KRIGING WITH ESTIMATED PARAMETERS

Richard L. Smith

Department of Statistics and Operations Research

University of North Carolina

Chapel Hill

Iowa State University

September 12 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 background about the $PM_{2.5}$ application:

R.L. Smith, S. Kolenikov and L.H. Cox (2003), Spatio-temporal modeling of PM2.5 data with missing values. *J. Geophys. Res.*, **108**(D24), 9004, doi:10.1029/2002JD002914, 2003.
http://www.stat.unc.edu/postscript/rs/Smith-JGR-2003.pdf

I Motivating Example: Interpolation of Air Pollution Data

II Universal Kriging, REML Estimation and Bayesian Spatial Statistics

III Existing Results on “Kriging With Estimated Parameters”

IV The Approach Based on Second-Order Asymptotics

V Example Based on $\mathrm{PM}_{2.5}$ Monitors in North Carolina

I. Motivating Example: Interpolation of Air Pollution Data

The problems of spatial interpolation and network design are introduced through an example using the EPA fine particulates ($PM_{2.5}$) network. The analysis is taken from Smith, Kolenikov and Cox (2003).

Current EPA standard for $PM_{2.5}$:

- Twenty-four hour average $PM_{2.5}$ not to exceed 65 $\mu g/m^3$ for a three-year average of annual 98th percentiles at any population-oriented monitoring site in a monitoring area.

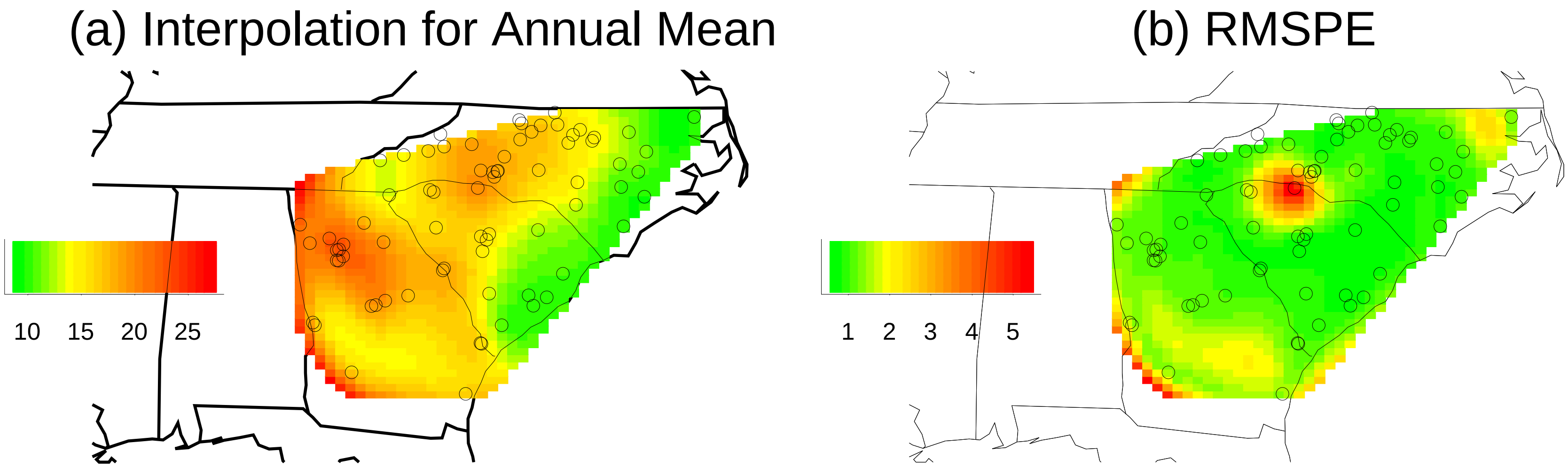
- Three-year annual average $PM_{2.5}$ not to exceed 15 $\mu g/m^3$ concentrations from a single community-oriented monitoring site or the spatial average of eligible community-oriented monitoring sites in a monitoring area.

The data (compiled from a larger data set) consisted of weekly average $PM_{2.5}$ levels during 1999 at 74 EPA stations in NC, SC, GA. We fitted a model of the form

$$y_{xt} \;=\; f(x) + g(t) + \eta_{xt}$$

in which y_{xt} is the square root of $PM_{2.5}$ in location x in week t, $f(x)$ and $g(t)$ are fixed functions of time and space respectively (both represented as linear regression functions), and η_{xt} is a random error (independent in time but dependent in space).

The model parameters were estimated by maximum likelihood, and kriging (described later) was used to interpolate η_{xt} at locations x off the network. Based on that, a map was constructed of the interpolated annual average for each point in the study region, together with an estimated root mean square prediction error.



(a) Interpolation for Annual Mean

(b) RMSPE

Fig. 1. (a) Interpolated surface for annual mean $PM_{2.5}$ in $\mu g/m^3$, and monitor locations (circles). (b) Root mean squared prediction error for the surface in (a). Adapted from Smith, Kolenikov and Cox (2003).

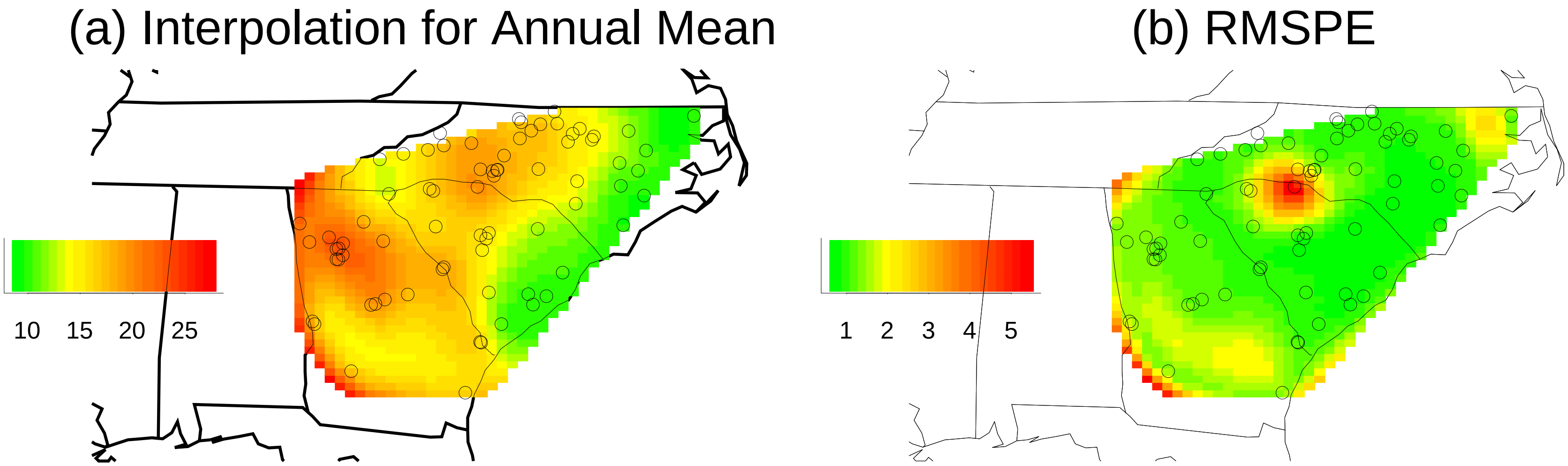


Fig. 2. (a) Interpolated surface for annual mean $PM_{2.5}$ in $\mu g/m^3$, and monitor locations (circles). (b) Root mean squared prediction error for the surface in (a). Similar to Fig. 1 but uses Bayesian interpolation.

Questions for this talk:

There are many questions along the lines of what is the right model for this data set, but in this talk, I focus on two more theoretical aspects:

1. *Computation of RMSPEs.* The calculations above used the conventional formulae for kriging prediction errors that you can find in any text on geostatistics, but these ignore model estimation errors. Is this appropriate, and how could we improve on this approach?
2. *Design of the network.* The EPA is constantly adding or deleting stations as it attempts to provide better coverage or to reduce costs. How could it do this most efficiently?

The two questions are linked because accurate determination of kriging errors is critical in assessing the design.

II. 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

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

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, and β is a p-dimensional vector of unknown regression coefficients. For the moment, we assume V, w and v_0 are known.

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 PM$_{2.5}$ data analysis actually assumed

$$Var\{Y_i - Y_j\} = \theta_1 + \theta_2 d_{ij}^{\theta_3} \qquad (\theta_1 > 0, \theta_2 > 0, 0 < \theta_3 \leq 2).$$

This is of *intrinsically stationary* form and $C_\theta(d)$ does not exist, though the model can easily be transformed into one for which the methods of this talk apply.

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:

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

Maximum Likelihood and REML Estimation

Model of form

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

A classical method of estimation is the *method of maximum likelihood* (MLE), in which the parameters β and θ are chosen to maximize the joint density of Y given β and θ (the likelihood function).

In practice, once θ is specified, the MLE $\widehat{\beta}$ can be calculated by elementary algebra (the generalized least squares estimator). Therefore, in practice MLE is computed by maximizing the *profile log likelihood*

$$-\frac{1}{2}\log|V(\theta)| - \frac{G^2(\theta)}{2}$$

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.

An alternative to MLE is to use the *restricted likelihood function*. As defined originally by Patterson and Thompson (1971), if X is $n \times p$, we define a $(n-p)$-dimensional vector of contrasts $A^T Y$, where A is $n \times (n-p)$ has rank $n-p$ and satisfies $A^T X = 0$. The restricted likelihood is the density of $A^T Y \sim N[0, A^T V A]$.

Harville (1974) gave an equivalent Bayesian definition and derived the formula

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

The resulting estimator is called REML. 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

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

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 REML estimator $\widehat{\theta}$ is the value of θ that maximizes $\ell_n(\theta)$.

The conventional kriging formula uses the predictive density

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

also known as *estimative density* (Aitchison) or *plug-in rule*.

In contrast to this, we write (2) as $\tilde{\psi}(Y_0 \mid Y)$, which Aitchison called the *predictive density*.

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.

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

Designing Monitor Networks

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) discussed the idea of using Bayesian prediction intervals as a basis for network design, arguing that Bayesian intervals take account of the uncertainty of θ and therefore should be superior to the predictive approach. However, they rejected this as too computationally intensive, and instead proposed a two-stage criterion (more later).

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 second major theme of this talk is to propose an approximate approach that reduces the first two steps to a simple algebraic formula for the expected length of a Bayesian prediction interval.

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. Stein (1999) and Zhu and Stein (2004) defined

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

as a measure of the uncertainty in σ_0^2, and they suggested that some linear combination of V_1 and $\frac{V_2}{\sigma_0^2}$ would best measure the overall uncertainty. In particular, they suggested

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

as a suitable combined criterion. However, it's not clear exactly why this particular linear combination is appropriate.

Bayesian Approaches Based on Reference Priors

For a model with parameters (β, σ^2, ρ) where σ^2 is the marginal variance and ρ is the range parameter, Berger, De Oliveira and Sansó (2001) derived the "reference prior" in the form

$$\pi(\beta, \sigma^2, \rho) \propto \frac{1}{\sigma^2}\left[\mathrm{tr}\left(W\frac{\partial V}{\partial \rho}W\frac{\partial V}{\partial \rho}\right) - \frac{1}{n-p}\left\{\mathrm{tr}\left(W\frac{\partial V}{\partial \rho}\right)\right\}^2\right]^{1/2}. \quad (3)$$

They compare this with alternative definitions of reference prior and Jeffreys prior. Equation (3) is derived as the Jeffreys prior based on the restricted likelihood.

Paulo (2005) extended their result to a general multi-parameter spatial likelihood.

Both papers used simulation to demonstrate that the method would produce good frequentist coverage probability.

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

$$\tilde{\psi}(z \mid Y) = \frac{\int e^{\ell_n(\theta)+Q(\theta)}\psi(z \mid Y, \theta)d\theta}{\int e^{\ell_n(\theta)+Q(\theta)}d\theta} \qquad (4)$$

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

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

and

$$E\left\{z_P^*(Y) - z_P(Y \mid \theta)\right\} \tag{6}$$

(5) is called the coverage probability bias (CPB). (6) 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 $n \to \infty$ and we employ the summation convention.

Results

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

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

We can also evaluate the expected length of a prediction interval using the formulas

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

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

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:

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

In practice it seems possible to ignore the last line (Abt correction to the Harville-Jeske-Zimmerman-Cressie formula).

Application to Network Design

Suppose we are interested in constructing a network to optimize the prediction of a specific quantity Y_0 (e.g. a population-weighted exposure to a pollutant). Suppose we use $z_P^\dagger$ 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. Example Based on PM$_{2.5}$ Monitors in North Carolina

Assume the objective is to estimate population-weighted daily average. 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 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.

Assume number of monitors remains fixed at the current number of 38.

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 two-sided tail probabilities $P = 0.1, 0.01, 0.001$ are used as design criteria, and a simulated annealing algorithm is used to find the designs given in Figure 3.

Optimal Designs Under Four Criteria

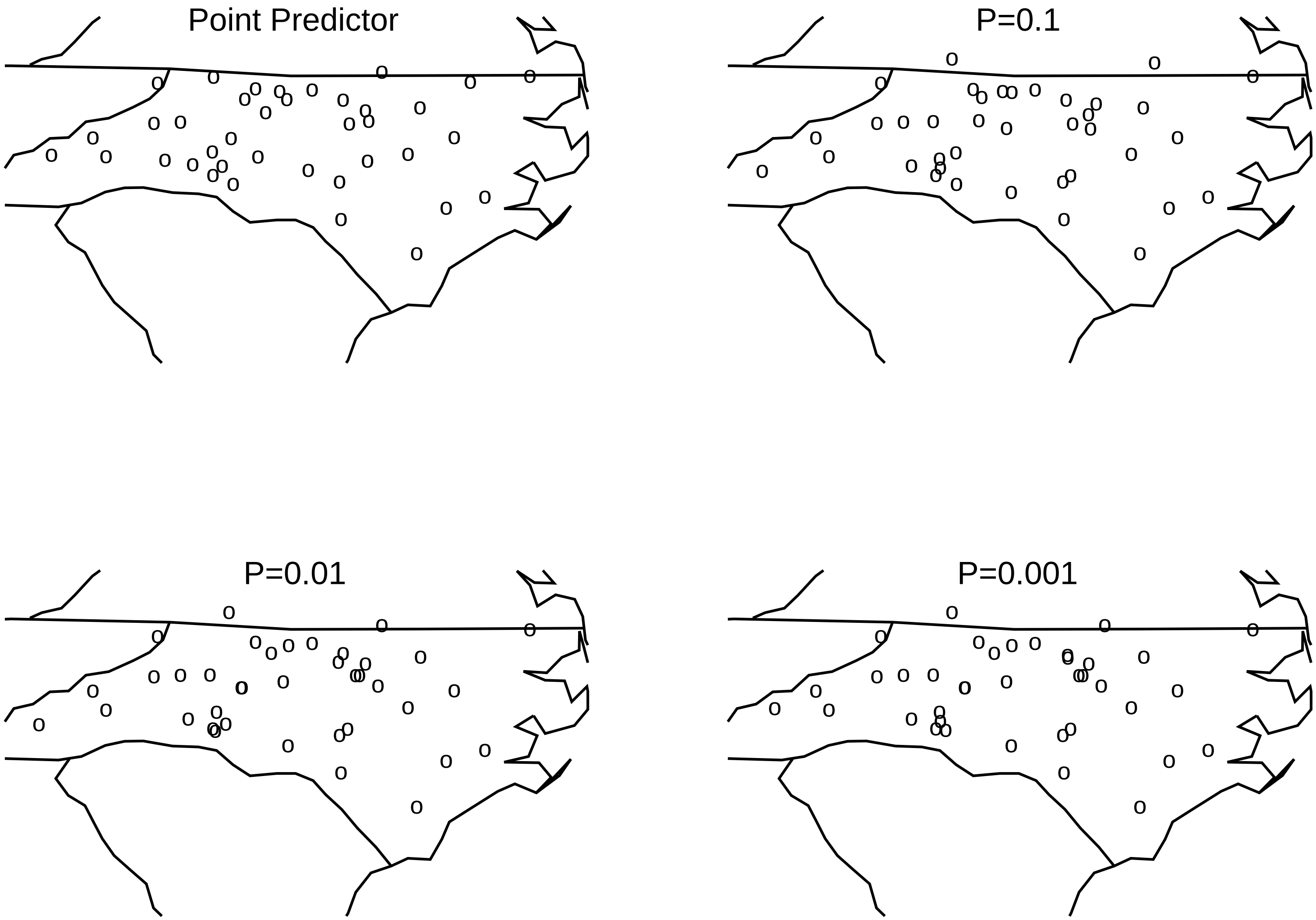


Fig. 3. Four designs selected using criteria of this talk (calculations due to Zhengyuan Zhu)

All four designs tend to place monitors in regions of high population density (as does the current EPA network, Fig. 1) 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.