Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 25–49
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Approximating posterior probabilities in a linear model with possibly noninvertible moving average errors

Suriani Pokta[a], Jeffrey D. Hart[b,*]

[a]*Allergan, Inc. P.O. Box 19534 Irvine, CA 92623, USA*
[b]*Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA*

Received 1 February 2006
Available online 18 April 2007

Abstract

The method of Laplace is used to approximate posterior probabilities for a collection of polynomial regression models when the errors follow a process with a noninvertible moving average component. These results are useful in the problem of period-change analysis of variable stars and in assessing the posterior probability that a time series with trend has been overdifferenced. The nonstandard covariance structure induced by a noninvertible moving average process can invalidate the standard Laplace method. A number of analytical tools is used to produce corrected Laplace approximations. These tools include viewing the covariance matrix of the observations as tending to a differential operator. The use of such an operator and its Green's function provides a convenient and systematic method of asymptotically inverting the covariance matrix.

In certain cases there are two different Laplace approximations, and the appropriate one to use depends upon unknown parameters. This problem is dealt with by using a weighted geometric mean of the candidate approximations, where the weights are completely data-based and such that, asymptotically, the correct approximation is used. The new methodology is applied to an analysis of the prototypical long-period variable star known as Mira.

AMS 1991 subject classification: 62C10; 62G07; 62G20; 62M10

Keywords: BIC; Information matrix; Likelihood analysis; Model selection; Overdifferencing; Noninvertible moving average process

* Corresponding author. Fax: +1 979 845 3144.
E-mail address: hart@stat.tamu.edu (J.D. Hart).

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.04.004

1. Introduction

This paper develops Laplace approximations to the posterior probabilities of polynomial regression models with errors equal to the sum of independent processes, one of which is stationary autoregressive and the other a noninvertible moving average. This error structure finds application in at least two problems of scientific importance. First of all, it arises in the problem of period change analysis of variable stars, where it was essentially proposed by Eddington and Plakidis [1] and Sterne [23], and refined by Lombard [14]. In this context, the autoregressive part of the model corresponds to random variation intrinsic to times between successive maximum (or minimum) brightnesses of a star, and the noninvertible moving average arises from differences between errors made in recording times of maximum (or minimum) brightness. The application of statistical methods to testing for period changes in variable stars has a long history in the astronomy literature; see, for example, Sterne and Campbell [24], Isles and Saw [4], Lombard [14], Percy and Colivas [18], Koen and Lombard [11,12] and Hart et al. [3]. To our knowledge, only frequentist-type tests have been used to detect period changes in variable stars. Results in the current paper provide an apparatus for performing Bayesian tests of period change.

The current method of choice for approximating posterior probabilities in analytically intractable Bayesian models is Markov chain Monte Carlo (MCMC). However, the setting of Koen and Lombard [12] and Hart et al. [3] provides an example of when an alternative method, such as that of Laplace, is desirable. In these papers, a period change test is applied to each of over 375 variable stars. Using MCMC methods in such a setting would be extremely time consuming. Typically, some human intervention is required to insure that the MCMC output is mixing adequately and/or that adequate burn-in time has been achieved (Gilks et al., [2]). Carrying out this exercise for *hundreds* of different data sets is impractical at best.

A Bayesian test of no systematic change in periods may be conducted by determining if the posterior probability of models with polynomial degree higher than 0 is sufficiently large. As such, our results have implications on regression model selection when the errors have a noninvertible moving average component. A widely used criterion, first proposed by Schwarz [22], for model selection is BIC. Schwarz [22] showed that for certain exponential family models his criterion approximates the log of the posterior probability of each model. Therefore, BIC approximately corresponds to the Bayesian procedure of selecting the model with highest posterior probability. BIC has been extended and studied by a number of authors. Kashyap [7] noted that Schwarz's approximation to a posterior probability can be viewed as a special case of Laplace's method and gives a more accurate approximation by including more terms in the expansion. Kass and Wasserman [10] point out that BIC is more directly related to the log of the Bayes factor than to the log of the posterior probabilities. They showed that for a particular class of reference priors the log of the Bayes factor is approximated by BIC with error of order $O_p(n^{-1/2})$ instead of the more typical $O_p(1)$.

Laplace's [13] method provides an analytical approximation to integrals that take a particular form. A recent review of asymptotic expansion of integrals including Laplace's method is given by Olver [16]. Practitioners often provide no justification for the validity of Laplace's approximation to a posterior integral. Moreover, the necessary regularity conditions, such as those derived in Kass et al. [8] and Johnson [6], are typically derived for i.i.d. observations. The data of our model are neither independent nor identically distributed, and hence the validity of the usual approximation is in question. We will show that BIC and the usual first order Laplace approximation [9,20] do not always provide a good approximation to the log of the Bayes factor for our models. In those

cases where the usual approximation breaks down, we develop a modified Laplace approximation that is asymptotically more accurate.

A second setting where noninvertible moving averages arise is in time series analysis when a series has been over-differenced. Data differencing has long been a tool in econometrics for inducing stationarity of an error series [19,21]. Suppose that the errors of an observed time series follow an ARIMA process with nonseasonal differencing order d, which may be unknown. If the data are differenced $d+1$ times, i.e., they are overdifferenced, the resulting error series has a noninvertible moving average component. Tsay [25] provides two other reasons why noninvertible moving averages are important, and proposes frequentist tests of the hypothesis that the data have a noninvertible component. The results in this paper provide a general and computationally simple Bayesian alternative to such tests.

The remainder of the paper proceeds as follows. In Section 2 our model is defined and the problem of interest stated. Section 3 is devoted to an analysis of the likelihood for this model and the development of Laplace approximations to posterior probabilities. An important part of this section is analyzing the asymptotic behavior of information matrices as the sample size tends to infinity. This involves viewing these matrices as tending to differential operators. The use of differential operators and their Green's functions provides a convenient and systematic method to asymptotically invert information matrices. We also describe in Section 3 how our results can be used to assess the probability of overdifferencing. In Section 4 we present an analysis of data from the long-period variable star Mira, which is one of the 392 data sets of Koen and Lombard [12]. It is shown that our modified Laplace approximation is superior to a BIC approximation of posterior probabilities.

2. Model formulation and definitions

Given observations $Y_1, \ldots, Y_n$ at evenly spaced time points $1, \ldots, n$, consider the model

$$Y_j = \mu_m(j) + I_j + \epsilon_j - \epsilon_{j-1}, \quad j = 1, \ldots, n, \tag{2.1}$$

where μ_m is an mth degree polynomial accounting for systematic variation in the observations, and $\{I_j : j = 1, \ldots, n\}$ and $\{\epsilon_j : j = 0, \ldots, n\}$ are independent, mean 0 error processes. It is assumed that the I_j's follow a first order autoregressive, AR(1), process, i.e.,

$$I_j = \rho I_{j-1} + Z_j, \quad j = 2, \ldots, n,$$

where $|\rho| < 1$ and $Z_2, \ldots, Z_n$ are i.i.d. normal random variables with mean 0 and finite variance σ_Z^2. The variance of I_j is denoted σ_I^2 and equals $\sigma_Z^2/(1-\rho^2)$.

The ϵ_j's are assumed to be independent normal random variables with mean 0 and finite variance, and are allowed to be heteroscedastic in the following way:

$$\mathrm{Var}(\epsilon_j) = v\left(x_j; \boldsymbol{\beta}\right) = \exp[2(\beta_0 + \beta_1 x_j)], \quad j = 1, \ldots, n, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $x_j = j/n$, $j = 1, \ldots, n$.

Remarks about model (2.1):

1. The most general version of model (2.1) is motivated by the period-change problem discussed in Section 1. In that setting, each Y_j is the observed length of time between successive maximum brightnesses of a given variable star, μ_m accounts for systematic variation in these times, $I_1, \ldots, I_n$ are errors intrinsic to the star, and $\epsilon_0, \ldots, \epsilon_n$ are errors made in measuring the times of maximum brightness. Of interest is testing whether or not there is systematic variation in

the times between maximum brightness, which is equivalent to testing whether or not the polynomial degree m is 0.

2. The heteroscedastic error model (2.2) is motivated by the analysis in Hart et al. [3], where it is noted that residual variance for most stars tends to decrease monotonically over time. This is consistent with the fact that methods of measuring times of maximum brightness have improved over the time period in which the data have been observed.
3. The connection of model (2.1) to the overdifferencing problem may be described as follows. Suppose one observes a time series $U_j = r(j) + \gamma_j$, $j = 1, \ldots, n + d + 1$, where r is a polynomial of degree $m + d + 1$ and $\{\gamma_j : j = 1, 2, \ldots\}$ is a Gaussian $ARIMA(0, d, 0)$ process [15]. If the data $U_1, \ldots, U_{n+d+1}$ are differenced $d + 1$ times, the result is a series of n observations identical in distribution to those of model (2.1) with $\sigma_I^2 = 0$ and $\beta_1 = 0$. If the U_j's are differenced d times, the resulting errors are i.i.d. Gaussian. As will be shown in Section 3.7, these two facts entail that our methodology can be used to approximate the posterior probability that a series with $ARIMA(0, d, 0)$ errors has been overdifferenced.

Application of our results to the overdifferencing problem will be discussed in Section 3.7. Until that point, all our discussion pertains to Laplace approximations in the period-change context of Remark 1.

In the period-change problem, the case $\sigma_I^2 = 0$ is of crucial importance since it necessitates modified Laplace approximations. Deciding whether the error term I_j is present or absent will be a part of the model selection process. A model will be described by a pair $M = (m, h)$ where m is the degree of the fitted polynomial and h is a binary variable such that

$$h = \begin{cases} 0 & \text{if } \sigma_I^2 \text{ is assumed to be 0,} \\ 1 & \text{if } \sigma_I^2 \text{ is assumed to be positive.} \end{cases}$$

The observations $Y_1, \ldots, Y_n$ are distributed multivariate normal with means

$$E(Y_j) = \theta_0 + \theta_1 \frac{j}{n} + \cdots + \theta_m \left(\frac{j}{n}\right)^m, \quad j = 1, \ldots, n. \tag{2.3}$$

We will consider models for which $0 \leqslant m \leqslant m_{\max}$. Let Θ_m denote the parameter space of $\boldsymbol{\theta}_m = (\theta_0, \ldots, \theta_m)$ for the degree m model. The covariance matrix $\boldsymbol{\Sigma}$ of $Y_1, \ldots, Y_n$ is given by

$$\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma_Z^2/(1-\rho^2) + v(x_j; \boldsymbol{\beta}) + v(x_{j-1}; \boldsymbol{\beta}), & i = j, \\ \rho\sigma_Z^2/(1-\rho^2) - v(\min(x_i, x_j); \boldsymbol{\beta}), & |i-j| = 1, \\ \rho^{|i-j|}\sigma_Z^2/(1-\rho^2), & |i-j| > 1, \end{cases} \tag{2.4}$$

where the function v is defined by (2.2). Let $\boldsymbol{\eta}$ denote the covariance parameters for the model, i.e., $\boldsymbol{\eta} = \boldsymbol{\beta}$ if $h = 0$ and $\boldsymbol{\eta} = (\sigma_I^2, \rho, \boldsymbol{\beta})$ if $h = 1$. Let Ω_h denote the parameter space for $\boldsymbol{\eta}$ when the model indicator is h. For brevity we will omit the subscript h whenever it is clear from the context.

The likelihood is

$$f(\mathbf{y}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} (\det(\boldsymbol{\Sigma}))^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_m)' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}_m)\right),$$

where the elements of $\boldsymbol{\mu}_m = \boldsymbol{\mu}_m(\boldsymbol{\theta}_m)$ are defined by (2.3) and of $\boldsymbol{\Sigma}$ by (2.4).

Let α_M denote the prior probability of model M. We assume that the mean parameters $\boldsymbol{\theta}_m$ and the covariance parameters $\boldsymbol{\eta}$ are *a priori* independent. Then the prior has the form

$$\pi(\boldsymbol{\theta}_m, \boldsymbol{\eta}|M) = \pi_m(\boldsymbol{\theta}_m)\pi_h(\boldsymbol{\eta}).$$

Let $Z(\boldsymbol{y})$ be the marginal density of $\boldsymbol{Y}$. The posterior probability of model M given the data is then

$$\pi(M|\boldsymbol{y}) = \frac{\alpha_M}{Z(\boldsymbol{y})} \int_{\Omega_h} \int_{\Theta_m} f(\boldsymbol{y}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma})\pi_m(\boldsymbol{\theta}_m)\pi_h(\boldsymbol{\eta})\, d\boldsymbol{\theta}_m\, d\boldsymbol{\eta}.$$

In general, it is not possible to evaluate this integral exactly, and hence we consider a Laplace approximation in the next section.

3. Approximation of posterior probabilities using Laplace's method

Throughout Section 3 it is assumed that the model, (m, h), whose posterior probability we are computing is such that $m \geqslant m_0$, where m_0 is the polynomial degree of the *true* model. This case will suffice since, as argued by Kass and Vaidyanathan [9], the posterior probability of a model with $m < m_0$ is exponentially small (asymptotically) in comparison to ones with $m \geqslant m_0$. It follows that any of the approximations we consider will be extremely small for $m < m_0$.

We may express $\boldsymbol{\mu}_m$ as $\boldsymbol{X}\boldsymbol{\theta}_m$, where $\boldsymbol{X}$ is an $n \times (m+1)$ design matrix and we have suppressed the dependence of $\boldsymbol{X}$ on m and n. The posterior probability of model M given the data is

$$\pi(M|\boldsymbol{y}) = \frac{\alpha_M}{Z(\boldsymbol{y})(2\pi)^{n/2}} \int_{\Omega_h} (\det(\boldsymbol{\Sigma}))^{-1/2}\pi_h(\boldsymbol{\eta}) \times \int_{\Theta_m} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_m)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_m)\right)\pi_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m\, d\boldsymbol{\eta}.$$

Since our data $\boldsymbol{Y}$ are normally distributed, the parameters $\boldsymbol{\theta}_m$ which only influence the mean of $\boldsymbol{Y}$ and the parameters $\boldsymbol{\eta}$ which only influence the covariance matrix of $\boldsymbol{Y}$ are orthogonal. Let $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\theta}}_m$ denote the MLEs for these quantities for model M. Note that

$$\hat{\boldsymbol{\theta}}_m = (\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y},$$

which is also a generalized least-squares estimator of $\boldsymbol{\theta}_m$. The information matrices for $\boldsymbol{\eta}$ and $\boldsymbol{\theta}_m$ are, respectively,

$$\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}} = \left(\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\eta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial\boldsymbol{\Sigma}}{\partial\eta_j}\right)\right)$$

and

$$\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}.$$

If the hypotheses necessary for the Laplace approximation hold, then the resulting approximation to $Z(\boldsymbol{y})\pi(M|\boldsymbol{y})$ is

$$\frac{\alpha_M}{(2\pi)^{n/2}}\left(\det\hat{\boldsymbol{\Sigma}}\right)^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right) \times \pi_m(\hat{\boldsymbol{\theta}}_m)\pi_h(\hat{\boldsymbol{\eta}})(2\pi)^{(m+3+2h)/2}\left(\det\boldsymbol{I}_{\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\eta}}}\right)^{-1/2}\left(\det\boldsymbol{I}_{\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}}\right)^{-1/2}\left(1 + O_p(1/n)\right). \quad (3.1)$$

We note that $Z(\boldsymbol{y})$ is the actual marginal distribution of $\boldsymbol{Y}$, which depends upon the integrals that we are approximating. Once we compute our approximations to $Z(\boldsymbol{y})\pi(M|\boldsymbol{y})$, they may be summed over M to obtain $\hat{Z}(\boldsymbol{y})$, an approximation to $Z(\boldsymbol{y})$.

There are technical conditions which must be met for the Laplace approximation in (3.1) to be valid. In particular, it is necessary that the eigenvalues of $\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}}$ and $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ tend to infinity as n tends to infinity. It will be seen that the asymptotic formulas for the posterior probability will depend on the asymptotic form of these matrices.

Laplace's method was recently applied to variance component models by Pauler et al. [17]. Their work dealt with the situation that causes difficulties in our model, namely that a variance component can be 0. However, we are unable to apply their technique to our model since we cannot assume that the cubic term in our asymptotic expansion is negligible at the boundary.

3.1. Information matrix of $\boldsymbol{\theta}$

To study the asymptotic behavior of $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}$, we take the following approach. Any vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$, such as a column of the design matrix $\boldsymbol{X}$, can be viewed as a step function on $(0, 1]$ by identifying $\boldsymbol{v}$ with the function $f_{\boldsymbol{v}}(t) = v_i$ for $(i-1)/n < t \leqslant i/n$, $i = 1, \ldots, n$. The dot product of two vectors is then interpreted as integration via the formula

$$\boldsymbol{v} \cdot \boldsymbol{w} = n \int_0^1 f_{\boldsymbol{v}}(t) f_{\boldsymbol{w}}(t)\, dt.$$

This interpretation makes it possible to identify the limit of a sequence of vectors of length n as n tends to infinity with a function on $[0, 1]$. For example, for our design matrix $\boldsymbol{X}$ the ith column is $\boldsymbol{X}_i = ((1/n)^i, (2/n)^i, \ldots, ((n-1)/n)^i, (n/n)^i)'$ and $\lim_{n\to\infty} f_{\boldsymbol{X}_i}(t) = t^i$. For piecewise smooth regression models, including polynomials and Fourier series, the columns of the design matrix have a nice limiting behavior.

Similarly if $\boldsymbol{A} = (a_{i,j})$ is an $n \times n$ matrix, then we can interpret $\boldsymbol{A}$ as a piecewise constant function $a(s, t)$ on $(0, 1] \times (0, 1]$ by setting $a(s, t) = a_{i,j}$ for $(i-1)/n < s \leqslant i/n$ and $(j-1)/n < t \leqslant j/n$. Hence, matrix multiplication becomes integration as well. Specifically, if the matrix $\boldsymbol{A}$ corresponds to the function $a(s, t)$, the matrix $\boldsymbol{B}$ corresponds to $b(s, t)$, and the vector $\boldsymbol{v}$ corresponds to the function $f_{\boldsymbol{v}}(t)$, then the vector $\boldsymbol{Av}$ corresponds to

$$f_{\boldsymbol{Av}}(s) = n \int_0^1 a(s, t) f_{\boldsymbol{v}}(t)\, dt$$

and the matrix $\boldsymbol{AB}$ to the function

$$n \int_0^1 a(s, \tau) b(\tau, t)\, d\tau.$$

Unfortunately, for most of the covariance matrices we are interested in, taking this limit will require rescaling by a power of n and interpreting the limit as a distribution on $[0, 1] \times [0, 1]$ rather than a function. For example, the $n \times n$ identity matrix $\boldsymbol{I}_n$ corresponds to the function $i_n(s, t)$ which is 1 if $s, t \in ((i-1)/n, i/n]$ for some i and zero otherwise. Hence, $\lim_{n\to\infty} n i_n(s, t) = \delta(s - t)$ where δ denotes the Kronecker delta function. This will be abbreviated to $\boldsymbol{I}_n = \delta(s - t)/n + O(n^{-2})$.

The covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{A} + \boldsymbol{B}$ can be broken into the two parts $\boldsymbol{A}$ and $\boldsymbol{B}$. Here $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the parts of $\boldsymbol{\Sigma}$ coming from the ϵ_j's and I_j's, respectively. Specifically, we have

$\boldsymbol{A} = (a_{i,j})$ where

$$a_{i,j} = \begin{cases} v(x_j;\boldsymbol{\beta}) + v(x_{j-1};\boldsymbol{\beta}), & i = j, \\ -v(\min(x_i, x_j);\boldsymbol{\beta}), & |i-j| = 1, \\ 0, & |i-j| > 1. \end{cases} \tag{3.2}$$

The elements $(\boldsymbol{A}^{-1})_{i,j}$ of $\boldsymbol{A}^{-1}$ are given explicitly by

$$\frac{e^{\beta_1/n - 2\beta_0}}{e^{\beta_1/n} - e^{-\beta_1/n}} \times \frac{e^{2(1+1/n-\max(x_i,x_j))\beta_1} - e^{2(1+1/n-x_i-x_j)\beta_1} - 1 + e^{-2\min(x_i,x_j)\beta_1}}{e^{2(1+1/n)\beta_1} - 1}.$$

A correct expression for this quantity when $\beta_1 = 0$ can be obtained by using L'Hôpital's rule.

The parameter β_0 that determines the variance of the first measurement error ϵ_0, should not depend on the number of observations n. Furthermore, β_1 is assumed not to depend on n since otherwise the variances would change dramatically between the first and last observations and hence only a small fraction of the data would actually contribute to our parameter estimates. This scaling seems to be borne out by the data.

It is easily shown that as $n \to \infty$, $(1/n)\boldsymbol{A}^{-1}$ converges to the function

$$g(s,t) = \frac{e^{-2\beta_0}}{2b(e^{2b}-1)} \left(e^{2b(1-\max(s,t))} - e^{2b(1-s-t)} - 1 + e^{-2b\min(s,t)} \right). \tag{3.3}$$

Alternatively, (3.3) can be derived without explicitly inverting $\boldsymbol{A}$ using techniques that would be helpful for a large number of covariance structures. Consider multiplying the matrices $\boldsymbol{A}$ by a sequence of vectors $\boldsymbol{v}$ which converge to the smooth function $f_{\boldsymbol{v}}(t)$. Then, ignoring boundary effects or assuming $f_{\boldsymbol{v}}(0) = f_{\boldsymbol{v}}(1) = 0$, we compute

$$\lim_{n\to\infty} n^2 f_{\boldsymbol{A}\boldsymbol{v}}(t) = -e^{2\beta_0} \frac{d}{dt} \left(e^{2bt} \frac{df_{\boldsymbol{v}}(t)}{dt} \right). \tag{3.4}$$

Thus $\lim_{n\to\infty} n^2\boldsymbol{A}$ can be interpreted as a differential operator. The inverse to a differential operator is the corresponding Green's function. Specifically, suppose we have a sequence of vectors $\boldsymbol{w}$ converging to $f_{\boldsymbol{w}}(t)$. Since $(1/n)\boldsymbol{A}^{-1}$ converges to $g(s,t)$, $(1/n^2)\boldsymbol{A}^{-1}\boldsymbol{w}$ converges to

$$h(t) = \int_0^1 g(t,\tau) f_{\boldsymbol{w}}(\tau)\, d\tau. \tag{3.5}$$

Hence by the identification of $\boldsymbol{A}$ with a differential operator in (3.4), we see that $n^2\boldsymbol{A}(1/n^2)\boldsymbol{A}^{-1}\boldsymbol{w} = \boldsymbol{w}$ will converge to

$$-e^{2\beta_0} \frac{d}{dt} \left(e^{2bt} \frac{dh(t)}{dt} \right) = f_{\boldsymbol{w}}(t). \tag{3.6}$$

Combining (3.5) and (3.6) gives

$$-e^{2\beta_0} \frac{\partial}{\partial s} \left(e^{2bs} \frac{\partial g(s,t)}{\partial s} \right) = \delta(s-t).$$

Thus, $g(s,t)$ is Green's function for the differential operator corresponding to $n^2\boldsymbol{A}$. Conversely we could have used this method to find the asymptotic behavior of $\boldsymbol{A}^{-1}$. We first identify $n^2\boldsymbol{A}$

with the differential operator using (3.4), then directly compute Green's function $g(s,t)$ for this differential operator on [0, 1] with the boundary conditions $g(0,t) = g(1,t) = 0$. Thus it follows that $(1/n)\boldsymbol{A}^{-1}$ converges to this Green's function.

The second part $\boldsymbol{B} = (b_{i,j})$ of the covariance matrix $\boldsymbol{\Sigma}$ is given by

$$b_{i,j} = \sigma_I^2 \rho^{|i-j|}.$$

Based on the observed data it appears most reasonable to assume that ρ does not vary with n, and therefore the absolute values of entries of $\boldsymbol{B}$ decrease rapidly as they move away from the diagonal. For vectors that tend to smooth functions (the only type we need to consider), entries near the diagonal have almost the same effect as diagonal entries (with errors of order n^{-1}). Ignoring boundary effects that are also $O(1/n)$, the row sums of $\boldsymbol{B}$ are

$$\sum_{j=-\infty}^{\infty} \rho^{|j|} = \frac{1+\rho}{1-\rho}.$$

Suppose the vector $\boldsymbol{v}$ represents a smooth function f in the sense that $\boldsymbol{v}' = (f(1/n), f(2/n), \ldots, f(1))$. Then we have

$$\left\| \left(\boldsymbol{B} - \sigma_I^2 \frac{1+\rho}{1-\rho} \boldsymbol{I}_n \right) \boldsymbol{v} \right\| \leqslant O(\|\boldsymbol{v}\|/n),$$

which will be abbreviated as

$$\boldsymbol{B} = \sigma_I^2 \frac{1+\rho}{1-\rho} \boldsymbol{I}_n (1 + O(1/n)). \tag{3.7}$$

Hence if $\sigma_I^2 > 0$,

$$\boldsymbol{B}^{-1} = \sigma_I^{-2} \frac{1-\rho}{1+\rho} \boldsymbol{I}_n (1 + O(1/n)). \tag{3.8}$$

When considering $\boldsymbol{I}_{\theta,\theta} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, a natural measure of the size of a matrix $\boldsymbol{A}$ is the matrix norm $\max_{\{\boldsymbol{v}:\|\boldsymbol{v}\|\neq 0\}} \|\boldsymbol{A}\boldsymbol{v}\|/\|\boldsymbol{v}\|$, where the maximum is taken over vectors $\boldsymbol{v}$ which represent smooth functions. Thus the matrix $\boldsymbol{A}$ of (3.2) has size $O(n^{-2})$, $\boldsymbol{A}^{-1}$ has size $O(n^2)$, and provided $\sigma_I^2 > 0$, $\boldsymbol{B}$ and $\boldsymbol{B}^{-1}$ have size $O(1)$. Thus if $\sigma_I^2 > 0$, $\boldsymbol{A}$ is much smaller than $\boldsymbol{B}$ and $\boldsymbol{\Sigma} \approx \boldsymbol{B} = O(1)$, but if $\sigma_I^2 = 0$, then $\boldsymbol{\Sigma} = \boldsymbol{A} = O(n^{-2})$. This difference in scales results in differences in the Laplace approximations.

First suppose $\sigma_I^2 > 0$. The columns of the $\boldsymbol{X}$ matrix converge to functions on [0, 1]; therefore $\boldsymbol{X}$ converges to a row vector of functions:

$$\lim_{n\to\infty} \boldsymbol{X} = (f_0(t)\ f_1(t)\ \cdots\ f_m(t)).$$

Since

$$\boldsymbol{\Sigma} = \boldsymbol{B} + O(n^{-2}) = \sigma_I^2 \frac{1+\rho}{1-\rho} \boldsymbol{I}_n (1 + O(1/n)),$$

we have

$$\lim_{n\to\infty} n\boldsymbol{\Sigma}^{-1} = \sigma_I^{-2} \frac{1-\rho}{1+\rho} \delta(s-t)$$

and hence for $0 \leqslant i, j \leqslant m$

$$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})_{i,j} \sim n^2 \int_0^1 \int_0^1 f_i(s) n^{-1} \sigma_I^{-2} \frac{1-\rho}{1+\rho} \delta(s-t) f_j(t)\, ds\, dt$$

$$= n\sigma_I^{-2} \frac{1-\rho}{1+\rho} \int_0^1 f_i(t) f_j(t)\, dt. \tag{3.9}$$

If the functions f_i are linearly independent on $[0, 1]$, as they are in any nonredundant regression model, then the matrix with (i, j) entry $\int_0^1 f_i(t) f_j(t) dt$ is positive definite. It follows from (3.9) that all the eigenvalues of $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$ will be large for large n, as required. Furthermore,

$$\log \det(\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}) = (m+1)\log(n) + O(1),$$

which is consistent with the standard BIC formula. For our specific case of $f_i(t) = t^i$, we have

$$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})_{i,j} \sim n\sigma_I^{-2} \frac{1-\rho}{1+\rho} \cdot \frac{1}{i+j+1}$$

and hence

$$\det(\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}})^{-1/2} \approx n^{-(m+1)/2} \left(\frac{\sigma_I^2(1+\rho)}{1-\rho} \right)^{(m+1)/2} \prod_{i=1}^{m} (2i+1)^{1/2} \binom{2i}{i}.$$

Next suppose $\sigma_I^2 = 0$, in which case $\boldsymbol{\Sigma} = \boldsymbol{A}$. The $\boldsymbol{X}$ matrix is exactly as in the previous case but

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{-1} \sim n g(s, t),$$

where Green's function $g(s, t)$ is given in (3.3). Hence for $0 \leqslant i, j \leqslant m$ we have

$$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})_{i,j} \sim n^3 \int_0^1 \int_0^1 f_i(s) g(s, t) f_j(t)\, ds\, dt.$$

The matrices $\boldsymbol{L}_m = (\ell_{i,j})_{0 \leqslant i,j \leqslant m}$ with

$$\ell_{i,j} = \int_0^1 \int_0^1 s^i g(s, t) t^j\, ds\, dt$$

are positive definite. To see this, let $p(t)$ be a nonzero polynomial and let $q(t) = \int_0^1 g(s, t) p(s)\, ds$ be the unique solution to $-e^{2\beta_0} \frac{d}{dt}(e^{2bt} q'(t)) = p(t)$ with $q(0) = q(1) = 0$. Then $q(t)$ is not constant and hence

$$\int_0^1 \int_0^1 p(s) g(s, t) p(t)\, ds\, dt = \int_0^1 q(t) p(t)\, dt$$

$$= -e^{2\beta_0} \int_0^1 q(t) \frac{d}{dt} e^{2bt} q'(t)\, \mathrm{d}t$$

$$= e^{2\beta_0} \int_0^1 e^{2bt} (q'(t))^2\, \mathrm{d}t > 0.$$

It follows that all the eigenvalues of $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ tend to infinity (but like n^3) as n tends to infinity. Hence

$$\log \det(\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}) = 3(m+1)\log n + \log \det(\boldsymbol{L}_m) + O(1/n).$$

Importantly, this differs from the standard BIC formula. Because of the negative correlation between adjacent observations, the data contain more information about the regression coefficients $\boldsymbol{\theta}_m$ than one might naively expect.

3.2. Information matrix of $\boldsymbol{\eta}$

Next we consider the asymptotic behavior of $\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}}$. When the covariance matrix $\boldsymbol{\Sigma}$ of a multivariate normal distribution depends on parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$, the (i, j) entry of the information matrix is

$$\boldsymbol{I}_{\eta_i,\eta_j} = \frac{1}{2}\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\,\frac{\partial\boldsymbol{\Sigma}}{\partial\eta_i}\,\boldsymbol{\Sigma}^{-1}\,\frac{\partial\boldsymbol{\Sigma}}{\partial\eta_j}\right) = -\frac{1}{2}\,\mathrm{tr}\left(\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\eta_i}\,\frac{\partial\boldsymbol{\Sigma}}{\partial\eta_j}\right). \tag{3.10}$$

Consider the case where $\sigma_I^2 = 0$, i.e., $h = 0$. In calculating the posterior probability of an $h = 1$ model, we need the full 4×4 information matrix even though the truth is $h = 0$. The development in Section 3.1 must be applied with care to this situation. The earlier discussion was for $\boldsymbol{A}$ and $\boldsymbol{B}$ or their inverses applied to vectors $\boldsymbol{v}$ which tend to a smooth function $f_{\boldsymbol{v}}$. The columns of $\boldsymbol{\Sigma}^{-1} = \boldsymbol{A}^{-1}$ after rescaling tend to continuously differentiable functions, but not twice differentiable functions. Thus we cannot expect to treat $\boldsymbol{A}$ as a second order differential operator. But the asymptotic behavior of $\boldsymbol{B}$ only requires the function $f_{\boldsymbol{v}}$ to be Lipschitz continuous, and therefore the discussion of Section 3.1 still applies.

From (3.7), to leading order the contribution of $\boldsymbol{B}$ depends only on $\tau^2 = \sigma_I^2(1+\rho)/(1-\rho)$. We therefore use $(\tau^2, \rho, \beta_0, \beta_1)$ as our parameters, in which case

$$\frac{\partial\boldsymbol{\Sigma}}{\partial\tau^2} = \frac{1-\rho}{1+\rho}\,(\rho^{|i-j|}) = \boldsymbol{I}_n + O(1/n), \tag{3.11}$$

and

$$\frac{\partial\boldsymbol{\Sigma}}{\partial\rho} = -\tau^2\,\frac{2}{(1+\rho)^2}\,(\rho^{|i-j|}) + \tau^2\,\frac{1-\rho}{1+\rho}(|i-j|\rho^{|i-j|-1}). \tag{3.12}$$

Away from the boundaries (which are $O(1/n)$ corrections), the row sums of the matrix $\partial\boldsymbol{\Sigma}/\partial\rho$ tend to zero. Hence, when applied to a sequence of vectors $\boldsymbol{v}$ which tend to a differentiable function $f_{\boldsymbol{v}}(t)$ we have

$$\frac{\partial\boldsymbol{\Sigma}}{\partial\rho}\boldsymbol{v} = O(\tau^2/n). \tag{3.13}$$

For the parameter β_0

$$\frac{\partial\boldsymbol{\Sigma}}{\partial\beta_0} = 2\boldsymbol{A} = 2\boldsymbol{\Sigma}. \tag{3.14}$$

Since $\boldsymbol{A}$ is tridiagonal, so is $\frac{\partial\boldsymbol{\Sigma}}{\partial\beta_1} = \frac{\partial\boldsymbol{A}}{\partial\beta_1}$ and

$$\frac{1}{2}\left(\frac{\partial\boldsymbol{\Sigma}}{\partial\beta_1}\right)_{i,i} = x_{i-1}v(x_{i-1};\boldsymbol{\beta}) + x_i v(x_i;\boldsymbol{\beta}) \tag{3.15}$$

and

$$\left(\frac{\partial \boldsymbol{\Sigma}}{\partial \beta_1}\right)_{i,i+1} = \left(\frac{\partial \boldsymbol{\Sigma}}{\partial \beta_1}\right)_{i+1,i} = -2x_i v(x_i; \boldsymbol{\beta}). \tag{3.16}$$

Also note the useful identity

$$\boldsymbol{A}_{i,i}^{-1} - 2\boldsymbol{A}_{i,i+1}^{-1} + \boldsymbol{A}_{i+1,i+1}^{-1} = \frac{1}{v(x_i; \boldsymbol{\beta})} - \frac{(1 - e^{2\beta_1/n})e^{-2\beta_0 + 2(1-2x_i)\beta_1}}{e^{2(1+1/n)\beta_1} - 1}$$

which follows by direct computation.

If $\sigma_I^2 = 0$, then $\tau^2 = 0$, $\partial\boldsymbol{\Sigma}/\partial\rho = 0$ and all entries of the information matrix corresponding to ρ are of course zero. However, we will want to apply this discussion to the case where σ_I^2, and hence τ^2, is small but positive. We thus need to compute the magnitude of the (ρ, ρ) entry in this case, though we will not need the off-diagonal ρ entries. Plugging formulas (3.11)–(3.16) into (3.10) leads to the following:

$$\boldsymbol{I}_{\tau^2,\tau^2} \sim \frac{n^4}{2}\left(\frac{e^{4b} + e^{-4b} - 16e^{2b} - 16e^{-2b} + 30 + 48b^2}{192b^4(e^{2b} - 1)^2}\right), \tag{3.17}$$

$$\boldsymbol{I}_{\tau^2,\beta_0} \sim n^2\left(\frac{e^{2b} - e^{-2b} - 4b}{8b^2(e^{2b} - 1)}\right), \tag{3.18}$$

$$\boldsymbol{I}_{\tau^2,b} \sim \frac{n^2}{2}\left(\frac{e^{6b} - (4b^2 + b + 1)e^{4b} - e^{2b} + 1 + b}{4b^3e^{2b}(e^{2b} - 1)^2}\right), \tag{3.19}$$

$$\boldsymbol{I}_{\rho,\rho} = O(n^2\tau^4), \quad \boldsymbol{I}_{\beta_0,\beta_0} = 2n, \tag{3.20}$$

$$\begin{aligned}\boldsymbol{I}_{\beta_0,\beta_1} &= n + \frac{e^{2(n+1)\beta_1} + 1}{e^{2(n+1)\beta_1} - 1} - \frac{2(e^{2n\beta_1} - 1)}{n(e^{2(n+1)\beta_1} - 1)(1 - e^{-2\beta_1})} \\ &= n + O(1),\end{aligned} \tag{3.21}$$

and

$$\begin{aligned}\boldsymbol{I}_{\beta_1,\beta_1} &= -\frac{e^{2\beta_0}}{n^2}\left\{ne^{2n\beta_1}\frac{\partial \boldsymbol{A}_{n,n}^{-1}}{\partial b} + \sum_{i=1}^{n-1}\frac{\partial}{\partial b}\left(\boldsymbol{A}_{i,i}^{-1} - 2\boldsymbol{A}_{i,i+1}^{-1} + \boldsymbol{A}_{i+1,i+1}^{-1}\right)ie^{2i\beta_1}\right\} \\ &= \frac{2n}{3} + O(1).\end{aligned} \tag{3.22}$$

Formulas (3.17)–(3.19) are indeterminate if $\beta_1 = 0$ since then both numerator and denominator are zero. This apparent singularity is removed by use of L'Hôpital's rule.

3.3. Posterior probability that $\sigma_I^2 = 0$ when in fact σ_I^2 is 0

Suppose the true model is an $h = 0$ model and that we wish to calculate the posterior probability of an $h = 0$ model. Then formulas (3.20)–(3.22) show that the information matrix for $\boldsymbol{\eta}$ is

$$\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}} = n\begin{pmatrix} 2 & 1 \\ 1 & 2/3 \end{pmatrix} + O(1).$$

All the eigenvalues of this matrix are large as n tends to infinity and therefore the likelihood will be sharply peaked about the MLEs with the dominant contribution to the posterior probability coming from $\boldsymbol{\eta}$ with $\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\| = O(n^{-1/2})$. Since

$$E\left(\frac{\partial^3 \log L(\boldsymbol{\eta}|Y)}{\partial \eta_i^3}\right) = -\frac{3}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \eta_i}\boldsymbol{\Sigma}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}}{\partial \eta_i^2}\right) + 2\mathrm{tr}\left(\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \eta_i}\right)^3\right)$$

and similarly for mixed partials, it is straightforward to show that these expected values are also $O(n)$. Thus the cubic term in the Taylor expansion of the log-likelihood is of order $n\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\|^3$. Thus in the relevant range $\|\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\| = O(n^{-1/2})$, the cubic correction is $O(n^{-1/2})$ and hence negligible. Laplace's method thus applies and we obtain

$$\begin{aligned}\pi(M|\boldsymbol{y}) &= \frac{\alpha_M}{Z(\boldsymbol{y})(2\pi)^{n/2}}(\det\hat{\boldsymbol{\Sigma}})^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right)\\ &\quad\times\pi_m(\hat{\boldsymbol{\theta}}_m)\pi_h(\hat{\boldsymbol{\eta}})(2\pi)^{(m+3+2h)/2}(\det \boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}})^{-1/2}(\det \boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}})^{-1/2}\left(1+O_p(1/n)\right)\\ &= \frac{\sqrt{3}\,\alpha_M n^{-(3m+5)/2}}{Z(\boldsymbol{y})(2\pi)^{(n-m-3)/2}}(\det\hat{\boldsymbol{\Sigma}})^{-1/2}\pi_m(\hat{\boldsymbol{\theta}}_m)\pi_h(\hat{\boldsymbol{\eta}})\det(\boldsymbol{L}_m)^{-1/2}\\ &\quad\times\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right)\left(1+O_p(1/n)\right). \qquad (3.23)\end{aligned}$$

3.4. Posterior probability that $\sigma_I^2 > 0$ when $\sigma_I^2 = 0$

If the truth is $h = 0$ and we are computing the posterior probability of an $h = 1$ model, then the Laplace approximation breaks down in a number of ways. First as we see from (3.12), the ρ pieces of the information matrix are zero. Physically this corresponds to the fact that if $\sigma_I^2 = 0$, then ρ does not affect the likelihood and we get no information about ρ. Thus the ρ part of the integral cannot be approximated using the Laplace method. Less obvious is that the τ^2 part of the integration cannot be done using the Laplace approximation either. Formula (3.17) shows that $\hat{\tau}^2 = O(n^{-2})$ and the dominant range for the integration will be $|\tau^2-\hat{\tau}^2| = O(n^{-2})$. Thus the dominant range of the integral will reach the boundary and boundary effects will be significant. If this were the only problem, then it could be handled using the results of Pauler et al. [17]. However, there is a further problem. In this range the coefficient of the cubic term in the Taylor expansion is

$$\begin{aligned}E\left(\frac{\partial^3 \log L(\boldsymbol{\eta}|Y)}{(\partial\tau^2)^3}\right) &= 2\,\mathrm{tr}\left(\boldsymbol{A}^{-3}\right)+O(n^4)\\ &= 2n^6\int_0^1\int_0^1\int_0^1 g(t_1,t_2)g(t_2,t_3)g(t_3,t_1)\,dt_1\,dt_2\,dt_3+O(n^4)\\ &= O(n^6).\end{aligned}$$

In the dominant range, this means that the cubic term is $O(1)$ and not negligible, and so naive application of Laplace's method will give an inaccurate approximation to the posterior probability.

To obtain an accurate approximation in this case a little more care is needed. Since the (τ, τ) term of the information scales like n^4 and the (β_0, β_0) and (β_1, β_1) terms scale like n, we would expect the off-diagonal terms (τ, β_0) and (τ, β_1) to scale like $n^{2.5}$. Since these entries actually scale like n^2, τ and (β_0, β_1) are asymptotically orthogonal. Thus we can split off the integration over β_0 and β_1 and perform it first. Further, the dominant contribution comes from $\tau^2 = O(n^{-2})$ and hence $\boldsymbol{I}_{\rho,\rho} = O(n^{-2})$. So, there is actually no information about ρ over the entire range of integration. We may thus ignore the dependence of the likelihood on ρ in this range, and the ρ integral is almost trivial.

Let $\tau^2 = \phi^2/n^2$. If $\boldsymbol{v}$ is a sequence of vectors which converges to a smooth function $f_{\boldsymbol{v}}(t)$ as n tends to infinity, then

$$n^2\boldsymbol{\Sigma}\boldsymbol{v} = n^2(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v} \to -e^{2\beta_0}\frac{d}{dt}\left(e^{2bt}\frac{df_{\boldsymbol{v}}(t)}{dt}\right) + \phi^2 f_{\boldsymbol{v}}(t).$$

Thus $\boldsymbol{\Sigma}^{-1}$ will be asymptotic to $n\tilde{g}(s,t)$ where $\tilde{g}$ is Green's function for this differential operator with boundary conditions $\tilde{g}(s,0) = \tilde{g}(s,1) = 0$. The differential equation

$$-e^{2\beta_0}\frac{d}{dt}\left(e^{2bt}\frac{dy}{dt}\right) + \phi^2 y = 0$$

has solutions

$$y(t) = e^{-bt}K_1\left(\frac{\phi e^{-\beta_0 - bt}}{b}\right) \quad \text{and} \quad e^{-bt}K_2\left(\frac{\phi e^{-\beta_0-bt}}{b}\right),$$

where K_1 and K_2 are modified Bessel functions. Defining $\xi = \phi e^{-\beta_0}/\beta_1$, the Green's function is given by

$$\begin{aligned}\tilde{g}(s,t) = e^{-2\beta_0}&\frac{e^{-\beta_1(s+t)}}{\beta_1\left[K_1(\xi)K_2(\xi e^{-\beta_1}) - K_2(\xi)K_1(\xi e^{-\beta_1})\right]}\\ &\times\left[K_1(\xi e^{-\beta_1\max(s,t)})K_2(\xi e^{-\beta_1}) - K_2(\xi e^{-\beta_1\max(s,t)})K_1(\xi e^{-\beta_1})\right]\\ &\times\left[K_1(\xi)K_2(\xi e^{-\beta_1\min(s,t)}) - K_2(\xi)K_1(\xi e^{-\beta_1\min(s,t)})\right].\end{aligned}$$

Then, as for the case $\sigma_I^2 = 0$, we have

$$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})_{i,j} \sim n^3\int_0^1\int_0^1 s^i\tilde{g}(s,t)t^j\,ds\,dt = n^3\tilde{\ell}_{i,j}.$$

Define $\tilde{\boldsymbol{L}}_m = (\tilde{\ell}_{i,j})_{0\leqslant i,j\leqslant m}$ and let μ_h be the marginal prior for $(\tau^2, \beta_0, \beta_1)$. Integrating out $\boldsymbol{\theta}$, β_0, β_1, and ρ, and substituting $\tau^2 = \phi^2/n^2$ gives

$$\begin{aligned}\pi(M|\boldsymbol{y}) \approx\ &\frac{\sqrt{3}\alpha_M\pi_m(\hat{\boldsymbol{\theta}}_m)\mu_h(\hat{\tau}^2,\hat{\beta}_0,\hat{\beta}_1)}{Z(\boldsymbol{y})(2\pi)^{(n-m-3)/2}n^{(3m+9)/2}}\int_0^\infty\left(\det(\tilde{\boldsymbol{L}}_m)\det(\boldsymbol{\Sigma})\right)^{-1/2}\\ &\times\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right)d\phi^2.\end{aligned} \tag{3.24}$$

In this integral all parameters other than ϕ^2 are to be replaced by their MLEs; ρ should not contribute and may be set to zero. Note that in this case there is no penalty for the parameter ρ

but the penalty for σ_I^2, i.e., τ^2, more than compensates. The approximation is further complicated by the fact that the last integral is not Gaussian and cannot be done in closed form.

3.5. Posterior probabilities in the case where $\sigma_I^2 > 0$

Now suppose the true model is such that $\sigma_I^2 > 0$, implying that both the ϵ_j and I_j sources of variation are present. In this case any $h = 0$ model is incorrect and gives exponentially small values for the likelihood, the posterior probability and BIC. Kass and Vaidyanathan [9] argue that it is not necessary to approximate the posterior probability in this case. Nonetheless, it is not difficult to see that the appropriate expansion here takes precisely the same form as it does in the case where the truth is $h = 0$ and we are computing the posterior probability that $h = 0$. Of course, the asymptotic behavior of the MLEs is different since they no longer converge to the true parameter values, but the correct approximation to the posterior probability is still (3.23).

We turn now to the case where we wish to compute the posterior probability that $h = 1$ when the truth is $h = 1$. Let

$$\boldsymbol{S} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{n\times n}$$

denote the $n \times n$ cyclic shift matrix. Note that $\boldsymbol{S}^{-1} = \boldsymbol{S}^T = \boldsymbol{S}^{n-1}$ is the cyclic shift in the other direction. Cyclic matrices are polynomials in $\boldsymbol{S}$. Since cyclic matrices commute, they form a convenient subalgebra of all $n \times n$ matrices. If there were no heteroscedasticity in the model, i.e., if $\beta_1 = 0$, then except for negligible boundary effects, the variance components $\boldsymbol{A}$ and $\boldsymbol{B}$ and hence their sum $\boldsymbol{\Sigma}$ would be cyclic matrices. Explicitly

$$\begin{aligned} \boldsymbol{B} &\approx \sigma_I^2 \left(\boldsymbol{I}_n + \rho(\boldsymbol{S} + \boldsymbol{S}^{-1}) + \rho^2(\boldsymbol{S}^2 + \boldsymbol{S}^{-2}) + \cdots \right) \\ &= (1-\rho^2)\sigma_I^2 \left[(1+\rho^2)\boldsymbol{I}_n - \rho(\boldsymbol{S} + \boldsymbol{S}^{-1}) \right]^{-1}, \end{aligned}$$

and

$$\boldsymbol{A} \approx e^{2\beta_0}(2\boldsymbol{I}_n - \boldsymbol{S} - \boldsymbol{S}^{-1}).$$

If $\beta_1 \neq 0$, then $\boldsymbol{A}$ is not quite so simple. However, even in this case $\boldsymbol{A}$ is still a tridiagonal matrix and can be related to cyclic matrices.

Let $\boldsymbol{D} = (d_{i,j})$ be the $n \times n$ diagonal matrix with diagonal entries $d_{i,i} = \exp(2x_i\beta_1)$. The matrices $\boldsymbol{S}$ and $\boldsymbol{D}$ do not commute, but since the entries of $\boldsymbol{D}$ are slowly varying we have $\boldsymbol{S}^k\boldsymbol{D} \approx \boldsymbol{D}\boldsymbol{S}^k$ for $|k| \ll n$. Since $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are all concentrated near the diagonal only small powers of $\boldsymbol{S}$ will contribute in the formulas below and this will suffice. Thus we may carry out our calculations as though $\boldsymbol{D}$ and $\boldsymbol{S}$ commute. With this definition we have

$$\boldsymbol{A} \approx e^{2\beta_0}\boldsymbol{D}\left(2\boldsymbol{I}_n - \boldsymbol{S} - \boldsymbol{S}^{-1}\right)$$

and

$$\begin{aligned}\boldsymbol{\Sigma} &\approx e^{2\beta_0}\boldsymbol{D}\left(2\boldsymbol{I}_n - \boldsymbol{S} - \boldsymbol{S}^{-1}\right)\\ &\quad +(1-\rho^2)\sigma_I^2\left[(1+\rho^2)\boldsymbol{I}_n - \rho(\boldsymbol{S}+\boldsymbol{S}^{-1})\right]^{-1}. \end{aligned} \tag{3.25}$$

Defining $\boldsymbol{U} = \boldsymbol{S} + \boldsymbol{S}^{-1}$, it follows that

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &\approx \left((1+\rho^2)\boldsymbol{I}_n - \rho\boldsymbol{U}\right)\\ &\quad\times\left[(1-\rho^2)\sigma_I^2\boldsymbol{I}_n + 2(1+\rho^2)e^{2\beta_0}\boldsymbol{D} - (1+\rho)^2e^{2\beta_0}\boldsymbol{D}\boldsymbol{U} + \rho e^{2\beta_0}\boldsymbol{D}\boldsymbol{U}^2\right]^{-1}\\ &= \left((1+\rho^2)\boldsymbol{I}_n - \rho\boldsymbol{U}\right)\\ &\quad\times\left[\left((1-\rho^2)\sigma_I^2\boldsymbol{I}_n + 2(1+\rho^2)e^{2\beta_0}\boldsymbol{D}\right)(\boldsymbol{I}_n - \boldsymbol{R}_+\boldsymbol{U})(\boldsymbol{I}_n - \boldsymbol{R}_-\boldsymbol{U})\right]^{-1}, \end{aligned}$$

where $\boldsymbol{R}_\pm$ are the diagonal matrices given by

$$\begin{aligned}\boldsymbol{R}_\pm &= \left((1+\rho)^2e^{2\beta_0}\boldsymbol{D} \pm \sqrt{(1-\rho)^4e^{4\beta_0}\boldsymbol{D}^2 - 4\rho(1-\rho^2)\sigma_I^2e^{2\beta_0}\boldsymbol{D}}\right)\\ &\quad\times\left[2((1-\rho^2)\sigma_I^2\boldsymbol{I}_n + 2(1+\rho^2)e^{2\beta_0}\boldsymbol{D})\right]^{-1}. \end{aligned}$$

Let $\boldsymbol{C}_\pm$ be the diagonal matrices

$$\boldsymbol{C}_\pm = \left((1+\rho^2)\boldsymbol{R}_\pm - \rho\boldsymbol{I}_n\right)\left[\boldsymbol{R}_\pm - \boldsymbol{R}_\mp\right]^{-1}$$

and $\boldsymbol{F}_\pm = \boldsymbol{I}_n - 4\boldsymbol{R}_\pm^2$. Some algebraic manipulations and expansion of each of $\boldsymbol{I}_n - \boldsymbol{R}_\pm\boldsymbol{U}$ as a geometric series yields

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &\approx \left((1-\rho^2)\sigma_I^2\boldsymbol{I}_n + 2(1+\rho^2)e^{2\beta_0}\boldsymbol{D}\right)^{-1}\\ &\quad\times\left(\boldsymbol{C}_+[\boldsymbol{I}_n - \boldsymbol{R}_+\boldsymbol{U}]^{-1} + \boldsymbol{C}_-[\boldsymbol{I}_n - \boldsymbol{R}_-\boldsymbol{U}]^{-1}\right)\\ &= \left((1-\rho^2)\sigma_I^2\boldsymbol{I}_n + 2(1+\rho^2)e^{2\beta_0}\boldsymbol{D}\right)^{-1}\\ &\quad\times\sum_{k=-\infty}^{\infty}\left\{\boldsymbol{C}_+\boldsymbol{F}_+^{-1/2}\left(2\boldsymbol{R}_+\left[\boldsymbol{I}_n + \sqrt{\boldsymbol{F}_+}\right]^{-1}\right)^{|k|}\right.\\ &\quad\left.+\,\boldsymbol{C}_-\boldsymbol{F}_-^{-1/2}\left(2\boldsymbol{R}_-\left[\boldsymbol{I}_n + \sqrt{\boldsymbol{F}_-}\right]^{-1}\right)^{|k|}\right\}\boldsymbol{S}^k. \end{aligned} \tag{3.26}$$

Dropping the $\pm$ subscripts, the diagonal entries $r_{i,i}$ of $\boldsymbol{R}_+$ and $\boldsymbol{R}_-$ are the two roots of the quadratic equation

$$\left((1-\rho^2)\sigma_I^2 + 2(1+\rho^2)e^{2\beta_0}d_{i,i}\right)r_{i,i}^2 - (1+\rho)^2e^{2\beta_0}d_{i,i}r_{i,i} + \rho e^{2\beta_0}d_{i,i} = 0. \tag{3.27}$$

If the roots of this polynomial are complex conjugates, then their squared modulus is

$$|r_{i,i}|^2 = \frac{\rho e^{2\beta_0} d_{i,i}}{(1-\rho^2)\sigma_I^2 + 2(1+\rho^2)e^{2\beta_0} d_{i,i}}.$$

For $\sigma_I^2(1-\rho^2) > 0$, we conclude

$$|r_{i,i}|^2 < \frac{\rho}{2(1+\rho^2)} < \frac{1}{4}.$$

If the roots are real, then rearranging (3.27) gives

$$(2r_{i,i} - 1)((1+\rho^2)r_{i,i} - \rho) = -\frac{(1-\rho^2)\sigma_I^2 r_{i,i}^2}{e^{2\beta_0} d_{i,i}}.$$

For $\sigma_I^2(1-\rho^2) > 0$, the right-hand side of this equation is negative. Therefore, $r_{i,i}$ must lie strictly between the two roots of the quadratic on the left. Since $-\frac{1}{2} < \rho/(1+\rho^2) < \frac{1}{2}$, we conclude that $-\frac{1}{2} < r_{i,i} < \frac{1}{2}$. Combining these two cases, we see that every entry of the diagonal matrices $\boldsymbol{R}_\pm$ has magnitude strictly less than $\frac{1}{2}$. Therefore, the series in (3.26) all converge and the coefficients of $\boldsymbol{S}^k$ decay exponentially as $|k|$ tends to ∞. This justifies our claim above that $\boldsymbol{\Sigma}^{-1}$ is concentrated near the diagonal and hence our use of the approximation $\boldsymbol{S}^k\boldsymbol{D} \approx \boldsymbol{D}\boldsymbol{S}^k$ is legitimate. Since the coefficients decay exponentially, the coefficient of $\boldsymbol{S}^n = \boldsymbol{I}_n$ and powers of higher multiples of n are negligible and we can ignore them below.

In the limit as n tends to infinity, the diagonal matrices $\boldsymbol{D}$, $\boldsymbol{R}_\pm$, and $\boldsymbol{C}_\pm$ should be interpreted as converging to functions on $[0, 1]$. The diagonal matrix $\boldsymbol{D}$ converges to the function $d(t) = \exp(2bt)$. Let $r_\pm(t)$ be the limiting functions for $\boldsymbol{R}_\pm$. Then

$$r_\pm = \frac{(1+\rho)^2 e^{2\beta_0+2bt} \pm \sqrt{(1-\rho)^4 e^{4\beta_0+4bt} - 4\rho(1-\rho^2)\sigma_I^2 e^{2\beta_0+2bt}}}{2((1-\rho^2)\sigma_I^2 + 2(1+\rho^2)e^{2\beta_0+2bt})}.$$

The formulas (3.25) and (3.26) give

$$\boldsymbol{\Sigma} \sim \sum_{k=-\infty}^{\infty} \boldsymbol{F}_k \boldsymbol{S}^k \to \sum_{k=-\infty}^{\infty} f_k(t)\boldsymbol{S}^k$$

and

$$\boldsymbol{\Sigma}^{-1} \sim \sum_{k=-\infty}^{\infty} \boldsymbol{G}_k \boldsymbol{S}^k \to \sum_{k=-\infty}^{\infty} g_k(t)\boldsymbol{S}^k$$

for diagonal matrices $\boldsymbol{F}_k$ and $\boldsymbol{G}_k$ and limiting functions f_k and g_k:

$$f_k(t) = \begin{cases} \sigma_I^2 + 2e^{2\beta_0+2bt}, & k = 0, \\ \rho\sigma_I^2 - e^{2\beta_0+2bt}, & |k| = 1, \\ \rho^{|k|}\sigma_I^2, & |k| > 1, \end{cases}$$

$$g_k(t) = \frac{1}{\sqrt{(1-\rho)^4 e^{4\beta_0+4bt} - 4\rho(1-\rho^2)\sigma_I^2 e^{2\beta_0+2bt}}} \times \left\{ \frac{((1+\rho^2)r_+(t)-\rho)}{\sqrt{1-4r_+^2(t)}} \left(\frac{2r_+(t)}{1+\sqrt{1-4r_+^2(t)}} \right)^{|k|} + \frac{((1+\rho^2)r_-(t)-\rho)}{\sqrt{1-4r_-^2(t)}} \left(\frac{2r_-(t)}{1+\sqrt{1-4r_-^2(t)}} \right)^{|k|} \right\}.$$

Since $f_{-k} = f_k$ and $g_{-k} = g_k$

$$\begin{aligned} \boldsymbol{I}_{\eta_i,\eta_j} &= -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \eta_i}\frac{\partial \boldsymbol{\Sigma}}{\partial \eta_j}\right) \\ &\sim -\frac{1}{2}\operatorname{tr}\left(\sum_{k=-\infty}^{\infty}\frac{\partial \boldsymbol{G}_k}{\partial \eta_i}\boldsymbol{S}^k \sum_{\ell=-\infty}^{\infty}\frac{\partial \boldsymbol{F}_\ell}{\partial \eta_j}\boldsymbol{S}^\ell\right) \\ &\sim -\frac{1}{2}\sum_{k=-\infty}^{\infty}\sum_{\ell=-\infty}^{\infty}\operatorname{tr}\left(\frac{\partial \boldsymbol{G}_k}{\partial \eta_i}\left[\boldsymbol{S}^k\frac{\partial \boldsymbol{F}_\ell}{\partial \eta_j}\boldsymbol{S}^{-k}\right]\boldsymbol{S}^{k+\ell}\right). \end{aligned}$$

We will see below that this sum converges exponentially, therefore we need only consider terms with $|k|, |\ell| \ll n$. Hence $\boldsymbol{S}^k$ and $\frac{\partial \boldsymbol{F}_\ell}{\partial \eta_j}$ approximately commute.

Since $\boldsymbol{S}$ is a cyclic shift matrix, $\boldsymbol{S}^{k+\ell}$ has only zero entries on the diagonal unless $k+\ell$ is a multiple of n. Since $|k|, |\ell| \ll n$, the only case we need to consider is when $k+\ell = 0$. Plugging in these two observations gives

$$\begin{aligned} \boldsymbol{I}_{\eta_i,\eta_j} &\sim -\frac{1}{2}\sum_{k=-\infty}^{\infty}\operatorname{tr}\left(\frac{\partial \boldsymbol{G}_k}{\partial \eta_i}\frac{\partial \boldsymbol{F}_{-k}}{\partial \eta_j}\right) \\ &\sim -\frac{n}{2}\sum_{k=-\infty}^{\infty}\int_0^1\left(\frac{\partial g_k(t)}{\partial \eta_i}\frac{\partial f_k(t)}{\partial \eta_j}\right)dt. \end{aligned}$$

The functions g_k and f_k decay exponentially as $|k|$ tends to infinity. Hence this sum converges rapidly and we see that

$$\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}} = nK(\boldsymbol{\eta}) + O(1)$$

for some calculable 4×4 matrix $K(\boldsymbol{\eta})$. In particular all eigenvalues of the information matrix tend to infinity as n tends to infinity. Thus the integral representing the posterior probability of model M is peaked and the dominant contribution comes from $\boldsymbol{\eta}$ with $\|\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}\| = O(n^{-1/2})$. Similar arguments show that the cubic and higher order coefficients in the Taylor expansion of the log-likelihood are also $O(n)$. Thus they are negligible for $\boldsymbol{\eta}$ in the dominant range. Hence the

standard Laplace approximation applies in this case and we obtain

$$\pi(M|\boldsymbol{y}) = \frac{\alpha_M \pi_m(\hat{\boldsymbol{\theta}}_m)\pi_h(\hat{\boldsymbol{\eta}})}{n^{(m+5)/2} Z(\boldsymbol{y})(2\pi)^{(n-m-5)/2}} \prod_{i=1}^{m} (2i+1)^{1/2} \binom{2i}{i}$$
$$\times \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right) (\det \hat{\boldsymbol{\Sigma}} \det K(\hat{\boldsymbol{\eta}}))^{-1/2}$$
$$\times \left(\frac{\hat{\sigma}_I^2(1+\hat{\rho})}{1-\hat{\rho}}\right)^{(m+1)/2} \left(1 + O_p(1/n)\right). \tag{3.28}$$

3.6. Choice of an approximation

In practice it is not known whether or not $\sigma_{I,0}^2 = 0$, and hence it is not clear which approximation of $\pi((m,h)|\boldsymbol{y})$ is "correct". To deal with this problem we propose using a weighted geometric mean of approximations that are appropriate for the model under consideration. Define $h_{m,0}$ by

$$h_{m,0} = \begin{cases} 1 & \text{if the true degree-}m\text{ model has an intrinsic component,} \\ 0 & \text{otherwise.} \end{cases}$$

For $i, j = 0, 1$, let ζ_{ij} denote the appropriate approximation to the quantity $\pi((m,i)|\boldsymbol{y})Z(\boldsymbol{y})$ when $h_{m,0} = j$, where for simplicity we suppress dependence of the approximations on m. From Sections 3.3–3.5, $\zeta_{00}, \zeta_{01}, \zeta_{10}, \zeta_{11}$ are defined by (3.23), (3.23), (3.24) and (3.28), respectively. We consider approximations of the form

$$\hat{\pi}((m,i)|\boldsymbol{y}) = \zeta_{i0}^{\hat{p}_0} \cdot \zeta_{i1}^{1-\hat{p}_0} / \hat{Z}(\boldsymbol{y}), \quad i = 0, 1, \tag{3.29}$$

where $\hat{p}_0$ approximates $P(h = 0|m, \boldsymbol{y})$ and $\hat{Z}(y)$ is the appropriate normalizing constant.

In principle, it seems that $P(h = 0|m, \boldsymbol{y})$ would be just as difficult to approximate as $\hat{\pi}((m,0)|\boldsymbol{y})$. However, it turns out that a simple BIC approximation of $P(h = 0|m, \boldsymbol{y})$ suffices. This is because two models, namely $(m, 0)$ and $(m, 1)$, are being compared that have *the same value of m*. Sections 3.3–3.5 suggest the following two versions of BIC for the respective cases $h_{m,0} = 0$ and 1:

$$\mathrm{BIC}(m,h) = \begin{cases} 2\log \hat{L}_{m,h} - (3m+5)\log n, & h = 0, \\ 2\log \hat{L}_{m,h} - (3m+9)\log n, & h = 1, \end{cases} \tag{3.30}$$

and

$$\mathrm{BIC}(m,h) = \begin{cases} 2\log \hat{L}_{m,h} - (m+3)\log n, & h = 0, \\ 2\log \hat{L}_{m,h} - (m+5)\log n, & h = 1, \end{cases} \tag{3.31}$$

where $\hat{L}_{m,h}$ denotes maximized likelihood. The second of these is just ordinary BIC, while the former takes into account how the exponents of n in (3.23) and (3.24) differ from the classical setting.

A BIC approximation of $P(h = 0|m, \boldsymbol{y})$ is

$$\frac{\exp(\mathrm{BIC}(m, 0)/2)}{\exp(\mathrm{BIC}(m, 0)/2) + \exp(\mathrm{BIC}(m, 1)/2)},$$

which takes the form

$$\hat{p}_{0i} = \frac{1}{1 + n^{-2+i}\hat{L}_{m,1}/\hat{L}_{m,0}}, \quad i = 0, 1,$$

depending on whether one uses (3.30) or (3.31), respectively. Of these two, we prefer $\hat{p}_{00}$, for reasons that will become clear momentarily.

If $h_{m,0} = 0$, then $2\log \hat{L}_{m,1}/\hat{L}_{m,0}$ converges in distribution to a random variable having a χ^2_2 distribution. In this case it follows that $\hat{p}_{00} = 1 + O_p(n^{-2})$, whereas $\hat{p}_{01} = 1 + O_p(n^{-1})$. If $h_{m,1} = 1$, then

$$\hat{p}_{0i} = O_p(n^{2-i}\exp(-Cn)),$$

for some positive constant C, in which case $\hat{p}_{0i}$ converges to 0 exceptionally quickly whether $i = 0$ or 1. The preceding facts together lead us to propose $\hat{p}_{00}$ for use in practice.

3.7. Assessing the posterior probability of overdifferencing

Let us now suppose we are in the setting of Remark 3. The data have been differenced a given number of times, and we wish to compute the posterior probability of two possibilities for the error model. One possibility is that the errors $\epsilon_1, \ldots, \epsilon_n$ are i.i.d., meaning that the data have been differenced the correct number of times. In the other case the errors are $\epsilon_1 - \epsilon_0, \ldots, \epsilon_n - \epsilon_{n-1}$, i.e., the data have been overdifferenced. (We assume that the data have not been underdifferenced, a condition which is usually easy to identify.) In either of these two cases, there is only one unknown error parameter, $\mathrm{Var}(\epsilon_i) = \exp(2\beta_0)$.

We wish to approximate the posterior probabilities of the events E_0 and E_1, which denote that the errors are i.i.d. and that the data have been overdifferenced, respectively. This is done in much the same way as were the previous approximations in this section, with the main difference being that now the information matrices are simpler in form. Let $\pi((m, E)|\boldsymbol{y})$ denote the posterior probability that m is the correct polynomial degree and E is the truth, $E = E_0, E_1$. Approximating $\pi((m, E_1)|\boldsymbol{y})$ is virtually the same as approximating the posterior probability of an $h = 0$ model as in Sections 3.3 and 3.5. The only difference is that in the present case the errors are assumed to be homoscedastic. Whether or not E_1 is the truth, an appropriate Laplace approximation is

$$\hat{\pi}((m, E_1)|\boldsymbol{y}) = \frac{\alpha(m, E_1)n^{-(3m+4)/2}}{\sqrt{2}\,\hat{Z}(\boldsymbol{y})(2\pi)^{(n-m-2)/2}}\pi_m(\hat{\boldsymbol{\theta}}_m)\varpi(\hat{\beta}_0)\det(\boldsymbol{L}_m)^{-1/2}$$

$$\times(\det\hat{\boldsymbol{\Sigma}})^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_m)\right),$$

where $\alpha(m, E_1)$ is the prior probability of (m, E_1), ϖ is the prior density of β_0, $\boldsymbol{\Sigma} \equiv \boldsymbol{A}$ with $\beta_1 = 0$, and $\boldsymbol{L}_m$ is defined as in Section 3.1 with $\beta_1 = 0$.

Approximating $\pi((m, E_0)|\boldsymbol{y})$ will usually be straightforward since E_0 corresponds to i.i.d. errors. Using a conjugate prior will actually allow exact determination of the requisite integral.

Writing $\sigma = \exp(\beta_0)$, a reasonable approximation for a more general prior is

$$\hat{\pi}((m, E_0)|\boldsymbol{y}) = \frac{\alpha(m, E_0)n^{-(m+1)/2}}{\hat{Z}(\boldsymbol{y})(2\pi)^{(n-m-1)/2}}(\det \boldsymbol{R})^{-1/2}\pi_m(\tilde{\boldsymbol{\theta}}_m) \times \int_0^\infty \exp\left(-\frac{n\hat{\sigma}^2}{2\sigma^2}\right)\sigma^{-(n-m)}\varpi(\log\sigma)\,d\sigma, \tag{3.32}$$

where $\tilde{\boldsymbol{\theta}}_m$ is the least-squares estimate of $\boldsymbol{\theta}_m$, $\hat{\sigma}^2 = n^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}_m)'(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\theta}}_m)$, and $\boldsymbol{R}$ is the $(m+1)\times(m+1)$ matrix with (i, j) element equal to $(i+j-1)^{-1}$. This approximation results from using a Laplace approximation for the integral with respect to $\boldsymbol{\theta}_m$ in the expression that defines $\pi((m, E_0)|\boldsymbol{y})$. If need be, numerical integration can be used to approximate the integral in (3.32).

Having obtained $\hat{\pi}((m, E_1)|\boldsymbol{y})$ for each m, an approximation of the posterior probability of overdifferencing may be obtained by summing $\hat{\pi}((m, E_1)|\boldsymbol{y})$ over all m.

4. An analysis of variable star data

We now apply the approximations developed in Section 3 to data from the prototypical long-period variable star known as Mira. The first step in this comparison is to choose explicit priors. First, consider the prior probabilities $\alpha_{(m,h)}$ on the models. There seems no *a priori* reason to prefer one covariance structure over the other. Therefore, we took the priors to be independent of h. We wish to use model (2.1) to test whether or not there is systematic variation, i.e., a trend, in the Y_js. Therefore, we assigned a combined prior of $\frac{1}{2}$ (or $\frac{1}{4}$ each) to the two no-trend ($m = 0$) models. For the remaining polynomial degrees, we took $m_{\max} = 15$ and chose a prior proportional to $1/m$, which is Jeffreys' noninformative prior for an integer parameter [5]. Normalizing these gives

$$\alpha_{(m,h)} = \begin{cases} 1/4 & \text{if } m = 0, \\ 0.0753413946/m & \text{if } m = 1, \ldots, 15. \end{cases}$$

For the priors on the model parameters, we assume the mean parameters $\boldsymbol{\theta}_m$ are *a priori* independent of the covariance parameters $\boldsymbol{\eta}$, and use a multivariate normal prior for $\boldsymbol{\theta}_m$. The choice of mean and covariance of this normal prior will be discussed below. Use of a normal prior for $\boldsymbol{\theta}_m$ was in part motivated by the fact that it allows the computation of the posterior probability integral over $\boldsymbol{\theta}_m$ to be done in closed form.

The covariance parameters β_0 and β_1 are determined by measurement error, whereas the parameters ρ and σ_Z^2 are determined by intrinsic variation due to the star. Therefore we assumed that (β_0, β_1) and (ρ, σ_Z^2) were *a priori* independent. For (β_0, β_1) we chose a bivariate normal prior, i.e., $(\beta_0, b) \sim N_2(\boldsymbol{\gamma}, \boldsymbol{V})$.

The parameter ρ is taken to be *a priori* uniformly distributed on $[-1, 1]$. In the AR(1) model, we have

$$I_j = \rho I_{j-1} + Z_j, \quad j = 2, \ldots, n,$$

where $Z_2, \ldots, Z_n$ are i.i.d. $N(0, \sigma_Z^2)$. Here ρ represents the carry-over from the previous observation and Z_j a new random effect, and hence we chose to have ρ and σ_Z^2 be *a priori* independent.

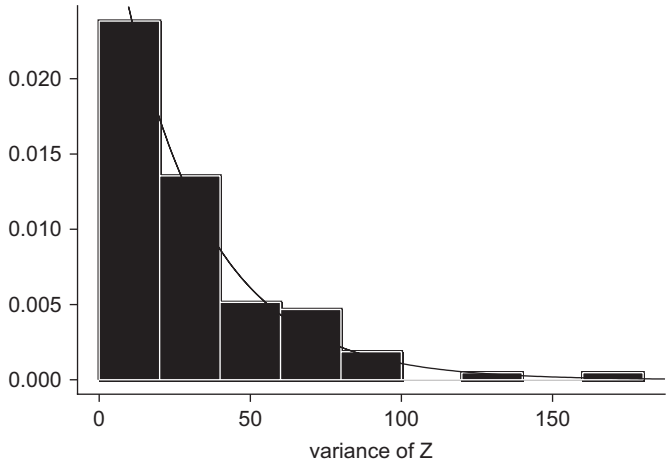

Fig. 1. Histogram of observed $\hat{\sigma}_Z^2$ and fitted density.

A histogram of MLEs of σ_Z^2 for all 378 stars in the database is shown in Fig. 1. This histogram motivated us to use the fitted exponential density in Fig. 1 as a prior for σ_Z^2.

The mean vectors and covariance matrices for the multivariate normal priors of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{15}$ were also determined empirically. For each of the 378 stars, we chose a model $(\hat{m}_i, \hat{h}_i)$ by naive application of BIC (which does not require specifying a prior) and computed the maximum likelihood estimators of the parameters for the selected model. The prior mean v_m and prior covariance $\boldsymbol{W}_m$ for $\boldsymbol{\theta}_m$ were chosen to be the sample mean and sample covariance of $\hat{\boldsymbol{\theta}}_m$ for all stars having $\hat{m} = m$. Two exceptions to this rule were deemed necessary. The number of stars with $\hat{m} = 14$ and 15 was too small to give a positive definite sample covariance $\boldsymbol{W}_m$. Therefore data for $\hat{m} = 14$ and 15 were pooled to give estimates for $\boldsymbol{W}_{14}$ and $\boldsymbol{W}_{15}$. Also one star with only 32 observations had $\hat{m} = 15$. Such a large polynomial degree hardly seems warranted on the basis of 32 observations, and hence this star was treated as an outlier and excluded from the computations.

Similarly the prior mean $\boldsymbol{\gamma}$ and prior covariance $\boldsymbol{V}$ for (β_0, β_1) were taken to be the sample mean and sample covariance of $(\hat{\beta}_0, \hat{\beta}_1)$ for all stars, excluding three outliers. A scatterplot of maximum likelihood estimates of $(\hat{\beta}_0, \hat{b})$ for all 378 stars in the database is shown in Fig. 2.

The integral with respect to $\boldsymbol{\theta}_m$ in $\pi(M|\boldsymbol{y})Z(\boldsymbol{y})$ can be done in closed form, since the integrand is proportional to a multivariate normal density. The remaining 2- or 4-variate integral was approximated using importance sampling based on 10,000 i.i.d. observations from a multivariate normal distribution with the same mode and Hessian at the mode as the integrand.

One approximation of $\pi(M|\boldsymbol{y})$ we wish to consider is that provided by the standard BIC, i.e.,

$$\pi(M|\boldsymbol{y}) \approx \frac{e^{\mathrm{BIC}_M/2}}{\sum_{M'} e^{\mathrm{BIC}_{M'}/2}},$$

where

$$\mathrm{BIC}_M = 2\log L_M(\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\eta}}|\boldsymbol{y}) - (m + 3 + 2h)\log n.$$

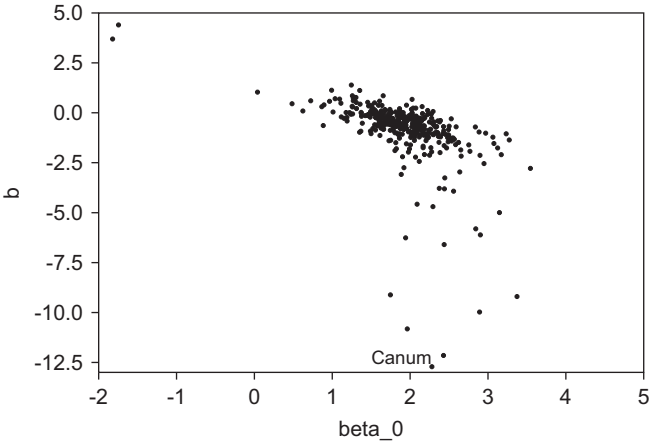

Fig. 2. Scatterplot of $\hat{b}$ vs. $\hat{\beta}_0$.

A second approximation of interest is a modified version of BIC with a corrected penalty term. Define

$$\text{mod-BIC}_M = 2 \log L_M(\hat{\boldsymbol{\theta}}_m, \hat{\boldsymbol{\eta}} | \boldsymbol{y}) - k_M \log n,$$

where

$$k_M = \begin{cases} 3m + 5 & \text{if } h = 0, \\ m + 5 & \text{if } h = 1 \text{ and } \hat{\sigma}_Z^2 \geqslant 0.001, \\ 3m + 9 & \text{if } h = 1 \text{ and } \hat{\sigma}_Z^2 < 0.001. \end{cases}$$

These penalties have been chosen based on the powers of n in (3.23), (3.24), and (3.28). The cutoff $\hat{\sigma}_Z^2 < 0.001$ for deciding when to use (3.24) is somewhat arbitrary. However, data sets with $\hat{\sigma}_Z^2 < 0.001$ invariably had values of $\hat{\sigma}_Z^2$ much smaller than 0.001.

The third approximation is the standard Laplace approximation (3.1), where the information matrices $\boldsymbol{I}_{\boldsymbol{\theta},\boldsymbol{\theta}}$ and $\boldsymbol{I}_{\boldsymbol{\eta},\boldsymbol{\eta}}$ are estimated as the negative of the Hessian of the log-likelihood at the MLE. Finally, "Corrected Laplace" in Table 1 is an asymptotically correct version of the Laplace approximation that uses the weighted geometric mean discussed in Section 3.6.

The "exact" posterior probabilities and the five approximations were computed for Mira. Plots of the observed pseudo-periods and a sixth degree polynomial fit are shown in Fig. 3. "Exact" posterior probabilities and the five approximates are given in Table 1. Plots of posterior probability as a function of polynomial degree are given in Fig. 4.

The model with highest posterior probability is $(m, h) = (6, 0)$ but models with nearby degrees are nearly as likely. Standard BIC correctly selected the model with highest posterior probability, but did not provide a good estimate of the posterior probabilities. This failure of Standard BIC is in agreement with our theoretical results as derived in Section 3.

Modified BIC provided by far the poorest estimates of the posterior probabilities and is maximized at $m = 0$. An explanation of this performance is that the constant terms derived in Section 3 are not included in the Modified BIC. These constant terms are quite large. For example,

Table 1
Posterior probabilities for Mira

m	h	"Exact"	Standard BIC	Modified BIC	Standard Laplace	Corrected Laplace
4	0	0.0042	0.0016	0.0000	0.0039	0.0029
5	0	0.2445	0.3160	0.0000	0.2563	0.2257
6	0	0.2577	0.4831	0.0000	0.2864	0.2697
7	0	0.1373	0.0586	0.0000	0.1450	0.1429
8	0	0.2487	0.0936	0.0000	0.2037	0.2234
9	0	0.0173	0.0297	0.0000	0.0025	0.0030
10	0	0.0340	0.0036	0.0000	0.0016	0.0022
11	0	0.0077	0.0007	0.0000	0.0041	0.0012
12	0	0.0093	0.0001	0.0000	0.0077	0.0000
13	0	0.0004	0.0000	0.0000	0.0000	0.0000
0	1	0.0145	0.0000	0.6063	0.0784	0.0660
1	1	0.0000	0.0000	0.3486	0.0000	0.0000
2	1	0.0014	0.0000	0.0401	0.0079	0.0081
3	1	0.0008	0.0000	0.0049	0.0027	0.0029
4	1	0.0078	0.0000	0.0000	0.0000	0.0037
5	1	0.0024	0.0042	0.0000	0.0000	0.0174
6	1	0.0034	0.0064	0.0000	0.0000	0.0158
7	1	0.0018	0.0008	0.0000	0.0000	0.0083
8	1	0.0046	0.0012	0.0000	0.0000	0.0062
9	1	0.0005	0.0004	0.0000	0.0000	0.0001
10	1	0.0012	0.0000	0.0000	0.0000	0.0000
11	1	0.0002	0.0000	0.0000	0.0000	0.0001
12	1	0.0003	0.0000	0.0000	0.0000	0.0001

The various methods are explained in the text. Values of (m, h) for which each probability was 0.0000 have been excluded from the table.

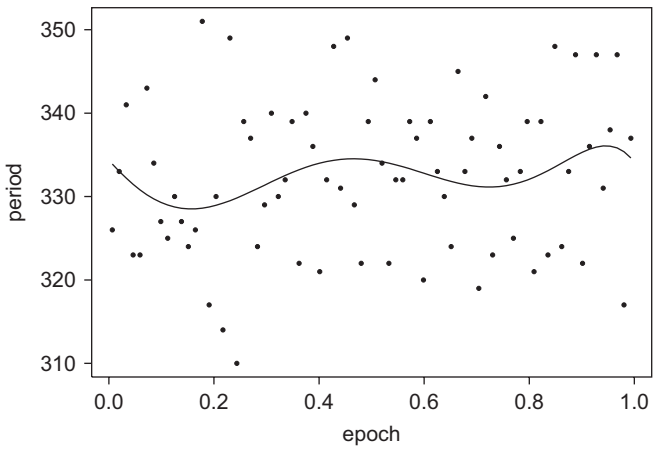

Fig. 3. Data and sixth degree polynomial fit for Mira.

in (3.28) the constant term includes

$$\prod_{i=1}^{m} (2i+1)^{1/2} \binom{2i}{i}$$

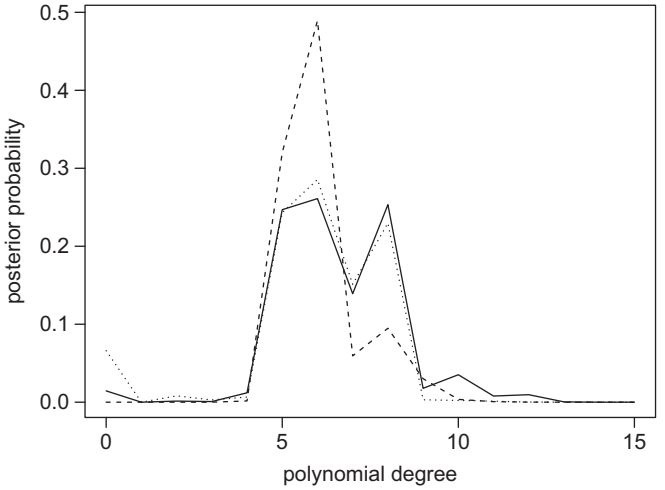

Fig. 4. Approximate posterior probabilities for Mira. The solid, dotted and dashed lines correspond to "exact," corrected Laplace and standard BIC, respectively.

which grows roughly like $2^{m^2/2}$. Further, the Modified BIC has a much smaller penalty for $h = 1$ models with $\hat{\sigma}_Z^2 > 0.001$. As a result it is strongly biased towards the $h = 1$ models with low polynomial degree. Obviously, correcting the usual BIC in this model is not a simple matter of adjusting the penalty term.

The standard and corrected Laplace approximations gave comparable results, although corrected Laplace was on average closer to the "exact" probabilities than was standard Laplace. Both provided more accurate estimates of the posterior probabilities than their BIC counterparts.

The Standard and Modified BICs both provided poor estimates of the posterior probabilities and hence their use for this purpose is not recommended. However, Standard BIC does seem to provide a fairly good criterion for model selection. This justifies our method of estimating priors, wherein we used parameter estimates corresponding to models that maximized BIC.

All the estimates of the posterior probabilities required dramatically less computation time than the "exact" posterior probabilities. Computing "exact" posterior probabilities for all 378 variable stars in the data would be a prohibitively lengthy calculation.

References

[1] A.S. Eddington, L. Plakidis, Irregularities of period of long-period variable stars, Monthly Notices Roy. Astronomical Soc. 90 (1929) 65–71.

[2] W.R. Gilks, S. Richardson, D.J. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC, Boca Raton, FL, 1996.

[3] J.D. Hart, C. Koen, F. Lombard, An analysis of pulsation periods of long-period variable stars, Department of Statistics, Texas A&M University, ⟨http://www.stat.tamu.edu/~hart/HKL5.rev.pdf⟩.

[4] J.E. Isles, D.B.R. Saw, Mira stars—I: R Ari, R Aur, R Boo and S Boo, J. British Astronomical Assoc. 97 (1987) 106–116.

[5] H. Jeffreys, Theory of Probability, third ed., Oxford University Press, London, 1961.

[6] R.A. Johnson, An asymptotic expansion for posterior distributions, Ann. Math. Statist. 38 (1967) 1899–1906.

[7] R.L. Kashyap, Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Trans. Pattern Anal. Mach. Intell. 4 (1982) 99–104.

[8] R.E. Kass, L. Tierney, J.B. Kadane, The validity of posterior expansions based on Laplace's method, in: S. Geisser, J.S. Hodges, S.J. Press, A. Zellner (Eds.), Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard, North-Holland, Amsterdam, 1990, pp. 473–488.
[9] R.E. Kass, S.K. Vaidyanathan, Approximate Bayes factors and orthogonal parameters with application to testing equality of two binomial proportions, J. Roy. Statist. Soc. Ser. B 54 (1992) 129–144.
[10] R.E. Kass, L. Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, J. Amer. Statist. Assoc. 90 (1995) 928–934.
[11] C. Koen, F. Lombard, The analysis of indexed astronomical time series—VII. Simultaneous use of times of maxima and minima to test for period changes in long-period variables, Monthly Notices Roy. Astronomical Soc. 325 (2001) 1124–1132.
[12] C. Koen, F. Lombard, The analysis of indexed astronomical time series—IX. A period change test, Monthly Notices Roy. Astronomical Soc. 353 (2004) 98–104.
[13] L.M. Laplace, Théorie Analytique des Probabilités, third ed., Courcier, Paris (Reprinted in Complete Works, vol. 7. Gauthier-Villars, Paris, 1886), 1820.
[14] F. Lombard, An alternative to O–C analysis of variable star periods, Monthly Notices Roy. Astronomical Soc. 294 (1998) 657–666.
[15] H.J. Newton, Timeslab: A Time Series Analysis Laboratory, Wadsworth & Brooks/Cole, Belmont, CA, 1988.
[16] F. Olver, Asymptotics and Special Functions, Academic Press, New York, 1997.
[17] D.K. Pauler, J.C. Wakefield, R.E. Kass, Bayes factors and approximations for variance component models, J. Amer. Statist. Assoc. 94 (1999) 1242–1253.
[18] J.R. Percy, T. Colivas, Long term changes in Mira stars. I. Period fluctuations in Mira stars, Pub. Astronomical Soc. Pacific 111 (1999) 94–97.
[19] C.I. Plosser, G.W. Schwert, Estimation of a non-invertible moving average process: the case of over-differencing, J. Econometrics 6 (1977) 199–224.
[20] A.E. Raftery, Approximate Bayes factors and accounting for model uncertainty in generalized linear models, Biometrika 83 (1996) 251–266.
[21] J.D. Sargan, A. Bhargava, Maximum likelihood estimation of regression models with first order moving average errors when the root lies on the unit circle, Econometrica 51 (1983) 799–820.
[22] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[23] T.E. Sterne, The errors of period of variable stars, Harvard College Observatory Circular 387 (1934) 1–23.
[24] T.E. Sterne, L. Campbell, Changes of period in variable stars of long period, Proc. Nat. Acad. Sci. 23 (1937) 115–117.
[25] R.S. Tsay, Testing for noninvertible models with applications, J. Business Economic Statist. 11 (1993) 225–233.