ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Presentation for GSP/NCAR, February 2002

Covariance Estimation of Geostatistical Data

Reinhard Furrer, IMA, SB, EPF-Lausanne

Outline of the Presentation

- Motivation
- The estimator $\hat{\mathbf{U}}$
- Asymptotic considerations
- Eigenvalues and eigenvectors of $\hat{\mathbf{U}}$
- Simulations
- Application to ozone data
- Conclusion and outlook

Motivation

Define a spatio-temporal process

$$\left\{Z(\mathbf{x},t) : \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^d, d \geq 1, t \in \mathcal{T}\right\}$$

with covariograms

$$c\big(\mathbf{x}_i - \mathbf{x}_j, t_r - t_s\big) = \text{Cov}\big(Z(\mathbf{x}_i,t_r), Z(\mathbf{x}_j,t_s)\big) = (\mathbf{C}_{rs})_{ij}$$

We aim to estimate

$$\text{Var}\big(Z(\mathbf{x},t)\big) = \mathbf{\Sigma} = (\sigma_{rs}) = \big((\mathbf{C}_{rs})_{ii}\big)$$

used for (F)PCA, for example.

The Estimator $\hat{\mathbf{U}}$

We suppose that we observe the process $Z(\cdot,\cdot)$ at $\mathbf{x}_i$, $i = 1,\ldots,n$ for all times t_r, $r = 1,\ldots,p$.

Write $\mathbf{Z}(\mathbf{x}_i,\cdot) = \big(Z(\mathbf{x}_i,t_1),\ldots,Z(\mathbf{x}_i,t_p)\big)^{\mathsf{T}}$, then

$$\hat{\mathbf{U}} = \frac{1}{n-1}\sum_{i=1}^{n}\big(\mathbf{Z}(\mathbf{x}_i,\cdot) - \bar{\mathbf{Z}}\big)\big(\mathbf{Z}(\mathbf{x}_i,\cdot) - \bar{\mathbf{Z}}\big)^T$$

Question: $\mathsf{E}(\hat{\mathbf{U}}) \overset{?}{=} \mathbf{\Sigma}$

Bias of the Estimator

Straightforward computation leads to

$$\mathsf{E}(\hat{U}_{rs}) = \sigma_{rs} - \frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{n}(\mathbf{C}_{rs})_{ij}$$

- Highly positive correlated variables lead to a higher bias
- Nugget effect has no influence on the bias
- The bias is bounded

For Gaussian processes we also have an explicit expression of $\text{Cov}(\hat{U}_{rs},\hat{U}_{uv})$ in terms of $(\mathbf{C}_{rs})_{ij}$.

Asymptotic Considerations

- Infill asymptotics ⇝ limiting bias

  $b^{\{n\}}(c^\circ_{rs}) = \mathsf{E}(\hat{U}_{rs}) - \sigma_{rs} \qquad b^{\text{lim}}(c^\circ_{rs}) = \lim_{n\to\infty} b^{\{n\}}(c^\circ_{rs})$

- Increasing-domain asymptotics ⇝ asymptotic bias

  $b^{(n)}(c^\circ_{rs}) = \mathsf{E}(\hat{U}_{rs}) - \sigma_{rs} \qquad b^{\text{asy}}(c^\circ_{rs}) = \lim_{n\to\infty} n b^{(n)}(c^\circ_{rs})$

Bias for Regular Grids

Proposition: Let $Z(\cdot,\cdot)$ be an isotropic second-order stationary process. Suppose the $n = \prod_{k=1}^{d} n_k$ locations form a regular grid in $d \geq 1$ dimensions. Let

$$g_1(q) = 2\sum_{l=1}^{n_1-1}(n_1 - l)c^{\circ}_{rs}\big(h\sqrt{q^2+l^2}\big)$$

$$g_k(q) = 2\sum_{l=1}^{n_k-1}(n_k - l)\Big(g_{k-1}\big(\sqrt{q^2+l^2}\big) + c^{\circ}_{rs}\big(h\sqrt{q^2+l^2}\big)\prod_{s=1}^{k-1} n_s\Big) + n_k g_{k-1}(q)$$

then we have

$$b(c^{\circ}_{rs}) = \mathrm{E}(\widehat{U}_{rs}) - \sigma_{rs} = -\frac{1}{n(n-1)} g_d(0)$$

Limiting Bias in one dimension

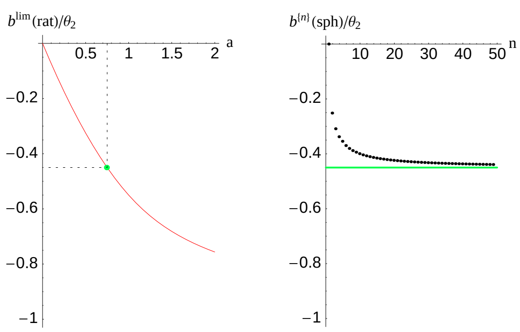
Suppose spatial process on a regular transect $[0,\delta]$, then

$$b^{\mathrm{lim}}(c^{\circ}_{rs}) = -\frac{2}{\delta^2}\int_0^{\delta}(\delta - x)c^{\circ}_{rs}(x)\,\mathrm{d}x$$

- with a spherical covariogram $\mathrm{sph}(h,\theta_1,\theta_2,\theta_3)$:

$$b^{\mathrm{lim}}(\mathrm{sph}) = \begin{cases} \theta_2\big(\frac{a^2}{5} - \frac{3a}{4}\big) & \text{if } a \leq 1 \\ \theta_2\big(\frac{1}{2a} - \frac{1}{20a^3} - 1\big) & \text{if } a > 1 \end{cases}$$

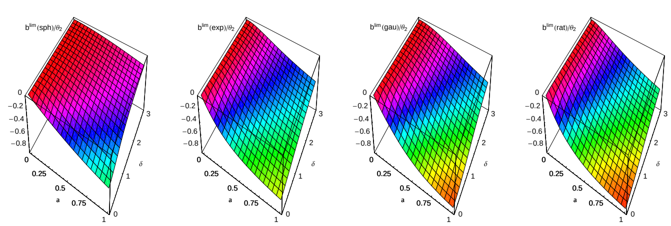
  where a is the ratio of the range and the domain.
- with an exponential covariogram $\mathrm{expo}(h,\theta_1,\theta_2,\theta_3)$:
  $b^{\mathrm{lim}}(\mathrm{expo}) = -2\theta_2\big(a + (\mathrm{e}^{-1/a} - 1)a^2\big)$
- with a Gaussian covariogram $\mathrm{gau}(h,\theta_1,\theta_2,\theta_3)$:
  $b^{\mathrm{lim}}(\mathrm{gau}) = -\theta_2 a\big(a(\mathrm{e}^{-1/a^2} - 1) + 2\sqrt{\pi}\Phi(\sqrt{2}/a) - \sqrt{\pi}\big)$
- with rational a quadratic covariogram $\mathrm{rat}(h,\theta_1,\theta_2,\theta_3)$
  $b^{\mathrm{lim}}(\mathrm{rat}) = -\theta_2 a\big(2\arctan(1/a) + a\log(a^2/1 + a^2)\big)$



Limiting Bias in two dimensions

For a spatial process on a regular two-dimensional grid $[0,1]\times[0,\delta]$ with underlying

- spherical covariogram
  $b^{\mathrm{lim}}(\mathrm{sph}) =$
- exponential covariogram
  $b^{\mathrm{lim}}(\mathrm{expo}) =$
- Gaussian covariogram
  $b^{\mathrm{lim}}(\mathrm{gau}) =$
- rational quadratic covariogram
  $b^{\mathrm{lim}}(\mathrm{rat}) =$



Asymptotic Bias

Proposition: Let $Z(\cdot,\cdot)$ be an isotropic second-order stationary process.

- If $c^{\circ}_{rs}(h) = \mathcal{O}\big(h^{-d(1+\epsilon)}\big)$ then $b^{(n)}(c^{\circ}_{rs}) = \mathcal{O}(1/n)$.
- If $c^{\circ}_{rs}(h) = \mathcal{O}\big(h^{-d}\big)$ then $b^{(n)}(c^{\circ}_{rs}) = \mathcal{O}\big(\log(n)/n\big)$.

The asymptotic bias is approximatively

- linear in $\theta_2\theta_3/h$ in one dimension
- linear in $\theta_2(\theta_3/h)^2$ in two dimensions

Asymptotic Covariance

Proposition: Let $Z(\cdot,\cdot)$ be an isotropic second-order stationary Gaussian process or fourth-order stationary process.
Then $\text{Cov}^{(n)}(\widehat{U}_{rs}, \widehat{U}_{uv}) = \mathcal{O}(1/n)$.

Expressions for the asymptotic covariance get long (if the fourth-order covariograms are not known).

Asymptotic Distribution

Proposition: Let $Z(\cdot,\cdot)$ be an isotropic fourth-order stationary process. If the underlying covariogram structure has finite range.
Then $\sqrt{n-1}\,\text{vec}(\widehat{\mathbf{U}}^{(n)} - \mathbf{\Sigma})$ converges in distribution a multinormal distribution.

For processes with infinite range, we have to use mixing theory.

Confidence Cones

Tyler (1981) constructs confidence (hyper)cones for sets of eigenvectors of random matrices.

Our setting satisfies Tyler's assumptions.
The confidence cone for $\widehat{\mathbf{v}}_i$ is given by

$$\Big\{\gamma\mathbf{a} \mid \gamma \in \mathbb{R}, \mathbf{a} \in \mathbb{R}^p : \quad \widehat{\mathbf{v}}^{\mathsf{T}}\mathbf{a} = 1 \text{ and } n\mathbf{a}^{\mathsf{T}}\big(\widehat{\mathbf{\Lambda}}^{(n)}\big)^{+}\mathbf{a} < q\chi^2_{p-1}(1-\alpha)\Big\}$$

where

$$\widehat{\mathbf{\Lambda}}^{(n)} = \big(\widehat{\mathbf{v}}_i{}^{\mathsf{T}} \otimes (\widehat{\mathbf{U}}^{(n)} - \widehat{\lambda}_i\mathbf{I})^{+}\big)\widehat{\mathbf{\Upsilon}}^{(n)}\big(\widehat{\mathbf{v}}_i \otimes (\widehat{\mathbf{U}}^{(n)} - \widehat{\lambda}_i\mathbf{I})^{+}\big)$$
$$\widehat{\mathbf{\Upsilon}}^{(n)} = \text{Var}\big(\text{vec}(\widehat{\mathbf{U}}^{(n)})\big)$$

Example

Covariogram	Situation 1	Situation 2
$c^\circ_{11}(h)$	$\text{sph}(h,0,3,7)$	$\text{sph}(h,0,8,22)$
$c^\circ_{22}(h)$	$\text{sph}(h,0,3,5)$	$\text{sph}(h,0,3,5)$
$c^\circ_{33}(h)$	$\text{sph}(h,0,3,5)$	$\text{sph}(h,0,3,5)$
$c^\circ_{12}(h)$	$\text{sph}(h,0,2,4)$	$\text{sph}(h,0,2,4)$
$c^\circ_{13}(h)$	$\text{sph}(h,0,1,3)$	$\text{sph}(h,0,1,3)$
$c^\circ_{23}(h)$	$\text{sph}(h,0,1,2)$	$\text{sph}(h,0,1,2)$

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{U}} = \begin{pmatrix} 2.407 & 1.854 & 0.960 \\ 1.854 & 2.667 & 0.985 \\ 0.960 & 0.985 & 2.667 \end{pmatrix}$$

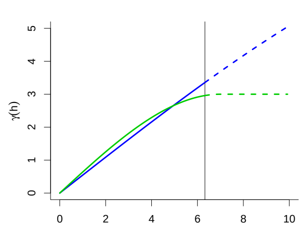
$$\mathbf{\Sigma} = \begin{pmatrix} 8 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{U}} = \begin{pmatrix} 2.768 & 1.854 & 0.960 \\ 1.854 & 2.667 & 0.985 \\ 0.960 & 0.985 & 2.667 \end{pmatrix}$$

$$\lambda_1 = 5.562 \quad \widehat{\lambda}_1 = 4.929 \quad \mathbf{v}_1 = \begin{pmatrix} 0.628 \\ 0.628 \\ 0.460 \end{pmatrix} \quad \widehat{\mathbf{v}}_1 = \begin{pmatrix} 0.599 \\ 0.638 \\ 0.484 \end{pmatrix}$$

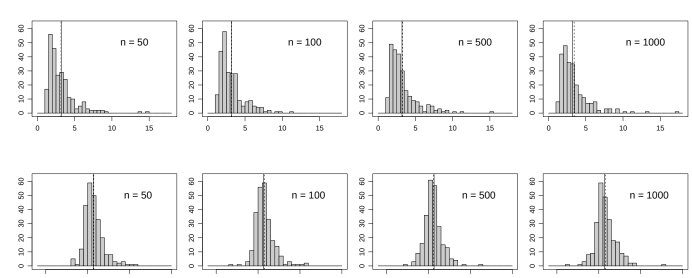
angle$(\mathbf{v}_1, \widehat{\mathbf{v}}_1) \approx 2.2°$, which is negligible: p-value= 97%

$$\lambda_1 = 9.806 \quad \widehat{\lambda}_1 = 5.080 \quad \mathbf{v}_1 = \begin{pmatrix} 0.916 \\ 0.342 \\ 0.210 \end{pmatrix} \quad \widehat{\mathbf{v}}_1 = \begin{pmatrix} 0.632 \\ 0.621 \\ 0.464 \end{pmatrix}$$

angle$(\mathbf{v}_1, \widehat{\mathbf{v}}_1) \approx 27.3°$, which is significant: p-value= 3%

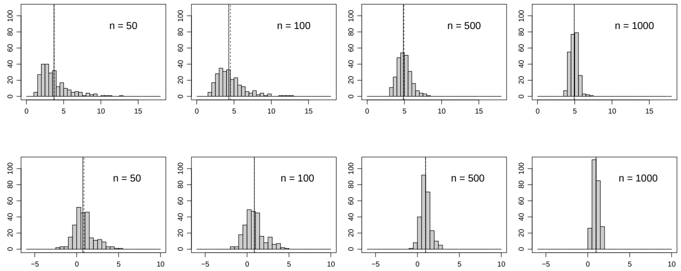


Limiting distribution



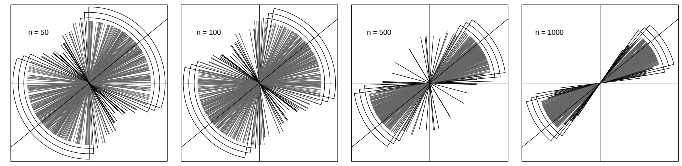
Histograms of $\widehat{U}^{\{n\}}_{11}$ and $\widehat{U}^{\{n\}}_{11}$ based on 250 samples of a Gaussian process with underlying covariograms $c^\circ_{11}(h) = \text{sph}(h,1,4,0.75n)$, $c^\circ_{12}(h) = \text{sph}(h,0,1,0.5n)$ and $c^\circ_{22}(h) = \text{sph}(h,0,1,0.5n)$.

Asymptotic distribution



Histograms of $\hat{U}_{11}^{(n)}$ and $\hat{U}_{12}^{(n)}$ based on 250 samples of a Gaussian process with covariograms $c_{11}^{\circ}(h) = \text{sph}(h, 1, 4, 25)$, $c_{12}^{\circ}(h) = \text{sph}(h, 0, 1, 20)$ and $c_{22}^{\circ}(h) = \text{sph}(h, 0, 1, 20)$.

Confidence Cones



Confidence cones for a Gaussian process with covariograms $c_{11}^{\circ}(h) = \text{sph}(h, 1, 4, 25)$, $c_{12}^{\circ}(h) = \text{sph}(h, 0, 1, 20)$ and $c_{22}^{\circ}(h) = \text{sph}(h, 0, 1, 20)$.

Irregular Locations

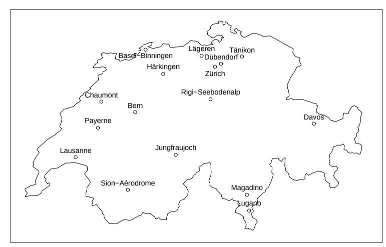
We perturb each location of a regular 20×20 grid.

Variogram	Distribution	$b^{(n)}(c_{rs}^{\circ}(\lVert \mathbf{y}_i - \mathbf{y}_j \rVert))$ MRE (%)	STD ($\times 10^{-3}$)
sph(h, 0, 4, 9)	$\mathcal{U}(-0.2, 0.2)$	−0.040	1.288
	$\mathcal{N}(0, 0.04/3)$	−0.019	1.246
sph(h, 0, 4, 15)	$\mathcal{U}(-0.5, 0.5)$	−0.154	2.840
	$\mathcal{N}(0, 0.25/3)$	−0.125	2.734
expo(h, 0, 4, 3)	$\mathcal{U}(-0.2, 0.2)$	0.001	1.171
	$\mathcal{N}(0, 0.04/3)$	0.007	1.125
expo(h, 0, 4, 5)	$\mathcal{U}(-0.5, 0.5)$	−0.100	2.430
	$\mathcal{N}(0, 0.25/3)$	−0.095	2.468

We sample $\tilde{n}$ locations of a regular 20×20 grid and adjust range with $\sqrt{\tilde{n}/n} \cdot \theta_3$.

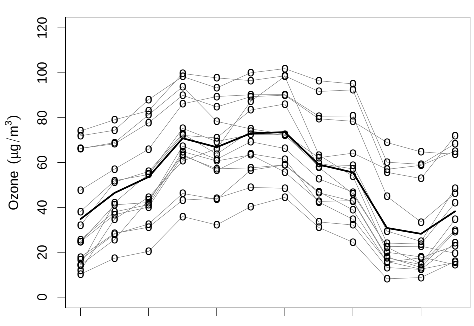
Variogram	$\tilde{n} = 196$ MRE (%)	STD ($\times 10^{-2}$)	$\tilde{n} = 289$ MRE (%)	STD ($\times 10^{-2}$)
sph(h, 0, 4, 9)	1.851	2.446	0.705	1.542
sph(h, 0, 4, 15)	0.486	2.605	0.188	1.563
expo(h, 0, 4, 3)	2.151	2.303	0.788	1.424
expo(h, 0, 4, 5)	0.803	1.913	0.392	1.178

Application: Ozone Data



- 16 'National Air Pollution Monitoring Network' stations in Switzerland
- Ozone level is automatically measured in μg/m^3 every half hour
- Aggregated monthly data for 1999

Mean monthly ozone values



Mean monthly ozone values of the 16 NABEL monitoring stations

Model

The transformed measures are modeled with an isotropic second-order stationary process $Z(\mathbf{x}_i, t_r)$, $i = 1, \ldots, 16$ and $r = 1, \ldots, 12$ and fitted covariograms

$$c_{rs}^{\circ}(h) = \begin{cases} 1.85 & \text{if } h = 0 \\ \dfrac{0.80}{1 + 12.8|t_r - t_s|} \exp\left(-\dfrac{0.16h^2}{1 + 12.8|t_r - t_s|}\right) & \text{otherwise} \end{cases}$$

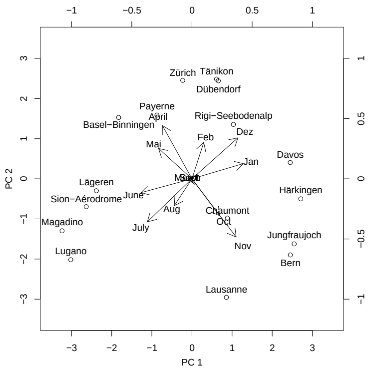
$\hat{\mathbf{U}}$ is estimated and corrected for bias to obtain $\hat{\mathbf{\Sigma}}$.

The principal components are derived from the eigenvectors of $\hat{\mathbf{\Sigma}}$.

The angles between the first four eigenvectors of the matrices $\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{U}}$ are 0.426°, 1.461°, 1.373° and 2.073°.

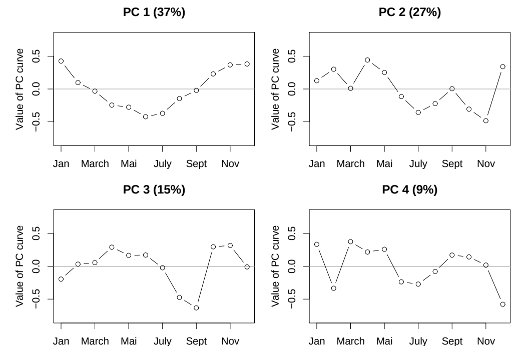
Principal components	1	2	3	4	5	6	7	8	...
Variation	0.37	0.27	0.15	0.09	0.05	0.02	0.02	0.01	...
Cumulative variation	0.37	0.63	0.78	0.87	0.92	0.94	0.96	0.97	...

Biplot



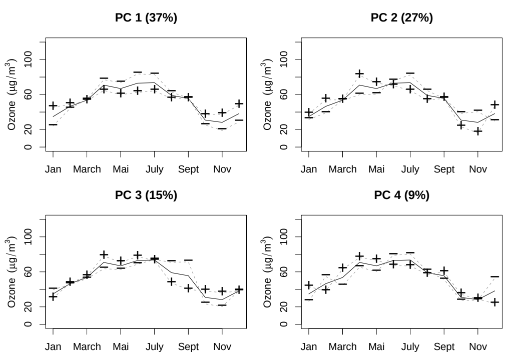
Biplot of the first two principal components of the ozone dataset

Principal Components



First four principal components curves of the ozone dataset

Mean Ozone Curves



Mean ozone curves and the effects of adding (+) and subtracting (−) a suitable multiple of each principal component

Conclusion and Outlook

- Fast and accurate bias correction
- Approximation remains valid for non regular grids
- For spatio-temporal or multivariate processes
- Asymptotic distribution for processes with infinite ranges
- Apply to other datasets

References

BUWAL (2000). *NABEL, Luftbelastung 1999, Schriftenreihe Umwelt Nr. 316*. Bundesamt für Umwelt, Wald und Landschaft, Switzerland.

Furrer (1998). Principal component analysis of Lake Geneva sediments. In Buccianti, Nardi and Potenza, editors, *Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology*, 421–426. Frede Editore, Napoly.

Furrer (1999). Covariance estimation under spatial dependence. In Mardia, Aykroyd, and Dryden, editors, *Proceedings in Spatial Temporal Modelling and its Applictions*, 137–140. Leeds University Press.

Ramsay and Silverman (1997). *Functional Data Analysis*. Springer-Verlag, New York.

Tyler (1981). Asymptotic inference for eigenvectors. *The Annals of Statistics*, **9**, 725–736.