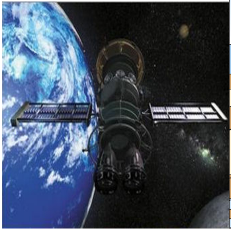
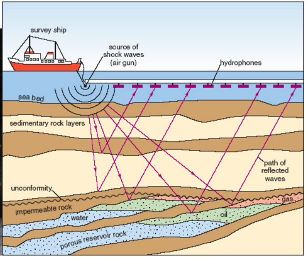
Schedule

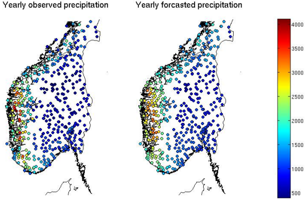
- 17 Jan: Gaussian processes (Jo Eidsvik)
- 24 Jan: Hands-on project on Gaussian processes (Team effort, work in groups)
- 31 Jan: Latent Gaussian models and INLA (Jo Eidsvik)
- 7 Feb: Hands-on project on INLA (Team effort, work in groups)
- 12-13 Feb: Template model builder. (Guest lecturer Hans J Skaug)
- To be decided...

Large spatial (spatio-temporal) datasets

Gaussian processes are very commonly used in practice.





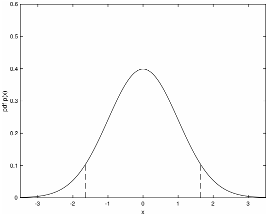


Other contexts

- Genetic data
- Functional data
- Response surfaces modeling and optimization

Extremely common as building block in several machine learning applications.

Univariate Gaussian distribution



$Y \sim N(\mu, \sigma^2)$.

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right), \quad y \in \mathbb{R}.$$

$$Z = \frac{Y-\mu}{\sigma}, \quad Y = \mu + \sigma Z.$$

Z is standard normal, mean 0 and variance 1.

Multivariate gaussian distribution

Size $n \times 1$ vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$

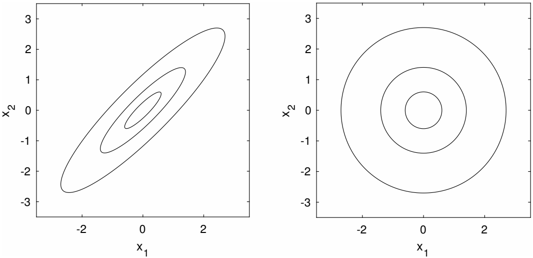
$$p(\boldsymbol{y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad \boldsymbol{Y} \in \mathbb{R}^n.$$

Mean is $E(\boldsymbol{Y}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Variance-covariance matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \ldots & \Sigma_{1,n} \\ \ldots & \ldots & \ldots \\ \Sigma_{n,1} & \ldots & \Sigma_{n,n} \end{bmatrix},$$

$\Sigma_{i,i} = \sigma_i^2 = \mathrm{Var}(Y_i)$, $\Sigma_{i,j} = \mathrm{Cov}(Y_i, Y_j)$, $\mathrm{Corr}(Y_i, Y_j) = \Sigma_{i,j}/(\sigma_i\sigma_j)$.

Illustrations $n = 2$ - correlation 0.9 (left), independent (right).



Joint for blocks

$\boldsymbol{Y}_A = (Y_{A,1}, \ldots, Y_{A,n_A})$, $\boldsymbol{Y}_B = (Y_{B,1}, \ldots, Y_{B,n_B})$, joint Gaussian with mean $(\boldsymbol{\mu}_A, \boldsymbol{\mu}_B)$, covariance:

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_A, \boldsymbol{\mu}_B), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{A,B} \\ \boldsymbol{\Sigma}_{B,A} & \boldsymbol{\Sigma}_B \end{bmatrix},$$

Conditioning

$$
\begin{aligned}
E(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{Y}_B - \boldsymbol{\mu}_B), \\
\mathrm{Var}(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_{B,A}.
\end{aligned}
$$

Mean is linear in conditioning variable (data).
Variance is not dependent on data.

Illustration

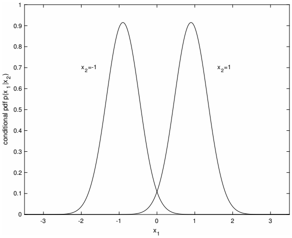


Figure: Conditional pdf for Y_1 when $Y_2 = 1$ or $Y_2 = -1$.

Transformation

$$\boldsymbol{Z} = \boldsymbol{L}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}), \quad \boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z}, \quad \boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'.$$

$\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ are independent standard normal, mean 0 and variance 1.

Cholesky factorization

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{1,1} & \dots & \Sigma_{1,n} \\ \dots & \dots & \dots \\ \Sigma_{n,1} & \dots & \Sigma_{n,n} \end{bmatrix} = \boldsymbol{L}\boldsymbol{L}',$$

Lower triangular matrix

$$\boldsymbol{L} = \begin{bmatrix} L_{1,1} & 0 & \dots & 0 \\ L_{2,1} & L_{2,2} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ L_{n,1} & L_{n,2} & \dots & L_{n,n} \end{bmatrix},$$

Cholesky - example

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} 1 & 0 \\ 0.9 & 0.44 \end{bmatrix}.$$

Consider sampling from joint $p(y_1, y_2) = p(y_1)p(y_2|y_1)$:

- Sample from $p(y_1)$ is constructed by $Y_1 = \mu_1 + L_{1,1}Z_1$.
- Sample from $p(y_2|y_1)$ is constructed by $Y_2 = \mu_2 + L_{2,1}Z_1 + L_{2,2}Z_2 = \mu_2 + L_{2,1}\frac{Y_1 - \mu_1}{L_{1,1}} + L_{2,2}Z_2$

Gaussian process

For any set of time locations $t_1, \ldots, t_n$. $Y(t_1), \ldots, Y(t_n)$ is jointly Gaussian.
Mean $\mu(t_i)$, $i = 1, \ldots, n$.

$$\mathbf{\Sigma} = \left[\begin{array}{ccc} \Sigma_{1,1} & \ldots & \Sigma_{1,n} \\ \ldots & \ldots & \ldots \\ \Sigma_{n,1} & \ldots & \Sigma_{n,n} \end{array} \right],$$

Covariance function

The covariance tends to decay with distance:

$$\Sigma_{i,j} = \gamma(|t_i - t_j|),$$

for some covariance function γ.
Examples:

$$\gamma_{\text{exp}}(h) = \sigma^2 \exp(-\phi h)$$

$$\gamma_{\text{mat}}(h) = \sigma^2 (1 + \phi h) \exp(-\phi h)$$

$$\gamma_{\text{gauss}}(h) = \sigma^2 \exp(-\phi h^2)$$

Illustration of covariance function

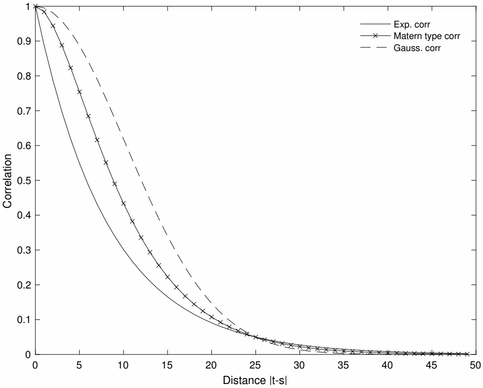
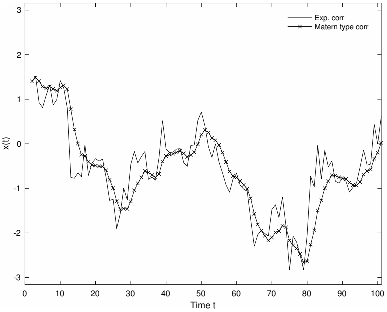


Illustration of samples of Gaussian processes



Markov property

The exponential correlation function gives a Markov process.
For any $t > s > u$,

$$p(y(t)|y(s), y(u)) = p(y(t)|y(s)).$$

(Proof by trivariate distribution, and conditioning.)

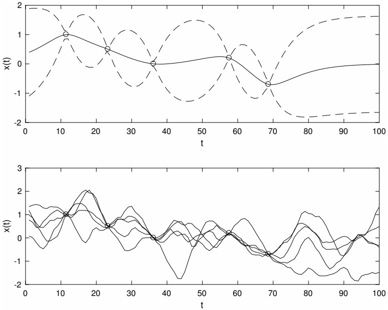
$$\begin{aligned} E(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{Y}_B - \boldsymbol{\mu}_B), \\ \mathrm{Var}(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_{B,A}. \end{aligned}$$

Conditional formula

$$
\begin{aligned}
E(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{Y}_B - \boldsymbol{\mu}_B), \\
\mathrm{Var}(\boldsymbol{Y}_A|\boldsymbol{Y}_B) &= \boldsymbol{\Sigma}_A - \boldsymbol{\Sigma}_{A,B}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\Sigma}_{B,A}.
\end{aligned}
$$

- Expectation linear in data.
- Variance only dependent on data locations, not data.
- Expectation close to conditioning variables near data locations, goes to $\boldsymbol{\mu}_A$ far from data.
- Variance small near data locations, goes to $\boldsymbol{\Sigma}_A$ far from data.
- Close data locations are not double data.

Illustration of conditioning in Gaussian processes



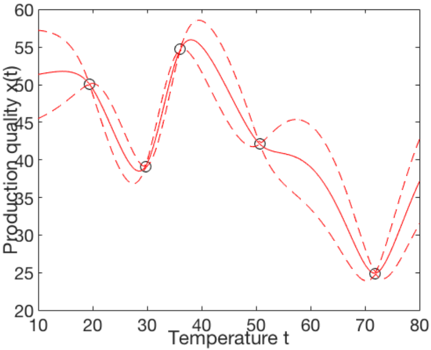
Application of Gaussian processes: function optimization

Several applications involve very time-demanding or costly experiment.
Which configurations of inputs give highest output?
Goal is to find the optimal input without too many trials / tests.

Approach: Fit a GP to the function, based on the evaluation points and results. This allows fast consideration of which evaluations points to choose!

Production quality : Current prediction

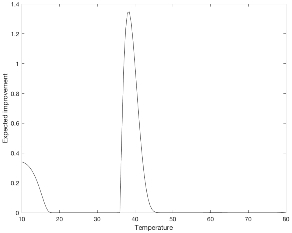
Goal: Find the best temperature input, to give optimal production. Which temperature to evaluate next? Experiment is costly, want to do few evaluations.



Expected improvement

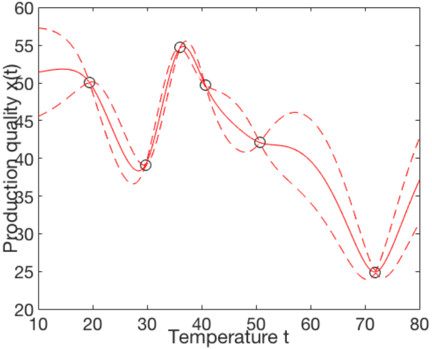
Maximum so far $Y^* = \max(\boldsymbol{Y})$.

$$EI_n(s) = E(max(Y(s) - Y^*, 0)|\boldsymbol{Y})$$



Sequential uncertainty reduction

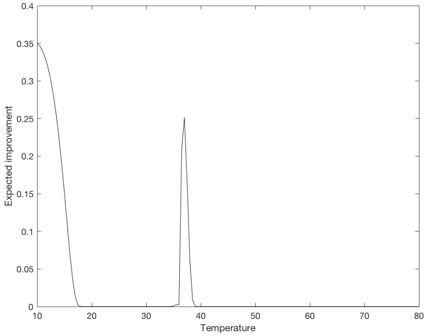
Perform test at $t = 40$, result is $Y(40) = 50$.



Sequential uncertainty reduction and optimization

Maximum so far $Y^* = \max(\boldsymbol{Y}, Y_{n+1})$.

$$EI_{n+1}(s) = E(max(Y(s) - Y^*, 0)|\boldsymbol{Y}, Y_{n+1})$$



Project on function optimization using EI next week

Analytical solutions to parts of computational challenge.

Precision matrix Q

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_A & \boldsymbol{Q}_{A,B} \\ \boldsymbol{Q}_{B,A} & \boldsymbol{Q}_B \end{bmatrix}.$$

$\boldsymbol{Q}$ holds the conditional variance structure.

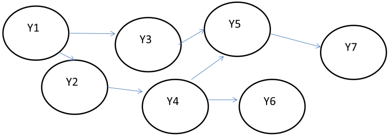
Interpretation of precision

$$\boldsymbol{Q}_A^{-1} = \mathrm{Var}(\boldsymbol{Y}_A | \boldsymbol{Y}_B),$$

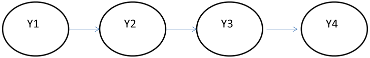
$$\mathrm{E}(\boldsymbol{Y}_A | \boldsymbol{Y}_B) = \boldsymbol{\mu}_A - \boldsymbol{Q}_A^{-1} \boldsymbol{Q}_{A,B} (\boldsymbol{Y}_B - \boldsymbol{\mu}_B),$$

(Proof by $\boldsymbol{Q}\boldsymbol{\Sigma} = \boldsymbol{I}$.
Or by writing out quadratic form and $p(\boldsymbol{Y}_A | \boldsymbol{Y}_B) \propto p(\boldsymbol{Y}_A, \boldsymbol{Y}_B)$.)

Sparse precision matrix Q



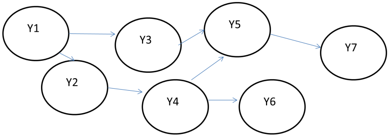
$$p(Y_7 \mid Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) = p(Y_7 \mid Y_5)$$



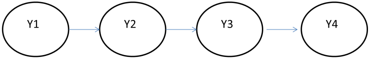
$$p(Y_i \mid Y_1, \ldots, Y_{i-1}) = p(Y_i \mid Y_{i-1})$$

- For graphs the precision matrix is sparse.
- $Q_{ij} = 0$ if nodes i and j are not neighbors. Conditionally independent.
- $Q_{i,i\pm 2} = 0$ for exponential covariance function on a regular grid in time.

Sparse precision matrix $\boldsymbol{Q}$



$$p(Y_7 \mid Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) = p(Y_7 \mid Y_5)$$



$$p(Y_i \mid Y_1, \ldots, Y_{i-1}) = p(Y_i \mid Y_{i-1})$$

This sparseness means that several techniques from numerical analysis can be used. Solve $\boldsymbol{Qb} = \boldsymbol{a}$ quickly for $\boldsymbol{b}$.

Cholesky factorization of $\boldsymbol{Q}$

Common method for sampling and evalution:

$$\boldsymbol{Q} = \begin{bmatrix} Q_{1,1} & \dots & Q_{1,n} \\ \dots & \dots & \dots \\ Q_{n,1} & \dots & Q_{n,n} \end{bmatrix} = \boldsymbol{L}_Q \boldsymbol{L}_Q',$$

Lower triangular matrix

$$\boldsymbol{L}_Q = \begin{bmatrix} L_{Q,1,1} & 0 & \dots & 0 \\ L_{Q,2,1} & L_{Q,2,2} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ L_{Q,n,1} & L_{Q,n,2} & \dots & L_{Q,n,n} \end{bmatrix},$$

The Cholesky factor is often sparse, but not as sparse as $\boldsymbol{Q}$, because it holds only partial conditional structure, according to an ordering. Sparsity is maintained for exponential covariance function in time dimension (Markov).

Sampling and evaluation using $\boldsymbol{L}_Q$

$$\boldsymbol{Q} = \begin{bmatrix} Q_{1,1} & \dots & Q_{1,n} \\ \dots & \dots & \dots \\ Q_{n,1} & \dots & Q_{n,n} \end{bmatrix} = \boldsymbol{L}_Q \boldsymbol{L}_Q',$$

$$\boldsymbol{L}_Q \boldsymbol{Y} = \boldsymbol{Z}.$$

(Previously, for covariance we had $\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{Z}$.)

$$\log|\boldsymbol{Q}| = 2\log|\boldsymbol{L}_Q| = 2\sum_i L_{Q,ii}$$

GMRF for spatial applications.

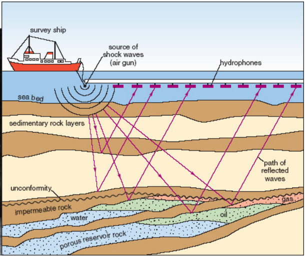
A Markovian model can be constructed for a spatial Gaussian processes (Lindgren et al., 2011).

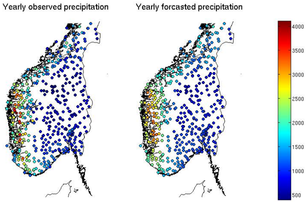
The spatial process is viewed as a stochastic partial differential equation, and the solution is embedded in a triagularized graph over a spatial domain.

More later (31 Jan).

Regression for spatial datasets

Applications with trends and residual spatial variation.





Spatial Gaussian regression model

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

1. $Y(s)$ response variable at location /location-time position $\boldsymbol{s}$.
2. β regression effects. $\boldsymbol{X}(s)$ covariates at $\boldsymbol{s}$.
3. $w(s)$ structured (space-time correlated) Gaussian process.
4. $\epsilon(s)$ unstructured (independent) Gaussian measurement noise.

Gaussian model

Model: $Y(s) = \boldsymbol{X}(s)\beta + w(s) + \epsilon(s)$.
Data at n locations: $\boldsymbol{Y} = (Y(s_1), \ldots, Y(s_n))'$.
Main goals are:

- Parameter estimation
- Prediction

Gaussian model

Likelihood for parameter estimation:

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

$\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{C} = \boldsymbol{\Sigma} + \tau^2\boldsymbol{I}_n$
$\text{Var}(\boldsymbol{w}) = \boldsymbol{\Sigma}$, $\text{Var}(\epsilon(s_i)) = \tau^2$ for all i.
$\boldsymbol{\theta}$ include parameters of the covariance model.

Maximum likelihood

MLE:

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

Analytical derivatives

Formulas for matrix derivatives.

$$
\begin{aligned}
\boldsymbol{Q}(\theta) &= \boldsymbol{C}^{-1} \\
\hat{\beta} &= [\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{Y}, \\
\boldsymbol{Z} &= \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} \\
\frac{d\log|\boldsymbol{C}|}{d\theta_r} &= \mathrm{trace}(\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}) \\
\frac{d\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}}{d\theta_r} &= -\boldsymbol{Z}'\boldsymbol{Q}\frac{d\boldsymbol{C}}{d\theta_r}\boldsymbol{Q}\boldsymbol{Z}.
\end{aligned}
$$

Score and Hessian for $\boldsymbol{\theta}$

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

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

Updates for each iteration

$$\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta}_p)$$

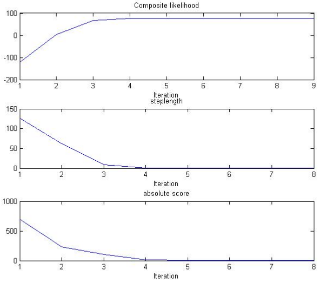
$$\hat{\beta}_p = [\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{Y},$$

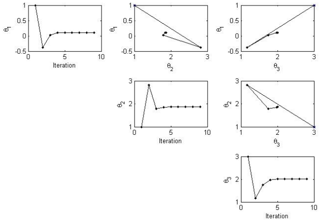
$$\hat{\boldsymbol{\theta}}_{p+1} = \hat{\boldsymbol{\theta}}_p - E\left(\frac{d^2 l(\boldsymbol{Y}; \hat{\beta}_p, \hat{\boldsymbol{\theta}}_p)}{d\boldsymbol{\theta}^2}\right)^{-1} \frac{dl(\boldsymbol{Y}; \hat{\beta}_p, \hat{\boldsymbol{\theta}}_p)}{d\boldsymbol{\theta}},$$

Iterative scheme usually starts from preliminary guess, obatined via summary statistics.

Illustration maximization

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





Asymptotic properties

$$\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}, G^{-1}).$$

$$G = G(\hat{\boldsymbol{\theta}}) = -E\left(\frac{d^2 l}{d\boldsymbol{\theta}^2}\right).$$

Prediction from joint Gaussian formulation

Prediction

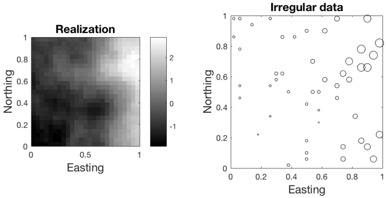
$$\hat{Y}_0 = E(Y_0 | \boldsymbol{Y}) = \boldsymbol{X}_0 \hat{\beta} + \boldsymbol{C}_{0,.} \boldsymbol{C}^{-1} (\boldsymbol{Y} - \boldsymbol{X} \hat{\beta}).$$

$\boldsymbol{C}_{0,.}$ is size $1 \times n$ vector of cross-covariances between prediction site $\boldsymbol{s}_0$ and data sites.
Prediction variance

$$\text{Var}(Y_0 | \boldsymbol{Y}) = C_0 - \boldsymbol{C}_{0,.} \boldsymbol{C}^{-1} \boldsymbol{C}'_{0,.}.$$

Synthetic data

Consider unit square. Create grid of $25^2 = 625$ locations. Use 49 data randomly assigned, or along center line (two designs).

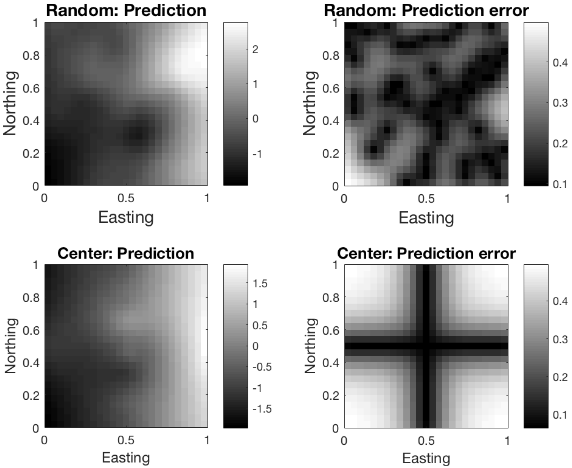


Covariance $C(h) = \tau^2 I(h=0) + \sigma^2(1+\phi h)\exp(-\phi h)$, $h = |\boldsymbol{s}_i - \boldsymbol{s}_j|$.
$\boldsymbol{\theta}$ include transformations of: σ, τ and ϕ.

Likelihood optimization

True parameters $\boldsymbol{\beta} = (-2, 3, 1)$, $\boldsymbol{\theta} = (0.25, 9, 0.0025)$.
Random design:
$\boldsymbol{\beta} = [-2(0.486), 3.43(0.552), 0.812(0.538)]$
$\boldsymbol{\theta} = [\ 0.298(0.118), 7.89(1.98), 0.00563(0.00679)]$
Center design:
$\hat{\boldsymbol{\beta}} = [-2.06(0.576), 3.4(0.733), 0.353(0.733)]$
$\hat{\boldsymbol{\theta}} = [0.255(0.141), 7.19(1.97), 0.00283(0.00128)]$

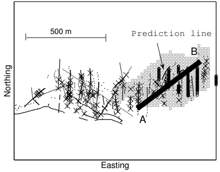
Predictions



Project on spatial Gaussian regression model next week

Mining example

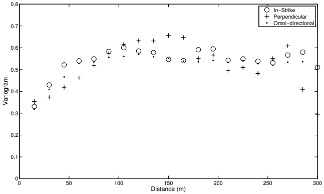
Data giving oxide grade information at $n = 1600$ locations.



Use spatial statistics to predict oxide grade. Question: Will mining be profitable? Should one gather more data before making mining decision?

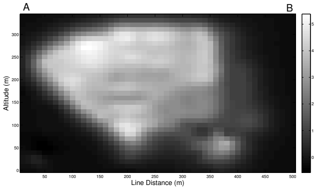
Mining example

Model: $Y(s) = \boldsymbol{X}(s)\beta + w(s) + \epsilon(s)$.
Data at n borehole locations used to fit model parameters by MLE.
Starting values from variogram (related to sample covariance).

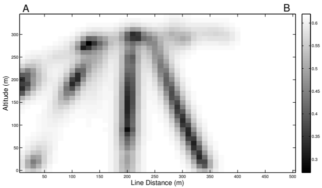


Mining example

Current prediction



,



Mining example

Develop or not? Decision is done using expected profits.
Value based on current data:

$$\mathrm{PV} = \max(p_y, 0), \quad p_y = \boldsymbol{r}^t \boldsymbol{\mu}_{|y} - \boldsymbol{k}^t \mathbf{1},$$

Will more data influence the mining decision? Expected value with more data:

$$\mathrm{PoV} = \int_{\boldsymbol{y}^n} \max(p_{y,y^n}, 0)\pi(\boldsymbol{y}^n|\boldsymbol{y})d\boldsymbol{y}^n, \quad p_{y,y^n} = \boldsymbol{r}^t \boldsymbol{\mu}_{|y,y^n} - \boldsymbol{k}^t \mathbf{1},$$

Is more data valuable. Compute the value of information (VOI):

$$\mathrm{VOI} = \mathrm{PoV} - \mathrm{PV}.$$

If VOI exceeds the cost of data, it is worthwhile gathering this information.

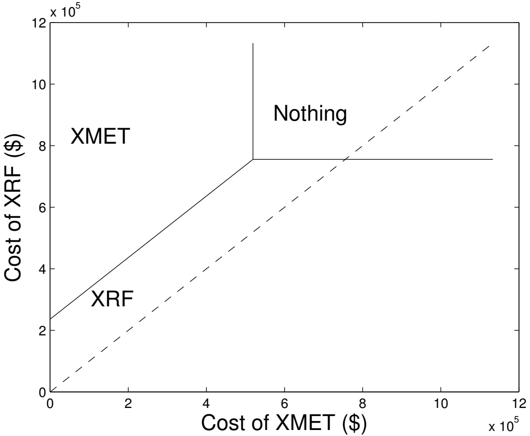
(Computations partly analytical, similar to EI.)

Mining example

Use this to compare two data types in planned boreholes:

- XRF : scanning in a lab (expensive, accurate).
- XMET : handheld device for grade scanning at the mining site (inexpensive, inaccurate).

Mining example



Schedule

- 17 Jan: Gaussian processes
- 24 Jan: Hands-on project on Gaussian processes
- 31 Jan: Latent Gaussian models and INLA.
- 7 Feb: Hands-on project on INLA
- 12-13 Feb: Template model builder. (Guest lecturer Hans J Skaug)

Schedule

Project (24 Jan) will include

- Expected improvement for function maximization.
- Parameter estimation in spatial regression model (forest example).