ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Statistics & Probability Letters III (IIII) III–III

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

A characterization of consistency of model weights given partial information in normal linear models

Hubert Wong[a,*], Bertrand Clarke[b,1]

[a] *Department of Health Care and Epidemiology, University of British Columbia, Vancouver, BC, Canada*
[b] *Department of Statistics, University of British Columbia, Vancouver, BC, Canada*

Received April 2003

Abstract

We characterize the consistency of posterior model probabilities that are computed conditional on affine functions of the outcome variable for normal linear models.

Keywords: Consistency; Posterior probabilities; Model averaging; Partial information

1. Introduction

How much information do we really need for the posterior probabilities of models to be consistent? As a generality, if one chooses $\alpha\%$ of the data at random then consistency is assured, for $\alpha > 0$. Consistency can be assured when conditioning on finitely many statistics if the statistics are sufficient. Outside of special cases such as these, consistency given only partial information is difficult to establish.

Here, we restrict our attention to the case of normal linear models and characterize when posterior probabilities conditional on affine functions of the outcome variable are consistent. That is, we characterize when the posterior probability given the affine function goes to zero for all wrong models and to one for the right model.

The result here is of potential interest in three broad areas. First, it arose in a model averaging context in which the goal was to ensure that predictions made for time-point $n + 1$ would respond

* Corresponding author.
E-mail address: hubert@hivnet.ubc.ca (H. Wong).
[1] Partial support was provided by NSERC Discovery/Operating Grants RGPIN 261834-03 (H. Wong) and OGP-0138122 (B. Clarke).

0167-7152/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2004.01.011

to the past predictive performance of the models up to time-point n. This led to partial information model weights obtained as posterior probabilities computed conditional on past prediction residuals rather than full data (Wong, 2000; Wong and Clarke, 2004). Our main result here permits determination of when such conditioning gives consistency. These partial information model weights can also be used to discredit poor models in a Bayesian model averaging (BMA). In standard BMA, the Bayes factors for comparing pairs of models are used to exclude models which fit the data poorly relative to other models (for a recent review, see Clyde, 1999). This approach is often applied in combination with the Occam's window argument (Madigan and Raftery, 1994) in which a simple model with a larger posterior probability than a more complex model discredits the latter. By conditioning on past performance, partial information model weights respond specifically to information on predictive accuracy rather than simply model fit.

Second, selecting one model from a finite list of models is, formally, much the same as estimating a discrete parameter. As in the continuous case, we want to know when posterior probabilities converge and the rates at which they do. We conjecture that as in other discrete cases the rate will be exponential but have not investigated this.

Finally, posterior distributions have natural interpretations in information theory, see Cover and Thomas (1991, Chapters 5, 8, 13, 14). These arise because posterior convergence results automatically lead to conjectures for the asymptotic behavior of the Shannon mutual information, SMI. In the simplest case, consider models $P_1, \ldots, P_\theta, \ldots, P_M$ with a distribution W on θ. The SMI between Θ and Y where Y comes from one of the P_θ's is $I(\Theta; Y) = H(\Theta) - H(\Theta|Y)$, where $H(\Theta|Y)$ is the conditional entropy of $(\Theta|Y)$ which reduces to the entropy H when Y does not appear. When Y is a finite dimensional statistic for which our theorem gives consistency, we have that $H(\Theta|Y)$ goes to zero. This shows that the capacity achieving distribution on Θ is 'maxent' and continues to be uniform.

More generally, the conditioning permitted in our theorem can be regarded as partial information, for use in data compression and transmission in multi-user networks.

We first state our main theorem for the case of two nested models. We state and prove an extension to two non-nested models as a corollary. A second corollary extends the main theorem to any finite number of models. Section 2 sets up the notation and states our results formally. In addition, we give an example illustrating an application of the main theorem. Section 3 discusses the conditions used in the characterization and suggests when the conditions may be simplified. The formal proof of our main result is given in the appendix.

2. Consistency of model weights

Let $\mathbf{Y}_{(n)} = (Y_1, Y_2, \ldots, Y_n)$ denote the vector of outcomes. For each outcome Y_i, let $\mathbf{X}_i$ denote the p-vector of all candidate predictors and let $\mathbf{X}_{(n)}$ denote the $n \times p$ matrix with row i equal to $\mathbf{X}_i$. We consider normal linear models, indexed by $k = 1, \ldots, K$, of the form

$$\mathbf{Y}_{(n)}|\mathbf{Z}_{k,(n)}, \beta_k \sim \mathcal{N}(\mathbf{Z}_{k,(n)}\beta_k, \sigma^2\mathbf{I}),$$

where $\mathbf{Z}_{k,(n)}$ is the sub-matrix of $\mathbf{X}_{(n)}$ that contains only the columns associated with the p_k predictors that are included in model k. For simplicity, assume that σ^2 is known. Let β denote the parameter vector for the full model, containing all candidate predictors $\mathbf{Z}_{(n)}$. The set of all candidate

predictors is the union of the sets of predictors used in the K models. Let π, the prior distribution on β, be $\mathcal{N}(\mathbf{b}, \Gamma)$. Without loss of generality, it is sufficient to consider the case where Γ is diagonal. If Γ is not diagonal, the problem can be transformed to the diagonal case by re-parameterizing the model $\mathbf{Z}_{(n)}\beta$ as $\mathbf{Z}^*_{(n)}\beta^*$ and solving in terms of $\mathbf{Z}^*_{(n)} = \mathbf{Z}_{(n)}\Gamma^{1/2}$ and $\beta^* = \Gamma^{-1/2}\beta$. Assume the prior π_k for β_k is obtained by restricting π to the components that are in model k, i.e., $\beta_k \sim \mathcal{N}(\mathbf{b}_k, \Gamma_k)$, where $\mathbf{b}_k$ and Γ_k are the appropriate subsets of $\mathbf{b}$ and Γ, respectively. The distribution for $\mathbf{Y}_{(n)}$ under model k, after mixing over the prior for β_k, is $\mathcal{N}(\nu_k, \Psi_k)$, where

$$\begin{aligned} \nu_k &= \mathbf{Z}_{k,(n)}\mathbf{b}_k, \\ \Psi_k &= \sigma^2\mathbf{I} + \mathbf{Z}_{k,(n)}\Gamma_k\mathbf{Z}^{\mathrm{T}}_{k,(n)}. \end{aligned} \tag{1}$$

The "true" model is the smallest sub-model that contains all predictors whose coefficients are non-zero in the data generator.

For prior weights $\alpha_o = (\alpha_{1,o}, \ldots, \alpha_{K,o})$ on the models under consideration and a statistic $\mathbf{S}_n = \mathbf{S}_n(\mathbf{Y}_{(n)})$, Bayes rule for updating α_o to get the posterior weights $\alpha(\mathbf{S}_n) = (\alpha_1(\mathbf{S}_n), \ldots, \alpha_K(\mathbf{S}_n))$, gives, for each model k,

$$\alpha_k(\mathbf{S}_n) = \frac{\alpha_{k,o} m_k(\mathbf{S}_n)}{\sum_{i=1}^{K} \alpha_{i,o} m_i(\mathbf{S}_n)}, \tag{2}$$

where $m_k(\mathbf{S}_n)$ is the density of $\mathbf{S}_n$ after mixing over the prior for β_k. If model k is true, then the posterior weights will be consistent if and only if $\alpha_i(\mathbf{S}_n) \to 0, \forall i \neq k$, or equivalently,

$$\frac{\alpha_k(\mathbf{S}_n)}{\alpha_i(\mathbf{S}_n)} = \frac{\alpha_{k,o}}{\alpha_{i,o}} \frac{m_k(\mathbf{S}_n)}{m_i(\mathbf{S}_n)} \to \infty, \quad \forall i \neq k.$$

Thus, consistency is determined by the asymptotic behavior of the marginal density ratios $m_k(\mathbf{S}_n)/m_i(\mathbf{S}_n)$. The symbol $\to$ indicates convergence in probability.

We characterize the conditions under which the posterior weights are consistent when the partial information statistic $\mathbf{S}_n$ is affine in $\mathbf{Y}_{(n)}$, i.e.,

$$\mathbf{S}_n = \mathbf{U}^{\mathrm{T}}(\mathbf{Y}_{(n)} + \mathbf{c}),$$

where $\mathbf{U}$ and $\mathbf{c}$ are arbitrary functions of $\mathbf{X}_{(n)}$ but do not depend on $\mathbf{Y}_{(n)}$. (For notational simplicity, the dependence of $\mathbf{U}$ and $\mathbf{c}$ on n has been suppressed.) Since we are interested in $\mathbf{S}_n$ only for the σ-field it generates, without loss of generality we assume $\mathbf{U}$ is of full rank.

From the properties of the multivariate normal distribution and the fact that $\mathbf{S}_n$ is affine in $\mathbf{Y}_{(n)}$, $\mathbf{S}_n$ is distributed under model k as a $\mathcal{N}(\mu_k, \Sigma_k)$ with mean and variance

$$\begin{aligned} \mu_k &= \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}\mathbf{b}_k + \mathbf{U}^{\mathrm{T}}\mathbf{c}, \\ \Sigma_k &= \sigma^2\mathbf{U}^{\mathrm{T}}\mathbf{U} + \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}\Gamma_k\mathbf{Z}^{\mathrm{T}}_{k,(n)}\mathbf{U}. \end{aligned} \tag{3}$$

Hence the densities in (2) are given by

$$m_k(\mathbf{S}_n) = (2\pi)^{-J/2}|\Sigma_k|^{-1/2} \exp\left\{-\tfrac{1}{2}(\mathbf{S}_n - \mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mathbf{S}_n - \mu_k)\right\}.$$

We begin with a characterization of consistency for comparing two nested models. Let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote, respectively, the minimum and the maximum eigenvalues of the matrix $\mathbf{A}$.

Theorem 2.1. *Let j and k index two models with model j nested within model k. Partition the predictors in model k as $\mathbf{Z}_{k,(n)} = (\mathbf{Z}_{j,(n)} | \tilde{\mathbf{Z}}_{k,(n)})$ where $\tilde{\mathbf{Z}}_{k,(n)}$ denotes the predictors that are in model k but not model j. Define*

$$\mathbf{G} \equiv \Sigma_j^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{k,(n)} = (\mathbf{U}^{\mathrm{T}} \Psi_j \mathbf{U})^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{k,(n)}. \tag{4}$$

The condition

$$\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G}) \to \infty \tag{5}$$

is:

(i) *if model j is true, necessary and sufficient for $\alpha_j \to 1$, and*
(ii) *if model k is true, necessary for $\alpha_k \to 1$ and is sufficient if in addition*

$$\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G}) \leqslant R\lambda_{\min}(\mathbf{G}^{\mathrm{T}}\mathbf{G}) \tag{6}$$

for all n and some constant R.

Proof. The proof is somewhat long and is given in the appendix. □

The need for (6) ruins the symmetry of the result; we discuss this point in the concluding section. The general case of two non-nested models is covered by the following:

Corollary 2.1. *Let j and k index a pair of non-nested models where model k is true. Let the index c denote the model that contains only the predictors $\mathbf{Z}_{c,(n)}$ that are common to both models. Let the index f denote the full model that contains all predictors $\mathbf{Z}_{(n)}$ used in either model. Partition $\mathbf{Z}_{(n)} = (\mathbf{Z}_{c,(n)} | \tilde{\mathbf{Z}}_{j,(n)} | \tilde{\mathbf{Z}}_{k,(n)})$ where $\tilde{\mathbf{Z}}_{j,(n)}$ and $\tilde{\mathbf{Z}}_{k,(n)}$ are the predictors that are unique to models j and k, respectively. Define*

$$\mathbf{G}_* \equiv \Sigma_c^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{k,(n)} = (\mathbf{U}^{\mathrm{T}} \Psi_c \mathbf{U})^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{k,(n)}, \tag{7}$$

$$\mathbf{G}_\dagger \equiv \Sigma_k^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{j,(n)} = (\mathbf{U}^{\mathrm{T}} \Psi_k \mathbf{U})^{-1/2} \mathbf{U}^{\mathrm{T}} \tilde{\mathbf{Z}}_{j,(n)}. \tag{8}$$

Then at least one of the two conditions

$$\lambda_{\max}(\mathbf{G}_*^{\mathrm{T}}\mathbf{G}_*) \to \infty, \tag{9}$$

$$\lambda_{\max}(\mathbf{G}_\dagger^{\mathrm{T}}\mathbf{G}_\dagger) \to \infty, \tag{10}$$

is necessary for $\alpha_k(\mathbf{S}_n) \to 1$. Sufficiency is obtained if either (9) holds or both (10) and the condition

$$\lambda_{\max}(\mathbf{G}_\dagger^{\mathrm{T}}\mathbf{G}_\dagger) \leqslant R\lambda_{\min}(\mathbf{G}_\dagger^{\mathrm{T}}\mathbf{G}_\dagger) \tag{11}$$

for all n and some constant R, holds.

Proof. Model k is nested within model f with model k the true model. Model j is nested within model f with model f the true model. Consistency holds if and only if

$$\frac{m_k(\mathbf{S}_n)}{m_j(\mathbf{S}_n)} = \frac{m_k(\mathbf{S}_n)}{m_f(\mathbf{S}_n)} \frac{m_f(\mathbf{S}_n)}{m_j(\mathbf{S}_n)} \to \infty. \tag{12}$$

Necessity: If both (9) and (10) fail to hold then by the theorem, both $m_k(\mathbf{S}_n)/m_f(\mathbf{S}_n)$ and $m_f(\mathbf{S}_n)/m_j(\mathbf{S}_n)$ are bounded.

Sufficiency: It is intuitively clear that both $m_k(\mathbf{S}_n)/m_f(\mathbf{S}_n)$ and $m_f(\mathbf{S}_n)/m_k(\mathbf{S}_n)$ are bounded away from 0 in probability since it is not possible to discredit the true model (i.e., the model in the numerator in each ratio) in favor of an incorrect model. The formal verification is given at the end of proof of the theorem. Hence, it is sufficient to have one of the ratios go to ∞. If (9) holds, then by the theorem, $m_k(\mathbf{S}_n)/m_f(\mathbf{S}_n) \to \infty$. If both (10) and (11) hold then by the Theorem, $m_f(\mathbf{S}_n)/m_j(\mathbf{S}_n) \to \infty$. □

These results extend immediately to the case of more than two models:

Corollary 2.2. *Let i, $i = 1,\ldots,K$ index a collection of models with k as the index to the true model. Then $\alpha_k(\mathbf{S}_n) \to 1$ iff for each pair of models (k,i), $i = 1,\ldots,K$, $i \neq k$, the conditions in the Theorem (for nested models) or the first Corollary (for non-nested models), are satisfied.*

We illustrate an application of the Theorem in the following example.

Example. Wong and Clarke (2004) consider a model averaging approach in which $\mathbf{S}_n$ consists of some of the *predictuals*, $Y_t - \hat{Y}_{k,t}$, $t < n$, where $\hat{Y}_{k,t}$ is the prediction issued by model k at time t. A specific scenario was the averaging of two nested linear models, differing by a single predictor variable. Suppose $\mathbf{S}_n$ is set to be a fraction of the most recent predictuals from the smaller model. Does such a set of predictuals contain sufficient information to obtain consistency in the model weights? The expressions for $\mathbf{U}$ in this situation are complicated making analytic evaluation of the asymptotic behaviour of $\mathbf{G}^T\mathbf{G}$ (which in this scenario is a scalar) difficult. However, for a given sequence it is straightforward to evaluate the required quantities numerically. Fig. 1 plots the value of $\mathbf{G}^T\mathbf{G}$ over time for 3 randomly generated sequences of data. In the first row of the figure, $\mathbf{S}_n$ was set to be the complete data, or equivalently, to the set of all past predictuals. For this specification, it is easy to verify analytically that the value of $\mathbf{G}^T\mathbf{G}$ increases monotonically at a linear rate and so consistency is assured. In the second row, $\mathbf{S}_n$ was set to the last $n/2$ (rounded down) predictuals. Now, $\mathbf{G}^T\mathbf{G}$ does not increase monotonically but the trend continues to be a linear increase, albeit at a slower rate. Thus, the theorem suggests that this specification also gives consistency. The same conclusion is reached if $\mathbf{S}_n$ is set to the last $n/4$ predictuals with yet an even slower rate of increase than that observed from using the last $n/2$ predictuals and there is greater variability (third row). In the bottom row, $\mathbf{S}_n$ consists of only the 5 most recent predictuals at each time-point. In this case, the plots suggest that consistency would not obtain.

3. Discussion

We have shown that consistency of model weights is characterized by the asymptotic behaviour of the minimum and maximum eigenvalues of $\mathbf{G}^T\mathbf{G}$ (and $\mathbf{G}_*^T\mathbf{G}_*$ and $\mathbf{G}_\dagger^T\mathbf{G}_\dagger$ if needed). As noted in the example, analytic evaluation of these eigenvalues is difficult but one can obtain a sense of the large sample behaviour in any given instance as long as it is straightforward to compute the eigenvalues numerically.

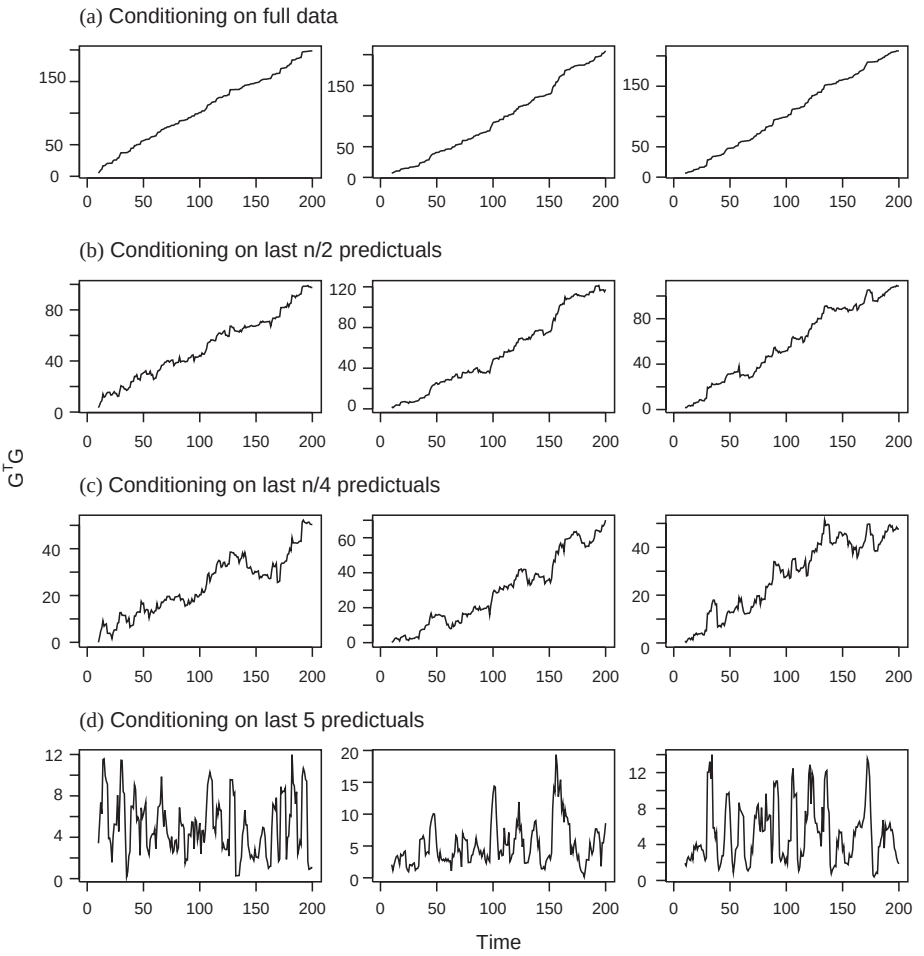

Fig. 1. Time evolution of the value of $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ for 3 randomly generated sequences (one per column) obtained by setting $\mathbf{S}_n$ equal: (a) full data; (b) the most recent $n/2$ predictuals; (c) the most recent $n/4$ predictuals; and (d) the most recent 5 predictuals.

The extra condition (6) needed for sufficiency is a nuisance. We suspect that the condition always holds as long as $\tilde{p}$, the number of predictors in $\tilde{\mathbf{Z}}_{k,(n)}$, does not exceed the dimension of $\mathbf{U}$. We observe that $\mathbf{G}^{\mathrm{T}}\mathbf{G} = \tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U}(\mathbf{U}^{\mathrm{T}}\Psi_j\mathbf{U})^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)} = \mathbf{Z}^{\mathrm{T}}\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{Z}$ where $\mathbf{Z} = \Psi_j^{-1/2}\tilde{\mathbf{Z}}_{k,(n)}$ and $\mathbf{W} = \Psi_j^{1/2}\mathbf{U}$. The matrix $\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}$ is idempotent and so has eigenvalues taking values of 0 or 1 only. Hence, if this matrix has rank $\tilde{p}$ or greater so that $\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\Psi_j^{-1}\tilde{\mathbf{Z}}_{k,(n)}$ is full rank, it is plausible to expect that the relationship between the eigenvalues of $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$ is unchanged by inclusion of $\mathbf{W}(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}$ in the centre. If this holds, then whether or not (6) holds does not depend on $\mathbf{U}$ and (6) reduces to the condition $\lambda_{\max}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})/\lambda_{\min}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}) < R$, $\forall n$ and some constant R. This condition is the same as that for consistency given full data (which is pre-supposed as satisfied).

Appendix—proof of Theorem 2.1

Let $\lambda_i(\mathbf{A})$ denote the ith eigenvalue of the matrix $\mathbf{A}$. Let $\|\mathbf{x}\|_{\mathbf{A}}^2 \equiv \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$. For any non-negative definite matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, we write $\mathbf{A} \leqslant \mathbf{B}$ iff $\|\mathbf{x}\|_{\mathbf{A}}^2 \leqslant \|\mathbf{x}\|_{\mathbf{B}}^2$ for all vectors $\mathbf{x}$. The following matrix inequality will be useful: For compatibly dimensioned matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{D}$ with $\mathbf{A}$ and $\mathbf{D}$ positive definite,

$$\mathbf{B}(\mathbf{D}^{-1} + \mathbf{B}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}} \leqslant \mathbf{A}. \tag{13}$$

Let

$$M = \frac{m_j(\mathbf{S}_n)}{m_k(\mathbf{S}_n)}$$

denote the ratio of the marginal densities for $\mathbf{S}_n$ from the two models and to define

$$-2\log M = \Delta - \log D,$$

where

$$\begin{aligned} D &= \frac{|\Sigma_k|}{|\Sigma_j|} = |\Sigma_j^{-1/2}\Sigma_k\Sigma_j^{-1/2}|, \\ \Delta &= \|\mathbf{S}_n - \mu_j\|_{\Sigma_j^{-1}}^2 - \|\mathbf{S}_n - \mu_k\|_{\Sigma_k^{-1}}^2 \\ &= \|\mathbf{S}_n - \mu_j\|_{\Sigma_j^{-1}-\Sigma_k^{-1}}^2 - 2(\mathbf{S}_n - \mu_j)^{\mathrm{T}}\Sigma_k^{-1}(\mu_j - \mu_k) - \|\mu_j - \mu_k\|_{\Sigma_k^{-1}}^2. \end{aligned}$$

Denote the true parameter value by $\beta_o = (\beta_o^*, \tilde{\beta})$ where $\tilde{\beta} \equiv \mathbf{0}$ when model j is true. Let

$$\begin{aligned} \mu_o &= \mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}\beta_o + \mathbf{U}^{\mathrm{T}}\mathbf{c} \\ &= \mathbf{U}^{\mathrm{T}}(\mathbf{Z}_{j,(n)}\beta_o^* + \tilde{\mathbf{Z}}_{k,(n)}\tilde{\beta}) + \mathbf{U}^{\mathrm{T}}\mathbf{c} \end{aligned}$$

denote the expected value of $\mathbf{S}_n$ under the data generator.

Let $\mathbf{E}_i$ and $\mathbf{V}_i$ denote, respectively, the expection and variance operators under model i with parameter value β_o. Applying the well-known results for the mean and variance of a normal quadratic form, we obtain

$$\begin{aligned} \mathbf{E}_i\Delta &= [\mathrm{tr}(\Sigma_j^{-1}\Sigma_i) + \|\mu_o - \mu_j\|_{\Sigma_j^{-1}\Sigma_i\Sigma_j^{-1}}^2] - [\mathrm{tr}(\Sigma_k^{-1}\Sigma_i) + \|\mu_o - \mu_k\|_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}^2] \\ &= \mathrm{tr}(\Sigma_j^{-1}\Sigma_i - \Sigma_k^{-1}\Sigma_i) + \|\mu_o - \mu_j\|_{\Sigma_j^{-1}\Sigma_i\Sigma_j^{-1}}^2 - \|\mu_o - \mu_k\|_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}^2 \end{aligned} \tag{14}$$

and

$$\begin{aligned} \mathbf{V}_i\Delta &\leqslant 2[\mathbf{V}_i(\|\mathbf{S}_n - \mu_j\|_{\Sigma_j^{-1}-\Sigma_k^{-1}}^2) + \mathbf{V}_i(2(\mathbf{S}_n - \mu_k)^{\mathrm{T}}\Sigma_k^{-1}(\mu_j - \mu_k))] \\ &= 4[\mathrm{tr}((\Sigma_j^{-1} - \Sigma_k^{-1})\Sigma_i)^2 + 2\|\mu_o - \mu_j\|_{(\Sigma_j^{-1}-\Sigma_k^{-1})\Sigma_i(\Sigma_j^{-1}-\Sigma_k^{-1})}^2 \\ &\quad + \|\mu_j - \mu_k\|_{\Sigma_k^{-1}\Sigma_i\Sigma_k^{-1}}^2], \end{aligned}$$

where the inequality follows from the relation $\mathbf{V}(A + B) \leqslant 2(\mathbf{V}(A) + \mathbf{V}(B))$.

Lemma 3.1. *The quantities* $\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}}$, $\|\mu_o - \mu_j\|^2_{\Sigma_k^{-1}}$, *and* $\|\mu_j - \mu_k\|^2_{\Sigma_k^{-1}}$ *are bounded irrespective of whether model k or model j is true.*

Proof. Substituting for Σ_k^{-1} using (3) and applying (13),

$$\begin{aligned}\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}} &= \|\beta_o - \mathbf{b}_k\|_{\mathbf{Z}^{\mathrm{T}}_{k,(n)}\mathbf{U}\Sigma_k^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}}\\ &\leqslant \|\beta_o - \mathbf{b}_k\|_{\Gamma_k^{-1}}.\end{aligned}$$

The expression on the final line involves only constants and hence is bounded. The proofs for the other two quantities follow analogously. □

Lemma 3.2. *When model j is true,* $\|\mu_o - \mu_j\|^2_{\Sigma_j^{-1}}$ *is bounded.*

Proof. Using the fact that $\tilde{\beta} = \mathbf{0}$, substituting for Σ_j^{-1} using (1), and applying (13),

$$\begin{aligned}\|\mu_o - \mu_j\|^2_{\Sigma_j^{-1}} &= \|\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}(\beta_o^* - \mathbf{b}_j)\|_{\Sigma_j^{-1}}\\ &= \|\beta_o^* - \mathbf{b}_j\|_{\mathbf{Z}^{\mathrm{T}}_{k,(n)}\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}}\\ &\leqslant \|\beta_o^* - \mathbf{b}_k\|_{\Gamma_j^{-1}}.\end{aligned}$$

The expression on the last line involves only constants and hence is bounded. □

Let Γ_k be partitioned as

$$\Gamma_k = \begin{pmatrix} \Gamma_j & \mathbf{0} \\ \mathbf{0} & \tilde{\Gamma} \end{pmatrix}.$$

Lemma 3.3. *Let* $\mathbf{H} = \mathbf{G}\tilde{\Gamma}\mathbf{G}^{\mathrm{T}}$ (*where* $\mathbf{G}$ *is given by* (41)). *The following are equivalent*:

(i) $\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G}) \to \infty$,

(ii) $\lambda_{\max}(\mathbf{H}) \to \infty$

(iii) $\mathrm{tr}(\mathbf{H}) \to \infty$. (15)

Proof. (i) ⇔ (ii): Note $\lambda_{\min}(\tilde{\Gamma})\mathbf{G}\mathbf{G}^{\mathrm{T}} \leqslant \mathbf{H} \leqslant \lambda_{\max}(\tilde{\Gamma})\mathbf{G}\mathbf{G}^{\mathrm{T}}$ implies $\lambda_{\min}(\tilde{\Gamma})\lambda_{\max}(\mathbf{G}\mathbf{G}^{\mathrm{T}}) \leqslant \lambda_{\max}(\mathbf{H}) \leqslant \lambda_{\max}(\tilde{\Gamma})\lambda_{\max}(\mathbf{G}\mathbf{G}^{\mathrm{T}})$. The result then follows from the fact that the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are equal for any matrix $\mathbf{A}$. (ii) ⇔ (iii): $\mathbf{H}$ has at most $\tilde{p}$ non-zero eigenvalues and $\mathrm{tr}(\mathbf{H})$ equals the sum of its eigenvalues. □

With the foregoing preliminaries, we can now prove Theorem 2.1.

(i) *If model j is true*:

Clearly, $\alpha_j \to 1 \Leftrightarrow \alpha_k \to 0 \Leftrightarrow (-2\log M) \to -\infty$. It is sufficient to show that $D \to \infty \Leftrightarrow$ (15) holds while both $\mathbf{E}_j\Delta$ and $\mathbf{V}_j\Delta$ are bounded irrespective of whether (15) holds.

Observe that

$$\Sigma_k = \Sigma_j + \mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}\tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U} \tag{16}$$

implies $D = |\mathbf{I} + \mathbf{H}|$. Clearly, $\mathbf{I} + \mathbf{H}$ has at most $\tilde{p}$ eigenvalues greater than one and the remaining eigenvalues are equal to one. Since the determinant of a matrix equals the product of its eigenvalues, it follows that (1) $D \geqslant 1$, and (2) by Lemma 3.3 $D \to \infty \Leftrightarrow$ (15) holds.

Setting $i = j$ in (14) and simplifying gives

$$\mathbf{E}_j\Delta = \mathrm{tr}(\mathbf{I} - \Sigma_j^{-1}\Sigma_k) + \|\mu_o - \mu_j\|^2_{\Sigma_j^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_j\Sigma_k^{-1}}. \tag{17}$$

The first term in (17) is bounded since $\mathrm{tr}(\mathbf{I} - \Sigma_k^{-1}\Sigma_j) = \mathrm{tr}(\Sigma_k^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}\tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U}) = \mathrm{tr}(\tilde{\Gamma}^{1/2}\tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_k^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}^{1/2}) \leqslant \mathrm{tr}(\tilde{\Gamma}^{1/2}\tilde{\Gamma}^{-1}\tilde{\Gamma}^{1/2}) = \tilde{p}$ where the inequality follows from applying (13) after substituting for Σ_k using (16). The second term is bounded by Lemma 3.2 and the third term is bounded by Lemma 3.1 since $\Sigma_j \leqslant \Sigma_k$ implies $\|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_j^{-1}\Sigma_k^{-1}} \leqslant \|\mu_o - \mu_k\|_{\Sigma_k^{-1}}$.

Setting $i = j$ in (15) and simplifying gives

$$\mathbf{V}_j\Delta \leqslant 4[\mathrm{tr}[(\mathbf{I} - \Sigma_j^{-1}\Sigma_k)^2] + 2\|\mu_o - \mu_j\|^2_{(\Sigma_j^{-1} - \Sigma_k^{-1})\Sigma_j^{-1}(\Sigma_j - \Sigma_k^{-1})} + \|\mu_j - \mu_k\|^2_{\Sigma_k^{-1}\Sigma_j\Sigma_k^{-1}}].$$

The first term is bounded since $\mathrm{tr}(\mathbf{A}^2) \leqslant (\mathrm{tr}(\mathbf{A}))^2$. Since $(\Sigma_j^{-1} - \Sigma_k^{-1})\Sigma_j^{-1}(\Sigma_j - \Sigma_k^{-1}) = \Sigma_j^{-1/2}(\mathbf{I} - (\mathbf{I} + \mathbf{H})^{-1})\Sigma_j^{-1/2} \leqslant \Sigma_j^{-1}$, applying Lemma 3.2 shows that the second term is bounded. Finally, the third term is bounded by Lemma 3.1 and the fact $\Sigma_j \leqslant \Sigma_k$.

(ii) *If model k is true*:

(a) *Necessity*: Since $\alpha_k \to 1 \Leftrightarrow (-2\log M) \to \infty$, it is sufficient to show that both $\mathbf{E}_k(-2\log M)$ and $\mathbf{V}_k(-2\log M)$ are bounded when (15) fails to hold. This task reduces to showing that both $\mathbf{E}_k\Delta$ and $\mathbf{V}_k\Delta$ are bounded since it has already been seen that $\log(D)$ is bounded if (15) fails to hold. So suppose (15) does not hold.

Setting $i = k$ in (14) and simplifying gives

$$\mathbf{E}_k\Delta = \mathrm{tr}(\mathbf{H}) + \|\mu_o - \mu_j\|^2_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}} - \|\mu_o - \mu_k\|^2_{\Sigma_k^{-1}}. \tag{18}$$

The first term is bounded by supposition and the third term is bounded according to Lemma 3.1. To show that the second term is bounded, let $c = 1 + \lambda_{\max}(\mathbf{H})$ and observe that $\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1} = \Sigma_j^{-1/2}(\mathbf{I} + \mathbf{H})\Sigma_j^{-1/2} \leqslant c\Sigma_j^{-1}$. Then

$$\begin{aligned}
\|\mu_o - \mu_j\|_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}} &\leqslant c\|\mu_o - \mu_j\|_{\Sigma_j^{-1}} \\
&= c\|\mathbf{Z}_{j,(n)}(\beta_o^* - \mathbf{b}_j) + \tilde{\mathbf{Z}}_{k,(n)}\tilde{\beta}\|_{\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}} \\
&\leqslant c\|\mathbf{Z}_{j,(n)}(\beta_o^* - \mathbf{b}_j)\|_{\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}} + c\|\tilde{\mathbf{Z}}_{k,(n)}\tilde{\beta}\|_{\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}} \\
&= c\|\beta_o^* - \mathbf{b}_j\|_{\mathbf{Z}_{j,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{j,(n)}} + c\|\tilde{\beta}\|_{\mathbf{G}^{\mathrm{T}}\mathbf{G}} \\
&\leqslant c\|\beta_o^* - \mathbf{b}_j\|_{\Gamma_j^{-1}} + c\|\tilde{\beta}\|_{\mathbf{G}^{\mathrm{T}}\mathbf{G}}.
\end{aligned} \tag{19}$$

The second inequality follows from Cauchy–Schwarz and the third one from an application of (13). Both of the terms in (19) are bounded by supposition.

Setting $i=k$ in (15) and simplifying gives

$$\begin{aligned}\mathbf{V}_k\Delta &\leqslant 4\left[\mathrm{tr}(\mathbf{H}^2)+2\|\mu_o-\mu_j\|^2_{(\Sigma_j^{-1}-\Sigma_k^{-1})\Sigma_k(\Sigma_j^{-1}-\Sigma_k^{-1})}+\|\mu_j-\mu_k\|^2_{\Sigma_k^{-1}}\right]\\ &\leqslant 4\left[\mathrm{tr}(\mathbf{H}^2)+2\|\mu_o-\mu_j\|_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}}+\|\mu_j-\mu_k\|^2_{\Sigma_k^{-1}}\right],\end{aligned}\tag{20}$$

where the last inequality follows from the fact $(\Sigma_j^{-1}-\Sigma_k^{-1})\Sigma_k(\Sigma_j^{-1}-\Sigma_k^{-1})\leqslant\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}$. All of the terms in (20) are bounded by supposition or have already have been shown to be bounded.

(b) *Sufficiency*: It is sufficient to show (15) and (16) imply $\mathbf{E}_k(-2\log M)\to\infty$ and $\mathbf{V}_k(-2\log M)/\mathbf{E}_k^2(-2\log M)\to 0$ since Chebyshev's inequality implies $-2\log M\to\infty$.

Clearly, from (18),

$$\begin{aligned}\mathbf{E}_k(-2\log M)&=\mathrm{tr}(\mathbf{H})+\|\mu_o-\mu_j\|^2_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}}-\|\mu_o-\mu_k\|^2_{\Sigma_k^{-1}}-\log|\mathbf{I}+\mathbf{H}|\\ &\geqslant\|\mu_o-\mu_j\|^2_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}}-\|\mu_o-\mu_k\|^2_{\Sigma_k^{-1}}\end{aligned}\tag{21}$$

since $\mathrm{tr}(\mathbf{H})-\log|\mathbf{I}+\mathbf{H}|=\sum(\lambda_i(\mathbf{H})-\log(1+\lambda_i(\mathbf{H}))\geqslant 0$. By Lemma 3.1, $\|\mu_o-\mu_k\|^2_{\Sigma_k^{-1}}$ is bounded so can be ignored. We now show that $\|\mu_o-\mu_j\|^2_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}}$ increases at a rate of at least $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2))$. Since $\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}=\Sigma_j^{-1}+\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}\tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1}$, we have

$$\begin{aligned}\|\mu_o-\mu_j\|^2_{\Sigma_j^{-1}\Sigma_k\Sigma_j^{-1}}&\geqslant\|\mu_o-\mu_j\|^2_{\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}\tilde{\mathbf{Z}}_{k,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1}}\\ &=\|\beta_o^*\|^2_{\mathbf{Z}_{j,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1/2}\mathbf{H}\Sigma_j^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{j,(n)}}+\|\tilde{\beta}\|^2_{\mathbf{G}^{\mathrm{T}}\mathbf{G}\tilde{\Gamma}\mathbf{G}^{\mathrm{T}}\mathbf{G}}\\ &\quad-2\beta_o^{*T}\mathbf{Z}_{j,(n)}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{Z}}_{k,(n)}\tilde{\Gamma}\mathbf{G}^{\mathrm{T}}\mathbf{G}\tilde{\beta}.\end{aligned}\tag{22}$$

The second term in (22) is at least $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2))$ since

$$\begin{aligned}\|\tilde{\beta}\|^2_{\mathbf{G}^{\mathrm{T}}\mathbf{G}\tilde{\Gamma}\mathbf{G}^{\mathrm{T}}\mathbf{G}}&\geqslant\lambda_{\min}(\tilde{\Gamma})\|\tilde{\beta}\|^2_{(\mathbf{G}^{\mathrm{T}}\mathbf{G})^2}\\ &\geqslant\lambda_{\min}(\tilde{\Gamma})\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2)\|\tilde{\beta}\|^2.\end{aligned}$$

In contrast, the first term in (22) increases at a rate of at most $\mathrm{O}(\lambda_{\min}(\mathbf{G}^{\mathrm{T}}\mathbf{G}))$ since

$$\begin{aligned}\|\beta_o^*\|^2_{\mathbf{Z}_{k,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1/2}\mathbf{H}\Sigma_j^{-1/2}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{k,(n)}}&\leqslant\lambda_{\max}(\mathbf{G}\tilde{\Gamma}\mathbf{G}^{\mathrm{T}})\|\beta_o^*\|^2_{\mathbf{Z}_{j,(n)}^{\mathrm{T}}\mathbf{U}\Sigma_j^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Z}_{j,(n)}}\\ &\leqslant\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G})\lambda_{\max}(\tilde{\Gamma})\|\beta_o^*\|^2_{\Gamma_j^{-1}}\\ &\leqslant R\lambda_{\min}(\mathbf{G}^{\mathrm{T}}\mathbf{G})\lambda_{\max}(\tilde{\Gamma})\|\beta_o^*\|^2_{\Gamma_j^{-1}},\end{aligned}$$

where condition (6) has been used in the last inequality. Hence, the second term in (22) dominates the first term. By the Cauchy–Schwarz inequality, the first term also dominates the third term. Therefore $\mathbf{E}_k(-2\log M)\to\infty$ at a rate of at least $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2))$.

To show $\mathbf{V}_k(-2\log M)$ increases at a rate of at most $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2)$, it is sufficient to show that the first term in (20) is at most $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2)$ since it has already been shown that the third term in (20) is bounded and the second term is at most $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2))$. But $\mathrm{tr}(\mathbf{H}^2)\leqslant(\mathrm{tr}(\mathbf{H}))^2\leqslant(\lambda_{\max}(\tilde{\Gamma})\,\mathrm{tr}(\mathbf{G}^{\mathrm{T}}\mathbf{G}))^2\leqslant(\lambda_{\max}(\tilde{\Gamma})\tilde{p}\lambda_{\max}(\mathbf{G}^{\mathrm{T}}\mathbf{G}))^2\leqslant(\lambda_{\max}(\tilde{\Gamma})\tilde{p}R\lambda_{\min}(\mathbf{G}^{\mathrm{T}}\mathbf{G}))^2$ is $\mathrm{O}(\lambda_{\min}((\mathbf{G}^{\mathrm{T}}\mathbf{G})^2))$. □

Note that the proof yields also: (i) when model j is true, M is bounded away from 0 in probability (since Δ is bounded and $D \geqslant 1$), and (ii) when model k is true, M is bounded away from ∞ in probability (since by (21) the mean of $-2\log M$ is bounded away from $-\infty$ and the density of $-2\log M$ is unimodal). That is, the true model can never be discredited by an incorrect model. These properties are used in the proof of the first Corollary.

References

Clyde, M.A., 1999. Bayesian model averaging and model search strategies. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, Vol. 6. Oxford University Press, Oxford, pp. 157–185.

Cover, T., Thomas, J., 1991. Elements of Information Theory. Wiley, New York.

Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Amer. Statist. Assoc. 89, 1535–1546.

Wong, H., 2000. Small sample improvement over Bayes prediction under model uncertainty. Ph.D. Thesis, Department of Statistics, University of British Columbia.

Wong, H., Clarke, B., 2004. Improvement over Bayes prediction in small samples in the presence of model uncertainty. Can. J. Stat., to be published.