and the* θ*-reference prior* $\pi^\theta(\theta, \lambda)$ *is defined as its corresponding intrinsic limit.*

A heuristic justification of Theorem 11 is now provided. Under the stated conditions, given k independent observations from model $\mathcal{M}$, the *conditional* posterior distribution of λ given θ is asymptotically normal with precision $k\, h_{\lambda\lambda}(\theta, \hat{\lambda}_k)$, and the *marginal* posterior distribution of θ is asymptotically normal with precision $k\, v^{-1}_{\theta\theta}(\hat{\theta}_k, \hat{\lambda}_k)$; thus, using Theorem 9, $\pi(\lambda \,|\, \theta) \propto h^{1/2}_{\lambda\lambda}(\theta, \lambda)$, which is Equation (30). Moreover, using Theorem 3,

$$\pi_k(\theta) \propto \exp\Big\{ \iint p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta) \log[\mathrm{N}\{\theta \,|\, \hat{\theta}_k, k^{-1/2} v^{1/2}_{\theta\theta}(\hat{\theta}_k, \hat{\lambda}_k)\}] \, d\hat{\theta}_k \, d\hat{\lambda}_k \Big\} \tag{32}$$

where, if $\pi(\lambda \,|\, \theta)$ is proper, the integrated model $p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta)$ is given by

$$p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta) = \int_\Lambda p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda. \tag{33}$$

Introducing (33) into (32) and using the fact that $(\hat{\theta}_k, \hat{\lambda}_k)$ is a consistent estimator of (θ, λ)—so that as $k \to \infty$ integration with $p(\hat{\theta}_k, \hat{\lambda}_k \,|\, \theta, \lambda)$ reduces

to substitution of $(\hat{\theta}_k, \hat{\lambda}_k)$ by (θ, λ)—directly leads to Equation (31). If $\pi(\lambda \,|\, \theta)$ is not proper, it is necessary to integrate in an increasing sequence $\{\Lambda_i\}_{i=1}^{\infty}$ of subsets of Λ such that the restriction $\pi_i(\lambda \,|\, \theta)$ of $\pi(\lambda \,|\, \theta)$ to Λ_i is proper, obtain the sequence of reference priors which correspond to these restricted models, and then take limits to obtain the required result. □

Notice that under appropriate regularity conditions (see *e.g.*, Bernardo and Smith (1994, Sec. 5.3) and references therein) the joint posterior distribution of $\{\theta, \lambda\}$ is asymptotically normal with precision matrix $n\, I(\hat{\theta}_n, \hat{\lambda}_n)$, where $I(\theta)$ is Fisher information matrix; in that case, the asymptotic dispersion matrix in Theorem 11 is simply $V(\theta, \lambda) = I^{-1}(\theta, \lambda)/n$.

Theorem 12 (Reference priors under factorization) *In the conditions of Theorem 11, if* (i) *θ and λ are variation independent—so that Λ does not depend on θ—and* (ii) *both $h_{\lambda\lambda}(\theta, \lambda)$ and $v_{\theta\theta}(\theta, \lambda)$ factorize, so that*

$$v_{\theta\theta}^{-1/2}(\theta, \lambda) \propto f_\theta(\theta)\, g_\theta(\lambda), \quad h_{\lambda\lambda}^{1/2}(\theta, \lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda), \tag{34}$$

then the θ-reference prior is simply $\pi^\theta(\theta, \lambda) = f_\theta(\theta)\, g_\lambda(\lambda)$, even if the conditional reference prior $\pi(\lambda \,|\, \theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ is improper.

If $h_{\lambda\lambda}^{1/2}(\theta, \lambda)$ factorizes as $h_{\lambda\lambda}^{1/2}(\theta, \lambda) = f_\lambda(\theta) g_\lambda(\lambda)$, then the conditional reference prior is $\pi(\lambda \,|\, \theta) \propto f_\lambda(\theta) g_\lambda(\lambda)$ and, normalizing, $\pi(\lambda \,|\, \theta) = c_1\, g_\lambda(\lambda)$, which does not depend on θ. If, furthermore, $v_{\theta\theta}^{-1/2}(\theta, \lambda) = f_\theta(\theta) g_\theta(\lambda)$ and Λ does not depend on θ, Equation (31) reduces to

$$\pi(\theta) \propto \exp\Big\{ \int_\Lambda c_1\, g_\lambda(\lambda) \log[f_\theta(\theta) g_\theta(\lambda)]\, d\lambda \Big\} = c_2\, f_\theta(\theta)$$

and, hence, the reference prior is $\pi^\theta(\theta, \lambda) = \pi(\lambda \,|\, \theta)\, \pi(\theta) = c\, f_\theta(\theta)\, g_\lambda(\lambda)$. □

Example 16 *Inference on the univariate normal parameters.* Let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of a random sample of fixed size n from a normal distribution $\mathrm{N}(x \,|\, \mu, \sigma)$. The information matrix $I(\mu, \sigma)$ and its inverse matrix are respectively

$$I(\mu, \sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix}, \quad V(\mu, \sigma) = I^{-1}(\mu, \sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Hence, $i_{\sigma\sigma}^{1/2}(\mu, \sigma) = \sqrt{2}\, \sigma^{-1} = f_\sigma(\mu)\, g_\sigma(\sigma)$, with $g_\sigma(\sigma) = \sigma^{-1}$, so that $\pi(\sigma \,|\, \mu) = \sigma^{-1}$. Similarly, $v_{\mu\mu}^{-1/2}(\mu, \sigma) = \sigma^{-1} = f_\mu(\mu)\, g_\sigma(\sigma)$, with $f_\mu(\mu) = 1$, and thus $\pi(\mu) = 1$. Therefore, using Theorem 11 the μ-reference prior is $\pi^\mu(\mu, \sigma) = \pi(\sigma \,|\, \mu)\, \pi(\mu) = \sigma^{-1}$ for all $n \geq 2$. For $n = 1$ the posterior distribution is not proper, the function $h(\mu, \sigma) = \sigma^{-1}$ is *not* a permissible prior, and a reference prior does not exist. Besides, since $I(\mu, \sigma)$ is diagonal, the σ-reference prior is $\pi^\sigma(\mu, \sigma) = f_\sigma(\sigma)\, g_\mu(\mu) = \sigma^{-1}$, the same as $\pi^\mu(\mu, \sigma)$.

Consider now the case where the quantity of interest is *not* the mean μ or the standard deviation σ, but the *standardized* mean $\phi = \mu/\sigma$ (or, equivalently, the coefficient of variation σ/μ). Fisher's matrix in terms of the parameters ϕ and σ is $I(\phi, \sigma) = J^t\, I(\mu, \sigma)\, J$, where $J = (\partial(\mu, \sigma)/\partial(\phi, \sigma))$ is the Jacobian of the inverse transformation, and this yields

$$I(\phi, \sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix}, \quad V(\phi, \sigma) = \begin{pmatrix} 1 + \frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma \\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Thus, $i^{1/2}_{\sigma\sigma}(\phi, \sigma) = \sigma^{-1}(2+\phi^2)^{1/2}$, and $v^{-1/2}_{\phi\phi}(\phi, \sigma) = (1 + \frac{1}{2}\phi^2)^{-1/2}$. Hence, using Theorem 11, $\pi^{\phi}(\phi, \sigma) = (1 + \frac{1}{2}\phi^2)^{-1/2}\sigma^{-1}$ $(n \geq 2)$. In the original parametrization, this is $\pi^{\phi}(\mu, \sigma) = (1 + \frac{1}{2}(\mu/\sigma)^2)^{-1/2}\sigma^{-2}$, which is *very* different from $\pi^{\mu}(\mu, \sigma) = \pi^{\sigma}(\mu, \sigma) = \sigma^{-1}$. The reference posterior of the quantity of interest ϕ after data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ have been observed is

$$\pi(\phi \,|\, \boldsymbol{x}) \propto (1 + \tfrac{1}{2}\phi^2)^{-1/2}\, p(t \,|\, \phi) \tag{35}$$

where $t = (\sum x_j)/(\sum x_j^2)^{1/2}$, a one-dimensional statistic whose sampling distribution, $p(t \,|\, \mu, \sigma) = p(t \,|\, \phi)$, only depends on ϕ. Thus, the reference prior algorithm is seen to be consistent under marginalization.

The reference priors $\pi^{\mu}(\mu, \sigma) = \sigma^{-1}$ and $\pi^{\sigma}(\mu, \sigma) = \sigma^{-1}$ for the normal location and scale parameters obtained in the first part of Example 16 are just a particular case of a far more general result:

Theorem 13 (Location-scale models) *If $\mathcal{M}$ is a* location-scale *model, so that for some function f, $\mathcal{M} \equiv \sigma^{-1} f\{(x-\mu)/\sigma\}$, $x \in \mathcal{X}$, $\mu \in \mathbb{R}, \sigma > 0\}$, and $\mathcal{P}_0$ is the class of all continuous, strictly positive (joint) priors for (μ, σ), then a reference prior for either μ or σ, if it exists, is of the form*

$$\pi^{\mu}(\mu, \sigma \,|\, \mathcal{M}, \mathcal{P}_0) = \pi^{\sigma}(\mu, \sigma \,|\, \mathcal{M}, \mathcal{P}_0) \propto \sigma^{-1}.$$

For a proof, which is based on the form of the relevant Fisher matrix, see Fernández and Steel (1999b). □

When the quantity of interest and the nuisance parameter are *not* variation independent, derivation of the reference prior requires special care. This is illustrated in the example below:

Example 17 *Product of positive normal means.* Let data consist of two independent random samples $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{y} = \{y_1, \ldots, y_m\}$ from $N(x \,|\, \alpha, 1)$ and $N(y \,|\, \beta, 1)$, $\alpha > 0$, $\beta > 0$, so that the assumed model is

$$p(\boldsymbol{x}, \boldsymbol{y} \,|\, \alpha, \beta) = \textstyle\prod_{i=1}^{n} N(x_i \,|\, \alpha, 1) \prod_{j=1}^{m} N(y_j \,|\, \beta, 1), \quad \alpha > 0,\ \beta > 0,$$

and suppose that the quantity of interest is the product of the means,

$\theta = \alpha\beta$, a frequent situation in physics and engineering. Reparametrizing in terms of the one-to-one transformation $(\theta, \lambda) = (\alpha\beta, \alpha/\beta)$, Fisher matrix $I(\theta, \lambda)$ and its inverse matrix $V(\theta, \lambda)$ are,

$$I = \begin{pmatrix} \frac{m+n\lambda^2}{4\theta\lambda} & \frac{1}{4}\left(n - \frac{m}{\lambda^2}\right) \\ \frac{1}{4}\left(n - \frac{m}{\lambda^2}\right) & \frac{\theta\,(m+n\lambda^2)}{4\lambda^3} \end{pmatrix}, \quad V = \begin{pmatrix} \theta(\frac{1}{n\lambda} + \frac{\lambda}{m}) & \frac{1}{n} - \frac{\lambda^2}{m} \\ \frac{1}{n} - \frac{\lambda^2}{m} & \frac{\lambda(m+n\lambda^2)}{nm\theta} \end{pmatrix}.$$

and, therefore, using (30),

$$\pi(\lambda \,|\, \theta) \propto I_{22}(\theta, \lambda)^{1/2} \propto \theta^{1/2}(m + n\lambda^2)^{1/2}\lambda^{-3/2}. \tag{36}$$

The natural increasing sequence of subsets of the original parameter space, $\Omega_i = \{(\alpha, \beta);\ 0 < \alpha < i,\ 0 < \beta < i\}$, transforms, in the parameter space of λ, into the sequence $\Lambda_i(\theta) = \{\lambda;\ \theta\, i^{-2} < \lambda < i^2\, \theta^{-1}\}$. Notice that this depends on θ, so that θ and λ are *not* variation independent and, hence, Theorem 12 *cannot* be applied. Renormalizing (36) in $\Lambda_i(\theta)$ and using (31), it is found that, for large i,

$$\begin{aligned} \pi_i(\lambda \,|\, \theta) &= c_i(m, n)\, \theta^{1/2}(m + n\lambda^2)^{1/2}\lambda^{-3/2} \\ \pi_i(\theta) &= c_i(m, n) \int_{\Lambda_i(\theta)} (m + n\lambda^2)^{1/2}\lambda^{-3/2}\ \log\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{-1/2} d\lambda, \end{aligned}$$

where $c_i(m, n) = i^{-1}\sqrt{nm}/(\sqrt{m} + \sqrt{n})$, which leads to the θ-reference prior $\pi^\theta(\theta, \lambda) \propto \theta^{1/2}\lambda^{-1}\left(\frac{\lambda}{m} + \frac{1}{n\lambda}\right)^{1/2}$. In the original parametrization, this corresponds to

$$\pi^\theta(\alpha, \beta) \propto (n\alpha^2 + m\beta^2)^{1/2}, \quad n \geq 1,\ m \geq 1 \tag{37}$$

which depends on the sample sizes through the ratio m/n. It has already been stressed that the reference prior depends on the experimental design. It is therefore not surprising that, if the design is unbalanced, the reference prior depends on the ratio m/n which controls the level of balance. Notice that the reference prior (37) is very different from the uniform prior $\pi^\alpha(\alpha, \beta) = \pi^\beta(\alpha, \beta) = 1$, which should be used to make reference inferences about either α or β.

It will later be demonstrated (Example 22) that the prior $\pi^\theta(\alpha, \beta)$ found above provides approximate agreement between Bayesian credible regions and frequentist confidence intervals for θ (Berger and Bernardo, 1989); indeed, this prior was originally suggested by Stein (1986) (who only considered the case $m = n$) to obtain such approximate agreement. Efron (1986) used this problem as an example in which *conventional* objective Bayesian theory encounters difficulties since, even within a fixed model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, the "correct" objective prior depends on the particular function $\phi = \phi(\boldsymbol{\theta})$ one

desires to estimate. For the reference priors associated to generalizations of the product of normal means problem, see Sun and Ye (1995, 1999).

3.9 Many parameters

Theorems 11 and 12 may easily be extended to any number of nuisance parameters. Indeed, let data $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ consist of a random sample of size n from a model $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\omega}),\, \boldsymbol{y} \in \boldsymbol{\mathcal{Y}},\, \boldsymbol{\omega} = \{\omega_1, \ldots, \omega_m\}, \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, let ω_1 be the quantity of interest, assume regularity conditions to guarantee that, as $n \to \infty$, the joint posterior distribution of $\boldsymbol{\omega}$ is asymptotically normal with mean $\hat{\boldsymbol{\omega}}_n$ and dispersion matrix $V(\hat{\boldsymbol{\omega}}_n)/n$, and let $H(\boldsymbol{\omega}) = V^{-1}(\boldsymbol{\omega})$. It then follows that, if $V_j(\boldsymbol{\omega})$ is the $j \times j$ upper matrix of $V(\boldsymbol{\omega})$, $j = 1, \ldots, m$, $H_j(\boldsymbol{\omega}) = V_j^{-1}(\boldsymbol{\omega})$ and $h_{jj}(\boldsymbol{\omega})$ is the lower right (j, j) element of $H_j(\boldsymbol{\omega})$, then

(1) the *conditional* posterior distribution of ω_j given $\{\omega_1, \ldots, \omega_{j-1}\}$, is asymptotically normal with precision $n\, h_{jj}(\hat{\boldsymbol{\omega}}_n)$, $(j = 2, \ldots, m)$ and
(2) the *marginal* posterior distribution of ω_1 is asymptotically normal with precision $n\, h_{11}(\hat{\boldsymbol{\omega}}_n)$.

This may be used to extend the algorithm described in Theorem 11 to sequentially derive $\pi(\omega_m \,|\, \omega_1, \ldots, \omega_{m-1})$, $\pi(\omega_{m-1} \,|\, \omega_1, \ldots, \omega_{m-2}), \ldots, \pi(\omega_2 \,|\, \omega_1)$ and $\pi(\omega_1)$; their product yields the reference prior associated to the particular ordering $\{\omega_1, \omega_2, \ldots, \omega_m\}$. Intuitively, this is a mathematical description of a situation where, relative to the particular design considered $\mathcal{M}$, one maximizes the missing information about the parameter ω_1 (that of higher inferential importance), but also the missing information about ω_2 given ω_1, that of ω_3 given ω_1 and ω_2,... and that of ω_m given ω_1 to ω_{m-1}. As in sequential decision theory, this must be done backwards. In particular, to maximize the missing information about ω_1, the prior which maximizes the missing information about ω_2 given ω_1 has to be derived first.

The choice of the ordered parametrization, say $\{\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$, precisely describes the particular prior required, namely that which sequentially maximizes the missing information about the θ_j's in order of inferential interest. Indeed, "diffuse" prior knowledge about a particular sequence $\{\theta_1(\boldsymbol{\omega}), \theta_2(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ may be very "precise" knowledge about another sequence $\{\phi_1(\boldsymbol{\omega}), \phi_2(\boldsymbol{\omega}), \ldots, \phi_m(\boldsymbol{\omega})\}$ unless, *for all* j, $\phi_j(\boldsymbol{\omega})$ is a one-to-one function of $\theta_j(\boldsymbol{\omega})$. Failure to recognize this fact is known to produce untenable results; famous examples are the paradox of Stein (1959) (see Example 19 below) and the marginalization paradoxes (see Example 16).

Theorem 14 (Reference priors under asymptotic normality) *Let data* $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ *consist of a random sample of size* n *from a statistical model* $\mathcal{M} \equiv \{p(\boldsymbol{y} \,|\, \boldsymbol{\theta}), \boldsymbol{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}, \boldsymbol{\theta} \in \boldsymbol{\Theta} = \prod_{j=1}^{m} \Theta_j\}$, *and let* $\mathcal{P}_0$ *be the class of all continuous priors with support* $\boldsymbol{\Theta}$. *If the posterior distribution of* $\boldsymbol{\theta}$ *is asymptotically normal with dispersion matrix* $V(\hat{\boldsymbol{\theta}}_n)/n$, *where* $\hat{\boldsymbol{\theta}}_n$ *is a consistent estimator of* $\boldsymbol{\theta}$, $H(\boldsymbol{\theta}) = V^{-1}(\boldsymbol{\theta})$, V_j *is the upper* $j \times j$ *submatrix of* V,

$H_j = V_j^{-1}$, and $h_{jj}(\boldsymbol{\theta})$ is the lower right element of H_j, then the $\boldsymbol{\theta}$-reference prior, associated to the ordered parametrization $\{\theta_1, \ldots, \theta_m\}$, is

$$\pi(\boldsymbol{\theta} \,|\, \mathcal{M}^n, \mathcal{P}_0) = \pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) \times \cdots \times \pi(\theta_2 \,|\, \theta_1)\, \pi(\theta_1),$$

with $\pi(\theta_m \,|\, \theta_1, \ldots, \theta_{m-1}) = h_{mm}^{1/2}(\boldsymbol{\theta})$ and, for $i = 1, \ldots, m-1$,

$$\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1}) \propto \exp\left\{ \int_{\boldsymbol{\Theta}^{j+1}} \prod_{l=j+1}^{m} \pi(\theta_l \,|\, \theta_1, \ldots, \theta_{l-1}) \log[h_{jj}^{1/2}(\boldsymbol{\theta})]\, d\boldsymbol{\theta}^{j+1} \right\}$$

with $\boldsymbol{\theta}^{j+1} = \{\theta_{j+1}, \ldots, \theta_m\}$, provided $\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})$ is proper for all j.
If the conditional reference priors $\pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})$ are not all proper, integration is performed on elements of an increasing sequence $\{\boldsymbol{\Theta}_i\}_{i=1}^{\infty}$ such that $\int_{\Theta_{ij}} \pi(\theta_j \,|\, \theta_1, \ldots, \theta_{j-1})\, d\theta_j$ is finite, to obtain the corresponding sequence $\{\pi_i(\boldsymbol{\theta})\}_{i=1}^{\infty}$ of reference priors for the restricted models. The $\boldsymbol{\theta}$-reference prior is then defined as their intrinsic limit.
If, moreover, (i) *Θ_j does not depend on $\{\theta_1, \ldots, \theta_{j-1}\}$, and* (ii) *the functions $h_{jj}(\theta, \lambda)$ factorize in the form*

$$h_{jj}^{1/2}(\boldsymbol{\theta}) \propto f_j(\theta_j)\, g_j(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_m), \quad j = 1, \ldots, m,$$

then the $\boldsymbol{\theta}$-reference prior is simply $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \prod_{j=1}^{m} f_j(\theta_j)$, even if the conditional reference priors are improper.

Under appropriate regularity conditions—see *e.g.*, Bernardo and Smith (1994, Theo. 5.14)—the posterior distribution of $\boldsymbol{\theta}$ is asymptotically normal with mean the mle $\hat{\boldsymbol{\theta}}_n$ and precision matrix $n\, I(\hat{\boldsymbol{\theta}}_n)$, where $I(\boldsymbol{\theta})$ is Fisher matrix,

$$i_{ij}(\boldsymbol{\theta}) = -\int_{\boldsymbol{\mathcal{Y}}} p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\, \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log[p(\boldsymbol{y} \,|\, \boldsymbol{\theta})]\, d\boldsymbol{y};$$

in that case, $H(\boldsymbol{\theta}) = n\, I(\boldsymbol{\theta})$, and the reference prior may be computed from the elements of Fisher matrix $I(\boldsymbol{\theta})$. Notice, however, that in the multivariate case, the reference prior does *not* yield Jeffreys multivariate rule (Jeffreys, 1961), $\pi^J(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$. For instance, in location-scale models, the (μ, σ)-reference prior and the (σ, μ)-reference prior are both $\pi^R(\mu, \sigma) = \sigma^{-1}$ (Theorem 13), while Jeffreys multivariate rule yields $\pi^J(\mu, \sigma) = \sigma^{-2}$. As a matter of fact, Jeffreys himself criticised his own multivariate rule. This is known, for instance, to produce both marginalization paradoxes Dawid, Stone and Zidek (1973), and strong inconsistencies (Eaton and Freedman, 2004). See, also, Stein (1962) and Example 23.

Theorem 14 provides a procedure to obtain the reference prior $\pi^{\boldsymbol{\theta}}(\boldsymbol{\theta})$ which corresponds to any *ordered parametrization* $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$. Notice that, within any particular multiparameter model

$$\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}), \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \quad \boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\} \in \boldsymbol{\Theta} \subset \mathbb{R}^k\},$$

the reference algorithm provides a (possibly different) joint reference prior

$$\pi^{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \pi(\phi_m \,|\, \phi_1, \ldots, \phi_{m-1}) \times \cdots \times \pi(\phi_2 \,|\, \phi_1)\, \pi(\phi_1),$$

for each possible ordered parametrization $\{\phi_1(\boldsymbol{\theta}), \phi_2(\boldsymbol{\theta}), \ldots, \phi_m(\boldsymbol{\theta})\}$. However, as one would hope, the results are coherent under monotone transformations of each of the $\phi_i(\boldsymbol{\theta})$'s in the sense that, in that case, $\pi^{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \pi^{\boldsymbol{\theta}}[\,\boldsymbol{\theta}(\boldsymbol{\phi})\,]|J(\boldsymbol{\phi})|$, where $J(\boldsymbol{\phi})$ is the Jacobian of the inverse transformation $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\phi})$, of general element $j_{ij}(\boldsymbol{\phi}) = \partial\theta_i(\boldsymbol{\phi})/\partial\phi_j$. This property of coherence under appropriate reparametrizations may be very useful in choosing a particular parametrization (for instance one with orthogonal parameters, or one in which the relevant $h_{jj}(\boldsymbol{\theta})$ functions factorize) which simplifies the implementation of the algorithm.

Starting with Jeffreys (1946) pioneering work, the analysis of the invariance properties under reparametrization of multiparameter objective priors has a very rich history. Relevant pointers include Hartigan (1964), Stone (1965, 1970), Zidek (1969), Florens (1978, 1982), Dawid (1983), Consonni and Veronese (1989b), Chang and Eaves (1990), George and McCulloch (1993), Datta and J. K. Ghosh (1995b), Yang (1995), Datta and M. Ghosh (1996), Eaton and Sudderth (1999, 2002, 2004) and Severini, Mukerjee and Ghosh (2002). In particular, Datta and J. K. Ghosh (1995b), Yang (1995) and Datta and M. Ghosh (1996) are specifically concerned with the invariance properties of reference distributions.

Example 18 *Multivariate normal data.* Let data consist of a size n random sample $\boldsymbol{x} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, $n \geq 2$, from an m-variate normal distribution with mean $\boldsymbol{\mu}$, and covariance matrix $\sigma^2\,\mathbf{I}_m$, $m \geq 1$, so that

$$I(\boldsymbol{\mu}, \sigma) = \begin{pmatrix} \sigma^{-2}\,\mathbf{I}_m & 0 \\ 0 & (2/m)\,\sigma^{-2} \end{pmatrix}$$

It follows from Theorem 14 that the reference prior relative to the natural parametrization $\boldsymbol{\theta} = \{\mu_1, \ldots, \mu_m, \sigma\}$ is $\pi^{\boldsymbol{\theta}}(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$, and also that the result does not depend on the order in which the parametrization is taken, since their asymptotic covariances are zero. Hence, $\pi^{\boldsymbol{\theta}}(\mu_1, \ldots, \mu_m, \sigma) \propto \sigma^{-1}$ is the appropriate prior function to obtain the reference posterior of any piecewise invertible function $\phi(\mu_j)$ of μ_j, and also to obtain the reference posterior of any piecewise invertible function $\phi(\sigma)$ of σ. In particular, the corresponding reference posterior for any of the μ_j's is easily shown to be the Student density

$$\pi(\mu_j \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathrm{St}\left\{\mu_j \;\middle|\; \overline{y}_j,\, s/\sqrt{(n-1)},\, m(n-1)\right\}$$

with $n\overline{y}_j = \sum_{i=1}^{n} y_{ij}$, and $nms^2 = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \overline{y}_j)^2$, which agrees with the standard argument according to which one degree of freedom should

be lost by each of the unknown means. Similarly, the reference posterior of σ^2 is the inverted Gamma

$$\pi(\sigma^2 \,|\, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = \mathrm{IGa}\{\sigma^2 \,|\, n(m-1)/2,\, nms^2/2\}$$

When $m = 1$, these results reduce to those obtained in Example 16.

Example 19 *Stein's paradox.* Let $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ be a random sample of size n from a m-variate normal $\mathrm{N}_m(\boldsymbol{x} \,|\, \boldsymbol{\mu}, I_m)$ with mean $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_m\}$ and unitary dispersion matrix. The reference prior which corresponds to any permutation of the μ_i's is uniform, and this uniform prior leads indeed to appropriate reference posterior distributions for any of the μ_j's, given by $\pi(\mu_j \,|\, \boldsymbol{x}) = \mathrm{N}(\mu_j \,|\, \overline{x}_j, 1/\sqrt{n})$. Suppose, however, that the quantity of interest is $\theta = \sum_j \mu_j^2$, the distance of $\boldsymbol{\mu}$ from the origin. As shown by Stein (1959), the posterior distribution of θ based on the uniform prior (or indeed any "flat" *proper* approximation) has very undesirable properties; this is due to the fact that a uniform (or nearly uniform) prior, although "noninformative" with respect to each of the individual μ_j's, is actually highly informative on the sum of their squares, introducing a severe bias towards large values of θ (Stein's paradox). However, the reference prior which corresponds to a parametrization of the form $\{\theta, \boldsymbol{\lambda}\}$ produces, for any choice of the nuisance parameter vector $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$, the reference posterior for the quantity of interest $\pi(\theta \,|\, \boldsymbol{x}) = \pi(\theta \,|\, t) \propto \theta^{-1/2} \chi^2(n\, t \,|\, m, n\, \theta)$, where $t = \sum_i \overline{x}_i^2$, and this posterior is shown to have the appropriate consistency properties. For further details see Ferrándiz (1985).

If the μ_i's were known to be related, so that they could be assumed to be exchangeable, with $p(\boldsymbol{\mu}) = \prod_{i=1}^m p(\mu_i \,|\, \boldsymbol{\phi})$, for some $p(\mu \,|\, \boldsymbol{\phi})$, one would have a (very) different (hierarchical) model. Integration of the μ_i's with $p(\boldsymbol{\mu})$ would then produce a model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\phi}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$ parametrized by $\boldsymbol{\phi}$, and only the corresponding reference prior $\pi(\boldsymbol{\phi} \,|\, \mathcal{M})$ would be required. See below (Subsection 3.12) for further discussion on reference priors in hierarchical models.

Far from being specific to Stein's example, the inappropriate behaviour in problems with many parameters of specific marginal posterior distributions derived from multivariate "flat" priors (proper or improper) is very frequent. Thus, as indicated in the introduction, uncritical use of "flat" priors (rather than the relevant reference priors), should be very strongly discouraged.

3.10 Discrete parameters taking an infinity of values

Due to the non-existence of an asymptotic theory comparable to that of the continuous case, the infinite discrete case presents special problems. However, it is often possible to obtain an approximate reference posterior by embedding the discrete parameter space within a continuous one.

Example 20 *Discrete parameters taking an infinite of values.* In the context of capture-recapture models, it is of interest to make inferences about the population size $\theta \in \{1, 2, \ldots\}$ on the basis of data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, which are assumed to consist of a random sample from

$$p(x \,|\, \theta) = \frac{\theta(\theta+1)}{(x+\theta)^2}\,, \quad 0 \le x \le 1.$$

This arises, for instance, in software reliability, when the unknown number θ of bugs is assumed to be a continuous mixture of Poisson distributions. Goudie and Goldie (1981) concluded that, in this problem, all standard non-Bayesian methods are liable to fail; Raftery (1988) finds that, for several plausible "diffuse looking" prior distributions for the discrete parameter θ, the corresponding posterior virtually ignores the data; technically, this is due to the fact that, for most samples, the corresponding likelihood function $p(\boldsymbol{x} \,|\, \theta)$ tends to one (rather than to zero) as $\theta \to \infty$. Embedding the discrete parameter space $\Theta = \{1, 2, \ldots\}$ into the continuous space $\Theta = (0, \infty)$ (since, for each $\theta > 0$, $p(x|\theta)$ is still a probability density for x), and using Theorem 9, the appropriate reference prior is

$$\pi(\theta) \propto i(\theta)^{1/2} \propto (\theta+1)^{-1}\theta^{-1},$$

and it is easily verified that this prior leads to a posterior in which the data are no longer overwhelmed. If the problem requires the use of discrete θ values, the discrete approximation $\Pr(\theta = 1 \,|\, \boldsymbol{x}) = \int_0^{3/2} \pi(\theta \,|\, \boldsymbol{x})\, d\theta$, and $\Pr(\theta = j \,|\, \boldsymbol{x}) = \int_{j-1/2}^{j+1/2} \pi(\theta \,|\, \boldsymbol{x})\, d\theta$, $j > 1$, may be used as an approximate discrete reference posterior, specially when interest mostly lies on large θ values, as it is typically the case.

3.11 Behaviour under repeated sampling

The frequentist coverage probabilities of the different types of credible intervals which may be derived from reference posterior distributions are sometimes identical, and usually very close, to their posterior probabilities; this means that, even for moderate samples, an interval with reference posterior probability q may often be interpreted as an *approximate* frequentist confidence interval with significance level $1 - q$.

Example 21 *Coverage in simple normal problems.* Consider again inferences about the mean μ and the variance σ^2 of a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ model. Using the reference prior $\pi^\mu(\mu, \sigma) \propto \sigma^1$ derived in Example 16, the reference posterior distribution of μ after a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ has been observed, $\pi(\mu \,|\, \boldsymbol{x}) \propto \int_0^\infty \prod_{j=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma)\, \pi^\mu(\mu, \sigma)\, d\sigma$, is the Student density $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}(\mu \,|\, \overline{x}, s/\sqrt{n-1}, n-1) \propto [s^2 + (\overline{x} - \mu)^2]^{-n/2}$, where $\overline{x} = \sum_j x_j/n$, and $s^2 = \sum_j (x_j - \overline{x})^2/n$. Hence, the reference *pos-*

terior of the standardized function of μ, $\phi(\mu) = \sqrt{n-1}\,(\mu - \overline{x})/s$ is standard Student with $n-1$ degrees of freedom. But, conditional on μ, the *sampling* distribution of $t(\boldsymbol{x}) = \sqrt{n-1}\,(\mu - \overline{x})/s$ is *also* standard Student with $n-1$ degrees of freedom. It follows that, for all sample sizes, posterior reference credible intervals for μ will numerically be identical to frequentist confidence intervals based on the sampling distribution of t. Similar results are obtained concerning inferences about σ: the reference posterior distribution of $\psi(\sigma) = ns^2/\sigma^2$ is a χ^2 with $n-1$ degrees of freedom but, conditional on σ, this is also the sampling distribution of $r(\boldsymbol{x}) = ns^2/\sigma^2$.

The *exact* numerical agreement between reference posterior credible intervals and frequentist confidence intervals shown in Example 21 is however the exception, not the norm. Nevertheless, for *large* sample sizes, reference credible intervals are always *approximate* confidence intervals.

More precisely, let data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ consist of n independent observations from $\mathcal{M} = \{p(x \,|\, \theta), x \in \mathcal{X}, \theta \in \Theta\}$, and let $\theta_q(\boldsymbol{x},\, p_\theta)$ denote the q quantile of the posterior $p(\theta \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \theta)\, p(\theta)$ which corresponds to the prior $p(\theta)$, so that

$$\Pr\left[\theta \leq \theta_q(\boldsymbol{x},\, p_\theta) \,|\, \boldsymbol{x}\right] = \int_{\theta \leq \theta_q(\boldsymbol{x},\, p_\theta)} p(\theta \,|\, \boldsymbol{x})\, d\theta = q.$$

Standard asymptotic theory may be used to establish that, for any sufficiently regular pair $\{p_\theta,\, \mathcal{M}\}$ of prior p_θ and model $\mathcal{M}$, the *coverage* probability of the region thus defined, $R_q(\boldsymbol{x},\, \theta,\, p_\theta) = \{\boldsymbol{x};\ \theta \leq \theta_q(\boldsymbol{x},\, p_\theta)\}$, converges to q as $n \to \infty$. Specifically, for all sufficiently regular priors,

$$\Pr\left[\theta_q(\boldsymbol{x}, p_\theta) \geq \theta \,|\, \theta\right] = \int_{R_q(\boldsymbol{x},\, \theta,\, p_\theta)} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x} = q + O(n^{-1/2}).$$

It has been found however that, when there are no nuisance parameters, the reference prior π^θ typically satisfies

$$\Pr\left[\theta_q(\boldsymbol{x}, \pi^\theta) \geq \theta \,|\, \theta\right] = q + O(n^{-1});$$

this means that the reference prior is often a *probability matching* prior, that is, a prior for which the coverage probabilities of *one-sided* posterior credible intervals are asymptotically closer to their posterior probabilities. Hartigan (1966) showed that the coverage probabilities of *two-sided* Bayesian posterior credible intervals satisfy this type of approximation to $O(n^{-1})$ for *all* sufficiently regular prior functions.

In a pioneering paper, Welch and Peers (1963) established that in the case of the one-parameter regular continuous models Jeffreys prior (which in this case, Theorem 9, is also the reference prior), is the only probability matching prior. Hartigan (1983, p. 79) showed that this result may be extended

to one-parameter discrete models by using continuity corrections. Datta and J. K. Ghosh (1995a) derived a differential equation which provides a necessary and sufficient condition for a prior to be probability matching in the multi-parameter continuous regular case; this has been used to verify that reference priors are typically probability matching priors.

In the nuisance parameter setting, reference priors are sometimes matching priors fro the parameter of interest, but in this general situation, matching priors may not always exist or be unique (Welch, 1965; Ghosh and Mukerjee, 1998). For a review of probability matching priors, see Datta and Sweeting (2005), in this volume.

Although the results described above only justify an *asymptotic* approximate frequentist interpretation of reference posterior probabilities, the coverage probabilities of reference posterior credible intervals derived from relatively small samples are also found to be typically close to their posterior probabilities. This is now illustrated within the product of positive normal means problem, already discussed in Example 17.

Example 22 *Product of normal means, continued.* Let available data $\boldsymbol{x} = \{x, y\}$ consist of one observation x from $\mathrm{N}(x \,|\, \alpha, 1)$, $\alpha > 0$, and another observation y from $\mathrm{N}(y \,|\, \beta, 1)$, $\beta > 0$, and suppose that the quantity of interest is the product of the means $\theta = \alpha\,\beta$. The behaviour under repeated sampling of the posteriors which correspond to both the conventional uniform prior $\pi^u(\alpha, \beta) = 1$, and the reference prior $\pi^\theta(\alpha, \beta) = (\alpha^2 + \beta^2)^{1/2}$ (see Example 17) is analyzed by computing the coverage probabilities $\Pr[R_q \,|\, \theta, p_\theta] = \int_{R_q(\boldsymbol{x}, \theta, p_\theta)} p(\boldsymbol{x} \,|\, \theta)\, d\boldsymbol{x}$ associated to the regions $R_q(\boldsymbol{x}, \theta, p_\theta) = \{\boldsymbol{x};\, \theta \leq \theta_q(\boldsymbol{x}, p_\theta)\}$ defined by their corresponding quantiles, $\theta_q(\boldsymbol{x}, \pi^u)$ and $\theta_q(\boldsymbol{x}, \pi^\theta)$. Table 1 contains the coverage probabilities of the regions defined by the 0.05 posterior quantiles. These have been numerically computed by simulating $4,000$ pairs $\{x, y\}$ from $\mathrm{N}(x \,|\, \alpha, 1)\mathrm{N}(y \,|\, \beta, 1)$ for each of the $\{\alpha, \beta\}$ pairs listed in the first column of the table.

Table 1 *Coverage probabilities of* 0.05*-credible regions for* $\theta = \alpha\,\beta$.

$\{\alpha, \beta\}$	$\Pr[R_{0.05} \,|\, \theta, \pi^u]$	$\Pr[R_{0.05} \,|\, \theta, \pi^\theta]$
$\{1, 1\}$	0.024	0.047
$\{2, 2\}$	0.023	0.035
$\{3, 3\}$	0.028	0.037
$\{4, 4\}$	0.033	0.048
$\{5, 5\}$	0.037	0.046

The standard error of the entries in the table is about 0.0035. It may be observed that the estimated coverages which correspond to the reference prior are appreciably closer to the nominal value 0.05 that those corresponding to the uniform prior. Notice that, although it may be shown that the reference prior *is* probability matching in the technical sense described

above, the empirical results shown in the Table do *not* follow from that fact, for probability matching is an *asymptotic* result, and one is dealing here with samples of size $n = 1$. For further details on this example, see Berger and Bernardo (1989).

3.12 Prediction and hierarchical models

Two classes of problems that are not specifically covered by the methods described above are hierarchical models and prediction problems. The difficulty with these problems is that the distributions of the quantities of interest must belong to specific families of distributions. For instance, if one wants to predict the value of y based on $\boldsymbol{x}$ when $(y, \boldsymbol{x})$ has density $p(y, \boldsymbol{x} \,|\, \boldsymbol{\theta})$, the unknown of interest is y, but its distribution is conditionally specified; thus, one needs a prior for $\boldsymbol{\theta}$, not a prior for y. Likewise, in a hierarchical model with, say, $\{\mu_1, \mu_2, \ldots, \mu_p\}$ being $\mathrm{N}(\mu_i \,|\, \theta, \lambda)$, the μ_i's may be the parameters of interest, but a prior is only needed for the hyperparameters θ and λ.

In hierarchical models, the parameters with conditionally known distributions may be integrated out (which leads to the so-called marginal overdispersion models). A reference prior for the remaining parameters based on this marginal model is then required. The difficulty that arises is how to then identify parameters of interest and nuisance parameters to construct the ordering necessary for applying the reference algorithm, the real parameters of interest having been integrated out.

A possible solution to the problems described above is to define the quantity of interest to be the conditional mean of the original parameter of interest. Thus, in the prediction problem, the quantity of interest could be defined to be $\phi(\boldsymbol{\theta}) = \mathrm{E}[y|\boldsymbol{\theta}]$, which will be either $\boldsymbol{\theta}$ or some transformation thereof, and in the hierarchical model mentioned above the quantity of interest could be defined to be $\mathrm{E}[\mu_i \,|\, \theta, \lambda] = \theta$. More sophisticated choices, in terms of appropriately chosen discrepancy functions, are currently under scrutiny.

Bayesian prediction with objective priors is a very active research area. Pointers to recent suggestions include Kuboki (1998), Eaton and Sudderth (1998, 1999) and Smith (1999). Under appropriate regularity conditions, some of these proposals lead to Jeffreys multivariate prior, $\pi(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$. However, the use of that prior may lead to rather unappealing predictive posteriors as the following example demonstrates.

Example 23 *Normal prediction.* Let available data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from $\mathrm{N}(x_j \,|\, \mu, \sigma)$, and suppose that one is interested in predicting a new, future observation x from $\mathrm{N}(x \,|\, \mu, \sigma)$. Using the argument described above, the quantity of interest could be defined to be $\phi(\mu, \sigma) = \mathrm{E}[x \,|\, \mu, \sigma] = \mu$ and hence (see Example 16) the appropriate reference prior would be $\pi^x(\mu, \sigma) = \sigma^{-1}$ $(n \geq 2)$. The corresponding joint reference posterior is $\pi(\mu, \sigma \,|\, \boldsymbol{x}) \propto \prod_{j=1}^{n} \mathrm{N}(x_j \,|\, \mu, \sigma)\, \sigma^{-1}$ and the posterior

predictive distribution is

$$\begin{aligned}\pi(x \,|\, \boldsymbol{x}) &= \int_0^\infty \int_{-\infty}^\infty \mathrm{N}(x \,|\, \mu, \sigma)\, \pi(\mu, \sigma \,|\, \boldsymbol{x})\, d\mu\, d\sigma \\ &\propto \{(n+1)s^2 + (\overline{x} - \mu)^2)\}^{-n/2}, \\ &\propto \mathrm{St}(x \,|\, \overline{x},\, s\{(n+1)/(n-1)\}^{1/2},\, n-1), \quad n \geq 2 \end{aligned} \tag{38}$$

where, as before, $\overline{x} = n^{-1} \sum_{j=1}^n x_j$ and $s^2 = n^{-1} \sum_{j=1}^n (x_j - \overline{x})^2$. As one would expect, the reference predictive distribution (38) is proper whenever $n \geq 2$: in the absence of prior knowledge, $n = 2$ is the minimum sample size required to identify the two unknown parameters.

It may be verified that the predictive posterior (38) has consistent coverage properties. For instance, with $n = 2$, the reference posterior predictive probability that a third observation lies within the first two is

$$\Pr[x_{(1)} < x < x_{(2)} \,|\, x_1, x_2] = \int_{x_{(1)}}^{x_{(2)}} \pi(x \,|\, x_1, x_2)\, dx = \tfrac{1}{3}\,,$$

where $x_{(1)} = \min[x_1, x_2]$, and $x_{(2)} = \max[x_1, x_2]$. This is consistent with the fact that, for all μ and σ, the frequentist coverage of the corresponding region of $\mathbb{R}^3$ is precisely

$$\int\int\int_{\{(x_1,x_2,x_3);\ x_{(1)} < x_3 < x_{(2)}\}} \prod_{i=1}^3 \mathrm{N}(x_j \,|\, \mu, \sigma)\, dx_1\, dx_2\, dx_3 = \tfrac{1}{3}\,. \tag{39}$$

In sharp contrast, if Jeffreys multivariate rule $\pi^J(\mu, \sigma) \propto |I(\mu, \sigma)|^{1/2} = \sigma^{-2}$ were used, the posterior predictive would have been a Student t centred at $\overline{x}$, with scale $s\{(n+1)/n\}^{1/2}$, and with n degrees of freedom, which is proper whenever $n \geq 1$. Thus, with $\pi^J(\mu, \sigma)$ as a prior, probabilistic predictions would be possible with only *one* observation, rather unappealing when no prior knowledge is assumed. Moreover, the probability that a third observation lies within the first two which corresponds to the prior $\pi^J(\mu, \sigma)$ is $1/2$, rather than $1/3$, a less than attractive result in view of (39).

For a recent predictive probability matching approach to objective predictive posteriors, see Datta, Mukerjee, Ghosh and Sweeting (2000).

4 Reference Inference Summaries

From a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is nothing but the posterior distribution of that quantity. Thus, given some data $\boldsymbol{x}$ and conditions C, *all* that can be said about any function $\boldsymbol{\theta}(\boldsymbol{\omega})$ of the parameters $\boldsymbol{\omega}$ which govern the model is contained in the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{x}, C)$, and *all* that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y} \,|\, \boldsymbol{x}, C)$. In fact (Bernardo, 1979a),

Bayesian inference may be described as a decision problem where the space of available actions is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution, while retaining as much of the information as possible. This is conventionally done by (i) providing values of the quantity of interest which, in the light of the data, are likely to be "close" to its true value, and (ii) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. In this section, objective Bayesian counterparts to these traditional inference problems of *estimation* and *testing*, which are based on the joint use of intrinsic loss functions and reference analysis, are briefly considered.

4.1 Point Estimation

Let $\boldsymbol{x}$ be the available data, which are assumed to consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\omega} \in \Omega\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be the quantity of interest. Without loss of generality, the original model $\mathcal{M}$ may be written as $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, in terms of the quantity of interest $\boldsymbol{\theta}$ and a vector $\boldsymbol{\lambda}$ of nuisance parameters. A *point estimate* of $\boldsymbol{\theta}$ is some value $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ which could possibly be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$.

Formally, to choose a point estimate for $\boldsymbol{\theta}$ is a *decision problem*, where the action space is the class Θ of possible $\boldsymbol{\theta}$ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\boldsymbol{\theta}}$ of some quantity $\boldsymbol{\theta}$ is a *decision* to act as if $\tilde{\boldsymbol{\theta}}$ were $\boldsymbol{\theta}$, not to assert something about the value of $\boldsymbol{\theta}$ (although desire to assert something simple may well be the main reason to obtain an estimate). To solve this decision problem it is necessary to specify a *loss function* $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\boldsymbol{\theta}}$, when it is actually $\boldsymbol{\theta}$. The expected posterior loss if $\tilde{\boldsymbol{\theta}}$ were used is $l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}] = \int_{\Theta} \ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \, d\boldsymbol{\theta}$, and the *Bayes estimate* is that $\tilde{\boldsymbol{\theta}}$ value which minimizes $l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}]$ in $\boldsymbol{\Theta}$. The *Bayes estimator* is the function of the data $\boldsymbol{\theta}^*(\boldsymbol{x}) = \arg\min_{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} l[\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}]$.

For any given model and data, the Bayes estimate depends on the chosen loss function. The loss function is context specific, and should generally be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged. These loss functions produce estimates which may be regarded as simple descriptions of the *location* of the posterior distribution. For example, if the loss function is quadratic, so that $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, then the Bayes estimate is the *posterior mean* $\boldsymbol{\theta}^* = E[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that

$\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ if $\tilde{\boldsymbol{\theta}}$ belongs to a ball of radius ϵ centred in $\boldsymbol{\theta}$ and $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ otherwise, then the Bayes estimate $\boldsymbol{\theta}^*$ tends to the *posterior mode* as the radius of the ball tends to zero, assuming that a unique mode exists.

If θ is univariate and the loss function is linear, so that $\ell(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $\ell(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, then the Bayes estimate is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the Bayes estimate is the *posterior median*. The results just described for univariate linear loss functions clearly illustrate the fact that *any* possible parameter value may turn out be the Bayes estimate: it all depends on the loss function describing the consequences of the anticipated uses of the estimate.

Conventional loss functions are typically *not* invariant under reparametrization. As a consequence, the Bayes estimator ϕ^* of a one-to-one transformation $\phi = \phi(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\phi(\boldsymbol{\theta}^*)$ (the *univariate* posterior median, which *is* coherent under reparametrization, is an interesting exception). Moreover, conventional loss functions, such as the quadratic loss, focus attention on the discrepancy between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, rather than on the more relevant discrepancy between the statistical *models* they label. The intrinsic discrepancy $\delta\{\mathcal{M}_{\tilde{\boldsymbol{\theta}}},\ p_{\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}}\}$ (Definition 1) directly measures how different the probability *model* $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ is from its closest approximation within the family $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}),\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and its value does not depend on the particular parametrization chosen to describe the problem.

Definition 8 (Intrinsic estimation) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$ be the restricted model $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, and let*

$$\delta\{\tilde{\boldsymbol{\theta}},\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \delta\{\mathcal{M}_{\tilde{\boldsymbol{\theta}}},\ p_{\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\lambda}}\} = \min_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \delta\{p(\boldsymbol{x}\,|\,\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\} \tag{40}$$

be the intrinsic discrepancy between the distribution $p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set of distributions $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$. The reference posterior expected intrinsic loss is

$$d(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{x}) = \mathrm{E}[\delta\,|\,\boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta\{\tilde{\boldsymbol{\theta}},\ (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}, \tag{41}$$

where $\pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,\boldsymbol{x}) \propto p(\boldsymbol{x}\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the reference posterior of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ when δ is the quantity of interest. Given $\boldsymbol{x}$, the intrinsic estimate $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{x})$ *is that value $\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ which minimizes the posterior reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}}\,|\,\boldsymbol{x})$. As a function of $\boldsymbol{x}$, $\boldsymbol{\theta}^*(\boldsymbol{x})$ is the* intrinsic estimator *of $\boldsymbol{\theta}$.*

The intrinsic estimate is well defined for any dimensionality, and it is coherent under transformations, in the sense that, if $\boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, then the intrinsic estimate $\boldsymbol{\phi}^*$ of $\boldsymbol{\phi}(\boldsymbol{\theta})$ is simply $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$. Under broad regularity conditions (Juárez, 2004), the intrinsic estimator is admissible under the in-

trinsic loss. Moreover, the reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is typically a convex function of $\tilde{\boldsymbol{\theta}}$ in a neighbourhood of its minimum, in which case the intrinsic estimate $\boldsymbol{\theta}^*$ is unique, and it is easily derived by either analytical or numerical methods.

Example 24 *Intrinsic estimation of a binomial parameter.* Consider estimation of a binomial proportion θ from r successes given n trials; the reference prior (see Example 12) is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, the corresponding reference posterior is $\pi(\theta \,|\, n, r) = \mathrm{Be}(\theta \,|\, r + \frac{1}{2}, n - r + \frac{1}{2})$, and the quadratic loss based estimator (the posterior mean) of θ is $\mathrm{E}[\theta \,|\, n, r] = (r + 1/2)/(n + 1)$. However, the quadratic loss based estimator of the log-odds $\phi(\theta) = \log[\theta/(1-\theta)]$, is $\mathrm{E}[\phi \,|\, n, r] = \psi(r + 1/2) - \psi(n - r + 1/2)$ (where $\psi(x) = d \log[\Gamma(x)]/dx$ is the *digamma* function), which is *not* equal to $\phi(\mathrm{E}[\theta \,|\, n, r])$.

On the other hand the intrinsic discrepancy between two binomial distributions with parameters θ and $\tilde{\theta}$ and the same value of n, the loss to be suffered if $\tilde{\theta}$ were used as a proxy for θ, is $\delta\{\tilde{\theta}, \theta \,|\, n\} = n\, \delta_1\{\tilde{\theta}, \theta\}$, where (see Example 1)

$$\begin{aligned} \delta_1\{\theta_i, \theta_j\} &= \min\{k(\theta_i \,|\, \theta_i), k(\theta_j \,|\, \theta_i)\}, \\ k(\theta_i \,|\, \theta_j) &= \theta_j \log[\,\theta_j/\theta_i\,] + (1-\theta_j) \log[(1-\theta_j)/(1-\theta_i)]. \end{aligned}$$

The intrinsic estimator $\theta^* = \theta^*(r, n)$ is obtained by minimizing the reference expected posterior loss

$$d(\tilde{\theta} \,|\, n, r) = \int_0^1 \delta(\tilde{\theta}, \theta \,|\, n)\, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, d\theta. \tag{42}$$

Since intrinsic estimation is coherent under reparametrization, the intrinsic estimator of, say, the log-odds is simply the log-odds of the intrinsic estimator of θ. The exact value of θ^* may be easily obtained by numerical methods, but a very good linear approximation, based on the reference posterior mean of the approximate location parameter (Definition 7) $\phi(\theta) = \int_0^\theta \theta^{-1/2}(1-\theta)^{-1/2}\, d\theta = \frac{2}{\pi} \arcsin \sqrt{\theta}$, is

$$\theta^*(r, n) \approx \sin^2\{\tfrac{\pi}{2} E[\phi \,|\, r, n]\} \approx (r + \tfrac{1}{3})/(n + \tfrac{2}{3}). \tag{43}$$

The linear approximation (43) remains good even for small samples and extreme r values. For instance, the exact value of the intrinsic estimator with $r = 0$ and $n = 12$ (see Example 28 later in this section) is $\theta^* = 0.02631$, while the approximation yields 0.02632.

Example 25 *Intrinsic estimation of normal variance.* The intrinsic discrepancy $\delta\{p_1, p_2\}$ between two normal densities $p_1(x)$ and $p_2(x)$, with $p_i(x) = \mathrm{N}(x \,|\, \mu_i, \sigma_i)$, is $\delta\{p_1, p_2\} = \min\{k\{p_1 \,|\, p_2\},\ k\{p_2 \,|\, p_1\}\}$, where the

relevant Kullback-Leibler directed divergences are

$$k\{p_i \,|\, p_j\} = \int_{\mathcal{X}} p_j(x) \log \frac{p_j(x)}{p_i(x)}\, dx = \frac{1}{2}\Big\{ \log \frac{\sigma_i^2}{\sigma_j^2} + \frac{\sigma_j^2}{\sigma_i^2} - 1 + \frac{(\mu_i - \mu_j)^2}{\sigma_i^2} \Big\}.$$

The intrinsic discrepancy between the normal $\mathrm{N}(x \,|\, \mu, \sigma)$ and the set of normals with standard deviation $\tilde{\sigma}$, $\mathcal{M}_{\tilde{\sigma}} \equiv \{\mathrm{N}(x \,|\, \tilde{\mu}, \tilde{\sigma}),\ \tilde{\mu} \in \mathbb{R}\}$ is achieved when $\tilde{\mu} = \mu$, and is found to be

$$\delta\{\mathcal{M}_{\tilde{\sigma}},\ N(x \,|\, \mu, \sigma)\} = \delta(\theta) = \begin{cases} \frac{1}{2}[\log \theta^{-1} + \theta - 1],\ \theta < 1 \\ \frac{1}{2}[\log \theta + \theta^{-1} - 1],\ \theta \geq 1 \end{cases}$$

which only depends on the ratio $\theta = \tilde{\sigma}^2/\sigma^2$. Since, for any fixed $\tilde{\sigma}$, the intrinsic discrepancy, $\delta\{\tilde{\sigma}, (\mu, \sigma)\} = \delta(\theta)$ is a one-to-one function of σ, the reference prior when δ is the quantity of interest is $\pi^{\delta}(\mu, \sigma) = \sigma^{-1}$, the same as if the quantity of interest were σ (see Example 16). The corresponding posterior distribution of $\theta = \tilde{\sigma}^2/\sigma^2$, after a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of fixed size $n \geq 2$ has been observed, is the gamma density $\pi(\theta \,|\, \boldsymbol{x}) = \mathrm{Ga}(\theta \,|\, (n-1)/2, ns^2/\tilde{\sigma}^2)$, where $s^2 = \sum_j (x_j - \overline{x})^2/n$. The intrinsic estimate of σ is that value σ^* of $\tilde{\sigma}$ which minimizes the expected posterior loss,

$$\int_0^{\infty} \delta(\theta)\, \pi(\theta \,|\, \boldsymbol{x})\, d\theta = \int_0^{\infty} \delta(\theta)\, \mathrm{Ga}(\theta \,|\, (n-1)/2, ns^2/\tilde{\sigma}^2)\, d\theta.$$

The exact value of $\sigma^*(\boldsymbol{x})$ is easily obtained by one-dimensional numerical integration. However, for $n > 2$, a very good approximation is given by

$$\sigma^* = \sqrt{\frac{\sum_j (x_j - \overline{x})^2}{n-2}} \tag{44}$$

which is larger than both the mle estimate s (which divides by n the sum of squares) and the squared root of the conventional unbiased estimate of the variance (which divides by $n-1$). A good approximation for $n = 2$ is $\sigma^* = (\sqrt{5}/2)|x_1 - x_2|$. Since intrinsic estimation is coherent under one-to-one reparametrizations, the intrinsic estimator of the variance is $(\sigma^*)^2$, and the intrinsic estimator of, say, $\log \sigma$ is simply $\log \sigma^*$.

Intrinsic estimation is a very powerful, general procedure for objective, invariant point estimation. For further discussion, see Bernardo and Juárez (2003).

4.2 Region (interval) estimation

To describe the inferential content of the posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the quantity of interest it is often convenient to quote regions $R \subset \boldsymbol{\Theta}$ of given (posterior) probability under $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$. Any subset of the parameter space

$R_q \subset \boldsymbol{\Theta}$ such that $\int_{R_q} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = q$, $0 < q < 1$, so that, given data $\boldsymbol{x}$, the true value of $\boldsymbol{\theta}$ belongs to R_q with probability q, is said to be a (posterior) *q-credible region* of $\boldsymbol{\theta}$. Credible regions are coherent under reparametrization; thus, for any q-credible region R_q of $\boldsymbol{\theta}$ a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$, $\boldsymbol{\phi}(R_q)$ is a q-credible region of $\boldsymbol{\phi}$. However, for any given q there are generally infinitely many credible regions.

Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are *not* coherent under reparametrization: the image $\boldsymbol{\phi}(R_q)$ of an HPD q-credible region R_q will be a q-credible region for $\boldsymbol{\phi}$, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. In one dimension, posterior quantiles are often used to derive credible regions. Thus, if $\theta_q = \theta_q(\boldsymbol{x})$ is the $100q\%$ posterior quantile of θ, then $R_q^l = \{\theta;\, \theta \leq \theta_q\}$ is a one-sided, typically unique q-credible region, and it is coherent under reparametrization. *Probability centred* q-credible regions of the form $R_q^c = \{\theta;\, \theta_{(1-q)/2} \leq \theta \leq \theta_{(1+q)/2}\}$ are easier to compute, and are often quoted in preference to HPD regions. However, centred credible regions are only really appealing when the posterior density has a unique interior mode, and have a crucial limitation: they are not uniquely defined in problems with more than one dimension.

For reasonable loss functions, a typically unique credible region may be selected as a *lowest posterior loss* (LPL) region, where all points in the region have smaller (posterior) expected loss than all points outside.

Definition 9 (Intrinsic credible region) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_{\tilde{\boldsymbol{\theta}}}$ be the restricted model $\mathcal{M}_{\tilde{\boldsymbol{\theta}}} \equiv \{p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ and let $\delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_{\tilde{\boldsymbol{\theta}},}$. An* intrinsic q-credible region $R_q^* = R_q^*(\boldsymbol{x}) \subset \boldsymbol{\Theta}$ *is a subset of the parameter space $\boldsymbol{\Theta}$ such that,*

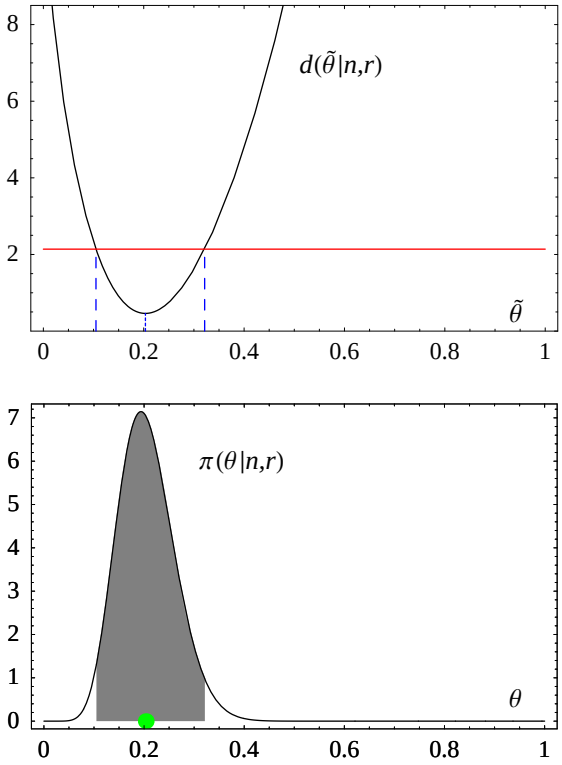
$$\int_{R_q^*(\boldsymbol{x})} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = q, \quad \forall \boldsymbol{\theta}_i \in R_q^*(\boldsymbol{x}),\ \forall \boldsymbol{\theta}_j \notin R_q^*(\boldsymbol{x}),\ d(\tilde{\boldsymbol{\theta}}_i \,|\, \boldsymbol{x}) \leq d(\tilde{\boldsymbol{\theta}}_j \,|\, \boldsymbol{x}),$$

where, as before, $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \mathrm{E}[\delta \,|\, \boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Lambda}} \delta\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$ is the reference posterior expected intrinsic loss.

Intrinsic credible regions are well defined for any dimensionality, and they are coherent under one-to-one transformations, in the sense that, if $\boldsymbol{\phi}\{\boldsymbol{\theta}\}$ is a one-to-one transformation of $\boldsymbol{\theta}$ and $R_q^* \subset \boldsymbol{\Theta}$ is an intrinsic q-credible region for $\boldsymbol{\theta}$, then $\boldsymbol{\phi}\{R_q^*\} \subset \boldsymbol{\Phi}$ is an intrinsic q-credible region for $\boldsymbol{\phi}$. As mentioned above, the reference expected intrinsic loss $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is often a convex function of $\tilde{\boldsymbol{\theta}}$; in that case, for each $q \in (0, 1)$ there is a unique (convex) intrinsic q-credible region.

Example 26 *Intrinsic binomial credible regions.* Let r be the number of successes observed in n independent Bernoulli trials with parameter θ.

Figure 4 *Intrinsic* 0.95*-credible region for a binomial parameter.*



As described in Example 24, the reference posterior expected intrinsic loss which corresponds to using $\tilde{\theta}$ instead of the actual (unknown) θ is the convex function $d\{\tilde{\theta} \,|\, n, r\}$ of Equation (42), which is represented in the upper panel of Figure 4 as a function of $\tilde{\theta}$, for $r = 10$ and $n = 50$. Using the invariance of the intrinsic loss with respect to one-to-one transformations, and a normal approximation to the posterior distribution of the approximate location parameter $\phi(\theta) = \frac{2}{\pi} \arcsin \sqrt{\theta}$, it is found that

$$d\{\tilde{\theta} \,|\, n, r\} \approx \tfrac{1}{2} + 2\, n \left(\arcsin \sqrt{\tilde{\theta}} - \arcsin \sqrt{(r + \tfrac{1}{3})/(n + \tfrac{2}{3})} \right)^2 .$$

A lowest posterior loss q-credible region consists of the set of $\tilde{\theta}$ points with posterior probability q and minimum expected loss. In this problem, the intrinsic q-credible region $R^*_q(r, n)$, is therefore obtained as the interval $R^*_q(r, n) = [\theta_a(r, n),\, \theta_b(r, n)]$ defined by the solution (θ_a, θ_b) to the system

$$\Big\{ d\{\theta_a \,|\, n, r\} = d\{\theta_b \,|\, n, r\}, \quad \int_{\theta_a}^{\theta_b} \pi(\theta \,|\, n, r)\, d\theta = q \Big\}.$$

In particular, the intrinsic 0.95-credible region is the set of $\tilde{\theta}$ points with

posterior expected loss smaller than 2.139 (shaded region in the lower panel of Figure 4), which is $R^*_{0.95} = \{\tilde{\theta};\ 0.105 \leq \tilde{\theta} \leq 0.321]$. Notice that this is neither a HPD interval nor a centred interval. The point with minimum expected loss is the intrinsic estimator, $\theta^* = 0.2034$. Since intrinsic estimation is coherent under one-to-one reparametrizations, the intrinsic estimator and the 0.95-intrinsic credible region of the log-odds, $\psi = \psi(\theta) = \log[\theta/(1-\theta)]$ are immediately derived as $\psi(\theta^*) = -1.365$ and $\psi(R^*_{0.95}) = [-2.144,\ -0.747]$.

It may be argued that, in practice, it is reasonable for credible regions to give privilege to the most probable values of the parameters, as HPD regions do. This is obviously incompatible with an invariance requirement, but it is interesting to notice that, in one-parameter problems, intrinsic credible regions are approximately HPD in the approximate location parametrization. Thus, in Example 26, the 0.95-credible region for the approximate location parameter, $\phi(\theta) = \frac{2}{\pi} \arcsin \sqrt{\theta}$, $\phi(R^*_{0.95}) = [0.210, 0.384]$, is nearly an HPD interval for ϕ.

4.3 Hypothesis Testing

Let $\boldsymbol{x}$ be the available data, which are assumed to consist of one observation from model $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, parametrized in terms of the vector of interest $\boldsymbol{\theta}$ and a vector $\boldsymbol{\lambda}$ of nuisance parameters. The posterior distribution $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ of the quantity of interest $\boldsymbol{\theta}$ conveys immediate intuitive information on the values of $\boldsymbol{\theta}$ which, given $\mathcal{M}$, might be declared to be *compatible* with the observed data $\boldsymbol{x}$, namely, those with a relatively high probability density. Sometimes, a *restriction*, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$, of the possible values of the quantity of interest (where $\boldsymbol{\Theta}_0$ may possibly consist of a single value $\boldsymbol{\theta}_0$) is suggested in the course of the investigation as deserving special consideration, either because restricting $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}_0$ would greatly simplify the model, or because there are additional, context specific arguments suggesting that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Intuitively, the (null) *hypothesis* $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ should be judged to be *compatible* with the observed data $\boldsymbol{x}$ if there are elements in $\boldsymbol{\Theta}_0$ with a relatively high posterior density. However, a more precise conclusion is typically required and this is made possible by adopting a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ is a *decision problem* where the action space $\mathcal{A} = \{a_0, a_1\}$ only contains two elements: to accept (a_0) or to reject (a_1) the proposed restriction.

To solve this decision problem, it is necessary to specify an appropriate loss function, $\ell(a_i, \boldsymbol{\theta})$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ of the vector of interest. Notice that this requires the statement of an *alternative* a_1 to accepting H_0; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined. Given data $\boldsymbol{x}$, the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting the null, $\int_{\boldsymbol{\Theta}} \ell(a_0, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, is larger than the expected posterior loss of

rejecting, $\int_{\boldsymbol{\Theta}} \ell(a_1, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_{\boldsymbol{\Theta}} [\ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})]\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \Delta\ell(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} > 0. \tag{45}$$

Therefore, only the loss difference $\Delta\ell(\boldsymbol{\Theta}_0, \boldsymbol{\theta}) = \ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})$, which measures the *advantage* of rejecting $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ as a function of $\boldsymbol{\theta}$, has to be specified: the hypothesis H_0 should be rejected whenever the expected advantage of rejecting is positive.

A crucial element in the specification of the loss function is a description of what is precisely meant by rejecting H_0. By assumption a_0 means to act *as if* H_0 were true, *i.e.*, as if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, but there are at least two options for the alternative action a_1. This may either mean (i) the *negation* of H_0, that is to act as if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ or, alternatively, it may rather mean (ii) to reject the simplification implied by H_0 and to keep the unrestricted model, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, which is true by assumption. Both options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis, where hypothesis testing procedures are typically used, are better described by the second alternative. Indeed, an established model, identified by $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$, is often embedded into a more general model, $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}\}$, constructed to include promising departures from H_0, and it is then required to verify whether presently available data $\boldsymbol{x}$ are still compatible with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, or whether the extension to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is really required.

The simplest loss structure has, for all values of the nuisance parameter vector $\boldsymbol{\lambda}$, a zero-one form, with $\{\ell(a_0, \boldsymbol{\theta}) = 0, \ell(a_1, \boldsymbol{\theta}) = 1\}$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and $\{\ell(a_0, \boldsymbol{\theta}) = 1, \ell(a_1, \boldsymbol{\theta}) = 0\}$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$, so that the *advantage* $\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ of rejecting H_0 is

$$\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \begin{cases} 1, \text{ if } \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0 \\ -1, \text{ if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_0. \end{cases} \tag{46}$$

With this (rather naïve) loss function it is immediately found that the optimal action is to reject H_0 if (and only if) $\Pr(\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0 \,|\, \boldsymbol{x}) > \Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$. Notice that this formulation *requires* that $\Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0) > 0$, that is, that the (null) hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and $\boldsymbol{\Theta}_0$ has zero measure (for instance if H_0 consists of a single point $\boldsymbol{\theta}_0$), this requires the use of a non-regular "sharp" prior concentrating a positive probability mass on $\boldsymbol{\theta}_0$. With no mention to the loss structure implicit behind, this solution was early advocated by Jeffreys (1961, Ch. 5). However, this is known to lead to the difficulties associated to Lindley's paradox (Lindley, 1957; Bartlett, 1957; Bernardo, 1980; Robert, 1993; Brewer, 2002).

The intrinsic discrepancy loss may also be used to provide an attractive general alternative to Bayesian hypothesis testing, the *Bayesian reference cri-*

terion, BRC (Bernardo, 1999a; Bernardo and Rueda, 2002). This follows from assuming that the loss structure is such that

$$\Delta\ell\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} - d^*, \quad d^* > 0, \tag{47}$$

where $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which describes as a function of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ the loss suffered by assuming that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, is the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$. The function $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, which is invariant under one-to-one reparametrization, is non-negative and it is zero if, and only if, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. The constant d^* is the (strictly positive) advantage of being able to work with the null model when it is true, measured in the same units as δ; the choice of d^*, in terms of posterior expected log-likelihood ratios, is discussed below.

Definition 10 (Intrinsic hypothesis testing: BRC) *Let available data $\boldsymbol{x}$ consist of one observation from $\mathcal{M} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}), \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$, let $\mathcal{M}_0$ be the restricted model $\mathcal{M}_0 \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\lambda}), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ and let $\delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ be the intrinsic discrepancy between the distribution $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and the set $\mathcal{M}_0$. The* Bayesian reference criterion (BRC) *rejects model $\mathcal{M}_0$ if the* intrinsic statistic *$d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$, defined as the reference posterior expected intrinsic loss, exceeds a critical value d^*. In traditional language, the null hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ is rejected if*

$$d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x}) = \mathrm{E}[\delta \,|\, \boldsymbol{x}] = \int_{\boldsymbol{\Theta}} \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda} > d^*,$$

where $\pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi^{\delta}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is the reference posterior of $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ when $\delta = \delta\{\boldsymbol{\Theta}_0, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ is the quantity of interest. The conventional value $d^* = \log(100)$ may be used for scientific communication.

As the sample size increases, the expected value of $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ may be regarded as a continuous, positive measure of the expected loss (in information units) from simplifying the model by accepting $\mathcal{M}_0$. In traditional language, $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ is a test statistic, and the BRC criterion rejects the null if this *intrinsic test statistic* $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ exceeds some *critical value* d^*. However, in sharp contrast to frequentist hypothesis testing, the critical value d^* is simply a utility constant which measures the number of *information units* which the decision maker is prepared to loose in order to be able to work with the null model H_0, *not* a function of sampling properties of the model.

The interpretation of the intrinsic discrepancy in terms of the minimum posterior expected likelihood ratio in favour of the true model (see Section 2) provides a direct *calibration* of the required critical value. Indeed, $d(\boldsymbol{\Theta}_0 \,|\, \boldsymbol{x})$ is the minimum posterior expected log-likelihood ratio in favour of the true model. For instance, values around $\log[10] \approx 2.3$ should be regarded as mild evidence against H_0, while values around $\log[100] \approx 4.6$ suggest strong evidence against the null, and values larger than $\log[1000] \approx 6.9$ may be safely

used to reject H_0. Notice that, in contrast to frequentist hypothesis testing, where it is hazily recommended to adjust the significance level for dimensionality and sample size, the intrinsic statistic in measured on an absolute scale which remains valid for *any* sample size and *any* dimensionality.

Example 27 *Testing the value of a normal mean.* Let data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a normal $\mathrm{N}(x \,|\, \mu, \sigma)$ distribution, and consider the "canonical" problem of testing whether or not these data are compatible with some specific sharp hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean. The intrinsic discrepancy is easily found to be

$$\delta(\mu_0, \mu \,|\, \sigma) = \frac{n}{2}\left(\frac{\mu - \mu_0}{\sigma}\right)^2, \tag{48}$$

a simple transformation of the standardized distance between μ and μ_0, which generalizes to $\delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}) = (n/2)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^t \Sigma^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)$, a linear function of the Mahalanobis distance, in the multivariate normal case.

Consider first the case where σ is assumed to be known. The reference prior for μ is then uniform; this is also the reference prior when the parameter of interest is δ, since $\delta(\mu_0, \mu)$ is a piecewise invertible function of μ (see Theorem 6). The corresponding posterior distribution, is $\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})$, $(n \geq 1)$. The expected value of $\delta(\mu_0, \mu)$ with respect to this posterior yields the corresponding intrinsic statistic,

$$d(\mu_0 \,|\, \boldsymbol{x}) = \tfrac{1}{2}(1 + z^2), \quad z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \tag{49}$$

a simple function of the standardized distance between the sample mean $\overline{x}$ and μ_0. As prescribed by the general theory, the expected value of $d(\mu_0, \,|\, \boldsymbol{x})$ under repeated sampling is one if $\mu = \mu_0$, and increases linearly with n otherwise. In this canonical example, to reject H_0 whenever $|z| > 1.96$ (the frequentist suggestion with the conventional 0.05 significance level), corresponds to rejecting H_0 whenever $d(\mu_0 \,|\, \boldsymbol{x})$ is larger than 2.42, a rather weak evidence, since this means that the posterior expected likelihood ratio against H_0 is only about $\exp[2.42] = 11.25$. Conversely, to reject whenever posterior expected likelihood ratio against H_0 is about 100, so that $d^* = \log[100] \approx 4.6$, is to reject whenever $|z| > 2.86$, which is close to the conventional 3σ rule often used by engineers. The extreme 6σ rule, apparently popular these days, would correspond to $d^* = 18.5 \approx \log[10^8]$.

If the scale parameter σ is also unknown, the intrinsic discrepancy is

$$\delta\{\mu_0, (\mu, \sigma)\} = \frac{n}{2} \log\left[1 + \left(\frac{\mu - \mu_0}{\sigma}\right)^2\right], \tag{50}$$

which is *not* the same as (48). The intrinsic test statistic $d(\mu_0, \boldsymbol{x})$ may then be found as the expected value of $\delta\{\mu_0, (\mu, \sigma)\}$ under the corresponding joint reference posterior distribution $\pi^\delta(\mu, \sigma \,|\, \boldsymbol{x})$ when δ is the quantity of

interest. After some algebra, the exact result may be expressed in terms of hypergeometric functions (Bernardo, 1999a), but is very well approximated by the simple function

$$d(\mu_0 \,|\, \boldsymbol{x}) \approx \frac{1}{2} + \frac{n}{2} \log\left(1 + \frac{t^2}{n}\right), \tag{51}$$

where t is the conventional statistic $t = \sqrt{n-1}\,(\overline{x} - \mu_0)/s$, written in terms of the sample variance $s^2 = \sum_j (x_j - \overline{x})^2/n$. For instance, for samples sizes 5, 30 and 1000, and using the threshold $d^* = \log[100]$, the null hypothesis $H_0 \equiv \{\mu = \mu_0\}$ would be rejected whenever $|t|$ is respectively larger than 4.564, 3.073, and 2.871.

Example 28 *A lady tasting tea.* A lady claims that by tasting a cup of tea made with milk she can discriminate whether milk has been poured over the tea infusion or the other way round, and she is able to give the correct answer in n consecutive trials. Are these results compatible with the hypothesis that she is only guessing and has been lucky? The example, a variation suggested by Neyman (1950, Sec. 5.2) to a problem originally proposed by Fisher (1935, Sec. 2.5), has often been used to compare alternative approaches to hypothesis testing. See Lindley (1984) for a subjectivist Bayesian analysis.

The intrinsic objective Bayesian solution is immediate from the results in Examples 24 and 26. Indeed, using Definition 10, if data are assumed to consist of n Bernoulli observations and r successes have been observed, the intrinsic statistic to test the precise null $\theta = \theta_0$ is

$$d(\theta_0 \,|\, r, n) = \int_0^1 \delta\{\theta_0, \theta \,|\, n\}\, \mathrm{Be}(\theta \,|\, r + \tfrac{1}{2}, n - r + \tfrac{1}{2})\, d\theta,$$

where $\delta\{\theta_0, \theta \,|\, n\}$ is given by (7). In this case, one has $r = n$ and $\theta_0 = \frac{1}{2}$. For the values $n = 8$, $n = 10$ and $n = 12$ traditionally discussed, the intrinsic test statistic, $d(\theta_0 \,|\, r, n)$, respectively yields the values $d(\frac{1}{2} \,|\, 8, 8) \approx 4.15$, $d(\frac{1}{2} \,|\, 10, 10) \approx 5.41$ and $d(\frac{1}{2} \,|\, 12, 12) \approx 6.70$. Since $\log[100] \approx 4.61$, the hypothesis of pure guessing would not be rejected with $n = 8$ with the conventional threshold $d^* = \log[100]$, but would be rejected with $n = 10$ successes (and *a fortiori* with $n = 12$). Actually, the value of $d(\frac{1}{2} \,|\, 8, 8)$ says that the observed data are only estimated to be about $\exp[4.15] \approx 64$ times more likely under the true model (unknown θ) that under the null model (no discrimination power, $\theta = \theta_0 = \frac{1}{2}$). However, with $n = 10$ and $n = 12$ the observed data are respectively estimated to be about 224 and 811 times more likely under the true model that under the null.

The Bayesian reference criterion may also be used with non-nested problems. Thus, given two alternative models for $\boldsymbol{x} \in \mathcal{X}$, $\mathcal{M}_1 = \{p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1),\, \boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1\}$ and $\mathcal{M}_2 = \{p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2),\, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2\}$, one may introduce the a new parameter α to define a mixture model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_1, \alpha) = p_1(\boldsymbol{x} \,|\, \boldsymbol{\theta}_1)^{\alpha}\, p_2(\boldsymbol{x} \,|\, \boldsymbol{\theta}_2)^{1-\alpha}$ (with either

a continuous $\alpha \in [0, 1]$ or, more simply, a discrete $\alpha \in \{0, 1\}$), and use BRC to verify whether $\mathcal{M}_1$, or $\mathcal{M}_2$, or both, are compatible with the data, assuming the mixture is. For further discussion on hypothesis testing and the development of the Bayesian reference criterion see Bernardo (1982, 1985a, 1999a), Bernardo and Bayarri (1985), Rueda (1992) and Bernardo and Rueda (2002).

5 Further Reading

Reference analysis already has a long history, but it still is a very active area of research. The original paper on reference analysis, (Bernardo, 1979b), is easily read and it is followed by a very lively discussion; Bernardo (1981), extends the theory to general decision problems; see also Bernardo and Smith (1994, Sec. 5.4.1) and Rabena (1998). Berger and Bernardo (1989, 1992c) contain crucial mathematical extensions. Bernardo (1997) is a non-technical analysis, in a dialogue format, of the foundational issues involved, and it is followed by a discussion. A textbook level description of reference analysis is provided in Bernardo and Smith (1994, Sec. 5.4); Bernardo and Ramón (1998) contains a simple introduction to reference distributions. BRC, the Bayesian reference criterion for hypothesis testing, was introduced by Bernardo (1999a) and further refined in Bernardo and Rueda (2002). Intrinsic estimation was introduced in Bernardo and Juárez (2003). Berger, Bernardo and Sun (2005) contains the last mathematical developments of reference theory at the moment of writing.

Papers which contain either specific derivations or new applications of reference analysis include, in chronological order of the first related paper by the same author(s), Bernardo (1977a,b, 1978, 1980, 1982, 1985a,b, 1999b), Bayarri (1981, 1985), Ferrándiz (1982, 1985), Sendra (1982), Eaves (1983a,b, 1985), Armero (1985), Bernardo and Bayarri (1985), Chang and Villegas (1986), Chang and Eaves (1990), Hills (1987), Mendoza (1987, 1988, 1990), Bernardo and Girón (1988), Lindley (1988), Berger and Bernardo (1989, 1992a,b,c), Clarke and Barron (1990, 1994), Polson and Wasserman (1990), Phillips (1991), Severini (1991, 1993, 1995, 1999), Ye and Berger (1991), Ghosh and Mukerjee (1992), Singh and Upadhyay (1992), Stephens and Smith (1992), Berger and Sun (1993), Clarke and Wasserman (1993), Dey and Peng (1993, 1995), Kuboki (1993, 1998), Liseo (1993, 2003, 2005), Ye (1993, 1994, 1995, 1998), Berger and Yang (1994), Kubokawa and Robert (1994), Sun (1994, 1997), Sun and Berger (1994, 1998), Yang and Berger (1994, 1997), Datta and J. K. Ghosh (1995a,b), Datta and M. Ghosh (1995a); Datta and M. Ghosh (1995b), Giudici (1995), Ghosh, Carlin and Srivastava (1995), du Plessis, van der Merwe and Groenewald (1995), Sun and Ye (1995, 1996, 1999), de Waal, Groenewald and Kemp (1995), Yang and Chen (1995), Bernard (1996), Clarke (1996), Ghosh and Yang (1996), Armero and Bayarri (1997), Fernández, Osiewalski and Steel (1997), Garvan and Ghosh (1997, 1999), Ghosal and Samanta (1997), Ghosal (1997, 1999), Sugiura and Ishibayashi (1997), Berger, Philippe and Robert

(1998), Bernardo and Ramón (1998), Chung and Dey (1998, 2002), Scholl (1998), Sun, Ghosh and Basu (1998), Philippe and Robert (1998), Berger, Liseo and Wolpert (1999), Burch and Harris (1999), Brewer (1999), Scricciolo (1999), Vernotte and Zalamansky (1999), Yuan and Clarke (1999), Berger, Pericchi and Varshavsky (1998), Lee (1998), Fernández and Steel (1998b, 1999a,b, 2000), Mendoza and Gutiérrez-Peña (1999), Mukerjee and Reid (1999, 2001), Aguilar and West (2000), Eno and Ye (2000, 2001), Elhor and Pensky (2000), Fernández and Steel (2000), Kim, Kang and Cho (2000), van der Linde (2000), Berger, de Oliveira and Sansó (2001), Fan (2001), Ghosh and Kim (2001), Ghosh, Rousseau and Kim (2001), Kim, Chang and Kang (1961), Kim, Kang and Lee (2001, 2002), Komaki (2001, 2004), Natarajan (2001), Rosa and Gianola (2001), Aslam (2002a,b), Daniels (2002), Datta, Ghosh and Kim (2002), Millar (2002), Philippe and Rousseau (2002), Pretorius and van der Merwe (2002), Tardella (2002), Consonni and Veronese (2003), Datta and Smith (1995a), Fraser, Reid, Wong and Yi (2003), Ghosh and Heo (2003a,b), Ghosh, Yin and Kim (2003), Gutiérrez-Peña and Rueda (2003), He (2003), Leucari and Consonni (2003), Lauretto, Pereira, Stern and Zacks (2003), Madruga, Pereira and Stern (2003), Ni and Sun (2003), Sareen (2003), Consonni, Veronese, and Gutiérrez-Peña (2004), Sun and Ni (2004), Grünwald and Dawid (2004), Roverato and Consonni (2004), Stern (2004a,b), van der Merwe and Chikobvu (2004) and Liseo and Loperfido (2005).

This chapter concentrates on reference analysis. It is known, however, that ostensibly different approaches to the derivation of objective priors often produce the same result, a testimony of the robustness of many solutions to the definition of what an appropriate objective prior may be in a particular problem. Many authors have proposed alternative objective priors (often comparing the resulting inferences with those obtained within the frequentist paradigm), either as general methods or as *ad hoc* solutions to specific inferential problems, and a few are openly critical with objective Bayesian methods. Relevant papers in this very active field of Bayesian mathematical statistics include (in chronological order of their first related paper) Laplace (1825), Jeffreys (1946, 1955, 1961), Perks (1947), Haldane (1948), Barnard (1952, 1988), Good (1952, 1969, 1981, 1986), Lindley (1958, 1961, 1965), Stein (1959, 1962, 1986), Welch and Peers (1963), Geisser and Cornfield (1963), Stone (1963, 1965, 1970, 1976), Box and Cox (1964), Hartigan (1964, 1965, 1966, 1971, 1996, 1998, 2004), Geisser (1965, 1979, 1980, 1984, 1993), Hill (1965), Novick and Hall (1965), Peers (1965, 1968), Stone and Springer (1965), Welch (1965), Freedman (1966, 1995), Jaynes (1968, 1976, 1982, 1985, 1989), Cornfield (1969), Novick (1969), Villegas (1969, 1971, 1977a,b, 1981, 1982), Zidek (1969), DeGroot (1970, Ch. 10), Kappenman, Geisser and Antle (1970), Kashyap (1971), Zellner (1971); Zellner (1977, 1983, 1986a, 1988, 1991, 1996, 1997), Box and Tiao (1973, Sec. 1.3), Piccinato (1973, 1977), Aitchison and Dunsmore (1975), Rai (1976), Akaike (1978, 1980a,b,c, 1983), Florens (1978, 1982), Heath and Sudderth (1978, 1989), Banerjee and Bhattacharyya (1979),

Berger (1979), Evans and Nigm (1980), Miller (1980), Zellner and Siow (1980), Pericchi (1981), Torgersen (1981), Fatti (1982), Gokhale and Press (1982), Eaton (1982, 1992), Dawid (1983, 1991) Hartigan (1983, Ch. 5) Rissanen (1983, 1986, 1987, 1988, 1989), Sono (1983), Gatsonis (1984), Inaba (1984), Csiszár (1985, 1991), DasGupta (1985), Fraser, Monette and Ng (1985), Poirier (1985, 1994), Spiegelhalter (1985), Sweeting (1985, 1994, 1995a,b, 1996, 2001), Chang and Villegas (1986), Efron (1986, 1993), Raftery and Akman (1986), Casella and Hwang (1987), Cifarelli and Regazzini (1987), Chaloner (1987), Cox and Reid (1987), Maryak and Spall (1987), Smith and Naylor (1987), Stewart (1987), Wallace and Freeman (1987), Agliari and Calvi-Parisetti (1988), Howlader and Weiss (1988), Raftery (1988), de Waal and Nel (1988), Hill and Spall (1988, 1994), Consonni and Veronese (1989a,b, 1992, 1993), Crowder and Sweeting (1989), Kass (1989, 1990), Erickson (1989), Pole and West (1989), Tibshirani (1989), Tiwari, Chib and Jammalamadaka (1989), Upadhyay and Pandey (1989), Berger and Robert (1900), Chang and Eaves (1990), Lee and Shin (1990), Spall and Hill (1990), Ogata (1990), DiCiccio and Martin (1991), Ibrahim and Laud (1991), Joshi and Shah (1991), Meeden and Vardeman (1991), Rodríguez (1991), Ghosh and Mukerjee (1991, 1992, 1993a,b, 1995a,b), Pericchi and Walley (1991), Crowder (1992), Eaves and Chang (1992), Fan and Berger (1992), Polson (1992), Sansó and Pericchi (1992, 1994), Tsutakawa (1992), Vaurio (1992), George and McCulloch (1993), Mukerjee and Dey (1993), Nicolaou (1993), DiCiccio and Stern (1994), Paris (1994), Upadhyay and Smith (1989), Belzile and Angers (1995), Datta and M. Ghosh (1995a), Mukhopadhyay and Ghosh (1995), Pericchi and Sansó (1995), Wasserman and Clarke (1995), de Alba and Mendoza (1996), Atwood (1996), Berger and Strawderman (1996), Datta (1996), Fraser and Reid (1996, 2002), Keyes and Levy (1996), Mengersen and Robert (1996), Reid (1996), Upadhyay, Agrawal and Smith (1989), Wasserman (1996, 2000), Clarke and Sun (1997), Ghosh and Meeden (1997), Ibrahim (1997), Fraser, McDunnough and Taback (1997), Moreno and Girón (1997), Mukhopadhyay and DasGupta (1997), Mukerjee and Ghosh (1997), Barron, Rissanen and Yu (1998), Chao and Phillips (1998, 2002), Diaconis and Freedman (1998), Eaton and Sudderth (1998, 1999, 2002, 2004), Fernández and Steel (1998a), Hadjicostas (1998), Ibrahim and Chen (1998), Natarajan and McCulloch (1998), Blyth and Smith (1998), Barron (1999), Daniels (1999, 2005), Fraser, Reid and Wu (1999), Marinucci and Petrella (1999), Pauler, Wakefield and Kass (2003), Wallace and Dowe (1999), Walker and Muliere (1999), Walker and Gutiérrez-Peña (1999), Chen , Ibrahim and Shao (2000), Datta, Mukerjee, Ghosh and Sweeting (2000), Datta, Ghosh and Mukerjee (2000), Ghosh, Chen, Ghosh and Agresti (2000a); Ghosh, Ghosh, Chen and Agresti (2000b), Kim and Ibrahim (2000), Kim Lee and Kang (2000), Lee and Chang (2000), Lee and Hwang (2000), Lunn, Thomas, Best and Spiegelhalter (2000), McCulloch, Polson and Rossi (2000), Mendoza and Gutiérrez-Peña (2000), Natarajan and Kass (2000), Oh and Kim (2000), Price and Bonett (2000), Rousseau (2000), Strawderman (2000), Wolfinger and Kass (2000), Brown, Cai and DasGupta (2001, 2002),

Delampady *et al.* (2001), Robert and Rosenthal (2001), Sun, Tsutakawa and He (2001), Upadhyay, Vasishta and Smith (2001), Cho and Baek (2002), Everson and Bradlow (2002), Fraser and Yi (2002), Ghosh and Samanta (2002), Hartigan and Murphy (2002), Meng and Zaslavsky (2002), Molitor and Sun (2002), Shieh and Lee (2002), Singh, Gupta and Upadhyay (2002), Severini, Mukerjee and Ghosh (2002), Lee (2003), Mukerjee and Chen (2003), Strachan and van Dijk (2003), Upadhyay and Peshwani (2003), Datta and Mukerjee (2004), Gutiérrez-Peña and Muliere (2004) and Hobert, Marchev and Schweinsberg (2004).

For reviews of many of these, see Dawid (1983), Bernardo and Smith (1994, Sec. 5.6.2), Kass and Wasserman (1996) and Bernardo (1997).

6 Acknowledgements

The author is indebted to many colleagues for helpful comments on an earlier draft of this paper. Special recognition is due, however, to the very detailed comments by my *maestro*, Professor Dennis Lindley.

Research supported by grant BMF 2002-2889 of the former *Ministerio de Ciencia y Tecnología*, Madrid, Spain.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

de Alba, E. and Mendoza, M. (1996). A discrete model for Bayesian forecasting with stable seasonal patterns. *Advances in Econometrics II* (R. Carter Hill, ed.) New York: JAI Press, 267–281.

Agliari, A., and Calvi-Parisetti, C. (1988). A g-reference informative prior: A note on Zellner's g-prior. *The Statistician* **37**, 271–275.

Aguilar, O. and West, M. (2000). Bayesian Dynamic factor models and portfolio allocation. *J. Business Econ./ Studies* **18**, 338–357.

Akaike, H. (1977). On Entropy Maximization Principle. *Applications of Statistics* (P. R. Krishnaiah, ed.) Amsterdam: North-Holland, 27-41.

Akaike, H. (1978). A new look at the Bayes procedure. *Biometrika* **65**, 53–59.

Akaike, H. (1980). The interpretation of improper prior distributions as limits of data dependent proper prior distributions. *J. Roy. Statist. Soc. B* **45**, 46–52.

Akaike, H. (1980b). Likelihood and the Bayes procedure. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 144–166 and 185–203 (with discussion).

Akaike, H. (1980c). Ignorance prior distribution of a hyperparameter and Stein's estimator. *Ann. Inst. Statist. Math.* **32**, 171–178.

Akaike, H. (1983). On minimum information prior distributions. *Ann. Inst. Statist. Math.* **35**, 139–149.

Armero, C. (1985). Bayesian analysis of M/M/1/∞/Fifo Queues. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 613–618.

Armero, C. and Bayarri, M. J. (1997). A Bayesian analysis of a queuing system with unlimited service. *J. Statist. Planning and Inference* **58**, 241–261.

Aslam, M. (2002a). Bayesian analysis for paired comparison models allowing ties and not allowing ties. *Pakistan J. Statist.* **18**, 53–69.

Aslam, M. (2002b). Reference Prior for the Parameters of the Rao-Kupper Model. *Proc. Pakistan Acad. Sci.* **39**, 215–224

Atwood, C. L. (1996). Constrained noninformative priors in risk assessment. *Reliability Eng. System Safety* **53**, 37–46.

Banerjee, A. K. and Bhattacharyya G. K. (1979). Bayesian results for the inverse Gaussian distribution with an application. *Technometrics* **21**, 247–251.

Barnard, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika* **39**, 155–150.

Barnard, G. A. (1988). The future of statistics: Teaching and research. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 17–24.

Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 27–52, (with discussion).

Barron, A., Rissanen, J. and Yu, B. (1998). The Minimum Description Length principle in coding and modelling. *IEEE Trans. Information Theory* **44**, 2743–2760.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534.

Bayarri, M. J. (1981). Inferencia Bayesiana sobre el coeficiente de correlación de una población normal bivariante. *Trab. Estadist.* **32**, 18–31.

Bayarri, M. J. (1985). Bayesian inference on the parameters of a Beta distribution. *Statistics and Decisions* **2**, 17–22.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. Published posthumously in *Phil. Trans. Roy. Soc. London* **53**, 370–418 and **54**, 296–325. Reprinted in *Biometrika* **45** (1958), 293–315, with a biographical note by G. A. Barnard.

Belzile, E. and Angers, J.-F. (1995). Choice of noninformative priors for the variance components of an unbalanced one-way random model. *Comm. Statist. Theory and Methods* **24** 1325–1341.

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *J. Amer. Statist. Assoc.* **95**, 1269–1276.

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84**, 200–207.

Berger, J. O. and Bernardo, J. M. (1992a). Ordered group reference priors with applications to a multinomial problem. *Biometrika* **79**, 25–37.

Berger, J. O. and Bernardo, J. M. (1992b). Reference priors in a variance components problem. *Bayesian Analysis in Statistics and Econometrics* (P. K. Goel and N. S. Iyengar, eds.). Berlin: Springer, 323–340.

Berger, J. O. and Bernardo, J. M. (1992c). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 35–60 (with discussion).

Berger, J. O., Bernardo, J. M. and Mendoza, M. (1989). On priors that maximize expected information. *Recent Developments in Statistics and their Applications* (J. P. Klein and J. C. Lee, eds.). Seoul: Freedom Academy, 1–20.

Berger, J. O., Bernardo, J. M. and Sun, D. (2005). Reference priors from first principles: A general definition. *Tech. Rep.*, SAMSI, NC, USA.

Berger, J. O., de Oliveira, V. and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96**, 1361–1374.

Berger, J. O., Liseo, B. and Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14**, 1–28.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**, 109–122.

Berger, J. O., Pericchi, L. R. and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā A* **60**, 307–321.

Berger, J. O., Philippe, A. and Robert, C. (1998). Estimation of quadratic functions: noninformative priors for noncentrality parameters. *Statistica Sinica* **8**, 359–376.

Berger, J. O. and Robert, C. P. (1990). Subjective hierarchical Bayes estimation of a multivariate mean: On the frequentist interface. *Ann. Statist.* **18**, 617–651.

Berger, J. O. and Strawderman, W. E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Ann. Statist.* **24**, 931–951.

Berger, J. O. and Sun, D. (1993). Bayesian analysis for the poly-Weibull distribution. *J. Amer. Statist. Assoc.* **88**, 1412–1418.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle* (2nd ed.). Hayward, CA: IMS.

Berger, J. O. and Yang, R. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10**, 461–482.

Berger, R. L. (1979). Gamma minimax robustness of Bayes rules. *Comm. Statist. Theory and Methods* **8**, 543–560.

Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernoulli process. *Amer. Statist.* **50**, 7–13.

Bernardo, J. M. (1977a). Inferences about the ratio of normal means: a Bayesian approach to the Fieller-Creasy problem. *Recent Developments in Statistics* (J. R. Barra, F. Brodeau, G. Romier and B. van Cutsem eds.). Amsterdam: North-Holland, 345–349.

Bernardo, J. M. (1977b). Inferencia Bayesiana sobre el coeficiente de variación: una solución a la paradoja de marginalización. *Trab. Estadist.* **28**, 23-80.

Bernardo, J. M. (1978). Unacceptable implications of the left Haar measure in a standard normal theory inference problem *Trab. Estadist.* **29**, 3–9.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.). Brookfield, VT: Edward Elgar, 1995, 229–263.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 605–647 (with discussion).

Bernardo, J. M. (1981). Reference decisions. *Symposia Mathematica* **25**, 85–94.

Bernardo, J. M. (1982). Contraste de modelos probabilísticos desde una perspectiva Bayesiana. *Trab. Estadist.* **33**, 16–30.

Bernardo, J. M. (1985a). Anlisis Bayesiano de los contrastes de hipótesis paramtricos. *Trab. Estadist.* **36**, 45–54.

Bernardo, J. M. (1985b). On a famous problem of induction. *Trab. Estadist.* **36**, 24–30.

Bernardo, J. M. (1997). Non-informative priors do not exist. *J. Statist. Planning and Inference* **65**, 159–189 (with discussion).

Bernardo, J. M. (1999a). Nested hypothesis testing: The Bayesian reference criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 101–130 (with discussion).

Bernardo, J. M. (1999b). Model-free objective Bayesian prediction. *Rev. Acad. Ciencias de Madrid* **93**, 295–302.

Bernardo, J. M. and Bayarri, M. J. (1985). Bayesian model criticism. *Model Choice* (J.-P. Florens, M. Mouchart, J.-P. Raoult and L. Simar, eds.). Brussels: Pub. Fac. Univ. Saint Louis, 43–59.

Bernardo, J. M. and Girón F. J. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 67–88 (with discussion).

Bernardo, J. M. and Juárez, M. (2003). Intrinsic estimation. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford: University Press, 465-476.

Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *The Statistician* **47**, 1–35.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *Internat. Statist. Rev.* **70**, 351–372.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

Blyth, S. and Smith, A. F. M. (1998). Bayesian meta-analysis of frequentist p-values. *Comm. Statist. Theory and Methods* **27**, 2707–2724.

Box, G. E. P. and Cox D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. B* **26**, 211–252 (with discussion).

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.

Boulton, D. M. and Wallace, C. S. (1970). A program for numerical classification. *The Computer J.* 13, 63–69.

Brewer, K. R. W. (1999). Testing a precise null hypothesis using reference posterior odds. *Rev. Acad. Ciencias de Madrid* **93**, 303–310.

Brewer, K. R. W. (2002). The Lindley and Bartlett paradoxes. *Pak./ J./ Statist.* **18**, 1–13.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation of a binomial proportion. *Statist. Sci.* **16**, 101-133, (with discussion).

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30**, 160–201.

Burch, B. D. and Harris, I. R. (1999). Bayesian estimators of the intraclass correlation coefficient in the one-way random effects model. *Comm. Statist. Theory and Methods* **28**, 1247–1272.

Casella, G. (1996). Statistical inference and Monte Carlo algorithms. *Test* **5**, 249–344 (with discussion).

Casella, G. and Hwang, J. T. (1987). Employing vague prior information in the construction of confidence sets. *J. Multivariate Analysis* **21**, 79–104.

Chaloner, K. (1987). A Bayesian approach to the estimation of variance components for the unbalanced one way random model. *Technometrics* **29**, 323–337.

Chang, T. and Eaves, D. M. (1990). Reference priors for the orbit of a group model. *Ann. Statist.* **18**, 1595–1614.

Chang, T. and Villegas, C. (1986). On a theorem of Stein relating Bayesian and classical inferences in group models. *Can. J. Statist.* **14**, 289–296.

Chao, J. C. and Phillips, P. C. B. (1998). Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior. *J. Econometrics* **87**, 49–86.

Chao, J. C. and Phillips, P. C. B. (2002). Jeffreys prior analysis of the simultaneous equations model in the case with $n + 1$. *J. Econometrics* **111**, 251–283.

Chen, M. H., Ibrahim, J. G. and Shao, Q. M. (2000). Power prior distributions for generalized linear models. *J. Statist. Planning and Inference* **84**, 121-137.

Cho, J. S. and Baek, S. U. (2002). Development of matching priors for the ratio of the failure rate in the Burr model. *Far East J. Theor. Stat.* **8**, 79–87.

Chung, Y. and Dey, D. K. (1998). Bayesian approach to estimation in intraclass correlation using reference priors. *Comm. Statist. Theory and Methods* **27**, 2241–2255.

Chung, Y. and Dey, D. K. (2002). Model determination for the variance component model using reference priors. *J. Statist. Planning and Inference* **100**, 49–65.

Cifarelli, D. M. and Regazzini, E. (1987). Priors for exponential families which maximize the association between past and future observations. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 83–95.

Cover, M. and Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Crowder, M. (1992). Bayesian priors based on a parameter transformation using the distribution function.*Ann. Inst. Statist. Math.* **44**, 405–416.

Crowder, M. and Sweeting, T. (1989). Bayesian inference for a bivariate binomial distribution. *Biometrika* **76**, 599–603.

Csiszár, I. (1985). An extended maximum entropy principle and a Bayesian justification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 83–98, (with discussion).

Csiszár, I. (1991). Why least squared and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.* **19**, 2032–2066.

Clarke, B. (1996). Implications of reference priors for prior information and for sample size. *J. Amer. Statist. Assoc.* **91**, 173–184.

Clarke, B. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Information Theory* **36**, 453–471.

Clarke, B. and Barron, A. R. (1994). Jeffreys' prior is asymptotically least favourable under entropy risk. *J. Statist. Planning and Inference* **41**, 37–60.

Clarke, B. and Sun, D. (1997). Reference priors under the chi-squared distance. *Sankhyā A* **59**, 215–231.

Clarke, B., and Wasserman, L. (1993). Noninformative priors and nuisance parameters, *J. Amer. Statist. Assoc.* **88**, 1427–1432.

Consonni, G. and Veronese, P. (1989a). Some remarks on the use of improper priors for the analysis of exponential regression problems. *Biometrika* **76**, 101–106.

Consonni, G. and Veronese, P. (1989b). A note on coherent invariant distributions as non-informative priors for exponential and location-scale families. *Comm. Statist. Theory and Methods* **18**, 2883–2907.

Consonni, G. and Veronese, P. (1992). Bayes factors for linear models and improper priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 587–594.

Consonni, G. and Veronese, P. (1993). Unbiased Bayes estimates and improper priors. *Ann. Inst. Statist. Math.* **45**, 303–315.

Consonni, G. and Veronese, P. (2003). Enriched conjugate and reference priors for the Wishart family on symmetric cones. *Ann. Statist.* **31**, 1491–1516.

Consonni, G., Veronese, P. and Gutiérrez-Peña, E. (2004). Reference priors for exponential families with simple quadratic variance function. *J. Multivariate Analysis* **88**, 335–364.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. B* **49**, 1–39 (with discussion).

Cornfield, J. (1969). The Bayesian outlook and its application. *Biometrics* **25**, 617–657, (with discussion).

Daniels, M. J. (1999). A prior for the variance in hierarchical models. *Can. J. Statist.* **27**, 567–578.

Daniels, M. J. (2005). A class of shrinkage priors for the dependence structure in longitudinal data. *J. Statist. Planning and Inference* **127**, 119-130.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* **89**, 553–566.

DasGupta, A. (1985). Bayes minimax estimation in multiparameter families when the parameter space is restricted to a bounded convex set. *Sankhyā A* **47**, 326–332.

Datta, G. S. (1996). On priors providing frequentist validity for Bayesian inference of multiple parametric functions. *Biometrika* **83**, 287–298.

Datta, G. S. and Ghosh, J. K. (1995a). On priors providing a frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

Datta, G. S. and Ghosh, J. K. (1995b). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114.

Datta, G. S. and Ghosh, M. (1995a). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90**, 1357–1363.

Datta, G. S. and Ghosh, M. (1995b). Hierarchical Bayes estimators of the error variance in one-way ANOVA models. *J. Statist. Planning and Inference* **45**, 399–411.

Datta, G. S. and Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24**, 141–159.

Datta, G. S., Ghosh, M. and Kim, Y.-H. (2002). Probability matching priors for one-way unbalanced random effect models. *Statistics and Decisions* **20**, 29–51.

Datta, G. S., Ghosh, M. and Mukerjee, R. (2000). Some new results on probability matching priors. *Calcutta Statist./ Assoc./ Bull.* **50**, 179–192. Corr: **51**, 125.

Datta, G. S., and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Berlin: Springer,

Datta, G. S., Mukerjee, R., Ghosh, M. and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity . *Ann. Statist.* **28**, 1414–1426.

Datta, G. S. and Smith, D. D. (2003). On property of posterior distributions of variance components in small area estimation. *J. Statist. Planning and Inference* **112**, 175–183.

Datta, G. S. and Sweeting, T. (2005). Probability matching priors. *In this volume.*

Dawid, A. P. (1980). A Bayesian look at nuisance parameters. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 167-203, (with discussion).

Dawid, A. P. (1983). Invariant prior distributions. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 228–236.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *J. Roy. Statist. Soc. B* **53**, 79—109 (with discussion).

Dawid, A. P. and Stone, M. (1972). Expectation consistency of inverse probability distributions. *Biometrika* **59**, 486–489.

Dawid, A. P. and Stone, M. (1973). Expectation consistency and generalized Bayes inference. *Ann. Statist.* **1**, 478–485.

Dawid, A. P., Stone, M. and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Statist. Soc. B* **35**, 189–233 (with discussion).

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.

Dey, D. K. and Peng, F. (1993). On the choice of prior for the Bayes estimation in accelerated life testing.*J. Statist. Computation and Simulation* **48**, 207–217.

Dey, D. K. and Peng, F. (1995). Elimination of nuisance parameters using information measure. *Parisankhyan Samikkha* **2**, 9–29.

Delampady, M., DasGupta, A., Casella, G. Rubin, H. and Strawderman, W. E. (2001). A new approach to default priors and robust Bayesian methodology. *Can. J. Statist.* **29**, 437–450.

Diaconis P. and Freedman, D. A. (1998). Consistency of Bayes estimates for nonparametric regression: Normal theory. *Bernoulli* **4**, 411–444.

DiCiccio, T. J. and Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to bayesian and conditional inference. *Biometrika* **78**, 891–902.

DiCiccio, T. J. and Stern, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihood. *J. Roy. Statist. Soc. B* **56**, 397–408.

Eaton, M. L. (1982). A method for evaluating improper prior distributions. *Statistical Decision Theory and Related Topics III* **1** (S. S. Gupta and J. O. Berger, eds.). New York: Academic Press,

Eaton, M. L. (1992). A statistical diptych: admissible inferences, recurrence of symmetric Markov chains. *Ann. Statist.* **20**, 1147–1179.

Eaton, M. L. and Freedman, D. A. (2004). Dutch book against some 'objective' priors. *Bernoulli* **10**, 861–872.

Eaton, M. L., Sudderth, W. D. (1998). A new predictive distribution for normal multivariate linear models. *Sankhyā A* **60**, 363–382.

Eaton, M. L., Sudderth, W. D. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli* **5**, 833–854.

Eaton, M. L., Sudderth, W. D. (2002). Group invariant inference and right Haar measure. *J. Statist. Planning and Inference* **103**, 87–99.

Eaton, M. L., Sudderth, W. D. (2004). Properties of right Haar predictive inference. *Sankhyā A* **66**, 487–512.

Eaves, D. M. (1983a). On Bayesian nonlinear regression with an enzyme example. *Biometrika* **70**, 373-379.

Eaves, D. M. (1983b). Minimally informative prior analysis of a non-linear model. *The Statistician* **32**, 117.

Eaves, D. M. (1985). On maximizing the missing information about a hypothesis. *J. Roy. Statist. Soc. B* **47**, 263-266.

Eaves, D. and Chang, T.(1992). Priors for ordered conditional variance and vector partial correlation *J. Multivariate Analysis* **41**, 43–55.

Efron, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40**, 1–11, (with discussion).

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

Elhor, A. and Pensky, M. (2000). Bayesian estimation of location of lightning events. *Sankhyā B* **62**, 202–206.

Eno, D. R. and Ye, K. (2000). Bayesian reference prior analysis for polynomial calibration models. *Test* **9**, 191–202.

Eno, D. R. and Ye, K. (2001). Probability matching priors for an extended statistical calibration problem. *Can. J. Statist.* **29**, 19–35.

Erickson, T. (1989). Proper posteriors from improper priors for an unidentified error-in-variables model. *Econometrica* **57**, 1299-1316.

Evans, I. G. and Nigm A. M. (1980). Bayesian prediction for two parameter Weibull lifetime models. *Comm. Statist. Theory and Methods* **9**, 649–658.

Everson, P. J. and Bradlow, E. T. (2002). Bayesian inference for the Beta-binomial distribution via polynomial expansions. *J. Comp. Graphical Statist.* **11**, 202–207.

Fan, T.-H. (2001). Noninformative Bayesian estimation for the optimum in a single factor quadratic response model. *Test* **10**, 225–240.

Fan, T.-H. and Berger, J. O. (1992). Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions.*Statistics and Decisions* **10**, 99–120.

Fatti, L. P. (1982). Predictive discrimination under the random effect model. *South African Statist. J.* **16**, 55–77.

Fernández, C., Osiewalski, J. and Steel, M. (1997). On the use of panel data in stochastic frontier models with improper priors.*J. Econometrics* **79**, 169–193.

Fernández, C. and Steel, M. (1998a). Reference priors for the non-normal two-sample problems. *Test* **7**, 179–205.

Fernández, C. and Steel, M. (1998b) On Bayesian modelling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93**, 359–371.

Fernández, C. and Steel, M. (1999b). On the dangers of modelling through continuous distributions: A bayesian perspective. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 213–238, /diss

Fernández, C. and Steel, M. (1999b). Reference priors for the general location-scale model. *Statistics and Probability Letters* **43**, 377–384.

Fernández, C. and Steel, M. (2000). Bayesian regression analysis with scale mixtures of normals. *J. Economic Theory* **16**, 80–101.

Ferrándiz, J. R. (1982). Una solución Bayesiana a la paradoja de Stein. *Trab. Estadist.* **33**, 31–46.

Ferrándiz, J. R. (1985). Bayesian inference on Mahalanobis distance: An alternative approach to Bayesian model testing. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 645–654.

de Finetti, B. (1970). *Teoria delle Probabilit* **1**. Turin: Einaudi. English translation as *Theory of Probability* **1** in 1974, Chichester: Wiley.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Florens, J.-P. (1978). Mesures à priori et invariance dans une expérience Bayésienne. *Pub. Inst. Statist. Univ. Paris* **23**, 29–55.

Florens, J.-P. (1982). Expériences Bayésiennes invariantes. *Ann. Inst. M. Poincaré* **18**, 309–317.

Fraser, D. A. S., McDunnough, P. and Taback, N. (1997). Improper priors, posterior asymptotic normality, and conditional inference. *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz* (N. L. Johnson and N. Balakrishnan, eds.) New York: Wiley, 563–569.

Fraser, D. A. S., Monette, G., and Ng, K. W. (1985). Marginalization, likelihood and structural models, *Multivariate Analysis* **6** (P. R. Krishnaiah, ed.). Amsterdam: North-Holland, 209–217.

Fraser, D. A. S., Reid, N. (1996). Bayes posteriors for scalar interest parameters. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 581–585.

Fraser, D. A. S. and Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Planning and Inference* **103**, 263–285.

Fraser, D. A. S., Reid, N., Wong, A. and Yi, G. Y. (2003). Direct Bayes for interest parameters . *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford: University Press, 529–534.

Fraser, D. A. S., Reid, N. and Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.

Fraser, D. A. S. and Yi, G. Y. (2002). Location reparametrization and default priors for statistical analysis. *J. Iranian Statist. Soc.* **1**, 55–78.

Freedman, D. A. (1966). A note on mutual singularity of priors. *Ann. Math. Statist.* **37**, 375–381.

Freedman, D. A. (1995). Some issues in the foundation of statistics. *Topics in the Foundation of Statistics* (B. C. van Fraassen, ed.) Dordrecht: Kluwer19–83. (with discussion).

Garvan, C. W. and Ghosh, M. (1997). Noninformative priors for dispersion models. *Biometrika* **84**, 976–982.

Garvan, C. W. and Ghosh, M. (1999). On the property of posteriors for dispersion models. *J. Statist. Planning and Inference* **78**, 229–241.

Gatsonis, C. A. (1984). Deriving posterior distributions for a location parameter: A decision theoretic approach. *Ann. Statist.* **12**, 958–970.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36**, 150–159.

Geisser, S. (1979). In discussion of Bernardo (1979b). *J. Roy. Statist. Soc. B* **41**, 136–137.

Geisser, S. (1980). A predictivist primer. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys* (A. Zellner, ed.). Amsterdam: North-Holland, 363–381.

Geisser, S. (1984). On prior distributions for binary trials. *J. Amer. Statist. Assoc.* **38**, 244–251 (with discussion).

Geisser, S. (1993). *Predictive inference: An introduction.* London: Chapman and Hall

Geisser, S. and Cornfield, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. B* **25**, 368–376.

George, E. I. and McCulloch, R. (1993). On obtaining invariant prior distributions. *J. Statist. Planning and Inference* **37**, 169–179.

Ghosal, S. (1997). Reference priors in multiparameter nonregular cases. *Test* **6**, 159–186.

Ghosal, S. (1999). Probability matching priors for non-regular cases. *Biometrika* **86**, 956–964.

Ghosal, S. and Samanta, T. (1997). Expansion of Bayes risk for entropy loss and reference prior in nonregular cases. *Statistics and Decisions* **15**, 129–140.

Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997). Non-informative priors via sieves and packing numbers. *Advances in Decision Theory and Applications* (S. Panchpakesan and N. Balakrisnan, eds.) Boston: Birkhauser, 119–132.

Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multivariate case. *J. Multivariate Analysis* **38**, 385–393.

Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 195–210 (with discussion).

Ghosh, J. K. and Mukerjee, R. (1993a). Frequentist validity of highest posterior density regions in the multiparameter case. *Ann. Math. Statist.* **45**, 293–302; corr: 602.

Ghosh, J. K. and Mukerjee, R. (1993b). On priors that match posterior and frequentist distribution functions. *Can. J. Statist.* **21**, 89–96.

Ghosh, J. K. and Mukerjee, R. (1995a). Frequentist validity of highest posterior density regions in the presence of nuisance parameters. *Statistics and Decisions* **13**, 131–139.

Ghosh, J. K. and Mukerjee, R. (1995b). On perturbed ellipsoidal and highest posterior density regions with approximate frequentist validity. *J. Roy. Statist. Soc. B* **57**, 761–769.

Ghosh, J. K. and Samanta, T. (2002). Nonsubjective Bayes testing – An overview. *J. Statist. Planning and Inference* **103**, 205–223.

Ghosh, M., Carlin, B. P. and Srivastava, M. S. (1995). Probability matching priors for linear calibration. *Test* **4**, 333–357.

Ghosh, M., Chen, M.-H., Ghosh, A. and Agresti, A. (2000a). Hierarchical Bayesian analysis of binary matched pairs data. *Statistica Sinica* **10**, 647–657.

Ghosh, M., Ghosh, A., Chen, M.-H. and Agresti, A. (2000b). Noninformative priors for one-parameter item response models. *J. Statist. Planning and Inference* **88**, 99–115.

Ghosh M. and Heo J. (2003). Default Bayesian priors for regression models with first-order autoregressive residuals. *J. Time Series Analysis* **24**, 269–282.

Ghosh, M., and Heo J. (2003). Noninformative priors, credible sets and bayesian hypothesis testing for the intraclass model. *J. Statist. Planning and Inference* **112**, 133–146.

Ghosh, M., and Meeden G. (1997). *Bayesian Methods for Finite Population Sampling* London: Chapman and Hall.

Ghosh, M. and Mukerjee, R. (1998). Recent developments on probability matching priors. *Applied Statistical Science III* (S. E. Ahmed, M. Ashanullah and B. K. Sinha eds.) New York: Science Publishers, 227–252.

Ghosh, M., Rousseau, J. and Kim, D. H. (2001). Noninformative priors for the bivariate Fieller-Creasy problem. *Statistics and Decisions* **19**, 277–288.

Ghosh, M. and Yang, M.-Ch. (1996). Non-informative priors for the two sample normal problem. *Test* **5**, 145–157.

Ghosh, M. and Kim, Y.-H. (2001). The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis. *Can. J. Statist.* **29**, 5–17.

Ghosh, M., Yin, M. and Kim, Y.-H. (2003). Objective Bayesian inference for ratios of regression coefficients in linear models. *Statistica Sinica* **13**, 409–422.

Giudici, P. (1995). Bayes factors for zero partial covariances. *J. Statist. Planning and Inference* **46**, 161–174.

Gokhale, D. V. and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *J. Roy. Statist. Soc. A* **145**, 237–249.

Good, I. J. (1952). Rational decisions. *J. Roy. Statist. Soc. B* **14**, 107–114.

Good, I. J. (1969). What is the use of a distribution? *Multivariate Analysis* **2** (P. R. Krishnaiah, ed.). New York: Academic Press, 183–203.

Good, I. J. (1981). Flexible priors for estimating a normal distribution. *J. Statist. Computation and Simulation* **13**, 151–153.

Good, I. J. (1986). Invariant hyperpriors. *J. Statist. Computation and Simulation* **24**, 323-324.

Goudie, I. B. J. and Goldie, C. M. (1981). Initial size estimation for the pure death process. *Biometrika* **68**. 543–550.

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. Statist.* **32**, 1367-1433.

Gutiérrez-Peña, E. (1992). Expected logarithmic divergence for exponential families. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 669–674.

Gutiérrez-Peña, E. and Muliere, P. (2004). Conjugate priors represent strong pre-experimental assumptions. *Scandinavian J. Statist.* **31**, 235–246.

Gutiérrez-Peña, E. and Rueda, R. (2003). Reference priors for exponential families. *J. Statist. Planning and Inference* **110**, 35–54.

Hadjicostas, P. (1998). Improper and proper posteriors with improper priors in a hierarchical model with a beta-binomial likelihood. *Comm. Statist. Theory and Methods* **27**, 1905–1914.

Hadjicostas, P. and Berry, S. M. (1999). Improper and proper posteriors with improper priors in a Poisson-gamma hierarchical model. *Test* **8**, 147–166.

Haldane, J. B. S. (1948). The precision of observed values of small frequencies. *Biometrika* **35**, 297–303.

Hartigan, J. A. (1964). Invariant prior distributions. *Ann. Math. Statist.* **35**, 836–845.

Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Statist.* **36**, 1137–1152.

Hartigan, J. A. (1966). Note on the confidence prior of Welch and Peers. *J. Roy. Statist. Soc. B* **28**, 55-56.

Hartigan, J. A. (1971). Similarity and probability. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 305–313 (with discussion).

Hartigan, J. A. (1983). *Bayes Theory*. Berlin: Springer.

Hartigan, J. A. (1996). Locally uniform prior distributions. *Ann. Statist.* **24**, 160–173.

Hartigan, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26**, 2083–2103.

Hartigan, J. A. (2004). Uniform priors on convex sets improve risk. *Statistics and Probability Letters* **67**, 285–288.

Hartigan, J. A. and Murphy, T. B. (2002). Inferred probabilities. *J. Statist. Planning and Inference* **105**, 23–34.

He, C. Z. (2003). Bayesian modelling of age-specific survival in bird nesting studies under irregular visits. *Biometrics* **59**, 962–973.

Heath, D. and Sudderth, W. (1978). On finitely additive priors, coherence, and extended admissibility. *Ann. Statist.* **6**, 333–345.

Heath, D. and Sudderth, W. (1989). Coherent inference from improper priors and from finitely additive priors. *Ann. Statist.* **17**, 907–919.

Hill, B. M. (1965). Inference about variance components in the one-way model. *J. Amer. Statist. Assoc.* **60**, 806–825.

Hill, S. D. and Spall, J. C. (1988). Shannon information-theoretic priors for state-space model parameters. *Bayesian Analysis of Time series and Dynamic Models* (J. C. Spall, ed.). New York: Marcel Dekker, 509–524.

Hill, S. D. and Spall, J. C. (1994). Sensitivity of a Bayesian analysis to the prior distribution. *IEEE Trans. Systems, Man and Cybernetics* **24**, 216–221.

Hills, S. E. (1987). Reference priors and identifiability problems in non-linear models. *The Statistician* **36**, 235–240.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91**, 1461–1473.

Hobert, J. P. and Casella, G. (1996). Functional compatibility, markov chains and Gibbs sampling with improper posteriors. *J. Comp. Graphical Statist.* **7**, 42-60.

Hobert, J. P., Marchev, D. and Schweinsberg, J. (2004). Stability of the tail markov chain and the evaluation of improper priors for an exponential rate parameter. *Bernoulli* **10**, 549–564.

Howlader, H. A. and Weiss, G. (1988). Bayesian reliability estimation of a two parameter Cauchy distribution. *Biom. J.* **30**, 329–337.

Ibrahim, J. G. (1997). On properties of predictive priors in linear models. *Amer. Statist.* **51**, 333–337.

Ibrahim, J. G. and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models. *Sankhyā B* **60**, 48–64.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys' prior. *J. Amer. Statist. Assoc.* **86**, 981–986.

Ibragimov, I. A. and Khasminskii, R. Z. (1973). On the information in a sample about a parameter. *Proc. 2nd Internat. Symp. Information Theory.* (B. N. Petrov and F. Csaki, eds.), Budapest: Akademiaikiadó, 295–309.

Inaba, T. (1984). Non-informative prior distribution in a simultaneous equations model. *J. Japan Statist. Soc.* **14**, 93–103.

Jaynes, E. T. (1968). Prior probabilities. *IEEE Trans. Systems, Science and Cybernetics* **4**, 227–291.

Jaynes, E. T. (1976). Confidence intervals vs. Bayesian intervals. *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* **2** (W. L. Harper and C. A. Hooker eds.). Dordrecht: Reidel, 175–257 (with discussion).

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proc. of the IEEE* **70**, 939–952.

Jaynes, E. T. (1985). Some random observations. *Synthèse* **3**, 115–138.

Jaynes, E. T. (1989). *Papers on Probability, Statistics and Statistical Physics*, 2nd ed. Dordrecht: Kluwer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A* **186**, 453–461.

Jeffreys, H. (1955). The present position in probability theory. *Brit. J. Philos. Sci.* **5**, 275–289.

Jeffreys, H. (1961). *Theory of Probability* (Third edition). Oxford: University Press.

Joshi, S. and Shah, M. (1991). Estimating the mean of an inverse Gaussian distribution with known coefficient of variation. *Comm. Statist. Theory and Methods* **20**, 2907–2912.

Juárez, M. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universitat de València, Spain.

Kashyap, R. L. (1971). Prior probability and uncertainty. *IEEE Trans. Information Theory* **14**, 641–650.

Kappenman, R. F., Geisser, S. and Antle, C. F. (1970). Bayesian and fiducial solutions to the Fieller Creasy problem. *Sankhyā B* **25**, 331–330.

Kass, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4**, 188–235, (with discussion).

Kass, R. E. (1990). Data-translated likelihood and Jeffreys' rule. *Biometrika* **77**, 107–114.

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**, 1343–1370. Corr: 1998 **93**, 412.

Keyes, T. K. and Levy, M. S. (1996). Goodness of prediction fit for multivariate linear models. *J. Amer. Statist. Assoc.* **91**, 191–197.

Kim, B. H., Chang, I. H. and Kang, C. K. (2001). Bayesian estimation for the reliability in Weibull stress-strength systems using noninformative priors. *Far East J. Theor. Stat.* **5**, 299–315.

Kim, D. H., Kang, S. G. and Cho, J. S. (2000). Noninformative Priors for Stress-Strength System in the Burr-Type X Model *J. Korean Statist. Soc.* **29**, 17–28.

Kim, D. H., Kang, S. G. and Lee, W. D. (2001). Noninformative priors for intraclass correlation coefficient in familial data. *Far East J. Theor. Stat.* **5**, 51–65.

Kim, D. H., Kang, S. G. and Lee, W. D. (2002). Noninformative priors for the power law process. *J. Korean Statist. Soc.* **31**, 17–31.

Kim, D. H., Lee, W. D. and Kang, S. G. (2000). Bayesian model selection for life time data under type II censoring. *Comm. Statist. Theory and Methods* **29**, 2865–2878.

Kim, S. W. and Ibrahim, J. G. (2000). On Bayesian inference for proportional hazards models using noninformative priors. *Lifetime Data Analysis* **6**, 331–341.

Komaki, F. (2001). A shrinkage predictive distribution for multivariate Normal observables. *Biometrika* **88**, 859–864.

Komaki, F. (2004). Simultaneous prediction of independent Poisson observables *Ann. Statist.* **32**, 1744-1769

Kubokawa, T. and Robert, C. P. (1994). New perspectives in linear calibration. *J. Multivariate Analysis* **51**, 178–200.

Kuboki, H. (1993). Inferential distributions for non-Bayesian predictive fit. *Ann. Inst. Statist. Math.* **45**, 567–578.

Kuboki, H. (1998). Reference priors for prediction. *J. Statist. Planning and Inference* **69**, 295–317.

Kullback, S. (1968). *Information Theory and Statistics* (2nd ed.). New York: Dover. Reprinted in 1978, Gloucester, MA: Peter Smith.

Kullback, S. (1983). Kullback information. *Encyclopedia of Statistical Sciences* **4** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 421–425.

Kullback, S. (1983). The Kullback-Leibler distance. *Amer. Statist.* **41**, 340–341.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Lafferty, J. and Wasserman, L. (2001). Iterative Markov chain Monte Carlo computation of reference priors and minimax risk. *Uncertainty in Artificial Intelligence* (UAI, Seattle).

Lane, D. and Sudderth, W. D. (1984). Coherent predictive inference. *Sankhyā A* **46**, 166–185.

Laplace, P. S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. Reprinted as *Oeuvres Complètes de Laplace* **7**, 1878–1912. Paris: Gauthier-Villars.

Laplace, P. S. (1825). *Essai Philosophique sur les Probabilitiés*. Paris: Courcier (5th ed). English translation in 1952 as *Philosophical Essay on Probabilities*. New York: Dover.

Lauretto, M., Pereira, C. A. B. Stern, J. M. and Zacks, S. (2003). Comparing parameters of two bivariate normal bistributions using the invariant FBST, Full Bayesian Significance Test. *Brazilian J. Probab. Statist.* **17**, 147–168.

Lee, H. K. H. (2003). A noninformative prior for neural networks. *Machine Learning* **50**, 197–212.

Lee, G. (1998). Development of matching priors for $P(X < Y)$ in exponential distributions. *J. Korean Statist. Soc.* **27**, 421–433.

Lee, J. C. and Chang, C. H. (2000). Bayesian analysis of a growth curve model with a general autoregressive covariance structure. *Scandinavian J. Statist.* **27**, 703–713.

Lee, J. C. and Hwang, R. C. (2000). On estimation and prediction for temporally correlated longitudinal data. *J. Statist. Planning and Inference* **87**, 87–104.

Lee, J. J. and Shin, W. S. (1990). Prior distributions using the entropy principles. *Korean J. Appl. Statist.* **3**, 91–104.

Leucari, V. and Consonni, G. (2003). Compatible priors for causal Bayesian networks. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). Oxford: University Press, 597-606.

van der Linde, A. (2000). Reference priors for shrinkage and smoothing parameters. *J. Statist. Planning and Inference* **90**, 245–274.

Lindley, D. V. (1956). On a measure of information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Lindley, D. V. (1958). Fiducial distribution and Bayes' Theorem. *J. Roy. Statist. Soc. B* **20**, 102–107.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *Proc. Fourth Berkeley Symp.* **1** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 453–468.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge: University Press.

Lindley, D. V. (1984). A Bayesian lady tasting tea. *Statistics: An Appraisal*. (H. A.David and H. T.David, eds.) Ames, IA: Iowa State Press455–479.

Lindley, D. V. (1988). Statistical inference concerning the Hardy-Weinberg equilibrium. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 307–326 (with discussion).

Liseo, B. (1993). Elimination of nuisance parameters with reference priors. *Biometrika* **80**, 295–304.

Liseo, B. (2003). Bayesian and conditional frequentist analyses of the Fieller's problem. A critical review. *Metron* **61**, 133–152.

Liseo, B. and Loperfido, N. (2005). A note on reference priors for the scalar skew-normal distribution. *J. Statist. Planning and Inference* (to appear).

Liseo, B. (2005). The problem of the elimination of nuisance parameters in a Bayesian framework. *In this volume*.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statist. Computing* **10**, 325–337.

Madruga, M. R., Pereira, C. A. B. and Stern, J. M. (2003). Bayesian evidence test for precise hypothesis. *J. Statist. Planning and Inference* **117**, 185–198.

Maryak, J. L. and Spall, J. C. (1987). Conditions for the insensitivity of the Bayesian posterior distribution to the choice of prior distribution. *Statistics and Probability Letters* **5**, 401–407.

Marinucci, D. and Petrella, L. (1999). A Bayesian proposal for the analysis of stationary and nonstationary AR(1) time series. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 821–828.

McCulloch, R. E., Polson, N. G. and Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99**, 173–193.

Meeden, G. and Vardeman, S. (1991). A non-informative Bayesian approach to interval estimation in finite population sampling. *J. Amer. Statist. Assoc.* **86**, 972–986.

Mendoza, M. (1987). A Bayesian analysis of a generalized slope ratio bioassay. *Probability and Bayesian Statistics* (R. Viertl, ed.). London: Plenum, 357–364.

Mendoza, M. (1988). Inferences about the ratio of linear combinations of the coefficients in a multiple regression problem. *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Oxford: University Press, 705–711.

Mendoza, M. (1990). A Bayesian analysis of the slope ratio bioassay. *Biometrika* **46**, 1059–1069.

Mendoza, M. (1994). Asymptotic normality under transformations. A result with Bayesian applications. *Test* **3**, 173–180.

Mendoza, M. and Gutiérrez-Peña, E. (1999). Bayesian inference for the ratio of the means of two normal populations with unequal variances. *Biom. J.* **41**, 133–147.

Mendoza, M. and Gutiérrez-Peña, E. (2000). Bayesian conjugate analysis of the Galton-Walton process. *Test* **9**, 149–171.

Meng, X.-L. and Zaslavsky, A. M. (2002). Single observation unbiased priors. *Ann. Statist.* **30**, 1345–1375.

Mengersen, K. L. and Robert, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 255-276 (with discussion).

Van der Merwe, A. J. and Chikobvu, D. (2004). Bayesian analysis of the process capability Index C_{pk}. *South African Statist. J.* **38**, 97–117.

Millar, R. B. (2002). Reference priors for Bayesian fisheries models. *Can. J. Fish. Aquat. Sci.* **59**, 1492–1502.

Miller, R. B. (1980). Bayesian analysis of the two-parameter gamma distribution. *Technometrics* **22**, 65-69.

Molitor, J. and Sun, D. (2002). Bayesian analysis under ordered functions of parameters. *Environ./ Ecological Statist.* **9**, 179–193.

Moreno, E. and Girón, F. J. (1997). Estimating with incomplete count data: a Bayesian approach. *J. Statist. Planning and Inference* **66**, 147–159.

Mukerjee R. and Chen Z. (2003). On Expected Lengths of Predictive Intervals *Scandinavian J. Statist.* **30**, 757–766

Mukerjee, R. and Dey, D. K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika* **80**, 499–505.

Mukerjee, R. and Ghosh, M. (1997). Second-order probability matching priors. *Biometrika* **84**, 970–975.

Mukerjee, R. and Reid, N. (1999). On a property of probability matching priors: Matching the alternative coverage probabilities. *Biometrika* **86**, 333–340.

Mukerjee, R. and Reid, N. (2001). Second-order probability matching priors for a parametric function with application to Bayesian tolerance limits. *Biometrika* **88**, 587–592.

Mukhopadhyay, S. and DasGupta, A. (1997). Uniform approximation of Bayes solutions and posteriors: Frequentist valid Bayes inference. *Statistics and Decisions* **15**, 51–73.

Mukhopadhyay, S. and Ghosh, M. (1995). On the uniform approximation of Laplace's prior by t-priors in location problems. *J. Multivariate Analysis* **54**, 284–294.

Natarajan, R. (2001). On the propriety of a modified Jeffreys's prior for variance components in binary random effects models. *Statistics and Probability Letters* **51**, 409–414.

Natarajan, R. and Kass, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.* **95**, 227–237.

Natarajan, R. and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data driven inference?. *J. Comp. Graphical Statist.* **7**, 267–277.

Neyman, J. (1950). *First Course in probability and Statistics*. New York: Holt.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Ni, S. X, and Sun, D. (2003). Noninformative priors and frequentist risks of Bayesian estimators of vector-autoregressive models. *J. Econom.* **115**, 159–197.

Nicolau, A. (1993). Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. *J. Roy. Statist. Soc. B* **55**, 377–390.

Novick, M. R. (1969). Multiparameter Bayesian indifference procedures. *J. Roy. Statist. Soc. B* **31**, 29–64.

Novick, M. R. and Hall, W. K. (1965). A Bayesian indifference procedure. *J. Amer. Statist. Assoc.* **60**, 1104–1117.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure. *Ann. Inst. Statist. Math.* **42**, 403–433.

Oh, M.-S. and Kim, Y. (2000). Jeffreys noninformative prior in Bayesian conjoint analysis. *J. Korean Statist. Soc.* **29**, 137–153.

Palmer, J. L. and Pettit, L. I. (1996). Risks of using improper priors with Gibbs sampling and autocorrelated errors. *J. Comp. Graphical Statist.* **5** 245 - 249.

Paris, J. B. (1994) *The Uncertain Reasoner's Companion: A Mathematical Perspective*, Cambridge: University Press.

Pauler, D. K., Wakefield, J. C. and Kass, R. E. (1999). Bayes factors and approximations for variance component models. *J. Amer. Statist. Assoc.* **94**, 1242–1253.

Peers, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. B* **27**, 9–16.

Peers, H. W. (1968). Confidence properties of Bayesian interval estimates. *J. Roy. Statist. Soc. B* **30**, 535–544.

Pericchi, L. R. (1981). A Bayesian approach to transformation to normality. *Biometrika* **68**, 35–43.

Pericchi, L. R. (2005). Model selection and hypothesis testing based on probabilities. *In this volume.*

Pericchi L.R. and Sansó B. (1995) A note on bounded influence in Bayesian analysis. *Biometrika* **82**, 223-225.

Pericchi, L. R. and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *Internat. Statist. Rev.* **59**, 1–23.

Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *J. Inst. Actuaries* **73**, 285–334 (with discussion).

Philippe, A. and Robert, C. P. (1998). A note on the confidence properties of reference priors for the calibration model. *Test* **7**, 147–160.

Philippe A. and Rousseau, J.. (2002). Non informative priors in the case of Gaussian long-memory processes. *Bernoulli* **8**, 451–473.

Phillips, P. C. B. (1991). To criticize the critics: an objective Bayesian analysis of stochastic trends. *J. Applied Econometrics* **6**, 333-473, (with discussion).

Piccinato, L. (1973). Un metodo per determinare distribuzioni iniziali relativamente non-informative. *Metron* **31**, 124–156.

Piccinato, L. (1977). Predictive distributions and non-informative priors. *Trans. 7th. Prague Conf. Information Theory* (M. Uldrich, ed.). Prague: Czech. Acad. Sciences, 399–407.

du Plessis, J. L., van der Merwe, A. J. and Groenewald, P. C. N. (1995). Reference priors for the multivariate calibration problem. *South African Statist. J.* **29**, 155–168.

Poirier, D. J. (1985). Bayesian hypothesis testing in linear models with continuously induced conjugate priors across hypotheses. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 711–722.

Poirier, D. J. (1994). Jeffreys prior for logit models. *J. Econometrics* **63**, 327–339.

Pole, A. and West, M. (1989). Reference analysis of the dynamic linear model. *J. Time Series Analysis* **10**, 13–147.

Polson, N. G. (1992). In discussion of Ghosh and Mukerjee (1992). *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 203–205.

Polson, N. G. and Wasserman, L. (1990). Prior distributions for the bivariate binomial. *Biometrika* **76**, 604–608.

Press, S. J. (1972). *Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference*. Second edition in 1982, Melbourne, FL: Krieger.

Pretorius, A. L. and van der Merwe A. J. (2002). Reference and probability-matching priors in Bayesian analysis of mixed linear models. *J. Anim. Breed. Genet.* **119**, 311–324.

Price, R. M., Bonett, D. G. (2000). Estimating the ratio of two Poisson rates. *Comput. Statist. Data Anal.* **34**, 345–356.

Rabena, M. T. (1998). Deriving reference decisions. *Test* **7**, 161–177.

Rai, G. (1976). Estimates of mean life and reliability function when failure rate is randomly distributed. *Trab. Estadist.* **27**: 247–251.

Raftery, A. E. (1988) Inference for the binomial N parameter: A hierarchical Bayes approach *Biometrika* **75**. 223–228.

Raftery, A. E. (1988) Analysis of a simple debugging model. *Appl. Statist.* **37**, 12–22.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73**, 85–89.

Reid, N. (1996). Likelihood and Bayesian approximation methods. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 349–366 (with discussion).

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416–431.

Rissanen, J. (1986). Stochastic complexity and modelling *Ann. Statist.* **14**, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. B* **49**, 223-239 and 252-265 (with discussion).

Rissanen, J. (1988). Minimum description length principle. *Encyclopedia of Statistical Sciences* **5** (S. Kotz, N. L. Johnson and C. B. Read, eds.). New York: Wiley, 523–527.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Enquiry.* Singapore: World Scientific Pub..

Roberts, G. O. and Rosenthal, J. S. (2001). Infinite hierarchies and prior distributions. *Bernoulli* **7**, 453–471.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica* **3**, 601–608.

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision* **40**, 192–214.

Rodríguez, C. C. (1991). From Euclid to entropy. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 343–348.

Rousseau, J. (2000). Coverage properties of one-sided intervals in the discrete case and application to matching priors. *Ann. Inst. Statist. Math.* **52**, 28–42.

Rosa, G. J. M. and Gianola, D. (2001). Inferences about the coefficient of correlation in the standard bivariate normal distribution. *South African Statist. J.* **35**, 69–93.

Roverato, A. and Consonni, G. (2004). Compatible Prior Distributions for DAG models *J. Roy. Statist. Soc. B* **66**, 47–61

Rueda, R. (1992). A Bayesian alternative to parametric hypothesis testing. *Test* **1**, 61–67.

Sansó, B. and Pericchi, L. R. (1992). Near ignorance classes of log-concave priors for the location model. *Test* **1**, 39–46.

Sansó, B. and Pericchi, L. R. (1994). On near ignorance classes. *Rev. Brasilera Prob./ Estatist.* **8**, 119–126.

Sareen, S. (2003). Reference Bayesian inference in nonregular models. *J. Econom.* **113**, 265–288.

Savage, L. J. (1954). *The Foundations of Statistics.* New York: Wiley. Second edition in 1972, New York: Dover.

Scholl, H. R. (1998). Shannon optimal priors on independent identically distributed statistical experiments converge weakly to Jeffreys' prior. *Test* **7**, 75–94.

Scricciolo, C. (1999). Probability matching priors: A review. *J. It. Statist. Soc.* **8**, 83–100.

Sendra, M. (1982). Distribución final de referencia para el problema de Fieller-Creasy. *Trab. Estadist.* **33**, 55–72.

Severini, T. A. (1991). On the relationship between Bayesian and non-Bayesian interval estimates. *J. Roy. Statist. Soc. B* **53**, 611–618.

Severini, T. A. (1993). Bayesian interval estimates which are also confidence intervals. *J. Roy. Statist. Soc. B* **53**, 611-618.

Severini, T. A. (1995). Information and conditional inference. *J. Amer. Statist. Assoc.* **90**, 1341-1346.

Severini, T. A. (1999). On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters. *Statistica Sinica* **9**, 713–724.

Severini, T. A., Mukerjee, R. and Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89**, 953–957.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423 and 623–656. Reprinted in *The Mathematical Theory of Communication* (Shannon, C. E. and Weaver, W., 1949). Urbana, IL.: Univ. Illinois Press.

Shieh, G. and Lee, J. C. (2002). Bayesian prediction analysis for growth curve model using noninformative priors. *Ann. Inst. Statist. Math.* **54**, 324–337.

Singh, U. and Upadhyay, S. K. (1992). Bayes estimators in the presence of a guess value. *Comm. Statist. Simul. and Comput.* **21**, 1181–1198.

Singh, U., Gupta, P. K. and Upadhyay, S. K. (2002). Estimation of exponentiated Weibull shape parameters under linex loss functions. *Comm. Statist. Simul. and Comput.* **31**, 523–537.

Smith, R. L. (1999). Bayesian and frequentist approaches to parametric predictive inference. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 589–612, (with discussion).

Smith, R. L. and Naylor, J. C. (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Appl. Statist.* **36**, 358–369.

Sono, S. (1983). On a noninformative prior distribution for Bayesian inference of multinomial distribution's parameters. *Ann. Inst. Statist. Math.* **35**, 167–174.

Spall, J. C. and Hill, S. D. (1990). Least informative Bayesian prior distributions for finite samples based on information theory. *IEEE Trans. Automatic Control* **35**, 580–583.

Spiegelhalter, D. J. (1985). Exact Bayesian inference on the parameters of a Cauchy distribution with vague prior information. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 743–749.

Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Statist.* **30**, 877–880.

Stein, C. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc. B* **24**, 265–296 (with discussion).

Stein, C. (1986). On the coverage probability of confidence sets based on a prior distribution. *Sequential Methods in Statistics, 3rd ed* (G. B. Wetherbil and K. D. Glazebrook, eds.) London: Chapman and Hall, 485—514.

Stephens, D. A. and Smith, A. F. M. (1992). Sampling-resampling techniques for the computation of posterior densities in normal means problems. *Test* **1**, 1–18.

Stern, J. M. (2004a). Paraconsistent sensitivity analysis for Bayesian significance tests. *Lecture Notes in Artificial Intelligence* **3171**, 134–143.

Stern, J. M. (2004b). Uninformative reference sensitivity in possibilistic sharp hypotheses tests. *Amer. Inst. Physics Proc.* 735, 581–588.

Stewart, W. E. (1987). Multiresponse parameter estimation with a new and noninformative prior. *Biometrika* **74**, 557–562.

Stone, M. (1959). Application of a measure of information to the design and comparison of experiments. *Ann. Math. Statist.* **30**, 55–70.

Stone, M. (1963). The posterior t distribution. *Ann. Math. Statist.* **34**, 568–573.

Stone, M. (1965). Right Haar measures for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **36**, 440–453.

Stone, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *Ann. Math. Statist.* **41**, 1939–1953.

Stone, M. (1976). Strong inconsistency from uniform priors. *J. Amer. Statist. Assoc.* **71**, 114–125 (with discussion).

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayesian inference in routine statistical problems. *Biometrika* **59**, 369–375.

Stone, M. and Springer, B. G. F. (1965). A paradox involving quasi prior distributions. *Biometrika* **52**, 623–627.

Strachan, R. W. and van Dijk, H. K. (2003). Bayesian model selection with an uninformative prior. *Oxford Bul. Econ. Statist.* **65**, 863–76.

Strawderman, W. E. (2000). Minimaxity. *J. Amer. Statist. Assoc.* **95**, 1364-1368.

Sugiura, N. and Ishibayashi, H. (1997). Reference prior Bayes estimator for bivariate normal covariance matrix with risk comparison. *Comm. Statist. Theory and Methods* **26**, 2203–2221.

Sun, D. (1994). Integrable expansions for posterior distributions for a two-parameter exponential family.*Ann. Statist.* **22**, 1808–1830.

Sun, D. (1997). A note on noninformative priors for Weibull distributions. *J. Statist. Planning and Inference* **61**, 319–338.

Sun, D. and Berger, J. O. (1994). Bayesian sequential reliability for Weibull and related distributions. *Ann. Inst. Statist. Math.* **46**, 221–249.

Sun, D. and Berger, J. O. (1998). Reference priors with partial information. *Biometrika* **85**, 55–71.

Sun, D., Ghosh, M. and Basu, A. P. (1998). Bayesian analysis for a stress-strength system under noninformative priors. *Can. J. Statist.* **26**, 323–332.

Sun, D. and Ni, S. (2004). Bayesian analysis of vector-autoregressive models with noninformative priors. *J. Statist. Planning and Inference* **121**, 291–309.

Sun, D., Tsutakawa, R. K. and He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica* **11**, 77–95.

Sun, D. and Ye, K. (1995). Reference prior Bayesian analysis for normal mean products. *J. Amer. Statist. Assoc.* **90**, 589–597.

Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika* **83**, 55-65.

Sun, D. and Ye, K. (1999). Reference priors for a product of normal means when variances are unknown. *Can. J. Statist.* **27**, 97–103.

Sweeting, T. J. (1985). Consistent prior distributions for transformed models. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), Amsterdam: North-Holland, 755–762.

Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model. *Biometrika* **71**, 127–134.

Sweeting, T. J. (1995a). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–24.

Sweeting, T. J. (1995b). A Bayesian approach to approximate conditional inference. *Biometrika* **82**, 25–36.

Sweeting, T. J. (1996). Approximate Bayesian computation based on signed roots of log-density rations. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 427–444, (with discussion).

Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88**, 657–675.

Tardella, L. (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika* **89**, 807–817.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Tiwari, R. C., Chib, S. and Jammalamadaka, S. R. (1989). Bayes estimation of the multiple correlation coefficient. *Comm. Statist. Theory and Methods* **18**, 1401–1413.

Torgersen, E. N. (1981). Measures of information based on comparison with total information and with total ignorance. *Ann. Statist.* **9**, 638–657.

Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British J. Math. Statist. Philosophy* **45**, 51–74.

Upadhyay, S. K. and Pandey, M. (1989). Prediction limits for an exponential distribution: a Bayes predictive distribution approach. *IEEE Trans. Reliability* **38**, 599–602.

Upadhyay, S. K. and Peshwani, M. (2003). Choice between Weibull and lognormal models: A simulation-based bayesian study. *Comm. Statist. Theory and Methods* **32**, 381–405.

Upadhyay, S. K. and Smith, A. F. M. (1994). Modelling complexities in reliability, and the role of simulation in Bayesian computation. *Internat. J. Continuing Eng. Education* **4**, 93–104.

Upadhyay, S. K., Agrawal, R. and Smith, A. F. M. (1996). Bayesian analysis of inverse Gaussian non-linear regression by simulation. *Sankhyā B* **58**, 363–378.

Upadhyay, S. K., Vasishta, N. and Smith, A. F. M. (2001). Bayes inference in life testing and reliability via Markov chain Montecarlo simulation. *Sankhyā A* **63**, 15–40.

Vaurio, J. K. (1992). Objective prior distributions and Bayesian updating. *Reliability Eng. System Safety* **35**, 55–59.

Vernotte, F. and Zalamansky, G. (1999). A Bayesian method for oscillator stability analysis. *IEEE Trans. Ultrasonics, Ferroelec./ and Freq./ Controls* **46**, 1545–1550.

Villegas, C. (1969). On the a priori distribution of the covariance matrix. *Ann. Math. Statist.* **40**, 1098–1099.

Villegas, C. (1971). On Haar priors. *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). Toronto: Holt, Rinehart and Winston, 409–414 (with discussion).

Villegas, C. (1977a). Inner statistical inference. *J. Amer. Statist. Assoc.* **72**, 453–458.

Villegas, C. (1977b). On the representation of ignorance. *J. Amer. Statist. Assoc.* **72**, 651–654.

Villegas, C. (1981). Inner statistical inference II. *Ann. Statist.* **9**, 768–776.

Villegas, C. (1982). Maximum likelihood and least squares estimation in linear and affine functional models. *Ann. Statist.* **10**, 256–265.

Villegas, C. (1990). Bayesian inference in models with Euclidean structures. *J. Amer. Statist. Assoc.* **85**, 1159–1164.

de Waal, D. J. and Nel, D. G. (1988). A procedure to select a ML-II prior in a multivariate normal case. *Comm. Statist. Simul. and Comput.* **17**, 1021–1035.

de Waal, D. J., and Groenewald, P. C. N. (1989). On measuring the amount of information from the data in a Bayesian analysis. *South African Statist. J.* **23**, 23–62, (with discussion).

de Waal, D. J., Groenewald, P. C. N. and Kemp, C. J. (1995). Perturbation of the Normal Model in Linear Regression. *South African Statist. J.* **29**, 105–130.

Wallace, C. S. and Dowe, D. L. (1999). Minimum message length and Kolmogorov complexity. *The Computer J.* 42, 270–283.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **49**, 240–265.

Walker, S. and Muliere, P, (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Ann. Statist.* **27**, 589–599.

Walker, S. G. and Gutiérrez-Peña, E. (1999). Robustifying Bayesian procedures. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 685–710, (with discussion).

Wasserman, L. (1996). The conflict between improper priors and robustness. *J. Statist. Planning and Inference* **52**, 1–15.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. B* **62**, 159–180.

Wasserman, L. and Clarke, B. (1995). Information tradeoff. *Test* **4**, 19–38.

Welch, B. L. (1965). On comparisons between confidence point procedures in the case of a single parameter. *J. Roy. Statist. Soc. B* **27**, 1–8.

Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on intervals of weighted likelihoods. *J. Roy. Statist. Soc. B* **25**, 318–329.

Wolfinger, R. D., Kass, R. E. (2000). Nonconjugate Bayesian analysis of variance component models. *Biometrics* **56**, 768–774.

Yang, R. (1995). Invariance of the reference prior under reparametrization. *Test* **4**, 83–94.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195–1211.

Yang, R. and Berger, J. O. (1997). A catalogue of noninformative priors. *Tech. Rep.*, Duke University, ISDS 97-42.

Yang, R and Chen, M.-H. (1995). Bayesian analysis of random coefficient regression models using noninformative priors. *J. Multivariate Analysis* **55**, 283–311.

Ye, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88**, 360–363.

Ye, K. (1994). Bayesian reference prior analysis on the ratio of variances for the balanced one-way random effect model. *J. Statist. Planning and Inference* **41**, 267–280.

Ye, K. (1995). Selection of the reference priors for a balanced random effects model. *Test* **4**, 1–17.

Ye, K. (1998). Estimating a ratio of the variances in a balanced one-way random effects model. *Statistics and Decisions* **16**, 163–180.

Ye, K. and Berger, J. O. (1991). Non-informative priors for inferences in exponential regression models. *Biometrika* **78**, 645–656.

Yuan, A. and Clarke, B. S. (1999). A minimally informative likelihood for decision analysis: Illustration and robustness, *Can. J. Statist.* **27**, 649–665.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley. Reprinted in 1987, Melbourne, FL: Krieger.

Zellner, A. (1977). Maximal data information prior distributions. *New Developments in the Applications of Bayesian Methods* (A. Ayka and C. Brumat, eds.). Amsterdam: North-Holland, 211–232.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *The Statistician* **32**, 23–34.

Zellner, A. (1986a). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 233–243.

Zellner, A. (1986b). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* **81**, 446–451.

Zellner, A. (1988). Optimal information processing and Bayes' theorem. *Amer. Statist.* **42**, 278–284 (with discussion).

Zellner, A. (1991). Bayesian methods and entropy in economics and econometrics. *Maximum Entropy and Bayesian Methods* (W. T. Grandy and L. H. Schick eds.). Dordrecht: Kluwer, 17–31.

Zellner, A. (1996). Models, prior information, and Bayesian analysis. *J. Econometrics* **75**, 51–68.

Zellner, A. (1997). *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*. Brookfield, VT: Edward Elgar.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypothesis. *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). Valencia: University Press, 585–603 and 618–647 (with discussion).

Zidek, J. V.(1969). A representation of Bayes invariant procedures in terms of Haar measure. *Ann. Inst. Statist. Math.* **21**, 291–308.