STATISTICS 890: SPRING 2009

SPATIAL STATISTICS

PART 2

NONSTATIONARY SPATIAL PROCESSES

11. Moving window methods
12. EOFs and extensions
13. Deformation approaches
14. Kernel approaches

MODELS DEFINED BY CONDITIONAL PROBABILITIES

15. Markov random fields as spatial models
16. Inference in Markov random fields
17. Examples of Markov random fields
18. Markov random fields as spatial priors

DESIGN OF MONITORING NETWORKS

19. Maximum entropy and Bayesian approaches to network design
20. Applications of optimal design theory
21. Other approaches to monitoring design
22. Designs for data assimilation

11. Moving Window Methods

Idea: to perform kriging at a sampling location s, we should restrict ourselves to a "window" of sampling stations close to s, within which it is reasonable to assume a homogeneous model. Compromise between stationary models and truly nonstationary approaches.

Here we follow primarily Haas (1995).

Specifying the model

Suppose we have a sample $\{Z(t_i, s_i)\}$ from a spatio-temporal process Z at n time-space points $\{(t_i, s_i),\ 1 \leq i \leq n\}$.

Specify two parameters:
•*time window* m_T
•*sampling fraction* f_c

To predict at (t_0, s_0):

• Restrict to observations within time window $(t_0 - \frac{m_T}{2}, t_0 + \frac{m_T}{2})$

• Within that time window, select spatial points in order of nearness to s_0, until $n_c = nf_c$ points have been selected.

• Fit these n_c data points to the regression model

$$Z(t, s) = \mu(t, s, \beta) + \psi(\mu(t, s, \beta))R(t, s)$$

where μ and ψ are parametric functions of β and R is a residual process,

• Covariance function for R assumed to be

$$C\{R(t_1, s_1), R(t_2, s_2)\}$$
$$= C_T(t_2 - t_1)C_S(s_2 - s_1)$$

C_T: temporal covariance function
C_S: spatial covariance function each stationary within window

Haas assumed "spherical" form for C_T, C_S (with geometric anisotropy?)

May need separate model for different seasons.

• Then perform kriging at (t_0, s_0), with prediction standard error $S_e(t_0, s_0)$.

Some complications

1. Selection of window size: Haas recommended cross-validation based on accuracy of prediction variances (rather than minimizing the prediction variances)

2. The covariances estimated by this approach do not lead to a positive definite covariance matrix over all the sites. Haas (1998) proposed a way round this using numerical analysis algorithms to find the nearest p.d. matrix in some metric, but this is not as good as finding an explicit stochastic model for the entire process.

Recent developments (Haas 2002)

Features:

- Multivariate processes
- Transformations to normality — represent as trend + residual
- Further transformations of residuals to remove skewness and kurtosis
- Possible long memory in time
- Local or global estimation

Some specifics

ith coordinate process W_i, write as $W_i = t_i^{-1}(Y_i)$ with Y_i continuous (possible for W_i to be discrete, e.g. counting process)

$Y_i = \mu_i + R_i$, μ_i parametric trend, R_i residual

Transform R_i to Z_i to remove skewness and kurtosis

$Cov\{Z_i(x_1, y_1, t_1), Z_j(x_2, y_2, t_2)\} = C_{ij}(g, h)$ where g is the spatial separation and h is the temporal separation

Separable covariance functions, temporal covariance includes a parametric long-memory component

LOMAP process: for prediction at a specified space-time point, restrict to cylindrical window as in earlier Haas papers. Window size selected by cross-validation

GLOMAP process: represent global covariance function as a finite sum of local covariances using kernel weights (cf. Section 14)

12. The EOF method and extensions

EOF=*empirical orthogonal functions*

Also called *Karhunen-Loève expansion* and many other things

In finite samples reduces to *principal components analysis*

Good background refs: Cohen and Jones (1969), North (1984). Follow Cohen and Jones initially

Formulation: independent observation vectors $(X_i(s), y_i),\ 1 \le i \le n$ where each X_i ranges over $s \in \mathcal{D}$ (bounded) and y_i is the variable we want to predict

(e.g. y_i=temperature at Washington airport, X_i=pressure field over the northern hemisphere)

Model

$$y_i = \int_{\mathcal{D}} X_i(s)B(s)ds + \epsilon_i, \quad 1 \le i \le n,$$

ϵ_i ind. of $X_i(s),\ s \in \mathcal{D}$.

LSE for $B(s)$ solves

$$\sum y_i X_i(t) = \int_{\mathcal{D}} \sum_i X_i(s)X_i(t)B(s)ds.$$

Assume y_i and $X_i(s)$ have mean 0.

$n \to \infty$:

$$n^{-1} \sum y_i X_i(t) \to C_y(t),$$
$$n^{-1} \sum_i X_i(x) X_i(t) \to C(s,t),$$

where $C_y(t)$ is the covariance of y_i and $X_i(t)$ and $C(s,t) = \text{Cov}\{X_i(s), X_i(t)\}$. Hence $B(s)$ solves

$$C_y(t) = \int_{\mathcal{D}} C(s,t) B(s) ds.$$

Karhunen-Loève expansion: solve

$$\int_{\mathcal{D}} C(s,t)\psi_\nu(t)dt = \lambda_\nu \psi_\nu(s),$$

$\nu = 1, 2, ...,$ to find eigenvalues λ, eigenfunctions ψ. Typically can find *complete orthonormal basis* $\{\psi_\nu\}$ such that

$$\int_{\mathcal{D}} \psi_\mu(s)\psi_\nu(s)ds = \begin{cases} 1 & \text{if } \nu = \mu \\ 0 & \text{if } \nu \neq \mu \end{cases}$$

and for any square integrable function g,

$$g(s) = \sum_\nu a_\nu \psi_v(s),$$

$$a_\nu = \int_{\mathcal{D}} g(s)\psi_\nu(s)ds.$$

In particular,

$$C(s,t) = \sum_\nu \lambda_\nu \psi_\nu(s)\psi_\nu(t).$$

Also have representation of process itself:

$$X(s) = \sum_{\nu} z_\nu \lambda_\nu^{1/2} \psi_\nu(s),$$

$z_\nu \sim N[0,1]$ (i.i.d.)

The solution to integral equation

$$C_y(t) = \int_{\mathcal{D}} C(s,t) B(s) ds$$

is given by

$$B(s) = \sum_{\nu} \beta_\nu \psi_\nu(s)$$

$$\beta_\nu = \frac{1}{\lambda_\nu} \int C_y(t) \psi_\nu(t) dt.$$

Also

$$y_i = \sum_{\nu} z_{i\nu} \lambda_\nu^{1/2} \beta_\nu + \epsilon_i$$

In practice, truncate the sum at $\nu = N$. Leads to *principal components regression* for y_i.

Also, in practice, we would only observe the $X_i(s)$ process at a finite set of locations s, and in this case, the whole analysis is equivalent to finding a principal components decomposition of the sample covariance matrix of X.

Summary

1. For general class of $C(s,t)$, find complete orthonormal basis of eigenfunctions ψ_ν with eigenvalues λ_ν; the covariance function and the process itself can be expressed as expansions in the ψ_ν.

2. Best predictor of y_i is of form $\int_{\mathcal{D}} X_i(s)B(s)ds$ where we can express $B(s)$ as an expansion in functions ψ_ν

3. In practice, use principal components

Combining stationary models and EOFs

Nychka and Saltzman (1998), Holland *et al.* (1999)

$$C(s_1, s_2) =$$

$$\sigma(s_1)\sigma(s_2)\Bigg[\rho\Bigg\{(1-\alpha)\delta(s_1 - s_2)$$

$$+ \alpha e^{-||s_1 - s_2||/\theta}\Bigg\} + \sum_{\nu=1}^{M} \lambda_\nu \psi_\nu(s_1)\psi_\nu(s_2)\Bigg],$$

$$Z(s) = \sigma(s)\Bigg\{(\alpha\rho)^{1/2} Z_0(s)$$

$$+ \sum_{\nu=1}^{M} a_\nu \lambda_\nu^{1/2} \psi_\nu(s)\Bigg\} + \epsilon(s)$$

where $\epsilon(s) \sim N[0, \sigma^2(s)(1-\alpha)\rho]$ independently at each site s.

Wavelet expansions

Nychka, Wikle and Royle (1999) used wavelet expansions in place of eigenfunction expansions on an $M \times N$ grid.

$$Z(s) = \sum_{\nu=1}^{MN} a_\nu \psi_\nu(s),$$

basis functions ϕ_ν given,
$(a_\nu) \sim N_{MN}[0, \Sigma_a]$. Then

$$\Sigma_Z = \Psi \Sigma_a \Psi^T.$$

Disadvantage: unlike eigenfunctions expansions, cannot assume Σ_a diagonal

Advantage: In practice, can get away with sparse matrix for Σ_a, and then there are significant computational advantages to choosing ψ_ν with convenient properties.

13. Deformation models

Covariance function and dispersion:

$$\begin{aligned} C(s_1, s_2) &= \text{ Cov } \{Z(s_1), Z(s_2)\}, \\ D(s_1, s_2) &= \text{ Var } \{Z(s_1) - Z(s_2)\}, \end{aligned}$$

If either $C(s_1, s_2)$ or $D(s_1, s_2)$ depends on s_1 and s_2 only through $||s_1 - s_2||$, the process is *homogeneous*.

Simple inhomogeneous models

$$D(s_1, s_2) = 2\gamma_0(||A_0(s_1 - s_2)||),$$

$$D(s_1, s_2) = 2\sum_{j=0}^{J-1} \gamma_j(||A_j(s_1 - s_2)||),$$

(geometric anisotropy, zonal anisotropy) too simple for real applications.

Guttorp-Sampson idea:

$$D(s_1, s_2) = 2\gamma_0(f(s_1), f(s_2))$$

with γ_0 some homogeneous semivariogram and f a nonlinear function from *G-space* to *D-space*.

Fitting methods

1. Original Sampson-Guttorp (1992) algorithm: three stage procedure involving multidimensional scaling to determine embedding of site locations in D-space, followed by smoothing splines to construct a smooth map, then estimation of γ_0.

2. Alternative due to Guttorp *et al.* (1994): Minimize

$$\sum_{i,j} \left(\frac{d_{ij} - \hat{d}_{ij}}{\hat{d}_{ij}} \right)^2 + \lambda \{ J(f_1) + J(f_2) \},$$

where d_{ij} is the empirical dispersion between sites i and j, $\hat{d}_{ij}$ is the modeled dispersion, f_1 and f_2 are the x and y coordinates of the deformation and

$$J(f_i) = \int \int \left\{ \left(\frac{\partial^2 f_i}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f_i}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f_i}{\partial y^2} \right)^2 \right\} dx dy.$$

3. Maximum likelihood (Mardia-Goodall 1993, Smith 1996): parameterize f_1 and f_2 in terms of *radial basis functions* and minimize

$$NLLH = \frac{N}{2} \log |\Sigma| + \frac{N-1}{2} \text{ tr } \left(\Sigma^{-1} \hat{\Sigma} \right)$$

where N is number of replications, Σ is modeled correlation matrix and $\hat{\Sigma}$ is sample correlation matrix.

For stationary covariance C_0 may take standard parametric form (e.g. Matérn) or else

$$C_0(h) = \sum_{c=1}^{C} \phi_c J_0(w_c h)$$

with J_0 a Bessel function, cf. Shapiro and Botha (1991).

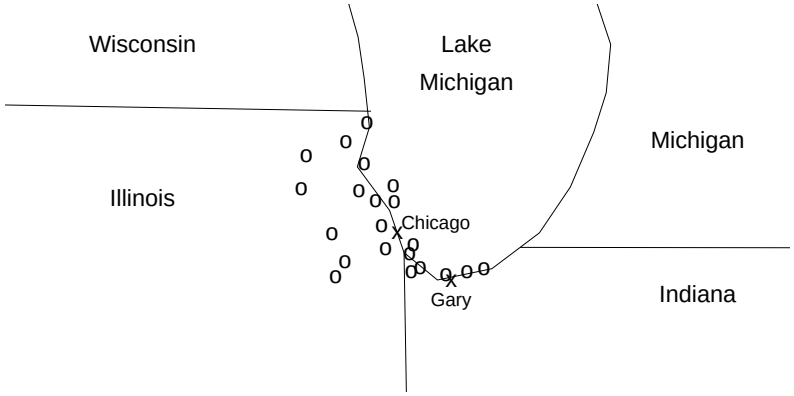
Key issues here:

(i). Number of RBF functions to include (AIC, CV,...)

(ii). Optimization of high-dimensional likelihood (multimodal)

4. There are also two recent Bayesian approaches due to Damian *et al.* (2001) and Schmidt and O'Hagan (2000). Both methods represent the G-space to D-space transformation as a stochastic process rather than a deterministic function, and use MCMC sampling ideas to fit the model.

Example (Smith 1996)



Ozone in Chicago. Estimated correlation matrix from 21 stations including 3 outlying stations

In this example we used the transformation $(x, y) \to (f^{(1)}(x, y), f^{(2)}(x, y))$ where

$$f^{(1)}(x, y) = b_1^2 x + \rho b_1 b_2 y + \sum_{i=1}^{m} \delta_i^{(1)} \eta_i(x, y),$$

$$f^{(2)}(x, y) = \rho b_1 b_2 x + b_2^2 y + \sum_{i=1}^{m} \delta_i^{(2)} \eta_i(x, y),$$

$$\eta_i(x, y) = r_i^2 \log r_i,$$

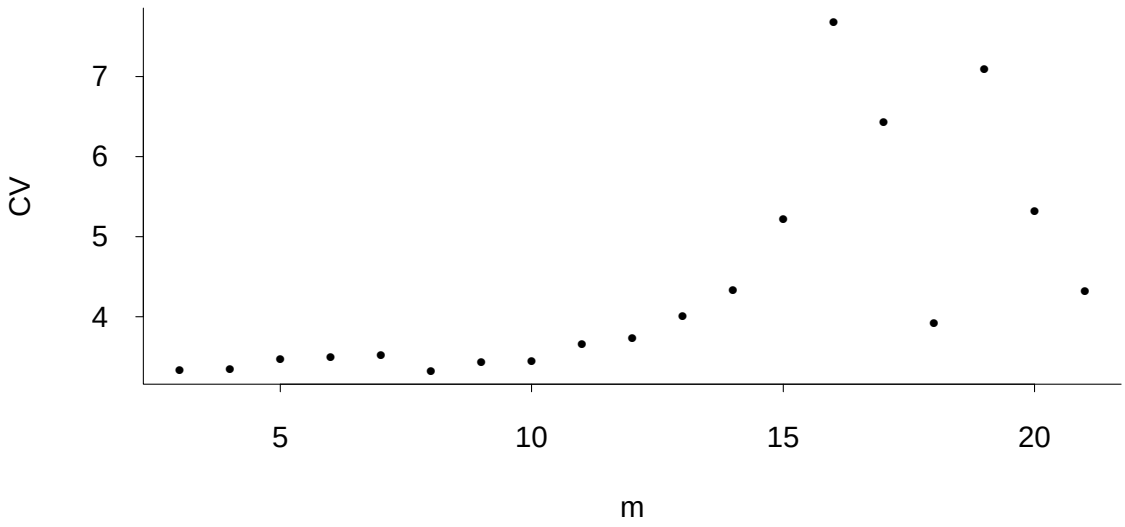
$$r_i = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

where $b_1 > 0$, $b_2 > 0, \sum \delta_i = \sum \delta_i x_i = \sum \delta_i y_i = 0$.

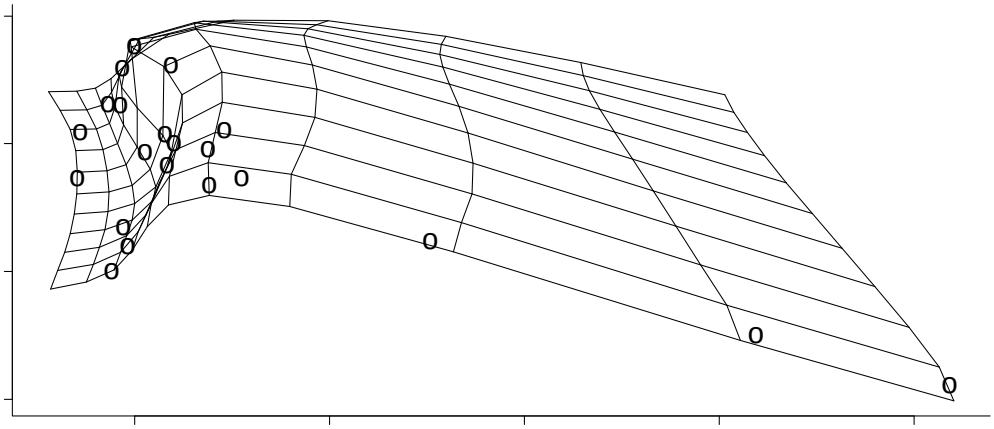
The centers (x_i, y_i) were the same as the actual stations, arranged in (some almost arbitrary) order, and the number of centers m was chosen by various criteria. The spatial model for the transformed data was assumed to be Matérn.

Log likelihoods and CV scores:

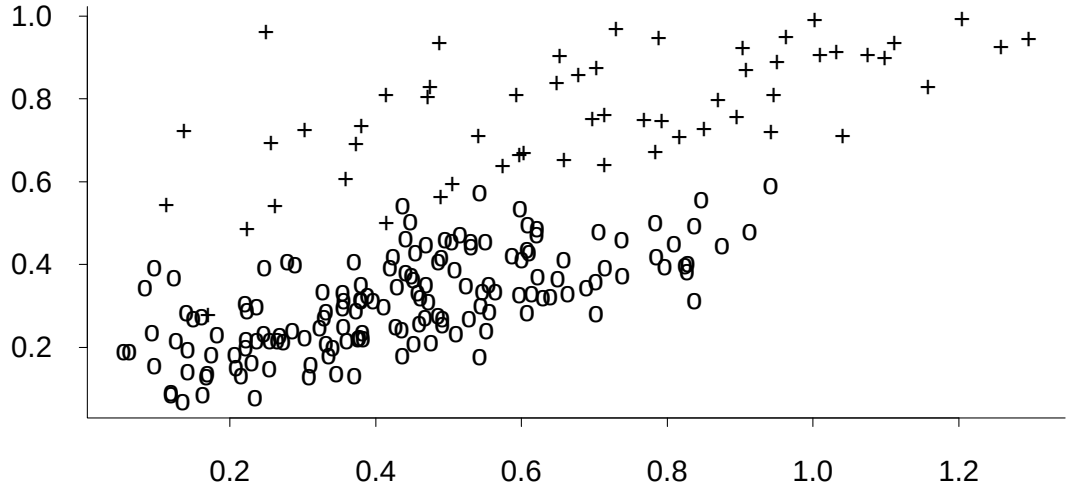
Number of centers m	LLH	CV
0	598.7	3.34
4	649.3	3.35
5	672.3	3.47
6	689.2	3.49
7	701.3	3.53
8	745.3	3.33
9	753.7	3.44
10	754.3	3.44
11	765.7	3.66
12	772.1	3.74
13	772.3	4.01
14	777.9	4.33
15	782.2	5.22
16	793.0	7.68
17	802.0	6.44
18	805.6	3.92



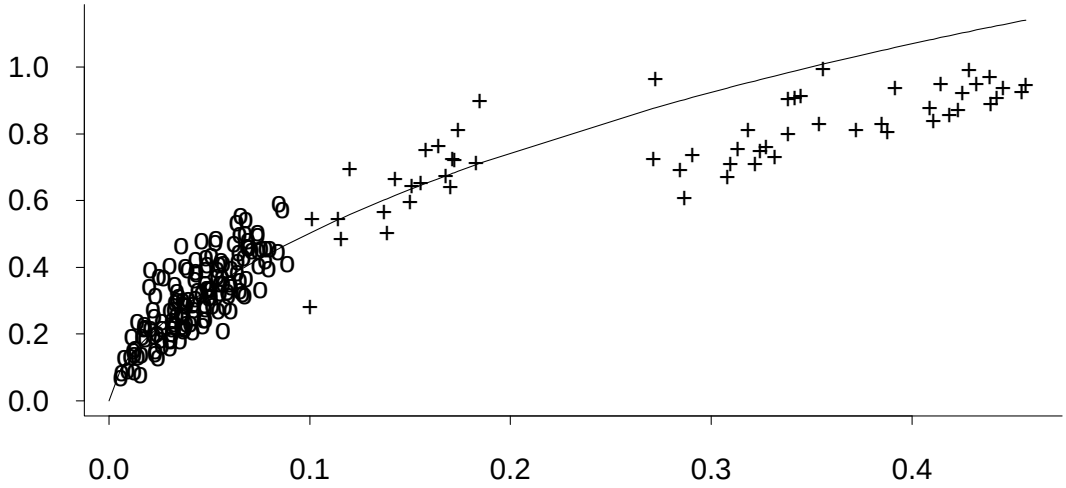
CV scores



D-space



Semivariogram plot in G-space.
+ means at least one station in the pair is one of the three outliers
o means neither station an outlier



Semivariogram plot in D space

14. Kernel-based models

Some refs:

Higdon (1998, 2001), Higdon *et al.* (1999), Barry *et al.* (1996), VerHoef and Barry (1998), VerHoef *et al.* (2000), Fuentes (2002), Fuentes and Smith (2001).

Basic idea:

$$z(s) = \int_{\mathcal{S}} K(u-s)w(u)du, \ s \in \mathcal{S},$$

where $K(\cdot)$ is a kernel and $w(\cdot)$ is white noise.

Covariance function:

$$\begin{aligned} C(h) &= \text{Cov}\{z(s), z(s-h)\} \\ &= \int K(u-h)K(u)du. \end{aligned}$$

Examples:

1. Gaussian case: if $K(u) \propto \exp\left(-\frac{1}{2}||u||^2\right)$ then $C(u) \propto \exp\left(-\frac{1}{4}||u||^2\right)$.

2. Suppose $C(u)$ is a d-dimensional Matérn covariance with parameters (ϕ, α, ν). Then $K(u)$ is also of Matérn structure with parameters $(\phi^{1/2}, \alpha, \frac{\nu}{2} - \frac{d}{4})$.

Here define Matérn (Fuentes, 2002) by

$$C_{\phi,\alpha,\nu}(u) = \frac{\pi^{d/2}\phi}{2^{\nu-1}\Gamma(\nu + d/2)\alpha^{2\nu}}(\alpha||u||)^{\nu}K_{\nu}(\alpha||u||)$$

with Fourier transform

$$\tilde{C}_{\phi,\alpha,\nu}(\omega) = \phi(\alpha^2 + ||\omega||^2)^{-\nu-d/2}.$$

3. Suppose $d = 2$ and $K(u) = \frac{2}{\pi}(1 - ||u||^2)I(||u|| < 1)$ (Epanechnikov kernel). If $||h|| = 2t$ for $0 \leq t < 1$, define

$$\begin{aligned} c_0 &= \frac{8}{15} - \frac{8t^2}{3}, \\ c_1 &= \frac{8t}{3}, \\ c_2 &= -\frac{16}{15} + \frac{8t^2}{3}, \\ c_3 &= -\frac{8t}{3}, \\ c_4 &= \frac{8}{15}. \end{aligned}$$

Also let

$$B_x(a, b) = \int_0^x t^{a-1}(1-t)^{b-1}dt.$$

Then

$$C(h) = \frac{16}{\pi^2} \sum_{k=0}^{4} c_k B_{1-t^2}\left(\frac{3}{2}, \frac{k}{2} + \frac{1}{2}\right).$$

Nonstationary extension:

$$z(s) = \int_{\mathcal{S}} K_s(u) w(u) du, \ s \in \mathcal{S}.$$

e.g. Higdon *et al.* (1999) derived an explicit formula for the covariance function of this process when

$$K_s(u) \propto \exp\left(-\frac{1}{2} u^T \Sigma(s)^{-1} u\right).$$

They used this in a hierarchical Bayesian context but there are a number of other possibiilities for building models from these foundations.

Higdon (2001) has considered discrete forms of the kernel model

$$z(s) = \sum_{j=1}^{m} w_j K(s - u_j),$$

with w_j i.i.d. $N(0, 1)$. When combined with an overall mean and random measurement error, this leads to the mixed models structure

$$y = \mu 1_n + Kw + \epsilon,$$

easily handled by REML techniques.

Extensions:

Multi-resolution models

Spatial-temporal processes

A specific model (Fuentes 2002):

$$z(s) = \int_{\mathcal{S}} K(s-u) z_{\theta(u)}(s) du, \qquad (\dagger)$$

where for each θ, $z_\theta(\cdot)$ is a stationary spatial process, but the nonstationarity comes from allowing $\theta(u)$ to vary with location u.

Two interpretations of (†):

1. For each θ write

$$z_{\theta(u)}(s) = \int e^{is^T x} \sqrt{f_{\theta(u)}(x)} w(x) dx,$$

in terms of a *common* white noise w. This process is of Higdon form with kernel

$$K_s(x) = e^{is^T x} \int K(s-u) \sqrt{f_{\theta(u)}(x)} du.$$

When $\theta(u)$ is a constant θ, $z(s)$ is just the process $z_\theta(s)$.

2. Let $z_{\theta(u)}(\cdot)$ be *independent* for each u. The covariance function of this process is

$$\begin{aligned}&\mathrm{Cov}\{z(s), z(s')\}\\ &= \int K(s-u)K(s'-u)C_{\theta(u)}(s-s')du.\end{aligned}$$

When $\theta(u)$ is constant, this is a stationary process, but not the stationary process with covariance C_θ.

In practice, often replace u by a discrete variable, the integral by a sum over a finite number of kernel-weighted covariance functions.

Application to Models-3 output on SO_2 concentrations and corresponding monitor data from CASTNet.

1. Discrete sum of $z_{\theta(u)}(\cdot)$ with kernel weights.

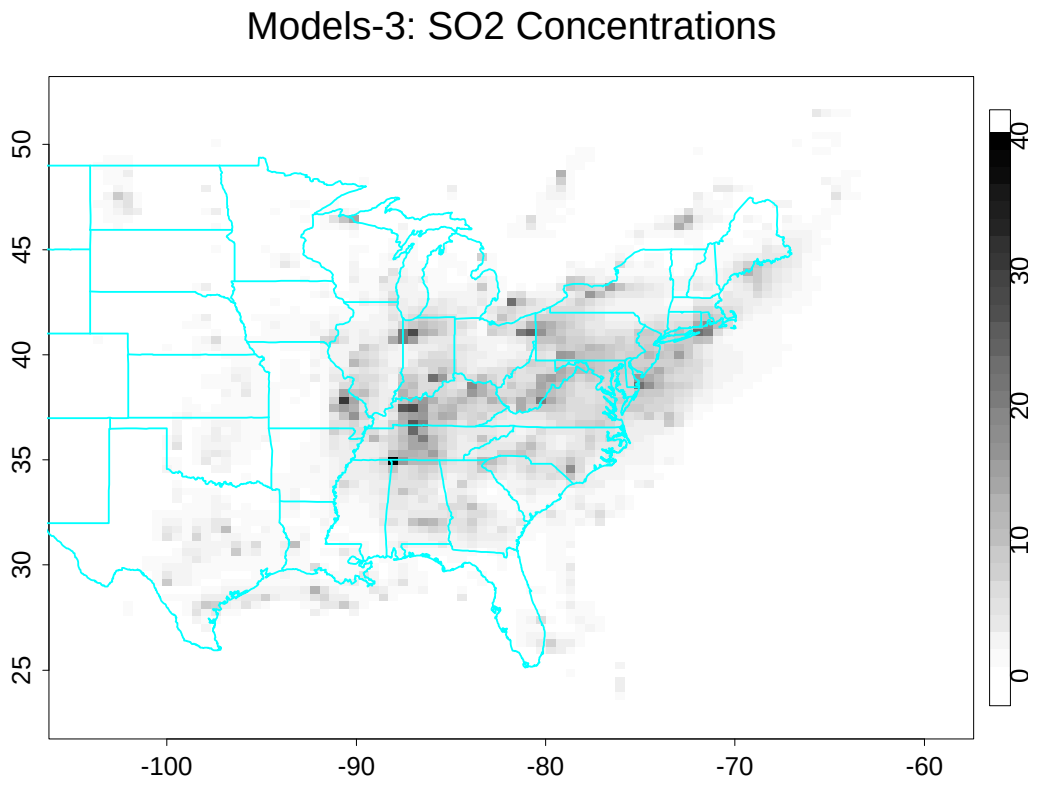
2. Each $z_{\theta(u)}$ a Matérn process, sill parameter varies across locations.

3. Hierarchical model for sill with latitude, longitude and random spatial components.

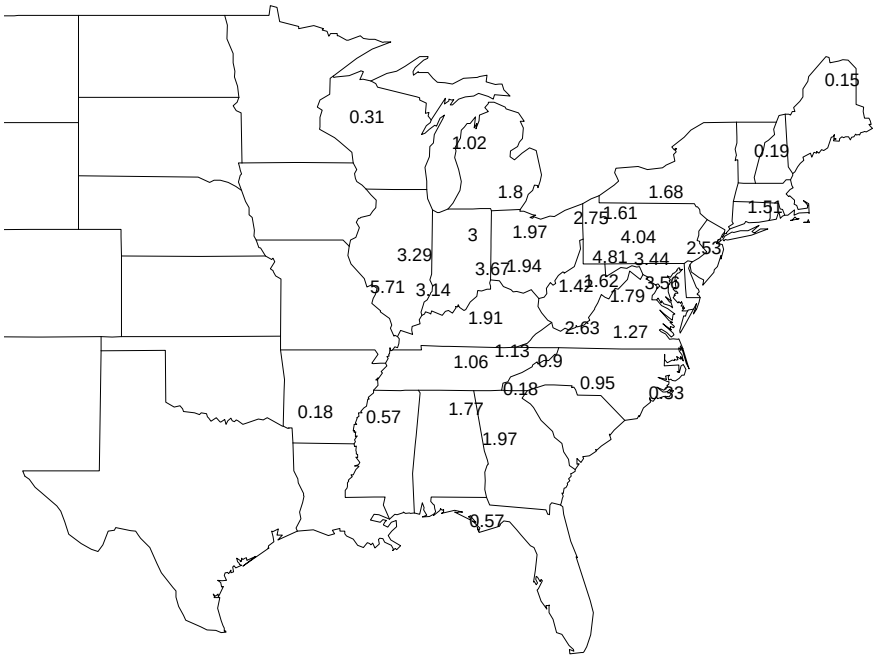
4. "Change of support": allow for discrepancy between grid cell and point data

5. Bayesian estimation and prediction

6. Results used to obtain predictive distributions for 6 sites — compare with monitor data at those sites.

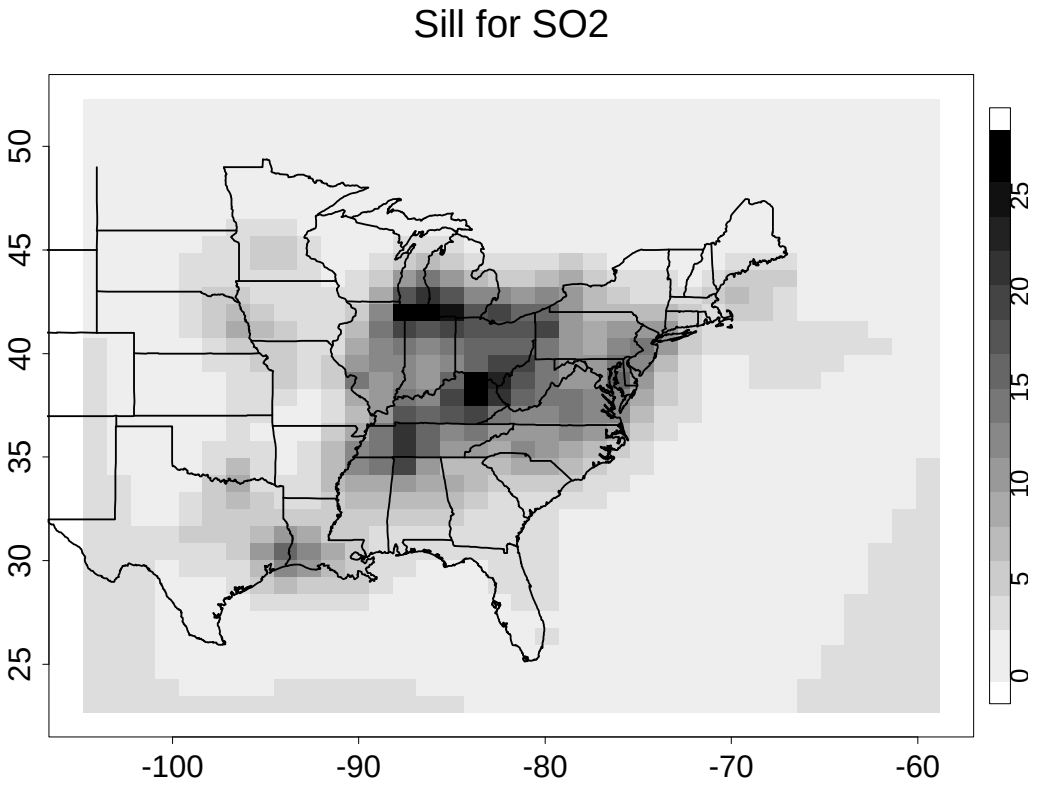


Models-3 output, mean SO_2 concentrations, week of July 11 1995

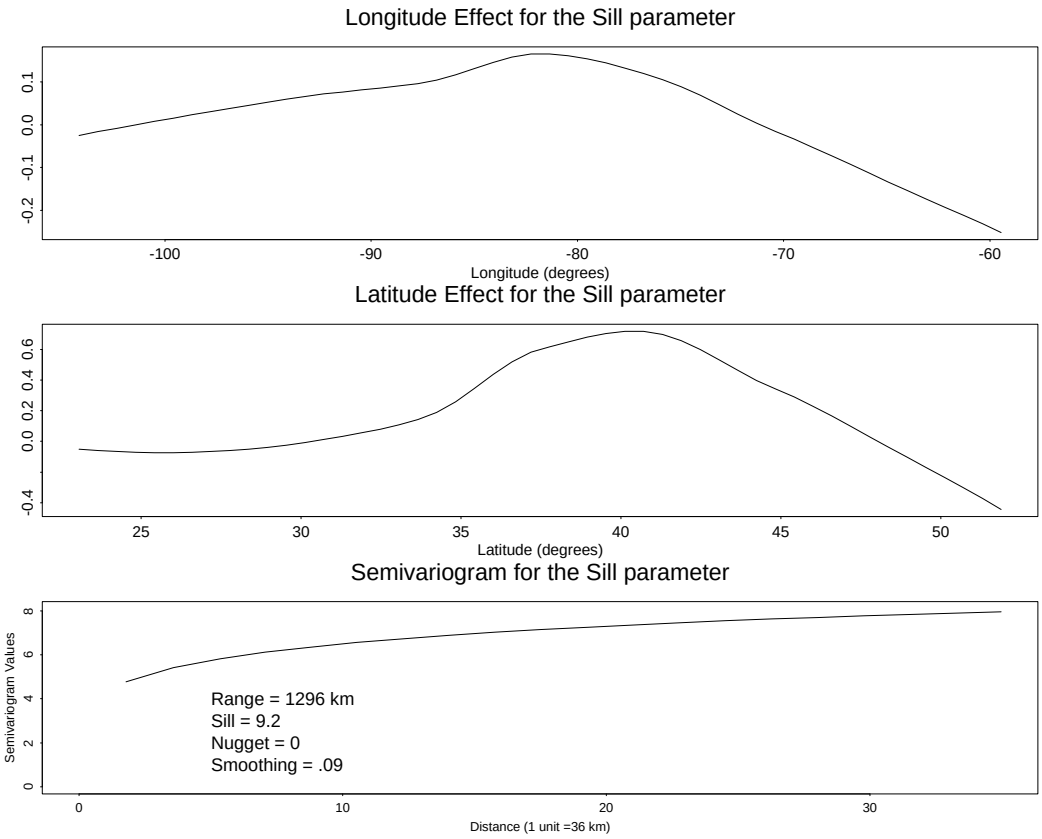


CASTNet data, mean SO_2 concentrations, week of July 11 1995

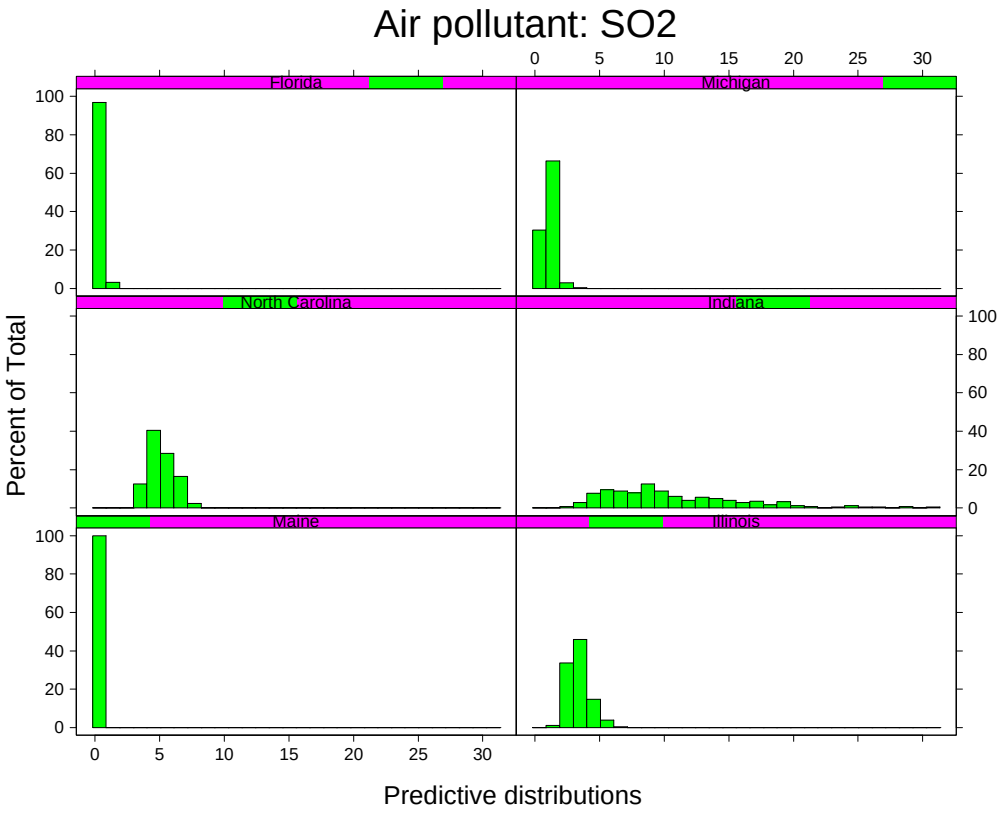
Mean SO_2 concentrations at selected sites, week of July 11 1995



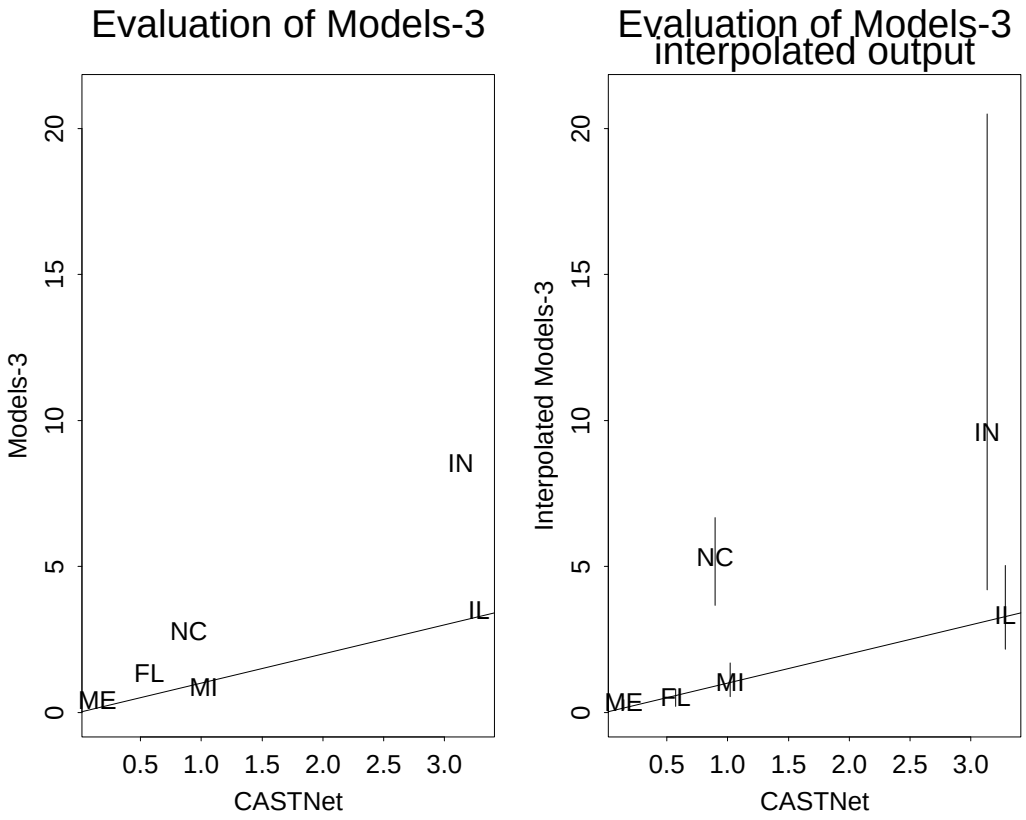
Modes of posterior distribution of Matérn sill parameter from Models-3 data



Spatial characteristics of Matérn still parameter



Predictive distributions of Models-3 concentrations at 6 selected sites

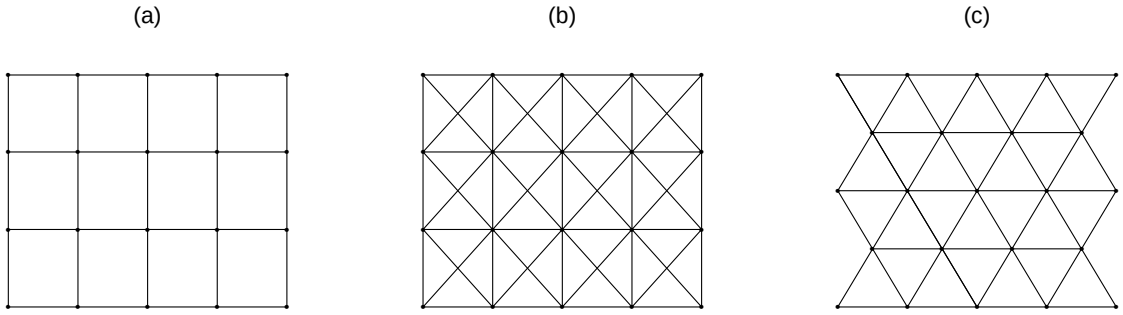


Comparisons of CASTNet predictions with Models-3 output: (L) crude comparison, (R) after allowing for change of support.

15. Markov random fields as spatial models

Principal reference: Besag (1974)

Consider data $\{X_i\}$ defined on a *lattice* with a *neighborhood structure*: N_i is set of neighbors of i (symmetric: $j \in N_i$ if and only if $i \in N_j$)



(a) Square lattice, first-order neighbor scheme.
(b) Square lattice, second-order neighbor scheme.
(c) Triangular lattice.

Could specify models in terms of joint probabilities $p(x_1, ..., x_n)$, or conditional probabilities of the form $p(x_i|x_j,\ j \neq i)$.

Example 1 (*auto-logistic* model): $X_1, ..., X_n$ are 0–1 random variables and

$$\begin{aligned}
&\Pr\{X_i = 1|X_j = x_j,\ j \neq i\} \\
&= \Pr\{X_i = 1|X_j = x_j,\ j \in N_i\} \\
&= \frac{\exp(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j)}{1 + \exp(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j)}.
\end{aligned}$$

Example 2 (*auto-normal*):

$$X_i|(X_j = x_j,\ j \neq i) \sim$$
$$N\left(\mu_i + \sum_{j \in N_i} \beta_{ij}(x_j - \mu_j), \sigma^2\right).$$

The auto-normal model should not be confused with

$$X_i = \mu_i + \sum_{j \in N_i} \beta_{ij}(X_j - \mu_j) + \epsilon_i,$$

$$\epsilon_i \text{ independent } N(0, \sigma^2),$$

known as the *simultaneous equation model*.

Question: How do we know these definitions are consistent, i.e. that there is a family of joint probabilities which generate the above conditional probabilities?

For auto-logistic, try

$$p(x_1, ..., x_n) \propto \exp\left(\sum_k \alpha_k x_k + \frac{1}{2} \sum_j \sum_{k \in N_j} \beta_{jk} x_j x_k \right).$$

Then

$$\frac{\Pr\{X_i = 1 | X_j = x_j,\ j \neq i\}}{\Pr\{X_i = 0 | X_j = x_j,\ j \neq i\}}$$

$$= \frac{p(x_1, ..., x_{i-1}, 1, x_{i+1}, ..., x_n)}{p(x_1, ..., x_{i-1}, 0, x_{i+1}, ..., x_n)}$$

$$= \exp\left\{ \alpha_i + \sum_{j \in N_i} \frac{\beta_{ij} + \beta_{ji}}{2} x_j \right\}.$$

If $\beta_{ij} = \beta_{ji}$, this is auto-logistic.

Similarly, for auto-normal case, if $\beta_{ij} = \beta_{ji}$ the joint density is

$$p(x_1, ..., x_n) \propto$$

$$\exp\left\{ -\frac{1}{2\sigma^2} \sum_{j,k} (x_j - \mu_j) b_{jk} (x_k - \mu_k) \right\}$$

where matrix B has entries $\{b_{jk}\}$ given by

$$b_{jk} = \begin{cases} 1 & \text{if } j = k, \\ -\beta_{jk} & \text{if } j \in N_k, \\ 0 & \text{otherwise.} \end{cases}$$

Result:

$$X \sim N[\mu, B^{-1}].$$

For simultaneous equation model, write

$$B(X - \mu) = \epsilon,$$

where $\epsilon \sim N[0, I]$. Then

$$X \sim N[\mu, (B^T B)^{-1}].$$

Do not require B symmetric in this case.

In general, it's not easy to check that a set of conditional probabilities is consistent with a family of joint probabilities, unless the latter can be explicitly constructed.

Brook's (1964) method: choose some reference state $(x_1^*, ..., x_n^*)$ and calculate

$$\begin{aligned}
&\frac{p(x_1, ..., x_n)}{p(x_1^*, ..., x_n^*)} \\
&= \prod_{i=0}^{n-1} \frac{p(x_1^*, ..., x_i^*, x_{i+1}, x_{i+2}, ..., x_n)}{p(x_1^*, ..., x_i^*, x_{i+1}^*, x_{i+2}, ..., x_n)} \\
&= \prod_{i=0}^{n-1} \frac{p(x_{i+1} | x_1^*, ..., x_i^*, x_{i+2}, ..., x_n)}{p(x_{i+1}^* | x_1^*, ..., x_i^*, x_{i+2}, ..., x_n)}.
\end{aligned}$$

This operation must be invariant to changes in the reference state and permutations of the indices.

The breakthrough in developing general classes of models came with the *Hammersley-Clifford Theorem* (1970). We state it in the form given by Besag (1974).

A Markov random field is characterized by the property that all conditional probabilities take the form

$$p(x_i|x_j,\ j \neq i) = p(x_i|x_j,\ j \in N_i)$$

where N_i denotes the set of neighbors of i under some lattice structure.

Define a *clique* to be any subset of sites with the property that each member of the clique is a neighbor of each other member. Assume (initially) that there are only finitely many values x_i available at each site, and that one of these is (arbitrarily) labelled 0.

Positivity: $p(\mathbf{x}) > 0$ for any state $\mathbf{x} = (x_1, ..., x_n)$.

Define

$$q(\mathbf{x}) = \log \left\{ \frac{p(\mathbf{x})}{p(\mathbf{0})} \right\}.$$

Then there exist functions $g_i(x_i)$, $g_{ij}(x_i, x_j)$, etc., such that

$$\begin{aligned} q(\mathbf{x}) = & \sum_i x_i g_i(x_i) + \sum_{i<j} x_i x_j g_{ij}(x_i, x_j) \\ & + \sum_{i<j<k} x_i x_j x_k g_{ijk}(x_i, x_j, x_k) \\ & + ... + x_1 x_2 ... x_n g_{12...n}(x_1, x_2, ..., x_n). \end{aligned}$$

The Hammersley-Clifford theorem is then:

For a Markov random field, $g_{ij...s}(x_i, x_j, ..., x_s)$ is non-zero if and only if $\{i, j, ..., s\}$ form a clique. Subject to this restriction, the g's are arbitrary.

Specific Spatial Models

General class of *auto-models*, for which

$$q(\mathbf{x}) = \sum_i x_i g_i(x) + \sum_{i<j} \beta_{ij} x_i x_j$$

and $\beta_{ij} = 0$ unless i and j are neighbors. Associated conditional probabilities are of form

$$\frac{\Pr\{X_i = x_i | X_j = x_j,\ j \neq i)}{\Pr\{X_i = 0 | X_j = x_j,\ j \neq i)}$$
$$= \exp\left[x_i \left\{ g_i(x_i) + \sum_j \beta_{ij} x_j \right\} \right],$$

in which $\beta_{ij} = 0$ unless i and j are neighbors, and also $\beta_{ij} = \beta_{ji}$ for all i, j.

Examples: auto-logistic, auto-normal, auto-Poisson, auto-exponential,...

e.g. auto-Poisson says $X_i|X_j = x_j,\ j \in N_i$ has a Poisson distribution with mean

$$\mu_i = \exp\left(\alpha_i + \sum_{j \in N_i} \beta_{ij} x_j\right).$$

However, for this to define a finite probability distribution we need $\beta_{ij} \leq 0$ for each (i, j) pair. This makes it an unsuitable model for count data with positive dependence. Fortunately, there are by now numerous alternatives available for that.

Models for square lattices

First-order model:

$$q(\mathbf{x}) = \alpha \sum x_{i,j} + \beta_1 \sum x_{i,j}x_{i+1,j} + \beta_2 \sum x_{i,j}x_{i,j+1}.$$

Second-order model:

$$q(\mathbf{x}) = \alpha \sum x_{i,j} + \beta_1 \sum x_{i,j}x_{i+1,j} + \beta_2 \sum x_{i,j}x_{i,j+1} + \gamma_1 \sum x_{i,j}x_{i+1,j+1} + \gamma_2 \sum x_{i,j}x_{i+1,j-1}$$

but this is not the most general form of model, since in this case there are cliques of three neighbors so one could include terms of the form $x_{i,j}x_{i-1,j}x_{i,j-1}$ etc.

For Gaussian models, one can re-express these models in terms of the conditional mean of a random variable given its neighbors, e.g.

$$\begin{aligned}\mathrm{E}\{X_{i,j}|X_{i',j'},\ (i',j') \neq (i,j)\} = \alpha \\ + \beta_1(X_{i-1,j} + X_{i+1,j}) \\ + \beta_2(X_{i,j-1} + X_{i,j+1})\end{aligned}$$

or

$$\begin{aligned}\mathrm{E}\{X_{i,j}|X_{i',j'},\ (i',j') \neq (i,j)\} = \alpha \\ + \beta_1(X_{i-1,j} + X_{i+1,j}) \\ + \beta_2(X_{i,j-1} + X_{i,j+1}) \\ + \gamma_1(X_{i-1,j-1} + X_{i+1,j+1}) \\ + \gamma_2(X_{i-1,j+1} + X_{i+1,j-1}).\end{aligned}$$

16. Inference in Markov random fields

Coding methods (Besag 1974)

Example: in a first-order lattice, the set of points (i, j) for which $i - j$ is odd is independent of the set of points for which it is even, so we can write down a conditional likelihood for the even points

$$\prod_{(i,j):\ i-j \text{ even}} p(x_{i,j} | x_{i',j'},\ (i', j') \neq (i, j))$$

and similarly for the odd points.

Pseudolikelihood

Besag (1975): combine odd and even coding likelihood to produce

$$\prod_{\text{all pairs } (i,j)} p(x_{i,j} | x_{i',j'},\ (i', j') \neq (i, j)).$$

Easy to compute but large-sample properties uncertain (much modern literature on this)

Exact and approximate MLEs for Gaussian processes

Difficulty in exact likelihood is $|B|$.

Whittle's approximation, modified by Besag: $\frac{1}{n}\log|B|$ is approximated by the coefficient of $z_1^0 z_2^0$ in the power series expansion

$$\log\left(1 - \sum_{j,k} \beta_{jk} z_1^j z_2^k\right).$$

Improvements: Guyon (1982), Dahlhaus and Künsch (1987)

Simulated maximum likelihood
(Penttinen 1984, Geyer-Thompson 1992)

Consider models of form

$$p(\mathbf{x};\theta) = C(\theta)F(\mathbf{x};\theta)$$

where F is a known function of data values $\mathbf{x}$ in terms of unknown parameters θ, and $C(\theta)$ is a normalizing constant defined by the property that the sum or integral of p is 1, but not directly computable.

Idea: fix some reference value θ_0, estimate ratios $C(\theta)/C(\theta_0)$ by simulation.

Fix θ_0 and let $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ denote M simulated realizations from the stochastic process when θ_0 is the true parameter. For the moment, we do not consider how such simulations might be generated. Also let $\mathbf{X}$ denote the actual data which are observed.

We then have that

$$\frac{1}{M}\sum_{m=1}^{M}\frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)}\cdot\frac{F(\mathbf{X};\theta_0)}{F(\mathbf{X};\theta)}$$

is an unbiased estimate of

$$\frac{C(\theta_0)}{C(\theta)}\cdot\frac{F(\mathbf{X};\theta_0)}{F(\mathbf{X};\theta)},$$

i.e., the likelihood ratio of θ_0 to θ.

To see this, the key step is the calculation

$$\begin{aligned}
&\mathrm{E}_{\theta_0}\left\{\frac{F(\mathbf{X}^{(m)};\theta)}{F(\mathbf{X}^{(m)};\theta_0)}\right\}\\
&=\sum_{\mathbf{x}}\frac{F(\mathbf{x};\theta)}{F(\mathbf{x};\theta_0)}\cdot C(\theta_0)F(\mathbf{x};\theta_0)\\
&=C(\theta_0)\sum_{\mathbf{x}}F(\mathbf{x};\theta)\\
&=\frac{C(\theta_0)}{C(\theta)}.
\end{aligned}$$

MCMLE scheme: generate a single sequence $\mathbf{X}^{(1)}, ..., \mathbf{X}^{(M)}$ from some given θ_0, then to minimize the sum (4.25) analytically with respect to θ. The simulation procedure is more efficient the closer θ_0 is to the true MLE $\widehat{\theta}$, so sometimes the procedure is repeated several times, using the estimate from one minimization at the initial θ_0 for the next. One good use of the pseudolikelihood method is to generate the initial θ_0.

Generation of initial simulations:

- Gibbs sampling
- Hastings-Metropolis
- Swendsen-Wang
- Perfect sampling

Only cover first two here.

Gibbs sampling. Start with arbitrary $\mathbf{x}^{(0)} = (x_1^{(0)}, ..., x_n^{(0)})$. Generate a new value of x_1, denoted $x_1^{(1)}$, from the conditional distribution of X_1 given $X_2 = x_2^{(0)}, ..., X_n = x_n^{(0)}$. Then generate a new value of x_2, denoted $x_2^{(1)}$, from the conditional distribution of X_2 given $X_1 = x_1^{(1)}, X_3 = x_3^{(0)}, ..., X_n = x_n^{(0)}$. Continue up to the generation of $x_n^{(1)}$ from the conditional distribution of X_n given $X_1 = x_1^{(1)}, ..., X_{n-1} = x_{n-1}^{(1)}$. This completes one ieration of the sampler. Then, starting from the new vector $\mathbf{x}^{(1)}$, return to x_1 and repeat the whole process to generate $\mathbf{x}^{(2)}$. Repeat many times.

The Hastings-Metropolis algorithm. Again we start with an arbitrary $\mathbf{x}^{(0)}$ and generate a new "trial value" $\mathbf{x}'$ from some distribution $q(\mathbf{x}'; \mathbf{x}^{(0)})$ which depends on $\mathbf{x}^{(0)}$. Typically, but not necessarily, $\mathbf{x}'$ is formed from $\mathbf{x}^{(0)}$ by just changing one component. Then form the ratio

$$\alpha = \frac{q(\mathbf{x}^{(0)}; \mathbf{x}')F(\mathbf{x}'; \theta_0)}{q(\mathbf{x}'; \mathbf{x}^{(0)})F(\mathbf{x}^{(0)}; \theta_0)}.$$

If $\alpha \geq 1$ then we accept $\mathbf{x}'$; in other words, set $\mathbf{x}^{(1)} = \mathbf{x}'$. If $\alpha < 1$, we perform an independent random drawing: with probability α, accept $\mathbf{x}'$ and set $\mathbf{x}^{(1)} = \mathbf{x}'$; otherwise, reject $\mathbf{x}'$ and set $\mathbf{x}^{(1)} = \mathbf{x}^{(0)}$.

17. Examples of Markov random fields

1. Mercer-Hall (1911) data (see also Whittle 1954, Besag 1974, Cressie 1993). Data show wheat yields on 500 plots arranged in a 20×25 array.

2. An example from current particulate matter research

Models fitted within S-PLUS

$$y \sim N[X\beta, S].$$

N a given neighborhood matrix, W a given diagonal matrix of weights.

CAR:

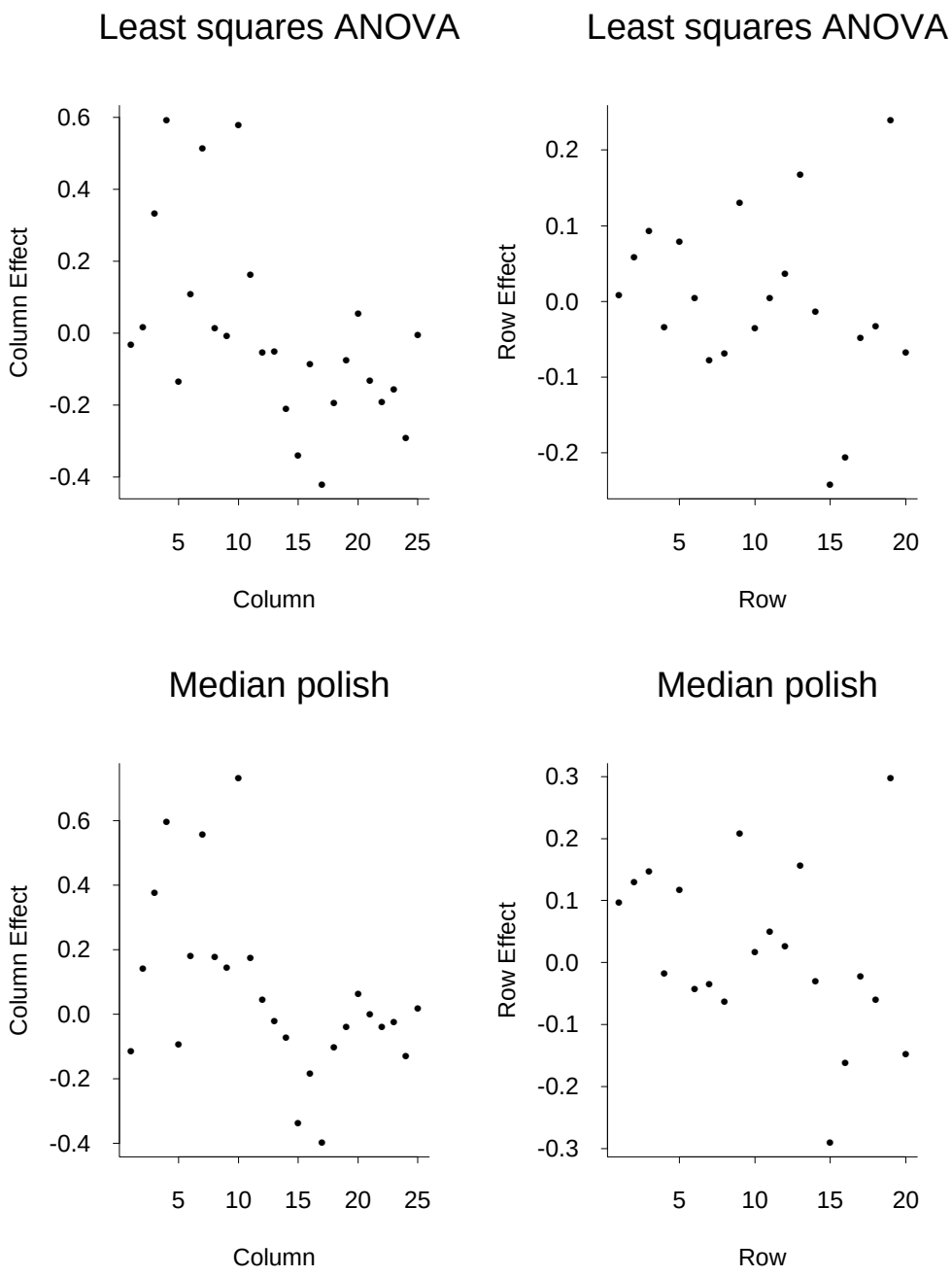
$$S = (I - \rho N)^{-1} W \sigma^2.$$

SAR:

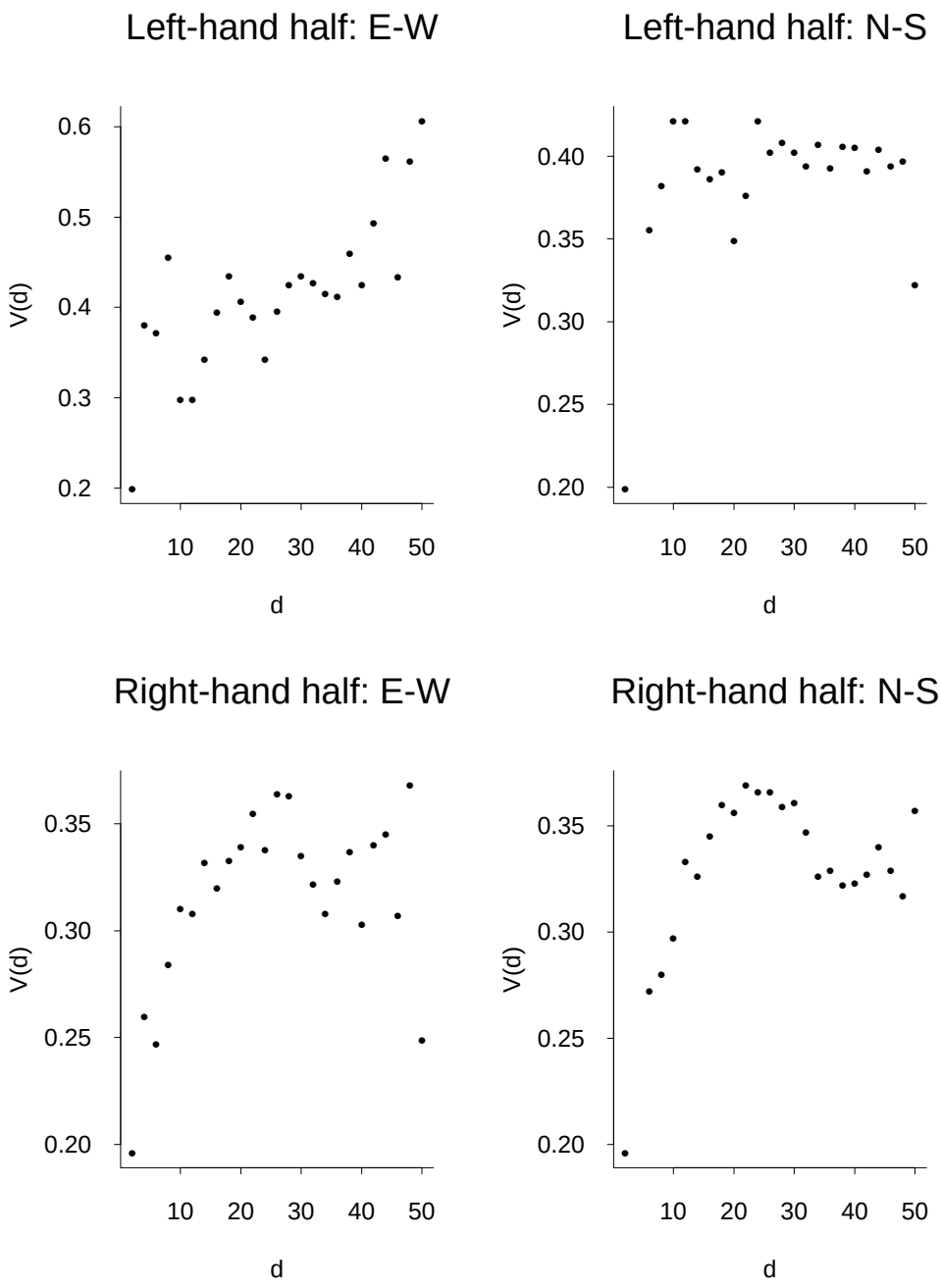
$$S = \{(I - \rho N)^T W^{-1} (I - \rho N)\}^{-1} \sigma^2.$$

MA:

$$S = \{(I + \rho N) W (I - \rho N)^T\} \sigma^2.$$



Row and columns effects for Mercer-Hall data, computed by least-squares ANOVA (top plots) and by median polish (bottom plots)



Variogram estimates for Mercer-Hall data, computed separately for E-W and N-S directions, and for the left-hand and right-hand halves of the data.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$
Coding 1	0.332	0.128
Coding 2	0.354	0.166
(S.E.	0.03	0.03)
Whittle	0.368	0.107
MLE	0.364	0.114
(S.E.	0.024	0.025)
Left half:		
MLE	0.400	0.000
(S.E.	0.029	0.033)
Right half:		
MLE	0.275	0.191
(S.E.	0.041	0.043)

Estimates first-order model using coding and Whittle methods (from Besag, 1974) and by exact MLE.

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
Code 1	0.344	0.043	0.079	–0.062
Code 2	0.318	0.085	0.016	0.011
Code 3	0.407	0.243	–0.067	–0.034
Code 4	0.361	0.236	–0.092	–0.041
(S.E.	0.05	0.06	0.07	0.06)
Whittle	0.381	0.160	–0.015	–0.056
MLE	0.380	0.171	–0.020	–0.060
(S.E.	0.024	0.046	0.037	0.035)
Left half:				
MLE	0.400	0.019	0.000	–0.025
(S.E.	0.029	0.075	0.055	0.055)
Right half:				
MLE	0.329	0.274	–0.042	–0.116
(S.E.	0.042	0.053	0.051	0.049)

Estimates for seond-order model in Mercer-Hall data.

Use of SAR modeling by the HEI Re-Analysis Project (Krewski *et al.* 2000):

151 cities — compute a standardized mortality rate for each city

Regress on city-wide covariates, including SO_4 and SO_2.

Spatial dependence: enclose each station within a Thiessen polygons, then define two stations to be "neighbors" if their Thiessen polygons are touching. Assume SAR structure.

Results

"Independent cities model" (allows random effects at city level, but not spatial dependence): RR due to SO_4 on its own is 1.25 (95% CI: 1.13–1.37) but when SO_4 and SO_2 are modeled together, the respective effects due to SO_4 and SO_2 are 1.13 (1.02–1.25) and 1.27 (1.15–1.40).

After fitting SAR model:

RR due to SO_4 alone is 1.20 (1.06–1.36)

For model including both SO_4 and SO_2, relative risks are 1.08 (0.91–1.28) and 1.31 (1.12–1.54)

18. Markov random fields as spatial priors

1. The Clayton-Kaldor model

2. The Cressie-Chan model

3. Besag, York and Mollié (1991) and other papers by Julian Besag and co-authors

4. Hierarchical models for disease rates (Bernardinelli and co-authors, Waller *et al.* 1997)

5. Model-based geostatistics (Diggle, Tawn and Moyeed 1998)

6. Comparisons: geostatistics vs. MRF

Clayton-Kaldor (1987): Number of cases S_i in a susceptible population of size n_i.

$\{\theta_i,\ 1 \le i \le N\}$ independent $Gam(\alpha, \nu)$,

$$f(\theta_i) \propto \theta_i^{\nu-1} \exp(-\alpha\theta_i).$$

Conditionally on θ_i,

$$S_i \sim Poi(n_i\theta_i)$$

Integrated distribution is *beta-binomial*:

$$\Pr\{S_i = k\} = \left(\frac{\alpha}{\alpha + E}\right)^{\nu} \left(\frac{n_i}{\alpha + n_i}\right)^{k} \frac{\Gamma(\nu + k)}{\Gamma(\nu)k!}.$$

Mean $n_i\nu/\alpha$, variance $n_i\nu/\alpha + n_i^2\nu/\alpha^2$.

$$\mathrm{E}\{\theta_i|S_i\} = \frac{\nu + O_i}{\alpha + n_i},$$

"Empirical Bayes" estimate for θ_i.

Estimation of α and ν: MLE or MoM.

Clayton-Kaldor model 2: Log-normal

$$S_i|\theta_i \sim Poi(n_i\theta_i),$$
$$\beta_i = \log\theta_i$$
$$(\beta_1, ..., \beta_n) \sim N(\mu, \Sigma).$$

No closed-form expression for joint distribution of S_i, but use EM algorithm (or Gibbs sampler...)

$$\mathrm{E}\{\beta_i|\beta_j,\ j \neq i\} =$$
$$\mu_i + \rho\sum_j w_{ij}(\beta_j - \mu_j),$$
$$\mathrm{Var}\{\beta_i|\beta_j,\ j \neq i\} = \sigma^2.$$

e.g. $w_{ij} = 1$ if $j \in N_i$, 0 otherwise.

$$\beta \sim N[\mu, \sigma^2(1-\rho W)^{-1}].$$

Cressie-Chan (1989)

S_i: the number of SIDS cases in county i
n_i: number of live births in county i
Freeman-Tukey transformation:

$$Y_i = \left(\frac{1000S_i}{n_i}\right)^{1/2} + \left(\frac{1000(S_i+1)}{n_i}\right)^{1/2},$$

$$\begin{aligned}
&\mathrm{E}\{Y_i|Y_j,\ j \in N_i\} \\
&= \mu_i + \sum_{j\in N_i} c_{ij}(Y_j - \mu_j), \\
&\mathrm{Var}\{Z_i|Z_j,\ j \in N_i\} = \tau_i^2,
\end{aligned}$$

equivalent to

$$Z \sim N[\mu, (I-C)^{-1}M],$$

with $M = \mathrm{diag}(\tau_1^2, ..., \tau_n^2)$.

Assume $c_{ij} = 0$ if $j \notin N_i$, and $c_{ij}\tau_j^2 = c_{ji}\tau_i^2$.

Here $\tau_i^2 \propto n_i^{-1}$ and a possible model for c_{ij} is

$$C_{ij} = \phi \cdot \{C(k) d_{ij}^{-k}\} (n_j/n_i)^{1/2},$$

where d_{ij} is distance between stations i and j, k is 0, 1 or 2, and $C(k) = i\{\min(d_{ij})\}^k$.

Besag, York and Mollié (1991)

(See also Besag *et al.* 1995, Besag and Higdon 1998)

Observations $y = \{y_k,\ k \in T\}$, states $x = \{x_i,\ i \in S\}$. Often, $S = T$ and

$$f(y|x) = \prod_k f_k(y_k|x_k).$$

If x is only unknown, inference based on the posterior density

$$p(x|y) \propto p(x)f(y|x).$$

"MAP estimate": choose x to maximize.

Further unknown parameter: distribution of x (and $y|x$) depends on other unknown parameters.

E.g. disease mapping problem,

$$\begin{aligned} y_i|x_i &\sim Poi[n_i e^{x_i}], \\ x &= t + u + v, \\ t &= X\beta, \\ u &\sim \text{ spatial process}, \\ v &\sim N[0, \lambda I_n]. \end{aligned}$$

Prior $p(u)$:

$$p(u) \propto \exp\left\{ -\sum_{i<j} w_{ij}\phi(u_i - u_j) \right\},$$

("pairwise difference prior"). Specific cases may be $\phi(z) = z^2/(2\kappa)$ or $\phi(z) = |z|/\kappa$. First case:

$$u_i|u_{-i} \sim N\left[\frac{\sum_{j\in N_i} w_{ij}u_j}{w_{i+}}, \frac{1}{\kappa^2 w_{i+}^2}. \right]$$

Posterior distribution of (u, v, κ, λ) given data y is of form

$$
\begin{aligned}
& p(u, v, \kappa, \lambda | y) \\
& \quad \propto \prod_{i=1}^{n} \frac{e^{-n_i e^{x_i}} (n_i e^{x_i})^{y_i}}{y_i!} \cdot \\
& \qquad \kappa^{-n/2} \exp\left\{ -\frac{1}{2\kappa} \sum_{i \sim j} (u_i - u_j)^2 \right\} \\
& \qquad \lambda^{-n/2} \exp\left\{ -\frac{1}{2\lambda} \sum v_i^2 \right\} \pi(\lambda, \kappa).
\end{aligned}
$$

where $\pi(\lambda, \kappa)$ denotes prior density.

Difficulties with integrability near $\lambda = 0$, $\kappa = 0$, so assume

$$
\begin{aligned}
& \pi(\lambda, \kappa) \propto \\
& \exp\left(-\frac{\epsilon}{2\kappa} - \frac{\epsilon}{2\lambda} \right).
\end{aligned}
$$

Hierarchical spatio-temporal models (Waller *et al.* 1997)

Environmental justice assessment involves

(1) exposure assessment at given locations,

(2) estimation of sociodemographic variables

(3) measuring disease incidence.

Focus on relation between (2) and (3).

$y_{i\ell t}$: disease incidence in region i, subgroup ℓ, time period t,

$$y_{i\ell t} \sim Poi[E_{i\ell t}\psi_{i\ell t}],$$

where $E_{i\ell t}$ is expected incidence under constant risk and $\psi_{i\ell t} = \exp(\mu_{i\ell t})$ is relative risk.

$$\mu_{i\ell t} = x_\ell^T \beta + z_i^T \omega + \theta_i^{(t)} + \phi_i^{(t)}$$

where first two terms represent fixed effects of time and region, $\theta_i^{(t)} \sim N[\kappa_\theta^{(t)}, 1/\tau^{(t)}]$ (i.i.d.),

$$\phi_i^{(t)} | \phi_{-i}^{(t)} \sim$$
$$N\left[\frac{1}{a_i}\sum w_{ij}\phi_j^{(t)}, \frac{1}{\lambda_t a_i^2}\right].$$

Here w_{ij} and a_i are defined by a lattice structure.

A number of similar models have been considered by Bernardinelli and co-authors (see references)

Heirarchical Bayesian structure — many issues associated with identifiability

MCMC algorithm for fitting

Use empirical predictive criteria for model selection, e.g.

$$\tilde{d}(y_{\text{new}}, y_{\text{obs}})$$
$$= 2\sum_{\ell}\left\{\left(y_{\ell,\text{obs}} + \frac{1}{2}\right)\log\left(\frac{y_{\ell,\text{obs}} + \frac{1}{2}}{y_{\ell,\text{new}} + \frac{1}{2}}\right)\right.$$
$$\left. - (y_{\ell,\text{obs}} - y_{\ell,\text{new}})\right\}.$$

Application. Follow lung cancer deaths for 21 years in 88 counties in Ohio. Classify by race (B/W) and gender. Particular interest in effect of a uranium recycling facility near Cincinnati.

Conclusions:

Increase in lung cancer rates over 21 year period

Increased clustering and increased heterogeneity over time

Increasing rates from west to east (effect of smoking?)

Increased rates near uranium facility

Model-based geostatistics (Diggle, Tawn and Moyeed 1998)

The idea: extend kriging to non-Gaussian processes

Example 1: Radionuclide concentration on Rongelap Island

Y_i: γ-ray count at local $i = 1, ..., 157$

$$Y_i \sim Poi[t_i \lambda(x_i)],$$

where t_i is observing time at location i and

$$\log \lambda(x) = \mu + S(x),$$

with $S(x)$ a zero-mean stationary Gaussian process.

Example 2: Campylobacter infections — record all locations (by postcode) of

incidences of a disease

$$Y_i \sim Bin[n_i, P(x_i)],$$
$$\log \frac{P(x)}{1 - P(x)} = \mu + S(x).$$

More generally: replace μ by $d(x)^T \beta$ to allow regression component at each site.

General structure

$S(x)$ a Gaussian spatial process of mean 0, covariance structure $\sigma^2 \rho(x - x')$,

Y_i depends on $S(x_i)$ through a mean M_i,

$h(M_i) = S(x_i) + d_i^T \beta$ for known link function h, known regressors d_i, unknown parameters β.

Observe Y_i at n current locations; may also want to predict S at m new locations, denoted S^*.

Joint density of Y is

$$\int \prod_{i=1}^{n} f_i(y_i|S(x_i))g_n(s)ds,$$

Joint density of Y and S^* is

$$\int \prod_{i=1}^{n} f_i(y_i|S(x_i))g_{m+n}(s, s^*)ds,$$

Predictive density of S^* given Y is ratio of last two quantities.

MCMC implementation: θ are parameters of Sm successively update (i) θ given S, (ii) S given Y, θ, β, (iii) β given Y, S, (iv) (once chain has reached equilibrium) S^* given S, θ. Hence estimate posterior distribution of parameters and Bayesian predictive distribution of S^*.

Comparison of geostatistics and MRF approaches to prediction of a PM_{10} field
(Cressie *et al.* 1999)

Pittsburgh area: 40 stations for PM_{10}. Use 1996 data, 27 stations. Focus on one particular day for prediction

Log transformation for closer fit to normal distribution. One station is a spatial outlier — drop.

Geostatistical modeling: Define principal axis directions (ENE–WSW, NNW–SSE), consider geometrically anisotropic spherical covariance structure. Fit by Cressie WLS method.

MRF modeling: Assume

$$\begin{aligned} &\log x_i | x_{-i} \sim N[A_i, \tau_i^2], \\ &A_i = \mu_i + \sum c_{ij}\{\log x_j - \mu_j\}, \qquad (*) \\ &c_{ij}\tau_j^2 = c_{ji}\tau_i^2. \end{aligned}$$

Assume τ_i constant,

$$c_{ij} = \frac{\eta \min\{d_{ij}\}}{d_{ij}},$$

where d_{ij} is distance and η is one of $\eta_1, ..., \eta_4$ according to direction sector.

Fit MLEs μ, τ^2, $\eta_1, ..., \eta_4$, use (*) to generate predictions off the grid.

Comparisons

Prediction error results by cross-validation:

	Bias	Var	MSE	r
G_0	-.54	27.4	27.7	.47
M_0	-.02	15.8	15.8	.71
M_1	.66	27.1	27.5	.54

G_0: geostat model without re-estimated parameters
M_0: MRF model without re-estimated parameters
M_1: MRF model with re-estimated parameters

19. Maximum entropy and Bayesian approaches to network design

19.1 Motivation for entropy criteria

19.2 Bayesian background

19.3 Bayesian spatial analysis

19.4 Hierarchical models

19.5 Discussion and extensions

19.6 Entropy-based criteria for design

19.7 A proposal for fully hierarchical models

19.1 Motivation for entropy criteria

Lindley (1956), Bernardo (1979)

Consider experiment E which will yield data X,

$$X \sim p(\cdot \mid \psi).$$

Measure of information contained in E:

$$I\{E, \pi\} = \int p_X(x) \int \pi(\psi \mid x) \cdot \\ \cdot \log \frac{\pi(\psi \mid x)}{\pi(\psi)} d\psi dx.$$

Decision-theoretic formulation

According to the Bayesian viewpoint, Ψ is a random variable, and the outcome of the experiment E may be represented by the statistician's reporting a probability distribution, $\pi^\dagger(\psi)$, to represent her "belief" about Ψ after conducting the experiment.

Utility function, $u(\pi^\dagger(\cdot), \psi)$ represents gain in reporting $\pi^\dagger$ when true value of Ψ is ψ.

Expected utility after experiment is

$$\int u(\pi^\dagger(\cdot), \psi)\pi(\psi \mid x)d\psi. \qquad (*)$$

Assume:

(a) u is *proper* if (*) is maximized over all probability distributions $\pi^\dagger$ by setting $\pi^\dagger(\psi) = \pi(\psi \mid x)$.

(b) u is *local* if $u(\pi^\dagger(\cdot), \psi)$ depends on $\pi^\dagger(\cdot)$ only through $\pi^\dagger(\psi)$.

Bernardo's theorem: *If u is proper and local, then it must be of the form*

$$u(\pi^\dagger(\cdot), \psi) = A \log \pi^\dagger(\psi) + B(\psi),$$

where A is constant and B is a function of ψ alone.

Interpretation

Expected utility before and after experiment are

$$\int \{A \log \pi(\psi) + B(\psi)\} \pi(\psi) d\psi,$$

$$\int \left\{ \int \{A \log \pi(\psi \mid x) + B(\psi)\} \cdot \right.$$

$$\left. \cdot \, \pi(\psi \mid x) d\psi \right\} p_X(x) dx,$$

and the difference is $AI\{E, \pi\}$.

19.2 Bayesian background

Multivariate and matrix normal:
$X \sim N_p(\mu, \Sigma)$ has density

$$(2\pi)^{-p/2}|\Sigma|^{-1/2}\cdot$$
$$\cdot \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}.$$

If X and M are $(p \times q)$ matrices and the covariances are of form

$$\mathrm{Cov}\{x_{ij}, x_{k\ell}\} = f_{ik}g_{j\ell},$$

then

$$X \sim N_{pq}(M, F \otimes G)$$

with density

$$(2\pi)^{-pq/2}|F|^{-q/2}|G|^{-p/2} \cdot \exp$$
$$\left[-\frac{1}{2}\mathrm{tr}\{F^{-1}(X-M)G^{-1}(X-M)^T\}\right]$$

Partitioned multivariate normal:

Suppose $X \sim N_p(\mu, \Sigma)$, partitioned so that

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then the conditional distribution of $X^{(1)}$ given $X^{(2)} = x_2$ is normal with mean

$$\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and variance

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Wishart distribution:

$$D \sim W_p(A, m)$$

if $D = \sum_{j=1}^{m} Z_j Z_j^T$, $Z_1, ..., Z_m$ indep. $N_p(0, A)$. Density of form

$$\frac{c_{p,m}|D|^{(m-p-1)/2}}{|A|^{m/2}} \exp\left\{-\frac{1}{2}\text{tr}(DA^{-1})\right\}$$

provided $m > p - 1$.

Inverse Wishart: for us, $\Sigma \sim IW_p(\Psi, m)$ means the same as $\Sigma^{-1} \sim W_p(\Psi^{-1}, m)$.

Matrix t: if

$$\Sigma^{-1} \sim W_p(P, m + p - 1),$$
$$T|\Sigma \sim N_{pq}[0, \Sigma \otimes Q],$$

the T is said to have a matrix t distribution (Dickey (1967), Johnson and Kotz (1972)), written

$$T \sim t(p, q; P, Q, m).$$

Density: given equivalently by

$$\pi^{-pq/2}\Gamma_q\left(\frac{m+p+q-1}{2}\right).$$
$$\left\{\Gamma_q\left(\frac{m+q-1}{2}\right)\right\}^{-1}|Q|^{(p+q-1)/2}.$$
$$|P|^{q/2}|Q+T^TPT|^{-(m+p+q-1)/2}$$

or

$$\pi^{-pq/2}\Gamma_q\left(\frac{m+p+q-1}{2}\right).$$
$$\left\{\Gamma_p\left(\frac{m+p-1}{2}\right)\right\}^{-1}.$$
$$|P|^{-(m+p-1)/2}|Q|^{-p/2}.$$
$$|P^{-1}+TQ^{-1}T^T|^{-(m+p+q-1)/2}.$$

Here

$$\Gamma_p\left(\frac{m}{2}\right)=\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\left(\frac{m+1-j}{2}\right)$$

Partitioned Wishart matrices

Suppose $C \sim W_p^{-1}(B, m)$, $p = a + b$ and C and B are decomposed as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

$$C_{1|2} = C_{11} - C_{12}C_{22}^{-1}C_{21},$$

$$\tau = C_{12}C_{22}^{-1},$$

$$B_{1|2} = B_{11} - B_{12}B_{22}^{-1}B_{21},$$

$$\eta = B_{12}B_{22}^{-1}.$$

Then

$$C_{22} \sim W_b^{-1}(B_{22}, m - a),$$

$$C_{1|2} \sim W_a^{-1}(B_{1|2}, m),$$

$$\tau | C_{1|2} \sim N_{ab}(\eta, C_{1|2} \otimes B_{22}^{-1}),$$

where, also, C_{22} is independent of $(C_{1|2}, \tau)$.

Bayesian multivariate regression

Suppose $y_1, ..., y_n$ independent with

$$y_j \sim N_p(Bx_j, \Sigma).$$

Prior for (B, Σ):

$$\begin{aligned} \Sigma &\sim W_p^{-1}(\Psi, m), \\ B|\Sigma &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}). \end{aligned}$$

Hyperparameters (m, Ψ, B^0, F).

Posterior distribution:

$$\begin{aligned} m &\to m + n, \\ \Psi &\to \Psi + H, \\ B^0 &\to B^*, \\ F &\to G, \end{aligned}$$

where

$$
\begin{aligned}
&G = S_{xx} + F, \\
&B^* = (S_{yx} + B^0 F)G^{-1}, \\
&H = S_{yy} - B^* S_{xy} - S_{yx} B^{*T} \\
&\quad + B^* S_{xx} B^{*T} + (B^* - B^0) F (B^* - B^0)^T
\end{aligned}
$$

Here

$$
\begin{aligned}
S_{yy} &= \sum y_j y_j^T, \\
S_{xy} &= \sum x_j y_j^T, \\
S_{yx} &= \sum y_j x_j^T, \\
S_{xx} &= \sum x_j x_j^T.
\end{aligned}
$$

Predictive distributions: Suppose K new observations are taken to form $p \times K$ matrix Y^* with corresponding $q \times K$ covariate matrix X^*.

$$Y^*|X^*, B, \Sigma \sim N_{pK}(BX^*, \Sigma \otimes I_K).$$

Combine this with

$$\begin{aligned} \Sigma|Y &\sim W_p^{-1}(\Psi + H, m + n) \\ B|\Sigma, Y &\sim N[B^*, \Sigma \otimes G^{-1}] \end{aligned}$$

we find as an intermediate step

$$\begin{aligned} Y^*|X^*, \Sigma, Y \sim & N[B^*X^*, \\ & \Sigma \otimes (I_k + X^{*T} G^{-1} X^*)], \end{aligned}$$

and so

$$\begin{aligned} (Y^* - B^*X^*)|X^*, Y \sim t(p, K; (\Psi + H)^{-1}, \\ I_K + X^{*T} G^{-1} X^*, m + n - p + 1). \end{aligned}$$

Information in MV normal and t distns

If a random variable X has density $f(x)$, define the information in X to be

$$I(X) = \int f(x) \log f(x) dx.$$

(*Note:* Entropy is $-I(X)$.)

If $X \sim N_p(\mu, \Sigma)$, then

$$I(X) = \mathrm{E}\{\log f(X)\} = \text{const} - \frac{1}{2} \log |\Sigma|.$$

If $T \sim t(p, q; P, Q, m)$, then

$$
\begin{aligned}
I(T) &= -\frac{pq}{2} \log \pi \\
&+ \log \left\{ \frac{\Gamma_q((m+p+q-1)/2)}{\Gamma_q((m+q-1)/2)} \right\} \\
&- \frac{p}{2} \log |Q| - \frac{q}{2} \log |P| \\
&- \frac{m+p+q-1}{2} \sum_{j=1}^{q} \left\{ \psi \left(\frac{m+p+q-j}{2} \right) \right. \\
&\left. - \psi \left(\frac{m+q-j}{2} \right) \right\}
\end{aligned}
$$

[$\psi(\cdot)$: digamma function,
Γ_q: multivariate gamma function.]

19.3 Bayesian spatial analysis

Le and Zidek (1992)

Assume $p = u + g$ locations with u "ungauged" and g "gauged".

$$y_j \sim N_p(Bx_j, \Sigma).$$

Partition

$$y_j = \begin{pmatrix} y_j^{(1)} \\ y_j^{(2)} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$
$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Write

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$
$$\tau = \Sigma_{12}\Sigma_{22}^{-1}.$$

Recall: conditional mean of $y_j^{(1)}$ given $y_j^{(2)}$ is $B_1 x_j + \tau(y_j^{(2)} - B_2 x_j)$ and conditional variance is $\Sigma_{1|2}$.

Also write Σ in the form

$$\Sigma = \begin{pmatrix} \Sigma_{1|2} + \tau\Sigma_{22}\tau^T & \tau\Sigma_{22} \\ \Sigma_{22}\tau^T & \Sigma_{22} \end{pmatrix}.$$

Joint prior:

$$\begin{aligned} \Sigma &\sim W_p^{-1}(\Psi, m), \\ B|\Sigma &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}). \end{aligned}$$

Equivalent to:

$$\begin{aligned} \Sigma_{22} &\sim W_g^{-1}(\Psi_{22}, m-u), \\ \Sigma_{1|2} &\sim W_u^{-1}(\Psi_{1|2}, m), \\ \tau|\Sigma_{1|2} &\sim N_{ug}(\eta, \Sigma_{1|2} \otimes \Psi_{22}^{-1}), \end{aligned}$$

where $(\Psi_{22}, \Psi_{1|2}, \eta)$ represents the same decomposition of the prior covariance matrix

Ψ as $(\Sigma_{22}, \Sigma_{1|2}, \tau)$ does of Σ; in particular $\eta = \Psi_{12}\Psi_{22}^{-1}$. Also Σ_{22} is *a priori* independent of $(\Sigma_{1|2}, \tau)$.

Observed data based on

$$y_j^{(2)}|B, \Sigma, x_j \sim N_g(B_2 x_j, \Sigma_{22}).$$

An elementary fact about Bayesian statistics: if the model parameter θ factorizes as (θ_1, θ_2) with θ_1 and θ_2 *a priori* independent, and if the distribution of the observed data D depends only on θ_2, then the posterior distributions of θ_1 and θ_2 are also independent, the posterior for θ_1 being the same as the prior and the posterior for θ_2 the same as if we did not consider θ_1 at all.

$$\frac{\pi_1(\theta_1)\pi_2(\theta_2)f(D|\theta_2)}{\int\int \pi_1(\theta_1)\pi_2(\theta_2)f(D|\theta_2)d\theta_1 d\theta_2}$$
$$= \pi_1(\theta_1) \cdot \frac{\pi_2(\theta_2)f(D|\theta_2)}{\int \pi_2(\theta_2)f(D|\theta_2)d\theta_2}.$$

Apply with

$$\theta_1 = (\Sigma_{1|2}, \tau, B_1 - B_1^0 - \tau(B_2 - B_2^0)),$$
$$\theta_2 = (B_2, \Sigma_{22}).$$

Then

$$\Sigma_{22}|D \sim W_g^{-1}(\Psi_{22} + H_{22}, m - u + n),$$
$$B_2|D, \Sigma_{22} \sim N_{gq}(B_2^*, \Sigma_{22} \otimes G^{-1}),$$

where

$$G = S_{xx} + F,$$
$$B_2^* = (S_{y_2x} + B_2^0 F)G^{-1},$$
$$H_{22} = S_{y_2y_2} - B_2^* S_{xy_2} - S_{y_2x} {B_2^*}^T + B_2^* S_{xx} {B_2^*}^T + (B_2^* - B_2^0)F(B_2^* - B_2^0)^T.$$

Key features of result:

(a) Those parts of the model that relate to $y^{(2)}$ are identical to the standard Bayesian calculations for multivariate regression. In particular, we may calculate a predictive distribution for a future observation ${y^{(2)}}^*$, which is multivariate t.

(b) The parameters relating to the conditional distribution of $y^{(1)}$ given $y^{(2)}$ are unaffected by the data, i.e. the posterior distributions are the same as the priors. In particular, this applies to the predictive conditional density of ${y^{(1)}}^*$ given ${y^{(2)}}^*$, which is multivariate t.

(c) Nevertheless, the *marginal* predictive density of ${y^{(1)}}^*$, obtained by multiplying (a) and (b) together and integrating out ${y^{(2)}}^*$, is affected by the data, because the marginal predictive density of ${y^{(2)}}^*$ is.

19.4 Hierarchical models

Need more structure on Ψ to get meaningful interpolations

Simple model:

$$\Psi = \begin{cases} \sigma^2 & \text{if } i = j, \\ \rho\sigma^2 & \text{if } i \neq j, \end{cases} .$$

Could also consider parametric structures on Ψ as in standard geostatistics (Switzer).

Suggests hierarchical models of form

$$\begin{aligned} y_j^{(2)} | B, \Sigma, x_j &\sim N_g(B_2 x_j, \Sigma_{22}), \\ \Sigma | \Psi, m &\sim W_p^{-1}(\Psi, m), \\ B | \Sigma, B^0, F &\sim N_{pq}(B^0, \Sigma \otimes F^{-1}), \end{aligned}$$

where B, Σ etc. depend on additional parameters θ and we assume prior

$$\theta \sim \pi(\theta).$$

Possible approaches now:

1. Ignore the prior on θ but obtain "type II MLE" for θ based on integrated likelihood. Could integrate out Σ analytically but they preferred to use the EM algorithm for this.

2. Fully Bayesian. Not yet used in practice.

Extensions

1. Brown, Le and Zidek (1994) consider the case of K pollutants measured at each site, with a prior covariance matrix Ψ of form

$$\Psi = \Lambda \otimes \Omega,$$

with Λ a $p \times p$ matrix of intersite covariances and Ω a $K \times K$ matrix of covariances between the variables.

Advantages: Simplifies specification of Ψ

Computational advantages if p and K both large

2. Le, Sun and Zidek (1997) extended this to *data missing by design*: suppose L of the gK possible site $\times$ pollutant combinations are never observed (g=number of sites with some data). They then proposed an extension of the EM algorithm to estimate the full $gK \times gK$ covariance matrix rather than just the $(gK - L) \times (gK - L)$ submatrix corresponding to the non-missing values.

19.5 Discussion and extensions

Advantages over traditional kriging:

• Predictive distributions take account of unknown parameters

• Less reliance on parametric models

• Allows full use of covariates

• Multivariate kriging as easy as univariate

Applications:

Air pollution network in Ontario
Brown, Le and Zidek (1994)
Le, Sun and Zidek (1997)

Used model from Sampson and Guttorp (1992) to model nonstationary covariance structure in Ψ (not fully Bayesian)

Sun, Zidek, Le and Özkaynak (2000):

Interpolate PM_{10} field from 10 stations in Vancouver

Used a CV approach to assess quality of spatial predictions

Not temporally independent: their method was to fit a common time series model at each site and apply spatial analysis to residuals

"Spatial leakage" phenomenon

If they assumed separability of the spatio-temporal covariance they could in principle avoid that by extending preceding analysis to include a temporal covariance as well as spatial covariance – not tried.

19.6 Entropy-based criteria for design

First formulation: Caselton-Zidek (1984)

Divide $p = u + g$ sites into two subsets, g sites gauged and rest ungauged. Y represents random field at all p sites, but subdivided into $Y^{(1)}$ and $Y^{(2)}$ corresponding to the ungauged and gauged sites respectively. Assume $Y \sim N_p(\mu, \Sigma)$, μ and Σ known.

Information criterion

$$I(Y^{(1)}|Y^{(2)}) - I(Y^{(1)}) = \frac{1}{2} \log \frac{|\Sigma_{11}|}{|\Sigma_{1|2}|}$$

Equivalent to minimizing

$$\sum \log(1 - \rho_i^2),$$

where $\rho_1^2, \rho_2^2, ...$ are the squared *canonical correlations* between U and G.

Reformulation: Caselton-Kan-Zidek (1992)

Definition of entropy (Jaynes 1963):

$$H(Y) = \mathrm{E}\left\{-\log \frac{f(Y)}{m(Y)}\right\},$$

(m a reference measure)

CZ criterion: choose U to minimize

$$H(U|G) - H(U).$$

However, CKZ criticized this as ignoring the information in G. Instead, minimize $H(U|G)$ directly. Because

$$H(U, G) = H(U|G) + H(G),$$

this is equivalent to:

Choose G to maximize $H(G)$.

MVN case: choose G to maximize $|\Sigma_{22}|$.

CKZ considered μ and Σ unknown but estimated through data on *complete* network. Prior of form

$$\Sigma \sim W_p^{-1}(\Psi, m),$$
$$\mu|\Sigma \sim N_p(\mu^0, f^{-1}\Sigma),$$

posterior of same form with (Ψ, m, μ^0, f) replace by $(\hat{\Psi}, \hat{m}, \hat{\mu}^0, \hat{f})$.

Predictive distribution: split future observation into ${y^{(1)}}^*$ ungauged and ${y^{(2)}}^*$ gauged, focus on prediction of ${y^{(2)}}^*$.

Predictive distribution of ${y^{(2)}}^* - \hat{\mu}_2^0$ is $t(g, 1; \hat{\Psi}_{22}^{-1}, 1 + \hat{f}^{-1}, m - u - g + 1)$ and the entropy is of the form

$$H(G) = \frac{1}{2} \log |\hat{\Psi}_{22}| + \text{ constants.}$$

Therefore, the network design problem reduces to choosing G to maximize $\log|\hat{\Psi}_{22}|$.

Application to the optimal reduction of a set of $p = 81$ stations in a wet deposition monitoring network. $n = 48$ months' data available — still caused problems with degrees of freedom for estimating Σ. Another paper by Wu and Zidek (1992) avoided that problem by dividing into regions of < 48 stations per region.

Algorithm for reducing network: drop one station at a time.

Later extended the formulation to include hierarchical models for Ψ: Brown-Le-Zidek (1994)

Incorporating costs: Zidek-Sun-Le (2000)

ZSL extended the CKZ methodology to incorporate costs of measurement

Choose new sites to maximize

$$\begin{aligned}&\log|\Phi_{\mathrm{add|g}}|\\&=\log|\Phi_{\mathrm{add}}-\Phi_{\mathrm{add,\ g}}\Phi_{\mathrm{g}}^{-1}\Phi_{\mathrm{g,\ add}}|\end{aligned}$$

Accounting for costs: if potential site s has entropy $E(s)$ and cost $C(s)$, choose s to maximize

$$O(s) = E(s) - DE\ C(s)$$

for some conversion factor DE.

Also mention new algorithms for finding combination of sites to maximize an entropy criterion (Ko *et al.* 1995, Anstreicher *et al.* 1996).

$$Y|\mu, \Sigma \sim f(Y|\mu, \Sigma),$$
$$(\mu, \Sigma)|\theta \sim g(\mu, \Sigma|\theta),$$
$$\theta \sim \pi(\theta).$$

Can find MCMC sample $\{\theta_a,\ 1 \leq a \leq A\}$ from the posterior density $\pi(\theta|Y)$.

The full predictive density is

$$\int\int\int f(y^{(2)^*}|\mu, \Sigma) g(\mu, \Sigma|\theta, Y) \pi(\theta|Y) d\mu d\Sigma d\theta$$

However, the integral with respect to μ and Σ is possible analytically, so reduce to

$$f_{\text{pred}}(y^{(2)^*}|Y) = \int f_{\text{pred}}(y^{(2)^*}|\theta, Y) \pi(\theta|Y) d\theta$$

which we would in practice estimate by

$$\hat{f}_{\text{pred}} = \frac{1}{A}\sum_{a=1}^{A} f_{\text{pred}}(y^{(2)^*}|\theta_a, Y)$$

We still need to evaluate entropy, however.

1. Generate $\theta_1, ..., \theta_A$ from posterior distribution of θ given Y.

2. Fix some reference value $\tilde{\theta}$, e.g. sample mean of $\theta_1, ..., \theta_A$. $\hat{\Psi}$ etc. computed using $\tilde{\theta}$.

3. Generate independent $z_1, ..., z_B$ from $N_g(0, I_g)$, $S_1, ..., S_B$ from $W_g^{-1}(I_g, \hat{m} - u)$.

4. For each $b \in \{1, ..., B\}$, define

$$\Sigma_{22}(b) = \hat{\Psi}_{22}^{1/2} S_b \hat{\Psi}_{22}^{1/2}$$
$$y^{(2)^*}(b) = \hat{\mu}_2^0 + (1 + \hat{f}^{-1})^{1/2} \Sigma_{22}^{1/2}(b) z_b$$

5. Calculate

$$\hat{H}(G) = -\frac{1}{B}\sum_{b=1}^{B} \log \hat{f}_{\text{pred}}(y^{(2)^*}(b)|Y) \cdot \cdot \frac{\hat{f}_{\text{pred}}(y^{(2)^*}(b)|Y)}{\hat{f}_{\text{pred}}(y^{(2)^*}(b)|\tilde{\theta}, Y)},$$

where $\hat{f}_{\text{pred}}(y^{(2)^*}|\tilde{\theta}, Y)$ is the analytic predictive density derived from the multivariate t distribution.

Note that we use the same random numbers for all possible G, to facilitate comparisons among different G of same size g.

20. Methods based on optimal design theory

20.1 The General Equivalence Theorem

Classical formulation of optimal design theory:

$$y_i = \sum_{j=1}^{p} f_j(x_i)\beta_j + \epsilon_i, \ \ i = 1, .., n$$

where f_j are known functions of design points x_i, $\beta_1, ..., \beta_p$ are unknown coefficients and, as usual in linear regression theory, ϵ_i are uncorrelated errors with mean 0 and common variance σ^2.

Write in form

$$Y_n = F_n\beta + \epsilon,$$

information matrix $\propto F_n^T F_n$.

Write in form

$$\frac{1}{n} F_n^T F_n = \int_{\mathcal{X}} f(x) f(x)^T d\xi_n(x)$$

with ξ_n the discrete measure that places mass $\frac{1}{n}$ at each x_i. Define

$$M(\xi) = \int_{\mathcal{X}} f(x) f(x)^T d\xi(x),$$

for any positive measure ξ. Optimal design criteria are of the following form: choose ξ to minimize $\Psi\{M(\xi)\}$ for some functional $\Psi\{\cdot\}$.

Examples:

- *D-optimality:* $\Psi = -\log |M(\xi)|$.
- *A-optimality:* $\Psi = \mathrm{tr}\{M(\xi)^{-1}\}$.
- *E-optimality:* Ψ is largest eigenvalue of $M(\xi)^{-1}$.

• *G-optimality:* $\Psi = \max_{x \in \mathcal{X}} d(x, \xi)$ where $d(x, \xi) = f(x)^T M(\xi)^{-1} f(x)$. (Prediction variance interpretation)

Note that

$$\int_{\mathcal{X}} d(x, \xi) d\xi(x) = p.$$

If we can find ξ^* for which $\max_{x \in \mathcal{X}} d(x, \xi^*) = p$, ξ^* must be G-optimal.

General equivalence theorem

Suppose δ_x is a unit point mass at x, and consider modifying ξ into

$$\xi'_{\alpha,x} = (1 - \alpha)\xi + \alpha\delta_x,$$

where $0 < \alpha < 1$. Then

$$M(\xi'_{\alpha,x}) = (1 - \alpha)M(\xi) + \alpha M(\delta_x).$$

The derivative of Ψ, in the direction δ_x, is

$$\phi(x, \xi) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[\Psi\{M(\xi'_{\alpha,x})\} - \Psi\{M(\xi)\} \right].$$

The General Equivalence Theorem asserts that the following are equivalent:

(1) ξ^* minimizes $\Psi\{M(\xi)\}$,

(2) $\phi(x, \xi^*) \geq 0$ for all x,

(3) $\phi(x, \xi^*)$ achieves its minimum at points of the design, i.e. at points x which have positive point measure under ξ^*.

Also leads to an algorithm: to find optimal design iteratively, increase mass of ξ at points where $\phi(x, \xi)$ is minimum

Example. Consider $\Psi = -\log |M(\xi)|$. Suppose $\xi = \xi_n$, the point measure with mass $\frac{1}{n}$

at each of $x_1, ..., x_n$. Let F_n be corresponding design matrix. Assume new design ξ_{n+1} created by adding point mass at x, so that $F_{n+1} = (\, F_n^T \quad f(x)\,)^T$ and

$$\xi_{n+1} = \frac{n}{n+1}\xi_n + \frac{1}{n+1}\delta_x.$$

We also have

$$F_{n+1}^T F_{n+1} = F_n^T F_n + f(x)f(x)^T.$$

Therefore, $|F_{n+1}^T F_{n+1}|$ is

$$\begin{aligned} &|F_n^T F_n| \cdot |I_p + (F_n^T F_n)^{-1} f(x)f(x)^T| \\ =&|F_n^T F_n|\{1 + f(x)^T (F_n^T F_n)^{-1} f(x)\} \end{aligned}$$

using the *determinant identity* $|I_n + B^T C| = |I_m + CB^T|$ applicable whenever B and C are both $m \times n$ matrices.

Since $|F_n^T F_n| = n^p |M(\xi_n)|$, we have

$$\begin{aligned}
&\Psi\{M(\xi_{n+1})\} - p\log(n+1) \\
&= \Psi\{M(\xi_n)\} - p\log n \\
&- \log\left\{1 + \frac{f(x)^T M(\xi)^{-1} f(x)}{n}\right\},
\end{aligned}$$

and hence

$$\begin{aligned}
&\lim_{n\to\infty} n[\Psi\{M(\xi_{n+1})\} - \Psi\{M(\xi_n)\}] \\
&= \lim_{n\to\infty} np\log\frac{n+1}{n} \\
&\qquad - \lim_{n\to\infty} n\log\left\{1 + \frac{d(x,\xi)}{n}\right\} \\
&= p - d(x,\xi_n).
\end{aligned}$$

Since, for large n, any design ξ may be approximated by one concentrated on n equally weighted points, we conclude

$$\phi(x,\xi) = p - d(x,\xi).$$

With this interpretation, condition (2) for the D-optimality of a design ξ^* reduces to

$$d(x, \xi) \leq p \text{ for all } x,$$

and we have already seen that this implies G-optimality of the design ξ. Therefore, provided we extend the notion of design to allow arbitrary measures ξ, the General Equivalence Theorem implies that D-optimality and G-optimality are equivalent.

20.2 The Fedorov-Müller approach

Ref: Fedorov and Müller (1989)

Assume

$$y_{it} = f(x_i)^T \theta_t + \epsilon_{it},\ 1 \le i \le n,\ 1 \le t \le T,$$

where y_{it} denotes time t and location x_i.

Random effects model: $\theta_t \sim N[\theta_0, D_0]$ independently at each t. So

$$y_t = F^T \theta_t + \epsilon_t,\ \mathrm{E}\{\theta_t\} = \theta_0,\ \mathrm{Cov}\{\theta_t\} = D_0,$$

or equivalently

$$\mathrm{E}\{y_t\} = \theta_0, \quad \mathrm{Cov}\{y_t\} = I + F^T D_0 F.$$

Three cases:

(a) θ_0, D_0 known, θ_t to be estimated for each t,

(b) D_0 known, θ_0 and each θ_t to be estimated,

(c) θ_0, D_0 both unknown.

Only consider (a) here. Optimal estimator of θ_t is

$$\hat{\theta}_t = (D_0^{-1} + nM)^{-1}(D_0^{-1}\theta_0 + Fy_t),$$

where $M = n^{-1}FF^T$. In this case,

$$\mathrm{E}\{(\hat{\theta}_t - \theta_t)(\hat{\theta}_t - \theta_t)^T\} = (D_0^{-1} + nM)^{-1}.$$

$$\Psi = -\log|D_0^{-1} + nM|.$$

By General Equivalence Theorem, observations should be placed at locations which maximize $\psi(x, \xi)$, where

$$\psi(x, \xi) = f(x)^T\{D_0^{-1} + nM(\xi)\}^{-1}f(x).$$

Alternatively: Use kriging to estimate at unobserved points, place observations where kriging variance is largest.

Since the latter approach is equivalent to finding the G-optimal design, the two methods are the same.

Moral of the story: Classical optimal design criteria can be used to solve network optimality problems.

Unfortunately, this result does not extend to more realistic spatial models.

20.3 Designs for estimating a regression function in a spatially correlated field

Ref: Müller (2000)

$$y_i = \eta(x_i, \beta) + \epsilon_i,$$

where $\eta(\cdot, \cdot)$ is a known nonlinear function dependent on an unknown parameter β, and $\mathrm{Cov}\{\epsilon_i, \epsilon_j\} = c(x_i, x_j)$ is *known*.

Information matrix:

$$M(A) = \sum_{x \in A} \sum_{x' \in A} \dot{\eta}(x) [C(A)^{-1}]_{x,x'} \dot{\eta}^T(x').$$

Design criterion:

$$\max_{A \subset \mathcal{X},\ n_A \leq n} \Phi\{M(A)\},$$

where n is the given number of monitors and $\mathcal{X}$ is the sampling space from which the values x_i must be selected.

Approximate information matrices.

Idea: original information matrix leads to functions of ξ which are not concave or differentiable, so must modify it.

Modification 1:

$$M^{(\epsilon)}(\xi) = \sum_{x \in S_\xi} \sum_{x' \in S_\xi} \dot{\eta}(x) \\ [C(S_\xi) + W^{(\epsilon)}(\xi)]^{-1}_{x,x'} \dot{\eta}^T(x'),$$

where $\epsilon > 0$ and $W^{(\epsilon)}(\xi)$ is a diagonal matrix with entries

$$[W^{(\epsilon)}(\xi)]_{x,x} = \log \left\{ \left(\frac{\xi^{(\epsilon)}}{\xi(x)} \right)^{\epsilon} \right\},$$

and $\xi^{(\epsilon)} = (\sum_{x \in S_\xi} \xi(x)^{1/\epsilon})^{\epsilon}$ is a continuous approximation to $\xi_{\max} = \max\{\xi(x), x \in S_\xi\}$. ($S_\xi$: support of ξ)

Idea: Gives high weight to designs with approximately equal masses on their support.

Modification 2:

$$M_\kappa^{(\epsilon)}(\xi) = \sum_{x \in S_\xi} \sum_{x' \in S_\xi} \dot{\eta}(x) \\ [C(S_\xi) + W_\kappa^{(\epsilon)}(\xi)]^{-1}_{x,x'} \dot{\eta}^T(x'),$$

where $W_\kappa^{(\epsilon)}(\xi)$ is a diagonal matrix with entries

$$[W_\kappa^{(\epsilon)}(\xi)]_{x,x} = \log\left\{\left(\frac{\xi_{\epsilon,\kappa}}{\xi_{\epsilon,\kappa}(x)}\right)^\epsilon\right\},$$

$$\xi_{\epsilon,\kappa}(x) = \{\kappa^{1/\epsilon} + \xi(x)^{1/\epsilon}\}^\epsilon - \kappa,$$

$$\xi_{\epsilon,\kappa} = \{\kappa^{1/\epsilon} + \sum_{x \in S_\xi} \xi(x)^{1/\epsilon}\}^\epsilon - \kappa.$$

Idea: Force weights to be at least $\kappa > 0$. In practice take $\kappa \approx \frac{1}{n}$.

Then construct iterative design procedure based on gradient of $|M_\kappa^{(\epsilon)}(\xi)|$

20.4 Other design objectives: Estimating the variogram

Criteria of Warrick and Myers (1987): Assume interpoint distances are grouped into classes for variogram estimation.

1. For each distance-angle class, the number of pairs should be as large as possible, particularly for short distances,

2. The average of the distances in each class should be close to the plotted distance,

3. The variance of the distances in each class should be small,

4. The average of the angles in each class should be close to the plotted angle,

5. The variance of the angles in each class should be small.

Müller and Zimmerman (1999) considered designs to maximize the information matrix for a parametric variogram model, when the estimator is variogram-based. Their procedure is complicated to implement. Here, we propose an alternative based on the MLE.

Suppose the current data are represented by an n-dimensional vector Y with mean $\mu(\theta)$ and covariance matrix $V(\theta)$, both parametric functions of some p-dimensional θ. Let $M(\theta)$ denote the Fisher information matrix, i.e. the matrix with entries $m_{rs}(\theta),\ 1 \leq r \leq p,\ 1 \leq s \leq p$, where

$$m_{rs}(\theta) = \mathrm{E}\left\{ \frac{\partial^2 \log f(Y;\theta)}{\partial\theta_r \partial\theta_s} \right\}.$$

Then

$$m_{rs}(\theta) = \frac{1}{2}\mathrm{tr}\left(\frac{\partial V}{\partial\theta_r} V^{-1} \frac{\partial V}{\partial\theta_s} V^{-1} \right) + \frac{\partial \mu^T}{\partial\theta_r} V^{-1} \frac{\partial \mu}{\partial\theta_s}.$$

Suppose we consider adding a new observation y^* at location x. The joint distribution of $\begin{pmatrix} Y^T & y^* \end{pmatrix}^T$ is normal with mean $\begin{pmatrix} \mu^T & \nu \end{pmatrix}^T$ and covariance matrix $\begin{pmatrix} V & \tau \\ \tau^T & \phi \end{pmatrix}$ where ν and ϕ are the mean and variance of y^* and τ is the vector of cross-covariances between Y and y^*. The conditional distribution of y^* given $Y = y$ is of the form $N(\beta, \alpha)$, where $\beta = \nu + \tau^T V^{-1}(y - \mu)$ and $\alpha = \phi - \tau^T V^{-1}\tau$. The (r, s) contribution to the information matrix from y^*, conditional on $Y = y$, is

$$\frac{1}{2\alpha^2}\frac{\partial\alpha}{\partial\theta_r}\frac{\partial\alpha}{\partial\theta_s} + \frac{1}{\alpha}\frac{\partial\beta}{\partial\theta_r}\frac{\partial\beta}{\partial\theta_s}.$$

However, β is a function of y, so we need to take a further expectation to evaluate the unconditional expectation. Defining $\omega = V^{-1}\tau$, this turns out to be

$$\frac{1}{2\alpha^2}\frac{\partial\alpha}{\partial\theta_r}\frac{\partial\alpha}{\partial\theta_s}+\frac{1}{\alpha}\left\{\left(\frac{\partial\nu}{\partial\theta_r}-\omega^T\frac{\partial\mu}{\partial\theta_r}\right)\cdot\right.$$
$$\left.\cdot\left(\frac{\partial\nu}{\partial\theta_s}-\omega^T\frac{\partial\mu}{\partial\theta_s}\right)+\frac{\partial\omega}{\partial\theta_r}V\frac{\partial\omega}{\partial\theta_s}\right\}$$

Collecting the entries into a matrix $U(\theta)$, we see that the new information matrix after adding y^* is of the form $M(\theta)+U(\theta)$. Choose new design point x to maximize

$$\Phi\{M(\theta)+U(\theta)\}-\Phi\{M(\theta)\}.$$

However, U is of the form WW^T where W is $p\times 3$. Then for D-optimality,

$$\begin{aligned}|M+U| &= |M+WW^T| \\ &= |M|\cdot|I_3+W^TM^{-1}W|,\end{aligned}$$

so the only determinant that needs to be re-evaluated for each x is that of a 3×3 matrix.

20.5 Combining estimative and predictive criteria

Work of Zhengyuan Zhu (2002)

Choose $S = \{s_1, ..., s_n\} \subset D$.

Process $Z(x) \sim N[\mu(x), \Sigma(\theta)]$.

Problems:

1. Optimal sampling design for prediction
2. Optimal sampling design for covariance parameter estimation

Initially concentrate on problem 2. Use $\log|\mathcal{I}|$ as design criterion, where $\mathcal{I}$ is the Fisher information matrix.

Assume $\mu(x) = \mu(M(x); \beta)$ as a linear or nonlinear regression function.

When $\mu = 0$, components of $\mathcal{I}$ are

$$I_{jk}(\theta) = \frac{1}{2}\text{tr}\{\Sigma^{-1}\Sigma_j\Sigma^{-1}\Sigma_k\},$$

where $\Sigma_j = \partial\Sigma/\partial\theta_j$.

Locally Optimal Designs

Criterion will minimize

$$V_0(S;\theta) = -\log|\mathcal{I}(\theta;S)|,$$

equivalent to D-optimality.

"Local design" minimizes V_0 for fixed θ using preliminary estimate $\hat{\theta}_0$.

"Minimax design" minimizes
$\max_{\theta\in\Theta} V_0(S;\theta)$.

Disadvantage — gives most attention to θ for which $V_0(S;\theta)$ is large for all S.

Instead, use relative efficiency

$$V_1(S,\theta) = -\log \frac{|\mathcal{I}(\theta;S)|}{|\mathcal{I}(\theta;S(\theta))|}$$

where $S(\theta)$ is locally optimal at θ.

Invariant to smooth nonlinear transformations of θ.

In practice, discretize Θ, maximize $V_1(S,\theta)$ over a finite set of θ.

Bayesian approach

Utility function $U(\theta, S, Z)$

If $U = V_1$ then

$$V_2(S) = -\int_\Theta \log \frac{|\mathcal{I}(\theta;S)|}{|\mathcal{I}(\theta;S(\theta))|} p(\theta) d\theta.$$

For general U, $U(S)$ is

$$\int\int U(\theta,S,Z)p(\theta|Z,S)p(Z|S)d\theta dZ.$$

Using Shannon information for U, leads to

$$U(S) = -\int_{\Theta} \log |\mathcal{I}(\theta; S)| p(\theta) d\theta$$

(Chalenor and Verdinelli 1995) — equivalent to $V_2(S)$ but preferable computationally.

Suppose the objective is to choose $S = \{s_1, ..., s_n\} \subset D$ to minimize some objective function $V(S)$. In practice we will restrict D to a finite set, say $|D| = N$, but this still leaves $\binom{N}{n}$ possible designs.

Ko *et al.* (1995) shows that for some design problems it is possible to obtain exact solution via a branch and bound algorithm, but their largest example had $N = 27$, $n = 13$, $\binom{27}{13} = 20,058,300$. We are interested in $N \approx 10^4$. Exact solution not possible.

Proposed method uses *simulated annealing*.

The idea of simulated annealing

Suppose we have a current design S.
Choose some trial new design S^* in a neighborhood of S.

1. If $V(S^*) \leq V(S)$, accept S^*.

2. If $V(S^*) > V(S)$, accept S^* with probability $\exp\left(\frac{V(S)-V(S^*)}{T}\right)$; otherwise stay at S and generate a new S^* for the next iteration.

T is called the "temperature" and should gradually be reduced as the number of iterations increases.

The steps $S \to S^*$ also decrease in size with the number of iterations — initially use a large neighborhood, then decrease.

Initial value of S is critical — optimize simultaneously from several starting values of S, then take the best among these.

Summary of Conclusions

The locally optimal design is highly dependent on θ

Minimax and Bayesian designs are better overall, but not as good as the locally optimal design for a specific θ.

Minimax and Bayesian designs both substantially outperform regular sampling designs.

Two-step design for prediction with estimated parameters

Select n sampling points from a region D.

Theoretical work by Su and Cambanis (1993)

Simulated annealing too complicated for $n \geq 100$.

Propose two-step procedure: use some sites to find best design for prediction with known covariances, then find best remaining sites (conditional on those already selected) to produce best design for estimation.

The proportion of sites in each stage is optimized to get the best prediction with estimated parameters.

Criteria: The goal is good prediction over the whole of D.

Let $V(x_0, S)$ be the optimization criterion for prediction at a specific location x_0.

Then consider integrated or maximum mean square (weighted) error:

$$A(S) = \int_D V(x, S)w(x)dx$$

or

$$M(S) = \sup\{V(x, S)w(x),\ x \in D\}.$$

Difficulty when parameters are unknown: want good point predictors *and* prediction intervals with accurate coverage probabilities.

"Plug-in" method vs. Bayesian prediction

Consider $\hat{Z}(\theta) = \lambda^T(\theta)Z$, BLUP for Z_0.

Then $\hat{Z}(\hat{\theta})$ is the plug-in or EBLUP.

Assume REML estimator — makes
$\hat{Z}(\hat{\theta}) - \hat{Z}(\theta)$ independent of $\hat{Z}(\theta) - Z_0$.

Then

$$
\begin{aligned}
&E\{(\hat{Z}(\hat{\theta}) - Z_0)^2\} \\
&= E\{(\hat{Z}(\theta) - Z_0)^2\} + E\{(\hat{Z}(\hat{\theta}) - \hat{Z}(\theta))^2\} \\
&= M(\theta) + E\{(\lambda^T(\hat{\theta})Z - \lambda^T(\theta)Z)^2\}
\end{aligned}
$$

Second term is

$$\approx \mathrm{tr}\left\{\mathcal{I}(\theta)^{-1}\left(\frac{\partial\lambda}{\partial\theta}\right)^{T}\Sigma(\theta)\left(\frac{\partial\lambda}{\partial\theta}\right)\right\}$$

where $\frac{\partial\lambda}{\partial\theta}$ is the matrix with entries $\frac{\partial\lambda_i}{\partial\theta_j}$.

Leads to

$$V_1(x_0, S) = M(\theta) + \mathrm{tr}\left\{\mathcal{I}(\theta)^{-1}\left(\frac{\partial\lambda}{\partial\theta}\right)^{T}\Sigma(\theta)\left(\frac{\partial\lambda}{\partial\theta}\right)\right\},$$

cf. Harville and Jeske (1992), Zimmerman and Cressie (1992).

$$
\begin{aligned}
& Var\{M(\hat{\theta})\} \\
& \approx \left(\frac{\partial M(\theta)}{\partial \theta}\right)^T \mathcal{I}(\theta)^{-1} \left(\frac{\partial M(\theta)}{\partial \theta}\right) \\
& = V_2(x_0, S).
\end{aligned}
$$

Possibility of using some linear combination of V_1 and V_2 as the design criterion, but it's not clear how to weight them.

An alternative way of characterizing the discrepancy between the predictive densities based on the true θ and the estimated $\hat{\theta}$ is to use Kullback-Leibler divergence,

$$\begin{aligned} & D(\theta, \hat{\theta}; Z(x_0)|Z) \\ & = E_\theta \left\{ \log \frac{p(Z(x_0)|Z, \theta)}{p(Z(x_0)|Z, \hat{\theta})} \right\}. \end{aligned}$$

Define

$$\mathcal{I}(\theta; W|X) = Cov\left\{ \frac{\partial}{\partial \theta} \log p(W|X, \theta), \frac{\partial}{\partial \theta} \log p(W|X, \theta)^T \right\}.$$

Then

$$\mathcal{I}(\theta; (W, X)) = \mathcal{I}(\theta|X) + \mathcal{I}(\theta; W|X).$$

Stein (1999) suggested

$$
\begin{aligned}
& D(\theta, \hat{\theta}; Z(x_0)|Z) \\
& \approx \frac{1}{2}\mathrm{tr}\{\mathcal{I}(\theta; Z)^{-1}\mathcal{I}(\theta; Z(x_0)|Z)\} \\
& \approx \frac{E\{(M(\hat{\theta}) - M(\theta))^2\}}{4M(\theta)^2} \\
& + \frac{E\{(\hat{Z}(\hat{\theta}) - \hat{Z}(\theta))^2\}}{2M(\theta)}.
\end{aligned}
$$

The first of these leads to design criterion

$$
\begin{aligned}
& V_3(x_0, S) = M(\theta)\Bigg[1+ \\
& \frac{c}{2}\mathrm{tr}\left\{\mathcal{I}(\theta; Z)^{-1}\mathcal{I}(\theta; Z(x_0)|Z)\right\}\Bigg].
\end{aligned}
$$

Suggest $c = 2$ when $V_3 \approx V_1 + \frac{V_2}{2M(\theta)}$. Call this the *estimation adjusted* (EA) criterion.

Bayesian criteria

Recall Bayesian predictive formula for prediction at x_0 given Z:

$$p(Z(x_0)|Z) = \int p(Z(x_0)|Z,\theta)p(\theta|Z)d\theta.$$

In principle one could use some criterion based on this predictive distribution (e.g. the expected width of a 95% posterior prediction interval) as a criterion for choosing a design.

However there are practical difficulties: the predictive density can only be evaluated using Monte Carlo methods, and in practice one would have to repeat the calculation for many x_0 to get a useful result.

Therefore, prefer to concentrate on simpler approximate criteria such as EA.

Two-step design

• For $p \in (0, 1)$, find $(1 - p)n$ design points to minimize mean squared prediction error assuming θ known

• Given these $(1-p)n$ design points, choose remaining np design points to minimize $-\log|\mathcal{I}|$ for the complete design, where $\mathcal{I}$ is the Fisher information matrix

• Repeat for several values of p and choose p to minimize an overall criterion for predictive error

In practice, can often use a regular design for the first step, and $p << 1$; then np is small enough to use simulated annealing in the second step.

21. Other approaches to network design

21.1 Haas's approach

Haas (1992) proposed procedure for optimally selecting new locations in a monitoring network, based on (a) the mean relative error of estimation (estimate standard error divided by estimate), and (b) the standard deviation of the relative error estimate at the subregion's center. Idea to balance prediction vs. estimation criteria.

21.2 Oehlert's approach

Oehlert (1993, 1995, 1996)

Features of his approach:

1. A novel spatial-temporal model for sulfate deposition. A natural candidate for Bayesian/MCMC analysis, though he did not use those ideas

2. Optimal design characterized in terms of either *regional* or *local* prediction errors

3. For network reduction, precise criterion does not matter too much because the method largely removes redundant station. For network addition, criterion is more important

4. Most ambitious exercise was to remove 100 stations from a 249-station network. The increased prediction variance using one-at-a-time optimal deletion was much smaller than using random deletions

21.3 Two methods by Nychka and Saltzman

Ref: Nychka and Saltzman (1998)

First approach: regression and variable selection

Useful when

(a) reducing an existing network given past data on the entire network,
(b) the variable predicted is something measurable in the past data (e.g. average over the whole current network)

In that case, it is possible to dispense with spatial modeling altogether and just use regression. The optimal design problem is then one of deciding which columns of the design matrix to keep, i.e. a problem of variable selection.

They used *leaps* (Furnival and Wilson 1974) and *lasso* (Tibshirani 1995) algorithms to do this.

Second approach: space-filling designs

Find designs to "fill space" in some suitable sense.

$$d_p(x, D) = \left(\sum_{u \in D} ||x - u||^p \right)^{1/p},$$

$$C_{p,q}(D) = \left(\sum_{x \in C} d_p(x, D)^q \right)^{1/q},$$

where $p < 0, q > 0$. Limits $p \to -\infty$ and $q \to \infty$ result in "minimax" criterion, i.e. minimize the maximum distance from any point in C to the nearest point in $D \subset C$.

Other metrics? e.g. Trujillo-Ventura and Ellis (1991) used

$$\sum_{i=1}^{N} \min_{j \neq i} |x_i - x_j|^{1/2}.$$

21.4 The Bayesian approach of P. Müller

e.g. Sansó and Müller (1997). Problem of reducing a set of 80 rainfall stations in Venezuela to one of 40 stations.

Decision-theoretic: utility function associated with network D of form

$$u(D, y) = C \sum_{i \in D^c} \mathcal{I}\left(|y_i - \hat{y}_i(y_D)| \leq \delta\right) - \sum_{i \in D} c_i + C_0,$$

where δ is a target for prediction accuracy, C is payoff for correct prediction, c_i is the cost of operating station i and C_0 is constant — need $u(D, y) > 0$ for all D and y.

Risk function:

$$U(D) = \mathrm{E}\{u(D, y)\}.$$

The idea: Embed this in a Bayesian estimation problem by treating the design as an unknown "parameter" and sampling from the density

$$h(D, \theta, y) \propto p(\theta)p(y|\theta)u(d, \theta, y).$$

We then obtain a "posterior distribution" of D by Monte Carlo sampling; the mode of this distribution will approximate the optimal D.

22. Designs for data assimilation

Berliner, Lu and Snyder (1999)

The problem:

X_0 (p-dimensional) represents "current state" of weather (time t_0). $p \approx 10^7$.

$X_1 = F_0(X_0)$ represents deterministic evolution of system to some future time point t_1. At this point we take some observations Y of dimension $q \approx 10^5$.

Real interest is in $X_2 = F_1(X_1)$ at some further time point t_2. Will use Y together with numerical model to predict X_2.

However, ≈ 50 of the observations in Y correspond to an aircraft flight and are therefore under the experimenter's control. The design problem concerns these observations.

The problem is highly nonlinear and chaotic, and of far too high a dimension to compute even linear solutions, and is therefore totally intractable by any normal measure of tractability.

A possible model

$$
\begin{aligned}
X_0 &\sim N_p(\mu_0, U_0), \\
X_1 &= A_0 X_0 + \epsilon_0,\ \epsilon_0 \sim N_p(0, V_0), \\
Y &= B X_1 + \eta,\ \eta \sim N_q(0, W), \\
X_2 &\sim A_1 X_1 + \epsilon_1,\ \epsilon_1 \sim N_p(0, V_1).
\end{aligned}
$$

B and W may depend on design D.

Comments:

1. BLS didn't have ϵ_0 and ϵ_1 but it costs nothing to include these, and they may be useful, e.g. if we can't in practice handle the full $p = 10^7$ dimensions

2. BLS included nonlinear response terms $F_0(X_0)$, $F_1(X_1)$ instead of A_0X_0, A_1X_1, but it is not clear that this is any better since they were eventually forced to linearize anyway

Solution: $X_1 \sim N_p(\mu_1, U_1)$ where

$$\mu_1 = A_0\mu_0,\ U_1 = V_0 + A_0U_0A_0^T.$$

We then have

$$\begin{aligned}
&Y \sim N_q(B\mu_1, W + BU_1B^T),\\
&X_2 \sim N_p(A_1\mu_1, V_1 + A_1U_1A_1^T),\\
&\mathrm{E}\{(Y - B\mu_1)(X_2 - A_1\mu_1)^T\} = BU_1A_1^T.
\end{aligned}$$

The joint distribution of Y and X_2 is

$$\begin{pmatrix} Y \\ X_2 \end{pmatrix} \sim N_{q+p}\left[\begin{pmatrix} B\mu_1 \\ A_1\mu_1 \end{pmatrix}, \begin{pmatrix} W + BU_1B^T & BU_1A_1^T \\ A_1U_1B^T & V_1 + A_1UA_1^T \end{pmatrix}\right].$$

Hence, the conditional distribution of X_2 given Y has mean

$$A_1\mu_1 + A_1U_1B^T(W + BU_1B^T)^{-1}(Y - B\mu_1),$$

and covariance matrix

$$\begin{aligned} S &= V_1 + A_1U_1A_1^T \\ &\quad - A_1U_1B^T(W + BU_1B^T)^{-1}BU_1A_1^T. \end{aligned}$$

Design problem: write $B = B_D$, $W = W_D$ and hence $S = S_D$ depending on design D, and hence maximize $\Phi(S_D)$ for some suitably chosen design criterion Φ. For example, $\Phi(S) = |S|$ would be D-optimality, $\Phi(S) = \mathrm{tr}(S)$ is A-optimality.

Possibilities for D-optimality: write

$$\begin{aligned} |S| &= |V_1 + A_1U_1A_1^T| \cdot |W + BU_1B^T|^{-1} \cdot \\ &\cdot |W + BU_1B^T - BU_1A_1^T \\ &\quad (V_1 + A_1U_1A_1^T)^{-1}A_1U_1B^T|. \end{aligned} \tag{$*$}$$

Reduces to $q \times q$ determinant.

If in fact Y is $(\, Y_1^T \quad Y_2^T \,)^T$ where Y_1 of dimension $q-d$ is fixed observations and only Y_2 of dimensions d is controlled, write

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

Then

$$W + BU_1B^T =$$

$$\begin{pmatrix} W_{11} + B_1U_1B_1^T & W_{12} + B_1U_1B_2^T \\ W_{21} + B_2U_1B_1^T & W_{22} + B_2U_1B_2^T \end{pmatrix}$$

and hence

$$|W + BU_1B^T| = |W_{11} + B_1U_1B_1^T| \cdot$$

$$\cdot |W_{22} + B_2U_1B_2^T - (W_{21} + B_2U_1B_1^T)$$

$$(W_{11} + B_1U_1B_1^T)^{-1}(W_{12} + B_1U_1B_2^T)|.$$

A similar simplification is possible for the third factor in (*). Hence in practice, the largest determinant we have to evaluate is of a $d \times d$ matrix.

Conclusions on Network Design

1. Max entropy approach is most sophisticated but there are gaps, e.g. no treatment of fully Bayesian hierarchical models and (related) no designs for estimating spatial covariance hyperparameters

2. Design optimality criteria simpler in conception but maybe they have gone further in actual implementation

3. *Ad hoc* approaches may be the most practical

4. Data assimilation problem has many possibilities for incorporating ideas introduced in "network design" contexts.

REFERENCES FOR PART 2

Allen, M.R. and Tett, S.F.B. (1999), Checking for model consistency in optimal fingerprinting. *Climate Dynamics* **15**, 419–434.

Almeida, M.P. and Gidas, B. (1993), A variational method for estimating the parameters of MRF from complete or incomplete data. *Annals of Applied Probability* **3**, 103–136.

Anstreicher, K.M., Fampa, M., Lee, J. and Williams, J. (1996), Using continuous nonlinear relaxations to solve constrained maximum-entropy sampling problems.

Atkinson, A.C. and Donev, A.N. (1992), *Optimum Experimental Designs.*
Oxford University Press.

Barry, R.P. and Ver Hoef, J.M. (1996), Blackbox kriging: spatial prediction without specifying variogram models. *Journal*

of Agricultural, Biological and Environmental Statistics **1** 297–322.

Bartlett, M.S. (1938), The approximate recovery of information from field experiments with large blocks. *J. Agric. Sci.* **28**, 418–427.

Bartlett, M.S. (1955), *An Introduction to Stochastic Processes.* Cambridge University Press, Cambridge.

Bartlett, M.S. (1971), Physical nearest-neighbour models and non-linear time series. *J. Appl. Probab.* **8**, 222–232.

Bartlett, M.S. (1976), *The Statistical Analysis of Spatial Pattern.* Chapman and Hall, London.

Bartlett, M.S. (1978), Nearest neighbour models in the analysis of field experiments (with discussion). *J.R. Statist. Soc. B* **40**, 147–175.

Berliner, L.M., Lu, Z.-Q. and Snyder, C. (1999), Statistical design for adaptive weather observations. *J. Atmos. Sci.* **56**, 2536–2552.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C. and Ghislandi, M. (1995), Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine* **14**, 2433–2443.

Bernardinelli, L., Clayton, D. and Montomoli, C. (1995), Bayesian estimates of disease maps: How important are priors?

Bernardinelli, L. and Montomoli, C. (1992), Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine* **11**, 983–1007.

Bernardinelli, L., Pascutto, C., Best, N.G. and Gilks, W.R. (1997), Disease mapping with errors in covariates. *Statistics in*

Medicine **16**, 741–752.

Bernardo, J.M. (1979), Expected information as expected utility. *Annals of Statistics* **7**, 686–690.

Besag, J.E. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion). *J.R. Statist. Soc. B* **36**, 192–236.

Besag, J.E. (1975), Statistical analysis of non-lattice data. *The Statistician* **24**, 179–195.

Besag, J.E. and Green, P.J. (1993), Spatial statistics and Bayesian computation. *J.R. Statist. Soc. B* **55**, 25–37.

Besag, J.E., Green, P., Higdon, D. and Mengersen, K. (1995), Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Besag, J.E. and Higdon, D. (1999),

Bayesian analysis of agricultural field experiments (with discussion). *J.R. Statist. Soc. B* **61**, 691–746.

Besag, J.E., York, J. and Mollié, A. (1991), Bayesian image restoration, with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.* **43**, 1–59.

Breslow, N.E. and Clayton, D.G. (1993), Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

Brook, D. (1964), On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481–483.

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

Caselton, W.F. and Zidek, J.V. (1984), Optimal monitoring network designs. *Statistics and Probability Letters* **2**, 223–227.

Caselton, W.F., Kan, L. and Zidek, J.V. (1992), Quality data networks that minimize entropy. Chapter 2 of *Statistics in the Environmental and Earth Sciences*, eds. A. Walden and P. Guttorp, Halsted Press, New York, pp. 10-38.

Chaloner, K. and Verdinelli, I. (1995), Bayesian experimental design: A review. *Statistical Science* **10**, 273–304.

Clayton, D.G.and Kaldor, J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671–681.

Clifford, P. (1990), Markov random fields in statistics. In *Disorder in Physical Systems: A Volume in Honour of John M. Hammers-*

ley, eds. G.R. Grimmett and D.J.A. Welsh, Oxford University Press, pp. 19–32.

Cohen, A. and Jones, R.H. (1969), Regression on a random field. *J. Amer. Statist. Assoc* **64**, 1172–1182.

Comets, F. (1992), On consistency of a class of estimators for exponential families of Markov random fields on a lattice. *Annals of Statistics* **20**, 455–468.

Comets, F. and Gidas, B. (1991), Asymptotics of maximum likelihood estimators for the Curie-Weiss model. *Annals of Statistics* **19**, 557–578.

Comets, F. and Gidas, B. (1992), Parameter estimation for Gibbs distributions from partially observed data. *Annals of Applied Probability* **2**, 142–170.

Cook, D.G. and Pocock, S.J. (1983), Multiple regression in geographical mortality

studies, with allowance for spatially correlated errors. *Biometrics* **39**, 361–371.

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

Cressie, N. and Chan, N.H. (1989), Spatial modeling of regional variables. *J. Amer. Statist. Assoc.* **84**, 393–401.

Cressie, N., Kaiser, M.S., Daniels, M.J., Aldworth, J., Lee, J., Lahiri, S.N. and Cox, L.H. (1999), Spatial analysis of particulate matter in an urban environment. In *geoENV II — Geostatistics for Environmental Applications*, eds. Gómez-Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 41–52.

Damian, D., Sampson, P.D. and Guttorp, P. (2001), Bayesian estimation of semi-parametric non-stationary spatial

covariance structures.
Environmetrics **12**, 161–178.

Dickey, J.M. (1967), Matricvariate generalizations of the multivariate t distribution. *Ann. Math. Statist.* **38**, 511–518.

Diggle, P.J., Tawn, J. and Moyeed, R.A. (1998), Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.

Donnelly, C.A. (1995), The spatial analysis of covatriates in a study of environmental epidemiology. *Statistics in Medicine* **14**, 2393–2409.

Donnelly, C.A., Ware, J.H. and Laird, N.M. (1994), Regression analysis fo spatially correlated data: The Kanawha County health study. In *Handbook of Statistics, Vol. 12: Environmental Statistics*, G.P. Patil and C.R. Rao (eds), North Holland Publishing Company, pp. 643–660.

Fedorov, V.V. (1972), *Theory of Optimal Experiments*. Academic Press, New York.

Fedorov, V. and Müller, W. (1989), Comparison of two approaches in the optimal design of an observation network. *Statistics* **20**, 339-351.

Fuentes, M. (2002), Spectral methods for nonstationary spatial processes.
Biometrika **89**, 197–210.

Fuentes, M. and Smith R.L. (2001), A new class of nonstationary spatial models.
Preprint, NCSU and UNC.

Furnival, G.M. and Wilson, R.W. Jr.(1974), Regression by leaps and bounds., *Technometrics* **16**, 499–511.

Gaver, D. and O'Muircheartaigh, I. (1987), Robust empirical Bayes analysis of event rates. *Technometrics* **29**, 1–15.

Gelfand, A.E. and Smith, A.F.M. (1990), Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 398-409.

Geman, S. and Geman, D. (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C.J. (1992), Practical Markov chain Monte carlo. *Statistical Science* **7**, 473–482.

Geyer, C.J. and Thompson, E.A. (1992), Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J.R. Statist. Soc. B 54*, 657–699.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (eds.) (1996), *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Guttorp, P., Meiring, W. and Sampson, P. (1994), A space-time analysis of ground-level ozone data. *Environmetrics* **5**, 241–254.

Guyon, X. (1982), Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika* **69**, 95–105.

Haas, T.C. (1990), Lognormal and moving-window methods of estimating acid deposition. *J. Amer. Statist. Assoc.* **85**, 950–963.

Haas, T.C. (1992), Redesigning continental-scale monitoring networks. *Atmos. Environment* **26A**, 3323–3333.

Haas, T.C. (1995), Local prediction of a spatio-temporal process with an application to wet sulfate deposition. *J. Amer. Statist. Assoc.* **90**, 1189–1199.

Haas, T.C. (1998), Statistical assessment of

spatio-temporal pollutant trends and meteorological transport models. *Atmospheric Environment* **32**, 1865–1879.

Haas, T.C. (2002), New systems for modeling, estimating and predicting a multivariate spatio-temporal process. *Environmetrics* **13**, 311–332.

Hasselmann, K. (1997), Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dynamics* **13**, 601–611.

Hastings, W.K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hegerl, G.C. and North, G.R. (1997), Comparison of statistical optimal approaches to detecting anthropogenic climate change. *Journal of Climate* **10**, 1125–1133.

Higdon, D. (1998), A process-convolution

approach to modeling temperatures in the north Atlantic Ocean. *J. Environ. Ecolo. Statist.* **5**, 173–190.

Higdon, D. (2001), Space and space-time modeling using process convolutions. Preprint, Duke University.

Higdon, D., Swall, J. and Kern, J. (1999), Non-stationary spatial modeling. In *Bayesian Statistics 6*, eds. J.M. Bernardo *et al.*, Oxford University Press, pp. 761–768.

Higham, N.J. (1988), Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications* **103**, 103–118.

Holland, D., Saltzman, N., Cox, L.H. and Nychka, D. (1999), Spatial prediction of sulfur dioxide in the eastern United States. In *geoENV II — Geostatistics for Environmental Applications*, eds. Gómez-

Hernández, J., Soares, A. and Froidevaux, R., Kluwer, Dordrecht, 65–76.

Jaynes, E.T. (1963), Information theory and statistical mechanics. In *Statistical Physics*, Vol. 3, K.W. Ford (ed.), Benjamin, New York, pp. 102–218.

Johnson, N.L. and Kotz, S. (1972), *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

Kiefer, J. and Wolfowitz, J. (1960), The equivalence of two extremum problems. *Canad. J. Math.* **12**, 363–366.

Ko, C.-W., Lee, J. and Queyranne, M. (1995), An exact algorithm for maximum entropy sampling. *Oper. Res.* **43**, 684–691.

Krewski, D., Burnett, R.T., Goldberg, M.S., Hoover, K., Siemiatycki, J., Jerrett, M., Abrahamowicz, M. and White, W.H.

(2000), *Reanalysis of the Harvard Six Cities Study and the American Cancer Society Study of Particulate Air Pollution and Mortality.* A Special Report of the Institute's Particulate Epidmiology Reanalysis Project. Health Effects Institute, Cambridge, MA.

Lark, R. (2002), Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma* **105**, 49–80.

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.

Le, N.D., Sun, W. and Zidek, J.V. (1997), Bayesian multivariate spatial interpolation with data missing by design. *J.R. Statist. Soc. B* **59**, 501–510.

Lindley, D.V. (1956), On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27**, 986–1005.

Mardia, K.V. and Goodall, C.R. (1993), Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao, Elsevier Science Publishers, pp. 347–386.

Meiring, W., Guttorp, P. and Sampson, P.D. (1998), Space-time estimation of grid-cell hourly ozone levels for assessment of a determinstic model. *Environmental and Ecological Statistics* **5**, 197–222.

Mercer, W.B. and Hall, M.A. (1911), The experimental error of fields trials. *J. Agric. Sci.* **4**, 54–62.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E.

(1953), Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Müller, P. (1999), Simulation based optimal design (with discussion). In *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, 459–474.

Müller, W.V. (2000), *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Second edition, Physica Verlag, Heidelberg.

Müller, W.V. and Zimmerman, D.L. (1999), Optimal designs for variogram estimation. *Environmetrics* **10**, 23–37.

North, G.R. (1984), Empirical orthogonal functions and normal modes. *J. Atmos. Sci.* **41**, 879–887.

Nychka, D. and Saltzman, N. (1998),

Design of air quality networks. In *Case Studies in Environmental Statistics*, eds. D. Nychka, W. Piegorsch and L.H. Cox, Lecture Notes in Statistics number 132, Springer Verlag, New York, pp. 51–76.

Nychka, D., Wikle, C. and Royle, J.A. (1999), Large spatial prediction problems and nonstationary random fields. Preprint, Geophysical Statistical Program, National Center for Atmospheric Research.

Oehlert, G.W. (1993), Regional trends in sulfate wet deposition. *J. Amer. Statist. Assoc.* **88**, 390–399.

Oehlert, G.W. (1995), The ability of wet decomposition networks to detect temporal trends. *Environmetrics* **6**, 327–339.

Oehlert, G.W. (1996), Optimal shrinking of a wet decomposition network. *Atmospheric Environment* **30**, 1347–1357.

Papadakis, J.S. (1937), Méthode statistique pour les expériences du champ. *Bull. Inst. Amél. Plantes à Salonique*, No. 23.

Penttinen, A. (1984), Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood method. *Jy. Stud. Comput. Sci. Econ. Statist.* **7**.

Sampson, P.D. and Guttorp, P. (1992), Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87**, 108-119.

Schmidt, A.M. and O'Hagan, A. (2000), Bayesian inference for nonstationary spatial covariance structure via spatial deformations. Preprint, University of Sheffield.

Shapiro, A. and Botha, J.D. (1991), Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.

Sansó, B. and Müller, P. (1997), Redesigning a network of rainfall stations. ISDS Discussion Paper 97-25, Duke University.

Silvey, S.D. (1980), *Optimal Design.* Chapman and Hall, London.

Smith, R.L. (1996), Estimating nonstationary spatial correlations. Preprint, University of North Carolina.

Su, Y. and Cambanis, S. (1993), Sampling designs for estimation of a random process. *Stochastic Processes and Their Applications* **46**, 47–89.

Sun, L., Zidek, J.V., Le, N.D. and Özkaynak, H. (2000), Interpolating Vancouver's daily ambient PM_{10} field. *Environmetrics* **11**, 651–663.

Swendsen, R.H. and Wang, J.-S. (1987), Nonuniversal critical dynamics in Monte

Carlo simulations. *Phys. Rev. Letters* **58**, 86–88.

Tibshirani, R. (1995), Regression selection and shrinkage via the *lasso*. *J.R. Statist. Soc. B* **58**, 267–288.

Van Groenigen, J.W. and Stein, A. (1998), Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality* **43**, 684–691.

Ver Hoef, J.M. and Barry, R.P. (1999), Constructing and fitting models for cokriging and multivariable spatial prediction. *J. Statist. Plann. Inference* **69**, 275–294.

Ver Hoef, J.M., Cressie, N. and Barry, R. (2000), Flaxible spatial models based on the fast Fourier transform (FFT) for cokriging. Technical report, Department of Statistics, Ohio State University.

Waller, L.A., Carlin, B.P., Xia, H. and Gelfand, A.E. (1997), Hierarchical spatio-temporal mapping of disease rates. *J. Amer. Statist. Assoc.* **92**, 607–617.

Warrick, A.W. and Myers, D.E. (1987), Optimization of sampling locations for variogram calculations. *Water Resources Research* **23**, 496–500.

Whittle, P. (1954), On stationary processes in the plane. *Biometrika* **41**, 434–439.

Wu, S. and Zidek, J.V. (1992), An entropy-based analysis of data from selected NADP/NTN network sites for 1983–1986. *Atmos. Environment* **26A**, 2089–2103.

Zhu, L., Carlin, B.P., English, P. and Scalf, R. (2000), Hierarchical modeling of spatio-temporally misaligned data: Relating traffic density to pediatric asthma hospitalizations. *Environmetrics* **11**, 43–61.

Zhu, Z. (2002), Optimal sampling design and parameter estimation of Gaussian random fields. PhD Thesis, Department of Statistics, University of Chicago.

Zidek, J.V., Sun, W. and Le, N.D. (2000), Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Applied Statistics* **49**, 63–79.

Zwiers, F. and von Storch, H. (1995), Taking serial correlation into account in tests of the mean. *Journal of Climate* **8**, 336–351.