Sampling and modelling of LARGE Gaussian processes

Colin Fox

http://www.physics.otago.ac.nz/

Outline

CD Sampling of Gaussian distributions:

- Conjugate vectors
- Direct samplers
- Conjugate direction sampling algorithm

Modelling covariance via PDEs in Hilbert space:

- A locally linear GMRF and the discrete PDE equivalent
- Covariance functions and Green's functions
- Non-stationary covariance model
- Comparison with Higdon/Gibbs/Paciorek construction

Conjugate Gradient Optimization

- Basic CG algorithm was introduced by Hestenes and Stiefel (1952)
- Solves the matrix equation $Ax = b$ where A is a symmetric positive definite (SPD) matrix
- Equivalently minimizes the quadratic form

$$Q(x) = \frac{1}{2} x^{\mathsf{T}} A x - b^{\mathsf{T}} x$$

  using a sequence of search directions derived from the gradient of Q projected onto the space conjugate to all previous directions
- Terminates in $\leq \dim(x)$ steps (exact arithmetic)
- Actually in number distinct eigenvalues of A (Jennings 1977)
- Roundoff error causes loss of conjugacy even earlier (Meurant 2006)
- Requires storage of two vectors, and one operation by A per iteration.
- Useful for LARGE problems, e.g. solving discrete PDE in 3D

Normal distributions and Gaussian random fields

Gaussian density with precision matrix $A \in \mathbb{R}^{m \times m}$, SPD, $\dim(x) = m$.

$$\pi(x) = \sqrt{\frac{\det(A)}{2\pi^m}} \exp\left\{-\frac{1}{2} x^\mathsf{T} A x\right\}$$

Covariance matrix is $\Sigma = A^{-1}$ is also SPD. Write $x \sim \mathcal{N}(0, \Sigma)$.

Gaussian random field over sites $s_1, s_2, \ldots, s_m \in \mathbb{R}^n$ is

$$x = \begin{pmatrix} x(s_1) \\ x(s_2) \\ \vdots \\ x(s_m) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \Sigma\right)$$

where

$$\Sigma_{ij} = R\left(\|s_i - s_j\|\right)$$

for a positive function R (guaranteeing Σ SPD).

Mutually conjugate vectors (and factorizations)

Definition 1 *A set of non-zero vectors* $p^{(i)}$, $i = 0, 1, \ldots, m-1$, *is mutually conjugate w.r.t.* A *(SPD) if*

$$p^{(i)\mathsf{T}} A p^{(j)} = 0 \quad \forall i \neq j.$$

Note: $d_i = p^{(i)\mathsf{T}} A p^{(i)} > 0 \; \forall i$, as A strictly PD. Hence, if $P \in R^{m \times m}$ is the matrix with columns $p^{(i)}$, $i = 0, 1, \ldots, m-1$, and $D = diag\,(d_0, d_1, \ldots, d_{m-1})$ (also SPD), then

$$P^{\mathsf{T}} A P = D.$$

If $z \sim \mathcal{N}(0, I_m)$ then $y = \sqrt{D^{-1}} z \sim \mathcal{N}\left(0, D^{-1}\right)$ and

$$x = Py \sim \mathcal{N}(0, \Sigma)$$

This follows since $P^{\mathsf{T}} A P = D$ so $A^{-1} = P D^{-1} P^{\mathsf{T}}$.

Can write $x = \sum_{i=0}^{m-1} \left(z_i / \sqrt{d_i}\right) p^{(j)}$. Since the z_i are i.i.d. $\sim \mathcal{N}(0, 1)$, this shows how a sample from π may be generated using a sequence of standard normal random numbers.

Some examples you know of

Cholesky factorizaton of Σ: Since $RR^\top = \Sigma$, $R^\top AR = I$, so the columns of R are mutually conjugate w.r.t. A.

Cholesky factorizaton of A: Since $LL^\top = A$, $\left(L^\top\right)^{-1} L^{-1} = A^{-1}$, so $\left(L^\top\right)^{-1} = R$ and so the columns of $\left(L^\top\right)^{-1}$ are mutually conjugate w.r.t. A.

Eigen decomposition of A or Σ: Consider

$$A = UDU^\top$$

where U is unitary (columns are normalized eigenvectors) and D is diagonal (eigenvalues on diagonal). Then $U^\top AU = D$, so the eigenvectors are mutually conjugate w.r.t. A.

These examples show that the algorithms based on the Cholesky and eigen factorizations are examples of the more general notion of mutually conjugate w.r.t. A, and show that these algorithms do indeed sample from π.

CD sampling algorithm for Gaussians

$\underline{\mathbf{x}, \mathbf{b} = \mathrm{cdsample}(A)}$

initialize $\mathbf{x}^0$ and $\mathbf{b}^0$ (with $A\mathbf{x}^0 \neq \mathbf{b}^0$). Let $\mathbf{r}^0 \leftarrow \mathbf{b}^0 - A\mathbf{x}^0$, $\mathbf{p}^0 \leftarrow \mathbf{r}^0$

for $k = 1$ to m do

$\mathbf{q}^{k-1} \leftarrow A\mathbf{p}^{k-1}$

$d^{k-1} \leftarrow \mathbf{q}^{(k-1)\mathsf{T}}\mathbf{p}^{k-1}$, $e^{k-1} \leftarrow \mathbf{q}^{(k-1)\mathsf{T}}\mathbf{x}^{k-1}/d^{k-1}$, $f^{k-1} \leftarrow \mathbf{p}^{(k-1)\mathsf{T}}\mathbf{b}^{k-1}/d^{k-1}$

draw $z^{k-1} \sim N(0,1)$, $\alpha^{k-1} \leftarrow z^{k-1}/\sqrt{d^{k-1}}$

$\mathbf{x}^k \leftarrow \mathbf{x}^{k-1} + \left(\alpha^{k-1} - e^{k-1}\right)\mathbf{p}^{k-1}$

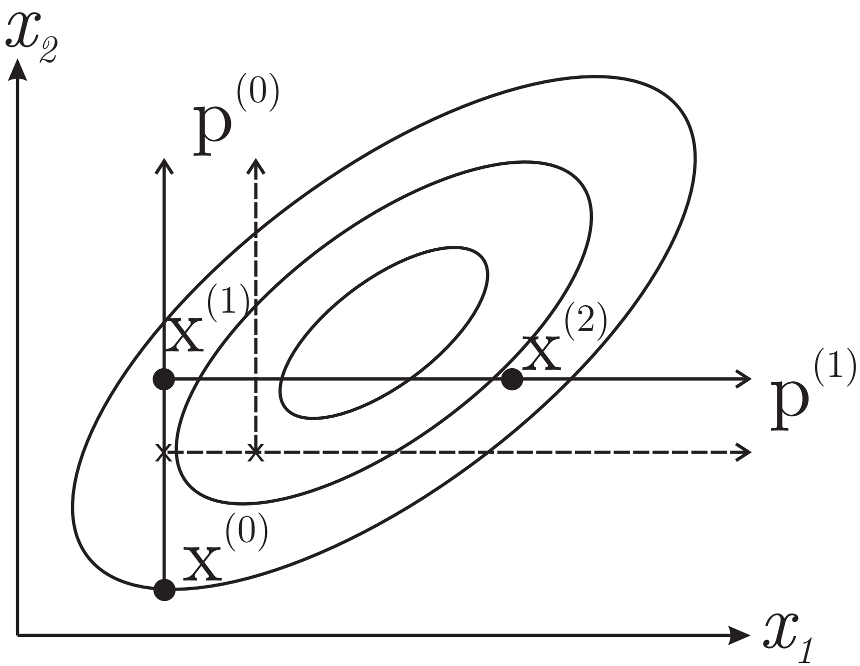
$\mathbf{b}^k \leftarrow \mathbf{b}^{k-1} + \left(\alpha^{k-1} - f^{k-1}\right)\mathbf{q}^{k-1}$

$\mathbf{r}^k \leftarrow \mathbf{r}^{k-1} - (f^{k-1} - e^{k-1})\mathbf{q}^{k-1}$

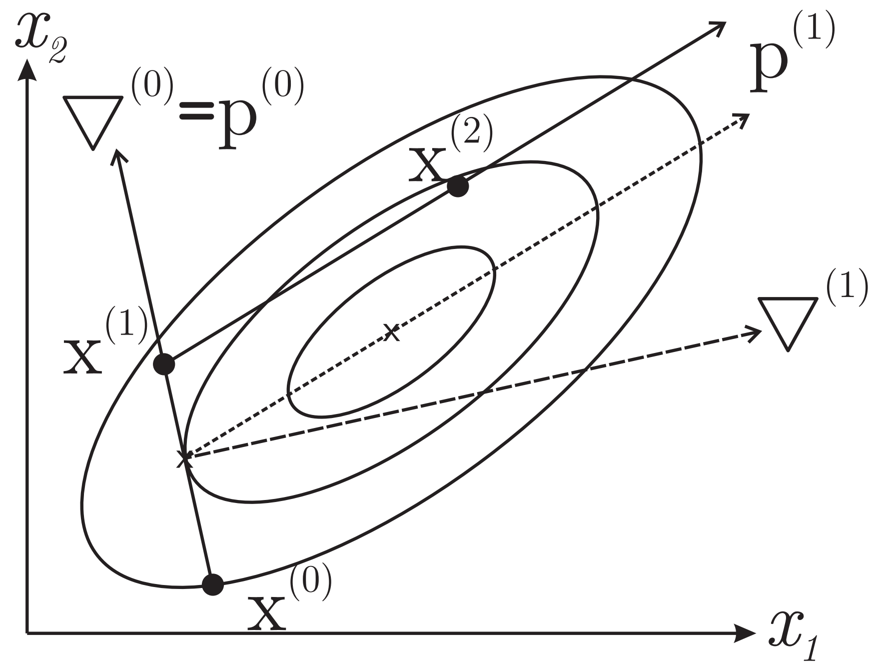
$$\mathbf{p}^k \leftarrow \mathbf{r}^k - \frac{\mathbf{r}^{k\mathsf{T}}\mathbf{q}^{k-1}}{d^{k-1}}\mathbf{p}^{k-1}$$

If the algorithm runs m iterations, i.e. $d^{k-1} \neq 0$ in each iteration, it returns $\mathbf{x} \sim \mathcal{N}\left(\mathbf{0}, A^{-1}\right)$ and $\mathbf{b} \sim \mathcal{N}\left(\mathbf{0}, A\right)$.

In pictures



Gibbs sampling & Gauss Seidel



CD sampling & CG optimization

Preconditioning A

Algorithm terminates ($\|r^{(k)}\| = 0$) after $k \leq$ number of distinct eigenvalues of A.

- If A is degenerate, use preconditioned precision matrix $U^\mathsf{T} A U$ to generate z and set $x = U^{-1} z$.
- Want $U^\mathsf{T} A U$ to have distinct (uniformly spaced) eigenvalues, and $x = U^{-1} z$ easy to solve.

$$U = \begin{pmatrix} 1 & U_1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & U_2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \ddots & \ddots & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & U_{m-2} & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & U_{m-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

where $U_i \overset{\text{i.i.d.}}{\sim} U(0,1)$.

An example $m = 10$

Exact $A^{-1} =$

$$\begin{pmatrix}
1.9786 & -1.0300 & 0.3454 & -0.2073 & 0.1523 & -0.0982 & 0.0531 & -0.0201 & 0.0006 & -0.0003 \\
-1.0300 & 1.0840 & -0.3635 & 0.2181 & -0.1603 & 0.1034 & -0.0560 & 0.0211 & -0.0007 & 0.0003 \\
0.3454 & -0.3635 & 1.5728 & -0.9439 & 0.6936 & -0.4474 & 0.2421 & -0.0915 & 0.0028 & -0.0014 \\
-0.2073 & 0.2181 & -0.9439 & 1.5555 & -1.1430 & 0.7372 & -0.3989 & 0.1508 & -0.0047 & 0.0023 \\
0.1523 & -0.1603 & 0.6936 & -1.1430 & 2.3520 & -1.5169 & 0.8209 & -0.3102 & 0.0096 & -0.0047 \\
-0.0982 & 0.1034 & -0.4474 & 0.7372 & -1.5169 & 1.7019 & -0.9210 & 0.3481 & -0.0108 & 0.0053 \\
0.0531 & -0.0560 & 0.2421 & -0.3989 & 0.8209 & -0.9210 & 1.2084 & -0.4567 & 0.0141 & -0.0069 \\
-0.0201 & 0.0211 & -0.0915 & 0.1508 & -0.3102 & 0.3481 & -0.4567 & 1.0006 & -0.0310 & 0.0152 \\
0.0006 & -0.0007 & 0.0028 & -0.0047 & 0.0096 & -0.0108 & 0.0141 & -0.0310 & 1.6747 & -0.8214 \\
-0.0003 & 0.0003 & -0.0014 & 0.0023 & -0.0047 & 0.0053 & -0.0069 & 0.0152 & -0.8214 & 1.0000
\end{pmatrix}$$

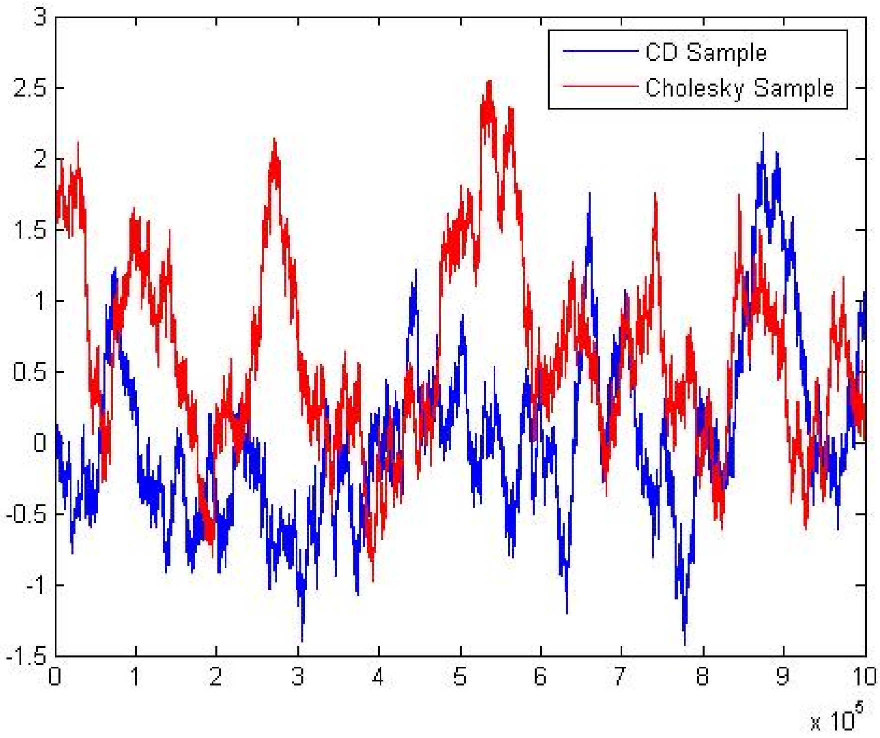
Estimated $(10^6$ samples)

$$\begin{pmatrix}
1.9767 & -1.0307 & 0.3472 & -0.2071 & 0.1495 & -0.0962 & 0.0527 & -0.0183 & 0.0026 & -0.0009 \\
-1.0300 & 1.0857 & -0.3640 & 0.2178 & -0.1593 & 0.1016 & -0.0544 & 0.0205 & -0.0008 & 0.0003 \\
0.3472 & -0.3640 & 1.5750 & -0.9455 & 0.6965 & -0.4484 & 0.2414 & -0.0900 & 0.0029 & -0.0010 \\
-0.2071 & 0.2178 & -0.9455 & 1.5550 & -1.1455 & 0.7386 & -0.3975 & 0.1488 & -0.0065 & 0.0038 \\
0.1495 & -0.1593 & 0.6965 & -1.1455 & 2.3564 & -1.5192 & 0.8203 & -0.3103 & 0.0112 & -0.0059 \\
-0.0962 & 0.1016 & -0.4484 & 0.7386 & -1.5192 & 1.7027 & -0.9221 & 0.3488 & -0.0117 & 0.0061 \\
0.0527 & -0.0544 & 0.2414 & -0.3975 & 0.8203 & -0.9221 & 1.2074 & -0.4557 & 0.0158 & -0.0087 \\
-0.0183 & 0.0205 & -0.0900 & 0.1488 & -0.3103 & 0.3488 & -0.4557 & 1.0005 & -0.0348 & 0.0167 \\
0.0026 & -0.0008 & 0.0029 & -0.0065 & 0.0112 & -0.0117 & 0.0158 & -0.0348 & 1.6744 & -0.8214 \\
-0.0009 & 0.0003 & -0.0010 & 0.0038 & -0.0059 & 0.0061 & -0.0087 & 0.0167 & -0.8214 & 1.0001
\end{pmatrix}$$

An example $m = 10^6$

GRF defined over 10^6 points on interval $[0, 1]$. Covariance function

$$R(x_1, x_2) = \exp\left\{-\frac{|x_1 - x_2|}{10}\right\}$$



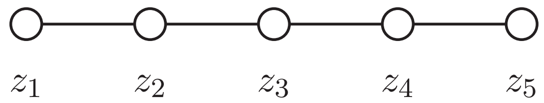
Covariance matrix is full, precision matrix is tridiagonal ...

Locally linear GMRF = discrete linear DE

Locally linear GMRF defined given neighbourhood system

$$x_i | x_{\partial_i} \sim \mathcal{N}\left(\overline{x}_{\partial_i}, 1/n_i\right)$$

$$W_{zij} = \begin{cases} n_i & \text{if } i = j \\ -1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$



$$W_z = \begin{pmatrix} 1 & -1 & \circ & \circ & \circ \\ -1 & 2 & -1 & \circ & \circ \\ \circ & -1 & 2 & -1 & \circ \\ \circ & \circ & -1 & 2 & -1 \\ \circ & \circ & \circ & -1 & 1 \end{pmatrix}$$

$$-u''(x) = f(x) \quad \text{for} \quad x \in (0,1),$$
$$u(0) = \varphi_0, \quad u(1) = \varphi_1$$

$$L_h = h^{-2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}$$

Dave Higdon *A Primer on Space-time Modeling from a Bayesian Perspective*

W. Hackbusch *Elliptic Differential Equations: Theory and Numerical Treatment*

BVP for $R = c\exp\{-|x|/r\}$

Consider covariance function

$$R(x_1, x_2) = c\exp\left\{-\frac{|x_1 - x_2|}{r}\right\}$$

where c is variance and r is length scale.

Stationary 'heat' $u(x)$ equation with loss over $\mathbb{R}$

$$\left[-\frac{d}{dx}\left(\frac{r}{2c}\frac{d}{dx}\right) + \frac{1}{2rc}\right] c\exp\left\{-\frac{|x-\xi|}{r}\right\} = \delta\left(x - \xi\right) \quad \forall x, \xi \in \mathbb{R}$$

- Differential operator and Green's operator are inverses (precision:covariance)
- c and r constant give convolution kernel over $\mathbb{R}$
- Differential operator is self-adjoint and positive $\forall c(x), r(x) > 0$

Finite domain $[a, b]$ with same covariance given by asserting exterior boundary integral equation (BIE)

$$u(a) = r(a)u'(a) \qquad -u(b) = r(b)u'(b)$$

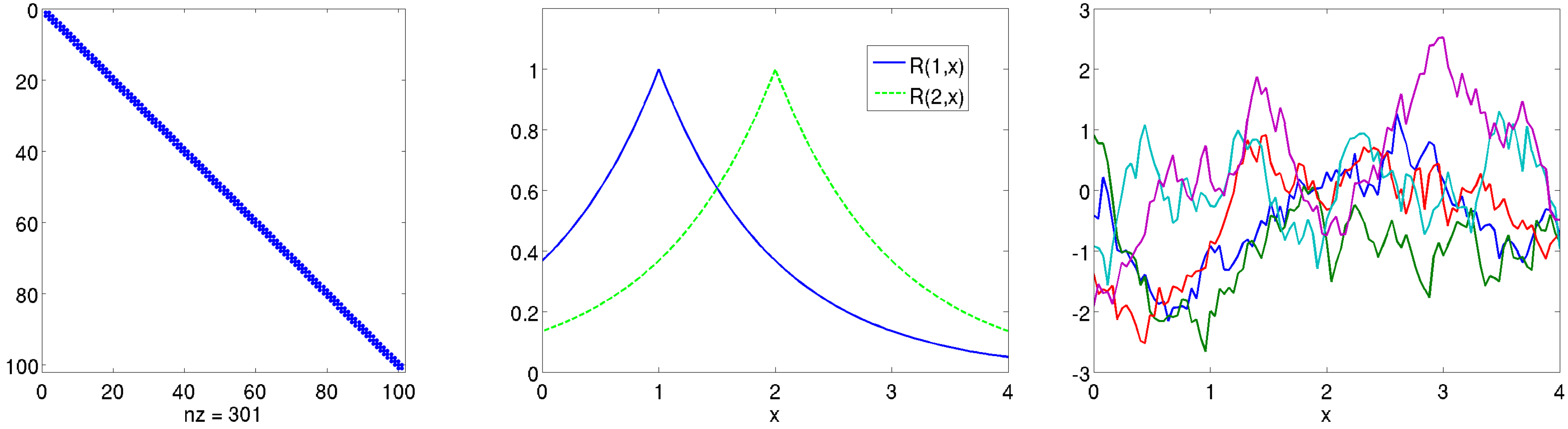
Variational form and FEM

The differential operator plus boundary conditions is equivalent to the variational form

$$\mathcal{I}(u) = \int_a^b \left(\frac{r}{4c} \left(\frac{du}{dx} \right)^2 + \frac{1}{4rc} u^2 \right) dx + \frac{u(a)^2}{4c} + \frac{u(b)^2}{4c}$$

- Discretize using piecewise linear basis functions (standard FEM)
- Gives tridiagonal stiffness plus damping matrix
- This is desired precision matrix A
- Sampling from $\mathcal{N}(0, A)$ is $O(m)$ using 'assembly by elements'
- Sampling from $\mathcal{N}(0, A^{-1})$ requires one linear solve
- Accuracy of precision:covariance relation better as $m \to \infty$ ... bigger IS better
- Irregular locations (unstructured mesh) simply accommodated

101 points in $[0, 4]$ with $c = 1$, $r = 1$



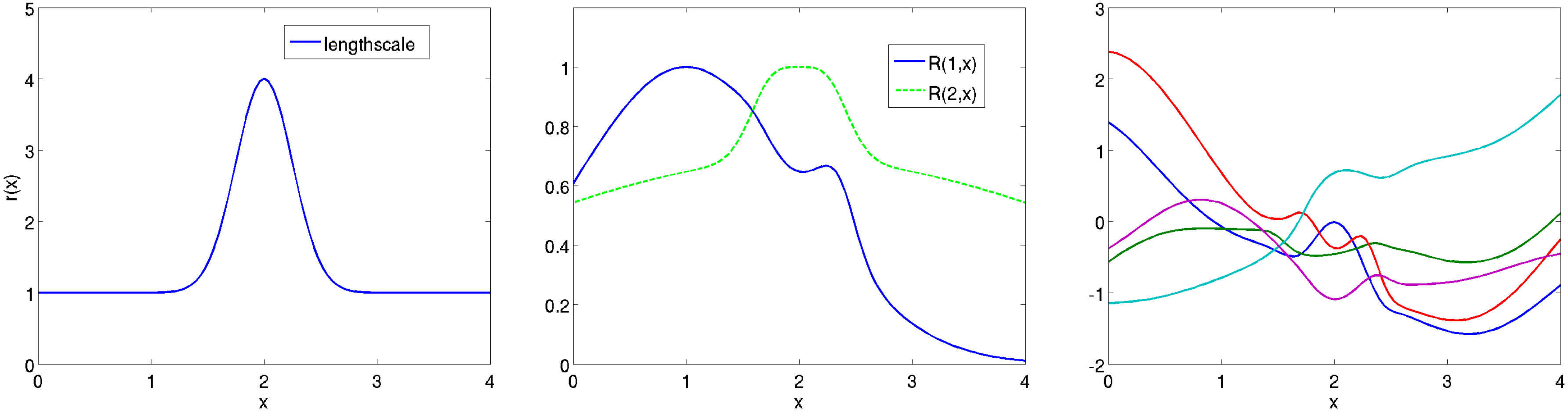
precision matrix | covariance functions | samples

Boundary terms give correct 'edge process'

Nonstationary construction of Higdon et al.

Mix (Gaussian) covariance functions according to

$$R(x_i, x_j) = 2^{n/2} \frac{|\Sigma_i|^{1/4}|\Sigma_j|^{1/4}}{|\Sigma_i + \Sigma_j|^{1/2}} \exp\left(-(x_i - x_j)^\mathsf{T} (\Sigma_i + \Sigma_j)^{-1} (x_i - x_j)\right)$$



length scale   covariance functions   samples

Mark Gibbs *Bayesian Gaussian Processes for Classification and Regression*

C. Paciorek *Nonstationary Gaussian Processes for Regression and Spatial Modelling*

Nonstationary covariance from transient heat equation

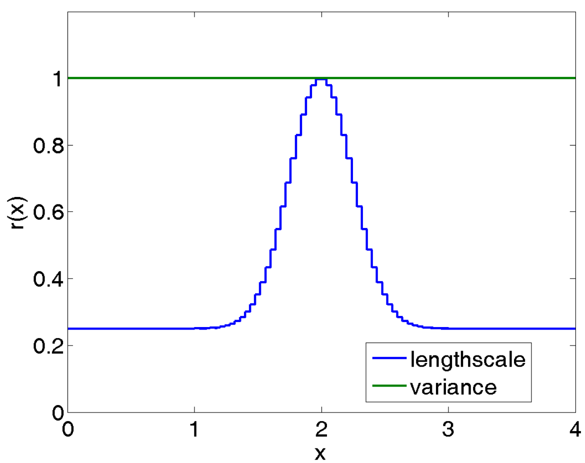
Differential operator

$$\frac{\partial}{\partial t} - \nabla \cdot (\alpha \nabla) + \beta$$

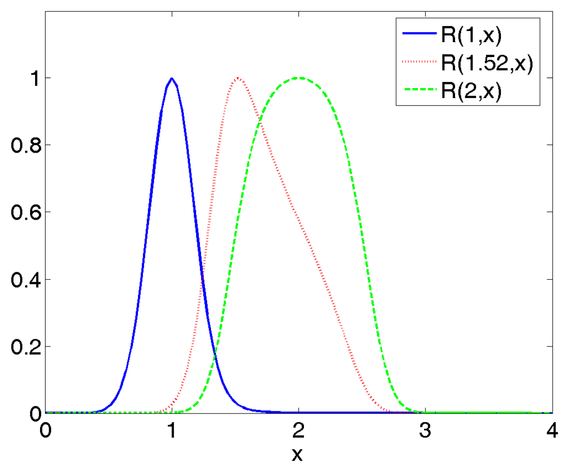
has Green's function

$$(4\pi\alpha t)^{-n/2} \exp\{-\beta t\} \exp\left\{-\frac{x^2}{4\alpha t}\right\} \quad t \geq 0$$

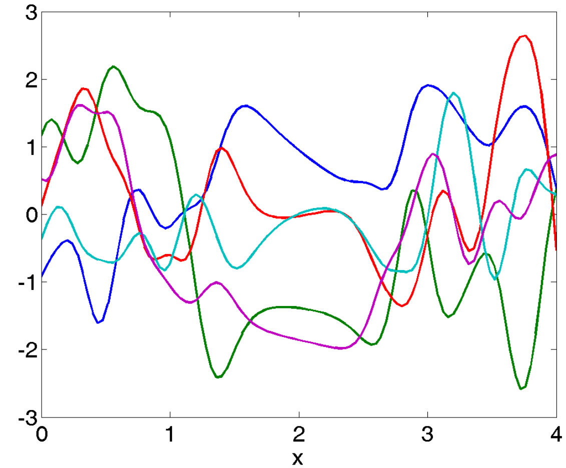
Simulate to $t = T$, with $\alpha(x), \beta(x)$ setting covariance and length scale



length scale



covariance functions



samples

Conclusions

- Conjugate direction algorithm samples high-dimensional ($m \approx 10^5$) Gaussian distributions
- FEM discretization of differential equations for the conservation, flow, loss of 'covarions' give nice non-stationary covariances and sparse precision matrices
- Hence provide

$$x(s)|c, r$$

  in hierachical models where variance c and length scale r set in a (hyper) spatial model
- Assembly of A by elements gives $O(m)$ sampling from $\mathcal{N}(0, A)$
- Sampling from $\mathcal{N}(0, A^{-1})$ has same cost as solving the PDE

- Fast methods from numerical linear algebra and FEM can be adapted to sampling and modelling LARGE Gaussian distributions