A HIERARCHICAL APPROACH TO COVARIANCE FUNCTION ESTIMATION FOR TIME SERIES

By Michael J. Daniels and Noel Cressie

Iowa State University, The Ohio State University

First Version received January 1999

Abstract. The covariance function in time series models is typically modelled via a parametric family. This ensures straightforward best linear prediction while maintaining positive-definiteness of the covariance function. We suggest an alternative approach, which will result in data-determined shrinkage towards this parametric model. Positive-definiteness is maintained by carrying out the shrinkage in the spectral domain. We offer both a fully Bayesian hierarchical approach and an approximate hierarchical approach that will be much simpler computationally. These are implemented on the frequently analysed Canadian lynx data and compared to other models that have been fitted to these data.

Keywords. Bayesian statistics, lynx data, MCMC, shrinkage, spectral density.

1. INTRODUCTION

Data are often collected over time or space. An important aspect of this type of data is dependence among the observations due to temporal or spatial proximity. To model accurately such processes, to predict into the future (time), or to predict at another location (space), this dependence must be carefully modelled and inferred. A simplifying assumption that is often used (and that we shall also use) is that of second-order stationarity, where the covariance between two observations is a function only of the time or displacement between the observations. Researchers then typically model the covariance function using parametric models, both for ease of prediction and to ensure that the covariance function is positive-definite.

In the context of multivariate analysis, Daniels and Kass (1999) discuss the idea of borrowing strength from a structured covariance matrix to obtain a more stable estimate of covariances in small samples. Their approach allows the data to determine the amount of shrinkage towards this structure. As a result, the uncertainty of not knowing the true structure is incorporated into the variability of the estimate, and precision is gained over the standard unshrunk estimate. Our intention in this paper is to extend this approach to the context of time series modelling and prediction. Thus, instead of fitting a specific parametric model to the covariance function, we develop instead a methodology for shrinkage towards a parametric model.

The aforementioned idea will be implemented as a fully Bayesian hierarchical model that will account for all uncertainties: the uncertainty in the estimation of the covariance function and all relevant uncertainties when carrying out predictions. To ensure the positive-definiteness of the inferred covariance function, we shall formulate our hierarchical model in the spectral domain. Then, we return to the time domain easily via Bochner's Theorem (Bochner, 1955). A non-parametric estimator for the covariance function for unequally spaced data, formulated within the spectral domain, was discussed in Hall *et al.* (1994). In a like manner, those authors ensure positive-definiteness by transforming back to the time domain from the spectral domain.

In Section 2, we present the Bayesian hierarchical model that is central to our approach. Section 3 discusses model fitting and computational issues. In Section 4, we apply our approach to the frequently analysed Canadian lynx data and compare our results to other models that have been fitted to these data. Finally, Section 5 contains conclusions and discussions.

2. MODEL

Let $\{Y_1, \ldots, Y_n\}$ be a time series of observations at times $t = 1, \ldots, n$, which we refer to subsequently as Y. We consider the following model:

$$Y_t = \mu_t + \delta_t \tag{1}$$

where $\mathrm{E}(Y_t) = \mu_t = X_t\beta$, X_t is a design matrix for the tth time point (in our example, $X_t \equiv 1$ so $\mu_t = \mu$ and we will use μ in subsequent discussions and notation), and $\{\delta_t\}$ is a second-order stationary process with covariance function C that captures the temporal dependence; $t = 1, \ldots, n$. Second-order stationarity implies that $\mathrm{var}(\delta_t) = \sigma^2$ and that the covariance between the process δ_t at times t_i and t_j can be written as

$$\mathrm{cov}(\delta_{t_i}, \delta_{t_j}) = C(|t_i - t_j|) \equiv \sigma^2\rho(|t_i - t_j|)$$

Consequently, $C(0) = \sigma^2$. We further assume that the $\{\delta_t\}$ are normally distributed. We refer to (1) and its accompanying assumptions as the *first stage* of our model.

A standard approach to modelling the covariance function is through assuming some parametric form for $C(\cdot)$, which ensures positive-definiteness and facilitates prediction. If we attempt to extend directly the approach of Daniels and Kass (1999) from the covariance matrix to the covariance function, we encounter problems ensuring positive-definiteness of the covariance function. To avoid this potential difficulty, we shall formulate our model in the spectral domain using the spectral density $f(\omega)$, where

$$C(h) = 2\sigma^2 \int_0^{\pi} f(\omega)\cos(\omega h)\mathrm{d}\omega; \qquad h = 0, 1, 2, \ldots \tag{2}$$

For n equally spaced observations, we can only estimate the spectral density f at m frequencies, where m is the smallest integer greater than or equal to $(n-1)/2$. Let

$$\hat{f}(\omega_i) = \frac{1}{n}\left[\left(\sum Y_t^* \sin(\omega_i t)\right)^2 + \left(\sum Y_t^* \cos(\omega_i t)\right)^2\right] \qquad i = 1, \ldots, m$$

where

$$Y_t^* = \frac{Y_t - \hat{\mu}_t}{\hat{\sigma}} \qquad \omega_i = \frac{2\pi i}{n}$$

and $\hat{\mu}_t$ and $\hat{\sigma}$ are the estimates obtained from the model (1). The next step is to shrink $\hat{f}$ (the periodogram) towards some prior parametric form (which may come from exploratory data analysis). By restricting the shrunken estimates to be non-negative, we can ensure that the resulting covariance function is positive-definite. To carry out the shrinkage in a statistically optimal fashion, we must first specify a prior distribution for the spectral density f.

2.1. *Prior distribution for components of f*

To specify a prior distribution on f, we need only specify a prior for the m components needed for the approximation, namely $f(\omega_i)$ with $\omega_i = 2\pi i/n$, $i = 1, \ldots, m$; note that $f(0)$ will be zero because $\{\delta_t\}$ has zero mean. Asymptotically, we know that for n odd, each periodogram ordinate $\hat{f}(\omega_i)$ will be approximately a scaled χ^2 random variable with 2 degrees of freedom and scale parameter $f(\omega_i)$, and for n even, the last periodogram ordinate, $\hat{f}(\omega_m)$, will be approximately a scaled χ^2 random variable with 1 degree of freedom. Work by Hawkins and Wixley (1986) has shown that the fourth root of χ_2^2 and χ_1^2 random variables are very well approximated by normal distributions. That is, approximately,

$$\hat{f}(\omega_i)^{1/4} \sim \mathrm{N}\left(\frac{f(\omega_i)^{1/4}}{M_i}, \frac{V_i f(\omega_i)^{1/2}}{M_i^2}\right) \qquad i = 1, \ldots, m \tag{3}$$

where for n odd, $M_i = 2^{1/4}\Gamma(1.25)$ and $V_i = 2^{1/2}\Gamma(1.5) - M_i^2$ for $i = 1, \ldots, m$; and for n even, $M_i = 2^{1/4}\Gamma(1.25)$ and $V_i = 2^{1/2}\Gamma(1.5) - M_i^2$ for $i = 1, \ldots, m-1$ and $M_m = 2^{1/4}\Gamma(0.75)/\Gamma(0.5)$ and $V_m = 2^{1/2}\Gamma(0.5) - M_m^2$. For subsequent calculations and approximations using (3), we substitute $\hat{f}$ for f in the variance term and for notational convenience, we drop the subscript on M and V. Equation (3) suggests a scale for the prior distribution for the unknown f, namely the fourth root. That is, we assume that

$$f(\omega)^{1/4} \sim (f^{\mathrm{p}}(\omega; \psi)^{1/4}, \tau^2) \tag{4}$$

where $\mathrm{TN}_q(\nu, \lambda^2)$ denotes a truncated normal distribution whose normal distribution with mean ν, variance λ^2 has been truncated below at q. Its density can be written as,

$$g(z) = \frac{\dfrac{\phi((z-\nu)/\lambda) I\{z \geqslant q\}}{\sqrt{\{2\pi\lambda^2\}}}}{\displaystyle\int_q^\infty \frac{\phi((z-\nu)/\lambda)}{\sqrt{\{2\pi\lambda^2\}}}\,\mathrm{d}z}$$

We refer to (4) as the *second stage* of our model.

In Equation (4), f^p is a specified parametric form for the spectral density and ψ is a vector containing the relevant hyperparameters. Non-informative priors for ψ will be chosen and a flat prior will be used for τ^2. If we ignore the truncation in (4), then, approximately,

$$f(\omega)^{1/4} \sim \mathrm{N}(f^{\mathrm{p}}(\omega;\psi)^{1/4}, \tau^2) \tag{5}$$

This model formulation results in shrinking the empirical covariance function towards a parametric structure, as discussed in Section 1. A simple heuristic argument for why this happens follows: From equations (3) and (5), the full conditional distribution

$$f(\omega)^{1/4}|\mu, \sigma^2, \psi, \tau^2, Y \equiv f(\omega)^{1/4}|\psi, \tau^2, \hat{f}$$

will be approximately normally distributed with (estimated) mean

$$m \equiv S\frac{\hat{f}^{1/4}}{M} + (1-S)f^p(\omega)^{1/4} \tag{6}$$

and (estimated) variance

$$v^2 \equiv \left\{\frac{M^2}{(V\hat{f}(\omega)^{1/2})} + \frac{1}{\tau^2}\right\}^{-1} \tag{7}$$

where

$$S \equiv \frac{\hat{f}(\omega)^{1/2}V/M^2}{\hat{f}(\omega)^{1/2}V/M^2 + \tau^2}$$

Thus, high spectral density values will be shrunk more towards the prior density ordinate. If we desired uniform shrinkage across the fourth-root of all frequencies, we might specify the variance term in (5) to be $\tau^2 f(\omega)^{1/2}$. In addition, we can control the amount of shrinkage by fixing the value of τ^2. However, we are really interested in f, not $f^{1/4}$. Using the same normal approximations (3) and (5), we can compute the posterior mean for $f(\omega)$ as,

$$\mathrm{E}(f(\omega)|\mu, \sigma^2, \psi, \tau^2, Y) = 3v^4 + 6v^2m^2 + m^4 \tag{8}$$

where m and v^2 are given by (6) and (7), respectively. So, although we observe linear shrinkage on the $f^{1/4}$-scale, we see non-linear shrinkage on the f-scale.

The shrinkage on the C (equivalently, ρ)-scale will be evident from the way we compute C (Section 2.2).

A number of remarks about the prior model follow. First, by specifying the model through the spectral density, we have avoided the problem of the covariance function not being positive-definite. A less parametric form for this prior might be to place a Dirichlet-process prior (Ferguson, 1973; MacEachern, 1994) on f instead of the normal prior.

2.2. *Computing C from f*

As we shall explicitly model ordinates of the spectral density, we need techniques to move from the spectral domain to the time domain. Bochner's Theorem (Bochner, 1955) allows us to write an integral that we shall evaluate numerically:

$$C(h) = 2\sigma^2 \int_0^{\pi} f(\omega)\cos(\omega h)\mathrm{d}\omega \qquad h = 0, 1, 2, \ldots \tag{9}$$

Because we are only modelling f at a discrete number of frequencies, we need to approximate this integral.

In terms of the correlation function, we seek to approximate the integral

$$\rho(h) = 2\int_0^{\pi} f(\omega)\cos(\omega h)\mathrm{d}\omega \qquad h = 0, 1, 2, \ldots \tag{10}$$

based on models of f evaluated at a discrete number of frequencies. We consider two approaches to computing $\rho(\cdot)$ based on (10) and the spectral density f evaluated at ω_i; $i = 1, \ldots, m$:

1. a discrete approximation, and
2. a numerical-integration approximation.

Both these approximations are more accurate for when the lag h is small (where we have more data) than when h is large (for which we have little information).

Discrete approximation A simple method to move from the spectral to the time domain, is to use a discrete approximation to the spectral cumulative distribution function by putting point masses at the identified frequencies. However, this only allows computation of $\rho(h)$ for $h = 0, 1, \ldots, m$, through

$$\rho(h) = \frac{\sum_{i=1}^{m} f(\omega_i)\cos(\omega_i h)}{\sum_{i=1}^{m} f(\omega_i)}$$

By evaluating f at an additional $n - m$ frequencies, obtained by linearly interpolating between the original m frequencies, we obtain

$$\rho_D(h) = \frac{\sum_{i=1}^{n} f(\omega_i)\cos(\omega_i h)}{\sum_{i=1}^{n} f(\omega_i)} \qquad h = 0, 1, 2, \ldots, n \tag{11}$$

where $\omega_i = \pi i/n$; $i = 1, \ldots, n$. The best linear predictor of a value at $n^* > n$, requires some covariances at lags larger than n. In this case, we can use (11) evaluated at $h = 0, 1, 2, \ldots, n^* - 1$, where $\omega_i = \pi i/(n^* - 1)$ and $f(\omega_i)$ can again be obtained by linearly interpolating between the original m frequencies; $i = 1, \ldots, n^* - 1$.

Numerical-integration approximation Various numerical techniques are available to approximate the integral (10). However, we are restricted by the fact that the function $f(\cdot)$ is only evaluated at m frequencies, $\omega_i = 2\pi i/n$; $i = 1, \ldots, m$. In what is to follow, we shall use Filon's numerical-integration approximation, since this is an approach designed for Fourier integrals (Filon, 1928). In using this approximation, we must be aware of three things:

1. The approximation to f must be non-negative.
2. f must integrate to 1.
3. The approximation only works for n a factor of 4.

Filon's general approach is as follows: Divide the interval $[a, b]$ into $2N$ subintervals of equal length l; that is, $l = (b - a)/(2N)$. The function $f(\cdot)$ will be approximated by a quadratic over each successive pair of subintervals. In our case, $N = n/4$, $[a, b] = [0, \pi]$, and all the observed $f(\cdot)$ at the endpoints of the subintervals are non-negative. Then, the approximate function can be evaluated analytically. Specific details on Filon's approximation are given in the Appendix.

It may occur that the quadratic approximation to $f(\cdot)$ over particular subintervals results in negative values. To adjust for this, we truncate the approximation at zero. The problem of f not integrating to unity can be handled by normalizing the integral above with $\int_a^b f(\omega)\mathrm{d}\omega$, where f is approximated as in the standard Filon formula. The Appendix should be consulted for the final modified approximation, denoted by $\rho_{\mathrm{M}}(\cdot)$.

3. COMPUTING THE SHRINKAGE ESTIMATORS

3.1. *Fully Bayesian analysis*

To simulate from the posterior distribution of $f, \mu, \sigma^2, \psi, \tau^2 | Y$, we use a Markov chain Monte Carlo (MCMC) method called the Gibbs sampler (Smith and Roberts, 1993). This involves iteratively simulating from each of the full conditional distributions derived from the posterior distribution. The full conditional distributions of $\sigma^2 | f, \mu, \psi, \tau^2, Y$ and $\tau^2 | f, \mu, \sigma^2, \psi, Y$ are both inverse gamma and the full conditional of $\mu | f, \sigma^2, \psi, \tau^2, Y$ is normal. We use a Metropolis–Hastings algorithm to sample from the full conditional distributions of $\psi | f, \mu, \sigma^2, \tau^2, Y$ and of (the components of) $f | \mu, \sigma^2, \psi, \tau^2, Y$. For ψ, we use a normal approximation to the full conditional as a candidate distribution in the Metropolis–Hastings algorithm. That is, conditional on the current values of f,

μ, σ^2, τ^2, we compute the mode and Hessian of the full conditional for ψ and use these values for the mean and variance for a normal candidate distribution. The corresponding acceptance probability α for the candidate value at the kth iteration will be $\alpha = \min(1, \alpha_0)$, where here

$$\alpha_0 = \frac{\dfrac{p(\psi^{(k)}|f, \mu, \sigma^2, \tau^2, Y)}{\hat{p}(\psi^{(k)}|f, \mu, \sigma^2, \tau^2, Y)}}{\dfrac{p(\psi^{(k-1)}|f, \mu, \sigma^2, \tau^2, Y)}{\hat{p}(\psi^{(k-1)}|f, \mu, \sigma^2, \tau^2, Y)}} \tag{12}$$

$p(\cdot)$ is the full conditional distribution of ψ, and $\hat{p}(\cdot)$ is the normal candidate distribution referred to above.

The major computational issue in our approach is efficient sampling from the full conditional distribution of the components of f, since each realization from $f|\mu$, σ^2, ψ, τ^2, Y requires inversion of an $n \times n$ matrix. As a result, we consider two possible candidate distributions for the Metropolis–Hastings algorithm that obviate the need for more than one $n \times n$ matrix inversion at each iteration of the Gibbs sampler. The first possibility is a normal (or t-) approximation to the full conditional distribution for the components of f obtained after replacing (1) with (3) and (4) with (5). The acceptance probability α for the candidate value at the kth iteration will be $\alpha = \min(1, \alpha_0)$, where here

$$\alpha_0 = \frac{\dfrac{p(f^{(k)}|\mu, \sigma^2, \psi, \tau^2, Y)}{\hat{p}(f^{(k)}|\mu, \sigma^2, \psi, \tau^2, Y)}}{\dfrac{p(f^{(k-1)}|\mu, \sigma^2, \psi, \tau^2, Y)}{\hat{p}(f^{(k-1)}|\mu, \sigma^2, \psi, \tau^2, Y)}} \tag{13}$$

$p(\cdot)$ is the full conditional distribution of the components of f, and $\hat{p}(\cdot)$ is the normal candidate distribution with mean given by (6) and variance given by (7). The second possibility is a random walk using a variance proportional to (6) but substituting in a fixed value for τ^2. A proposed choice for τ^2 will be discussed in the next section. The acceptance probability α for the candidate value at the kth iteration will be $\alpha = \min(1, \alpha_0)$, where here

$$\alpha_0 = \frac{p(f^{(k)}|\mu, \sigma^2, \psi, \tau^2, Y)}{p(f^{(k-1)}|\mu, \sigma^2, \psi, \tau^2, Y)} \tag{14}$$

We shall consider both of these possible candidate distributions in the example.

3.2. *An Approximate Analysis*

For long time series, where n is large, computations can become prohibitive. A model that approximates the various distributions given in Section 2 would be useful. We consider both fully Bayesian and empirical Bayesian (or frequentist) approximations that circumvent the computational difficulties. Intuitively, the sample size needed for these approximations to be accurate should increase with the strength of dependence in the time series.

Fully Bayesian approximation We first consider a fully Bayesian model in which inversion of the $n \times n$ covariance matrix at each iteration is not required. We replace the *first stage* (1), with the approximate model (3). Then, by combining (3) and (4), we have a simple hierarchical model (often called a normal-normal model; see Daniels and Kass, 1998) in which the data $\hat{f}^{1/4}$, and their mean, $f^{1/4}$ both follow normal distributions. This model can be fit using the Gibbs sampler in standard software such as BUGS (e.g., Spiegelhalter *et al.*, 1996). The full conditional of $f^{1/4}|\psi, \tau^2, \hat{f}$ will now be a truncated normal distribution, obtained by truncating a normal distribution, with mean and variance given by (6) and (7) respectively, at 0; the full conditionals of ψ and τ^2 will take the forms given in Section 3.1.

Empirical Bayesian approximation We attempt to simplify computations further by developing a point estimator from the approximate model (3) and the approximation (5) that can be computed directly without recourse to the Gibbs sampler. We consider the estimator for $f(\omega)$ given in (8), which is conditional on ψ and τ^2. A standard empirical Bayes approach is to plug in estimates of these parameters into (8). We estimate ψ with a maximum likelihood estimator obtained by fitting directly to the data the parametric model that yields f^{p} in (5). For the example in Section 4, f^{p} is the spectral density derived from an AR(2) process. To estimate τ^2, we use a moment estimator similar to the one used for the between-study variance in random-effects meta-analysis (Whitehead and Whitehead, 1991). Here we can assume the prior means $f^{\mathrm{p}}(\omega_i; \hat{\psi})^{1/4}$ are known, and we obtain

$$\hat{\tau}^2 = \max\left(0, \frac{\sum w_i(\hat{f}^{1/4}(\omega_i) - f^{\mathrm{p}}(\omega_i; \hat{\psi})^{1/4})^2 - m}{\sum\omega_i - \dfrac{\sum\omega_i^2}{\sum\omega_i}}\right) \tag{15}$$

where $\omega_i^{-1} = \widehat{\mathrm{var}}(\hat{f}(\omega_i)^{1/4})$ given by (3) with $\hat{f}$ substituted for f. Other estimates can be used for ψ and τ^2, but they lack the ease and simplicity of the aforementioned estimates.

Accuracy of the fully Bayesian approximation We now discuss a simple way to assess the accuracy of the approximation (3). We first run 10–20 iterations of the Gibbs sampler using the first candidate distribution for the full conditional of f (Section 3.1) and monitor the Metropolis–Hastings acceptance probabilities of f. Recall that these probabilities are derived from the quantity α_0 given by (13). The numerator and denominator of α_0 can be thought of as importance weights and the candidate densities can be viewed as importance-sampling densities for the exact full conditional distribution of $f|\mu, \sigma^2, \psi, \tau^2, Y$. If these candidate densities are also a good approximation to the exact model, the importance weights in both the numerator and the denominator will be close to one; thus, α_0 will also be close to one. So, by monitoring α_0 for a few iterations of the Gibbs sampler that was specified in Section 3.1, we can gauge the accuracy of the approximation. An example of the use of this diagnostic is given in Section 4.

4. EXAMPLE

In this section, we shall fit our hierarchical model (and the approximation) to the frequently analysed Canadian lynx data; see, for example, Moran (1953), Campbell and Walker (1977), Lim (1987) and Lin and Pourahmadi (1998). However, our goal here will not be to find the best model for the lynx data, but to illustrate the attractiveness of our approach in comparison to a non-hierarchical model fitted to these data.

The following two models are fitted to the logarithm of the Canadian lynx data. In both cases, we assume that $\mu_t \equiv \mu$. The first will consist of the original AR(2) model suggested by Moran (1953), along with a prior distribution directly on the AR(2) parameters. The second model is our fully Bayesian hierarchical model (shrinkage model), described in Section 2, where an AR(2) model is chosen for f^{p}.

Our goal here is improved prediction through better estimation of the covariance function. To assess how well our methods compare to existing methods, we use the first 80 years of the Canadian lynx data to fit models and the last 34 years to provide performance criteria.

We shall compute two quantities to evaluate the predictions. The first quantity is the posterior predictive mean squared-error (MSPE) and the second is the 95% posterior predictive prediction interval (PI). We shall estimate the posterior predictive MSPE averaged over the predictions at all 34 times, from M iterations of the MCMC algorithm, by using

$$\operatorname{ave}\left\{\frac{1}{M}\sum_{k=1}^{M}(Y^{(k)}_{\mathrm{pred},t} - Y_{\mathrm{true},t})^2 \qquad t = 81, \ldots, 114\right\} \tag{16}$$

where

$$\left\{\frac{1}{M}\sum_{k=1}^{M}(Y^{(k)}_{\mathrm{pred},t} - Y_{\mathrm{true},t})^2\right\}$$

is an estimate of the MSPE for time t, $Y^{(k)}_{\mathrm{pred},t}$ is the conditional expectation

$$E(Y_t | f^{(k)}, \mu^{(k)}, \sigma^{2(k)}, Y_j \qquad j = 1, \ldots, 80)$$

and $Y_{\mathrm{true},t}$ is the observed value of the process at time t; $t = 81, \ldots, 114$. Models with smaller posterior predictive MSPEs will be preferred. The posterior predictive PIs will be computed using the 0.025 and the 0.975 observed quantiles of the distributions of $Y_{\mathrm{pred},t}$; $t = 81, \ldots, 114$. The model with coverage of the posterior predictive PIs closer to 95% will be preferred. We note that these comparisons could also be done for other more complicated models that have been fitted to these data; such models would then play the role of f^{p} in *stage 2* of our hierarchical model.

The Markov chain of the MCMC algorithm using the first candidate distribution for the full conditional of f (Section 3.1) mixes rather slowly and the acceptance probabilities suggest that the normal approximation is not adequate here. As a result, we used the second candidate distribution (the random walk) discussed in Section 3.1.

Figure 1 shows the fitted correlation functions from the shrinkage model and from the Bayesian AR(2) model, and in addition, it shows pointwise 95% credible intervals from the shrinkage model. We choose the posterior mean of ρ as our estimate of the correlation function. This will be positive-definite, since any convex combination of positive-definite correlation functions is positive-definite. The figure shows how our estimator differs from that under the AR(2) model in terms of both location and size of peaks and valleys. Notice that the AR(2) correlation function lies within the 95% credible intervals of our model, which is perhaps not surprising since the prior f^{p} is an AR(2) model. Figure 2 displays twenty realizations from the posterior distribution of the correlation function from the shrinkage model. This portrays the uncertainty in the correlation function more accurately than the pointwise confidence intervals in Figure 1.

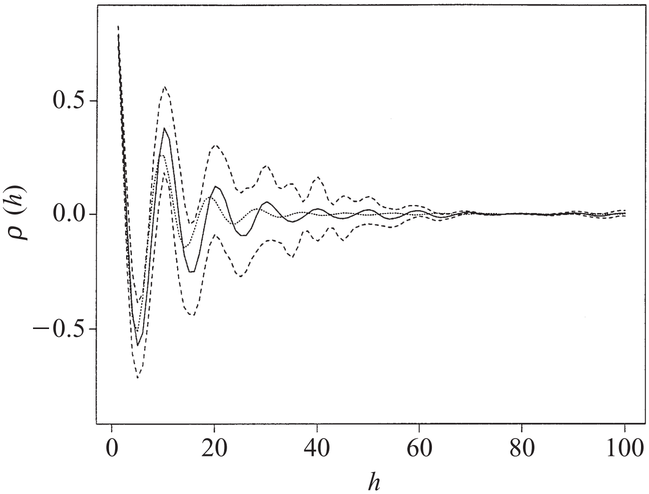

FIGURE 1. Shrinkage estimator (posterior mean) of the correlation function (solid line), Bayesian AR(2) estimator (posterior mean) of the correlation function (dotted line), and 95% credible intervals for the full Bayesian hierarchical model (A) based on the posterior distribution of ρ (dashed line)

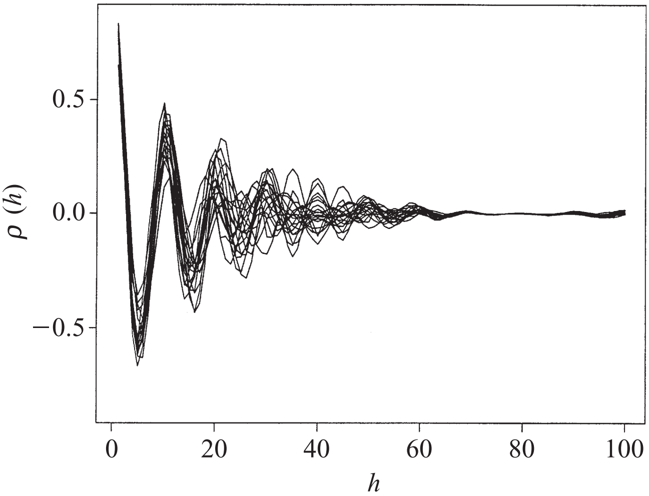

FIGURE 2. Twenty realizations from the posterior distribution of the correlation function for the fully Bayesian hierarchical model (A)

The posterior predictive MSPE averaged over all 34 times (given by equation (16)) was about 30% larger for the fitted AR(2) model, and the shrinkage model had smaller posterior predictive MSPE at 21 of the 34 times. The 95% posterior predictive PIs covered the true values 18 out of 34 times for our model, versus 12 out of 34 for the fitted AR(2) model. Notice that we are performing anywhere from a one-step-ahead forecast up to a 34-step-ahead forecast in this assessment. Clearly, the shrinkage estimator is resulting in much better predictions.

Although the normal approximation to the fourth root of the periodogram ordinates does not appear to be very accurate here, we nonetheless compared the predictions (on the log scale) derived from the approximate model, given by (3) and (5), to various other predictions using the average prediction error (APE) and the average squared prediction error (ASPE) averaged over all 34 time points,

$$\frac{1}{34}\sum_{t=81}^{114}(Y_{\mathrm{pred},t} - Y_{\mathrm{true},t}) \qquad \text{and} \qquad \frac{1}{34}\sum_{t=81}^{114}(Y_{\mathrm{pred},t} - Y_{\mathrm{true},t})^2$$

respectively. The other predictions used in the comparison were those derived from a non-Bayesian AR(2) model (using maximum likelihood) and from the posterior predictive means

$$\mathrm{E}(Y_t | Y_j,\, j = 1, \ldots, 80) \approx \frac{1}{m}\sum_{k=1}^{m} \mathrm{E}(Y_t | f^{(k)},\, \mu^{(k)},\, \sigma^{2(k)},\, Y_j,\, j = 1, \ldots, 80)$$

$$t = 81, \ldots, 114$$

assuming the two Bayesian models described at the beginning of this section. Table 1 contains the APE and the ASPE for all four models:

(A) Bayesian hierarchical model, discussed at the beginning of this section
(B) Bayesian AR(2) model, also discussed at the beginning of this section
(C) Empirical Bayesian approximate hierarchical model, discussed in Section 3.2
(D) Non-Bayesian (fitted by maximum likelihood) AR(2) model

Both shrinkage estimators, based on models (A) and (C), did better than the estimators based on the AR(2) models, namely models (B) and (D). In comparing models (A) and (B) at each $t = 81, \ldots, 114$, the shrinkage estimator from model (A) had better predictions at 27 out of the 34 time points. The analysis above suggests that the estimator based on model (C), the empirical Bayesian approximate hierarchical model, may have merit as a computationally convenient estimator, even if the model approximation is not very good (although this needs further study).

5. DISCUSSION

Our approach can be thought of as a means to smooth or shrink the spectral density towards a particular parametric form for the density. This approach, though computationally intensive, should often lead to better predictions and can be made much more computationally tenable using the approximate model, given by (3) and (5) (computationally, this only requires inversion of one $n \times n$ covariance matrix).

Current extensions of this research that we are working on include adding a measurement error component to the model, handling the case of unequally spaced observations, such as one encounters in longitudinal and spatial data, and using alternative approaches to compute the correlation function from the ordinates of the spectral density, including kernel smoothing. Other more general issues include what to use as a point estimate of the spectral density and/or covariance function; currently, we use the posterior means but other summaries of the posterior distribution might have better properties.

TABLE I
AVERAGE PREDICTION ERRORS (APE) AND AVERAGE SQUARED PREDICTION ERRORS (ASPE) FOR ALL FOUR MODELS

Model	APE	ASPE
(A) Bayesian hierarchical	−0.379	0.971
(B) Bayesian AR(2)	−0.462	1.49
(C) Empirical Bayes approx.	−0.440	1.32
(D) Non-Bayesian AR(2)	−0.491	1.54

ACKNOWLEDGMENTS

The authors' research was supported by the U.S. Environmental Protection Agency under co-operative agreement CR822919-01-0 with Iowa State University, and by the NSF under grant DMS9816630. The authors would like to thank Andrew Gelman for comments which improved the manuscript.

APPENDIX: FILON'S INTEGRAL

We wish to approximate the following integral:

$$\rho(h) = 2\int_0^{\pi} \cos(\omega h) f(\omega)\mathrm{d}\omega \qquad h = 0, 1, 2, \ldots \tag{17}$$

Filon's approximation takes the following form:

$$\rho_{\mathrm{F}}(h) = \sum_{r=1}^{N} \int_{\omega_{2r-2}}^{\omega_{2r}} \cos(\omega h) \tilde{f}(\omega)\mathrm{d}\omega \qquad h = 0, 1, 2, \ldots \tag{18}$$

where $\tilde{f}$ is a quadratic approximation to f over the interval $(\omega_{2r-2}, \omega_{2r})$ that goes through the points $f(\omega_{2r-2})$, $f(\omega_{2r-1})$, and $f(\omega_{2r})$; that is,

$$\tilde{f}(\omega) = A + B(\omega - \omega_{2r-1}) + D(\omega - \omega_{2r-1})^2$$

where

$$A = f(\omega_{2r-1})$$

$$B = \frac{f(\omega_{2r}) - f(\omega_{2r-2})}{2(\omega_{2r} - \omega_{2r-1})}$$

and

$$D = \frac{f(\omega_{2r}) + f(\omega_{2r-2}) - 2f(\omega_{2r-1})}{2(\omega_{2r} - \omega_{2r-1})^2}$$

For details, see Filon (1928).

To normalize the integral, we use the modified approximation,

$$\rho_{\mathrm{M}}(h) = \frac{\sum_{r=1}^{N} \int_{\omega_{2r-2}}^{\omega_{2r}} \cos(\omega h) \tilde{f}(\omega)\mathrm{d}\omega}{\sum_{r=1}^{N} \int_{\omega_{2r-2}}^{\omega_{2r}} \tilde{f}(\omega)\mathrm{d}\omega} \qquad h = 0, 1, 2, \ldots \tag{19}$$

where if there are intervals (c, d) within the subinterval $(\omega_{2r-2}, \omega_{2r})$ such that $\tilde{f} < 0$, we replace

$$\int_{\omega_{2r-2}}^{\omega_{2r}} \cos(\omega h) \tilde{f}(\omega)\mathrm{d}\omega \qquad \text{and} \qquad \int_{\omega_{2r-2}}^{\omega_{2r}} \tilde{f}(\omega)\mathrm{d}\omega$$

respectively, with,

$$\int_{\omega_{2r-2}}^{c} \cos(\omega h) \tilde{f}(\omega)\mathrm{d}\omega + \int_{d}^{\omega_{2r}} \cos(\omega h) \tilde{f}(\omega)\mathrm{d}\omega$$

and

$$\int_{\omega_{2r-2}}^{c} \tilde{f}(\omega)\mathrm{d}\omega + \int_{d}^{\omega_{2r}} \tilde{f}(\omega)\mathrm{d}\omega$$

REFERENCES

BOCHNER, S. (1955) *Harmonic Analysis and the Theory of Probability.* Berkeley and Los Angeles: University of California Press.

CAMPBELL, M. J. and WALKER, A. M. (1977) A survey of statistical work on the Mackenzie river series of annual Canadian lynx trappings for the years 1821–1934 and a new analysis. *Journal of the Royal Statistical Society, Series B* 40, 411–31.

DANIELS, M. J. and KASS, R. E. (1998) A note on first stage approximation in two stage hierarchical models. *Sankhya, Series B* 60, 19–30.

—— (1999) Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* 94, 1254–63.

FERGUSON, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1, 209–30.

FILON, L. N. G. (1928) On a quadrature formula for trigonometric integrals. *Proceedings of the Royal Society of Edinburgh* 49, 38–47.

HALL, P., FISHER, N. I. and HOFFMAN, B. (1994) On the nonparametric estimation of covariance functions. *The Annals of Statistics* 22, 2115–2134.

HAWKINS, D. M. and WIXLEY, R. A. J. (1986) A note on the transformation of chi-squared variables to normality. *The American Statistician* 40, 296–8.

LIM, K. S. (1987) A comparative study of various univariate time series models for Canadian lynx data. *Journal of Time Series Analysis* 8, 161–76.

LIN, T. C. and POURAHMADI, M. (1998) Nonparametric and non-linear models and data mining in time series: A case-study on the Canadian lynx data. *Applied Statistics* 47, 187–201.

MACEACHERN, S. N. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics, Part B – Simulation and Computation* 23, 727–41.

MORAN, P. A. P. (1953) The statistical analysis of the Canadian lynx cycle, I. *Australian Journal of Zoology* 1, 163–73.

SMITH, A. F. M. and ROBERTS, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 55, 3–23.

SPIEGELHALTER, D. J., BEST, N. G., GILKS, W. R. and INSKIP, H. (1996) Hepatitis B: A case study in MCMC methods. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter). London: Chapman and Hall, pp. 339–58.

WHITEHEAD, A. and WHITEHEAD, J. (1991) A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 10, 1665–77.