Bayes Factors and Marginal Distributions in Invariant Situations

James O. Berger, Luis R. Pericchi and Julia A. Varshavsky
Department of Statistics, Purdue University, USA
Departamento de Cómputo Científico y Estadística,
Universidad Simón Bolívar, Venezuela
Lilly Research Laboratories, Indiana, USA

Abstract

In Bayesian analysis with a "minimal" data set and common noninformative priors, the (formal) marginal density of the data is surprisingly often independent of the error distribution. This results in great simplifications in certain model selection methodologies; for instance, the Intrinsic Bayes Factor for models with this property reduces simply to the Bayes factor with respect to the noninformative priors. The basic result holds for comparison of models which are invariant with respect to the same group structure. Indeed the condition reduces to a condition on the distributions of the common maximal invariant. In these situations, the marginal density of a "minimal" data set is typically available in closed form, regardless of the error distribution. This provides very useful expressions for computation of Intrinsic Bayes Factors in more general settings. The conditions for the results to hold are explored in some detail for nonnormal linear models and various transformations thereof.

Key Words: Intrinsic Bayes Factor, Haar Measure, Group Invariance, Reference Prior, Maximal Invariant, Marginal Density

1 Introduction

We will freely use the language of group theory, referring to Eaton (1989) for definitions. Note, however, that the key examples can be understood without use of group theory; indeed, to broaden accessibility, the paper has purposefully been written with essentially the minimally needed group theory.

Suppose that a locally compact group G acts properly on the observational space $\mathcal{Y} \subseteq \mathbf{R}^n$. Consider the problem of choosing between two families of densities

$$M_1 : \{f_1(y|\theta_1), \theta_1 \in \Theta_1\} \text{ and } M_2 : \{f_2(y|\theta_2), \theta_2 \in \Theta_2\}, \tag{1}$$

where all densities are with respect to μ, a given relatively invariant Radon measure on $\mathcal{Y}$. Assume that Θ_1 and Θ_2 are isomorphic to G and that G acts transitively on the Θ_j in such a way that both M_1 and M_2 are invariant with respect to G (i.e. $P_j(gY \in A|\theta_j) = P_j(Y \in A|g\theta_j),\ A \in \mathcal{Y}$, $\theta_j \in \Theta_j$). Let $\mu^\star$ be a relatively invariant measure on G (and, hence, on the Θ_j).

Example 1.1 Suppose $y = (y_1, \ldots, y_n)$, where the y_i's are *i.i.d.* from either the density (on $\mathbf{R}^1$ and with respect to Lebesgue measure) $\sigma_1^{-1} p_1((y_i - \beta_1)/\sigma_1)$ or $\sigma_2^{-1} p_2((y_i - \beta_2)/\sigma_2)$. Thus, for $j = 1, 2,\ \theta_j = (\beta_j, \sigma_j)$ and

$$f_j(y|\theta_j) = \frac{1}{\sigma_j^n} \prod_{i=1}^{n} p_j \left(\frac{y_i - \beta_j}{\sigma_j} \right). \tag{2}$$

These are invariant with respect to the usual location-scale group $G = \{g_{b,c} : -\infty < b < \infty,\ c > 0\}$, where the group actions on Y and Θ_j are defined by

$$g_{b,c} y = cy + b \cdot \mathbf{1} \ \text{ and } \ g_{b,c} \theta_j = (c\beta_j + b, c\sigma_j), \tag{3}$$

where $\mathbf{1} \in \mathbf{R}^n$ denotes a vector of 1's. Here μ is Lebesgue measure on $\mathbf{R}^n$, and $\mu^\star$ is Lebesgue measure on $(-\infty, \infty) \times (0, \infty)$.

To decide between M_1 and M_2, the standard Bayesian approach is to choose proper prior densities $\pi_j(\theta_j)$ (with respect to $\mu^\star$, say) $j = 1, 2$, and determine the Bayes factor of M_1 to M_2,

$$B_{12} = m_1(y)/m_2(y),$$

where

$$m_j(y) = \int_{\Theta_j} f_j(y|\theta_j) \pi_j(\theta_j)\, \mu^\star(d\theta_j). \tag{4}$$

Improper prior densities cannot typically be used in computation of Bayes factors, because the m_j(y) are then determined only up to arbitrary multiplicative constants. This has been a major barrier to the development of general default Bayesian model selection methodology.

Recently, Berger and Pericchi (1996a, b) proposed the *intrinsic Bayes factor* approach to overcoming this difficulty. The approach is based on using part of the data as a "training sample" to convert default improper priors into proper distributions. Precursors or variants on such use of training samples can be found in Lempers (1971), Atkinson (1978), Geisser and Eddy (1979), Smith and Spiegelhalter (1982), and Gelfand, Dey and Chang (1992).

To define an *intrinsic Bayes factor*, one begins by choosing default prior densities for the θ_j. In this paper we will choose $\pi(\theta_j)$ to be $\nu_r(\theta_j)$, where ν_r is a right Haar density (with respect to $\mu^\star$) corresponding to the group action on Θ_j. Next, one defines a *minimal training sample* (MTS) to be a subset, $y(l)$, of the full data vector y such that $0 < m_j^r(y(l)) < \infty$, for $j = 1, 2$, and $m_j^r(y^\star) = \infty$ for either $j = 1$ or $j = 2$ when $y^\star$ is a subset of $y(l)$; here

$$m_j^r(y(l)) = \int_{\Theta_j} f_j(y(l)|\theta_j) \nu_r(\theta_j)\, \mu^\star(d\theta_j). \tag{5}$$

Thus an MTS is a subset of the data for which the posteriors corresponding to $\nu_r(\theta_j)$ would be proper, but for which any smaller set of data would not yield a proper posterior for at least one of the models. An *intrinsic Bayes factor* is then defined as

$$B_{12}^I = \frac{m_1^r(y)}{m_2^r(y)} \cdot \mathrm{Av} \left(\frac{m_2^r(y(l))}{m_1^r(y(l))} \right), \tag{6}$$

where "Av" refers to some type of average of the $m_2^r(y(l))/m_1^r(y(l))$, over all possible choices of minimal training samples $y(l)$. Common choices of "Av" are the arithmetic average, geometric average, and the median. For discussions of the appealing properties of B^I_{12}, see Berger and Pericchi (1996a).

Example 1.1 (continued). A right Haar density for a location-scale problem is $\nu_r((\beta,\sigma)) = 1/\sigma$. For a single observation, y_i, it is clear that

$$m^r_j(y_i) = \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma_j^2} p_j(\frac{y_i-\beta_j}{\sigma_j}) d\beta_j d\sigma_j = \infty.$$

For two distinct observations $y_i \neq y_k$, computation shows that

$$m^r_j((y_i,y_k)) = \int_0^\infty \int_{-\infty}^\infty \frac{1}{\sigma_j^3} p_j(\frac{y_i-\beta_j}{\sigma_j}) p_j(\frac{y_k-\beta_j}{\sigma_j}) d\beta_j d\sigma_j = \frac{1}{2|y_i-y_k|}, \tag{7}$$

regardless of p_j. Thus an MTS is any pair of distinct observations. Furthermore, (7) implies that, for any MTS $y(l)=(y_i,y_k)$, $m^r_2(y(l))/m^r_1(y(l)) = 1$. It is immediate that

$$B^I_{12} = m^r_1(y)/m^r_2(y). \tag{8}$$

In other words, the intrinsic Bayes factor corresponds to utilizing $\nu_r(\theta_j)$ directly as the prior distribution for both M_1 and M_2 and the full data. The training samples become irrelevant.

Equations (7) and (8) provided the original motivation for this paper. Computation of B^I_{12} can be onerous if averaging over all training samples is required, and it is not always clear which "average" should be chosen. But when $m^r_2(y(l))/m^r_1(y(l)) = 1$ for all training samples, these concerns become irrelevant and (8) becomes a compelling choice as a default Bayes factor. This motivated our search for general conditions under which (8) would hold.

Equation (7) is itself quite a curiosity, because of the lack of dependence of the answer upon p_j. (The equation can be established directly by changing variables to $v=(y_i-\beta_j)/\sigma_j$ and $w=(y_k-\beta_j)/\sigma_j$, and using exchangeability of V and W.) This equation is generalized in sections 4 and 5 to broad classes of linear and transformed linear models. These generalizations are of considerable importance in their own right, as they allow closed form computation of the $m^r_2(y(l))/m^r_1(y(l))$ for a wide variety of situations in which Θ_1 and Θ_2 are not isomorphic. This can be crucial to easy implementation of intrinsic Bayes factor methods, as numerical integration of a large number of $m_j(y(l))$ is otherwise needed. Note that the expressions obtained in sections 4 and 5 formed the basis for the methodology developed for normal linear models in Berger and Pericchi (1996b) and Varshavsky (1996).

Section 2 states a general theorem which forms the basis for the determination of when simplification (8) results. This is applied to the general linear model in section 3, and to ANOVA type problems in particular. Section 5 considers various transformed linear models, and presents an application of the ideas to a common model selection problem in reliability and survival analysis.

2 The Bayes Factor Under Right Haar Measure

For the situation in section 1, let $T(y)$ be a maximal invariant function acting on $\mathcal{Y}$, with respect to G. (Again, see Eaton (1989) for definitions.) Let $f^\star_j(\cdot)$ denote the density of $T(y)$ under Model j, and with respect to the measure μ_T induced by μ.

Example 1.1 (continued). One possible choice of a maximal invariant is

$$T(y) = \left(\frac{y_1 - y_n}{|y_{n-1} - y_n|}, \frac{y_2 - y_n}{|y_{n-1} - y_n|}, \ldots, \frac{y_{n-1} - y_n}{|y_{n-1} - y_n|} \right). \tag{9}$$

Note that this takes values in the space $\mathbf{R}^{(n-2)} \times \{-1, 1\}$, and μ_T is counting measure on $\{-1,1\}$ and Lebesgue measure on $\mathbf{R}^{(n-2)}$. For the special case $n = 2$ (corresponding to y being a minimal training sample), $T(y)$ is either -1 or 1. Furthermore, because of exchangeability of y_1 and y_2, it is then clear that, for either $j = 1$ or 2,

$$f_j^\star(-1) = f_j^\star(1) = 1/2. \tag{10}$$

For $n > 2$, $f_j^\star$ will typically vary with j. This means that *only* a minimal training sample will typically yield the simplification resulting in (8), providing a rather compelling reason to restrict attention to training samples that are minimal in such situations.

Theorem 2.1 *For the situation described above and when a right Haar density, $\nu_r(\cdot)$, is used as the prior density for both M_1 and M_2, the Bayes factor can be represented as*

$$B_{12} = \frac{f_1^\star(T(y))}{f_2^\star(T(y))}. \tag{11}$$

Proof: Since the base measure, μ, on $\mathcal{Y}$ is relatively invariant, it has a multiplier $\chi(\cdot)$. The Wijsman representation theorem (Wijsman (1990)) gives the following expression for the ratio of densities of the maximal invariants:

$$\frac{\int_G f_2(gy|\theta_2)\chi(g)\nu_l(dg)}{\int_G f_1(gy|\theta_1)\chi(g)\nu_l(dg)}, \tag{12}$$

where $\nu_l(\cdot)$ denotes a left invariant Haar measure on G. Hence, to prove the theorem, it suffices to show that

$$\int_G f_j(gy|\theta_j)\chi(g)\nu_l(dg) = \int_{\Theta_j} f_j(y|g\theta_j)\nu_r(d\theta_j). \tag{13}$$

Since the models (1) are invariant under G,

$$f_j(y|\theta_j) = f_j(gy|g\theta_j)\chi(g), \quad \forall y \in Y, g \in G, \theta_j \in \Theta_j. \tag{14}$$

Using (14) and the fact that $\nu_l(dg^{-1}) = \nu_r(dg)$ yields

$$\begin{aligned} \int_G f_j(gy|\theta_j)\chi(g)\nu_l(dg) &= \int_G f_j(y|g^{-1}\theta)\chi(g)\nu_l(dg) \\ &= \int_G f_j(y|g\theta)\nu_l(dg^{-1}) \\ &= \int_G f_j(y|g\theta)\nu_r(dg). \end{aligned}$$

Because G is isomorphic to Θ_j and ν_r is right invariant, (13), and hence the theorem follow immediately. $\square$

Example 1.1 (continued). From the earlier discussions, (10), and the above theorem, it is clear that, for any MTS $y(l)$,

$$B_{12}(y(l)) = \frac{1/2}{1/2} = 1.$$

This provides another route to (8). Note, further, that Theorem 2.1 shows that

$$B_{12}^{I} = f_1^{\star}(T(y))/f_2^{\star}(T(y)).$$

This may be useful for computation in some situations.

3 Application to the Linear Model

3.1 The Group Structure

Suppose the data $y = (y_1, \ldots, y_n)^t$ arises from a linear model

$$y = X\beta + \sigma\epsilon \ , \tag{15}$$

where $\epsilon=(\epsilon_1, \ldots, \epsilon_n)^t$ is a random vector having the density $f(u)$, $u \in \mathbf{R}^n$, X is an $n \times (m-1)$ known design matrix of full rank and $\beta = (\beta_1, \ldots, \beta_{m-1})^t$ and $\sigma > 0$ are unknown parameters.

Suppose that we are interested in choosing between two densities for the error ϵ,

$$M_1 : \epsilon \sim f_1(\cdot) \ \text{ and } \ M_2 : \epsilon \sim f_2(\cdot). \tag{16}$$

This is an important problem in model validation as well as model selection. The group of transformations that leaves these models invariant is

$$G = \{g_{b,c} : \ b = (b_1, \ldots, b_{m-1}), b \in \mathbf{R}^{m-1}, \ c > 0\} \tag{17}$$

that acts on $\mathcal{Y}$ via the group action

$$g_{b,c}(y) = cy + Xb \tag{18}$$

and acts on the parameter space $\Theta = \{(\beta, \sigma), \ \beta \in \mathbf{R}^{m-1}, \ \sigma > 0\}$, via

$$g_{b,c}(\beta, \sigma) = (c\beta + b, c\sigma).$$

Denote the matrix consisting of the first $n-m+1$ rows of X as X_1, and the matrix consisting of the last $m-1$ rows as X_2. Assume (without loss of generality) that the rank of X_2 equals $m-1$. Let $x_{(i)}$ be the ith row of X and $a_{(i)} = (a_{i,n-m+2}, \ldots, a_{i,n}) = x_{(i)}X_2^{-1}$. Then the following is true.

Proposition 3.1 *A maximal invariant under G is*

$$T(y) = \left(\frac{y_1 - \sum_{j=n-m+2}^{n} a_{1,j} \cdot y_j}{|y_{n-m+1} - \sum_{j=n-m+2}^{n} a_{n-m+1,j} \cdot y_j|}, \ \ldots, \ \frac{y_{n-m} - \sum_{j=n-m+2}^{n} a_{n-m,j} \cdot y_j}{|y_{n-m+1} - \sum_{j=n-m+2}^{n} a_{n-m+1,j} \cdot y_j|}, \right.$$
$$\left. \operatorname{sign}(y_{n-m+1} - \sum_{j=n-m+2}^{n} a_{n-m+1,j} \cdot y_j) \right). \tag{19}$$

Proof: See the Appendix.

3.2 Bayes Factors for Minimal Training Samples

Let us return to the question of computing the Bayes factor for a minimal training sample. A minimal training sample for the model (15), under the prior $\pi(\beta, \sigma) = 1/\sigma$ (which is equivalent to the right Haar density), consists of m observations $y(l) = (y_1(l), \ldots, y_m(l))^t$ such that the matrix $(X(l)\ y(l))$ is of rank m, where $X(l)$ is the respective $m \times (m-1)$ submatrix of the design matrix X. The maximal invariant in (19) is then

$$T(y(l)) = \text{sign}(y_1(l) - \sum_{j=2}^{m} a_j y_j(l)), \tag{20}$$

where $a = (a_2, \ldots, a_m) = x_1(l) X_2(l)^{-1}$. Note that this assumes only the values 1 and -1, corresponding to the sets $y_1(l) - \sum_{j=2}^{m} a_j y_j(l) > 0$ and $y_1(l) - \sum_{j=2}^{m} a_j y_j(l) < 0$, respectively. Thus

$$f_j^\star(1) = P_{f_j}(Y_1(l) - \sum_{j=2}^{m} a_j Y_j(l) > 0) = 1 - f_j^\star(0),$$

where P_f denotes probability with respect to the density f.

Theorem 3.1 *Suppose that it is desired to compare M_1 and M_2 using the prior density $\pi(\beta, \sigma) = 1/\sigma$ (which is a right Haar density) for each model. If a minimal training sample, $y(l)$, results in a maximal invariant in (20) which satisfies*

$$f_1^\star(1) = f_2^\star(1), \tag{21}$$

then the Bayes factor based on the minimal training sample is equal to one.

Proof: First note that the group G defined in (17) - (18) is a proper action on $\mathbf{R}^k, k \geq m$. The conclusion is immediate from Theorem 2.1.

□

The problem thus reduces to the question of when (21) holds. Let $|A|$ denote the determinant of a matrix A, $|A|^+$ the absolute value of $|A|$, and $\|a\|$ the Euclidean length of the vector a. Finally, define the set

$$H(l) = \{v \in \mathbf{R}^m : |(X(l)\ v)| = 0\}. \tag{22}$$

This is a hyperplane through the origin which splits the space $\mathbf{R}^m$ into

$$\Omega(l) = \{v \in \mathbf{R}^m : |(X(l)\ y(l))| / |(X(l)\ v)| > 0\} \quad \text{and} \quad \{\Omega(l)\}^c. \tag{23}$$

The following is true.

Theorem 3.2 *Equation (21) is satisfied if*

$$\int_{\Omega(l)} f_j(v)\, dv = \int_{\Omega(l)^c} f_j(v)\, dv. \tag{24}$$

Proof: Suppose, without loss of generality, that $|(X(l)\ y)| > 0$. Making the change of variables $v = (u - X(l)\beta)/\sigma,\ u = (u_1, \ldots, u_m) = (u_1, U_2)$, say, we get

$$\int_{\Omega(l)} f_j(v)dv \ = \ \int_{|(X(l)\ v)|>0} f_j(v)dv \ = \ \int_{\Omega(l)^*} (1/\sigma)^m \cdot f_j((u - X(l)\beta)/\sigma)\, du, \tag{25}$$

where

$$\Omega(l)^* = \ \{u : |(X(l)\ (u - X(l)\beta)/\sigma)| > 0\} \ = \ \{u : |(X(l)\ u)|/\sigma > 0\} \ = \ \{u : |(X(l)\ u)| > 0\}.$$

Now note that

$$\begin{aligned} |(X(l)\ u)| \ &= \ \left|\begin{pmatrix} x_1(l) & u_1 \\ X_2(l) & U_2 \end{pmatrix}\right| = - \left|\begin{pmatrix} X_2(l) & U_2 \\ x_1(l) & u_1 \end{pmatrix}\right| \\ &= \ -|X_2(l)| \cdot |u_1 - x_1(l) X_2(l)^{-1} U_2| \ = \ -|X_2(l)| \cdot (u_1 - \sum_{j=2}^{m} a_j u_j). \end{aligned}$$

Hence,

$$\int_{\Omega(l)} f_j(v)\, dv = \begin{cases} f_j^\star(-1) & \text{if } |X_2(l)| > 0 \\ f_j^\star(1) & \text{otherwise} \end{cases} .$$

Repeating the argument for $\int_{\Omega(l)^c} f_j(v)\, dv$, we get that it is equal to $f_j^\star(1)$, in the case $|X_2(l)| > 0$, and $f_j^\star(-1)$ otherwise. Hence equation (24) implies that $f_j^\star(1) = f_j^\star(-1) = 1/2$. This immediately yields (21).

$\square$

Remark 3.1 Condition (24) will clearly be satisfied if $f(v), v \in \mathbf{R}^m$, is symmetric about the origin, i.e. if

$$f(v) = f(-v). \tag{26}$$

Indeed, since $|(X(l)\ (-v))| = -|(X(l)\ v)|$ and $f(v) = f(-v)$, we then have

$$\int_{\Omega(l)} f(v)\, dv = \int_{\Omega(l)^c} f(v)\, dv \ .$$

While the symmetry of $f(v)$ about the origin is, of course, a much stronger condition than (24), it is extremely convenient for practical purposes for two reasons. First of all, it is often easy to check. For instance, it follows immediately that the theorem holds for models with *iid* errors coming from a symmetric distribution. It also holds, for example, for the non-*iid* case of $\epsilon \sim N(0, \Sigma)$. Secondly, it is convenient for use in computing intrinsic Bayes factors because (24) then holds for *any* minimal training sample, regardless of the associated $X(l)$. Indeed, it then follows directly that B_{12}^I is as in (8), which here can be written

$$B_{12}^I = \frac{\int\int f_1((y - X\beta)/\sigma)\sigma^{-(n+1)}\, d\beta\, d\sigma}{\int\int f_2((y - X\beta)/\sigma)\sigma^{-(n+1)}\, d\beta\, d\sigma}. \tag{27}$$

An important special case of a design matrix for which (24) holds is a design matrix that has two identical rows. We begin with a simple example.

Example 3.1 Consider the location-scale scenario discussed in Example 1.1. This is a special case of model (15) with $m = 2$, $X = (1, \ldots, 1)^t$, and $\beta = \mu$, and minimal training samples, $y(l)$, consist of any pair of distinct observations: $y(l) = (y_1(l), y_2(l))$. The design matrices, $X(l)$, are the same for all l: $X(l) = (1,1)^t$. Then $\{v \in \mathbf{R}^2 : |(X(l)\ v)| = 0\}$ becomes $\{(v_1, v_2) : v_1 - v_2 = 0\}$. Condition (24) will clearly be satisfied if $f_j(v_1, v_2)$ is symmetric about the line $v_1 = v_2$, i.e. if $f_j(v_1, v_2) = f_j(v_2, v_1)$; this holds if the ϵ_i, and hence, y_i, $i = 1, \ldots, n$, are exchangeable.

Another important situation that falls into this category is Analysis of Variance models. Indeed, consider an experiment with s factors, $A_1, A_2, \ldots, A_s$, having $a_1, a_2, \ldots, a_s$ levels, respectively. If $y_{i,k}$, $i = (i_1, i_2, \ldots, i_s)$, denotes the kth experimental observation ($k = 1, 2, \ldots, K_i$) on the combination of the i_1-st level of A_1, i_2-nd level of A_2, $\ldots$, i_s-th level of A_s, then the model can be written as

$$y_{i,k} = \mu_i + \sigma\epsilon_{i,k}, \tag{28}$$

where the $\epsilon_{i,k}$ are random variables with the joint density $f(u), u \in \mathbf{R}^n, n = \sum K_i$. By writing

$$y = \{y_{i,k},\ i_1 = 1, \ldots, a_1;\ \ldots;\ i_s = 1, \ldots, a_s;\ k = 1, \ldots, K_i\},$$

$$\beta = \{\mu_i,\ i_1 = 1, \ldots, a_1;\ \ldots;\ i_s = 1, \ldots, a_s\},$$

$$\epsilon = \{\epsilon_{i,k};\ i_1 = 1, \ldots, a_1;\ \ldots;\ i_s = 1, \ldots, a_s;\ k = 1, \ldots, K_i\},$$

the model (28) translates into the regression form (15) with $n \times (a_1 a_2 \cdots a_s)$ design matrix:

$$X = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & & & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}. \tag{29}$$

The size of the minimal training sample is $m = a_1 a_2 \cdots a_s + 1$, and hence any MTS design submatrix $X(l)$ will have 2 coinciding rows. (There can not be more than 2 identical rows since then $X(l)$ is not of full rank and the respective sample is not an MTS). Note that one can always rearrange the observations in the MTS in such a way that the two rows will follow each other. For an MTS, model (15) becomes

$$y(l) = X(l)\beta + \sigma\epsilon. \tag{30}$$

Theorem 3.3 *If the errors $\epsilon_{i,k}$, in the ANOVA model (28) corresponding to an MTS are exchangeable, then (24) holds for that minimal training sample. If all of the error terms in the model (28) are exchangeable, the statement of Theorem 3.1 holds for all minimal training samples, and B^I_{12} is given by (8) or (27).*

Proof: The pool of patterns for MTS design matrices for the model (28) (up to a permutation of observations in an MTS) is

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \quad \dots, \quad \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{31}$$

Note that each of the matrices (31) has two identical rows. It follows that respective hyperplanes take the form

$$\begin{aligned} H_1 &= \{(v_1, v_2, \dots, v_m) : \ v_1 = v_2\}, \\ H_2 &= \{(v_1, v_2, \dots, v_m) : \ v_2 = v_3\}, \\ &\vdots \\ H_{m-1} &= \{(v_1, v_2, \dots, v_m) : \ v_{m-1} = v_m\}. \end{aligned} \tag{32}$$

Indeed, suppose that rows i and $i+1$ of the matrix $X(l)$ are identical. Decompose $|(X(l)\ v)|$ by the last column. Then all the terms not involving v_i or v_{i+1} will contain determinants of matrices with 2 identical rows, and hence will be equal to zero. The term containing v_i will clearly have the same absolute value as the term containing v_{i+1}, but will differ in the sign. Hence the hyperplanes have the form (32).

For condition (24) to hold, it is clearly sufficient for $f(v)$, $v \in \mathbf{R}^m$, to be symmetric with respect to each one of these $m-1$ hyperplanes. This means that $f(v_1, v_2, v_3, \dots, v_m) = f(v_2, v_1, v_3, \dots, v_m)$, $f(v_1, v_2, v_3, \dots, v_m) = f(v_1, v_3, v_2, \dots, v_m)$, $\dots$, $f(v_1, v_2, \dots, v_{m-1}, v_m) = f(v_1, v_2, \dots, v_m, v_{m-1})$. The latter will follow if the error components in an MTS, are exchangeable. The rest of the statement of theorem follows immediately since exchangeability of the error terms in any MTS is implied by the exchangeabilty of *all* of the error terms in the model (28).

□

Remark 3.2 It is the special structure of ANOVA design matrices that allows one to employ the exchangeability condition. It is easy to construct an example of a design pattern $X(l)$ for which exchangeable errors coming from a nonsymmetric distribution will not imply (24). For this, it suffices to choose the design matrix $X(l)$ in such a way that it does not have an $(m-1) \times (m-1)$ zero minor, and then to take $\epsilon_1, \dots, \epsilon_m$ from $f(v)$ for which condition (24) does not hold. For instance, we can choose $y_1 = \mu + \sigma\epsilon_1$, $y_2 = 2\mu + \sigma\epsilon_2$, where ϵ_1, ϵ_2 are iid from an extreme value distribution $f(v_i) = \exp(-v_i) \cdot \exp(-\exp(-v_i))$, $i = 1, 2$. Then the left hand side of (24) becomes $P_f(V_2 > 2V_1)$, which is not equal to 1/2.

On the other hand, two independent observations, Y_1 and Y_2, with $Y_1 \sim N(\mu, \sigma)$, and $Y_2 \sim Cauchy(\mu, \sigma)$ provide an example where the symmetry condition (26) and hence (24) do hold, but exchangeability does not.

4 Marginal Distributions for Minimal Training Samples

The curious identity in (7) can also be generalized to structured linear models. This is of interest for model comparisons involving linear models having differing design matrices. The terms in the intrinsic Bayes factor in (6) that arise from the minimal training samples will then not equal one, but having a simple closed from expression for the $m_j^r(y(l))$ greatly reduces the computational burden in evaluating (6). Indeed, the expression (33) below was the basis for the formulae in Berger and Pericchi (1996b) and Varshavsky (1996) for determining intrinsic Bayes factors when comparing normal linear models. Note, however, that (6) will also hold for a wide variety of non-normal error structures. That closed form computation of the $m_j^r(y(l))$ is then still possible is quite surprising and extremely useful.

Theorem 4.1 *For the linear model in section 3, if condition (24) holds then the marginal density $m_j^r(y(l))$, with respect to the prior $\pi(\beta,\sigma) = 1/\sigma$, is*

$$m_j^r(y(l)) = \frac{1}{2\cdot(|X(l)^tX(l)|^+)^{1/2}\cdot\|y(l) - X(l)(X(l)^tX(l))^{-1}X(l)^ty(l)\|}\,. \tag{33}$$

Proof: Let

$$v = (v_1,\ldots,v_m)^t = (y(l) - X(l)\beta)^t/\sigma. \tag{34}$$

Letting $x_i = (x_{i,1},\ldots,x_{i,m-1})$ be the ith row of $X(l)$, the inverse of the Jacobian for the transformation to v is

$$\begin{aligned} J^{-1} &= \frac{dv_1,\ldots,dv_m}{d\beta_1,\ldots,d\beta_{m-1}d\sigma} \\ &= \frac{1}{\sigma^{m+1}}\left|\begin{array}{ccc} x_{1,1} & \ldots & x_{m,1} \\ \vdots & & \vdots \\ x_{1,m-1} & \ldots & x_{m,m-1} \\ y_1(l) - x_1\beta & \ldots & y_m(l) - x_m\beta \end{array}\right|^+ = \frac{1}{\sigma^{m+1}}\left|\begin{array}{ccc} x_{1,1} & \ldots & x_{m,1} \\ \vdots & & \vdots \\ x_{1,m-1} & \ldots & x_{m,m-1} \\ y_1(l) & \ldots & y_m(l) \end{array}\right|^+. \end{aligned}$$

Thus $J^{-1} = |(X(l)\ y(l))|^+/\sigma^{m+1}$. Note that $|(X(l)\ v)| = |(X(l)\ y(l))|/\sigma$. Using this, the marginal becomes

$$m_j^r(y(l)) = \int_{\mathrm{R}^{m-1}\times\mathrm{R}^+} f_j(\frac{y(l) - X(l)\beta}{\sigma})\cdot\frac{\pi(\beta,\sigma)}{\sigma^m}\,d\beta\,d\sigma \ = \ \frac{1}{|(X(l)\ y(l))|^+}\cdot\int_{\Omega(l)} f_j(v)\,dv.$$

Hence, in view of condition (24), we have $m_j^r(y(l)) = \left(2\cdot|(X(l)\ y(l))|^+\right)^{-1}$. To complete the proof, observe that

$$|(X(l)\ y(l))|^+ = (|X(l)^tX(l)|^+)^{1/2}\cdot\|y(l) - X(l)(X(l)^tX(l))^{-1}X(l)^ty(l)\|, \tag{35}$$

since $|(X(l)\ y(l))|^+$ is just the volume of the m dimensional hyperparallelepiped based on columns of $X(l)$ and vector $y(l)$ (see, for example, Shilov (1961)); $(|X(l)^tX(l)|^+)^{1/2}$ is the volume of its base and $\|y(l) - X(l)(X(l)^tX(l))^{-1}X(l)^ty(l)\|$ is the length of the height to the base.

□

5 Transformed Linear Models

Theorems 3.3 and 4.1 can be generalized to the case where the model is not of the form (15), but can be reduced to it by a suitable transformation. To fix notation, suppose, observations $z_i,\ i = 1,\ldots,n$, have a joint density $g(z|\rho)$, where $z \in \mathbf{R}^n$ and $\rho \in \mathbf{R}^m$. Suppose further that a minimal training sample, $z(l)$, under the prior $\pi(\rho)$ consists of m observations. Also assume that there exist transformations $y_i = h_i(z),\ i = 1,\ldots,n$, of the data, and $\beta_1 = \psi_1(\rho),\ \ldots,\ \beta_{m-1} = \psi_{m-1}(\rho),\ \sigma = \psi_m(\rho)$, of the parameters, such that the transformed data can be represented by the model (15) and the transformed prior density, $\pi(\beta,\sigma), \beta \in \mathbf{R}^{m-1}, \sigma \in \mathbf{R}^1$, is equal to $1/\sigma$. Then the following is true.

Proposition 5.1 *Suppose that the error density of the transformed data y as defined in (15) satisfies condition (24). (Note that Remark 3.1, Example 3.1, and Theorem 3.3 give conditions under which this is so). Then the marginal density of $z(l)$ is*

$$m(z(l)) = \frac{J(l)}{2\cdot(|X(l)^tX(l)|^+)^{1/2}\cdot\|h(z(l)) - X(l)(X(l)^tX(l))^{-1}X(l)^th(z(l))\|}\ , \tag{36}$$

where $J(l) = \frac{dh_1\ldots dh_m}{dz_1\ldots dz_m}$ is the Jacobian of the transformation $h = (h_1,\ldots h_m)$ from $z(l)$ to $y(l)$.

Proof: Straightforward change of variables. □

Typically, the choice of the transformation h of the data, if one exists, dictates the transformation of the parameters.

Example 5.1 Suppose $z_i,\ i = 1,\ldots,n$, are *iid* Weibull random variables distributed with the density $p(z_i|\alpha,\gamma) = \alpha\cdot z_i^{\alpha-1}\gamma^{-\alpha}\cdot\exp(-(z_i/\gamma)^\alpha)$. Suppose we are using the reference prior $\pi(\alpha,\gamma) = 1/(\alpha\gamma)$. (The reference prior, defined, for example, in Berger and Bernardo (1989), typically reduces to a right Haar measure in a group invariant model.) Employing Proposition 5.1, let $y_i = \log(z_i)$. Then the density of y_i is

$$f(y_i|\alpha,\gamma) = \alpha\cdot\exp\{(y_i - \log(\gamma))\alpha\}\cdot e^{-\exp\{(y_i-\log(\gamma))\alpha\}}\ .$$

If we now set $\beta = \log(\gamma)$ and $\eta = 1/\alpha$, the density $f(y_i|\beta,\eta)$ becomes a location-scale density with location parameter β and scale η. The Jacobian of the transformation to (η,β) is $\exp(\beta)/\eta^2$ and, hence, the transformed prior on the new parameters is $\pi(\beta,\eta) = 1/\eta$. Thus, by Proposition 5.1 and using Example 3.1,

$$m(z_i,z_j) = \frac{1}{2\cdot z_iz_j\cdot|log(z_i/z_j)|}.$$

Example 5.2 Reliability and life-testing studies often deal with sets of nonnegative data arising from a skewed distribution. The competing distributions in this case are often chosen to be two-parameter Weibull and Lognormal densities (McDonald *et al*, 1994); that is, the comparison

$$M_1 : p_1(z_i|\alpha,\gamma) = \alpha\cdot z_i^{\alpha-1}\gamma^{-\alpha}\exp(-(z_i/\gamma)^\alpha)$$

versus

$$M_2 : p_2(z_i|\mu,\sigma) = (1/(\sqrt{2\pi}\sigma z_i))\exp(-(log(z_i) - \mu)^2/(2\sigma^2))$$

for the observed iid data z_i, $i = 1, \ldots, n$, is of interest. Model M_1, with the reference prior $\pi_1(\alpha, \gamma) = 1/(\alpha\gamma)$ was considered in Example 5.1 above. Using the reference prior $\pi_2(\mu, \sigma) = 1/\sigma$ for the parameters of the second model and applying the same logarithmic transformation to the data, with the identity transformation for the parameters, Proposition 5.1 yields exactly the same marginal density for a minimal training sample. Hence, $B_{12}(z(l)) = 1$ for all training samples, and the intrinsic Bayes factor again simplifies as in *(8)*.

Example 5.3 An important class of examples is the class of generalized linear models, where the data z_i comes from $z = \exp(X\beta + \sigma\epsilon)$. Here we can again apply Proposition 5.1 with the logarithmic transformation of the data.

Acknowledgments. This work was supported by National Science Foundation Grants DMS-8923071 and DMS-9303556 and BID-Conicit. The authors thank Anirban DasGupta, ShyamalKumar Nariankadu, and J.K. Ghosh for many helpful discussions and insightful comments.

Appendix

Proof of Proposition 3.1.

To save on notation, let $k = m - 1$. The process of determining a maximal invariant will be carried out in two steps, corresponding to two subgroups that generate G:

$$G_1 : y \to y + Xb, \quad b = (b_1, \dots, b_k), \quad b \in \mathbf{R}^k,$$

$$G_2 : y \to cy, \quad c > 0.$$

(see Lehmann (1991)). The orbit containing the point $y_0 \in \mathbf{R}^n$ under G_1 is the set

$$\{y_0 + Xb, \ b \in \mathbf{R}^k\}. \tag{37}$$

To find a maximal invariant, let us choose a representative point O on this orbit. Let it be the point of intersection with the hyperplane $\{y_{n-k+1} = 0, \dots, y_n = 0\}$. Denote the matrix consisting of the first $n-k$ rows of X as X_1, and the matrix consisting of the last k rows as X_2. Likewise, let $y_0^{(1)} \in \mathbf{R}^{(n-k)}$ be a vector consisting of the first $(n-k)$ components of y_0, and $y_0^{(2)} \in \mathbf{R}^k$ a vector consisting of the last k components. Then we see that O corresponds to $b = -X_2^{-1} y_0^{(2)}$ and, thus, using (37), can be written as $(y_0 - X\,X_2^{-1} y_0^{(2)}) \ = \ (y_0^{(1)} - X_1 X_2^{-1} y_0^{(2)}, 0, \dots, 0)$. This leads to the maximal invariant (under G_1)

$$T_1(y_1, \dots, y_n) \ = \ y^{(1)} - X_1 X_2^{-1} y^{(2)}, \tag{38}$$

where $y^{(1)} = (y_1, \dots, y_{n-k})$, and $y^{(2)} = (y_{n-k+1}, \dots, y_n)$. Letting $x_{(i)}$ be the ith row of X and $a_{(i)} = (a_{i,n-k+1}, \dots, a_{i,n}) = x_{(i)} X_2^{-1}$, we can rewrite (38) component-wise as

$$\begin{aligned} T_1(y_1, \dots, y_n) \ &= \ (y_1 - \sum_{j=n-k+1}^{n} a_{1,j} \cdot y_j, \ \dots, \ y_{n-k} - \sum_{j=n-k+1}^{n} a_{n-k,j} \cdot y_j) \\ &= \ (z_1, \dots, z_{n-k}), \ \text{ say.} \end{aligned}$$

A maximal invariant with respect to G_2 acting on z is

$$T_2 = (z_1/|z_{n-k}|, \dots, z_{n-k-1}/|z_{n-k}|, \mathrm{sign}(z_{n-k})).$$

Switching back to the space of y's, we obtain the maximal invariant $T(y_1, \dots, y_n)$ under G:

$$T(y) = \left(\frac{y_1 - \sum_{j=n-k+1}^{n} a_{1,j} \cdot y_j}{|y_{n-k} - \sum_{j=n-k+1}^{n} a_{n-k,j} \cdot y_j|}, \ \dots, \ \frac{y_{n-k-1} - \sum_{j=n-k+1}^{n} a_{n-k-1,j} \cdot y_j}{|y_{n-k} - \sum_{j=n-k+1}^{n} a_{n-k,j} \cdot y_j|}, \right.$$

$$\left. \mathrm{sign}(y_{n-k} - \sum_{j=n-k+1}^{n} a_{n-k,j} \cdot y_j) \right).$$

$\square$

REFERENCES

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika* **65** 39-48.

Andersson, S. (1982). Distributions of maximal invariants using quotient measures. *Ann. Statist.* **10** 955-961.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York.

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 Pericchi, L.R. (1996a). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109-122.

Berger, J. O. and Pericchi, L.R. (1996b). The intrinsic Bayes factor for linear models (with discussion). In *Bayesian Statistics 5* (J. M. Bernardo *et al*, eds.). Oxford University Press, London.

Eaton, M. L. (1989). *Group Invariance Applications in Statistics.* Regional Conference Series in Probability and Statistics, v. 1., Inst. Math. Statist., Hayward, Ca.

Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153-160.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (J. M. Bernardo *et al*, eds.). Oxford University Press, London.

Lehmann, E. L. (1991). *Theory of Point Estimation.* Wadsworth & Brooks/Cole, Pacific Grove, Ca.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* University of Rotterdam Press, Rotterdam.

McDonald, G. C., Vance L. C. and Gibbons D. I. (1994). Some tests for discriminating between lognormal and Weibull distributions - an application to emission data. *Recent Advances in Life-Testing and Reliability* (N. Balakrishnan , ed.) CRC Press, Inc.

Shilov, G. E. (1961). *An Introduction to the Theory of Linear Spaces.* Prentice-Hall, Englewood Cliffs, N.J.

Spiegelhalter, D. (1975). Robust estimation of location using a finite mixture model. M.Sc. Thesis, University College, London.

Spigelhalter, D. J. and Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *J. Roy. Statist. Soc. Ser. B* **44** 377-387.

Varshavsky, J. A. (1996). Intrinsic Bayes factors for model selection with autoregressive data. In *Bayesian Statistics 5*, (J. M. Bernardo, *et al* eds.). Oxford University Press, London.

Wijsman, R. A. (1990). *Invariant Measures on Groups and Their Use in Statistics.* Lecture Notes-Monograph Series. v. 14, Inst. Math. Statist., Hayward, Ca.