BAYESIAN KRIGING AND BAYESIAN NETWORK DESIGN

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill

OBayes 5 Conference
Branson, Missouri
June 6 2005

Acknowledgements to:
Zhengyuan Zhu
Elizabeth Shamseldin
Stas Kolenikov
Joe Ibrahim

References

For a preliminary version of a paper on this talk is based, see:

Smith, R.L. and Zhu, Z. (2004), Asymptotic theory for kriging with estimated parameters and its application to network design. http://www.stat.unc.edu/postscript/rs/supp5.pdf

For a copy of these transparencies:

http://www.stat.unc.edu/faculty/rs/talks/obayes5.pdf

I Universal Kriging, REML Estimation and Bayesian Spatial Statistics

II Existing Results on “Kriging With Estimated Parameters”

III The Approach Based on Second-Order Asymptotics

IV Application to Network Design

V An Example

I. Universal Kriging, REML Estimation and Bayesian Spatial Statistics

We assume data follow a *Gaussian random field* with mean and covariance functions represented as functions of finite-dimensional parameters.

Define the prediction problem as

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

where Y is an n-dimensional vector of observations, Y_0 is some unobserved quantity we want to predict, X and x_0 are known regressors, β is a q-dimensional vector of unknown regression coefficients, and V, w and v_0 are functions of unknown finite-dimensional θ.

Model (1) arises in various contexts:

- Random effects ANOVA
- Time series models where the covariances are parametrically specified (e.g. ARIMA)
- Spatial statistics

The most widely used spatial models (stationary and isotropic) assume the covariance between components Y_i and Y_j is a function of the (scalar) distance between them, $C_\theta(d_{ij})$. An example is the *exponential power model*

$$C_\theta(d) = \sigma^2 \exp\left\{-\left(\frac{d}{\rho}\right)^\kappa\right\},$$

where $\theta = (\kappa, \sigma^2, \rho)$ with $0 < \kappa \leq 2,\ \sigma^2 > 0,\ \rho > 0$.

The key assumption is: the covariances are unknown in practice, but expressed as functions of finitely many parameters θ.

Universal Kriging

Assume model (1) where the covariances $V,\ w,\ v_0$ are known but β is unknown. The classical formulation of *universal kriging* asks for a predictor $\widehat{Y}_0 = \lambda^T Y$ that minimizes $\sigma_0^2 = E\left\{(Y_0 - \widehat{Y}_0)^2\right\}$ subject to the unbiasedness condition $E\left\{Y_0 - \widehat{Y}_0\right\} = 0$.

The classical solution:

$$\begin{aligned}
\lambda &= w^T V^{-1} + (x_0 - X^T V^{-1} w)^T (X^T V^{-1} X)^{-1} X^T V^{-1}, \\
\sigma_0^2 &= v_0 - w^T V^{-1} w + (x_0 - X^T V^{-1} w)^T (X^T V^{-1} X)^{-1} (x_0 - X^T V^{-1} w).
\end{aligned}$$

Estimation of θ

Consider the model $Y \sim N[X\beta,\ V(\theta)]$.

The *restricted likelihood* is the joint density of an $(n-q)$-dimensional set of contrasts (defined to be independent of β). It is also an *integrated likelihood* with respect to a flat prior on β (Harville 1974). It leads to the formula for the log RL:

$$\ell_n(\theta) = -\frac{1}{2}\log|V(\theta)| - \frac{1}{2}\log|X^T V(\theta)^{-1}X| - \frac{G^2(\theta)}{2},$$

where $G^2 = Y^T W Y$, $W = V^{-1} - V^{-1}X^T(XV^{-1}X^T)^{-1}XV^{-1}$, is the generalized residual sum of squares.

The REML estimator $\widehat{\theta}$ is defined to maximize $\ell_n(\theta)$ w.r.t. θ. It is usually considered superior to MLE, though the two estimators are equivalent to first-order asymptotics.

Bayesian Reformulation

Suppose (β, θ) have a joint prior density of the form $\pi(\theta) d\beta d\theta$ (constant in β).

The Bayesian predictive density of Y_0 given Y is

$$p(Y_0 \mid Y) \;=\; \frac{\int\int f(Y, Y_0 \mid \beta, \theta)\pi(\theta) d\beta d\theta}{\int\int f(Y \mid \beta, \theta)\pi(\theta) d\beta d\theta}.$$

After some algebraic manipulation, this may be rewritten

$$p(Y_0 \mid Y) \;=\; \frac{\int e^{\ell_n(\theta)}\psi(Y_0 \mid Y, \theta)\pi(\theta) d\theta}{\int e^{\ell_n(\theta)}\pi(\theta) d\theta} \qquad (2)$$

where

$$\psi(Y_0 \mid Y, \theta) = \frac{1}{\sqrt{2\pi}\sigma_0(\theta)} \exp\left\{-\frac{1}{2}\left(\frac{Y_0 - \lambda(\theta)^T Y}{\sigma_0(\theta)}\right)^2\right\}.$$

Two forms of predictive density

The conventional kriging formula uses the "plug-in" predictive density

$$\widehat{\psi}(Y_0 \mid Y) = \psi(Y_0 \mid Y, \widehat{\theta}).$$

The natural Bayesian solution uses (2) to define a predictive density, which we shall write $\tilde{\psi}(Y_0 \mid Y)$.

In subsequent discussion, we shall mostly use the predictive distribution function, i.e. redefine $\psi(z|Y,\theta) = \Phi\left(\frac{z-\lambda(\theta)^T Y}{\sigma_0(\theta)}\right)$ where $\Phi(\cdot)$ is the standard normal distribution function, but then define $\widehat{\psi}$ and $\tilde{\psi}$ in the same way.

The central question of this talk is then: *Is $\tilde{\psi}$ superior to $\widehat{\psi}$?* — and how is this influenced by the choice of prior?

III Existing Results on "Kriging With Estimated Parameters"

Suppose we apply universal kriging to predict Y_0 by $\lambda^T Y$, but estimate $\widehat{\lambda} = \lambda(\widehat{\theta})$ where $\widehat{\theta}$ is the MLE or REMLE.

Harville and Jeske (1992) and Zimmerman and Cressie (1992) proposed the following correction to the mean squared prediction error:

$$V_1 \;=\; E\left\{(Y_0 - \widehat{\lambda}^T Y)^2\right\} \;\approx\; \sigma_0^2 + \mathrm{tr}\left\{\mathcal{I}^{-1}\left(\frac{\partial\lambda}{\partial\theta}\right)^T V \left(\frac{\partial\lambda}{\partial\theta}\right)\right\}$$

where $\mathcal{I}$ is the observed information matrix for θ. This formula corrects for the error in specifying the kriging weights λ.

The derivation of this formula assumed that $\widehat{\theta} - \theta$ was independent of $Y_0 - \lambda^T Y$. Abt (1999) derived an improved formula without this assumption, but noted that in practice, the improvement made little difference to the result.

However, in calculating a prediction *interval* for Y_0, it is also necessary to consider the effect of σ_0^2 being unknown. Define

$$V_2 \;=\; \left(\frac{\partial \sigma_0^2}{\partial \theta}\right)^T \mathcal{I}^{-1} \left(\frac{\partial \sigma_0^2}{\partial \theta}\right)$$

Stein (1999) considered the KL divergence

$$D = \int \log \left\{ \frac{p(y_0 \mid Y, \theta)}{p(y_0 \mid Y, \widehat{\theta})} \right\} p(y_0 \mid Y, \theta) dy_0$$

and derived the approximation

$$D \approx \; \frac{1}{2\sigma_0^2} \mathrm{tr} \left\{ \mathcal{I}^{-1} \left(\frac{\partial \lambda}{\partial \theta}\right)^T V \left(\frac{\partial \lambda}{\partial \theta}\right) \right\} + \frac{V_2}{4\sigma_0^4}.$$

This led Zhu and Stein (2004) to suggest

$$V_3 \;=\; V_1 + \frac{1}{2} \cdot \frac{V_2}{\sigma_0^2}$$

could be a suitable *design criterion*. We return to this later.

Bayesian Approaches Based on Reference Priors

As shown by Berger, De Oliveira and Sansó (2001) and extended by Paulo (2003), the reference prior for a Bayesian approach is the same as the Jeffreys prior derived from the restricted likelihood. It is therefore given by

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

where $I(\theta)$ is Fisher information matrix, with entries $\kappa_{i,j}$ given by

$$\kappa_{i,j} = \text{trace}\left\{W\frac{\partial V}{\partial\theta^i}W\frac{\partial V}{\partial\theta^j}\right\},$$

where $W = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1}$.

These authors, as well as Stein (1999), all performed simulations to suggest that Bayesian prediction intervals would perform well if assessed by frequentist coverage probability. One of the aims of the present talk is to present a theoretical discussion of this issue.

IV The Approach Based on Second-Order Asymptotics

Long history —

- *Frequentist Asymptotics for Prediction* — Cox (1975), Barndorff-Nielsen and Cox (1996), Hall, Peng and Tajvidi (1999),...
- *Predictive Likelihood* — Lauritzen (1974), Hinkley (1979), Butler (1986), Davison (1986), Bjørnstad (1990),....
- *Decision Theoretic Approaches* — Aitchison (1975), Harris (1989), Komaki (1996), Smith (1999)
- *Matching Bayesian and Frequentist Inference* — Welch and Peers (1963),......., Datta and Mukerjee (2004 Springer-Verlag Monograph). See in particular, Datta, Mukerjee, M. Ghosh and Sweeting (2000, *Annals of Statistics*) for a "matching prior" approach to predictive inference.

With scattered exceptions, all of this literature applies only to the case of *independent* observations.

Notation

Define

$$\tilde{\psi}(z \mid Y) = \frac{\int e^{\ell_n(\theta)+Q(\theta)}\psi(z \mid Y,\theta)d\theta}{\int e^{\ell_n(\theta)+Q(\theta)}d\theta} \qquad (3)$$

where $e^{\ell_n(\theta)}$ is the restricted likelihood of θ, $Q(\theta) = \log \pi(\theta)$ and $\psi(z \mid Y, \theta) = \Phi\left(\frac{z-\lambda(\theta)^T Y}{\sigma_0(\theta)}\right)$. Also let $\tilde{\psi}^{-1}$ be inverse function, i.e. $\tilde{\psi}^{-1}(P \mid Y)$ is the value of z for which $\tilde{\psi}(z \mid Y) = P$.

For $P \in (0,1)$ define

$$\begin{aligned} z_P(Y \mid \theta) &= \lambda(\theta)^T Y + \sigma_0(\theta)\Phi^{-1}(P), \\ \hat{z}_P(Y) &= \hat{\lambda}^T Y + \hat{\sigma}_0\Phi^{-1}(P), \\ \tilde{z}_P(Y) &= \tilde{\psi}^{-1}(P \mid Y). \end{aligned}$$

For an estimator z_P^* (could be $\widehat{z}_P$ or $\widetilde{z}_P$) we would like to calculate

$$E\left\{\psi(z_P^*(Y) \mid Y, \theta) - \psi(z_P(Y| \theta) \mid Y, \theta)\right\} \tag{4}$$

and

$$E\left\{z_P^*(Y) - z_P(Y \mid \theta)\right\} \tag{5}$$

(4) is called the coverage probability bias (CPB). (5) leads to the expected length of a prediction interval (our proposed design criterion) because for a $100(P_2 - P_1)\%$ interval,

$$\begin{aligned} & E\left\{z_{P_2}^*(Y) - z_{P_1}^*(Y)\right\} \\ = \; & E\left\{z_{P_2} - z_{P_1}\right\} + E\left\{z_{P_2}^* - z_{P_2}\right\} - E\left\{z_{P_1}^* - z_{P_1}\right\} \\ = \; & \sigma_0\{\Phi^{-1}(P_2) - \Phi^{-1}(P_1)\} + E\left\{z_{P_2}^* - z_{P_2}\right\} - E\left\{z_{P_1}^* - z_{P_1}\right\} \end{aligned}$$

Define $U_i = \frac{\partial \ell_n(\theta)}{\partial \theta^i}$, $U_{ij} = \frac{\partial^2 \ell_n(\theta)}{\partial \theta^i \partial \theta^j}$, $U_{ijk} = \frac{\partial^3 \ell_n(\theta)}{\partial \theta^i \partial \theta^j \partial \theta^k}$.
The matrix with entries U_{ij} has an inverse with entries U^{ij}.

Other quantities $Q(\theta) = \log \pi(\theta)$, $\lambda(\theta)$, $\sigma_0(\theta)$. Suffixes denote partial differentiation, e.g. $Q_i = \frac{\partial Q}{\partial \theta^i}$, $\sigma_{0ij} = \frac{\partial^2 \sigma_0}{\partial \theta^i \partial \theta^j}$. Let

$$\begin{aligned} U_i &= n^{1/2} Z_i, \\ U_{ij} &= n^{1/2} Z_{ij} + n\kappa_{ij}, \\ U_{ijk} &= n^{1/2} Z_{ijk} + n\kappa_{ijk}, \end{aligned}$$

and define also $\kappa_{i,j} = n^{-1} E\left\{U_i U_j\right\} = -\kappa_{ij}$, $\kappa_{ij,k} = n^{-1} E\left\{U_{ij} U_k\right\}$. Suppose inverse of $\{\kappa_{i,j}\}$ matrix has entries $\{\kappa^{i,j}\}$. We assume all the Z quantities are $O_p(1)$ and all the κ quantities are $O(1)$ as $n \to \infty$ (*increasing domain asymptotics*) and we employ the summation convention.

Step 1: Taylor expansion of $\widehat{\psi}$

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

Follows well-known references on higher-order asymptotics of MLE, e.g. McCullagh (1987), Barndorff-Nielsen and Cox (1994).

Step 2: From $\widehat{\psi}$ to $\tilde{\psi}$

Using a Laplace approximation to the Bayesian integral, Lindley (1980) showed that

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

where the hats indicate that U_{ijk}, ψ_ℓ, etc., are evaluated at the REMLE $\theta = \widehat{\theta}$.

Hence

$$\begin{aligned}\tilde{\psi} - \widehat{\psi} \;&=\; \frac{1}{2n}\left\{\kappa_{ijk}\kappa^{i,j}\kappa^{k,\ell}\psi_\ell + (\psi_{ij} + 2\psi_i Q_j)\kappa^{i,j}\right\} \\ &\quad +O_p(n^{-3/2}).\end{aligned}$$

Together, these approximations give an expansions of either $\widehat{\psi}-\psi$ or $\tilde{\psi} - \psi$ in powers of $n^{-1/2}$.

Step 3: From a distribution function to its inverse (based on Cox (1975))

Suppose $\psi^*(z)$ (could be $\widehat{\psi}$ or $\tilde{\psi}$) is an estimator of the conditional prediction distribution function $\psi(z) = \psi(z \mid Y, \theta)$ that has an expansion

$$\psi^*(z) \;=\; \psi(z) + n^{-1/2}R + n^{-1}S + o(n^{-1}).$$

Define predictive quantile z_P^* by $\psi^*(z_P^*) = P$. Then

$$z_P^* - z_P \;=\; -n^{-1/2}\frac{R}{\psi'} - n^{-1}\left(\frac{RR'}{\psi'^2} - \frac{R^2}{\psi'^3} - \frac{S}{\psi'}\right) + o_p(n^{-1}), \quad (6)$$

$$\psi(z_P^*) - \psi(z_P) \;=\; -n^{-1/2}R - n^{-1}\left(\frac{RR'}{\psi'} - S\right) + o_p(n^{-1}). \quad (7)$$

Here primes denote differentiation with respect to z. By taking expectations in (6) and (7) respectively, we derive expressions for the expected length of a prediction interval and the coverage probability bias (CPB).

Step 4: Evaluate the expectations

Side comment: Up to this point, the calculations are the same as in the independent case. In particular, by taking expectations in (7), it should be possible to re-derive the results in Datta, Mukerjee, Ghosh and Sweeting (2000) (though they used a different method)

However, for dependent observations, these expectations are much harder to evaluate.

As an example of the terms to be evaluated, consider

$$\begin{aligned} E\left\{\frac{RR'}{\psi'}\right\} &= E\left\{\frac{\kappa^{i,j}Z_i\psi_j\kappa^{k,\ell}Z_k\psi'_\ell}{\psi'}\right\} \\ &= \kappa^{i,j}\kappa^{k,\ell}\phi(\Phi^{-1}(P)) \times \end{aligned}$$

$$E\left[Z_iZ_k\left\{\frac{\lambda_j^TY}{\sigma_0}+\frac{\sigma_{0j}}{\sigma_0}\Phi^{-1}(P)\right\}\left\{\frac{\sigma_{0\ell}}{\sigma_0}-\left(\frac{\lambda_\ell^TY}{\sigma_0}+\frac{\sigma_{0\ell}}{\sigma_0}\Phi^{-1}(P)\right)\Phi^{-1}(P)\right\}\right]$$

This requires evaluating moments of Y's of up to sixth order, but it is still an explicit calculation!

Results

$$
\begin{aligned}
& nE\left\{\psi(\widehat{z}_P(Y) \mid Y, \theta) - \psi(z_P(Y \mid \theta) \mid Y, \theta)\right\} \\
\sim \;& \phi(\Phi^{-1}(P))\Phi^{-1}(P)\left[-\frac{1}{2}\Phi^{-1}(P)^2 \kappa^{i,j}\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2}\right. \\
& +\kappa^{i,j}\kappa^{k,\ell}\left(\kappa_{jk,\ell} + \frac{1}{2}\kappa_{jk\ell}\right)\frac{\sigma_{0i}}{\sigma_0} + \frac{1}{2}\kappa^{i,j}\left\{\frac{\sigma_{0ij}}{\sigma_0} - \frac{\lambda_i^T V \lambda_j}{\sigma_0^2}\right\} \\
& \left.-\frac{1}{2}\kappa^{i,k}\kappa^{j,\ell}\cdot\frac{1}{n\sigma_0^2}\left(\lambda_i^T V \frac{\partial W}{\partial \theta^k} V \frac{\partial W}{\partial \theta^\ell} V \lambda_j + \lambda_i^T V \frac{\partial W}{\partial \theta^\ell} V \frac{\partial W}{\partial \theta^k} V \lambda_j\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
& nE\left\{\psi(\tilde{z}_P(Y) \mid Y, \theta) - \psi(z_P(Y \mid \theta) \mid Y, \theta)\right\} \\
\sim \;& \phi(\Phi^{-1}(P))\Phi^{-1}(P)\left[\kappa^{i,j}\kappa^{k,\ell}\left(\kappa_{jk,\ell} + \kappa_{jk\ell}\right)\frac{\sigma_{0i}}{\sigma_0}\right. \\
& -\kappa^{i,j}\left(\frac{\sigma_{0i}\sigma_{0j}}{\sigma_0^2} - \frac{\sigma_{0ij}}{\sigma_0}\right) + \kappa^{i,j}\frac{\sigma_{0i}}{\sigma_0}Q_j \\
& \left.-\frac{1}{2}\kappa^{i,k}\kappa^{j,\ell}\cdot\frac{1}{n\sigma_0^2}\left(\lambda_i^T V \frac{\partial W}{\partial \theta^k} V \frac{\partial W}{\partial \theta^\ell} V \lambda_j + \lambda_i^T V \frac{\partial W}{\partial \theta^\ell} V \frac{\partial W}{\partial \theta^k} V \lambda_j\right)\right].
\end{aligned}
$$

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

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

These results imply the existence of a "matching prior" for which the second-order CPB is 0. However we can also manipulate the asymptotic expressions to obtain a direct estimate of z_P with the same property:

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

IV. Application to Network Design

Large literature, many different approaches.

Recent work has focussed on contrast between two types of criterion:

- *Estimative* — e.g. choose the design to maximize the determinant of the Fisher information matrix of θ
- *Predictive* — focus on a specific Y_0, find a design to minimize σ_0. Note that this ignores the estimation of θ, in effect assuming θ known.

Zhu and Stein (2004) proposed a combined estimative and predictive criterion, using approximations derived by Stein (1999).

They also considered (but rejected as too computationally intensive) a direct Bayesian approach, choosing the optimal design to minimize the expected length(s) of a Bayesian prediction interval for the quantity (or quantities) being predicted.

Direct Bayesian Approach

- For any data set, use MCMC to construct the Bayesian predictive distribution
- For any given design, run the Bayesian analysis on simulated data sets to determine the expected length of Bayesian prediction intervals
- Use an optimization algorithm (e.g. simulated annealing) to find the optimal design

Direct implementation of this approach requires a lot of Monte Carlo simulation.

The new result is that these two criteria of Zhu and Stein are almost equivalent — taking the “direct Bayesian approach” but using asymptotic approximations, one derives a criterion for the optimal design very similar to the V_3 criterion given earlier.

Suppose we use an estimator of z_P whose second-order CPB is 0 (e.g. either the Bayes estimator with matching prior, or $z_P^\dagger$). In either case we have

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

The second line is Abt's refinement of the Harville-Jeske-Zimmerman-Cressie correction and will be ignored in subsequent discussion.

Use this to construct a two-sided prediction interval, with tail probability $1 - P$ in each tail. The approximate expected length of this prediction interval is

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

In the notation of Zhu and Stein (2004), the quantity under the square root sign is

$$V_4 = V_1 + \frac{\Phi^{-1}(P)^2}{4} \cdot \frac{V_2}{\sigma_0^2}.$$

Recall their own criterion was $V_3 = V_1 + \frac{1}{2} \cdot \frac{V_2}{\sigma_0^2}$.

Two Design Criteria

$$
\begin{aligned}
V_3 &= V_1 + \frac{1}{2} \cdot \frac{V_2}{\sigma_0^2} \quad \text{(Zhu and Stein)} \\
V_4 &= V_1 + \frac{\Phi^{-1}(P)^2}{4} \cdot \frac{V_2}{\sigma_0^2} \quad \text{(this talk)}
\end{aligned}
$$

The present formula V_4 has the unusual feature that the design might depend on the desired coverage probability of a prediction interval.

It is also tied directly to two specific methods of constructing a prediction interval whose second-order coverage probability bias is 0, whereas previous approaches have not shown how to construct such an interval.

V. An Example

In North Carolina there are 38 monitors for $\mathrm{PM}_{2.5}$ (fine particulate matter). Suppose we wanted to redesign the network for optimal estimation of population-weighted daily average. We use daily data from 2000. Assume individual days' data are independent replications of the model

$$Cov(y_i, y_j) = \begin{cases} \theta_1^2 & \text{if } i = j, \\ \theta_3 \theta_1^2 e^{-d_{ij}/\theta_2} & \text{if } i \neq j, \end{cases}$$

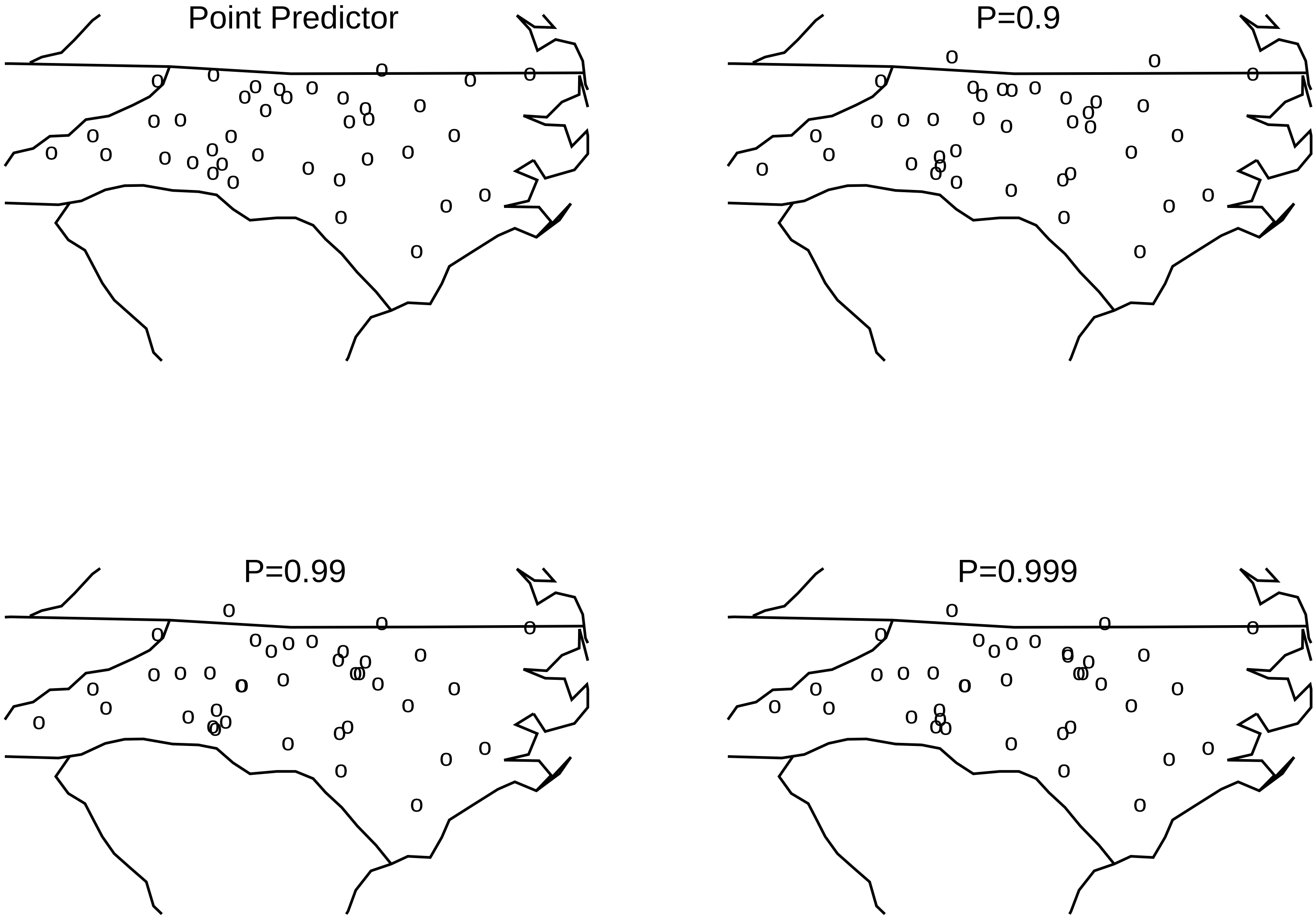
with y_i, y_j the $\mathrm{PM}_{2.5}$ at locations i and j, d_{ij} is distance (units of 100 km.), and we estimated $\theta_1 = 6.495$, $\theta_2 = 4.019$, $\theta_3 = .9423$. Treat this as the true model, but assume $\theta_1, \theta_2, \theta_3$ would have to be re-estimated on any given day.

Population-weighted averages were calculated using data from the 2000 U.S. census for the 809 zip code tabulation areas (ZCTA) in North Carolina. Select 38 ZCTA out of 809 to place the monitoring station to give most accurate prediction of the total population PM2.5 exposure defined as

$$y_0 = \sum_i p_i y_i,$$

where p_i is the population at the i'th ZCTA, and y_i is the PM2.5 level there. V_1 and V_4 with coverage probabilities $P = 0.9, 0.99, 0.999$ are used as design criteria, and a simulated annealing algorithm is used to find the optimal designs.

Optimal Designs Under Four Criteria



Four designs selected using criteria of this talk (calculations due to Zhengyuan Zhu)

All four designs tend to place monitors in regions of high population density (as does the current EPA network) but it is noticeable that the criterion V_4, especially for smaller P, tends to favor a network with clusters of nearby monitors, reflecting the role such clusters play in ensuring good estimation of model parameters.

Summary

1. The second-order coverage probability bias of the Bayes estimator of z_P is smaller than that of the plug-in estimator in the limit as $P \to 0$ or 1, regardless of the prior.

2. For the Bayesian predictive distribution there is a matching prior, i.e. one for which the second-order CPB of $\tilde{z}_P$ is 0.

3. However we can also achieve the same second-order properties directly, using the estimator $z_P^\dagger$.

4. For any of these estimators of predictive quantiles, we have an approximation for the expected length of a prediction interval, and this can be used as a design criterion.

5. In the case of an estimate whose second-order CPB is 0, we obtain a design criterion very similar to that of Zhu and Stein, but adapted to a specific construction of a prediction interval.