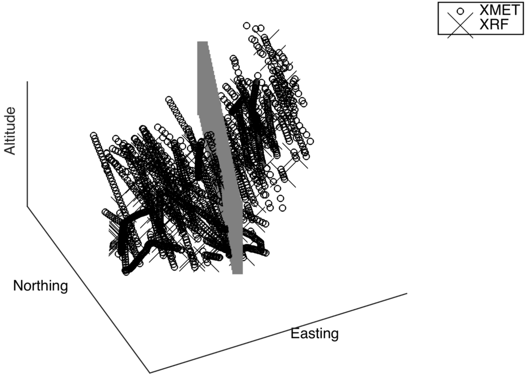
Gaussian processes: Kriging and parameter estimation

Jo Eidsvik

Department of Mathematical Sciences, NTNU, Norway

Grade prediction from boreholes



Kriging interpolation

$$Y^\star(\boldsymbol{s}_0) = \sum_{i=1}^{n} \alpha_i Y(\boldsymbol{s}_i) = \boldsymbol{\alpha}^t \boldsymbol{Y}$$

- Spatial interpolation from data $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$
- Best linear (spatial) predictor
- Kriging equals the optimal prediction for Gaussian model
- Unbiased and minimum prediction variance

Versions

- Simple kriging $E[Y(\boldsymbol{s})] = 0.$
- Ordinary kriging $E[Y(\boldsymbol{s})] = \mu$
- Universal kriging $E[Y(\boldsymbol{s})] = \boldsymbol{h}(\boldsymbol{s})\beta$
- Cokriging (Multivariate data $Y_1(\boldsymbol{s}), \ldots, Y_K(\boldsymbol{s})$)

Kriging derivation

$$\sigma^2_{s_0} = Var[Y^\star(s_0) - Y(s_0)] = E[Y^\star(s_0) - Y(s_0)]^2 - \{E[Y^\star(s_0)] - E[Y(s_0)]\}^2$$

'Mean Square Prediction Error' = 'Variance' + 'Bias squared'

$$\begin{aligned}\sigma^2_{s_0} &= E[Y^\star(s_0) - Y(s_0)]^2 = E[\sum_i \alpha_i Y(s_i) - Y(s_0)]^2 \qquad (1)\\ &= E[\sum_i \sum_j \alpha_j \alpha_i Y(s_i) Y(s_j) - 2Y(s_0) \sum_i \alpha_i Y(s_i) + Y(s_0)^2]\\ &= \sum_i \sum_j \alpha_i \alpha_i C(i,j) - 2 \sum_i \alpha_i C(0,i) + C(0,0)\\ &= \boldsymbol{\alpha}^t C \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^t \mathbf{C}_{0,\cdot} + C_0\end{aligned}$$

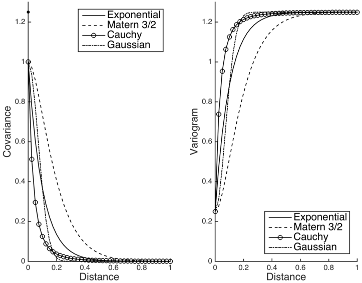
Kriging derivation

Minimizing prediction error as a function of the weights α_i.
Optimal weights - derivative is 0 at the minimum.

$$
\begin{aligned}
\frac{d\sigma^2_{s_0}}{d\boldsymbol{\alpha}} &= 2C\boldsymbol{\alpha} - 2\boldsymbol{C}_{0,\cdot} = 0 \\
\boldsymbol{\alpha} &= C^{-1}\boldsymbol{C}_{0,\cdot} \\
Y^\star(s_0) &= \boldsymbol{\alpha}^t\boldsymbol{Y} = \boldsymbol{C}^t_{0,\cdot}C^{-1}\boldsymbol{Y}
\end{aligned}
\tag{2}
$$

One can show the same with the variogram
$\gamma(\boldsymbol{h}) = \mathrm{Var}(Y(\boldsymbol{s}) - Y(\boldsymbol{s}+\boldsymbol{h}))$.

Covariance functions and variograms



Interpretation

$$Y^\star(\mathbf{s}_0) = \boldsymbol{\alpha}^t \mathbf{Y} = \mathbf{C}_{0,\cdot}^t C^{-1} \mathbf{Y}$$

- Unbiased linear predictor $E[Y(s)] = 0$ for all s.
- Weights depend on $Cov[Y(s_i), Y(s_0)]$: Closer sites get larger weight
- Weights depend on $Cov[Y(s_i), Y(s_j)]$: Clustered sites get less weight

Prediction variance

Plugging in optimal $\boldsymbol{\alpha}$ in $\sigma^2_{s_0}$.

$$\begin{aligned} \sigma^2_{s_0} &= \boldsymbol{\alpha}^t C \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^t \boldsymbol{C}_{0,\cdot} + C_0 \\ &= \boldsymbol{C}^t_{0,\cdot} C^{-1} C C^{-1} \boldsymbol{C}_{0,\cdot} - 2\boldsymbol{C}^t_{0,\cdot} C^{-1} \boldsymbol{C}_{0,\cdot} + C_0 \\ &= C_0 - \boldsymbol{C}^t_{0,\cdot} C^{-1} \boldsymbol{C}_{0,\cdot} \end{aligned} \quad (3)$$

- Prediction variance is smaller than C_0.
- Decrease in prediction variance is larger close to data sites: $\boldsymbol{C}_{0,\cdot}$ large.
- Prediction variance does not depend on data. It can be computed before the data acquisition.
- The spatial allocation of sites $s_1, \ldots, s_n$ is called 'spatial design'. This design impacts the prediction variance.

Spatial regression model

Model: $Y(\boldsymbol{s}) = \boldsymbol{h}(\boldsymbol{s})\boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.

1. $Y(\boldsymbol{s})$ response variable at position $\boldsymbol{s}$.
2. $\boldsymbol{\beta}$ regression effects. $\boldsymbol{h}(s)$ covariates at $\boldsymbol{s}$.
3. $w(\boldsymbol{s})$ zero-mean structured (spatially correlated) Gaussian process.
4. $\epsilon(s)$ zero-mean unstructured (independent) Gaussian measurement noise.

Gaussian model

Model: $Y(\boldsymbol{s}) = \boldsymbol{h}(\boldsymbol{s})\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.
Data at n locations: $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$.
Likelihood:

$$l(\boldsymbol{Y}; \beta, \boldsymbol{\theta}) = -\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{H}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{H}\beta)$$

$$\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{w}) + \mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{C} + \tau^2 \boldsymbol{I}$$

Maximum likelihood

$$(\hat{\theta}, \hat{\boldsymbol{\beta}}) = \text{argmax}_{\boldsymbol{\theta},\boldsymbol{\beta}}\{l(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta})\}.$$

$$\hat{\theta}_{p+1} = \hat{\theta}_p - E\left(\frac{d^2 l(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\theta}^2}\right)^{-1} \frac{dl(\boldsymbol{Y}; \hat{\boldsymbol{\beta}}_p, \hat{\theta}_p)}{d\boldsymbol{\theta}},$$

$\hat{\boldsymbol{\beta}}_p = \boldsymbol{A}^{-1}\boldsymbol{b}$, $\boldsymbol{A} = \boldsymbol{A}(\hat{\theta}_p)$, $\boldsymbol{b} = \boldsymbol{b}(\hat{\theta}_p)$.

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H} \\ \boldsymbol{b} &= \boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}. \end{aligned}$$

Analytical derivatives

Formulas for matrix derivatives.

$$
\begin{aligned}
\boldsymbol{Q} &= \boldsymbol{\Sigma}^{-1} \\
\frac{d \log |\boldsymbol{\Sigma}|}{d\theta_r} &= \operatorname{trace}(\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_r}) \\
\frac{d\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}}{d\theta_r} &= -\boldsymbol{Z}'\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_r}\boldsymbol{Q}\boldsymbol{Z}.
\end{aligned}
$$

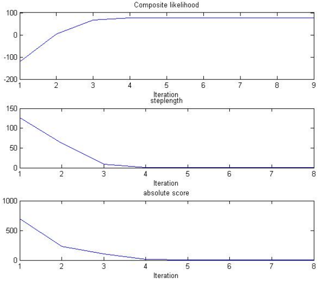
Score and Hessian

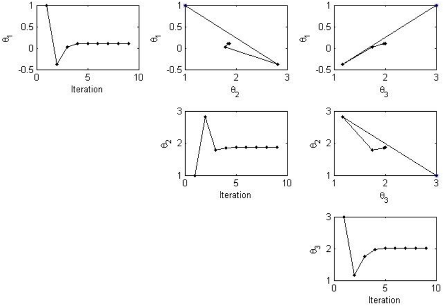
$$\frac{dl}{d\boldsymbol{\theta}_r} = -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_r}) + \frac{1}{2}\boldsymbol{Z}'\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_r}\boldsymbol{Q}\boldsymbol{Z},$$

$$E\left(\frac{d^2 l}{d\theta_r d\theta_s}\right) = -\frac{1}{2}\text{trace}(\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_s}\boldsymbol{Q}\frac{d\boldsymbol{\Sigma}}{d\theta_r})\}.$$

Illustration maximization

Exponential covariance with nugget effect. $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$: log **precision**, logistic **range**, log **nugget** precision.





Properties

- maximum likelihood estimators are asymptotically unbiased.
- maximum likelihood estimators attain asymptotically minimum variance
- maximum likelihood estimators are asymptotically Gaussian distributed.

Challenges

Model: $Y(\boldsymbol{s}) = \boldsymbol{H}(\boldsymbol{s})\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$.
Data at n locations: $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$.
Likelihood:

$$l(\boldsymbol{Y}; \beta, \boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{H}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{H}\beta)$$

Challenges:

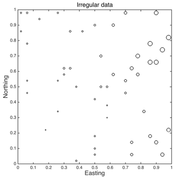
1. Build and store $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} = \boldsymbol{C} + \tau^2\boldsymbol{I}_n$
2. Compute $\log|\boldsymbol{\Sigma}|$
3. Compute $\boldsymbol{\Sigma}^{-1}$ or $(\boldsymbol{Y} - \boldsymbol{H}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{H}\beta)$

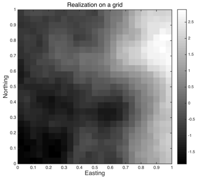
Possible solutions for large Gaussian models

- Approximate likelihood (Fuentes 2007).
- Basis representation (Banerjee et al. 2008; Cressie and Johanesson 2008).
- Markov representation (Lindgren et al. 2011).
- Tapered likelihood (Kaufman et al 2008).
- Composite likelihoods (Stein et al. 2004, Eidsvik et al 2014; Datta et al. 2016)
- Machine learning (Rasmussen and Williams 2006).
- Numerical linear algebra (Higham 2008, Aune et al., 2014).

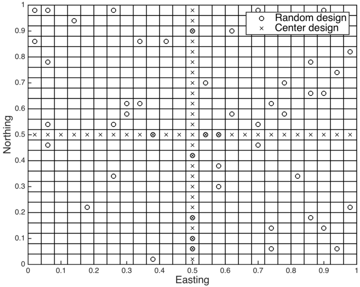
Example 1: Norwegian wood

$$Y(\boldsymbol{s}) = \boldsymbol{h}(\boldsymbol{s})\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$





Data designs



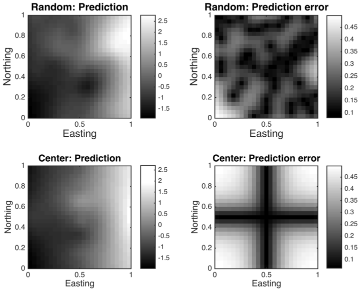
Estimation: MLE

Table: Estimates(standard error).

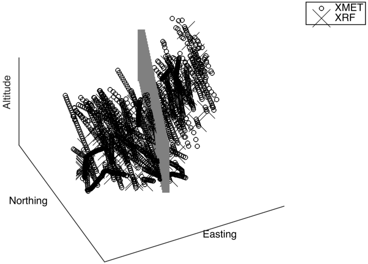
	β_0	β_1	β_2	σ^2	η	τ^2
Center	-2.1 (0.6)	3.4 (0.7)	0.4 (0.7)	0.3 (0.14)	7.2 (2.0)	0.002 (0.001)
Random	-2.0 (0.5)	3.4 (0.6)	0.8 (0.5)	0.3 (0.12)	7.9 (2.0)	0.005 (0.007)
Truth	-2	3	1	0.25	9	0.0025

Matern covariance function.

Predictions



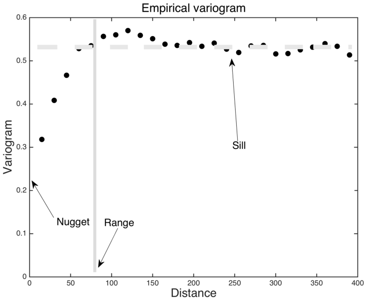
Example 2: Ore grade prediction in mining



Case

- Data at 1871 locations.
- Covariate is mineralization index (three possible classes) $h(\boldsymbol{s}) = [1, \text{min.ind}(\boldsymbol{s})]$
- Spatial covariance is modeled by exponential covariance model.

Variogram

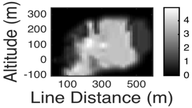

Empirical variogram
Variogram
Distance
Nugget
Range
Sill
0
0.1
0.2
0.3
0.4
0.5
0.6
0
50
100
150
200
250
300
350
400

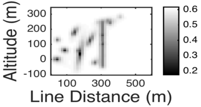
Parameter estimation

Maximum likelihood (10 iterations of Fisher scoring.)

- $\beta_1 = 1.32$ (higher grades with mineralization index).
- Correlation range 50 m, $\tau^2 = 0.45^2 = 0.2$, $\sigma^2 = 0.62^2 = 0.38$.

Predictions



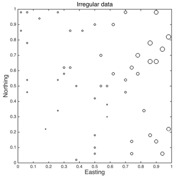


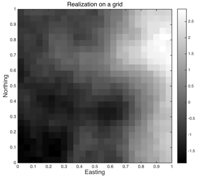
Exercise: Norwegian wood

$$Y(\boldsymbol{s}) = \boldsymbol{h}(\boldsymbol{s})\beta + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s})$$

$$\boldsymbol{Y} \sim N(\boldsymbol{H}\beta, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

Different covariate.





Spatio-temporal model

Model: $Y(\boldsymbol{s}, t) = \boldsymbol{h}(\boldsymbol{s}, t)\boldsymbol{\beta} + w(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t)$.

1. $Y(\boldsymbol{s}, t)$ response variable at position $\boldsymbol{s}$ at time t.
2. $\boldsymbol{\beta}$ regression effects. $\boldsymbol{h}(\boldsymbol{s}, t)$ covariates at $\boldsymbol{s}$ at time t.
3. $w(\boldsymbol{s}, t)$ zero-mean structured (spatio-temporally correlated) Gaussian process.
4. $\epsilon(\boldsymbol{s}, t)$ zero-mean unstructured (independent) Gaussian measurement noise.

Spatio-temporal statistics

Model: $Y(\boldsymbol{s}, t) = \boldsymbol{h}(\boldsymbol{s}, t)\boldsymbol{\beta} + w(\boldsymbol{s}, t) + \epsilon(\boldsymbol{s}, t)$.
Data at n_t locations at time t: $\boldsymbol{Y}_t = (Y(\boldsymbol{s}_1, t), \ldots, Y(\boldsymbol{s}_{n_t}, t))'$,
$t = t_1, t_2, \ldots, t_n$.
Goals could include:

- Estimate parameters: regression, noise structure in space and time, and noise of measurements.
- Characterize process in space and time: Smoothing, given all data. Filtering, given only current data. Prediction, given some data - look ahead in time. Interpolation (Kriging) in space.

Common assumptions

Covariates $h(\boldsymbol{s}, t)$ help include trends in space (say altitude, land-cover, etc.) or over time (hour, season, climate change, etc.), or coupling of space-time.

Covariance structure of $w(\boldsymbol{s}, t)$ is

- Stationary in space and time: $\mathrm{Var}(w(\boldsymbol{s}, t)) = \mathrm{Var}(w(\boldsymbol{s}', t'))$, $\mathrm{Corr}(w(\boldsymbol{s}, t), w(\boldsymbol{s}', t')) = \mathrm{Corr}(w(\boldsymbol{s} + \boldsymbol{s}_0, t + t_0), w(\boldsymbol{s}' + \boldsymbol{s}_0, t' + t_0))$.
- Separable in space and time: $\mathrm{Corr}(w(\boldsymbol{s}, t), w(\boldsymbol{s}', t')) = \mathrm{Corr}_s(w(\boldsymbol{s}, t), w(\boldsymbol{s}', t))\mathrm{Corr}_t(w(\boldsymbol{s}, t), w(\boldsymbol{s}, t'))$.

Autoregressive spatial process

Markov in time and stationary:

$$w(\boldsymbol{s}, t) = \phi w(\boldsymbol{s}, t-1) + \delta(\boldsymbol{s}, t), \quad \mathrm{Var}(w(\boldsymbol{s}, 0)) = \boldsymbol{\Sigma}_0, \quad \mathrm{Var}(\delta(\boldsymbol{s}, t)) = (1-\phi^2)\boldsymbol{\Sigma}_0$$

$\delta(\boldsymbol{s}, t)$ are independent in time.
This means that the one-step time correlation is ϕ (and separable from space).

Advection-diffusion equation

Structure process defined via a partial differential equation:

$$\frac{dw(\boldsymbol{s},t)}{dt} = -\boldsymbol{\mu}^t \nabla w(\boldsymbol{s},t) + \nabla \boldsymbol{S} \nabla w(\boldsymbol{s},t) + \delta(\boldsymbol{s},t),$$

$\boldsymbol{\mu}$ is advection (drift) term. $\boldsymbol{D}$ is diffusion term, δ is independent Gaussian noise.
Time-difference scheme gives Markovian process in time.

Advection-diffusion equation

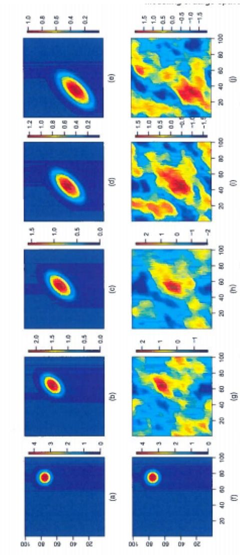


Fig. 1. Illustration of the SPDE (1) and the corresponding PDE (the drift vector points from north-east to south-west and the diffusive part exhibits anisotropy in the same direction; the same parameters are used for both the PDE and the SPDE, i.e. $\zeta = -\log(0.99)$, $\rho_1 = 0.06$, $\gamma = 3$, $\psi = \pi/4$, $\mu_x = -0.1$ and $\mu_y = -0.1$, and for the stochastic innovations, i.e. $\rho_0 = 0.05$ and $\sigma^2 = 0.7^2$; the colour scales are different in different panels): (a)–(e) a solution to the PDE which corresponds to the deterministic part of the SPDE without stochastic term $\epsilon(t, s)$; (f)–(j) one sample from the distribution specified by the SPDE with a fixed initial condition: (a), (f) $t = 1$; (b), (g) $t = 2$; (c), (h) $t = 3$; (d), (i) $t = 4$; (e), (j) $t = 5$

Exercise Spatial case : Simulate and re-estimate

$$\boldsymbol{Y} = \boldsymbol{H}\beta + \boldsymbol{L}N(0, I)$$

Cholesky matrix:

$$\boldsymbol{L}\boldsymbol{L}^t = \boldsymbol{\Sigma}(\theta)$$

- Fix parameters.
- Simulate a realization of $\boldsymbol{Y}$ at data locations jointly with variables $\boldsymbol{Y}_0$ at prediction locations.
- Fit parameters (maximum likelihood) from data $\boldsymbol{Y}$.
- Predict, $\hat{\boldsymbol{Y}}_0$ (with uncertainty) given data and parameters.

Exercise Spatial case : Joint Gaussian

$$\begin{pmatrix} \boldsymbol{Y}_0 \\ \boldsymbol{Y} \end{pmatrix} \sim N\left[\begin{pmatrix} \boldsymbol{H}_0 \\ \boldsymbol{H} \end{pmatrix}\beta, \begin{pmatrix} \boldsymbol{\Sigma}_0 & \boldsymbol{\Sigma}_{0,.} \\ \boldsymbol{\Sigma}_{.,0} & \boldsymbol{\Sigma} \end{pmatrix}\right]$$

$$[\boldsymbol{Y}_0|\boldsymbol{Y}] \sim N(\boldsymbol{H}_0\boldsymbol{\beta} + \boldsymbol{\Sigma}_{0,.}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{H}\beta), \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_{0,.}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{.,0})$$