Gaussian process regression

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

1. $Y(s)$ response variable at 'location' $\boldsymbol{s}$.
2. β regression effects. $\boldsymbol{X}(s)$ covariates at $\boldsymbol{s}$.
3. $w(s)$ structured (space-time correlated) Gaussian process with 0 mean.
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:

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

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

$$\boldsymbol{Q} = \boldsymbol{Q}(\theta_p)$$

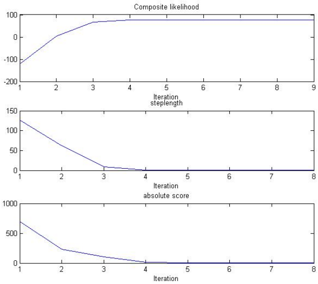
$$\hat{\beta}_p = [\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{Y},$$

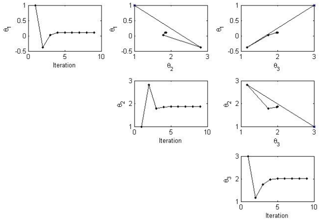
$$\hat{\boldsymbol{\theta}}_{p+1} = \hat{\boldsymbol{\theta}}_p - E\left(\frac{d^2 l(\boldsymbol{Y}; \hat{\beta}_p, \hat{\boldsymbol{\theta}}_p)}{d\boldsymbol{\theta}^2}\right)^{-1} \frac{dl(\boldsymbol{Y}; \hat{\beta}_p, \hat{\boldsymbol{\theta}}_p)}{d\boldsymbol{\theta}},$$

Iterative scheme usually starts from preliminary guess, obtained 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

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

Information matrix:

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

Prediction from joint Gaussian formulation

Prediction

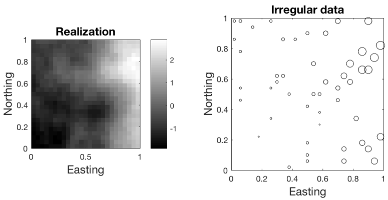
$$\hat{Y}_0 = E(Y_0|\boldsymbol{Y}) = \boldsymbol{X}_0\hat{\beta} + \boldsymbol{C}_{0,.}\boldsymbol{C}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}).$$

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

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

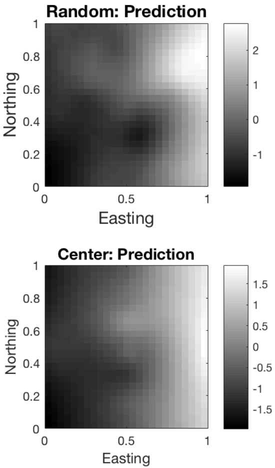
Synthetic data

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.

Predictions



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:
$\hat{\boldsymbol{\beta}} = [-2.06(0.576), 3.4(0.733), 0.353(0.733)]$
$\hat{\boldsymbol{\theta}} = [0.255(0.141), 7.19(1.97), 0.00283(0.00128)]$

Computational challenge for large n

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 $(\boldsymbol{Y} - \boldsymbol{X}\beta)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta)$
4. Factorize required matrices.

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

Possible solutions for large Gaussian models

- Approximate likelihood, Composite likelihoods.
- Basis representation.
- Markov representation.
- Predictive process models, sparse GPs.
- Tapered likelihood.
- Numerical linear algebra.

Composite likelihood

- Use pairs of joints, not full joint.

$$l_{cl}(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) = \sum_{i}\sum_{j>i} \log f(Y(s_i), Y(s_j); \boldsymbol{\beta},\boldsymbol{\theta})$$

- Fast calculations & Quantify loss in efficiency & Allows parallel computing.

M blocks.

$$\begin{aligned} l_{CL}(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) &= \sum_{k=1}^{M-1}\sum_{l>k} \log f(\boldsymbol{Y}_k, \boldsymbol{Y}_l; \boldsymbol{\beta},\boldsymbol{\theta}) \\ &= \sum_{k=1}^{M-1}\sum_{l>k} \{-\frac{1}{2}\log|\boldsymbol{\Sigma}_{kl}| - \frac{1}{2}\boldsymbol{Z}'_{kl}\boldsymbol{Q}_{kl}\boldsymbol{Z}_{kl}\}, \end{aligned}$$

$\boldsymbol{Z}_{kl} = (\boldsymbol{Y}_k, \boldsymbol{Y}_l)' - (\boldsymbol{X}_k, \boldsymbol{X}_l)'\beta$
$\boldsymbol{\Sigma}_{kl} = \boldsymbol{\Sigma}_{kl}(\theta)$ block-pair covariance. Size $(n_k + n_l) \times (n_k + n_l)$
$\boldsymbol{Q}_{kl} = \boldsymbol{\Sigma}_{kl}^{-1}$
$n = \sum_{k=1}^{M} n_k$

Asymptotic properties: Godambe sandwich

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

$$\begin{aligned} G = G(\hat{\theta}) &= H(\hat{\theta}) J^{-1}(\hat{\theta}) H(\hat{\theta}), \\ H(\hat{\theta}) &= -E\left(\frac{d^2 l_{CL}}{d\theta^2}\right), \quad J(\hat{\theta}) = Var\left(\frac{d l_{CL}}{d\theta}\right). \end{aligned}$$

Markov property

In the time domain, the Markov property holds if for any $t > s > u$,

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

The exponential correlation function gives a Markov process.
(Proof by trivariate distribution, and conditioning.)

Precision matrix Q : inverse covariance matrix

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

$\boldsymbol{Q}$ 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 $p(\boldsymbol{Y}_A | \boldsymbol{Y}_B) \propto p(\boldsymbol{Y}_A, \boldsymbol{Y}_B)$.)

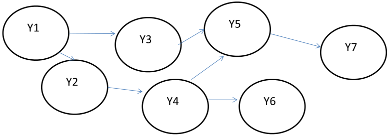
Algebraically equivalent forms

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

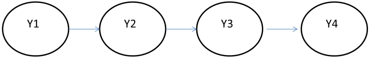
$$
\mathrm{E}(\boldsymbol{Y}_A|\boldsymbol{Y}_B) = \boldsymbol{\mu}_A - \boldsymbol{Q}_A^{-1}\boldsymbol{Q}_{A,B}(\boldsymbol{Y}_B - \boldsymbol{\mu}_B),
$$

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

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.
- $Q_{i,i+2} = 0$ for exponential covariance function on a regular grid in time.

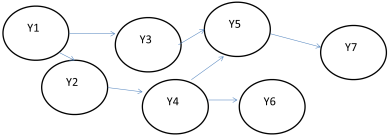
Conditional independence via $\boldsymbol{Q}$

All other variables than y_i are denoted $\boldsymbol{y}_{-i}$.
Neighborhood of node i is denoted $\mathcal{N}_i$.

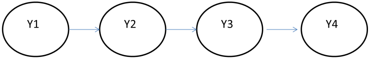
$$p(y_i|\boldsymbol{y}_{-i}) = p(y_i|y_j; j \in \mathcal{N}_i)$$

The neighborhood structure is given by the non-zero entries in $\boldsymbol{Q}$.

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 the partial (ordered) conditional structure, according to an ordering. This gives 'fill in'.
The ordering matters in how the fill-in takes place.

Sparse L_Q

L_Q is related to a recursion:

$$p(\boldsymbol{y}) = p(y_n)p(y_{n-1}|y_n)\dots p(y_1|y_2,\dots,y_n)$$

Which can be removed in the conditioning? If $L_{Q,i,j} = 0$, it can be removed.
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 (SPDE), and the solution is embedded in a triagularized graph over a spatial domain.
More later (23 Jan).