Karhunen-Loeve Expansion and Optimal Low-Rank Model for Spatial Processes

Hao Zhang

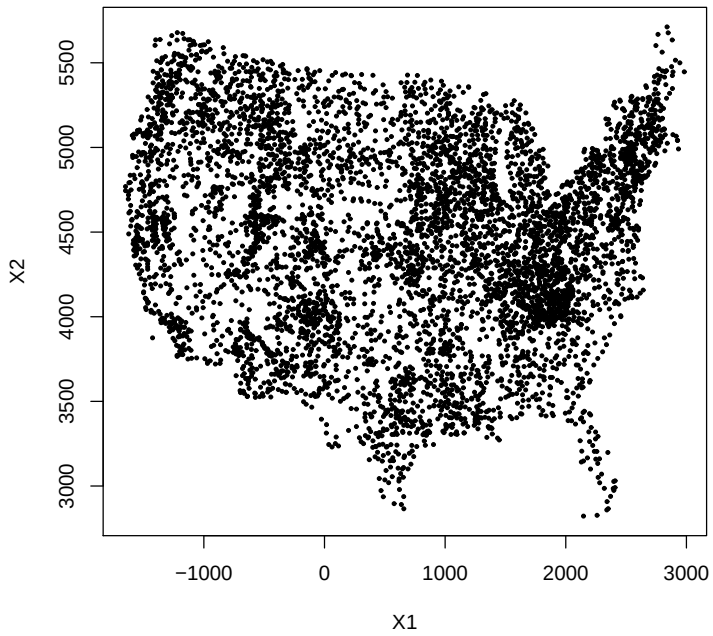
Department of Statistics
Department of Forestry and Natural Resources
Purdue University
zhanghao@purdue.edu

Outline

- Motivations of studying KL expansion in spatial statistics
- Numerical algorithms
- Some examples and results
- Applications in multivariate spatial processes

Spatial data

- $Y(\boldsymbol{s}_i), \boldsymbol{s}_i \in D \subset R^2$ or R^3.
- Correlated
- Massive
  - Satellite data
  - Remotely sensed data
  - Censor network
  - Multi-source
  - . . . . . .



Spatial process

Review the spatial data as a realization of an underlying spatial process:

$$\text{data} = \text{mean} + \text{error},$$

- mean function=deterministic, linear function of some covariates (i.e., lat and long).
- error process=2nd order process, often assumed to be stationary.
- Spatial correlation attributable to unobserved latent variables; accounting for the spatial variation unaccounted for by the mean.

I assume the mean is a constant (known or unknown).

Covariogram and Kriging

Covariogram:

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = Cov(Y(\boldsymbol{s}_1), Y(\boldsymbol{s}_2)).$$

Stationary if

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\mathbf{0}, \boldsymbol{s}_2 - \boldsymbol{s}_1).$$

Isotropic if

$$C(\boldsymbol{s}_1, \boldsymbol{s}_2) = C(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|).$$

Kriging=Best Linear Prediction:

Given $Y(\boldsymbol{s}_1), \cdots, Y(\boldsymbol{s}_n)$,

$$\begin{aligned} \hat{Y}(\boldsymbol{s}_0) &= E(Y(\boldsymbol{s}_0)) + \sum_{i=1}^{n} \alpha_i (Y(\boldsymbol{s}_i) - EY(\boldsymbol{s}_i)) \\ &= E(Y(\boldsymbol{s}_0)) + Cov(Y(\boldsymbol{s}), \boldsymbol{Y})[Var(\boldsymbol{Y})]^{-1}(\boldsymbol{Y} - E\boldsymbol{Y}). \end{aligned}$$

Large covariance matrix

— n is hundreds of thousands or millions.

— Data are correlated

— The inverse of $V_{n\times n}$ is needed in the full MLE or Bayesian inferences, and kriging, rending the traditional methods impractical:

$$L_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V_n(\theta)| - \frac{1}{2}(\boldsymbol{Y}_n - X\beta)'V_n^{-1}(\theta)(\boldsymbol{Y}_n - X\beta),$$

- The inverse matrix is needed for prediction.

Large storage and memory

- For an $n \times n$ matrix, the double precision storage is $8n^2$ bites.
  - When $n = 10^4$, the storage is 800MB!
  - When $n = 10^6$, the storage is 8000GB!
- Computation for $V_{n\times n}^{-1}$ is $O(n^3)$ unless $V_{n\times n}$ has a special structure.
- Therefore, for massive spatial data, traditional geostatistical methods may not be applicable and innovative methods are need.

Low rank models

$$Y(\boldsymbol{s}) = \mu + \boldsymbol{a}(\boldsymbol{s})'\boldsymbol{Z} + \epsilon(\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{a}(\boldsymbol{s}) = (a_1(\boldsymbol{s}), \cdots, a_m(\boldsymbol{s}))'$ is a function of $\boldsymbol{s}$ that may depend on some parameters θ, $\boldsymbol{Z} = (Z(\boldsymbol{u}_1), \cdots, Z(\boldsymbol{u}_m))'$ is a vector of latent variables with $\boldsymbol{u}_i$'s pre-chosen, and $\epsilon(\boldsymbol{s})$ is a white noise with variance τ^2 and is independent of $\boldsymbol{Z}$.

- Gaussian process convolution:
  $a_j(\boldsymbol{s}) = \exp(-\|\boldsymbol{s} - \boldsymbol{u}_j\|^2/\phi)/\sqrt{2\pi}$, $Z(\boldsymbol{u}_i)$'s are i.i.d.
- Fixed rank kriging: $a_j(\boldsymbol{s}) = b(\|\boldsymbol{s} - \boldsymbol{u}_j\|^2/\phi)$ for some known basis function b and $\boldsymbol{Z} = (Z(\boldsymbol{u}_1), \cdots, Z(\boldsymbol{u}_m))'$ is a realization of a latent process at locations $\boldsymbol{u}_j$, $j = 1, \cdots, m$.
- Predictive process: $\boldsymbol{a}(\boldsymbol{s}) = Var(\boldsymbol{Z})^{-1}Cov(\boldsymbol{Z}, Z(\boldsymbol{s}))$ where $Z(\boldsymbol{s})$ is a Gaussian stationary process with mean 0 and some parametric covariogram.

Covariance matrix of low rank

- Observing $\boldsymbol{Y} = (Y(\boldsymbol{s}_i), i = 1, \cdots, n)'$ from the low rank model, write

$$\boldsymbol{Y} = \mu \mathbf{1} + A\boldsymbol{Z} + \boldsymbol{\epsilon}.$$

- The covariance matrix of $\boldsymbol{Y}$ is

$$V = AV_zA' + \tau^2 I_n,$$

  where A is $n \times m$ and $V_z = Var(\boldsymbol{Z})$.

- Woodbury Formula:

$$V^{-1} = \tau^{-2}(I_n - A(\tau^2 V_z^{-1} + A'A)^{-1}A').$$

  Memory requirement is $8(n + nm + m^2)$ and linear in n.

- Sylvester's determinant theorem:

$$|V| = \tau^{2(n-m)}|V_z||\tau^2 V_z^{-1} + A'A|.$$

Optimal low rank models

Given a random field $Y(\boldsymbol{s}), \boldsymbol{s} \in D$, consider the best low rank model, i.e., to minimize

$$\int_D E\left(Y(\boldsymbol{s}) - \sum_{k=1}^{m} a_k(\boldsymbol{s}) Z_k\right)^2 d\boldsymbol{s},$$

over all functions $a_k(\boldsymbol{s})$ and random variables Z_k.

Solution: $a_k(\boldsymbol{s}) = \sqrt{\lambda_k} f_k(\boldsymbol{s})$ where λ_k and $f_k(\boldsymbol{s})$, $k = 1, \cdots, m$ are the first m largest eigenvalues and the corresponding eigenfunctions of the Karhunen-Loeve expansion of the random field.

K-L expansion

$$Y(\boldsymbol{s}) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} f_k(\boldsymbol{s}) Z_k, \boldsymbol{s} \in D$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, Z_k are white noises of unit variance and

$$\int_D f_j(\boldsymbol{s}) f_k(\boldsymbol{s}) d\boldsymbol{s} = \delta_{jk}.$$

Hence

$$C(\boldsymbol{s}, \boldsymbol{x}) = \sum_{k=1}^{\infty} \lambda_k f_k(\boldsymbol{s}) f_k(\boldsymbol{x})$$

and

$$\int_D C(\boldsymbol{s}, \boldsymbol{x}) f_k(\boldsymbol{x}) d\boldsymbol{x} = \lambda_k f_k(\boldsymbol{s}).$$

Truncated KL

$$\int_D E\left\{Y(\boldsymbol{s}) - \sum_{k=1}^{m} \sqrt{\lambda_k} f_k(\boldsymbol{s}) Z_k\right\}^2 d\boldsymbol{s}$$

$$\leq \int_D E\left\{Y(\boldsymbol{s}) - \sum_{k=1}^{m} a_k(\boldsymbol{s}) Z_k\right\}^2 d\boldsymbol{s}$$

for any functions $a_k(\boldsymbol{s})$ and zero-mean random variables Z_k where D is bounded.

Computation of eigenpairs

Given a covariance function $C(\boldsymbol{s}, \boldsymbol{x}), \boldsymbol{x}, \boldsymbol{s} \in D \subset R^d$, find $\lambda > 0$ and function $f(\boldsymbol{x})$ such that

$$\int_D C(\boldsymbol{s}, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \lambda f(\boldsymbol{s}).$$

Approximation:

- Choose a basis $\{\phi_i, i = 1, \cdots, m\} \subset L^2(D)$.
- Approximate an eigenfunction by

  $$f(\boldsymbol{s}) = \sum_{i=1}^m d_i \phi_i(\boldsymbol{s}),$$

  so that $\int_D C(\boldsymbol{s}, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} - \lambda f(\boldsymbol{s})$ is orthogonal to each basis function $\phi_i(\boldsymbol{x})$.

Galerkin approximation

Finding λ and $\boldsymbol{d}$:

- $$\int_D \left(\sum_{i=1}^m d_i \int_D C(\boldsymbol{s}, \boldsymbol{x}) \phi_i(\boldsymbol{s}) d\boldsymbol{x} - \lambda \sum_{i=1}^m d_i \phi_i(\boldsymbol{s}) \right) \phi_j(\boldsymbol{s}) d\boldsymbol{s} = 0, \quad \text{for } j = 1, \cdots, m,$$

- $$G\boldsymbol{d} = \lambda B \boldsymbol{d}, \tag{2}$$

  where G is $m \times m$ whose (i, j) element is $\int_D \int_D \phi_i(\boldsymbol{s}) C(\boldsymbol{s}, \boldsymbol{x}) \phi_j(\boldsymbol{x}) d\boldsymbol{s} d\boldsymbol{x}$, B is $m \times m$ whose (i, j) element is $\int_D \phi_i(\boldsymbol{s}) \phi_j(\boldsymbol{s}) d\boldsymbol{s}$, $\boldsymbol{d} = (d_1, \cdots, d_m)'$.

- Solve the generalized eigenvalue problem (2) to get λ and $\boldsymbol{d}$.
- The m^2 elements of G are each 4-dimensional integrals.

Ghanem and Spanos, 1991, Schwab and Todor, 2005, Phoon, et al. 2002, Huang, et al., 2001

Gaussian Quadrature Method

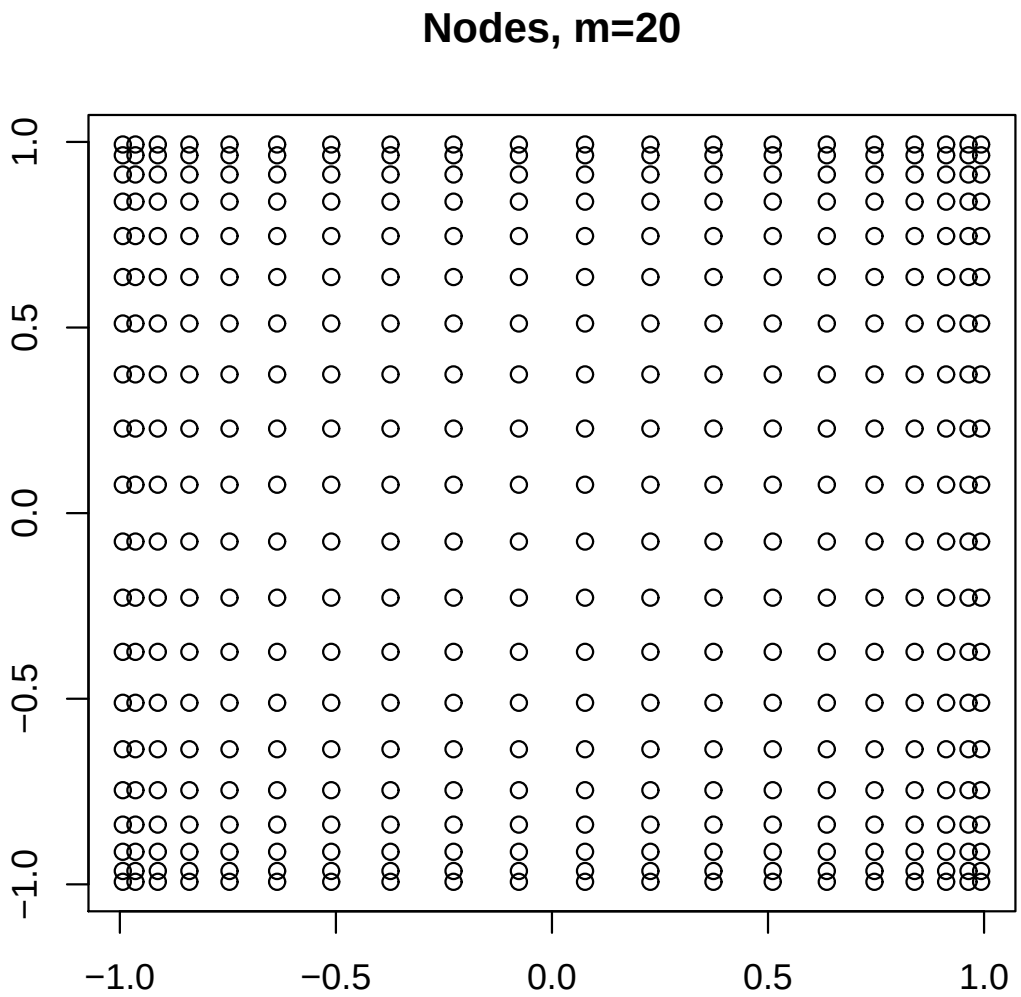
Gaussian Legendre quadrature:

$$\int_{-1}^{1} g(x)dx \approx \sum_{i=1}^{m} r_i g(x_i).$$

- The approximation is exact for polynomials of order $(2m-1)$ or less.
- The nodes x_i are the roots of Legendre polynomial of m order and the weights r_i also calculated from the Legendre polynomial.
- Extends to high dimension:

$$\int_{-1}^{1}\int_{1}^{1} g(y_1,y_2)dy_1dy_2 \approx \sum_{i=1}^{m}\sum_{j=1}^{m} r_i r_j g(x_i,x_j)$$

Example of Nodes



Gaussian Quadrature Method for KL

Given $C(\boldsymbol{s}, \boldsymbol{x})$, $\boldsymbol{s}, \boldsymbol{x} \in [-1, 1]^d$ ($d = 2$), find λ and $f(\boldsymbol{x})$ such

$$\int_{[-1,1]^d} C(\boldsymbol{s}, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} = \lambda f(\boldsymbol{x}).$$

Let u_i, $i = 1, \ldots, m$ be the m roots of Legendre polynomial and $\boldsymbol{u}_i$, $i = 1, \ldots, m^2$ denotes the pairs (u_j, u_k), $j, k = 1, \ldots, m$.
Find $\boldsymbol{f} = (f(\boldsymbol{u}_1), \ldots, f(\boldsymbol{u}_{m^2}))'$ by solving

$$\sum_{j=1}^{m^2} w_j C(\boldsymbol{u}_i, \boldsymbol{u}_j) f(\boldsymbol{u}_j) = \lambda f(\boldsymbol{u}_i)$$

where $w_j = u_k u_l$ if $\boldsymbol{u}_j = (u_k, u_l)$.

Gaussian Quadrature Method for KL

Write

$$V = (C(\boldsymbol{u}_i, \boldsymbol{u}_j)), W = \text{diag}(w_1, \ldots, w_{m^2}).$$

Then

$$VW\boldsymbol{f} = \lambda \boldsymbol{f}$$

A generalized eigenvalue problem. Hence λ is an eigenvalue of

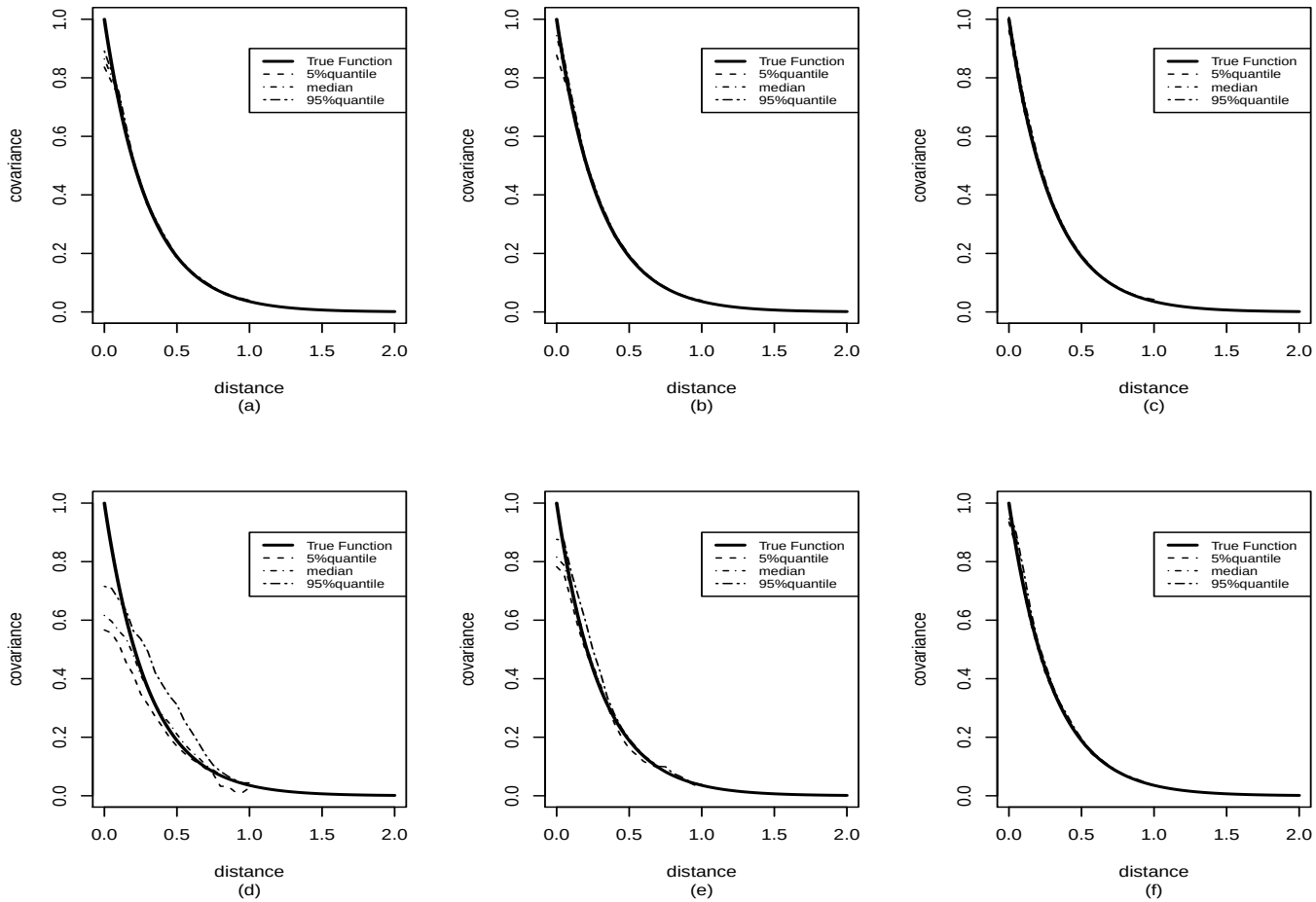
$$W^{1/2} V W^{1/2}$$

and $W^{1/2}\boldsymbol{f}$ is the corresponding eigenvector.

Polynomial Interpolation

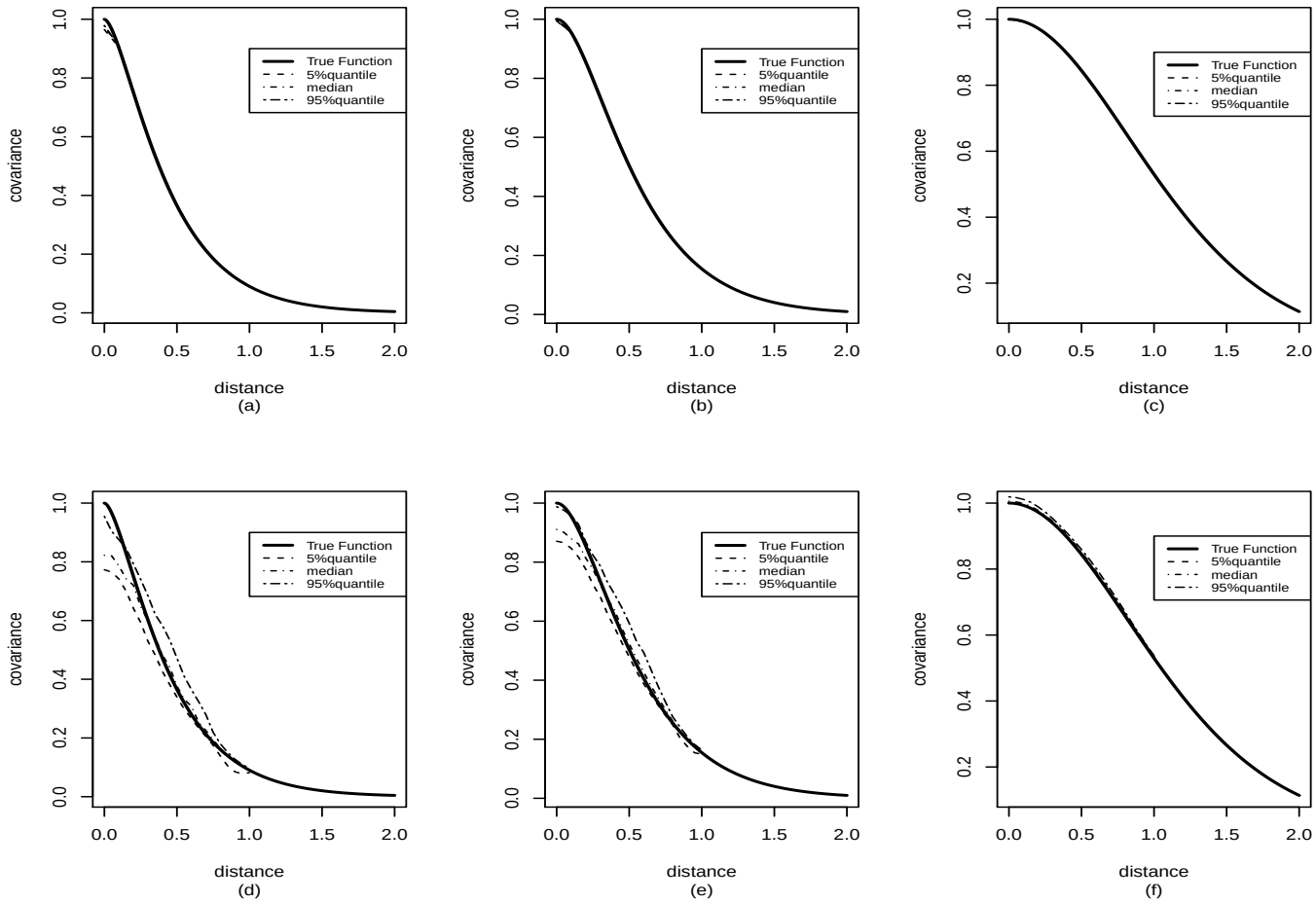
Having obtained $\boldsymbol{f}$, we now have $f(\boldsymbol{u}_i)$. Now approximate the eigenfunction by a polynomial through the Lagrange interpolation.

Example 1



Top: Gaussian quadrature; Bottom: Galerkin projection.

Example 2



Top: Gaussian quadrature; Bottom: Galerkin projection.

Simulation study

Objective: To compare the predictive performance of the different low-rank approximations

True model: Gaussian process $Y(\boldsymbol{s}), \boldsymbol{s} \in [0,1]^2$ with mean 0 and exponential covariance function $\exp(-h/0.3)$.

We use three low-rank models to approximate the true model: KL by Galerkin projection (GPKL), KL by Gaussian quadrature (GQKL) and the predictive process (Banerjee, et al., 2008).

Simulate from the *true* model on a 81*81 lattice, but use each of the low rank model to estimate model parameters.

Predictive Performance

MSE:

$$(1/n)\sum(Y(\boldsymbol{s}_i) - \hat{Y}(\boldsymbol{s}_i))^2.$$

Logarithmic score (LogS): Let $Z_i = (Y(\boldsymbol{s}_i) - \hat{Y}(\boldsymbol{s}_i))/\sigma_{-i}(\boldsymbol{s}_i)$.

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\log(2\pi\hat{\sigma}_{-i}(\boldsymbol{s}_i)^2) + \frac{1}{2}Z_i^2\right)$$

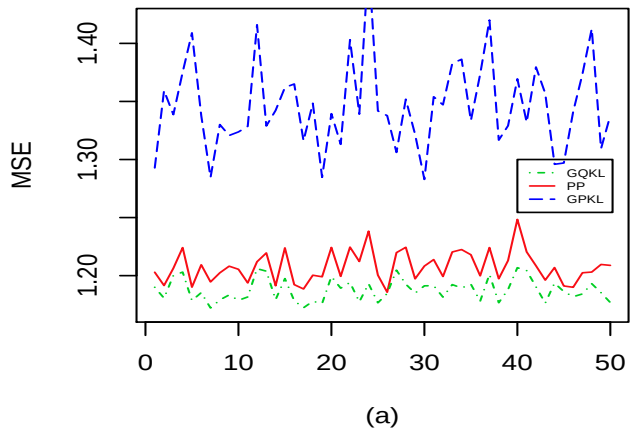
Continuous ranked probability score:

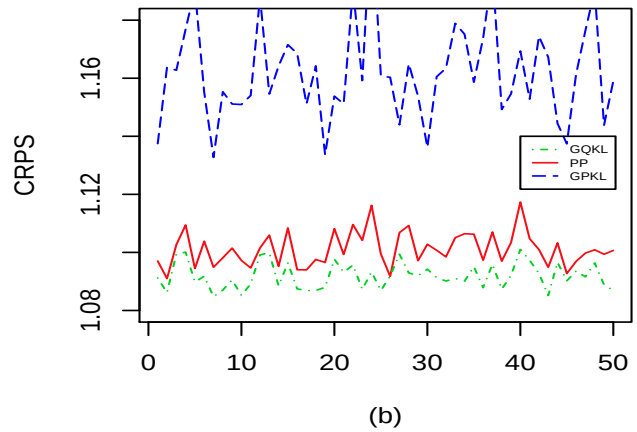
$$crps(F,x) = \int_{-\infty}^{\infty}\left(F(y) - 1\{y \geq x\}\right)^2 dy.$$

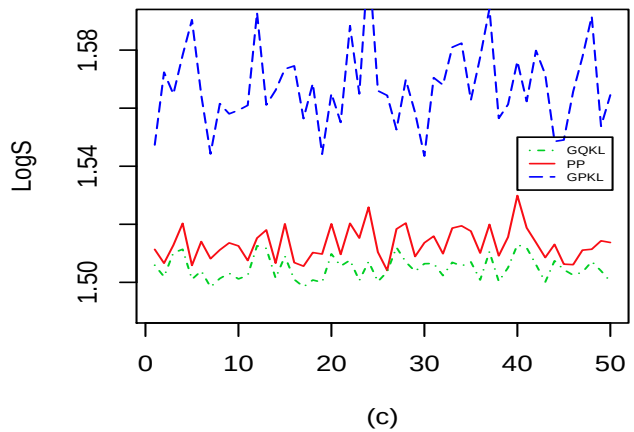
$$CRPS = \frac{1}{n}\sum_{i=1}^{n} crps(F_i, Y(\boldsymbol{s}_i)).$$

Gneiting, et al., 2006, 2007.

150 eigen terms

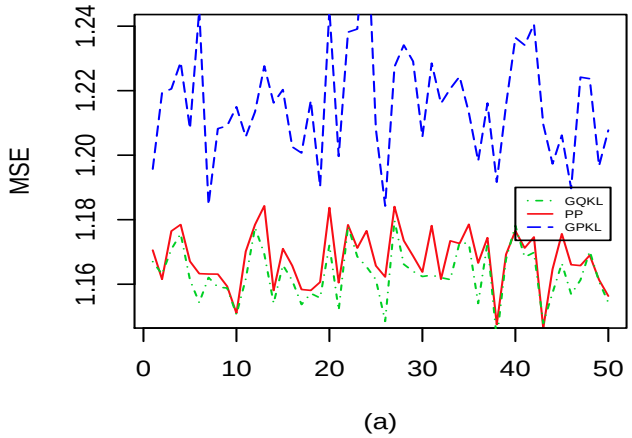


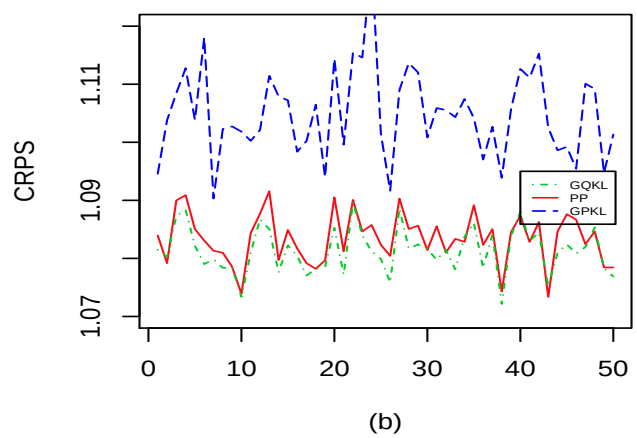


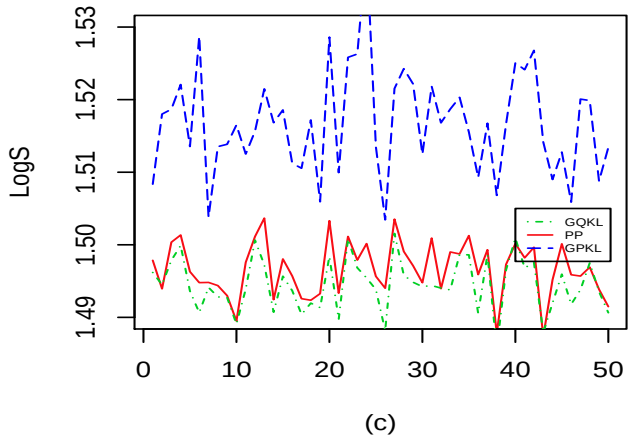


(a) MSE scores(b) CRPS scores(c) LogS scores using 150 eigen terms

300 eigen terms

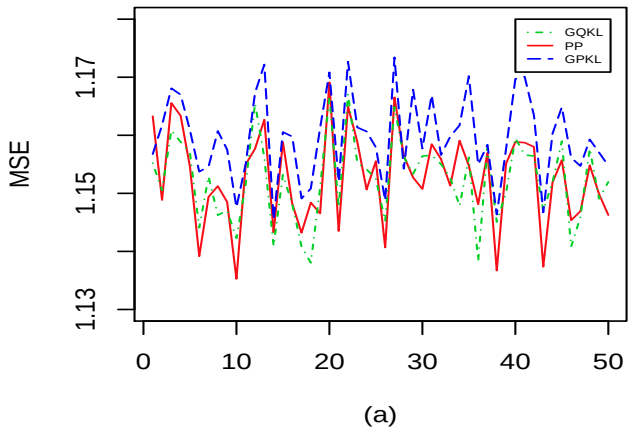


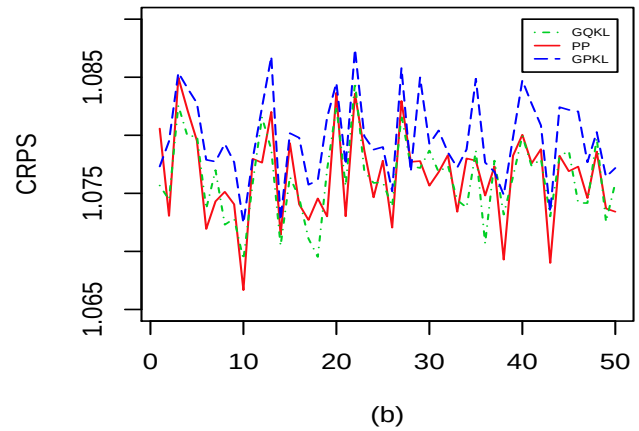


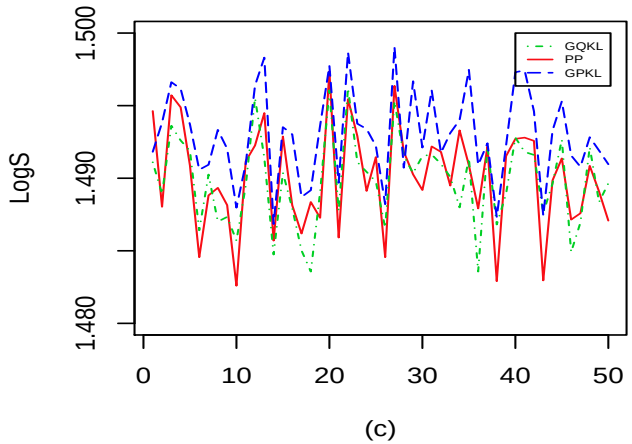


(a) MSE scores(b) CRPS scores(c) LogS scores using 300 eigen terms

600 eigen terms







(a) MSE scores(b) CRPS scores(c) LogS scores using 300 eigen terms

Multivariate covariogram

Two processes $Y_i(\boldsymbol{s})$, $i = 1, 2$.

- Marginal or direct covariogram: $C_{ii}(\boldsymbol{s}, \boldsymbol{x}) = Cov(Y_i(\boldsymbol{s}), Y_i(\boldsymbol{x}))$.
- Cross covariogram: $C_{ij}(\boldsymbol{s}, \boldsymbol{x}) = Cov(Y_i(\boldsymbol{s}), Y_j(\boldsymbol{x}))$, $i \neq j$.
- The matrix valued function $C(\boldsymbol{s}, \boldsymbol{x}) = (C_{ij}(\cdot, \cdot))$ must satisfy: For any $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$, $\sum_i \sum_j \alpha_i' C(\boldsymbol{s}_i, \boldsymbol{s}_j) \alpha_j \geq 0$ for any vectors $\alpha_i \in R^p$.
- It is not trivial to specify a cross covariogram.

Some existing models

- Proportional model: $C_{ij}(\boldsymbol{s}, \boldsymbol{x}) = \sigma_{ij}\rho(\boldsymbol{s}, xx)$.
- Linear model of coregionalization: Sum of proportional models.
- Convolution models: Gaspari and Cohn (1999), Gelfand et al. (2004)
- Asymmetric models: Apanasovich and Genton (2009), Li and Zhang (2010)
- Multivariate Matern model: Gneiting et al. (2010)

Limitation of existing models

- Estimation of parameters becomes more complex; constrained maximization.
- Do not allow arbitrary marginal covariogram models.
- There are cases when the multivariate covariogram model does not improve over the univariate model in predictive performance.

Valid cross covariogram for given marginals

The problem: Given two marginal covariograms $C_{ii}(\boldsymbol{s}, \boldsymbol{x})$, $\boldsymbol{s}, \boldsymbol{x} \in R^d$, $i = 1, 2,$, find the class of all valid cross covariograms.

A theoretical solution: Apply the Karhunen-Loeve expansion to the marginals to write

$$C_{ii}(\boldsymbol{s}, \boldsymbol{x}) = \sum_{k=1}^{\infty} \lambda_{ik} f_{ik}(\boldsymbol{s}) f_{ik}(\boldsymbol{x}), \boldsymbol{s}, \boldsymbol{x} \in D$$

where λ_k is a decreasing sequence of positive numbers and $\int_D f_{ik}(\boldsymbol{s})^2 d\boldsymbol{s} = 1$ and D is a compact subset of R^d. The class of all possible valid cross covariograms is

$$C_{ij}(\boldsymbol{s}, \boldsymbol{x}) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} r_{kl} \sqrt{\lambda_{ik} \lambda_{jl}} f_{ik}(\boldsymbol{s}) f_{jl}(\boldsymbol{x})$$

where $r_{kl} \in [-1, 1]$ are such that the matrix $(r_{kl})_{k,l=1}^{n}$ is non-negative definite for any n. Hence all cross covariograms are given by such an array $\{r_{kl}\}$.

Non-parametric cross covariogram

Consider a bivariate process. Write the KL expansion for each process

$$Y_i(\boldsymbol{s}) = \sum_{k=1}^{\infty} \sqrt{\lambda_{ik}} f_{ik}(\boldsymbol{s}) Z_{ik}, s \in D,$$

where $Z_{ik}, k = 1, 2, \cdots$ are a white noise $(i = 1, 2)$. In a special case, when Z_{1j} and Z_{2k} are either uncorrelated or perfectly correlated, we get the non-parametric cross covariogram

$$C_{12}(\boldsymbol{s}, \boldsymbol{x}) = \sum_{j,k:Cov(Z_{1j},Z_{2k})=1} \sqrt{\lambda_{1j}\lambda_{2k}} f_{1j}(\boldsymbol{s}) f_{2k}(\boldsymbol{x}).$$

Two practical steps

For the non-parametric cross covariogram to work in practice, we need

- An efficient algorithm to computer the eigenpairs $(\lambda_j, f_j(\boldsymbol{s}))$ for any given covariogram.
- Find the pairs (j, k) such that $Cov(Z_{1j}, Z_{2k}) = 1$.

Inferences

- Estimate parameters in each marginal model using only marginal data;
- Since cross covariogram involves no new parameters, we only need to identify the perfectly correlated pairs of Z_{1j} and Z_{2k}.
  - This is equivalent to specifying the correlation matrix $Corr(Z_1, Z_2)$ for $Z_i = (Z_{i1}, \cdots, Z_{im})'$ which has at most m 1s with the other elements being 0.
  - Start with $Corr(Z_1, Z_2) = 0$.
  - Replace one 0 element with 1 to optimize some objective function (i.e., likelihood, prediction score)
  - Fixing the non-zero elements, replace one more 0 element with 1 to optimize the objective function.
  - Iterate.
- The non-parameter cross covariogram leads to a much simplified estimation procedure.

An example

Wang and Zhang (2012): Let $W_i(\boldsymbol{s})$, $\boldsymbol{s} \in [0,1]^2$ be two latent processes with mean 0 and have a bivariate exponential covariogram, i.e.,

$$Cov(W_i(\boldsymbol{s}+\boldsymbol{h}), W_j(\boldsymbol{s})) = \sigma_{ij} \exp(-\|\boldsymbol{h}\|/\phi_{ij}).$$

We observe the following process

$$Y_i(\boldsymbol{s}) = W_i(\boldsymbol{s}) + \tau_i \epsilon_i(\boldsymbol{s})$$

at the 1225 equally spaced locations $\{(i/35, j/35), i, j = 1, \cdots, 35\}$. The model parameters for the covariogram are chosen to be: $\phi_{11} = 0.2, \phi_{22} = 0.3, \phi_{12} = 0.2, \sigma_{12} = 5.4$, $\sigma_{11} = \sigma_{22} = 9$ and $\tau_1^2 = \tau_2^2 = 1$.

Model fitting

Fit two models: The true model and the truncated KL expansion with m=600 terms.

For the true model, we estimate the parameters by maximizing the joint likelihood. For the truncated KL model, we estimate the marginal parameters by maximizing the marginal likelihood and identify the perfectly correlated pairs by maximizing the joint likelihood.

Predictive performance

Predictive scores: RMSE and logarithmic score:

$$MSE^{(i)} = \frac{1}{n}\sum_{k=1}^{n}\left(Y_i(\boldsymbol{s}_k) - \hat{Y}_i(\boldsymbol{s}_k)\right)^2$$

$$LogS^{(i)} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{2}\log(2\pi\hat{\sigma}_{ij}^2) + \frac{1}{2}z_{ij}^2\right), i = 1, 2$$

where $\hat{Y}_i(\boldsymbol{s}_k)$ denotes the cokriging prediction of $Y_i(\boldsymbol{s}_k)$ using data at all locations but $\boldsymbol{s}_k$; σ_{ik} the corresponding cokriging standard deviation and $z_{ik} = (Y_i(\boldsymbol{s}_k) - \hat{Y}_i(\boldsymbol{s}_k))/\sigma_{ik}$.

Among 100 simulations, the non-parametric cross covariogram model beats the fitted bivariate parametric model in each simulation according to the two predictive scores.