Gaussian processes in spatial statistics

Emiko Dupont

29 January 2018
What is a Gaussian process/Gaussian random field?

**Definition:**
Stochastic process \( \{ Z(s) \mid s \in D \} \), \( D \subset \mathbb{R}^d \) spatial domain

Any finite collection \( \{ Z(s_1), \ldots, Z(s_k) \} \) is multivariate normal:

\[
\begin{bmatrix}
Z(s_1) \\
\vdots \\
Z(s_k)
\end{bmatrix} \sim N
\begin{pmatrix}
\mu(s_1) \\
\vdots \\
\mu(s_k)
\end{pmatrix},
\begin{pmatrix}
\text{Cov}(Z(s_i), Z(s_j))
\end{pmatrix}
\]
What is a Gaussian process/Gaussian random field?

Definition:
Stochastic process \( \{Z(s) \mid s \in D\} \), \( D \subset \mathbb{R}^d \)

Any finite collection \( \{Z(s_1), \ldots, Z(s_k)\} \) is multivariate normal:

\[
\begin{bmatrix}
Z(s_1) \\
\vdots \\
Z(s_k)
\end{bmatrix}
\sim N
\left(
\begin{bmatrix}
\mu(s_1) \\
\vdots \\
\mu(s_k)
\end{bmatrix},
\begin{bmatrix}
\text{Cov}(Z(s_i), Z(s_j))
\end{bmatrix}
\right)
\]

Note:
In particular, \( Z(s) \sim N(\mu(s), \text{Var}(s)) \) for all \( s \in D \)
What is a Gaussian process/Gaussian random field?

Spatial field: \( \{Z(s) \mid s \in D\}, \ D \subset \mathbb{R}^2 \)
- White noise
  - \( Z(s) \sim_{\text{iid}} N(0, \sigma^2) \)
  - Any finite collection \( \{Z(s_1), \ldots, Z(s_k)\} \sim N(\mathbf{0}, \sigma^2 I) \)
What is a Gaussian process/Gaussian random field?

Spatial field: \( \{ Z(s) \mid s \in D \} \), \( D \subset \mathbb{R}^2 \)
- \( Z(s) \) = concentration of mineral at location \( s \)
  - \( \mu(s) = \mu \)
  - \( \text{Cov}(Z(s_1), Z(s_2)) = \exp(-|s_2 - s_1|^2 / R^2) \)

Emiko Dupont
Gaussian processes in spatial statistics
What is a Gaussian process/Gaussian random field?

Spatial field: \( \{ Z(i) \mid i = 1, \ldots, N \} \), \( N \) regions
- \( Z(i) = \) relative risk of lung cancer in region \( i \)
- Covariance: Neighbouring regions more similar than those far apart
What are Gaussian processes used for?

**Improve inference:**
- Identify spatial correlation structure/clustering
- More powerful inference by pooling data

**Prediction:** Given observations of $Z(s)$ at locations $s_1, \ldots, s_n$
- Estimate $\int_A Z(s) ds$ (e.g. total quantity of ore across region $A$)
- Reconstruct entire field $Z(s)$ (e.g. global sea surface temperature)
Applications

- geology (e.g. estimating mineral concentration for mining)
- environmental sciences (e.g. assessing time trends/spatial trends in flood risk/sea ice concentration/sea temperature...)
- ecology (e.g. assess fish stock to avoid overexploitation)
- epidemiology (e.g. understanding spatial distribution of diseases)
- econometrics (e.g. financial time series modelling)
- ...

Emiko Dupont
Gaussian processes in spatial statistics
Gaussian process models

What’s so special about Gaussians?

- A Gaussian is completely determined by its mean and covariance.
- Gaussians behave nicely under addition, conditioning etc.
- Gaussians are often good approximations of other distributions.
Gaussian process models

What’s so special about Gaussians?
- A Gaussian is completely determined by its mean and covariance
- Gaussians behave nicely under addition, conditioning etc.
- Gaussians are often good approximations of other distributions

Common assumption:
- **Isotropy**: Covariance depends only on $|s_1 - s_2|$

\[
\text{Cov}(Z(s_1), Z(s_2)) = C(|s_1 - s_2|)
\]

- Typically: nearby points are more similar than those far apart
Estimating the spatial structure

Given \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \).

**Assumption**: Mean and variance known up to unknown parameters.

**Goal**: Estimate parameters
Estimating the spatial structure

Given \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \).

Model:

\[
    z \mid \beta, \alpha, \theta \sim N(X \beta, \alpha V(\theta))
\]

\( X \) observed covariates at locations \( s_1, \ldots, s_n \)

For example:

\( Z(s) = \) sea surface temperature at location \( s \)
\( X = \) salinity at locations \( s_1, \ldots, s_n \)
Exponential covariance function with unknown range parameter \( \theta = R \)
Estimating the spatial structure

Given \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \).

**Model:**

\[
    z \mid \beta, \alpha, \theta \sim N(X\beta, \alpha V(\theta))
\]

\( X \) observed covariates at locations \( s_1, \ldots, s_n \)

**Parameter estimation:**

- Maximum likelihood: \((\hat{\beta}, \hat{\alpha}, \hat{\theta}) = \arg\max f(z \mid \beta, \alpha, \theta)\)
- Bayesian method: posterior \( \propto \) prior \( \times \) likelihood
Prediction: Kriging

Goal:
Given: \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \)
Predict \( z_0 = Z(s_0) \) in unobserved location \( s_0 \)
Prediction: Kriging

**Goal:**
Given: \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \)
Predict \( z_0 = Z(s_0) \) in unobserved location \( s_0 \)

**Assumption:** Covariance structure is known

**Model**

\[
\begin{align*}
z & \sim N(X \beta, \Sigma), \\
\tilde{z}_0 & \sim N(x_0^T \beta, \sigma_0^2), \\
\text{Cov}(z, \tilde{z}_0) &= \tau
\end{align*}
\]

\( x_0, X = \) observed covariates at locations \( s_0, s_1, \ldots, s_n \)
\( \beta = \) unknown coefficients of covariates
\( \sigma_0^2, \tau, \Sigma = \) known covariances
Prediction: Kriging

**Goal:**
Given: \( z = (z_1, \ldots, z_n) \) observations of \( Z(s) \) at locations \( s_1, \ldots, s_n \)
Predict \( z_0 = Z(s_0) \) in unobserved location \( s_0 \)

**Prediction:** Choose \( \hat{z}_0 = \lambda^T z \) so that

- \( \hat{z}_0 \) is unbiased (\( E(\hat{z}_0) = z_0 \))
- Mean squared prediction error \( E((z_0 - \hat{z}_0)^2) = \text{Cov}(\hat{z}_0) \) is minimised
Tools for estimation and prediction of Gaussian processes

**Frequentist methods**
- Directly optimise likelihood/REML/prediction error
- nlme (linear mixed model formulation of Gaussian process) (uses ML or REML)
- mgcv (GAM formulation) (uses penalised likelihood method)

**Bayesian methods**
- Markov Chain Monte Carlo (WinBUGS/JAGS/Stan)
- INLA for Gaussian Markov random fields (GMRFs) (uses integrated nested Laplace approximation)
References


