Gaussian processes in spatial statistics

Emiko Dupont

29 January 2018

What is a Gaussian process/Gaussian random field?

Definition:

Stochastic process $\{Z(s) \mid s \in D\}$, $\underbrace{D}_{\substack{\text{spatial}\\\text{domain}}} \subset \mathbb{R}^d$

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

$$\begin{bmatrix} Z(s_1) \\ \vdots \\ Z(s_k) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu(s_1) \\ \vdots \\ \mu(s_k) \end{bmatrix}, \begin{bmatrix} & & \\ & \mathrm{Cov}(Z(s_i), Z(s_j)) & \\ & & \end{bmatrix} \right)$$

What is a Gaussian process/Gaussian random field?

Definition:

Stochastic process $\{Z(s) \mid s \in D\}$, $\underbrace{D}_{\substack{\text{spatial}\\ \text{domain}}} \subset \mathbb{R}^d$

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

$$\begin{bmatrix} Z(s_1) \\ \vdots \\ Z(s_k) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu(s_1) \\ \vdots \\ \mu(s_k) \end{bmatrix}, \begin{bmatrix} \quad \mathrm{Cov}(Z(s_i), Z(s_j)) \quad \end{bmatrix} \right)$$

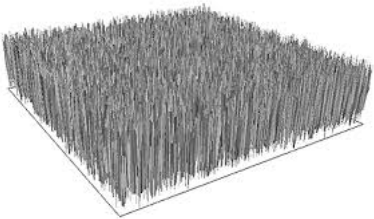
Note:

In particular, $Z(s) \sim N(\mu(s), \mathrm{Var}(s))$ for all $s \in D$

What is a Gaussian process/Gaussian random field?

Spatial field: $\{Z(s) \mid s \in D\}$, $D \subset \mathbb{R}^2$

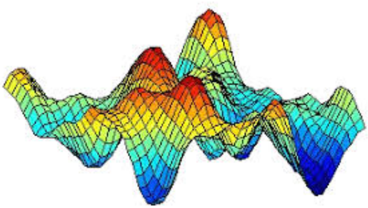
- White noise
  - $Z(s) \sim_{\text{iid}} N(0, \sigma^2)$
  - Any finite collection $\{Z(s_1), \ldots, Z(s_k)\} \sim N(\mathbf{0}, \sigma^2 I)$



What is a Gaussian process/Gaussian random field?

Spatial field: $\{Z(s) \mid s \in D\}$, $D \subset \mathbb{R}^2$

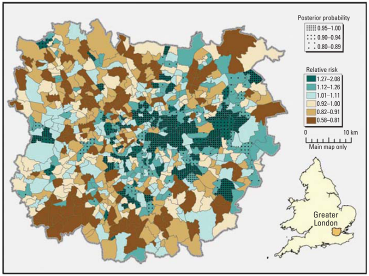
- $Z(s)$ = concentration of mineral at location s
  - $\mu(s) = \mu$
  - $\mathrm{Cov}(Z(s_1), Z(s_2)) = \exp(-|s_2 - s_1|^2 / \underbrace{R^2}_{\text{range parameter}})$



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)
- ...

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 $|s_1 - s_2|$

$$\mathrm{Cov}(Z(s_1), Z(s_2)) = \underbrace{C(|s_1 - s_2|)}_{\substack{\text{covariance function} \\ \text{e.g. exponential/spherical/Matern}}}$$

- 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:

$$z \mid \underbrace{\beta, \alpha, \theta}_{\substack{\text{unknown}\\ \text{parameters}}} \sim N(X\beta, \alpha V(\theta))$$

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:

$$z \mid \underbrace{\beta, \alpha, \theta}_{\substack{\text{unknown}\\\text{parameters}}} \sim N(X\beta, \alpha V(\theta))$$

X observed covariates at locations $s_1, \ldots, s_n$

Parameter estimation:

- Maximum likelihood: $(\hat{\beta}, \hat{\alpha}, \hat{\theta}) = \text{argmax}\ \underbrace{f(z \mid \beta, \alpha, \theta)}_{\text{likelihood of data}}$
- Bayesian method: posterior $\propto$ prior $\times$ likelihood

Prediction: Kriging

Goal:
Given: $z = (z_1, \dots, z_n)$ observations of $Z(s)$ at locations $s_1, \dots, s_n$
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

$$z \sim N(X\beta, \Sigma), \quad z_0 \sim N(x_0^T\beta, \sigma_0^2), \quad \mathrm{Cov}(z, z_0) = \tau$$

$x_0, X =$ 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 $(\mathrm{E}(\hat{z}_0) = z_0)$
- Mean squared prediction error $\mathrm{E}((z_0 - \hat{z}_0)^2) = \mathrm{Cov}(\hat{z}_0)$ 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

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of spatial statistics*, CRC press, 2010.

C. Paciorek, *Technical vignette 3: Kriging, interpolation, and uncertainty: Department of biostatistics*, Harvard School of Public Health, Version, 1 (2008).

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren, *Bayesian computing with inla: A review*, arXiv preprint arXiv:1604.00860, (2016).

R. L. Smith, *Environmental statistics*, Facultad de Ciencias Económicas, Universidad Nacional del Cuyo, 1999.

Y. Sun, B. Li, and M. G. Genton, *Geostatistics for large datasets*, in Advances and challenges in space-time modelling of natural events, Springer, 2012, pp. 55–77.

S. N. Wood, *Generalized additive models: an introduction with R*, CRC press, 2017.