Bayesian Inference for Model Choice

John S.J. Hsu and Thomas Leonard

This manuscript is currently being revised.
Once the revision is completed,
the revised manuscript will be posted on my webpage at
http://www.pstat.ucsb.edu/faculty/John_Hsu/

Bayesian Inference for Model Choice

John S.J. Hsu and Thomas Leonard

Two nested or non-nested candidate sampling models for an observed data set may be compared by consideration of summaries of a probability plot, which contrasts the posterior quantiles of the log-likelihoods under the two models. The procedures address both *preference inference* and *refutation inference*, and extensions to DIC and alternatives to AIC are developed. Preference inference favors models with more parameters, perhaps on a tentative basis when further data are anticipated, while refutation inference emphasizes parameter parsimony. A characterization relating to an α-profile motivates the comparison of the posterior medians of the log-likelihoods, when considering simple model preference. For nested models, a stronger *omega-preference* procedure is developed via a Bayes-frequency compromise. The Bayes-frequency performances of the different preference and refutation inference procedures are investigated when the models are nested. While attention is primarily confined to model inference within the linear paradigm, most of the methods are approximately applicable in a range of non-linear cases. A Gamma approximation to an *Upsilon distribution* facilitates a general approach, for the linear model with unknown variance. A data set for 71 hypertensive diabetic patients is analyzed, and a symptom of high blood pressure is related to four out of the eight explanatory variables available.

KEY WORDS: Model choice, DIC, AIC, BIC, Bayesian inference, Bayes-frequency properties, Posterior Bayes factor, Probability plot, Posterior quantiles, Measures of evidence, Lindley's paradox, Bayesian profiles, Omega-preference, Tau-dominance, Lambda-optimality, Linear model, Generalized linear models, Upsilon-distribution, Medical applications, Blood pressure.

John S.J. Hsu is Associate Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA. Thomas Leonard, formerly Professor of Statistics at the Universities of Wisconsin-Madison and Edinburgh, is retired and lives at 4/3 Hopetoun Crescent, Edinburgh, Scotland, EH7 4AY. The authors would like to thank Angelika van der Linde for her helpful advice on the literature.

1. INTRODUCTION

Consider two candidate sampling models M_1 and M_2 for an $n \times 1$ vector of observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. Under model M_i and for $i = 1, 2$, $\mathbf{y}$ is a realization of a random vector $\mathbf{Y}$ with joint density or probability mass function $\eta_i(\mathbf{y}|\boldsymbol{\theta}_i)$, where the functional form of η_i is specified, and $\boldsymbol{\theta}_i$ is a $p_i \times 1$ vector of unknown parameters, with $p_1 \leq p_2$. The vector $\boldsymbol{\theta}_i$ is taken to possess some prior distribution, given M_i, under each choice of sampling model. Attention is for simplicity confined to situations where, for $i = 1, 2$, the posterior distribution given M_i of the likelihood

$$\ell_i = \ell_i(\boldsymbol{\theta}_i|\mathbf{y}) = \eta_i(\mathbf{y}|\boldsymbol{\theta}_i) \tag{1.1}$$

possesses a continuous density over some interval with positive support. This condition ensures that all internal posterior percentiles of the ℓ_i are uniquely defined. It is also assumed that each ℓ_i has a finite maximum, given the observed $\mathbf{y}$.

Aitkin (1991, 1997) measures the evidence in the data in favor of M_1 when compared with M_2 by the *posterior Bayes factor*

$$B^* = \ell_1^*/\ell_2^* \ , \tag{1.2}$$

where, for $i = 1, 2$, ℓ_i^* denotes the posterior expectation given M_i of ℓ_i. Aitkin shows that in some nested cases his criterion calibrates well with frequency probabilities. He hence avoids many of the pitfalls surrounding Lindley's paradox (for example, Lindley, 1957; Shafer, 1982; Sellke *et al.*, 2001) that are experienced by ordinary Bayes factors. The latter can in particular heavily favor M_1 in nested situations where a classical significance test heavily favors M_2. Aitkin also circumvents the high sensitivity (for example, O'Hagan, 1995) of ordinary Bayes factors to the choices of the prior distributions of the $\boldsymbol{\theta}_i$, for example to the tail behavior of prior densities. These problems can become particularly acute when the prior information is vague, in which case the ℓ_i^* can be much less sensitive.

While ordinary Bayes factors can be developed via elementary applications of Bayes' theorem, with positive prior probabilities (not depending upon the $\boldsymbol{\theta}_i$) assigned to M_1 and M_2, posterior Bayes factors are not open to possible justifications of this type. The comparison of summaries of the conditional posterior distributions of the ℓ_i, without reference to prior probabilities on the M_i, has however been considered by many authors. Rather than comparing the ℓ_i^*, Dempster (1974), Spiegelhalter *et al.* (2002), van der Linde (2005), and many others, address the DIC criteria

$$\phi_i^* = E[\phi_i|\mathbf{y}, M_i] \qquad (i = 1, 2) \ , \tag{1.3}$$

where $\phi_i = \log \ell_i$.

In many examples, in particular under the special formulation described in section 2, ϕ_i^* either equals or may be approximated for large n by $\hat{\phi}_i - \frac{1}{2}p_i$, where $\hat{\phi}_i$ denotes the maximized log-likelihood. This expression may be contrasted with the AIC criteria (Akaike, 1978),

$$\hat{\phi}_i^* = \hat{\phi}_i - p_i \qquad (i = 1, 2) \ , \tag{1.4}$$

which introduce heavier penalties for each parameter in the models. The criteria in (1.4) are more appealing when judging whether to refute the model of lower dimensionality, rather than when just making a statement of model preference. The parallel problems of *preference inference* and *refutation inference* will be addressed throughout the current paper.

It is however useful to compare the entire posterior distributions of ϕ_1 and ϕ_2. Together with any related alternatives to penalized likelihoods, these will, in many standard examples, depend approximately upon just n, the $\hat{\phi}_i$ and the p_i, and not unduly upon further measures of model complexity. For $i = 1, 2$, let

$$\Omega_i(d) = p(\phi_i > d | \mathbf{y}, M_i) \ , \tag{1.5}$$

and let $\tilde{\phi}_i$ denote the posterior median of ϕ_i. Model M_2 will be said to be *MIC preferred* to model M_1 if $\tilde{\phi}_2 > \tilde{\phi}_1$. Furthermore, M_1 can be regarded as refutable against M_2 for low values of

$$\alpha_1 = \Omega_1(\tilde{\phi}_2) \ , \tag{1.6}$$

together with high values of

$$\alpha_2 = \Omega_2(\tilde{\phi}_1) \ . \tag{1.7}$$

One sensible loose sense measure of the evidence in the data in favor of M_2 when compared with M_1 is provided by the *α-profile*

$$\alpha = \begin{cases} \alpha_2 & \text{if } \alpha_2 \geq 0.5 \\ 1 - \alpha_1 & \text{if } \alpha_2 < 0.5 \ , \end{cases} \tag{1.8}$$

which contrasts with the suggestion by Good(1991) that logarithms of ordinary Bayes factors provide the only valid measures of evidence. Good's formulation does not address Lindley's paradox, or over-sensitivity problems with respect to the prior assumptions. His specially axiomatized measures of evidence can however be readily updatable in the light of further data. The α-profile is less readily updatable since the effect of fresh data should be evaluated after updating the entire posterior distributions of the log-likelihoods. This extra complexity preserves information, and is not necessarily a disadvantage, given the other summaries of these distributions which are available.

Consider now the implications of the equivalence relation

$$\alpha_1 \leq \alpha_2 \Leftrightarrow \tilde{\phi}_1 \leq \tilde{\phi}_2 \quad , \tag{1.9}$$

with $\alpha_1 = \alpha_2$ if and only if $\tilde{\phi}_1 = \tilde{\phi}_2$. The equivalence in (1.9) would not generally be true if $\tilde{\phi}_1$ and $\tilde{\phi}_2$ were replaced in (1.6) and (1.7) by other posterior measures of location of the ϕ_i. This characterization motivates comparisons of the $\tilde{\phi}_i$, rather than the ϕ_i^* or the ℓ_i^*, when addressing preference inference. Comparisons of the posterior medians of the $t(\phi_i)$, where t is strictly increasing, are of course transformation-invariant.

Whether or not preference inference should be completed by simple comparisons of posterior measures of location of the ϕ_i provides another key issue. In particular, if $\tilde{\phi}_1$ is slightly less than $\tilde{\phi}_2$, then M_1 might still be preferable to M_2, if ϕ_1 possesses less posterior variability. In section 2, a Bayes-frequency compromise will address this problem when M_1 is nested into M_2.

Johnson (2005) recommends some useful alternatives to ordinary Bayes factors, when M_1 is nested into M_2, by contrasting the distribution given M_1 of a single dimensional test statistic which is ancillary under M_1, with its marginal distribution given M_2, under a prior distribution which expresses uncertainty about the hypothesis that M_1 is true. He estimates a hyperparameter by its marginal maximum likelihood estimate, yielding an estimated *dimensionality-reduced* Bayes factor with appealing properties when $p_2 - p_1$ is large. See section 7 for further discussion. Other contributions to the general area include Han and Carlin (2001), Berger and Guglielmi (2001), Lee and Berger (2001), Berger and Pericchi (2004) and Barbieri and Berger (2004).

2. A QUANTILE-QUANTILE APPROACH

For $i = 1, 2$, and for $0 \leq \beta_i < 1$, let $b_i = \lambda_i(\beta_i)$ denote the $(1-\beta_i)$th quantile of ϕ_i, where $\lambda_i = \Omega_i^{-1}$ is the inverse of the function in (1.5). Model M_1 should be refuted against model M_2 if $b_2 \geq b_1$ for a convincing selection of choices of β_1 and β_2. However

$$b_2 \geq b_1 \Leftrightarrow (\beta_1, \beta_2) \in \Lambda \quad , \tag{2.1}$$

where

$$\Lambda = \{(\beta_1, \beta_2) \in Q_U : \lambda_2(\beta_2) \geq \lambda_1(\beta_1)\} \quad , \tag{2.2}$$

with $Q_U = \{(\beta_1, \beta_2) : 0 \leq \beta_1, \beta_2 < 1\}$. Then $(\beta_1, \beta_2) \in \Lambda$ provides convenient shorthand notation for saying that "the $100(1-\beta_2)th$ posterior percentile given M_2 of ϕ_2 is not exceeded by the $100(1-\beta_1)th$ posterior percentile given M_1 of ϕ_1."

In order to refute M_1 in favor of M_2 we will require that $(\beta_1, \beta_2) \in \Lambda$ for low enough values of β_1, and high enough values of β_2. We will also require that $(\beta, \beta) \in \Lambda$ for an appropriately large region of β values. The Λ region should provide a convincingly prevalent subset of Q_U, with enough concentration towards the upper left closure point $(\beta_1, \beta_2) = (0, 1)$ of this unit quadrant.

Consider, the situation where, for each $\boldsymbol{\theta}_i$ in p_i dimensional Euclidean space R^{p_i}, the log-likelihoods satisfy

$$\phi_i = \hat{\phi}_i - \frac{1}{2} w_i \qquad (i = 1, 2) \quad , \tag{2.3}$$

with $\hat{\phi}_i$ denoting the ith maximized log-likelihood,

$$w_i = (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i)^T \mathbf{C}_i^{-1} (\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i) \qquad (i = 1, 2) \tag{2.4}$$

and $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{C}_i$ representing the corresponding maximum likelihood vectors and likelihood dispersion matrices. This standard representation holds exactly whenever, the sampling distribution of $\mathbf{Y}$ given $\boldsymbol{\theta}_i$ and M_i is multivariate normal for $i = 1, 2$, with

$$\mathbf{Y} | \boldsymbol{\theta}_i, M_i \sim N[\mathbf{X}_i \boldsymbol{\theta}_i, \mathbf{D}_i] \tag{2.5}$$

where the $\mathbf{X}_i$ and $\mathbf{D}_i$ are specified $n \times p_i$ and $n \times n$ matrices, and the $\mathbf{X}_i^T \mathbf{D}_i^{-1} \mathbf{X}_i$ are non-singular. Furthermore, with appropriate parametric normalizing parametrizations, the representation holds approximately when the sample sizes are finite for many other models, for example, for generalized linear models with logit or logarithmic link functions for multinomial, Poisson, or exponential data.

When (2.3) and (2.4) hold, prior information regarding the $\boldsymbol{\theta}_i$ can be readily incorporated via assumptions of multivariate normal prior distributions for the $\boldsymbol{\theta}_i$. Whether or not to complicate the issue of sampling model comparison with two confounding choices of prior distribution is open to discussion. However, if the prior distribution given M_i of each $\boldsymbol{\theta}_i$ is uniform over R^{p_i}, then the posterior distribution given M_i of w_i in (2.4) is quite simply chi-squared $\chi^2_{p_i}$ with p_i degrees of freedom. We therefore follow Aitken by proceeding with these technically convenient assumptions.

In the exact special case, the uniform prior distributions yield

$$\Lambda = \{(\beta_1, \beta_2) \in Q_U : T \geq \tau_{p_2}(\beta_2) - \tau_{p_1}(\beta_1)\} \quad , \tag{2.6}$$

where

$$T = 2(\hat{\phi}_2 - \hat{\phi}_1) \tag{2.7}$$

denotes the observed log-likelihood ratio statistic and for $i = 1, 2$, $\tau_{p_i} = \Psi_{p_i}^{-1}$ is the inverse of the cumulative distribution function (c.d.f) Ψ_{p_i} of a $\chi^2_{p_i}$ distribution. For $i = 1, 2$, the posterior medians of the ϕ_i are $\tilde{\phi}_i = \hat{\phi}_i - \frac{1}{2} med_{p_i}$, where $med_{p_i} = \tau_{p_i}(0.5)$. The probabilities in (1.6) and (1.7) satisfy

$$\alpha_1 = \Psi_{p_1}(med_{p_2} - T) \quad , \tag{2.8}$$

and

$$\alpha_2 = \Psi_{p_2}(T + med_{p_1}) \quad . \tag{2.9}$$

In situations where M_1 is nested into M_2 and T is a realization given M_1 of a $\chi^2_{p_2-p_1}$ variate, the expression for α_2 in (2.9) may be contrasted with the observed Type I frequency success probability

$$\gamma^* = \Psi_{p_2-p_1}(T) \tag{2.10}$$

for the likelihood ratio test. The representation in (2.10) applies exactly when (2.5) holds, with M_1 is nested into M_2 and $\mathbf{D}_1 = \mathbf{D}_2 = \sigma^2 \mathbf{I}_n$, for some common specified variance σ^2, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. The result also holds exactly under a variety of more general covariance assumptions. Approximate justifications of (2.10) for a variety of other nested models, including generalized linear models, are discussed by Dobson (1990, p. 61). As p_2 gets large and for any fixed p_1, the probabilities in (2.9) and (2.10) calibrate perfectly when regarded as functions of T. However $\alpha_2 < \gamma^*$ for any finite p_2 whenever $T > med_{p_2} - med_{p_1}$, i.e., $\tilde{\phi}_2 > \tilde{\phi}_1$. Nevertheless, α_2 and γ^* are well enough calibrated for finite p_2 to preclude an extreme analogue of Lindley's paradox when employing α in (1.8) as a measure of evidence.

The Λ region in (2.6) consists of all points in Q_U falling on or below an upper boundary U of posterior probabilities, defined by the curve

$$\beta_2 = \Psi_{p_2}\{T + \tau_{p_1}(\beta_1)\} \qquad (0 \leq \beta_1, \beta_2 < 1) \quad , \tag{2.11}$$

which is strictly increasing between $(\beta_1, \beta_2) = \{0, \Psi_{p_2}(T)\}$ and the upper right closure point $(\beta_1, \beta_2) = (1, 1)$. When $\tilde{\phi}_2 > \tilde{\phi}_1$, so that $T > med_{p_2} - med_{p_1}$, $(\beta_1, \beta_2) = (\frac{1}{2}, \frac{1}{2})$ falls below U and $(\beta_1, \beta_2) = (\frac{1}{2}, \alpha_2)$ provides the central point where U intersects the vertical $\beta_2 = \frac{1}{2}$ line. When $\tilde{\phi}_1 > \tilde{\phi}_2$, $(\beta_1, \beta_2) = (\frac{1}{2}, \frac{1}{2})$ falls above U, and the boundary point $(\beta_1, \beta_2) = (\alpha_1, \frac{1}{2})$ becomes more relevant. The α-profile in (1.8) therefore always represents some central point on U.

In Fig 1, the quadrant Q_U is depicted together with the curve in (2.11) when $p_1 = 2$, $p_2 = 10$, and $T = 11$. It is useful in visual terms to include the $\beta_1 = \frac{1}{2}$, $\beta_2 = \frac{1}{2}$, $\beta_2 = \beta_1$ and $\beta_2 = 1 - \beta_1$ lines, and to highlight the points $(0.314, 0.686)$, $(\frac{1}{2}, \alpha_2) = (0.500, 0.740)$ and

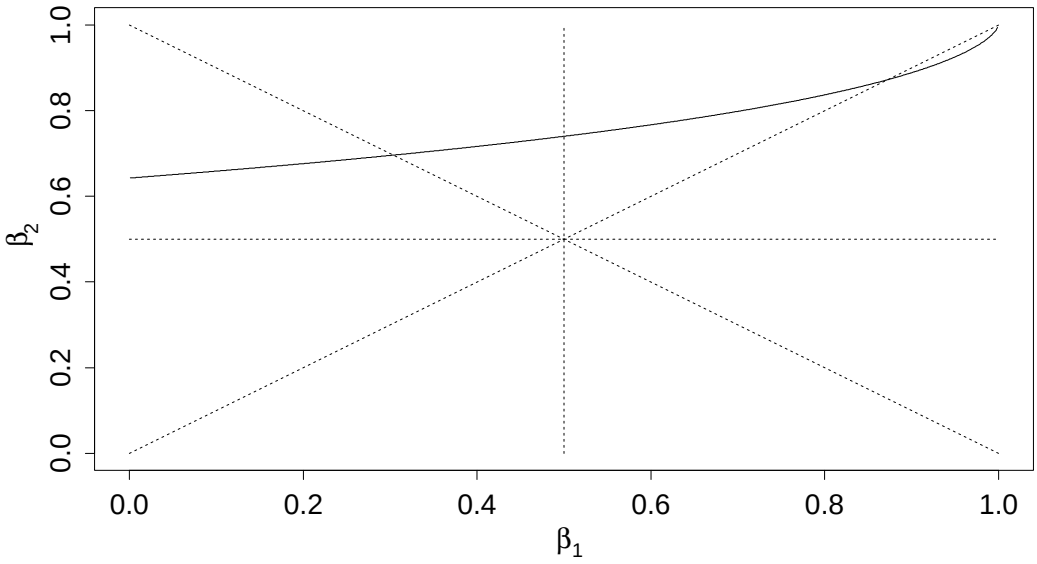

Figure 1: An upper boundary U

$(0.872, 0.872)$ of intersection of U with this *Union Jack* configuration. The U boundary will always intersect the $\beta_2 = 1 - \beta_1$ line at a unique point, and a unique point of intersection of the U boundary with the $\beta_2 = \beta_1$ line is guaranteed, when $T \geq 0$. A visual plot of U, together with experience from other data sets, and consideration of other models, may be used more generally to judge whether to refute M_1 against M_2. Guidance is provided by the α-profile, together with two further profiles *qsp* and *bsp* which respectively relate to the points where U intersects the $\beta_2 = 1 - \beta_1$ and $\beta_2 = \beta_1$ lines.

Definition 1: The *quantile significance profile qsp* is the largest β^* such that $(1 - \beta, \beta) \in \Lambda$ for all $\beta \leq \beta^*$.

Definition 2: The *Bayesian significance profile bsp*, if this exists, is the largest β^* such that $(\beta_1, \beta_2) \in \Lambda$ for all $\beta_1 \geq \beta^*$ and $\beta_2 \leq \beta^*$.

Under the special formulation in (2.3) and (2.4), *qsp* may be obtained by solving (2.11) with $\beta_1 = 1 - \beta^*$ and $\beta_2 = \beta^*$, and consequently uniquely satisfies the equation

$$T = \tau_{p_2}(\beta^*) - \tau_{p_1}(1 - \beta^*) \tag{2.12}$$

in β^*. If $T \geq 0$, then *bsp* solves (2.11) with $\beta_1 = \beta^*$ and $\beta_2 = \beta^*$ and therefore uniquely satisfies the equation

$$T = \tau_{p_2}(\beta^*) - \tau_{p_1}(\beta^*) \tag{2.13}$$

in β^*. Then *bsp* provides the largest β^* such that $(\beta^*, \beta^*) \in \Lambda$, and relates to a concept of *partial stochastic dominance*.

If $\tilde{\phi}_2 > \tilde{\phi}_1$ then $qsp < \alpha_2 < bsp$ and high values of the three key profiles *qsp*, α_2 and *bsp* indicate possible refutation of M_1 in favor of M_2. However moderately high values of *qsp* and

α_2 should be regarded as more convincing than extremely high values of *bsp*. For example, if $qsp = 0.9$ then the tenth posterior percentile of ϕ_2 equals the ninetieth posterior percentile of ϕ_1. If also $bsp = 0.99$ then the first posterior percentile of ϕ_1 and ϕ_2 are equal, and any posterior percentile of ϕ_2 above the first exceeds the corresponding posterior percentile of ϕ_1. If say, $\alpha_2 = 0.92$, then a larger value of *bsp* exceeding 0.99, for example, $bsp = 0.9999$, would not appear to provide substantial further evidence against M_1.

We now consider a rather heuristic idea, which will be justified in sections 4 and 7 by its Bayes-frequency properties. When M_1 is nested into M_2, we recommend seeking a critical value of T_o of T equating the Type I frequency probability γ^* in (2.10) with the value of *bsp* satisfying (2.13) when $T = T_o$. The corresponding common value of γ^* and *bsp* will be referred to as the *matched significance probability msp*.

When $\tilde{\phi}_2 > \tilde{\phi}_1$, it would be impossible to instead match γ^* with *qsp* or α_2 since $qsp < \alpha_2 < \gamma^*$. However, $bsp > \gamma^*$ whenever $\gamma^* > msp$, and $bsp < \gamma^*$ whenever $\gamma^* < msp$. Furthermore *msp* satisfies the algebraically appealing equation

$$\tau_{p_2-p_1}(\beta^*) = \tau_{p_2}(\beta^*) - \tau_{p_1}(\beta^*) \tag{2.14}$$

in β^*. The solution for *msp* may be interpreted as uniquely satisfying the property that M_2 should be preferred to M_1, if and only if both γ^* in (2.10) and *bsp* exceed *msp*. All properties and inequalities relating γ^*, *bsp* and *msp* may be clarified upon plotting and comparing the intersecting and strictly increasing graphs of γ^* and *bsp* against critical values T_o of T. These graphs also motivate a Bayes-frequency compromise, which will be shown in sections 3, 7, and 8 to be slightly stronger than MIC preference. The following definition is also intended to apply more generally, to situations where the Type I frequency probability in (2.10) assumes different forms.

Definition 3: If M_1 is nested into M_2 and under assumptions where *msp* uniquely exists, M_2 is *omega-preferred* to M_1 if the observed γ^* exceeds *msp*.

When considering either nested or non-nested models, a concept of *tau-dominance*, as defined below, may be considered for refutation inference. When the models are non-nested, the availability of *qsp* and the α-profile compensates for the lack of a single frequency probability that addresses the evidence in the data in a simple manner. If M_1 is nested into M_2 then *bsp* may also be usefully reported.

Definition 4: Model M_2 *tau-dominates model* M_1 *with quantile significance profile qsp* if *qsp* exceeds 0.75.

The curve in Fig 1 yields the values $qsp = 0.686$, $\alpha = 0.740$, and $bsp = 0.872$ for the three Bayesian profiles. As $qsp \leq 0.75$, model M_2 does not tau-dominate model M_1. However,

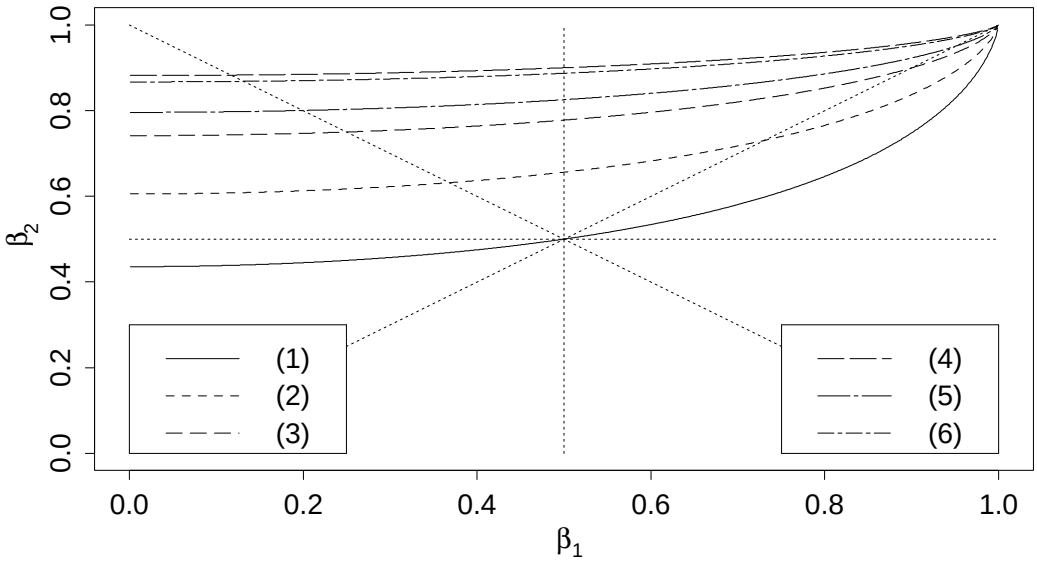

Figure 2: Upper boundaries for six critical values T_o of T ($\nu = 4$)

$msp = 0.669$ when $p_1 = 2$ and $p_2 = 10$, and the frequency success probability for $T = 11$ is $\gamma^* = \Psi(11) = 0.798$. As $\gamma^* > msp$, or equivalently $bsp > msp$, model M_2 is omega-preferred to model M_2. Model M_2 is also MIC preferred to model M_1, since the critical value for MIC preference is $T_o = med_{10} - med_2 = 7.966$, corresponding to $\gamma^* = 0.562$. The critical value for omega-preference is $T_o = 9.141$.

When considering both preference and refutation inference for l models $M_1, M_2, \ldots, M_l$, with respective number of parameters $p_1, p_2, \ldots, p_l$, a preliminary scan is recommended, before applying the preceding ideas to the detailed comparisons of important pairs of models. For each $i = 1, 2, \ldots, l$, the three posterior quantiles $QIC1_i$, $QIC2_i$, and $QIC3_i$ of the $-2\phi_i$ may be reported and contrasted with the criteria $DIC^* = -2\phi_i^*$ and $AIC^* = -2\hat{\phi}_i + 2p_i$. The models may then be evaluated according to the MIC preference and tau-dominance criteria. Model M_i tau-dominates model M_k whenever $QIC3_i < QIC1_k$.

3. INVESTIGATING EQUALITY

Under model M_2 take the elements Y_i of $\mathbf{Y}$ to be independent and respectively normally $\mathrm{N}(\theta_i, \sigma^2)$ distributed, for $i = 1, 2, \ldots, n$, where σ^2 is specified. Let M_1 represent the reduced form of M_2 which takes each θ_i to equal a common unknown value θ. Let $T = \sum(y_i - \bar{y})^2/\sigma^2$ denote the log-likelihood ratio statistic which under M_1 is the realization of a χ^2_ν variate, with $\nu = n - 1$. The six curves in Fig 2 describe the upper boundary U in (2.11) for the following six critical values of T_o for T when $p_1 = 1$ and $p_2 = n = 5$ and under appropriate uniform prior assumptions for the unknown parameters:
(1) The MIC critical value, $T_o = med_n - med_1$,

(2) The omega-preference value, $T_o = \tau_n(\beta^*) - \tau_1(\beta^*)$, where β^* satisfies (2.14),
(3) The tau-dominance critical value, $T_o = \tau_n(0.75) - \tau_1(0.25)$,
(4) The $\alpha_2 = 0.90$ critical value, $T_o = \tau_n(0.90) - med_1$,
(5) The $bsp = 0.95$ value, $T_o = \tau_n(0.95) - \tau_1(0.95)$,
(6) The $bsp = 0.99$ value, $T_o = \tau_n(0.99) - \tau_1(0.99)$.

Table 1: Bayesian Profiles and Type I Frequency Success Probabilities

	T_o	qsp	α_2	bsp	γ^*
(1)	3.897	0.500	0.500	0.500	0.580
(2)	5.176	0.632	0.656	0.730	0.730
(3)	6.524	0.750	0.778	0.899	0.837
(4)	8.781	0.883	0.900	0.994	0.933
(5)	7.229	0.800	0.825	0.950	0.876
(6)	8.451	0.868	0.887	0.990	0.924
(7)	9.488	0.909	0.923	0.998	0.950
(8)	13.277	0.979	0.983	1.000	0.990

These six values for T_o are listed in the second column of Table 1, followed by the critical values for the 5% and 1% level tests in rows (7) and (8). In the third to sixth columns, the corresponding values of the three profiles qsp, α_2 and bsp, and the $\gamma^* = \Psi_\nu(T_o)$, are listed. The critical value $T_o = 3.897$, corresponding to MIC preference, yields curve (1), which traverses the central point $(\beta_1, \beta_2) = (\frac{1}{2}, \frac{1}{2})$ of Q_U. Curve (2) is slightly more closely directed towards the upper left vertex $(\beta_1, \beta_2) = (0, 1)$. This curve corresponds to $T_o = 5.176$ and omega preference. This larger value for T_o indirectly takes into account the spread as well as the locations of the posterior distributions of the log-likelihoods, and is proposed for preference inference.

Values of T exceeding $T = T_o = 6.524$ would provide clearer evidence to refute M_1 in favor of M_2. This critical value relates to the $qsp = 0.75$ curve (3) which traverses the point $(\beta_1, \beta_2) = (\frac{1}{4}, \frac{3}{4})$ and roughly speaking splits the upper left subquadrant of Q_U into halves. The concept of tau-dominance appears to be plausible for mild refutation inference.

The critical values $T_o = 7.229$ and $T_o = 8.451$ yield $bsp = 0.95$ and 0.99 and curves (5) and (6) respectively. These facilitate stronger Bayesian inferences regarding the refutation of M_1 in favor of M_2. They correspond to respective significance probabilities $1 - \gamma^* = 0.124$ and 0.076 and contrast with the critical values $T_o = 9.488$ and $T_o = 13.277$ for the classical

5% and 1% fixed size log-likelihood ratio tests. The critical value when $\alpha_2 = 0.90$ is instead $T_o = 8.781$ and this yields curve (4) in Fig 2.

Table 2: Critical Values T_o for Varying ν

ν	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	2ν
1	0.93	1.54	2.67	4.15	2.15	2.58	3.84	6.63	2
2	1.91	2.80	4.01	5.80	3.97	4.71	5.99	9.21	4
3	2.90	4.00	5.28	7.32	5.65	6.64	7.81	11.34	6
4	3.90	5.18	6.52	8.78	7.23	8.45	9.49	13.28	8
5	4.89	6.33	7.74	10.19	8.75	10.18	11.07	15.09	10
6	5.89	7.47	8.94	11.56	10.23	11.84	12.59	16.81	12
7	6.89	8.60	10.12	12.91	11.67	13.46	14.07	18.48	14
8	7.89	9.72	11.29	14.23	13.08	15.03	15.51	20.09	16
9	8.89	10.84	12.45	15.53	14.47	16.57	16.92	21.67	18
10	9.89	11.94	13.60	16.82	15.83	18.09	18.31	23.21	20
15	14.88	17.42	19.27	23.09	22.45	25.37	25.00	30.58	30
20	19.88	22.82	24.83	29.16	28.83	32.30	31.41	37.57	40
30	29.88	33.49	35.79	40.97	41.14	45.56	43.77	50.89	60
60	59.88	65.02	67.94	75.06	76.39	82.96	79.08	88.38	120
100	99.88	106.53	110.08	119.13	121.61	130.35	124.34	135.82	200

The values of $\nu = n - 1$ in Table 2 are also the critical values of T for *simple DIC preference*. These slightly exceed the MIC preference values (1). The values of 2ν in the last column are the critical values for T corresponding to the AIC criteria. These may are useful for refutation inference, in particular when $\nu \leq 30$, though they become quite large in comparison with the other entries when ν is very large. The omega-preference values (2) fall between the MIC values (1) and the tau-dominance critical values (3). The $bsp = 0.95$ values (5) are always less than the critical values (7) for the 5% likelihood ratio test, and the $bsp = 0.99$ values (6) may be similarly compared with the values (8) for the 1% test. The $\alpha_2 = 0.90$ values (4) are, for small ν, larger than both the $bsp = 0.95$ values and the 5% test critical values, but become more conservative for large ν.

4. BAYES-FREQUENCY CALCULATIONS

In the special case discussed in section 3 and with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)^T$, a Type II

frequency success probability

$$\kappa^* = p(T \geq T_o|\boldsymbol{\theta}, M_2) \tag{4.1}$$

may be calculated for any critical value T_o, where T now represents a non-central chi-squared variate, with ν degrees of freedom and non-centrality parameter $\psi = \sum(\theta_i - \bar{\theta})^2/2\sigma^2$. It can however be easier to calculate an average Type II probability $\kappa = E[\kappa^*]$, where the expectation is with respect to an appropriate mixing distribution on θ_i, in this case not the prior distribution. If the θ_i are independent and $N(\mu, \sigma_o^2)$ distributed, then the marginal distribution given M_2 of $\sigma^2 T/(\sigma^2 + \sigma_o^2)$ is central χ_ν^2 so that

$$\kappa = \kappa(\xi) = 1 - \Psi_\nu[T_o/\{1 + (\xi^2/med_\nu\}] \qquad (0 < \xi < \infty) \ , \tag{4.2}$$

where $\xi^2 = \sigma_o^2 med_\nu/\sigma^2$ is the difference between the marginal medians of T, given M_2, and given M_1. The ξ parametrization sensibly scales the curves in (4.2) for varying ν.

When seeking choices of T_o which balance the Type I success probability $\gamma^* = \Psi_\nu(T_o)$ and the average Type II success probabilities, it is useful to plot the curves

$$MINSP(\xi) = min\{\gamma^*, \kappa(\xi)\} \tag{4.3}$$

and

$$AVESP(\xi) = \{\gamma^* + \kappa(\xi)\}/2 \ , \tag{4.4}$$

for $0 < \xi < \infty$. In Figs 3 and 4, these curves are contrasted when $\nu = 4$ for three critical choices T_o of T, namely (a) the MIC value 3.897 (b) the omega-preference value 5.176 and (c) the tau-dominance value 6.524. Omega-preference yields superior $MINSP$ values (≥ 0.580) when compared with MIC for all values of ξ such that $\xi \geq 1.642$, and superior $AVESP$ values (≥ 0.514) when compared with MIC for all values of ξ such that $\xi \geq 0.961$. Overall comparisons of curves (a) and (b) in Figs 3 and 4 suggests that omega-preference yields slightly more reasonable and more balanced average success probability properties than simple MIC preference. This provides a Bayes-frequency justification for omega-preference.

Omega-preference yields superior $MINSP$ values (≤ 0.730) when compared with tau-dominance for all values of ξ such that $\xi \leq 2.727$, and superior $AVESP$ values (≤ 0.704) when compared with tau-dominance for all values of ξ such that $\xi \leq 0.564$. Overall comparisons of curves (b) and (c) of Figs 3 and 4 suggest that omega-preference and tau-dominance possess equitable average success probability properties. Both criteria should perhaps be considered when contemplating simple model preference, though tau-dominance provides more evidence for refutation of the simpler model. In Table 3, the points of intersection of the (a) and (b) curves and (b) and (c) curves are reported for both $MINSP$ and $AVESP$

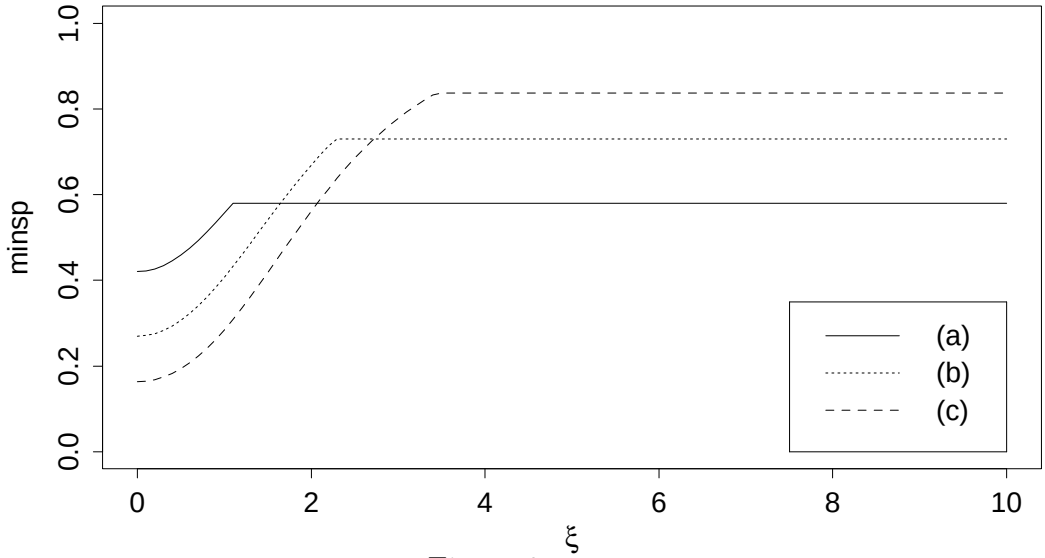

Figure 3: MINSP ($\nu = 4$)

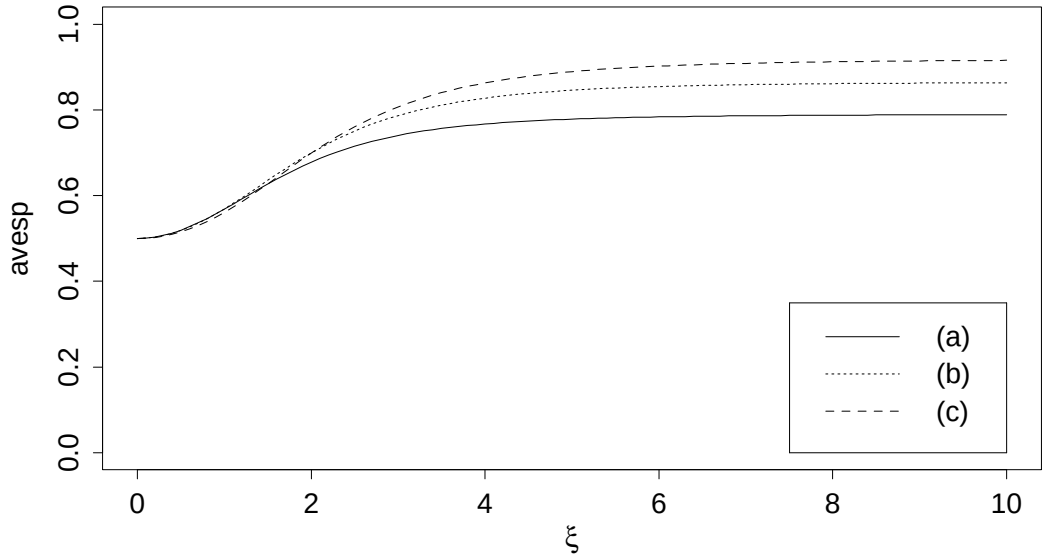

Figure 4: AVESP ($\nu = 4$)

and a range of values of ν. Comparisons of the unreported full average success probability curves would justify extending our preceding conclusions for $\nu = 4$ to general ν.

Similar $AVESP$ and $MINSP$ comparisons may be made for a variety of critical values for refutation inference. For example, when $\nu = 9$, the tau-dominance criterion possesses superior $AVESP$ when compared with the $bsp(95\%)$ criterion for all values of ξ whenever $\xi < 3.819$ and $AVESP < 0.811$ and superior $MINSP$ whenever $\xi < 3.376$ and $MINSP < 0.811$. Again when $\nu = 9$, the $bsp(95\%)$ criterion possesses superior $AVESP$ when compared with the fixed size 5% level significance test whenever $\xi < 4.977$ and $AVESP < 0.893$ and superior $MINSP$ whenever $\xi < 4.495$ and $MINSP < 0.895$. Qualitatively similar conclusions hold for general ν and for fixed levels of the criteria other than 5%. Further

Table 3: Intersection Points of Success Probability Curves

ν	(1)	(2)	(3)	(4)
1	(1.812,0.665)	(3.981,0.785)	(0.467,0.547)	(1.388,0.686)
2	(1.615,0.615)	(2.900,0.753)	(0.710,0.556)	(1.727,0.697)
3	(1.616,0.593)	(2.749,0.739)	(0.855,0.561)	(1.907,0.702)
4	(1.642,0.580)	(2.727,0.730)	(0.961,0.564)	(2.036,0.704)
5	(1.674,0.571)	(2.740,0.725)	(1.045,0.567)	(2.139,0.706)
6	(1.705,0.564)	(2.767,0.721)	(1.116,0.568)	(2.225,0.707)
7	(1.736,0.559)	(2.799,0.717)	(1.177,0.570)	(2.301,0.708)
8	(1.765,0.556)	(2.834,0.715)	(1.230,0.571)	(2.368,0.708)
9	(1.793,0.552)	(2.868,0.713)	(1.279,0.572)	(2.430,0.709)
10	(1.820,0.549)	(2.903,0.711)	(1.322,0.573)	(2.486,0.709)
15	(1.936,0.540)	(3.063,0.706)	(1.498,0.576)	(2.715,0.711)
20	(2.033,0.535)	(3.202,0.702)	(1.631,0.578)	(2.893,0.712)
30	(2.187,0.528)	(3.431,0.698)	(1.831,0.581)	(3.166,0.713)
60	(2.506,0.520)	(3.915,0.694)	(2.214,0.584)	(3.707,0.714)
100	(2.789,0.515)	(4.349,0.691)	(2.535,0.585)	(4.173,0.714)

(1) $MINSP$, MIC and omega-preference
(2) $MINSP$, omega-preference and tau-dominance
(3) $AVESP$, MIC and omega-preference
(4) $AVESP$, omega-preference and tau-dominance

comparisons indicate that the $AIC = 2\nu$ criterion does not perform particularly well when ν is very large. For example, when $\nu = 100$ the $\alpha_2(0.99\%)$ criterion $T_o = 135.82$ possesses superior $AVESP$ when compared with $AIC = 200$ when $\xi < 13.575$ and $AVESP < 0.982$, and superior $MINSP$ when $\xi < 13.575$ and $MINSP < 0.990$.

A Bayes-frequency alternative λ to the Bayesian profiles *qsp*, α_2, and *bsp* is now introduced. For any fixed ξ, $AVESP$ in (4.4) is maximized with respect to T_o whenever

$$T_o = \nu(1 + R^{-1}) \log(1 + R)$$

where $R = \xi^2/med_v = \sigma_o^2/\sigma^2$. Let $MAXAVESP(\xi)$ denote the corresponding maximizing value of $AVESP$.

Definition 5: The critical value of T_o is *lambda-optimal at level* λ if T_o maximizes $AVESP(\xi)$ with respect to T_o for some ξ, and $MAXAVESP(\xi) = \lambda$.

The lambda-optimal critical value for T_o may be plotted as a strictly increasing function of ξ, and the corresponding values of λ calculated. Therefore λ may be plotted against T_o. The results in Table 4 indicate close calibrations between the values of the Bayesian profiles qsp and α_2 and the Bayes-frequency criterion λ, when $\lambda \geq 0.750$. Both qsp and α_2 are quite close to λ even when $\nu = 1$ and the accuracy of calibration increases as ν increases. Indeed, the interval (qsp, α_2) always contains λ. This result is unexpected given the very different formulations for λ and our other criteria.

A similar Bayes-frequency formulation, but based upon the $MINSP$ criterion leads to slightly less convincing calibrations with qsp and α_2. Indeed the preceding close calibrations of qsp and α_2 with the $AVESP$ criterion, do not extend to the more general situations described in section 2, when $p_1 \geq 2$ (see section 7). However, λ will be more generally recommended, together with qsp and α_2, as a useful criterion for refutation inference.

5. THE UPSILON DISTRIBUTION

Consider the standard linear model where $\mathbf{Y}|\boldsymbol{\theta}, \sigma^2 \sim N[\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I}_n]$, with $\mathbf{R} = \mathbf{X}^T\mathbf{X}$ non-singular, $\boldsymbol{\theta}$ denoting an unknown $p \times 1$ vector of parameters, and σ^2 also unknown. If the prior distribution of $(\boldsymbol{\theta}, \log\sigma^2)$ is uniform on R^{p+1}, then the conditional posterior distribution of $\boldsymbol{\theta}$ given σ^2 is $N[\hat{\boldsymbol{\theta}}, \sigma^2\mathbf{R}^{-1}]$, where $\hat{\boldsymbol{\theta}}$ is the least squares vector. Furthermore, the posterior distribution of $U = S^2/\sigma^2$ is χ^2_ν, where $\nu = n - p$ and S^2 denotes the observed residual sum of squares. If ϕ and $\hat{\phi}$ are respectively the log-likelihood, and maximized log-likelihood, of $\boldsymbol{\theta}$ and σ^2, then $\Upsilon = 2(\hat{\phi} - \phi)$ satisfies

$$\Upsilon = U + \sigma^{-2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T\mathbf{R}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) - n\log U + n\log n - n \quad , \tag{5.1}$$

so that, in the posterior assessment,

$$\Upsilon \overset{d}{\sim} n\log n - n - n\log\chi^2_\nu + \chi^2_\nu + \chi^2_p \quad , \tag{5.2}$$

where χ^2_ν and χ^2_p are independent chi-squared variates, with the designated degrees of freedom.

Definition 6: A random variable Υ satisfying (5.2) with $\nu = n - p$ is said to possess an Υ-*distribution with sample size* n *and* ν *degrees of freedom.* If ν is set equal to n in (5.2), and the χ^2_p term omitted, then Υ possesses an Υ-distribution with sample size n and n degrees of freedom.

Table 4: Bayes-Frequency Calibrations

	ξ	T_o	λ	qsp	α_2	bsp	γ^*
	2.003	2.54	0.750	0.735	0.777	0.988	0.889
	2.782	3.06	0.800	0.791	0.828	0.999	0.920
$\nu = 1$	4.092	3.73	0.850	0.848	0.877	1.000	0.947
	6.784	4.67	0.900	0.904	0.923	1.000	0.969
	15.316	6.26	0.950	0.956	0.965	1.000	0.988
	92.277	9.84	0.990	0.993	0.994	1.000	0.998
	2.403	6.33	0.750	0.735	0.763	0.881	0.824
	2.897	7.02	0.800	0.786	0.812	0.937	0.865
$\nu = 4$	3.532	7.87	0.850	0.839	0.861	0.977	0.904
	4.456	9.04	0.900	0.893	0.909	0.996	0.940
	6.190	10.96	0.950	0.948	0.956	1.000	0.973
	11.486	15.16	0.990	0.990	0.992	1.000	0.996
	2.817	13.39	0.750	0.738	0.758	0.829	0.797
	3.287	14.33	0.800	0.788	0.807	0.888	0.842
$\nu = 10$	3.843	15.48	0.850	0.840	0.856	0.938	0.884
	4.577	17.01	0.900	0.893	0.905	0.977	0.926
	5.772	19.45	0.950	0.947	0.953	0.997	0.965
	8.565	24.58	0.990	0.990	0.991	1.000	0.994
	3.226	24.62	0.750	0.741	0.756	0.804	0.783
	3.711	25.86	0.800	0.791	0.805	0.861	0.829
$\nu = 20$	4.265	27.36	0.850	0.842	0.854	0.914	0.874
	4.961	29.32	0.900	0.894	0.903	0.959	0.918
	6.026	32.40	0.950	0.947	0.952	0.991	0.961
	8.238	38.71	0.990	0.989	0.991	1.000	0.993
	4.573	109.86	0.750	0.745	0.753	0.773	0.765
	5.172	112.40	0.800	0.795	0.802	0.827	0.813
$\nu = 100$	5.823	115.40	0.850	0.846	0.852	0.879	0.861
	6.597	119.26	0.900	0.897	0.901	0.928	0.908
	7.686	125.14	0.950	0.948	0.951	0.972	0.955
	9.650	136.69	0.990	0.990	0.990	0.998	0.991

For $\nu < n$, the c.d.f. of this Υ^n_ν distribution is

$$\Upsilon^n_\nu(d) = E\Psi_p[n \log U - U + d - \log n + n] \qquad (0 < d < \infty) \tag{5.3}$$

where the expectation should be taken with respect to a χ^2_ν distribution for U. Following standard asymptotics surrounding the approximations discussed in section 2, Υ will converge in distribution to a χ^2_{p+1} variate as $n \to \infty$, with p fixed. However, for finite n the posterior distribution of $\boldsymbol{\theta}$ and $\log \sigma^2$ is only roughly multivariate normal, and the distribution of Υ may not be closely approximated by χ^2_{p+1}.

The distribution of Υ may be more accurately approximated, when n is finite and $\nu < n$, by a Gamma distribution with mean

$$E^*[\Upsilon] = n(\log n - \log \nu + \nu^{-1}) \tag{5.4}$$

and variance

$$var^*[\Upsilon] = 2\nu^{-1}n(p + \nu^{-1}n) \quad . \tag{5.5}$$

which may be justified via a finite ν entropy argument described by Leonard and Hsu (1999, p. 36). They show that the distribution of $\log \chi^2_\nu$ is approximately normal with saddle-point accuracy, with mean

$$E^*[\log \chi^2_\nu] = \log \nu - \nu^{-1} \tag{5.6}$$

and variance $2\nu^{-1}$. The exact mean and variance of $\log \chi^2_\nu$ may be expressed in terms of the digamma and trigamma functions, and the variance is twice the derivative with respect to ν of the mean. Differentiating (5.6) motivates the adjusted approximation

$$var^*[\log \chi^2_\nu] = 2(\nu^{-1} + \nu^{-2}) \tag{5.7}$$

to the variance. Furthermore, $cov(\chi^2_\nu, \log \chi^2_\nu)$ is exactly equal to two. Combining these results gives from (5.2), and after some slight algebraic rearrangements, the expressions in (5.4) and (5.5). A Gamma distribution with this mean and variance is recommended as an approximation to the Υ^n_ν distribution.

The expressions in (5.4) and (5.5) respectively converge to $p+1$ and $2(p+1)$ as $n \to \infty$, with p fixed, as required by the χ^2_{p+1} limit. In the special case where the mean vector of $\mathbf{y}$ is specified and only σ^2 is unknown, ν in (5.2) should be set equal to n, and the χ^2_p term omitted. In this case, setting $p = 0$ in the preceding Gamma approximation suggests a finite sample χ^2_1 approximation to a Υ^n_n distribution.

The c.d.f of the Gamma approximation gives three decimal point accuracy when compared with the exact c.d.f. for many values of n and ν, while a χ^2_{p+1} approximation does

not provide such accuracy. This accuracy is for example available even when $(p, n) = (0, 40), (15, 20), (30, 40)$, or $(50, 60)$, and for all values of p and n considered in sections 7 and 8. All subsequent numerical calculations in this paper will be reported subject to Gamma approximations of this type.

6. COMPARING LINEAR MODELS

Consider now the comparisons of models M_1 and M_2, where, for $i = 1, 2$, $\mathbf{Y}$ given M_i follows the specification in (2.5), with $\mathbf{D}_i = \sigma_i^2 \mathbf{I}_n$, where σ_i^2 is unknown, and $\mathbf{X}_i^T \mathbf{X}_i$ is nonsingular. The observed log-likelihood ratio statistic T for comparing M_1 and M_2 may be expressed in the form

$$T = n \log(1 + \delta F) \quad , \tag{6.1}$$

where $\delta = \omega/\nu_2$ and

$$F = \delta^{-1}(S_1^2 - S_2^2)/S_2^2 \quad , \tag{6.2}$$

with $\omega = p_2 - p_1$, $\nu_2 = n - p_2$ and where, for $i = 1, 2$, S_i^2 is the residual sum of squares under model M_i. If M_1 is nested into M_2 the random analogue $\tilde{F}$ of (6.2) is $F_{\nu_2}^{\omega}$ distributed given M_1, i.e., F-distributed with ω and ν_2 degrees of freedom.

For $i = 1, 2$, we take the prior distribution given M_i of $(\theta_i, \log \sigma_i^2)$ to be uniform over R^{p_i+1}. Then, following the representations of section 5, the log-likelihoods ϕ_i of the $\boldsymbol{\beta}_i$ and σ_i^2 may be expressed in the forms

$$\phi_i = \hat{\phi}_i - \frac{1}{2}\Upsilon_i \quad , \tag{6.3}$$

where, for $i = 1, 2$, the posterior distribution of Υ_i is $\Upsilon_{\nu_i}^n$ with $\nu_i = n - p_i$. The maximized log-likelihood $\hat{\phi}_i$ satisfies $-2\hat{\phi}_i = n \log S_i^2 + \kappa_n$, where $\kappa_n = n \log(2\pi) - n \log n + n$. The Gamma representation described in section 5 may be employed when approximating the posterior means, quantiles, and c.d.f.'s of each Υ_i in (6.3). The DIC criterion in (1.3) satisfies, to a close approximation,

$$-2\phi_i^* = -2\hat{\phi}_i + n(\log n - \log \nu_i + \nu_i^{-1}) \quad , \tag{6.4}$$

and the Λ region in (2.2) consists of all points in Q_U falling on or below an upper boundary U satisfying

$$\beta_2 = \Upsilon_{\nu_2}^n\{T + \tau_{p_1}(\beta_1)\} \quad (0 \le \beta_1, \beta_2 < 1) \quad , \tag{6.5}$$

where, for $i = 1, 2$, τ_{p_i} is redefined to be the inverse of the c.d.f. $\Upsilon_{\nu_i}^n$ of a $\Upsilon_{\nu_i}^n$ variate.

With the recently redefined notation, *qsp* and *bsp* still solve the equations for β^* in (2.12) and (2.13). When M_1 is nested into M_2. the Type I success probability for a critical value T_o of T satisfies

$$\gamma^* = F^{\omega}_{\nu_2}(F_o) \quad , \tag{6.6}$$

where $F_o = \delta^{-1}\{\exp(T_o/n) - 1\}$ and $F^{\omega}_{\nu_2}$ is the c.d.f. of the corresponding F-distribution. Consequently, the matched significance probability *msp*, introduced in section 2, now satisfies the equation

$$n \log\{1 + \delta f^{\omega}_{\nu_2}(\beta^*)\} = \tau_{p_2}(\beta^*) - \tau_{p_1}(\beta^*) \tag{6.7}$$

in β^*, where $f^{\omega}_{\nu_2}$ is the inverse of the c.d.f. in (6.6). The concept of omega-preference, as introduced in Definition 3 of section 2, can therefore be applied to the current situation.

7. FURTHER BAYES-FREQUENCY CALCULATIONS

Under the formulation of section 6 consider the nested situation where ${\boldsymbol\theta_2}^T = ({\boldsymbol\theta_1}^T, {\boldsymbol\theta_*}^T)^T$ and ${\mathbf{X}_2}^T = ({\mathbf{X}_1}^T, {\mathbf{X}_*}^T)^T$ and $\boldsymbol\theta_*$ and $\mathbf{X}_*$ denote an $\omega \times 1$ vector and $n \times \omega$ matrix, with $\omega = p_2 - p_1$. Given $\boldsymbol\theta_2, \sigma_2^2$, and M_2, $(S_1^2 - S_2^2)/\sigma_2^2$ is a realization of a non-central chi-squared variate with ω degrees of freedom, and non-centrality parameter

$$\psi = \boldsymbol\theta_*^T \mathbf{W} \boldsymbol\theta_* / 2\sigma_2^2 \quad , \tag{7.1}$$

with

$$\mathbf{W} = \mathbf{X}_*^T(\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1)\mathbf{X}_* \quad , \tag{7.2}$$

where the algebraic arrangements may be confirmed upon equating the expectation of the corresponding quadratic form in $\mathbf{Y}$ when $\mathbf{Y}|\boldsymbol\theta_2, \sigma_2^2 \sim N[\mathbf{X}_2\boldsymbol\theta_2, \sigma_2^2]$ with the expectation $\omega + 2\psi$ of the non-central chi-squared distribution. The preceding partitioning of $\boldsymbol\theta_2$ and $\mathbf{X}_2$ gives, after some algebraic cancellation, the required result.

Assume further that, given σ_2^2 and M_2, $\boldsymbol\theta_2$ possesses the mixing distribution

$$\boldsymbol\theta_2|\sigma_2^2, M_2 \sim N[\boldsymbol\mu, \rho\sigma_2^2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}] \quad , \tag{7.3}$$

for some specified ρ, where the last ω elements of $\boldsymbol\mu$ are zero. Then

$$\boldsymbol\theta_*|\sigma_2^2, M_2 \sim N[\mathbf{0}_\omega, \rho\sigma_2^2\mathbf{W}^{-1}] \quad , \tag{7.4}$$

where $\mathbf{0}_\omega$ is the $\omega \times 1$ vector of zeros and $\mathbf{W}$ in (7.2) is by implication non-singular. Therefore $2\psi/\rho$ possesses, given σ_2^2 and M_2, a central χ^2_ω distribution. Standard moment generating

function arguments based upon the well-known Poisson mixture representation of the non-central chi-squared distribution may be used to show that given σ_2^2 and M_2, but unconditionally on $\boldsymbol{\theta}_2$, $(S_1^2 - S_2^2)/(1+\rho)\sigma_2^2$, follows a central χ_ω^2 distribution. Given σ_2^2 and M_2, S_2^2/σ_2^2 follows a central $\chi_{\nu_2}^2$ distribution, so that the random analogue $\tilde{F}$ of (6.2) satisfies

$$(1+\rho)^{-1}\tilde{F}|\sigma_2^2, M_2 \sim F_{\nu_2}^{\omega} \quad , \tag{7.5}$$

a central F-distribution. The preceding derivation simplifies the development by Johnson (2005, section 3), who then implicitly employs the point estimate $\hat{\rho} = \max(F-1, 0)$ for ρ in his estimated dimensionality-reduced Bayes factor. As $\hat{\rho}$ maximizes the marginal likelihood curve of ρ, given F and M_2, no account is taken of the dispersion of this curve. This dispersion can be substantial for small to moderate ω, even when ν_2 is large.

Johnson's estimated weight of evidence implicitly equals unity for all $F \leq 1$, thus placing equal emphasis on M_1 and M_2 in situations where M_1 should be preferred. As $\hat{\rho}$ is not marginally consistent given M_2 for ρ, as $\nu_2 \to \infty$ with ω fixed, it is difficult to interpret his estimated weight of evidence when $F > 1$. In particular, values of the estimated criterion which appear to substantively favor M_2 may instead be inconclusive. However, as the unestimated weight of evidence for M_2 against M_1 is proportional as a function of ρ to the marginal likelihood of ρ given F and M_2, the unestimated criterion may be valuably plotted against ρ, with very useful potential interpretations. As $\nu_2 \to \infty$, with ω fixed, this criterion approaches $(1+\rho)^{-\frac{1}{2}\omega}\exp\{\frac{1}{2}\omega\rho F/(1+\rho)\}$. Similarly appealing criteria are available in many other situations, if an appropriate test statistic, that is at least approximately ancillary given M_1 can be constructed.

As another consequence of (7.5), and paralleling the developments in section 4, we can set $\rho = \xi^2/med_\omega$ and consider the average Type II success probabilities

$$\kappa(\xi) = 1 - F_{\nu_2}^{\omega}[F_o/\{1 + (\xi^2/med_\omega)\}] \quad , \tag{7.6}$$

for critical values F_o of F. The minimum and average success probabilities in (4.3) and (4.4) may now be calculated with this choice of $\kappa(\xi)$ and upon replacing γ^* by the Type I success probability in (6.6). Under the assumption in (7.3) these calculations do not depend upon the choice of $\mathbf{X}_2$. The lambda-optimality criterion described in Definition 5 of section 4 may also be employed.

In Figs 5 and 6 the $MINSP$ and $AVESP$ curves are contrasted when $n = 100$, $p_1 = 7$, and $p_2 = 8$ and for (a) the MIC preference critical value $T_o = 1.088$ (b) the omega-preference critical value $T_o = 1.253$ and (c) the tau-dominance critical value $T_o = 6.647$ The slightly superior Bayes-frequency properties for omega-preference, when compared with

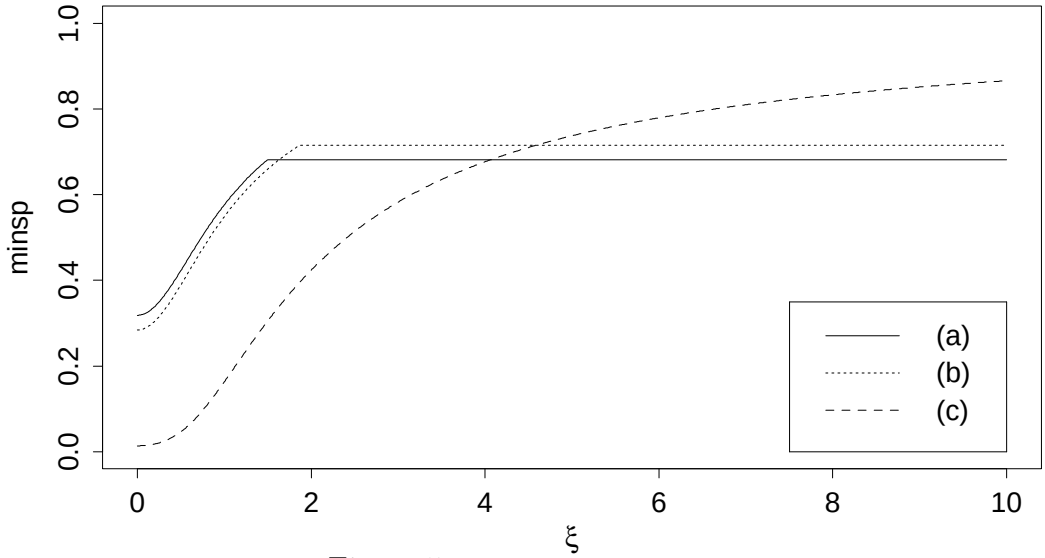

Figure 5: MINSP ($n = 100, p_1 = 7, p_2 = 8$)

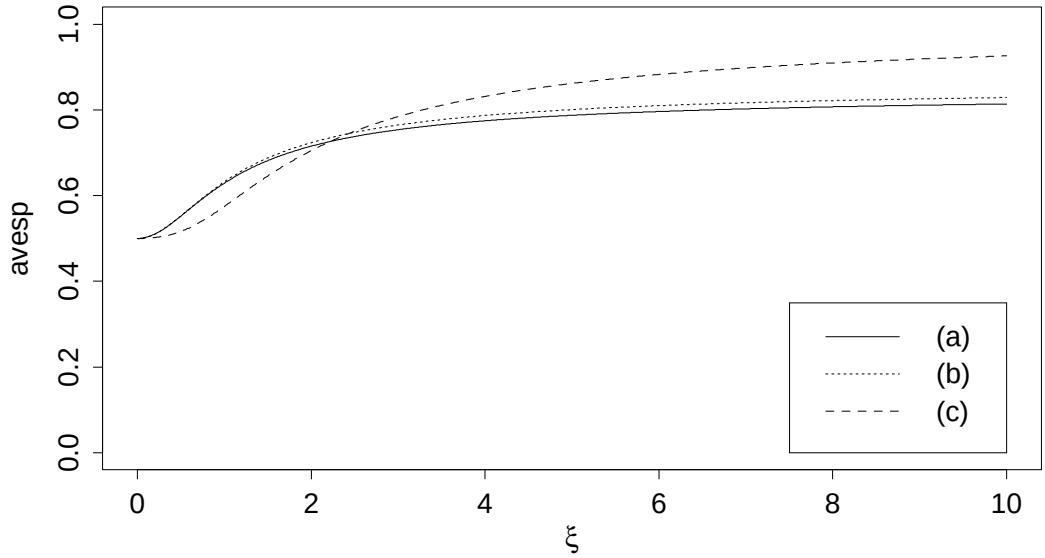

Figure 6: AVESP ($n = 100, p_1 = 7, p_2 = 8$)

MIC preference and as reported in section 4, extend to the current situation. As illustrated by the intersection points in Table 5, this conclusion is also true for general p_1 and p_2 unless n is too small. Similar Bayes-frequency calculations may be completed for the exact nested situations described in section 2. For the p_1 and p_2 of section 2 just apply the current numerical procedures with n very large, and with our current p_1 and p_2 reduced to $p_1 - 1$ and $p_2 - 1$. For the special case described in section 3, set the current p_1 and p_2 equal to zero and $n - 1$ respectively.

In Fig 7 the values of qsp, α_2, bsp, γ^* and λ are plotted against T_o in our current situation when $n = 100$, $p_1 = 7$ and $p_2 = 8$. While λ no longer calibrates well with qsp and α_2, this Bayes-frequency refutation criterion compromises between α_2 and γ^*. Graphs of $AVESP$

Table 5: Intersection Points of Success Probability Curves for Selected p_1 and p_2

n	p_1	p_2	(1)	(2)	(3)	(4)
20	7	8	(1.644,0.676)	(4.631,0.719)	(0.405,0.536)	(2.493,0.746)
60	7	8	(1.633,0.681)	(4.596,0.716)	(0.296,0.521)	(2.437,0.745)
100	7	8	(1.631,0.682)	(4.594,0.716)	(0.279,0.519)	(2.435,0.745)
100	5	9	(1.393,0.591)	(2.994,0.644)	(0.632,0.530)	(2.498,0.730)
100	6	7	(1.640,0.681)	(4.508,0.718)	(0.289,0.520)	(2.337,0.741)
100	6	8	(1.412,0.630)	(3.337,0.673)	(0.442,0.524)	(2.463,0.742)
100	5	6	(1.651,0.681)	(4.414,0.721)	(0.300,0.522)	(2.228,0.737)
100	5	7	(1.424,0.630)	(3.268,0.677)	(0.459,0.526)	(2.374,0.738)
100	5	8	(1.394,0.606)	(3.059,0.656)	(0.558,0.528)	(2.443,0.734)
100	5	9	(1.393,0.591)	(2.994,0.644)	(0.632,0.530)	(2.498,0.730)

N.B. The footnotes to Table 3 also apply to Table 5. The selected p_1 and p_2 values are relevant to the medical study in section 8.

and $MINSP$ may be calculated and contrasted by the user for specified out-off values for the various refutation criteria, as now illustrated.

The BIC model selection criterion, namely a $(p_2 - p_1) \log n$ critical value for T, is shown by Schwarz (1978), in generalizable special cases, to result from an asymptotic form ($n \to \infty$) of the logarithms of ordinary Bayes factors. However, BIC possesses unconvincing Bayes-frequency properties, when n is large enough, for any fixed p_1 and p_2, when compared with our Bayesian criteria, based upon the U curve. These properties result from the stabilization as n becomes large and for any fixed $T = T_o$ of the U curve and the corresponding $AVESP$ and $MINSP$ curves.

In contrast the BIC critical value becomes arbitrarily large. For example, when $p_1 = 2$ $p_2 = 6$ and $n = 10,000$, the BIC critical value of $T_o = 36.841$ compares with the $\alpha_2 = 0.99$ critical value of $T_o = 16.115$. The $\alpha_2 = 0.99$ criterion yields superior $AVESP$ whenever $\xi < 26.59$ and $AVESP < 0.9982$, and superior $MINSP$ whenever $\xi < 28.14$ and $MINSP < 0.9971$. Such computations arguably resolve Lindley's paradox, by decisively deciding the case against ordinary Bayes factors.

The posterior probability β_2 in (6.5) provides an alternative refutation criterion, for any specified choice of β_1. In particular, when $\beta_1 = 0.75$, $\beta_2 = \alpha_2^*$ denotes the posterior probability, given M_2, that ψ_2 exceeds the first posterior quantile of ψ_1, given M_1. When

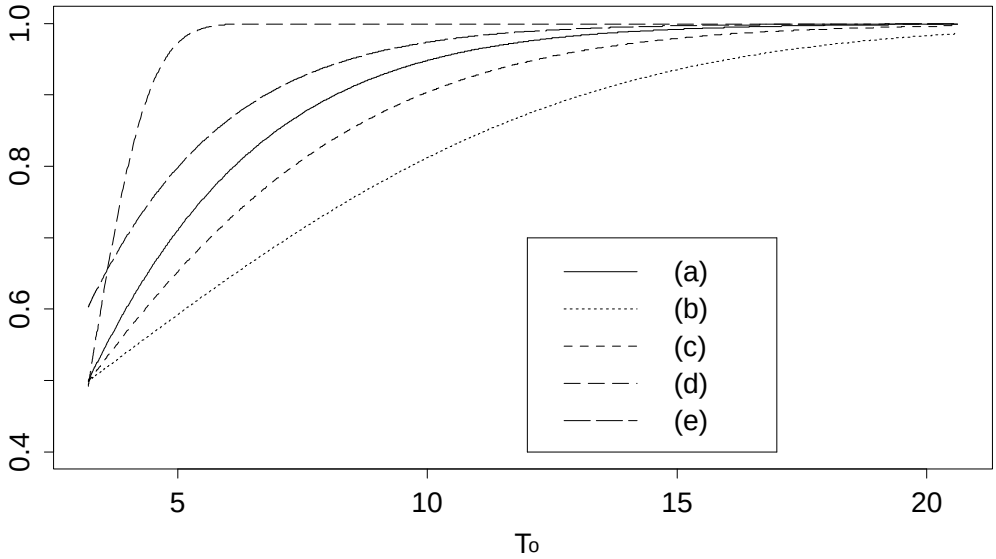

Figure 7: Bayesian and Bayes-Frequency Criteria ($n = 100, p_1 = 7, p_2 = 8$, (a) λ (b) qsp (c) α_2 (d) bsp (e) γ^*)

$\alpha_2^* \geq 0.75$, comparisons of α_2^* with the typically somewhat smaller Bayes-frequency criterion λ and with γ^* (see Table 7 of section 8) add credence to both α_2^* and λ as refutation criteria.

8. MEDICAL EXAMPLE

Feher *et al.* (1988), in a study of ongoing clinical relevance, consider the dependence of high-density lipoprotein (HDL) on eight explanatory variables, for a sample of $n = 71$ hypertensive diabetic patients. The HDL level is well-known to be negatively associated with blood pressure, high cholesterol and a risk of heart disease. The eight explanatory variables considered were:

B: the presence (+1) or absence (-1) of a beta-blocker,

D: consumption (+1) or non-consumption (-1) of alcohol,

S: smoking (+1) or non-smoking (-1),

A: age, W: weight, T: triglyceride level,

C: C-peptide level, and G: blood glucose level.

Logarithms were taken of each of the five continuous explanatory variables A, W, T, C, and G, before rescaling each variable, by subtracting the appropriate sample mean, and dividing by the corresponding sample standard deviation. Numerous regression models were considered for $Y = \text{logHDL}$, each including a constant term K and an error term represented symbolically by E. With A, W, T, C, and G now denoting the transformed variables, particular attention was paid to seventeen models, all special cases of the formulation introduced in section 5, and represented symbolically (+ denotes inclusion), as follows

M_1: $Y = K + B + T + E$,
M_2: $Y = K + B + T + S + A + E$,
M_3: $Y = M_2 + D + W + C + G$,
M_4: $Y = M_2 + G$,
M_5: $Y = M_2 + BS$,
M_6: $Y = M_2 + SA$,
M_7: $Y = M_2 + SA + BS$,
M_8: $Y = M_2 + SA + BS + TA$,
M_9: $Y = M_8 + TS$,
M_{10}: $Y = M_9 + BT$,
M_{11}: $Y = M_{10} + BA$,
M_{12}: $Y = M_8 + SAT$,
M_{13}: $Y = M_8 + SAB$,
M_{14}: $Y = M_8 + SAB + SAT$,
M_{15}: $Y = M_{14} + SBT$,
M_{16}: $Y = M_{15} + ABT$, and
M_{17}: $Y = M_7 + SAB$.

For example, M_{17} complements model M_7 with the triple interaction term SAB, or equivalently the four factor main effects model M_2 with two double interaction terms SA and BS, together with SAB. The choice of this set of models was partly motivated by consideration of standard regression output, and partly by the need to detect interaction effects of potential clinical relevance. The full eight factor main effects model M_3 yields respective (classical or Bayesian) two-sided p-values of 0.000, 0.000, 0.001, 0.009,, 0.523, 0.723, 0.747, and 0.815 for the presence of the T, A, S, G, D, E, and W main effects terms. This strongly suggests reduction of M_3 to M_2, as estimated by

$$Y = 1.900 - 0.449T - 0.245A - 0.304B - 0.179S + E \ , \tag{8.1}$$

with residual mean squared error 0.329.

The estimated coefficients in (8.1) are the corresponding posterior means, under the prior assumptions of section 5 and the reciprocal of the residual mean squared error is the posterior mean of the reciprocal of the residual variance. It is unnecessary to add squared terms for the measurement variables T and A as this complication would yield p-values 0.555 and 0.891 for the corresponding coefficients. The full double interaction model M_{11} that incorporates the four significant factors in M_2 yields respective p-values 0.044, 0.194, 0.240, 0.384, 0.414, and 0.824 for the presence of the interaction terms SA, BS, TA, TS, BT, and TA. This indicates the plausibility of M_6, which compounds M_2 with an interaction term SA that possesses estimated coefficient -0.155.

The preliminary model scan proposed in the last paragraph of section 3 is reported in Table 6, together with the residual sums of squares R_i^2 (see fourth column), for M_1 to M_{17}. Simple model preference, as suggested by the smallest values of both the DIC_i^* and $QIC2_i = -2MIC_i$, indicates the choice M_8. This three-fold double interaction model is MIC-preferred to any triple interaction model containing just multiples of T, A, B, and S,

Table 6: Preliminary Model Scan

M_i	ν_i	p_i	R_i^2	DIC_i^*	AIC_i^*	$QIC1_i$	$QIC2_i$	$QIC3_i$
M_1	68	3	0.514	122.65	127.57	121.54	123.02	125.10
M_2	66	5	0.631	105.18	111.98	103.58	105.56	108.16
M_3	62	9	0.636	108.65	119.01	106.26	109.07	112.52
M_4	65	6	0.634	105.77	113.48	103.96	106.16	108.99
M_5	65	6	0.641	104.23	111.95	102.42	104.62	107.45
M_6	65	6	0.653	101.81	109.53	100.00	102.20	105.03
M_7	64	7	0.665	100.40	109.02	98.39	100.80	103.85
M_8	63	8	0.675	99.51	109.01	97.31	99.92	103.18
M_9	62	9	0.679	99.77	110.13	97.38	100.19	103.65
M_{10}	61	10	0.683	100.03	111.23	97.46	100.46	104.11
M_{11}	60	11	0.683	101.14	113.17	98.41	101.59	105.43
M_{12}	62	9	0.679	99.71	110.07	97.32	100.13	103.59
M_{13}	62	9	0.677	100.08	110.44	97.69	100.50	103.95
M_{14}	61	10	0.684	99.78	110.98	97.21	100.21	103.86
M_{15}	60	11	0.686	100.58	112.61	97.85	101.03	104.87
M_{16}	59	12	0.686	101.73	114.57	98.82	102.18	106.21
M_{17}	63	8	0.671	100.36	109.85	98.15	100.77	104.02

for example M_{12} to M_{17}. Moreover $QIC3_8 < QIC1_2$, so that M_8 tau-dominates its main effects sub-model M_2. The MIC-preferred choice M_8 may be estimated by

$$Y = 1.900 - 0.479T - 0.167A - 0.324B - 0.167S - 0.155SA - 0.114BS - 0.105TA + E, \quad (8.2)$$

with residual mean squared error 0.238.

The p-value for the presence of the four main effects and three interaction terms in M_8 are respectively 0.000, 0.000, 0.000, 0.016, 0.0419, 0.101, and 0.184. The inclusion of TA is further open to question as M_8 is only marginally MIC and DIC preferred to M_7. However, M_8 is also omega-preferred to any of its sub-models, as illustrated below, thus more strongly supporting the MIC and DIC choice. While this does not necessarily imply that the BS and TA interactions are of clinical significance, they should certainly be considered in any further study, together with the more statistically significant SA interaction.

Ten nested models comparisons are analyzed in detail in Table 7. The entries in the fifth column describe the values α_0 of the posterior probability β_2 in (6.5) when $\beta_1 = 0$, while the

Table 7: Detailed Nested Model Comparisons

Comparison	p_1	p_2	T	F	α_0	qsp	α_2	α_2^*	bsp	λ	γ^*	msp
M_2 vs M_3	5	9	0.968	0.213	0.000	0.331	0.192	0.420	0.000	0.500	0.070	0.644
M_7 vs M_8	7	8	2.008	1.807	0.007	0.542	0.578	0.788	0.997	0.653	0.816	0.716
M_5 vs M_7	6	7	4.927	4.599	0.206	0.688	0.795	0.908	1.000	0.891	0.964	0.718
M_5 vs M_8	6	8	6.935	3.232	0.310	0.719	0.825	0.919	1.000	0.904	0.954	0.673
M_6 vs M_7	6	7	2.508	2.301	0.032	0.572	0.628	0.816	1.000	0.720	0.866	0.718
M_6 vs M_8	6	8	4.515	2.068	0.104	0.611	0.686	0.843	1.000	0.772	0.865	0.673
M_2 vs M_6	5	6	4.450	4.205	0.247	0.682	0.784	0.900	1.000	0.872	0.956	0.721
M_2 vs M_7	5	7	6.958	3.295	0.416	0.738	0.841	0.925	1.000	0.908	0.957	0.677
M_2 vs M_8	5	8	8.966	2.827	0.505	0.766	0.863	0.934	1.000	0.918	0.954	0.656
M_7 vs M_{17}	7	8	1.164	1.041	0.001	0.502	0.503	0.740	0.546	0.510	0.688	0.716

N.B. The entries to the second and the third columns now denote the respective numbers of linear parameters for the models in that comparison.

entries α_2^* in the eighth column ($\beta_1 = 0.75$) were introduced in the last paragraph of section 7. The values of msp in the final column do not depend on the observed values of F. Some Bayes-frequency calculations relevant to these comparisons were summarized in Figs 5, 6, and 7, and Table 5.

Consider firstly the comparison of M_7 and M_8. As $\gamma^* = 0.816 > msp = 0.716$, M_8 is omega-preferred to M_7. Similarly M_8 is omega-preferred to its submodels M_2, M_5, and M_6. Any more complex model such as M_{13} which contains M_8 but is not MIC-preferred to M_8, cannot be omega-preferred to M_8. As M_{17} is MIC preferred to M_7, $\alpha_2 = 0.503$ must exceed 0.5. However, as $\gamma^* = 0.688 < msp = 0.716$, M_{17} is not omega-preferred to M_7, It is nevertheless important to directly contrast the non-nested models M_8 and M_{17} in detail. While frequency criteria are unavailable, treating M_{17} as the first model and M_8 as the second model gives $T = 0.844$, $\alpha_0 = 0.000$, $qsp = 0.539$, $\alpha_2 = 0.575$, and $\alpha_2^* = 0.797$. Switching the order of the models gives $\alpha_0 = 0.000$, $qsp = 0.461$, $\alpha_2 = 0.420$, and $\alpha_2^* = 0.696$. Model M_8 is noticeably preferable.

Quite different conclusions should be drawn from refutation inference (even though M_8 yields the smallest AIC^* value in Table 6). For example, none of the submodels M_5, M_6, and M_7 are tau-dominated by M_8, and each yields a qsp value substantially smaller than

0.75. The key comparisons concern M_2 with M_7 and M_2 with M_8. The respective γ^* values 0.957 and 0.954 suggest refutation of M_2 at the 5% level. However the α_2^* values 0.925 and 0.934 correspond to slightly unconvincing points when $\beta_1 = 0.75$ on the respective U-curves.

With also $\alpha_0 = 0.416$, $qsp = 0.738 < 0.75$, and $\alpha_2 = 0.841$, the U curve for the M_2 versus M_7 comparison, will not be remarkably well directed towards the upper left $(\beta_1, \beta_2) = (0, 1)$ vertex of Q_U. With $\alpha_0 = 0.505$, $qsp = 0.766$, and $\alpha_2 = 0.863$ for the M_2 versus M_8 comparison, the U curve will be only slightly more directed towards this vertex. The Bayesian inferences suggest that there is insufficient information in the current data to justify refuting M_2 against either M_7 or M_8 The Bayesian case for refuting M_2 against the SA interaction model M_6 is even less convincing. The values of λ for the comparisons of M_2 with M_6, M_7, and M_8 contrast with the γ^*, and more clearly distinguish between these three alternative models, in a manner consistent with the fully Bayesian inferences. After consideration of a number of further models, we conclude that the four factor main effects model M_2, estimated in (8.1) is an appropriate parameter-parsimonious model choice. Residual analyses would further validate and contrast the models in (8.1) and (8.2).

9. FUTURE WORK

For any specified value of β_1 sufficiently exceeding 0.5 the Bayesian refutation criterion in (6.5) matches the frequency probability γ^* in (6.5) whenever

$$\tau_{p_2}\{F^{\omega}_{\nu_2}(F_0)\} = n \log\{1 + \delta F_0\} + \tau_{p_1}(\beta_1) \quad . \tag{9.1}$$

If F_0 satisfies (9.1), then F_0 provides a possible critical value for refutation inference, with Type I frequency success probability $msp^* = F^{\omega}_{\nu_2}(F_0)$ depending upon p_1 and p_2. The solution for F_0 may then be interpreted as uniquely satisfying the property that M_1 should be refuted in favor of M_2 if and only if both γ^* and β_2 exceed msp^*. This property may be clarified upon comparing the graphs of γ^* and β_2 against F_0. When $\beta_1 = 0.75$ the ten model comparisons in Table 7 yield the respective msp^* values 0.999, 0.964, 0.966, 0.988, 0.966, 0.988, 0.969, 0.990, 0.997, and 0.964. Very detailed Bayes-frequency investigations of this possibility are however required. Possible extensions of the close Bayes-frequency calibrations in Table 4 to more general situations, by comparisons of various Bayes profiles and generalizations of λ, can also be investigated.

When M_1 and M_2 are non-nested, tau-dominance, and specified cut-off values of qsp, the α profile, and other points on the U-boundary are open to scrutiny via large-scale simulations of $\mathbf{y}$ given M_1 and M_2, for selected choices of the parameter values. Given observed data, the frequency properties may be estimated using bootstrap techniques. The authors are currently

investigating Bayesian inferences for the comparison of linear and non-linear random effects models with their fixed effects counterparts, and this work will be reported elsewhere. The investigation of zero associations in contingency tables, the semi-parametric smoothing of regression functions, and the identification of wavelets, provide just a few further examples of application. With the possible exception of the matching procedures, all of the currently suggested ideas may be applied exactly to most situations, for example using MCMC.

In section 7 we showed that some standard measures of evidence, namely the logarithms of ordinary Bayes factors, can possess quite disappointing limiting Bayes-frequency properties, for example when contrasted with our recommended Bayesian criteria. This is a key conclusion with broad social implications. These aspects should also be emphasised by other authors and applied workers.

REFERENCES

Aitkin, M. (1991), "Posterior Bayes factors" (with discussion), *Journal of the Royal Statistical Society*, B 53, 111-142.

Aitkin, M. (1997), "The calibration of P-values, posterior Bayes factors and the *AIC* from the posterior distribution of the likelihood," *Statistics and Computing*, 7, 253-261.

Akaike, H. (1978), "A Bayesian analysis of the minimum AIC procedure," *Annals of the Institute of Statistical Mathematics*, 30(A), 9-14.

Barbieri, M.M. and Berger, J.O. (2004), "Optimal predictive model selection," *The Annals of Statistics*, 32, 870-897.

Berger, J.O. and Guglielmi, A. (2001), "Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives," *Journal of the American Statistical Association*, 96, 174-184.

Berger, J.O. and Pericchi, L.R. (2004), "Training samples in objective Bayesian model selection," *The Annals of Statistics*, 32, 841-869.

Dempster, A.P. (1974) "The direct use of likelihood for significance testing," in *Proceedings of the Conference on the Foundational Questions of Statistical Inference*, eds. O. Barndorff-Nielsen, P. Blaesild, and G. Schou, University of Aarhus, Demark, pp. 335-352.

Dobson, A.J. (1990), *An Introduction to Generalized Linear Models.* New York: Chapman & Hall.

Feher, M.D., Rains, S.G.H., Richmond, W., Torrend, D., Wilson, G., Wadsworth, J., Sever, P.S., and Elkeles, R.S. (1988), "β-blockers, Lipoproteins, and noninsulin-dependent diabetes," *Postgraduate Medical Journal*, 64, 926-930.

Good, I.J. (1991), "Weight of evidence and the likelihood ratio," in *The Use of Statistics in Forensic Science*, eds. C.G.G. Aitken and D.A. Stoney, Ellis Horwood, Chichester, 85-106.

Han, C. and Carlin, B.P. (2001), "Markov chain Monte Carlo methods for computing Bayes factors: A comparative review," *Journal of the American Statistical Association*, 96, 1122-1132.

Johnson, V.E. (2005), "Bayes factors based on test statistics," *Journal of the Royal Statistical Society*, B 67, 689-701.

Lee, J. and Berger, J.O. (2001), "Semiparametric Bayesian analysis of selection models," *Journal of the American Statistical Association*, 96, 1397-1409.

Leonard, T. and Hsu, J.S.J. (1999), *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. New York: Cambridge University Press.

Lindley, D.V. (1957), "A statistical paradox," *Biometrika*, 44, 187-192.

O'Hagan, A. (1995), "Fractional Bayes factors for model comparisons" (with discussion), *Journal of the Royal Statistical Society*, B 57, 99-138.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

Sellke, T. Bayarri, M.J., and Berger, J. O. (2001), "Calibration of p-values for testing precise null hypotheses," *American Statistician*, 55, 62-71.

Shafer, G. (1982), "Lindley's paradox" (with discussion), *Journal of the American Statistical Association*, 77, 325-351.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit" (with discussion), *Journal of the Royal Statistical Society*, B 64, 583-639.

van der Linde, A. (2005), "DIC in variable selection," *Statistica Neerlandica*, 59. 45-56.