Repetition: Gaussian process

A Gaussian process is a generalisation of a multivariate Gaussian distribution to infinitely many variables.

A Gaussian distribution is fully specified by a mean vector, $\boldsymbol{\mu}$, and a covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

A Gaussian process is fully specified by a mean function and a covariance function $\gamma(x, x')$:

$$Y(x) \sim GP(\mu(x), \gamma(x, x'))$$

Gaussian process (GP)

For any set $x_1, \ldots, x_n$ we have that $Y(x_1), \ldots, Y(x_n)$ is jointly multivariate Gaussian with mean $\mu(x_i)$, $i = 1, \ldots, n$ and

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

Here, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ can represent time points, spatial locations, etc. The covariance matrix $\boldsymbol{\Sigma}$ is thereby constructed from the covariance function.

Gaussian processes in applications

Gaussian processes or Gaussian random field models are a key construction in geostatistics and often used as a building block in hierarchical models.

- Radioactivity counts: Poisson. The log intensity can be modeled as a GP.
- Number of days with rain, or number of infections: Binomial. The logit probability is modelled as a GP.

The spatial regression model for continuous response data relies on Gaussian variables with linear asscociation.

Gaussian process regression

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

1. $Y(s)$ response variable at 'location' $\boldsymbol{s}$.
2. $\boldsymbol{\beta}$ regression effects. $\boldsymbol{X}(s)$ covariates at $\boldsymbol{s}$.
3. $w(s)$ structured (space-time correlated) GP with 0 mean.
4. $\epsilon(s)$ unstructured (independent) Gaussian measurement noise with 0 mean and $\text{Var}(\epsilon(s)) = \tau^2$. (Nugget effect)

Gaussian model

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

Assume that we can observe the spatial process $Y(s)$ and the associated covariates $\boldsymbol{X}'(s)$ at n 'locations': $s_1, \ldots, s_n$.

Main goals are:

- Parameter estimation
- Prediction

Gaussian model

Log-likelihood for parameter estimation:

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

with $\boldsymbol{C} = \boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} + \tau^2 \boldsymbol{I}_n$

$\boldsymbol{\theta}$ 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.

$$\begin{aligned}
\boldsymbol{Q}(\boldsymbol{\theta}) &= \boldsymbol{C}^{-1} \\
\hat{\beta} &= [\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{Y},
\end{aligned}$$

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

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

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

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

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

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

Prediction from joint Gaussian formulation

Prediction

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

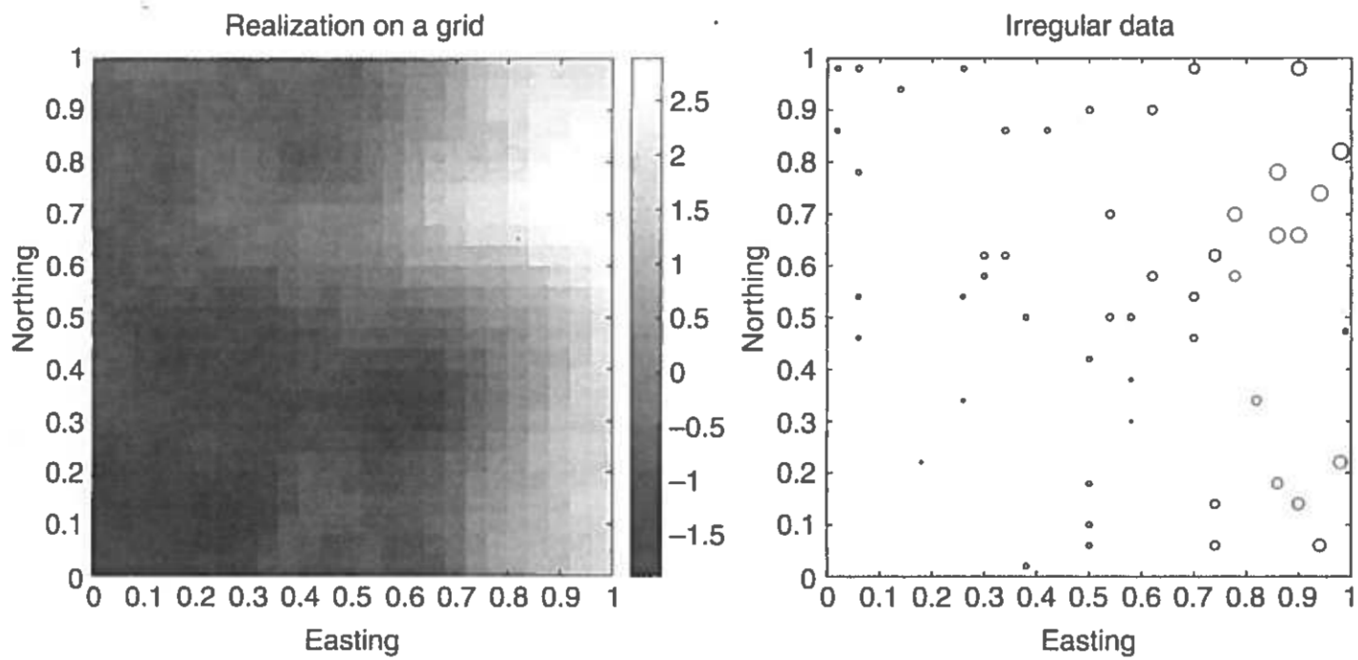
$\boldsymbol{C}_{0,\cdot}$ is size $N \times n$ matrix of cross-covariances between prediction sites $\boldsymbol{s}_0 = (s_{0,1}, \ldots, s_{0,N})$ and the n observation sites.

Prediction variance

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

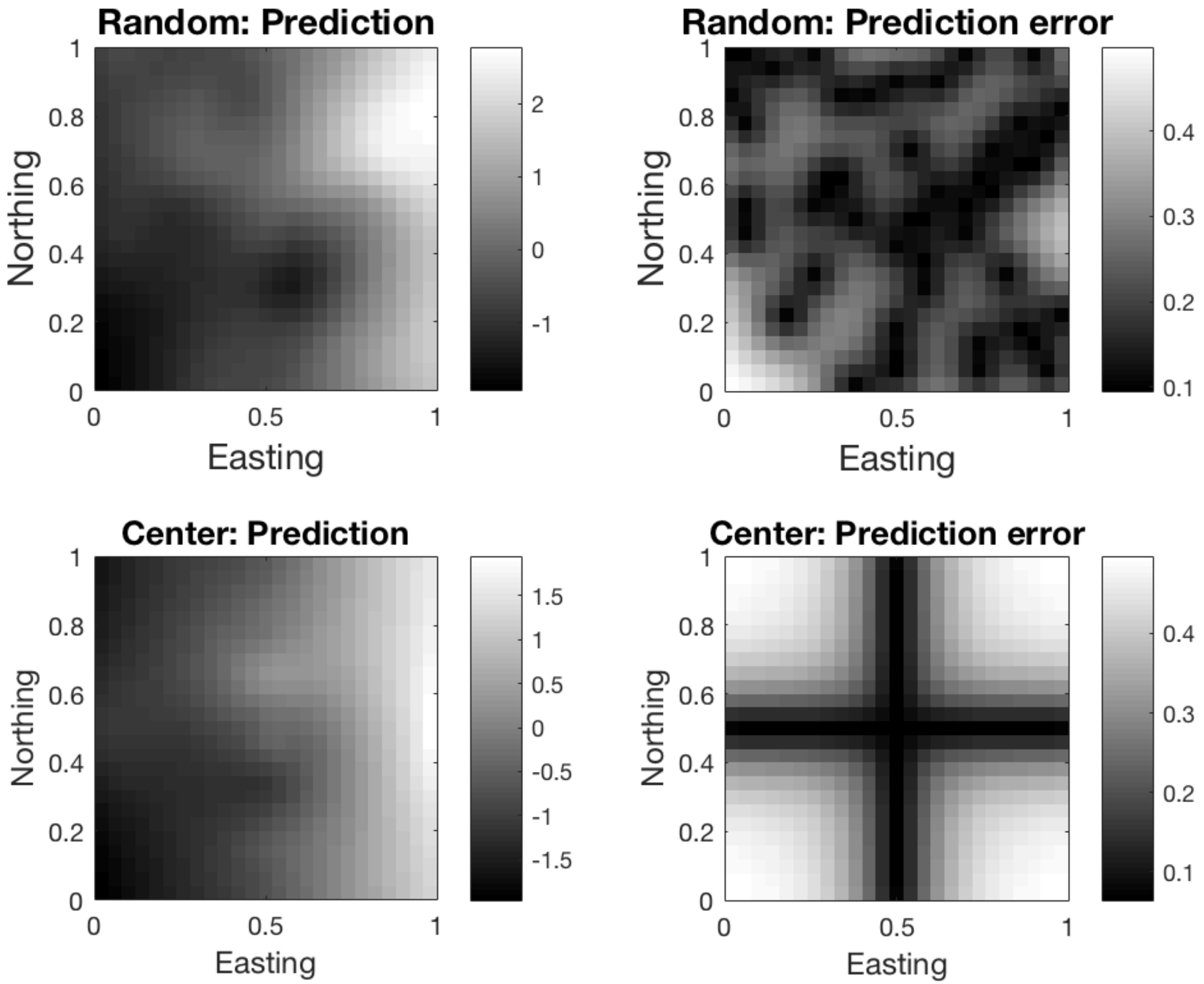
Synthetic data

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



Covariance $C(h) = \tau^2 I(h = 0) + \sigma^2(1 + \phi h)\exp(-\phi h)$ with $h = |\boldsymbol{s}_i - \boldsymbol{s}_j|$

Predictions



Likelihood optimization

True parameters:

$\boldsymbol{\beta} = (\beta_{\text{Intercept}} = -2, \beta_{\text{easting}} = 3, \beta_{\text{northing}} = 1)^\top,$

$\boldsymbol{\theta} = (\sigma^2 = 0.25, \phi = 9, \tau^2 = 0.0025).$

Parameter estimates based on random design:

$\hat{\boldsymbol{\beta}} = [-2(0.486), 3.43(0.552), 0.812(0.538)]$

$\hat{\boldsymbol{\theta}} = [\ 0.298(0.118), 7.89(1.98), 0.00563(0.00679)]$

Parameter estimates based on center design:

$\hat{\boldsymbol{\beta}} = [-2.06(0.576), 3.4(0.733), 0.353(0.733)]$

$\hat{\boldsymbol{\theta}} = [0.255(0.141), 7.19(1.97), 0.00283(0.00128)]$

Computational challenge for large n

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

In general, the computational cost is $O(n^3)$.

Outlook: Next lecture

Geir-Arne Fuglstad will present us next week various techniques to deal with the large n problem in spatial statistics.

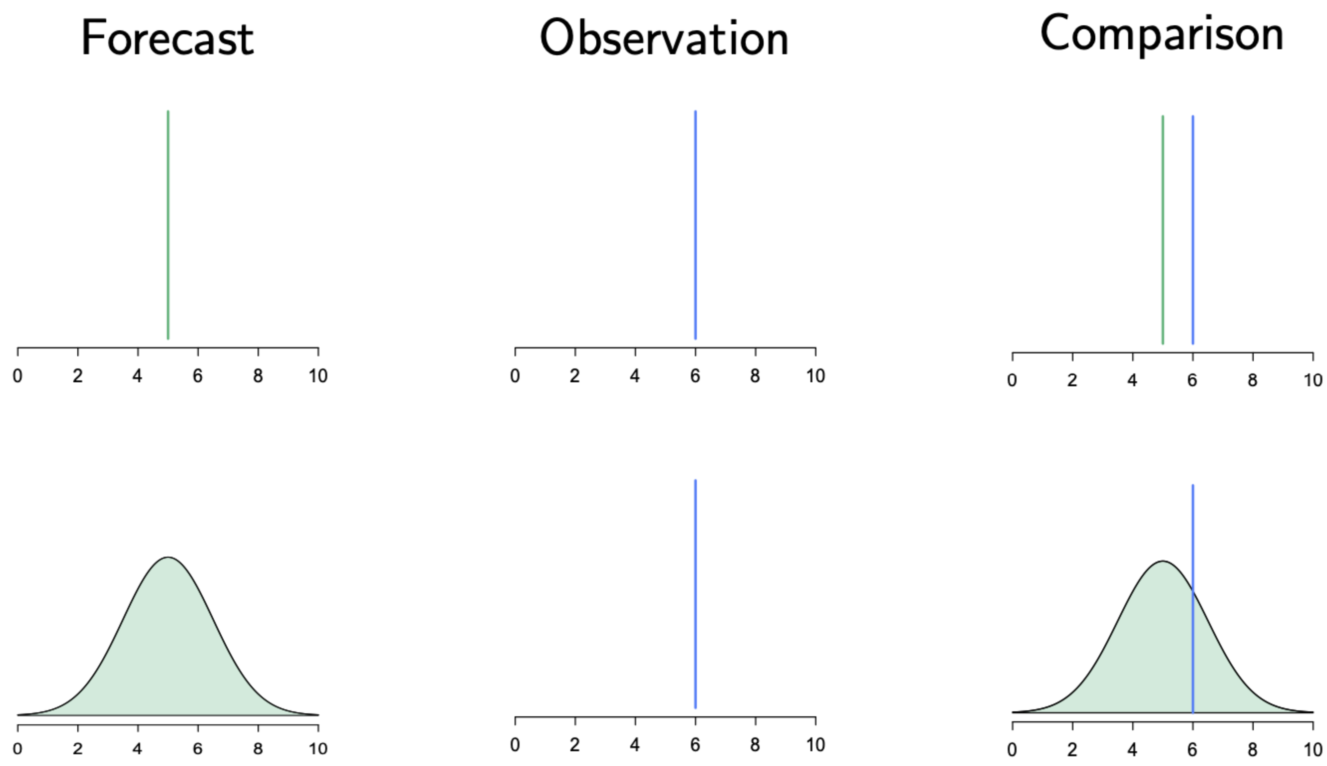
Assessment of predictions

We have showed how to generate predictions. However, how can we assess the performance of our prediction model, how do we choose between two covariance functions for example?

We will discuss scoring rules, for which one of the main references is:

> *Tilmann Gneiting & Adrian E Raftery (2007) Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102:477, 359-378*

Point versus probabilistic predictions



Point forecasts

The forecast x is evaluated against the observation y using a scoring function such as:

- Squared error: $S(x, y) = (x - y)^2$
- Absolute error: $S(x, y) = |x - y|$

Different methods that issue forecasts at a finite set of locations are compared using their mean scores.

Probabilistic forecasts

Using point forecasts we ignore the uncertainty that comes along with our predictions. Thus, ideally we would like to evaluate a probabilistic forecast F.

Two main evaluation criteria:

- Sharpness: Refers to the concentration of the predictive distribution F
- Calibration: Agreement of the true values and the predictive distribution

Proper scoring rules assess calibration and sharpness, and are often negatively oriented, i.e. the lower, the better.

Popular examples

The logarithmic score is strictly proper and defined as

$$\mathrm{LogS}(F, y) = -\log(f(y))$$

The continuous ranked probability score (CRPS) is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - I\{y \leq z\})^2 dz$$

If the first moment of F is finite the CRPS can be written as.

$$\mathrm{CRPS}(F, y) = \frac{1}{2} E_F|X - X'| - E_F|X - y|$$

where X and X' are independent copies of a random variable with distribution function F. The CRPS generalizes the absolute error for deterministic forecasts.

Assessment using simulated forecast distributions

For example using MCMC, at each iteration i, $i = 1, \ldots, M$ we get parameter samples $\boldsymbol{\theta}^{(i)}$, which can be used to additionally draw a value of $y^{(i)}$. The CRPS can then be computed via

$$CRPS(\hat{F}, y) = \frac{1}{M} \sum_{i=1}^{M} |y^{(i)} - y| - \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} |y^{(i)} - y^{(j)}|$$

Monte-Carlo estimation of the logarithmic score is possible if at least $f(y|\boldsymbol{\theta})$ is available:

$$\hat{f}(y) = \frac{1}{M} \sum_{i=1}^{M} f(y|\boldsymbol{\theta}^{(i)})$$