ASYMPTOTIC THEORY FOR KRIGING WITH ESTIMATED PARAMETERS AND ITS APPLICATION TO NETWORK DESIGN

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
rls@email.unc.edu

July 27, 2004

Abstract

A second-order expansion is established for predictive distributions in Gaussian processes with estimated covariances. Particular focus is on estimating quantiles of the predictive distribution and their subsequent application to prediction intervals. Two basic approaches are considered, (a) a "plug-in" approach using the restricted maximum likelihood estimate of the covariance parameters, (b) a Bayesian approach using general priors. Calculations of "coverage probability bias" show that the Bayesian approach is superior in the tails of the predictive distributions, regardless of the prior. However they also imply the existence of a "matching prior" for which the second-order coverage probability bias vanishes. Previously suggested frequentist corrections do not have this property, but we use our results to suggest a new frequentist approach that does. We also compute the expected length of a Bayesian prediction interval, suggesting that this might be used as a design criterion combining recent "estimative" and "predictive" approaches to network design. A surprising parallel emerges with the recent two-stage estimative-predictive approach of Zhu and Stein.

Keywords: Design of monitoring networks; Predictive inference; Second-order asymptotics; Spatial statistics.

1 Introduction

Many of the techniques of spatial statistics are built around the assumption that a set of observations — for example, measurement of atmospheric pollutants or meteorological variables at a finite network of stations, or some transformation of those variables — are samples from a Gaussian random field. The mean of the random field may be constant or may be representable as a linear combination of covariates, while the covariance function is chosen from a parametric family of positive-definite covariances. *Kriging* is a technique for predicting unobserved values of the random field through linear combinations of the observed variables. The weights are chosen to minimize the mean squared prediction error subject to an unbiasedness constraint. Ordinary kriging is applied when the mean of the process is an unknown constant, and universal kriging when it is a linear

combination of covariates. Often the covariance structure of the process is represented in terms of the variogram instead of the covariance function itself, which is a little more general because the variogram may exist in certain circumstances when the ordinary covariance function does not, but we shall not consider that as a separate case of the present paper. Numerous books, e.g. Ripley (1981), Cressie (1993), Chilès and Delfiner (1999), Stein (1999) have presented all these concepts in detail.

One widely recognized difficulty with these methods is that the usual formula for the mean squared prediction error of a kriging predictor does not take into account the estimation of the covariance model parameters. Typically in geostatistics, the estimation of the covariance model parameters is performed first, then the estimated model is used to construct the predictor, but in the second stage, the covariance parameters are treated as if they were known. Because of this, it is widely assumed that the prediction standard errors derived through kriging are underestimates of the true prediction standard deviations, even when the form of the model is correct. Zimmerman and Cressie (1992) and Stein (1999) have proposed approximate techniques for estimating the mean squared prediction error with estimated model parameters, but these techniques are somewhat *ad hoc*, and it is usually still assumed that the predictive distribution is normal. Bayesian methods were first proposed for these problems in the early 1990s (see e.g. Le and Zidek 1992, Handcock and Stein 1993, Brown, Le and Zidek 1994) and are often believed to be superior to standard kriging methods, because they take into account the uncertainty of model parameters and also do not rely on any normality assumptions about the posterior distributions of model parameters and predictions. However, despite much speculation about the issue (e.g. Stein 1999, Berger, De Oliveira and Sanso 2001), there is no proof in general that Bayesian methods are superior when assessed, for example, by how closely the true coverage probability of a prediction interval matches the nominal coverage probability.

The present paper examines these issues through second-order asymptotics. For reasons that will become clear in Section 2, our preferred method of estimation in spatial processes is restricted maximum likelihood (REML), and the "plug-in" approach to prediction uses ordinary or universal kriging, substituting the covariance model parameters by their REML estimates. We compare prediction intervals constructed by these methods with Bayesian prediction intervals. The latter may be simplified by using a Laplace approximation to the integral, and one of the by-products of the paper is a method of constructing approximate Bayesian prediction intervals without using Markov Chain Monte Carlo methods. Our main results, however, are approximate formulae for the coverage probability bias of both plug-in and Bayesian prediction intervals, and for the expected length of a prediction interval. The latter is potentially valuable as a design criterion. The coverage probability results do not produce a universal conclusion that either predictor is better than the other, but we show that as the desired coverage probability approaches 1, the second-order approximation to the Bayesian coverage probability bias is always smaller than that of the plug-in approach. This result holds regardless of the prior density, provided it satisfies some smoothness conditions. However, we also examine the possibility that the prior may be chosen so that the second-order coverage probability bias vanishes entirely, the case of a so-called "matching prior".

Although the paper was motivated by the problem of spatial prediction, there is nothing explicitly "spatial" about the results. The mathematical framework is that of a Gaussian process with mean linearly dependent on covariates, whose covariance function depends on a finite-dimensional vector of unknown parameters. This may include time series models such as ARMA processes, or variance components models such as mixed effects ANOVA. Also in the spatial case, although it

is common to assume that spatial covariances (or variograms) are stationary and isotropic, those assumptions are not essential for our methods either. The results may also be applied to nonstationary processes or spatial-temporal models provided the covariance functions are parametric.

2 Mathematical framework

Suppose we have an n-dimensional vector of observations Y, and are interested in predicting some unobserved scalar value of the random field Y_0. In practice we may well be interested in predicting more than one Y_0 from a given Y, but we do not treat that case separately, assuming that it suffices to consider one Y_0 at a time. The joint density of Y and Y_0 is assumed to be of the form

$$\begin{pmatrix} Y \\ Y_0 \end{pmatrix} \sim N\left[\begin{pmatrix} X\eta \\ x_0^T\eta \end{pmatrix}, \begin{pmatrix} V(\theta) & w^T(\theta) \\ w(\theta) & v_0(\theta) \end{pmatrix}\right] \tag{1}$$

where X and x_0 are known matrices of regressors of dimensions $n \times q$ and $q \times 1$ respectively, and η is an unknown $q \times 1$ vector of regression coefficients. The ordinary kriging model is the special case of (1) where η is of dimension 1, $X = \mathbf{1}^T$ ($\mathbf{1}$ is a column vector of ones) and $x_0 = 1$. We assume $V(\theta)$, $w(\theta)$ and $v(\theta)$ are covariance elements that are all known functions of an unknown p-dimensional parameter vector θ. Where there is no ambiguity, we shall simply write V, w and v without indicating explicitly the dependence on θ. Define $\ell_n(\theta)$ to be the restricted log likelihood function

$$e^{\ell_n(\theta)} = (2\pi)^{-(n-q)/2}|X^TX|^{1/2}|V|^{-1/2}|X^TV^{-1}X|^{-1/2}\exp\left(-\frac{G^2}{2}\right), \tag{2}$$

where $G^2 = G^2(\theta) = Y^T\left\{V^{-1} - V^{-1}X(X^TVX)^{-1}X^TV^{-1}\right\}Y$ is the generalized residual sum of squares. Also let

$$\lambda = V^{-1}w + V^{-1}X(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w), \tag{3}$$

$$\sigma_0^2 = v_0 - w^TV^{-1}w + (x_0^T - w^TV^{-1}X)(X^TV^{-1}X)^{-1}(x_0 - X^TV^{-1}w). \tag{4}$$

Lemma 1. Assume (1) holds and let $f_n(Y;\eta,\theta)$, $f_{n+1}(Y,Y_0;\eta,\theta)$ denote respectively the density of Y, and the joint density of Y and Y_0. Then

$$\int f_n(Y;\eta,\theta)d\eta = |X^TX|^{-1/2}e^{\ell_n(\theta)}, \tag{5}$$

$$\int f_{n+1}(Y,Y_0;\eta,\theta)d\eta = |X^TX|^{-1/2}e^{\ell_n(\theta)} \cdot \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \lambda^TY}{\sigma_0}\right)^2\right\} \tag{6}$$

Remark. Equation (5) is due to Harville (1974); (6) follows by an extension of Harville's argument. The detailed proof is in Section 8.1.

In the case that θ is known and η unknown with a uniform (improper) prior density, (5) and (6) show that the predictive density of Y_0 given Y is

$$\frac{\int f_{n+1}(Y,Y_0;\eta,\theta)d\eta}{\int f_n(Y;\eta,\theta)d\eta} = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \lambda^TY}{\sigma_0}\right)^2\right\} \tag{7}$$

— in other words, a normal density with mean $\lambda^T Y$ and variance σ_0^2. Thus in this case the Bayesian predictor coincides with universal kriging, usually derived by the frequentist argument of choosing λ to minimize $E\{(Y_0 - \lambda^T Y)^2\}$ subject to the unbiasedness constraint $E\{Y_0 - \lambda^T Y\} = 0$. This simple equivalence of Bayesian and frequentist arguments shows that, in this case, Bayesian prediction intervals have exactly the correct coverage probabilities.

Now let us consider the more complicated case that θ is unknown, with prior density $\pi(\theta)$. We continue to assume a uniform prior for η (independent of θ), but we allow $\pi(\theta)$ to be arbitrary provided it exists and is differentiable over a region that includes the true θ. The posterior predictive density of Y_0 given Y in this case is

$$\frac{\int\int f_{n+1}(Y, Y_0; \eta, \theta) d\eta d\theta}{\int\int f_n(Y; \eta, \theta) d\eta d\theta} = \frac{\int e^{\ell_n(\theta)} \cdot \frac{1}{\sqrt{2\sigma_0^2}} \exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \lambda^T Y}{\sigma_0}\right)^2\right\} \cdot \pi(\theta) d\theta}{\int e^{\ell_n(\theta)} \cdot \pi(\theta) d\theta}. \tag{8}$$

Alternatively, if we are interested in the predictive distribution function, we may define

$$\psi(z \ ; \ Y, \theta) = \Phi\left(\frac{z - \lambda^T Y}{\sigma_0}\right)$$

where Φ is the standard normal distribution function, and (8) leads to the predictive distribution function

$$\tilde{\psi}(z \ ; \ Y) = \frac{\int e^{\ell_n(\theta) + Q(\theta)} \psi(z \ ; \ Y, \theta) d\theta}{\int e^{\ell_n(\theta) + Q(\theta)} d\theta} \tag{9}$$

where we have written $Q(\theta) = \log \pi(\theta)$ and, here and subsequently, we use a tilde to denote a Bayesian estimate or predictor.

In contrast to (9), we also consider the plug-in predictor

$$\hat{\psi}(z \ ; \ Y) = \psi(z \ ; \ Y, \hat{\theta}) \tag{10}$$

where $\hat{\theta}$ is the REML estimator, i.e. the value of θ that maximizes $\ell_n(\theta)$.

3 Laplace approximation

First, we introduce some notation. We use superscripts to denote components of θ, such as θ^i for the ith component. Where we use scalar functions of θ, such as $\psi(z \ ; \ Y, \theta)$ (with z and Y held constant for the time being) or $Q(\theta)$, subscripts will indicate differentiation with respect to the components of θ. Thus $Q_i = \frac{\partial Q}{\partial \theta^i}$, $\psi_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}$, etc. We also define $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta^i}$, $U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}$, $U_{ijk} = \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}$. All of these quantities are functions of a particular θ, and when we evaluate them at the REML estimator $\hat{\theta}$, we denote this with a hat, e.g. $\hat{U}_i$, $\hat{\psi}_{ij}$, etc. To avoid making the notation still more complicated, we do not indicate explicitly that all these quantities also depend on n, but it is assumed that $\ell_n(\theta)$ and all of its derivatives are of $O_p(n)$, to be consistent with regular maximum likelihood theory for i.i.d. observations, while quantities such as Q and ψ, and their derivatives, are of $O_p(1)$. In the context of asymptotic theory for spatial processes, these conditions assume that we are working in the framework of "increasing domain asymptotics", as

defined by Mardia and Marshall (1984), rather than the alternative "infill asymptotics" that have been popularized by Stein (1999). We have $\hat{U}_i = 0$ (because the REML estimator is defined as the local maximum of ℓ_n), while the values of $-\hat{U}_{ij}$ form the observed information matrix. We assume the latter matrix is invertible and denote the inverse matrix with superscripts rather than subscripts. In other words, if G is the $p \times p$ matrix whose (i,j) entry is $\hat{U}_{ij}$, then G^{-1} exists and its (i,j) entry is $\hat{U}^{ij}$. Throughout we use the summation convention, that where a repeated index appears as both a subscript and a superscript in the same formula, summation over that index is implicit.

With these notations, two applications of formulae (8.3.50)–(8.3.55) in Chapter 8 of Bleistein and Handelsman (1986), to the numerator and denominator of (9), lead to the result

$$\tilde{\psi} - \hat{\psi} \;=\; \frac{1}{2}\hat{U}_{ijk}\hat{\psi}_\ell \hat{U}^{ij}\hat{U}^{k\ell} - \frac{1}{2}(\hat{\psi}_{ij} + 2\hat{\psi}_i\hat{Q}_j)\hat{U}^{ij} + O_p(n^{-2}). \tag{11}$$

It should be noted that (11) involves Taylor expansion of both the numerator and denominator of (9) about the REML estimator $\hat{\theta}$. An alternative approximation, due to Tierney and Kadane (1986), assumes that the integrands in the numerator and denominator of (9) are separately maximized with respect to θ; the approximation to $\tilde{\psi}$ is then the ratio of the maximized integrands. As shown by Tierney and Kadane, this is accurate to $O_p(n^{-2})$ without any additional correction terms based on Taylor expansion. Compared with (11), the Tierney-Kadane method avoids the need to evaluate U_i, U_{ij}, etc., which can be cumbersome. However for the purposes of the present paper, in particular the theoretical calculations of Sections 4 and 5, the explicit expression (11) is more useful.

We can also apply these arguments to the inverse of the predictive distribution function. Suppose we want to find $z_P = z_P(Y)$ to solve the equation $\psi(z_P\,;\, Y, \theta) = P$ for some given P. The plug-in estimator in this case is defined by setting (10) equal to P; this leads to $\hat{z}_P = \hat{\lambda}^T Y + \hat{\sigma}_0 \Phi^{-1}(P)$ where the hats over λ and σ_0 denote that they are evaluated at $\hat{\theta}$. The Bayes estimator $\tilde{z}_P$ is defined to be the P-quantile of the Bayesian predictive distribution function; in other words, choose $\tilde{z}_P$ to solve $\tilde{\psi}(\tilde{z}_P\,;\, Y) = P$. We use primes on ψ to denote differentiation with respect to z and subscripts,as previously, for differentiation with respect to components of θ. With these definitions, Taylor expansion of ψ about $\hat{z}_P$ leads to the approximation

$$\tilde{z}_P - \hat{z}_P \;=\; -\frac{\tilde{\psi}(\hat{z}_P) - \hat{\psi}(\hat{z}_P)}{\hat{\psi}'(\hat{z}_P)} + O_p(n^{-2}). \tag{12}$$

To evaluate $\tilde{\psi}(\hat{z}_P) - \hat{\psi}(\hat{z}_P)$, we either use directly the approximation (11) with $z = \hat{z}_P$, or else apply the Tierney-Kadane approximation as noted in the discussion immediately following (11).

For the evaluation of derivatives of ℓ_n that are used in (11), see Section 8.2.

To summarize this section, formula (11) for the predictive distribution function or (12) for its inverse provide an alternative to the popular Markov chain Monte Carlo (MCMC) methods for Bayesian computation, that have the advantage of being explicit formulae, not requiring consideration of such issues as the number of MCMC iterations which, despite much research, remains a source of possible large bias in routine implementation of MCMC. For the calculation of prediction intervals, a natural way to proceed is to select two values of P, say P_1 and P_2, such that $P_2 - P_1$ is the desired coverage probability. For example, for a 95% prediction interval, obvious values are $P_1 = 0.025,\ P_2 = 0.975$. Then $(\hat{z}_{P_1}, \hat{z}_{P_2})$ defines the plug-in prediction interval based on the

REML estimator, and $(\tilde{z}_{P_1}, \tilde{z}_{P_2})$ is its Bayesian analogue. Equation (12) may be used to define an approximation, accurate to $O_p(n^{-2})$, to the Bayesian interval. The question that now arises is to what extent either of these intervals actually achieves its nominal coverage probability, in the sense that in repeated sampling, the probability that $\hat{z}_{P_1} < Y_0 < \hat{z}_{P_2}$ or $\tilde{z}_{P_1} < Y_0 < \tilde{z}_{P_2}$ comes close to $P_2 - P_1$. These questions may be approached by calculating the *coverage probability bias*, to which we now turn.

4 Asymptotic approximation to the coverage probability bias

We need some further notation. Write $U_i = n^{1/2} Z_i$, $U_{ij} = n\kappa_{ij} + n^{1/2} Z_{ij}$, $U_{ijk} = n\kappa_{ijk} + n^{1/2} Z_{ijk}$ where $\kappa_{ij}, \kappa_{ijk}$ are non-random and Z_i, Z_{ij}, Z_{ijk} are random with mean 0; it is part of the assumption that all these quantities are $O(1)$ or $O_p(1)$ as $n \to \infty$. Also let $\kappa_{i,j} = E\{Z_i Z_j\}$, $\kappa_{ij,k} = E\{Z_{ij} Z_k\}$. By a standard identity, $\kappa_{i,j} = -\kappa_{ij}$ and is the (i,j) entry of the (normalized) Fisher information matrix; we assume this matrix is invertible with inverse entries $\kappa^{i,j}$. Explicit formulae exist for calculating these quantities; see Section 8.2. In this section, when quantities such as U_{ijk} or ψ_{ij} are indicated without hats, it is assumed that they are evaluated at the true θ. Let $W = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}$.

In this notation, a standard Taylor expansion of ℓ_n yields the approximation

$$\begin{aligned} \hat{\psi} - \psi &= n^{-1/2} \kappa^{i,j} Z_i \psi_j + n^{-1} \left(\kappa^{i,j} \kappa^{k,\ell} Z_{ik} Z_j \psi_\ell + \frac{1}{2} \kappa^{i,r} \kappa^{j,s} \kappa^{k,t} \kappa_{ijk} Z_r Z_s \psi_t + \frac{1}{2} \kappa^{i,j} \kappa^{k,\ell} Z_i Z_k \psi_{j\ell} \right) \\ &\quad + O_p(n^{-3/2}). \end{aligned} \tag{13}$$

Equation (11) shows that

$$\tilde{\psi} - \hat{\psi} = \frac{1}{2n} \left\{ \kappa_{ijk} \kappa^{i,j} \kappa^{k,\ell} \psi_\ell + (\psi_{ij} + 2\psi_i Q_j) \kappa^{i,j} \right\} + O_p(n^{-3/2}). \tag{14}$$

Equations (13) and (14) provide approximations for both $\hat{\psi} - \psi$ and $\tilde{\psi} - \psi$, accurate to $O_p(n^{-1})$, and these are the basis for all subsequent asymptotic manipulations.

Now consider the inverse of the predictive distribution function. We write $\psi = \psi(z \ ; \ Y, \theta)$ to indicate explicitly the dependence on z and Y, and also let $\psi^*(z \ ; \ Y)$ denote an estimator of $\psi(z \ ; \ Y, \theta)$; ψ^* could be either $\hat{\psi}$ or $\tilde{\psi}$. Assume ψ^* has an expansion

$$\psi^*(z \ ; \ Y) = \psi(z \ ; \ Y, \theta) + n^{-1/2} R(z, Y) + n^{-1} S(z, Y) + o_p(n^{-1}), \tag{15}$$

We define the true and estimated P-quantiles of the predictive distribution, $z_P = z_P(Y, \theta)$ and $z_P^* = z_P^*(Y)$, by the equations $\psi^*(z_P^* \ ; \ Y) = \psi(z_P \ ; \ Y, \theta) = P$. Then further asymptotic arguments show that

$$\begin{aligned} z_P^* - z_P &= -n^{-1/2} \frac{R(z_P, Y)}{\psi'(z_P \ ; \ Y, \theta)} + n^{-1} \left\{ \frac{R(z_P, Y) R'(z_P, Y)}{\psi'^2(z_P \ ; \ Y, \theta)} - \frac{1}{2} \frac{R^2(z_P, Y) \psi''(z_P \ ; \ Y, \theta)}{\psi'^3(z_P \ ; \ Y, \theta)} \right. \\ &\quad \left. - \frac{S(z_P, Y)}{\psi'(z_P \ ; \ Y, \theta)} \right\} + o_p(n^{-1}). \end{aligned} \tag{16}$$

and

$$\begin{aligned}\psi(z_P^*\ ;\ Y,\theta) - \psi(z_P\ ;\ Y,\theta) &= -n^{-1/2}R(z_P,Y) + n^{-1}\left\{\frac{R(z_P,Y)R'(z_P,Y)}{\psi'(z_P\ ;\ Y,\theta)} - S(z_P,Y)\right\} \\ &\quad + o_p(n^{-1}). \end{aligned} \tag{17}$$

Where there is no danger of confusion we omit the arguments $z_P,\ Y,\ \theta$.

The argument leading to (16) and (17) is essentially that of Cox (1975), but because the precise form of the result is different from Cox's, we provide an independent derivation in Section 8.3.

The expected value of (17) will be called the *coverage probability bias*; it represents the discrepancy between $\Pr\{Y_0 \le z_P^* \mid Y,\theta\}$ and the target probability P. The expected value of (16) is also of interest, in connection with the expected length of a Bayesian prediction interval; this is further discussed in Section 5. By (13) and (14), for both $\hat{z}_P$ and $\tilde{z}_P$, $R = \kappa^{i,j}Z_i\psi_j$. For $\hat{z}_P$, we have $S = \kappa^{i,j}\kappa^{k,\ell}Z_{ik}Z_j\psi_\ell + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\psi_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,\ell}Z_iZ_k\psi_{j\ell}$; we subsequently denote this by S_1. For $\tilde{z}_P$, the corresponding expression is $S_2 = S_1 + \frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,\ell}\psi_\ell + \left(\frac{1}{2}\psi_{ij} + \psi_iQ_j\right)\kappa^{i,j}$.

After considerable manipulations, given in detail in Section 8.4, the coverage probability bias in $\hat{z}_P$ reduces to

$$\begin{aligned} & nE\{\psi(\hat{z}_P\ ;\ Y,\theta) - \psi(z_P\ ;\ Y,\theta)\} \\ \approx\ & \phi(\Phi^{-1}(P))\Phi^{-1}(P)\Big\{\kappa^{i,j}\kappa^{k,\ell}\frac{\sigma_{0\ell}}{\sigma_0}\left(\kappa_{ik,j} + \frac{1}{2}\kappa_{ijk}\right) \\ & + \frac{1}{2}\kappa^{i,j}\left(\frac{\sigma_{0ij}}{\sigma_0} - \frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\Phi^{-1}(P)^2\right) - \frac{1}{2\sigma_0^2}\kappa^{i,j}\lambda_i^TV\lambda_j \\ & - \frac{1}{2n\sigma_0^2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^TV\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell + \lambda_j^TV\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right)\Big\}. \end{aligned} \tag{18}$$

The corresponding result for $\tilde{z}_P$ is

$$\begin{aligned} & nE\{\psi(\tilde{z}_P\ ;\ Y,\theta) - \psi(z_P\ ;\ Y,\theta)\} \\ \approx\ & \phi(\Phi^{-1}(P))\Phi^{-1}(P)\Big\{\kappa^{i,j}\kappa^{k,\ell}\frac{\sigma_{0\ell}}{\sigma_0}(\kappa_{ik,j} + \kappa_{ijk}) \\ & + \kappa^{j,\ell}\left(\frac{\sigma_{0j\ell}}{\sigma_0} - \frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}\right) + \kappa^{i,j}Q_j\frac{\sigma_{0i}}{\sigma_0} \\ & - \frac{1}{2n\sigma_0^2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^TV\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell + \lambda_j^TV\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right)\Big\}. \end{aligned} \tag{19}$$

We now make some comments about the general form of the results (18) and (19). Nothing in the results up to this point suggests any reason why either one of (18) or (19) should dominate the other universally. In particular, it is entirely plausible that there are situations when the plug-in approach leads to smaller coverage probability bias than the Bayesian approach. However, the most interesting cases are as $P \to 0$ or $P \to 1$ — the limiting cases when we want to be nearly certain that our prediction interval covers the true value. These cases are symmetric so we consider only $P \to 1$. In this case, (18) shows that the dominant term in the coverage probability bias is

$$-\frac{1}{2}\phi(\Phi^{-1}(P))\Phi^{-1}(P)^3\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}$$

whereas in (19), all the terms are of $O(\phi(\Phi^{-1}(P))\Phi^{-1}(P))$. Since $\Phi^{-1}(P) \to \infty$ as $P \to 1$, this suggests that the coverage probability bias of the plug-in predictor is bigger by $O(\Phi^{-1}(P)^2)$, as $P \to 1$, compared with the Bayesian predictor. Also this result does not depend on any particular choice of the prior density since the role of $Q_j = \frac{\partial}{\partial\theta^j}\log\pi(\theta)$ is relatively unimportant for this comparison.

However the form of (19) suggests another possibility: if we can choose π so that

$$\begin{aligned}&\kappa^{i,j}\kappa^{k,\ell}\frac{\sigma_{0\ell}}{\sigma_0}(\kappa_{ik,j}+\kappa_{ijk})+\kappa^{j,\ell}\left(\frac{\sigma_{0j\ell}}{\sigma_0}-\frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}\right)+\kappa^{i,j}Q_j\frac{\sigma_{0i}}{\sigma_0}\\ &-\frac{1}{2n\sigma_0^2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell+\lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right) \;=\; 0. \qquad (20)\end{aligned}$$

then the second-order coverage probability bias of the Bayesian predictor is 0.

Equation (20) is in the form of a first-order linear partial differential equation, of a structure that typically arises in the literature of matching priors for confidence intervals, see e.g. Datta and Ghosh (1995). Levine and Casella (2003) have collected and reviewed several methods for numerical solution of such equations. For reasons to be explained in Section 6, it may not be worth the effort actually to solve these equations, but the existence of a matching prior is still an important result qualitatively. For instance, even if we do not try to find a matching prior, we could still compare and rank different priors (such as the flat prior, the Jeffreys prior, or one of the different forms of reference prior developed by Berger *et al.* (2001)) according to how close they come to satisfying (20).

5 Expected length of a Bayesian prediction interval

We now give the corresponding calculations based on (16). Detailed calculations, given in Section 8.5, show that

$$nE\{\hat{z}_P(Y)-z_P(Y,\theta)\} \;\approx\; \Phi^{-1}(P)\left\{\kappa^{i,j}\kappa^{k,\ell}\sigma_{0\ell}\left(\kappa_{ik,j}+\frac{1}{2}\kappa_{ijk}\right)+\frac{1}{2}\kappa^{i,j}\sigma_{0ij}\right\} \qquad (21)$$

and

$$\begin{aligned}&nE\{\tilde{z}_P(Y)-z_P(Y,\theta)\} \;\approx\; \Phi^{-1}(P)\Big\{\kappa^{i,j}\kappa^{k,\ell}\sigma_{0\ell}(\kappa_{ik,j}+\kappa_{ijk})\\ &+\kappa^{i,j}\left(\sigma_{0ij}-\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}\right)+\kappa^{i,j}Q_j\sigma_{0i}+\frac{1}{2}\Phi^{-1}(P)^2\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}+\frac{1}{2}\kappa^{i,j}\frac{\lambda_i^T V\lambda_j}{\sigma_0}\Big\}. \qquad (22)\end{aligned}$$

In this case, the asymptotics as $P \to 1$ are the other way round from (18) and (19): the order of magnitude is $O(\Phi^{-1}(P))$ for $\hat{\psi}$ but $O\left(\Phi^{-1}(P)^3\right)$ for $\tilde{\psi}$. This, however, is what we would expect: the Bayesian prediction interval achieves more accurate coverage probability than the plug-in interval, but at the cost that it is a longer interval. This discrepancy is reflected in the $O\left(\Phi^{-1}(P)^3\right)$ term in (22.

We now discuss the consequences of these formulae for the length of a prediction interval. Suppose we choose P_1 and P_2 so that $P_2 - P_1$ is the desired coverage probability. For example,

for a 95% prediction interval we would most likely choose $P_1 = 0.025,\ P_2 = 0.975$. The Bayesian prediction interval is $(\tilde{z}_{P_1}, \tilde{z}_{P_2})$, and its expected length is

$$\begin{aligned} E\{\tilde{z}_{P_2} - \tilde{z}_{P_1}\} &= E\{z_{P_2} - z_{P_1}\} + E\{\tilde{z}_{P_2} - z_{P_2}\} - E\{\tilde{z}_{P_1} - z_{P_1}\} \\ &= \sigma_0\{\Phi^{-1}(P_2) - \Phi^{-1}(P_1)\} + E\{\tilde{z}_{P_2} - z_{P_2}\} - E\{\tilde{z}_{P_1} - z_{P_1}\}. \end{aligned} \quad (23)$$

Equation (23) might be used as a basis for network design. Suppose we are choosing locations for a network whose purpose is to predict a quantity Y_0. For example, the observations $Y_1, ..., Y_n$ might be measurements of atmospheric particulate matter at n air pollution monitors, and Y_0 might be the average (or a population-weighted average) over a region, that could be used as a predictor of human health effects. In this context, it is desirable to estimate Y_0 as accurately as possible, but because precise measurement is not possible, a prediction interval for Y_0, based on spatial interpolation from $Y_1, ..., Y_n$, is desirable as a means of reflecting uncertainty in the interpolation. Based on the results of Section 4, we propose a Bayesian prediction interval to minimize the coverage probability bias. When choosing among different possible layouts of the network, we propose minimizing (23) for given P_1 and P_2.

The structure of (23) is in two parts. If we were solely interested in the first term, we would choose the design to minimize σ_0, the predictive standard error of Y_0 assuming θ is known. In recent research such as Zhu (2002), this known as a "predictive criterion" for network design. The second and third terms in (23) account for the error due to estimation of θ. Choosing the design to optimize parameter estimation leads to so-called estimative criteria for design. The merit of (23) is that it is a combined criterion that accounts for both estimation and prediction. Using (22) to approximate (23) provides a practical means of evaluating this design criterion. It remains to investigate its practical implications.

6 Frequentist corrections to plug-in prediction intervals

A major focus of our results up to this point has been that the plug-in approach to prediction intervals typically underestimates the variability of the predictive distribution, and Bayesian methods may correct for that by reducing the coverage probability bias. But the disadvantages of the plug-in procedure are well known, and previous research has led to corrections derived from a frequentist perspective. It is therefore natural to ask to what extent the properties of Bayesian procedures compare with those of frequentist corrections to the plug-in approach.

The best known frequentist correction was derived by Harville and Jeske (1992) and also by Zimmerman and Cressie (1992). Starting with the identity

$$E\left\{(\hat{\lambda}^T Y - Y_0)^2\right\} = E\left\{(\lambda^T Y - Y_0)^2\right\} + E\left\{(\hat{\lambda}^T Y - \lambda^T Y)^2\right\}, \quad (24)$$

these authors used a first-order Taylor expansion to write the second term in (24) approximately as

$$E\left\{(\hat{\theta}^i - \theta^i)(\hat{\theta}^j - \theta^j)\lambda_i^T Y Y^T \lambda_j\right\} \quad (25)$$

and then, effectively assuming $\hat{\theta}$ independent of Y, replaced (25) by

$$n^{-1}\kappa^{i,j}\lambda_i^T V \lambda_j.$$

The corrected expression for (24) is then

$$E\left\{(\hat{\lambda}^T Y - Y_0)^2\right\} \approx \sigma_0^2 + n^{-1}\kappa^{i,j}\lambda_i^T V \lambda_j. \tag{26}$$

One possible approach to a predictive distribution would simply use (26) in place of σ_0^2 for the predictive variance, but otherwise assume normality. Thus, in place of $\hat{z} = \hat{\lambda}^T Y + \hat{\sigma}_0 \Phi^{-1}(P)$, we write

$$\begin{aligned} z_P^* &= \hat{\lambda}^T Y + \sqrt{\hat{\sigma}_0^2 + n^{-1}\hat{\kappa}^{i,j}\hat{\lambda}_i^T \hat{V} \hat{\lambda}_j}\,\Phi^{-1}(P) \\ &= \hat{z}_P + \frac{1}{2n\sigma_0}\hat{\kappa}^{i,j}\hat{\lambda}_i^T \hat{V}\hat{\lambda}_j \Phi^{-1}(P) + o_p(n^{-1}) \end{aligned} \tag{27}$$

where, as usual, hats over various terms indicate that the terms in question are to be evaluated at the REML estimator $\hat{\theta}$. For theoretical calculations based on (27), we can ignore this distinction (i.e. assume $\kappa^{i,j}$, λ_i, etc., are evaluated at the true θ) since this will not affect the $O_p(n^{-1})$ properties of the procedure.

In the notation of (16), this amounts to replacing the quantity S (which we have called S_1, in the case of the plug-in predictor) by $S_1 - \phi(\Phi^{-1}(P))\Phi^{-1}(P) \cdot \frac{\kappa^{i,j}\lambda_i^T V \lambda_j}{2\sigma_0^2}$, where we have also used (46). Therefore by (17), the asymptotic coverage probability bias is increased by $n^{-1}\phi(\Phi^{-1}(P))\Phi^{-1}(P) \cdot \frac{\kappa^{i,j}\lambda_i^T V \lambda_j}{2\sigma_0^2}$. Thus (18) becomes

$$\begin{aligned} & nE\{\psi(z_P^*\ ;\ Y, \theta) - \psi(z_P\ ;\ Y, \theta)\} \\ \approx\ & \phi(\Phi^{-1}(P))\Phi^{-1}(P)\Big\{\kappa^{i,j}\kappa^{k,\ell}\frac{\sigma_{0\ell}}{\sigma_0}\left(\kappa_{ik,j} + \frac{1}{2}\kappa_{ijk}\right) \\ & + \frac{1}{2}\kappa^{i,j}\left(\frac{\sigma_{0ij}}{\sigma_0} - \frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\Phi^{-1}(P)^2\right) \\ & - \frac{1}{2n\sigma_0^2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V \frac{\partial W}{\partial \theta^i} V \frac{\partial W}{\partial \theta^k} V \lambda_\ell + \lambda_j^T V \frac{\partial W}{\partial \theta^k} V \frac{\partial W}{\partial \theta^i} V \lambda_\ell\right)\Big\}. \end{aligned} \tag{28}$$

In other words, the Harville-Jeske-Zimmerman-Cressie correction eliminates one of the bias terms in (18), but leaves the rest intact, including the one that is dominant as $P \to 1$.

Abt (1999) derived an improved version of (25) that does not assume $\hat{\theta}$ and Y independent. As shown in Section 8.6, a more refined approximation to (25) is

$$n^{-1}\kappa^{i,j}\lambda_i^T V \lambda_j + n^{-2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V \frac{\partial W}{\partial \theta^i} V \frac{\partial W}{\partial \theta^k} V \lambda_\ell + \lambda_j^T V \frac{\partial W}{\partial \theta^k} V \frac{\partial W}{\partial \theta^i} V \lambda_\ell\right), \tag{29}$$

a result that is presumably equivalent to Abt's though a precise correspondence between (29) and Abt's result has not been established.

Therefore, a more refined version of the Harville-Jeske-Zimmerman-Cressie correction would replace σ_0^2 by

$$\sigma_0^2 + n^{-1}\kappa^{i,j}\lambda_i^T V \lambda_j + n^{-2}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V \frac{\partial W}{\partial \theta^i} V \frac{\partial W}{\partial \theta^k} V \lambda_\ell + \lambda_j^T V \frac{\partial W}{\partial \theta^k} V \frac{\partial W}{\partial \theta^i} V \lambda_\ell\right)$$

in the construction of the prediction interval. By the same argument as led to (28), the coverage probability bias is now

$$nE\{\psi(z_P^* \,;\, Y,\theta) - \psi(z_P \,;\, Y,\theta)\} \;\approx\; \phi(\Phi^{-1}(P))\Phi^{-1}(P)\Big\{\kappa^{i,j}\kappa^{k,\ell}\frac{\sigma_{0\ell}}{\sigma_0}\left(\kappa_{ik,j}+\frac{1}{2}\kappa_{ijk}\right) + \frac{1}{2}\kappa^{i,j}\left(\frac{\sigma_{0ij}}{\sigma_0} - \frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\Phi^{-1}(P)^2\right). \tag{30}$$

This still does not eliminate the dominant (as $P \to 1$) term in (18).

Therefore we consider a different approach: instead of trying to refine (18) using previously defined modifications, we use (18) to suggest a new one. This is similar in spirit to the asymptotic frequentist approach to prediction problems of Barndorff-Nielsen and Cox (1996). Consider the estimator

$$\begin{aligned} z_P^\dagger \;=\; & \hat{z}_P - n^{-1}\Phi^{-1}(P)\Big\{\hat{\kappa}^{i,j}\hat{\kappa}^{k,\ell}\hat{\sigma}_{0\ell}\left(\hat{\kappa}_{ik,j}+\frac{1}{2}\hat{\kappa}_{ijk}\right) \\ & + \frac{1}{2}\hat{\kappa}^{i,j}\left(\hat{\sigma}_{0ij} - \frac{\hat{\sigma}_{0i}\hat{\sigma}_{0j}}{\hat{\sigma}_0}\Phi^{-1}(P)^2\right) - \frac{1}{2\hat{\sigma}_0}\hat{\kappa}^{i,j}\hat{\lambda}_i^T\hat{V}\hat{\lambda}_j \\ & - \frac{1}{2n\hat{\sigma}_0}\hat{\kappa}^{i,j}\hat{\kappa}^{k,\ell}\left(\hat{\lambda}_j^T\hat{V}\frac{\partial\hat{W}}{\partial\theta^i}\hat{V}\frac{\partial\hat{W}}{\partial\theta^k}\hat{V}\hat{\lambda}_\ell + \hat{\lambda}_j^T\hat{V}\frac{\partial\hat{W}}{\partial\theta^k}\hat{V}\frac{\partial\hat{W}}{\partial\theta^i}\hat{V}\hat{\lambda}_\ell\right)\Big\}. \end{aligned} \tag{31}$$

In (16), we now have $S = S_1 + \frac{\phi(\Phi^{-1}(P))\Phi^{-1}(P)}{\sigma_0}\Big\{\kappa^{i,j}\kappa^{k,\ell}\sigma_{0\ell}\left(\kappa_{ik,j}+\frac{1}{2}\kappa_{ijk}\right)$
$+\frac{1}{2}\kappa^{i,j}\left(\sigma_{0ij} - \frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)^2\right) - \frac{1}{2\sigma_0}\kappa^{i,j}\lambda_i^T V\lambda_j - \frac{1}{2n\sigma_0}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell + \lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right)\Big\}$.
After combining this with (18) we now find (by construction) that all the second-order coverage probability bias terms cancel and therefore (31) has coverage probability bias 0, to $O_p(n^{-1})$.

Thus we conclude that there are two ways to achieve an estimate of the predictive distribution with second-order coverage probability bias of 0. The first is the Bayesian predictive distribution based on the matching prior found by solving (20). The second approach is direct, by (31). As a practical matter, direct application of (31) seems to be the simpler approach to compute, and would seem to render it unnecessary to solve the differential equation (20) in practice. Nevertheless, the existence of a matching prior is important because it shows that the artificial-looking predictor (31) is equivalent to a Bayesian predictor, which provides further justification of our overall emphasis on Bayesian methods.

7 Relationship to design criteria of Zhu and Stein

Considering the discussion of the previous section, suppose $z_P^\dagger$ is an estimate of the P-quantile of the predictive distribution, for which the second-order coverage probability bias is 0. As just noted, such an estimate may be calculated directly from (31), or indirectly by solving (20) for the matching prior and computing a Bayesian predictive distribution. If the first approach is taken, then a combination of (21) with (31) leads to

$$\begin{aligned} nE\{z_P^\dagger - z_P\} \;\approx\; & \Phi^{-1}(P)\Big\{\frac{1}{2}\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)^2 + \frac{1}{2\sigma_0}\kappa^{i,j}\lambda_i^T V\lambda_j \\ & + \frac{1}{2n\sigma_0}\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell + \lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right)\Big\}. \end{aligned} \tag{32}$$

It may easily be checked that we get the same result from (22) if we use the matching prior defined by (20).

Equation (32) has an interesting interpretation. The second term corresponds to the Harville-Jeske-Zimmerman-Cressie correction to the prediction variance, and the third term is Abt's refinement. If we assume that Abt's refinement is negligible in comparison with the main term of the Harville-Jeske-Zimmerman-Cressie correction (Abt's own simulations supported this), then we may ignore the third term in (32).

To apply (32) to the length of a prediction interval, we must combine it with (23). Assume $P_1 = P,\ P_2 = 1 - P$ for an equal-tailed prediction interval. Then the length of the prediction interval is

$$2\sigma_0\Phi^{-1}(P)\left[1 + n^{-1}\left\{\frac{1}{2}\Phi^{-1}(P)^2\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2} + \frac{1}{2}\kappa^{i,j}\frac{\lambda_i^T V\lambda_j}{\sigma_0^2}\right\}\right] \tag{33}$$

so if we square (33), ignoring the multiplier $2\Phi^{-1}(P)$, we want to minimize

$$\sigma_0^2 + n^{-1}\kappa^{i,j}\lambda_i^T V\lambda_j + n^{-1}\Phi^{-1}(P)^2\kappa^{i,j}\sigma_{0i}\sigma_{0j}. \tag{34}$$

This has an interesting parallel with the approach of Zhu and Stein (2004). They defined quantities $V_1 = \sigma_0^2 + n^{-1}\kappa^{i,j}\lambda_i^T V\lambda_j$ and $V_2 = n^{-1}\frac{\partial\sigma_0^2}{\partial\theta^i}\kappa^{i,j}\frac{\partial\sigma_0^2}{\partial\theta^j} = 4\sigma_0^2 n^{-1}\kappa^{i,j}\sigma_{0i}\sigma_{0j}$ in the present notation. They interpret V_1 as the Harville-Jeske-Zimmerman-Cressie approximation to the prediction error variance and V_2 as a term reflecting the uncertainty of estimating σ^2. In practice, V_2 is replaced by $\frac{V_2}{\sigma_0^2}$ to achieve scale invariance. As a combined criterion, Zhu and Stein suggested $V_3 = V_1 + \frac{V_2}{2\sigma_0^2}$ but also suggested that other linear combinations of V_1 and $\frac{V_2}{\sigma_0^2}$ could be considered.

Equation (34) is equivalent to

$$V_3 = V_1 + \frac{\Phi^{-1}(P)^2}{4}\cdot\frac{V_2}{\sigma_0^2} \tag{35}$$

which, surprisingly, is of the same structure as the Zhu-Stein criterion, and could even coincide with it (if we chose P such that $\Phi^{-1}(P) = \sqrt{2}$). Equation (35) carries the implication that the optimal design might depend on the coverage probability of the prediction interval. This, however, may not be unreasonable: the closer P gets to 1, the more critical the "estimative" properties of the design become, and this affects the weightings in (35).

In summary: if we apply a second-order correction to the coverage probability bias, either directly through (31) or indirectly with a matching prior, then (35) becomes a suitable design criterion, paralleling that of Zhu and Stein (2004).

8 Detailed derivations

8.1 Proof of Lemma 1

Following the derivation of Harville (1974), the restricted likelihood may be defined as the density of $U = A^T Y$, where A is an $n\times(n-q)$ matrix of rank $n-q$ such that $A^T X = 0,\ A^T A = I,\ AA^T = I - X(X^T X)^{-1}X^T$. Also let $\hat\eta = (X^T V^{-1}X)^{-1}X^T V^{-1}Y$ be the generalized least squares estimator of η. As shown by Harville, the transformation $Y \to (U, \hat\eta)$ has Jacobian $|X^T X|^{-1/2}$, so

$$f(Y) = |X^T X|^{-1/2}e^{\ell_n(U)}\cdot(2\pi)^{-q/2}|X^T V^{-1}X|^{1/2}\exp\left\{-\frac{1}{2}(\hat\eta-\eta)^T X^T V^{-1}X(\hat\eta-\eta)\right\}.$$

Equation (5) follows at once (Harville's result). For (6), consider the variable $T = Y_0 - \lambda^T Y$. The conditional distribution of T given Y is normal with mean $x_0^T\eta + w^T V^{-1}(Y - X\eta) - \lambda^T Y = -(x_0^T - w^T V^{-1}X)(\hat{\eta} - \eta)$ and variance $v - w^T V^{-1} w$. Therefore, the joint density of Y and Y_0 is

$$|X^T X|^{-1/2} f(U, \hat{\eta}, T) = |X^T X|^{-1/2} e^{\ell_n(U)} \cdot (2\pi)^{-q/2} |X^T V^{-1} X|^{1/2} \exp\left\{-\frac{1}{2}(\hat{\eta} - \eta)^T X^T V^{-1} X (\hat{\eta} - \eta)\right\} \cdot$$
$$(2\pi)^{-1/2}(v - w^T V^{-1} w)^{-1/2} \exp\left[-\frac{1}{2(v - w^T V^{-1} w)}\{T + (x_0^T - w^T V^{-1} X)(\hat{\eta} - \eta)\}^2\right].$$

The result (6) follows by standard manipulations.

8.2 Derivatives of $\ell_n(\theta)$

This section concerns the evaluation of the quantities $U_i,\ U_{ij},\ U_{ijk}$, that are needed in (11) and subsequently.

From (2) we have, ignoring constants not depending on θ,

$$\begin{aligned} \ell_n &= -\frac{1}{2}\log|V| - \frac{1}{2}\log|X^T V^{-1} X| - \frac{1}{2} Y^T W Y \\ &= -\frac{1}{2}\log|V| - \frac{1}{2}\log|X^T V^{-1} X| - \frac{1}{2} e_\alpha e_\beta w^{\alpha\beta} \end{aligned}$$

where the matrix $W = V^{-1} - V^{-1}X(X^T V^{-1} X^{-1}) X^T V^{-1}$ has entries $\{w^{\alpha\beta}\}$ and we exploit the fact that $Y^T W Y = e^T W e$ where $e = Y - X\eta$ has entries $e_\alpha,\ 1 \le \alpha \le n$. Note that we use greek letters to denote individual components of the observation vector to avoid confusion with letters $i, j, k, ...$ that are used for components of θ, but the summation convention applies the same way.

Exploiting identities such as $\frac{\partial \log|V|}{\partial \theta^i} = \mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \theta^i}\right)$ (Mardia, Kent and Bibby (1979), Sections A.2.3 and A.9) and with some algebraic manipulation, we deduce

$$U_i = \frac{1}{2} v_{\alpha\beta}\frac{\partial w^{\alpha\beta}}{\partial \theta^i} - \frac{1}{2} e_\alpha e_\beta \frac{\partial w^{\alpha\beta}}{\partial \theta^i}, \tag{36}$$

Further differentation leads to

$$U_{ij} = \frac{1}{2}\frac{\partial v_{\alpha\beta}}{\partial \theta^j}\frac{\partial w^{\alpha\beta}}{\partial \theta^i} + \frac{1}{2} v_{\alpha\beta}\frac{\partial^2 w^{\alpha\beta}}{\partial \theta^i \partial \theta^j} - \frac{1}{2} e_\alpha e_\beta \frac{\partial^2 w^{\alpha\beta}}{\partial \theta^i \partial \theta^j}, \tag{37}$$

$$\begin{aligned} U_{ijk} = {} & \frac{1}{2}\frac{\partial^2 v_{\alpha\beta}}{\partial \theta^j \partial \theta^k}\frac{\partial w^{\alpha\beta}}{\partial \theta^i} + \frac{1}{2}\frac{\partial v_{\alpha\beta}}{\partial \theta^j}\frac{\partial^2 w^{\alpha\beta}}{\partial \theta^i \partial \theta^k} + \frac{1}{2}\frac{\partial v_{\alpha\beta}}{\partial \theta^k}\frac{\partial^2 w^{\alpha\beta}}{\partial \theta^i \partial \theta^j} \\ & + \frac{1}{2} v_{\alpha\beta}\frac{\partial^3 w^{\alpha\beta}}{\partial \theta^i \partial \theta^j \partial \theta^k} - \frac{1}{2} e_\alpha e_\beta \frac{\partial^3 w^{\alpha\beta}}{\partial \theta^i \partial \theta^j \partial \theta^k}. \end{aligned} \tag{38}$$

We also note the identities

$$\frac{\partial W}{\partial \theta^i} = -W\frac{\partial V}{\partial \theta^i} W, \tag{39}$$

$$\frac{\partial^2 W}{\partial \theta^i \partial \theta^j} = W\frac{\partial V}{\partial \theta^i} W \frac{\partial V}{\partial \theta^j} W + W\frac{\partial V}{\partial \theta^j} W \frac{\partial V}{\partial \theta^i} W - W\frac{\partial^2 V}{\partial \theta^i \theta^j} W, \tag{40}$$

$$\frac{\partial^3 W}{\partial \theta^i \partial \theta^j \partial \theta^k} = -W\frac{\partial V}{\partial \theta^i} W \frac{\partial V}{\partial \theta^j} W \frac{\partial V}{\partial \theta^k} W - W\frac{\partial V}{\partial \theta^i} W \frac{\partial V}{\partial \theta^k} W \frac{\partial V}{\partial \theta^j} W - W\frac{\partial V}{\partial \theta^j} W \frac{\partial V}{\partial \theta^i} W \frac{\partial V}{\partial \theta^k} W$$

$$
\begin{aligned}
&-W\frac{\partial V}{\partial\theta^j}W\frac{\partial V}{\partial\theta^k}W\frac{\partial V}{\partial\theta^i}W - W\frac{\partial V}{\partial\theta^k}W\frac{\partial V}{\partial\theta^i}W\frac{\partial V}{\partial\theta^j}W - W\frac{\partial V}{\partial\theta^k}W\frac{\partial V}{\partial\theta^j}W\frac{\partial V}{\partial\theta^i}W\\
&+W\frac{\partial^2 V}{\partial\theta^i\partial\theta^j}W\frac{\partial V}{\partial\theta^k}W + W\frac{\partial^2 V}{\partial\theta^i\partial\theta^k}W\frac{\partial V}{\partial\theta^j}W + W\frac{\partial^2 V}{\partial\theta^j\partial\theta^k}W\frac{\partial V}{\partial\theta^i}W\\
&+W\frac{\partial V}{\partial\theta^i}W\frac{\partial^2 V}{\partial\theta^j\partial\theta^k}W + W\frac{\partial V}{\partial\theta^j}W\frac{\partial^2 V}{\partial\theta^i\partial\theta^k}W + W\frac{\partial V}{\partial\theta^k}W\frac{\partial^2 V}{\partial\theta^i\partial\theta^j}W\\
&-W\frac{\partial^3 V}{\partial\theta^i\partial\theta^j\partial\theta^k}W \qquad (41)
\end{aligned}
$$

which greatly aid the computation of (36)–(38), given that it is usually straightforward to differentiate the components of V analytically but much harder to do so for W.

In this section, we also indicate how to calculate the quantities κ_{ij}, κ_{ijk}, $\kappa_{ij,k}$ introduced in Section 4. In U_{ij} and U_{ijk}, the last two terms are respectively $n^{1/2}Z_{ij}, n^{1/2}Z_{ijk}$ and the remainder are respectively $n\kappa_{ij}, n\kappa_{ijk}$. Thus κ_{ij} and κ_{ijk} are calculated directly from (37) and (38), using also (39)–(41). For $\kappa_{ij,k}$, we have

$$
\begin{aligned}
n\kappa_{ij,k} &= \frac{1}{4}E\left\{\left(v_{\alpha\beta}\frac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j} - e_\alpha e_\beta\frac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j}\right)\left(v_{\gamma\delta}\frac{\partial w^{\gamma\delta}}{\partial\theta^k} - e_\gamma e_\delta\frac{\partial w^{\gamma\delta}}{\partial\theta^k}\right)\right\}\\
&= \frac{1}{4}(v_{\alpha\gamma}v_{\beta\delta}+v_{\alpha\delta}v_{\beta\gamma})\frac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j}\frac{\partial w^{\gamma\delta}}{\partial\theta^k}\\
&= \frac{1}{2}v_{\alpha\gamma}v_{\beta\delta}\frac{\partial^2 w^{\alpha\beta}}{\partial\theta^i\partial\theta^j}\frac{\partial w^{\gamma\delta}}{\partial\theta^k} \qquad (42)
\end{aligned}
$$

where in the middle of the calculation we used

$$
E\{e_\alpha e_\beta e_\gamma e_\delta\} = v_{\alpha\beta}v_{\gamma\delta}+v_{\alpha\gamma}v_{\beta\delta}+v_{\alpha\delta}v_{\beta\gamma}. \qquad (43)
$$

8.3 Asymptotics for the inverse predictive distribution function

Since $\psi^*(z_P^*\ ;\ Y)$ and $\psi(z_P\ ;\ Y,\theta)$ both equal P, we have

$$
\begin{aligned}
0 &= \psi^*(z_P^*\ ;\ Y) - \psi(z_P\ ;\ Y,\theta)\\
&= \psi^*(z_P^*\ ;\ Y) - \psi^*(z_P\ ;\ Y) + \psi^*(z_P\ ;\ Y) - \psi(z_P\ ;\ Y,\theta)\\
&= (z_P^*-z_P)\psi^{*\prime} + \frac{1}{2}(z_P^*-z_P)^2\psi^{*\prime\prime} + n^{-1/2}R + n^{-1}S + o_p(n^{-1}) \qquad (44)\\
&= (z_P^*-z_P)\psi' + n^{-1/2}(z_P^*-z_P)R' + \frac{1}{2}(z_P^*-z_P)^2\psi''\\
&\quad + n^{-1/2}R + n^{-1}S + o_p(n^{-1}). \qquad (45)
\end{aligned}
$$

Here, (44) is a combination of (15) with Taylor expansion of ψ^* (we assume $z_P^*-z_P$ is of $O_p(n^{-1/2})$ so that terms of order $(z_P^*-z_P)^3$ and higher may be neglected); while (45) follows on assuming that (15) may be differentiated term by term with respect to z.

We solve (45) in two stages: setting equal to 0 the $O_p(n^{-1/2})$ terms $(z_P^*-z_P)\psi' + n^{-1/2}R$, we deduce the first-order approximation $z_P^*-z_P = -n^{-1/2}R/\psi' + o_p(n^{-1/2})$. Next, assuming $z_P^*-z_P = -n^{-1/2}R/\psi' + n^{-1}\epsilon$ and substituting back in (45) to solve for ϵ, we deduce (16).

To deduce (17) from (16), we take a two-term Taylor expansion

$$
\psi(z_P^*\ ;\ Y,\theta) - \psi(z_P\ ;\ Y,\theta) = (z_P^*-z_P)\psi' + \frac{1}{2}(z_P^*-z_P)^2\psi'' + o_p(n^{-1})
$$

substituting the expansion of $z_P^* - z_P$ from (16) and neglecting all terms smaller than $O_p(n^{-1})$.

8.4 Proof of (18) and (19)

First we give explicit formulae for the derivatives of ψ. Recall that $\psi(z\ ;\ Y, \theta) = \Phi\left(\frac{z-\lambda^T Y}{\sigma_0}\right)$. we write $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ for the standard normal density and note $\phi'(z) = -z\phi(z)$. Calculating various derivatives of ψ and substituting $z = z_P = \lambda^T Y + \sigma_0 \Phi^{-1}(P)$, we have

$$\begin{aligned}
\psi' &= \frac{1}{\sigma_0}\phi(\Phi^{-1}(P)), && (46)\\
\psi'' &= -\frac{1}{\sigma_0^2}\Phi^{-1}(P)\phi(\Phi^{-1}(P)), && (47)\\
\psi_i &= -\left\{\frac{\lambda_i^T Y}{\sigma_0} + \frac{\sigma_{0i}}{\sigma_0}\Phi^{-1}(P)\right\}\phi(\Phi^{-1}(P)), && (48)\\
\psi_{ij} &= -\left\{\frac{\lambda_{ij}^T Y}{\sigma_0} - \frac{\sigma_{0j}\lambda_i^T Y}{\sigma_0^2} - \frac{\sigma_{0i}\lambda_j^T Y}{\sigma_0^2} + \frac{\sigma_{0ij}}{\sigma_0}\Phi^{-1}(P) - \frac{2\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\Phi^{-1}(P)\right\}\phi(\Phi^{-1}(P))\\
&\quad -\left\{\frac{\lambda_i^T Y}{\sigma_0} + \frac{\sigma_{0i}}{\sigma_0}\Phi^{-1}(P)\right\}\left\{\frac{\lambda_j^T Y}{\sigma_0} + \frac{\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)\right\}\Phi^{-1}(P)\phi(\Phi^{-1}(P)), && (49)\\
\psi_i' &= \left[-\frac{\sigma_{0i}}{\sigma_0^2} + \left\{\frac{\lambda_i^T Y}{\sigma_0} + \frac{\sigma_{0i}}{\sigma_0}\Phi^{-1}(P)\right\}\cdot\frac{1}{\sigma_0}\Phi^{-1}(P)\right]\phi(\Phi^{-1}(P)). && (50)
\end{aligned}$$

To evaluate the expectation of (17) we need the expectations of R, $\frac{RR'}{\psi'}$ and S, where $R = \kappa^{i,j} Z_i \psi_j$, $S = S_1 = \kappa^{i,j}\kappa^{k,\ell} Z_{ik} Z_j \psi_\ell + \frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk} Z_r Z_s \psi_t + \frac{1}{2}\kappa^{i,j}\kappa^{k,\ell} Z_i Z_k \psi_{j\ell}$.

First we consider $E\{R\}$. Noting that $n^{1/2} Z_i = U_i$ is given by (36) and ψ_i is given by (48), we have $E\{U_i\} = E\{Y_\alpha U_i\} = 0$ for any α, hence $E\{R\} = 0$.

Next, consider

$$\begin{aligned}
E\left\{\frac{RR'}{\psi'}\right\} &= E\left\{\frac{\kappa^{i,j} Z_i \psi_j \kappa^{k,\ell} Z_k \psi_\ell'}{\psi'}\right\}\\
&= \kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\cdot\\
&\quad \cdot E\left[Z_i Z_k\left\{\frac{\lambda_j^T Y}{\sigma_0} + \frac{\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)\right\}\left\{\frac{\sigma_{0\ell}}{\sigma_0} - \left(\frac{\lambda_\ell^T Y}{\sigma_0} + \frac{\sigma_{0\ell}}{\sigma_0}\Phi^{-1}(P)\right)\Phi^{-1}(P)\right\}\right] \quad (51)
\end{aligned}$$

combining (46), (48) and (50).

To evaluate (51), we need expressions for $E\{Z_i Z_k Y_\epsilon\}$ and $E\{Z_i Z_k Y_\epsilon Y_\zeta\}$ where Y_ϵ and Y_ζ are arbitrary components of Y. By (36),

$$\begin{aligned}
E\{Z_i Z_k Y_\epsilon\} &= \frac{1}{n}E\{U_i U_k Y_\epsilon\}\\
&= \frac{1}{4n}E\left\{\left(v_{\alpha\beta}\frac{\partial w^{\alpha\beta}}{\partial\theta^i} - e_\alpha e_\beta\frac{\partial w^{\alpha\beta}}{\partial\theta^i}\right)\left(v_{\gamma\delta}\frac{\partial w^{\gamma\delta}}{\partial\theta^k} - e_\gamma e_\delta\frac{\partial w^{\gamma\delta}}{\partial\theta^k}\right)Y_\epsilon\right\}\\
&= \kappa_{i,k}(X\eta)_\epsilon \quad (52)
\end{aligned}$$

where we write $Y_\epsilon = (X\eta)_\epsilon + e_\epsilon$ and exploit the fact that all terms including e_ϵ translate to odd-order moments of zero-mean Gaussian variables and are therefore 0.

Similarly, we have

$$\begin{aligned}
E\{Z_iZ_kY_\epsilon Y_\zeta\} &= \frac{1}{n}E\{U_iU_kY_\epsilon Y_\zeta\} \\
&= \frac{1}{4n}E\left\{\left(v_{\alpha\beta}\frac{\partial w^{\alpha\beta}}{\partial\theta^i} - e_\alpha e_\beta\frac{\partial w^{\alpha\beta}}{\partial\theta^i}\right)\left(v_{\gamma\delta}\frac{\partial w^{\gamma\delta}}{\partial\theta^k} - e_\gamma e_\delta\frac{\partial w^{\gamma\delta}}{\partial\theta^k}\right)Y_\epsilon Y_\zeta\right\} \\
&= \kappa_{i,k}(X\eta)_\epsilon(X\eta)_\zeta \\
&\quad + \frac{1}{4n}E\left\{\left(v_{\alpha\beta}\frac{\partial w^{\alpha\beta}}{\partial\theta^i} - e_\alpha e_\beta\frac{\partial w^{\alpha\beta}}{\partial\theta^i}\right)\left(v_{\gamma\delta}\frac{\partial w^{\gamma\delta}}{\partial\theta^k} - e_\gamma e_\delta\frac{\partial w^{\gamma\delta}}{\partial\theta^k}\right)e_\epsilon e_\zeta\right\}. \quad (53)
\end{aligned}$$

For the second term in (53), we need moments of up to sixth order, recalling (43) and

$$\begin{aligned}
E\{e_\alpha e_\beta e_\gamma e_\delta e_\epsilon e_\zeta\} &= v_{\alpha\beta}v_{\gamma\delta}v_{\epsilon\zeta} + v_{\alpha\beta}v_{\gamma\epsilon}v_{\delta\zeta} + v_{\alpha\beta}v_{\gamma\zeta}v_{\delta\epsilon} + v_{\alpha\gamma}v_{\beta\delta}v_{\epsilon\zeta} + v_{\alpha\gamma}v_{\beta\epsilon}v_{\delta\zeta} + v_{\alpha\gamma}v_{\beta\zeta}v_{\delta\epsilon} \\
&\quad + v_{\alpha\delta}v_{\beta\gamma}v_{\epsilon\zeta} + v_{\alpha\delta}v_{\beta\epsilon}v_{\gamma\zeta} + v_{\alpha\delta}v_{\beta\zeta}v_{\gamma\epsilon} + v_{\alpha\epsilon}v_{\beta\gamma}v_{\delta\zeta} + v_{\alpha\epsilon}v_{\beta\delta}v_{\gamma\zeta} + v_{\alpha\epsilon}v_{\beta\zeta}v_{\gamma\delta} \\
&\quad + v_{\alpha\zeta}v_{\beta\gamma}v_{\delta\epsilon} + v_{\alpha\zeta}v_{\beta\delta}v_{\gamma\epsilon} + v_{\alpha\zeta}v_{\beta\epsilon}v_{\gamma\delta}. \quad (54)
\end{aligned}$$

Combining (43) and (54), we have

$$\begin{aligned}
E\{(v_{\alpha\beta} - e_\alpha e_\beta)(v_{\gamma\delta} - e_\gamma e_\delta)e_\epsilon e_\zeta\} &= v_{\alpha\gamma}v_{\beta\delta}v_{\epsilon\zeta} + v_{\alpha\gamma}v_{\beta\epsilon}v_{\delta\zeta} + v_{\alpha\gamma}v_{\beta\zeta}v_{\delta\epsilon} + v_{\alpha\delta}v_{\beta\gamma}v_{\epsilon\zeta} + v_{\alpha\delta}v_{\beta\epsilon}v_{\gamma\zeta} \\
&\quad + v_{\alpha\delta}v_{\beta\zeta}v_{\gamma\epsilon} + v_{\alpha\epsilon}v_{\beta\gamma}v_{\delta\zeta} + v_{\alpha\epsilon}v_{\beta\delta}v_{\gamma\zeta} + v_{\alpha\zeta}v_{\beta\gamma}v_{\delta\epsilon} + v_{\alpha\zeta}v_{\beta\delta}v_{\gamma\epsilon}.
\end{aligned}$$

Hence (53) reduces to

$$\begin{aligned}
E\{Z_iZ_kY_\epsilon Y_\zeta\} &= \kappa_{i,k}(X\eta)_\epsilon(X\eta)_\zeta + \frac{1}{4n}\frac{\partial w^{\alpha\beta}}{\partial\theta^i}\frac{\partial w^{\gamma\delta}}{\partial\theta^k}(v_{\alpha\gamma}v_{\beta\delta}v_{\epsilon\zeta} + v_{\alpha\gamma}v_{\beta\epsilon}v_{\delta\zeta} + v_{\alpha\gamma}v_{\beta\zeta}v_{\delta\epsilon} \\
&\quad + v_{\alpha\delta}v_{\beta\gamma}v_{\epsilon\zeta} + v_{\alpha\delta}v_{\beta\epsilon}v_{\gamma\zeta} + v_{\alpha\delta}v_{\beta\zeta}v_{\gamma\epsilon} + v_{\alpha\epsilon}v_{\beta\gamma}v_{\delta\zeta} + v_{\alpha\epsilon}v_{\beta\delta}v_{\gamma\zeta} + v_{\alpha\zeta}v_{\beta\gamma}v_{\delta\epsilon} + v_{\alpha\zeta}v_{\beta\delta}v_{\gamma\epsilon}) \\
&= \kappa_{i,k}(X\eta)_\epsilon(X\eta)_\zeta + v_{\epsilon\zeta}\kappa_{i,k} + \frac{1}{n}\left\{V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\right\}_{\epsilon\zeta} + \frac{1}{n}\left\{V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\right\}_{\epsilon\zeta} \quad (55)
\end{aligned}$$

where $\left\{\ \cdot\ \right\}_{\epsilon\zeta}$ denotes the (ϵ,ζ) entry of the matrix enclosed in brackets, and we have used the fact that $\frac{1}{2n}\frac{\partial w^{\alpha\beta}}{\partial\theta^i}v_{\alpha\gamma}\frac{\partial w^{\gamma\delta}}{\partial\theta^k}v_{\beta\delta} = \frac{1}{2n}\frac{\partial w^{\alpha\beta}}{\partial\theta^i}v_{\alpha\delta}\frac{\partial w^{\gamma\delta}}{\partial\theta^k}v_{\beta\gamma}$ is one of several equivalent expressions for $\kappa_{i,k}$.

We use (52) and (55) to simplify (51). We write $\lambda_j^T Y = \lambda_j^\epsilon Y_\epsilon$ where λ_j^ϵ is the ϵth component of λ_j. We also note that because $\lambda^T X = x_0^T$ (independent of θ), we also have $\lambda_j^T X = 0$. Because of this, the terms involving $(X\eta)_\epsilon$ or $(X\eta)_\zeta$ in (52) and (55) become 0 when substituted into (51). We write

$$\begin{aligned}
E\left\{\frac{RR'}{\psi'}\right\} &= \kappa^{j,\ell}\frac{\sigma_{0j}}{\sigma_0}\frac{\sigma_{0\ell}}{\sigma_0}\phi(\Phi^{-1}(P))\Phi^{-1}(P)(1-\Phi^{-1}(P)^2) \\
&\quad -\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{\lambda_j^\epsilon\lambda_\ell^\zeta}{\sigma_0^2}\left[v_{\epsilon\zeta}\kappa_{i,k} + \frac{1}{n}\left\{V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\right\}_{\epsilon\zeta} + \frac{1}{n}\left\{V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\right\}_{\epsilon\zeta}\right] \\
&= \kappa^{j,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\left\{\frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}(1-\Phi^{-1}(P)^2) - \frac{1}{\sigma_0^2}\lambda_j^T V\lambda_\ell\right\} \\
&\quad -\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{1}{n\sigma_0^2}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell + \lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right). \quad (56)
\end{aligned}$$

Continuing the calculation of (17), we must next evaluate $E\{S_1\}$, which is the sum of three terms:

$$E\left\{\kappa^{i,j}\kappa^{k,\ell}Z_{ik}Z_j\psi_\ell\right\}, \tag{57}$$

$$\frac{1}{2}E\left\{\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}Z_rZ_s\psi_t\right\}, \tag{58}$$

$$\frac{1}{2}E\left\{\kappa^{i,j}\kappa^{k,\ell}Z_iZ_k\psi_{j\ell}\right\}. \tag{59}$$

First consider (57). By (36), (37) and (48), this reduces to the expectation of

$$-\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))Z_{ik}Z_j\left\{\frac{\lambda_\ell^TY}{\sigma_0}+\frac{\sigma_{0\ell}}{\sigma_0}\Phi^{-1}(P)\right\}.$$

In this expression, all the terms involving Y have expectation 0, for essentially the same reasons as in (51), so (57) reduces to

$$-\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{\sigma_{0\ell}}{\sigma_0}\kappa_{ik,j}. \tag{60}$$

Next consider (58). By (36) and (48), this is the expectation of

$$-\frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}\phi(\Phi^{-1}(P))Z_rZ_s\left\{\frac{\lambda_t^TY}{\sigma_0}+\frac{\sigma_{0t}}{\sigma_0}\Phi^{-1}(P)\right\}.$$

Once again the terms involving Y have expectation 0, and the rest reduce to

$$\begin{aligned}&-\frac{1}{2}\kappa^{i,r}\kappa^{j,s}\kappa^{k,t}\kappa_{ijk}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{\sigma_{0t}}{\sigma_0}\kappa_{r,s}\\ =\;&-\frac{1}{2}\kappa^{i,j}\kappa^{k,\ell}\kappa_{ijk}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{\sigma_{0\ell}}{\sigma_0}.\end{aligned} \tag{61}$$

Finally we evaluate (59). By (36), (49), this is the expectation of

$$\begin{aligned}&-\frac{1}{2}\kappa^{i,j}\kappa^{k,\ell}Z_iZ_k\cdot\\ &\cdot\Bigg[\left\{\frac{\lambda_{j\ell}^TY}{\sigma_0}-\frac{\sigma_{0\ell}\lambda_j^TY}{\sigma_0^2}-\frac{\sigma_{0j}\lambda_\ell^TY}{\sigma_0^2}+\frac{\sigma_{0j\ell}}{\sigma_0}\Phi^{-1}(P)-\frac{2\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}\Phi^{-1}(P)\right\}\phi(\Phi^{-1}(P))\\ &+\left\{\frac{\lambda_j^TY}{\sigma_0}+\frac{\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)\right\}\left\{\frac{\lambda_\ell^TY}{\sigma_0}+\frac{\sigma_{0\ell}}{\sigma_0}\Phi^{-1}(P)\right\}\Phi^{-1}(P)\phi(\Phi^{-1}(P))\Bigg].\end{aligned} \tag{62}$$

Once again the linear terms in Y have expectation 0. Those terms in (62) that do not depend on Y have expectation

$$-\frac{1}{2}\kappa^{j,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\left\{\frac{\sigma_{0j\ell}}{\sigma_0}-2\frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}+\frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0^2}\Phi^{-1}(P)^2\right\}. \tag{63}$$

The quadratic terms in Y within (62) reduce to

$$-\frac{1}{2}\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\frac{\lambda_j^\epsilon\lambda_\ell^\zeta}{\sigma_0^2}Z_iZ_kY_\epsilon Y_\zeta$$

and after taking expectations using (55), that becomes

$$-\frac{1}{2\sigma_0^2}\kappa^{j,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\lambda_j^T V\lambda_\ell$$
$$-\frac{1}{2n\sigma_0^2}\kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P))\Phi^{-1}(P)\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell+\lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right). \quad (64)$$

Thus, (59) is the sum of (63) and (64).

Based on (17), the approximation to $nE\{\psi(\hat{z}_P\ ;\ Y,\theta)-\psi(z_P\ ;\ Y,\theta)\}$ is (56) minus the sum of (60), (61), (63) and (64). After collecting terms together, we get the result (18).

To derive (19), we start with (18) and add

$$E\{S_1-S_2\} \ = \ E\left\{-\frac{1}{2}\kappa_{ijk}\kappa^{i,j}\kappa^{k,\ell}\psi_\ell-\frac{1}{2}\psi_{ij}\kappa^{i,j}-\psi_iQ_j\kappa^{i,j}\right\}. \quad (65)$$

We need the expectations of ψ_i and ψ_{ij}. However from (48), (49),

$$\begin{aligned}E\{\psi_i\} &= -\frac{\sigma_{0i}}{\sigma_0}\phi(\Phi^{-1}(P))\Phi^{-1}(P),\\ E\{\psi_{ij}\} &= \left\{-\frac{\sigma_{0ij}}{\sigma_0}+\frac{2\sigma_{0i}\sigma_{0j}}{\sigma_0^2}-\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\Phi^{-1}(P)^2-\frac{1}{\sigma_0^2}\lambda_i^TV\lambda_j\right\}\phi(\Phi^{-1}(P))\Phi^{-1}(P).\end{aligned}$$

Hence we evaluate (65). Adding the result to (18), we deduce (19).

8.5 Derivation of (21) and (22)

From (16) and the fact that $E\{R\}=0$, we have

$$nE\{z_P^*-z_P\}\approx E\left\{\frac{RR'}{\psi'^2}-\frac{S}{\psi'}-\frac{1}{2}\frac{R^2\psi''}{\psi'^3}\right\}. \quad (66)$$

The first two terms are the same as in (18) or (19), multiplied by $\frac{1}{\psi'}=\frac{\sigma_0}{\phi(\Phi^{-1}(P))}$. The third term in (66) is

$$\frac{1}{2}\frac{\sigma_0\Phi^{-1}(P)}{\phi(\Phi^{-1}(P))^2}E\{R^2\} \ = \ \frac{1}{2}\sigma_0\Phi^{-1}(P)\kappa^{i,j}\kappa^{k,\ell}E\left\{Z_iZ_k\left(\frac{\lambda_j^TY}{\sigma_0}+\frac{\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)\right)\left(\frac{\lambda_\ell^TY}{\sigma_0}+\frac{\sigma_{0\ell}}{\sigma_0}\Phi^{-1}(P)\right)\right\}.$$

As in several calculations in Section 8.4, all linear terms in Y have expectation 0. The terms that do not depend on Y reduce to

$$\frac{1}{2}\Phi^{-1}(P)^3\kappa^{i,j}\kappa^{k,\ell}\kappa_{i,k}\frac{\sigma_{0j}\sigma_{0\ell}}{\sigma_0}=\frac{1}{2}\Phi^{-1}(P)^3\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}$$

which the quadratic terms in Y reduce to

$$\frac{1}{2}\Phi^{-1}(P)\kappa^{i,j}\kappa^{k,\ell}\lambda_j^\epsilon\lambda_\ell^\zeta E\{Z_iZ_kY_\epsilon Y_\zeta\}.$$

Using (55), this becomes

$$\frac{1}{2}\Phi^{-1}(P)\kappa^{i,j}\frac{\lambda_i^T V\lambda_j}{\sigma_0}+\frac{1}{2n\sigma_0}\Phi^{-1}(P)\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell+\lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right).$$

Hence

$$\begin{aligned}E\left\{-\frac{1}{2}\frac{R^2\psi''}{\psi'^3}\right\} &= \frac{1}{2}\Phi^{-1}(P)^3\kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0}+\frac{1}{2}\Phi^{-1}(P)\kappa^{i,j}\frac{\lambda_i^T V\lambda_j}{\sigma_0}\\ &\quad+\frac{1}{2n\sigma_0}\Phi^{-1}(P)\kappa^{i,j}\kappa^{k,\ell}\left(\lambda_j^T V\frac{\partial W}{\partial\theta^i}V\frac{\partial W}{\partial\theta^k}V\lambda_\ell+\lambda_j^T V\frac{\partial W}{\partial\theta^k}V\frac{\partial W}{\partial\theta^i}V\lambda_\ell\right). \quad (67)\end{aligned}$$

Finally, combining (67) with (66) and the previous calculations of (18) and (19), we deduce (21) and (22).

8.6 Derivation of (29)

Using $\hat{\theta}^i-\theta^i\approx n^{-1/2}\kappa^{i,k}Z_k$, write (25) as $n^{-1}\kappa^{i,k}\kappa^{j,\ell}\lambda_i^\epsilon\lambda_j^\zeta E\{Z_kZ_\ell Y_\epsilon Y_\zeta\}$. Then apply (55). After some rearrangement of terms, equation (29) follows.

9 Summary and Conclusions

In this paper, we have considered the properties of predictive inference in spatial statistics, or more generally Gaussian processes whose mean is a combination of linear regressors and whose covariance is parametrically specified. We use REML estimators for the covariance parameters and linear predictors (universal kriging) to derive a predictive distribution for an unobserved variable of interest. Direct application of the REML estimator leads to the "plug-in" approach. As an alternative to that, we propose a Bayesian predictive distribution with arbitrary smooth prior density. We also compare with previous frequentist corrections due to Harville and Jeske, and Zimmerman and Cressie, and their refinement suggested by Abt. We compute the coverage probability bias associated with the P-quantile of the predictive distribition, and also the expected length of a prediction interval, suggesting that the latter may be used as a criterion for network design.

The key results of the paper are (18) and (19) for the coverage probability bias of the plug-in and Bayesian predictive distributions, and the results (21) and (22) which (together with (23)) define the expected length of a prediction interval. From the point of view of coverage probability bias, although there does not appear to be any universal comparison that Bayesian predictors are better than the plug-in approach, it does appear that in the tails of the predictive distribution, the Bayesian approach is superior, regardless of the prior. However, our results also suggest the existence of a matching prior, solving (20), for which the second-order coverage probability bias is 0. Previous frequentist corrections do not have this property, but we use our second-order results to suggest a new frequentist correction, equation (31), which does.

If we use an estimator of the predictive distribution for which the second-order coverage probability bias is 0, then the expected length of the prediction interval takes a particularly clear-cut form. When interpreted as a criterion for network design, we derive a combined "estimative" and "predictive" criterion parallelling recent work of Zhu and Stein. However the relative weights of the two components depend on P, i.e. on the coverage coefficient of the prediction interval. This appears to be a novel feature of the present approach.

The results are restricted to linear regressions and the REML estimator. Initial calculations suggest that because of the loss of orthogonality properties that the present paper has exploited, the results would be considerably more complicated in the case of nonlinear regressions or the ordinary maximum likelihood estimator. We have also not considered numerical consequences of our results, either for Bayesian inference or for network design, nor have we examined algorithms for finding network designs under the new criteria. All these remain promising possibilities for future research.

10 References

Abt, M. (1999), Estimating the prediction mean squared error in Gaussian stochastic processes with exponential correlation structure. *Scandinavian Journal of Statistics* **26**, 563–578.

Barndorff-Nielsen, O.E. and Cox, D.R. (1996), Prediction and asymptotics. *Bernoulli* **2**, 319–340.

Berger, J.O., De Oliveira, V. and Sansó, B. (2001), Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* **96**, 1361–1374.

Bleistein, N. and Handelsman, R.A. (1986), *Asymptotic Expansion of Integrals.* Second edition, Dover, New York.

Brown, P.J., Le, N.D. and Zidek, J.V. (1994a), Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics* **22**, 489–509.

Chilès, J.-P. and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty.* John Wiley, New York.

Cox, D.R. (1975), Prediction intervals and empirical Bayes confidence intervals. In *Perspectives in Probability and Statistics* (ed. J. Gani). Academic Press, London, pp. 47–55.

Cressie, N. (1993), *Statistics for Spatial Data.* Second edition, John Wiley, New York.

Datta, G.S. and Ghosh, M. (1995), Some remarks on noninformative priors. *Journal of the American Statistical Association* **90**, 1357–1363.

Handcock, M.S. and Stein, M. (1993), A Bayesian analysis of kriging. *Technometrics*, **35**, 403-410.

Harville, D.A. (1974), Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Harville, D.A. and Jeske, D.R. (1992), Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724-731.

Le, N.D. and Zidek, J.V. (1992), Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* **43**, 351–374.

Levine, R.A.and Casella, G. (2003), Implementing matching priors for frequentist inference. *Biometrika* **90**, 127–137.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate Analysis.* New York: Academic Press.

Mardia, K.V. and Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Ripley, B.D. (1981), *Spatial Statistics.* Wiley, New York.

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Tierney, L. and Kadane, J. (1986), Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81**, 82–86.

Zhu, Z. (2002), *Optimal Sampling Design and Parameter Estimation of Gaussian Random Fields.* PhD Thesis, Department of Statistics, University of Chicago.

Zhu, Z. and Stein, M.L. (2004), Two-step spatial sampling design for prediction with estimated parameters. Preprint, University of North Carolina at Chapel Hill.

Zimmerman, D.L. and Cressie, N. (1992), Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.