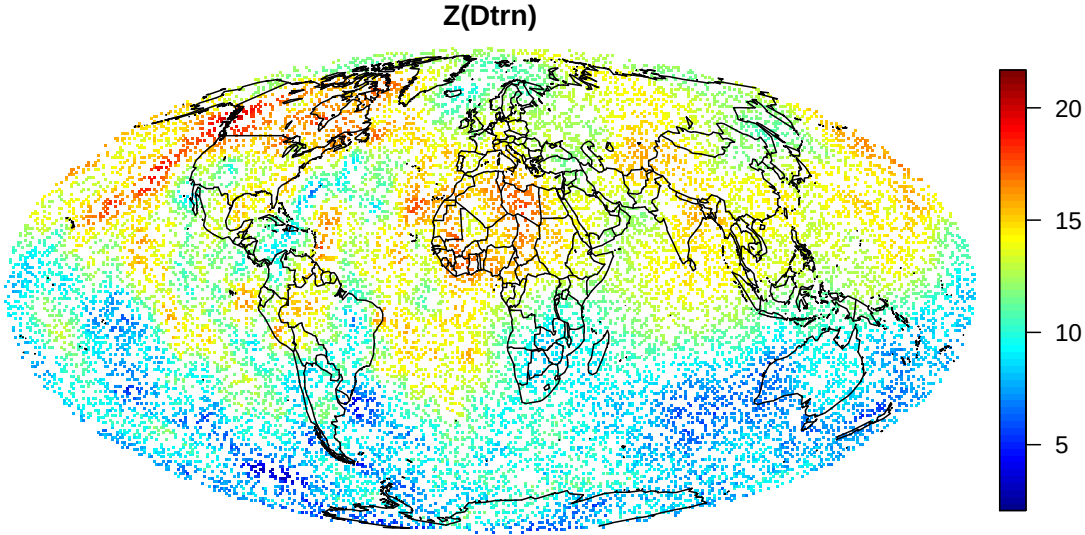
Numerical computation for statistics

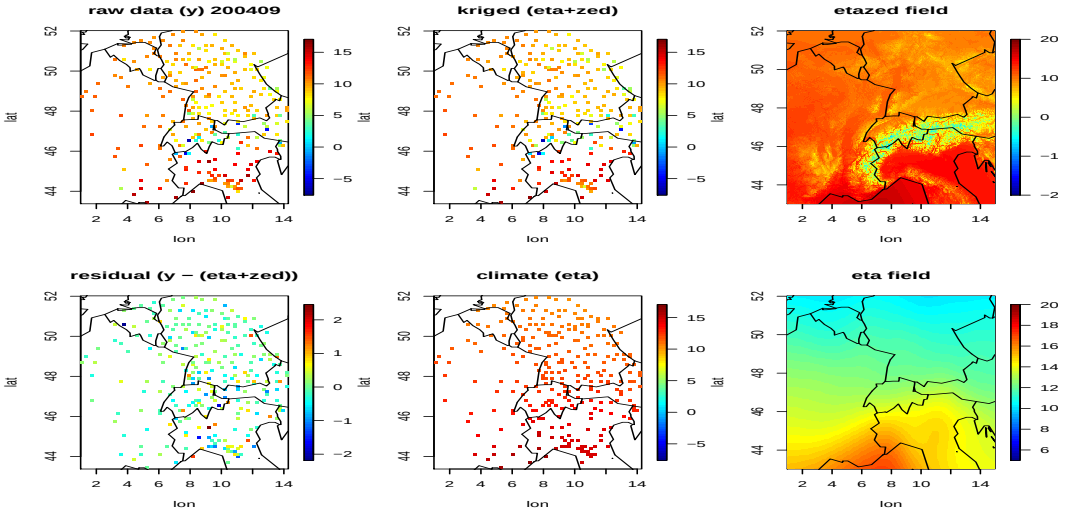
Finn Lindgren, University of Edinburgh, Scotland
http://www.maths.ed.ac.uk/~flindgre/cuso2019/
finn.lindgren@ed.ac.uk
(Parts 1 and 2)

CUSO, Villars-sur-Ollon, 1-4 September 2019

“Big” data



Sparse spatial coverage of temperature measurements



Regional observations: $\approx$ 20,000,000 from daily timeseries over 160 years

Overview (including parts not in this pdf)

- Spatial statistics (but perhaps not like you're used to if you've seen it before)
- From models to numerics with the help of Markov in space
- MCMC-free Bayesian inference with direct numerical approximations
- Assessing numerical and approximate methods; principled method assessment
- Scaling it up; Likelihood and covariance matrix for a 10^{11}-dimensional vector? No thank you!
- Some R demonstrations (`INLA`, `inlabru`, `excursions`)

Spatio-temporal modelling framework

Spatial statistics framework

- Spatial domain D, or space-time domain $D \times \mathbb{T}$, $\mathbb{T} \subset \mathbb{R}$.
- Random field $u(\boldsymbol{s})$, $\boldsymbol{s} \in D$, or $u(\boldsymbol{s}, t)$, $(\boldsymbol{s}, t) \in D \times \mathbb{T}$.
- Observations y_i. In the simplest setting, $y_i = u(\boldsymbol{s}_i) + \epsilon_i$, but more generally $y_i \sim \mathsf{GLMM}$, with $u(\cdot)$ as a structured random effect.
- Needed: models capturing stochastic dependence on multiple scales
- Partial solution: Basis function expansions, with large scale functions and covariates to capture static and slow structures, and small scale functions for more local variability

Two basic model and method components

- Stochastic models for $u(\cdot)$.
- Computationally efficient (i.e. avoid MCMC whenever possible) inference methods for the posterior distribution of $u(\cdot)$ given data $\boldsymbol{y}$.

Covariance functions and stochastic PDEs

The Matérn covariance family on $\mathbb{R}^d$

$$\mathrm{Cov}(u(\mathbf{0}), u(\boldsymbol{s})) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa \|\boldsymbol{s}\|)^\nu K_\nu(\kappa \|\boldsymbol{s}\|)$$

Scale $\kappa > 0$, smoothness $\nu > 0$, variance $\sigma^2 > 0$

Whittle (1954, 1963): Matérn as SPDE solution

Matérn fields are the stationary solutions to the SPDE

$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \alpha = \nu + d/2$$

$\mathcal{W}(\cdot)$ white noise, $\nabla \cdot \nabla = \sum_{i=1}^d \frac{\partial^2}{\partial s_i^2}$, $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}$

Gaussian random field (or Gaussian process)

A *Gaussian random field* $u : D \mapsto \mathbb{R}$ is defined via

$$\begin{aligned}
\mathsf{E}(u(\mathbf{s})) &= m(\mathbf{s}), \\
\mathsf{Cov}(u(\mathbf{s}), u(\mathbf{s}')) &= K(\mathbf{s}, \mathbf{s}'), \quad \text{(covariance kernel)} \\
\big[u(\mathbf{s}_i), i = 1, \ldots, n\big] &\sim \mathsf{N}(\boldsymbol{m} = \big[m(\mathbf{s}_i), i = 1, \ldots, n\big], \\
&\qquad \boldsymbol{\Sigma} = \big[K(\mathbf{s}_i, \mathbf{s}_j), i, j = 1, \ldots, n\big])
\end{aligned}$$

for all finite location sets $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, and $K(\cdot, \cdot)$ symmetric positive definite.

Generalised random field

A *generalised Gaussian random field* $u : D \mapsto \mathbb{R}$ is defined via a random measure, $\langle f, u \rangle_D = u^*(f) : H_{\mathcal{R}}(D) \mapsto \mathbb{R}$, $\mathcal{R}$ a covariance operator,

$$\mathsf{E}(\langle f, u \rangle_D) = \langle f, m \rangle_D = \int_D f(\mathbf{s}) m(\mathbf{s}) \, \mathrm{d}\mathbf{s},$$

$$\mathsf{Cov}(\langle f, u \rangle_D, \langle g, u \rangle_D) = \langle f, \mathcal{R} g \rangle_D \equiv \iint_{D \times D} f(\mathbf{s}) K(\mathbf{s}, \mathbf{s}') g(\mathbf{s}') \, \mathrm{d}\mathbf{s} \, \mathrm{d}\mathbf{s}',$$

$$\langle f, u \rangle_D \sim \mathsf{N}(\langle f, m \rangle_D, \langle f, \mathcal{R} f \rangle_D)$$

for all $f, g \in H_{\mathcal{R}}(D) \equiv \{f : D \mapsto \mathbb{R}; \langle f, \mathcal{R} f \rangle_D < \infty\}$.
This allows for singular covariance kernels $K(\cdot, \cdot)$.

White noise vs independent noise

Gaussian white noise on continuous domains

Standard Gaussian white noise $\mathcal{W}(\cdot)$ is a generalised random field, with

$$m(\mathbf{s}) = 0, \qquad K(\mathbf{s}, \mathbf{s}') = \delta_{\mathbf{s}}(\mathbf{s}'), \qquad \langle f, \mathcal{W} \rangle_D \sim \mathsf{N}(0, \langle f, f \rangle_D),$$

for all $f \in L_2(D)$. Since $\langle \delta_{\mathbf{s}}, \delta_{\mathbf{s}} \rangle_D = \infty$ for all $\mathbf{s} \in D$, $\mathcal{W}(\cdot)$ does not have pointwise meaning. We can only do calculus!

Independent Gaussian noise on continuous domains

Spatially independent Gaussian noise $w(\cdot)$ is a random field, with

$$m(\mathbf{s}) = 0, \qquad K(\mathbf{s}, \mathbf{s}') = \mathbf{1}_{\{\mathbf{s}=\mathbf{s}'\}}, \qquad w(\mathbf{s}) \sim \mathsf{N}(0, 1),$$

for all $\mathbf{s}, \mathbf{s}' \in D$. However, for every set $A \subset D$ with $|A|_{\mathsf{Leb(D)}} > 0$,

$$\mathsf{P}(\sup_{\mathbf{s} \in A} w(\mathbf{s}) = \infty) = \mathsf{P}(\inf_{\mathbf{s} \in A} w(\mathbf{s}) = -\infty) = 1,$$

and the generalised calculus is not applicable.

Spectral properties

Bochner's theorem on $\mathbb{R}^d$

A symmetric kernel $K(\mathbf{s}, \mathbf{s}')$, $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^d$, is a positive (semi-)definite stationary covariance kernel if and only if there exists a non-negative spectral measure $S^*(\boldsymbol{\omega})$ such that

$$K(\mathbf{s}, \mathbf{s}') = \int_{\mathbb{R}^d} \exp(i(\mathbf{s}' - \mathbf{s}) \cdot \boldsymbol{\omega}) \, \mathrm{d}S^*(\boldsymbol{\omega})$$

If the measure has a density $S(\boldsymbol{\omega})$,

$$K(\mathbf{s}, \mathbf{s}') = \int_{\mathbb{R}^d} \exp(i(\mathbf{s}' - \mathbf{s}) \cdot \boldsymbol{\omega}) S(\boldsymbol{\omega}) \, \mathrm{d}\boldsymbol{\omega}$$

$$S(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp(-i\mathbf{s} \cdot \boldsymbol{\omega}) K(0, \mathbf{s}) \, \mathrm{d}\mathbf{s}$$

White noise on $\mathbb{R}^d$ has spectral density $S_{\mathcal{W}}(\boldsymbol{\omega}) = 1/(2\pi)^d$.

Spectral properties

Spectral representation

Let $Z^*(\boldsymbol{\omega})$ be a complex Gaussian random measure on $D = \mathbb{R}^d$ with independent increments and

$$\overline{\mathrm{d}Z^*(\boldsymbol{\omega})} = \mathrm{d}Z^*(-\boldsymbol{\omega}), \qquad \mathsf{E}\left[\mathrm{d}Z^*(\boldsymbol{\omega})\right] = 0, \qquad \mathsf{E}\left[\mathrm{d}Z^*(\boldsymbol{\omega})\overline{\mathrm{d}Z^*(\boldsymbol{\omega})}\right] = \mathrm{d}S^*(\boldsymbol{\omega}).$$

Then

$$u(\mathbf{s}) = \int_{\mathbb{R}^d} \exp(i\mathbf{s}\cdot\boldsymbol{\omega})\,\mathrm{d}Z^*(\boldsymbol{\omega})$$

is a stationary Gaussian random field with spectral measure $S^*(\boldsymbol{\omega})$.

Let $\widehat{f}(\boldsymbol{\omega}) = (\mathcal{F}f)(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\exp(-i\mathbf{s}\cdot\boldsymbol{\omega})f(\mathbf{s})\,\mathrm{d}\mathbf{s}$.

Informally, $\widehat{u}(\boldsymbol{\omega})\,\mathrm{d}\boldsymbol{\omega} = \mathrm{d}Z^*(\boldsymbol{\omega})$, and the spectral density is $S_u(\boldsymbol{\omega}) = \mathsf{E}(|\widehat{u}(\boldsymbol{\omega})|^2)$.

Spectral properties

Spectra and linear differential operators

Differential operators can also be interpreted spectrally:

$\mathcal{L}f$	f	∇f	$-\nabla\cdot\nabla f$	$\mathcal{L}^{\alpha/2}f$
$\widehat{\mathcal{L}f} \equiv \mathcal{F}(\mathcal{L}f)$	$\widehat{f}$	$i\boldsymbol{\omega}\widehat{f}$	$\|\boldsymbol{\omega}\|^2\widehat{f}$	$\lvert\widehat{\mathcal{L}}\rvert^{\alpha/2}\widehat{f}$

The rightmost column is a *definition* of a fractional operator!

Exercise: Use the spectral field representation to derive the middle two results above.

Exercise: What would happen on a different manifold, such as the sphere? Hint: the harmonic functions in the Fourier transform are eigenfunctions of the Laplacian.

Spectral properties

For the Whittle-Matérn SPDE, informally,

$$
\begin{aligned}
(\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} u(\mathbf{s}) &= \mathcal{W}(\mathbf{s}) \\
(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha/2} \widehat{u}(\boldsymbol{\omega}) &= \widehat{\mathcal{W}}(\boldsymbol{\omega}) \\
\mathsf{E}(|(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha/2} \widehat{u}(\boldsymbol{\omega})|^2) &= \mathsf{E}(|\widehat{\mathcal{W}}(\boldsymbol{\omega})|^2) \\
(\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha} S_u(\boldsymbol{\omega}) &= S_{\mathcal{W}}(\boldsymbol{\omega}) \\
S_u(\boldsymbol{\omega}) &= \frac{1}{(2\pi)^d (\kappa^2 + \|\boldsymbol{\omega}\|^2)^{\alpha}}
\end{aligned}
$$

Whittle (1954, 1963) showed that $K(\mathbf{s}, \mathbf{s}') = (\mathcal{F}^{-1} S_u(\cdot))(\mathbf{s}' - \mathbf{s})$ is equal to the Matérn covariance (up to a known scaling constant), with smoothness $\nu = \alpha - d/2$.

Simple heat equation

For space-time fields, we write $u(\mathbf{s}, t)$, $(\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}$, and $S_u(\boldsymbol{k}, \omega)$, $(\boldsymbol{k}, \omega) \in \mathbb{R}^d \times \mathbb{R}$.
We drive a heat equation with a noise process $\mathcal{E}$ that is white noise in time and Matérn noise in space, with parameters matching the heat operator:

$$\left\{\gamma \frac{\partial}{\partial t} + \kappa^2 - \nabla_s \cdot \nabla_s\right\} u(\mathbf{s}) = \mathcal{E}(\mathbf{s}, t),$$
$$(\kappa^2 - \nabla_s \cdot \nabla_s)^{\alpha/2} \mathcal{E}(\mathbf{s}, t) = \mathcal{W}(\mathbf{s}, t).$$

The Fourier domain version is

$$\left\{i\gamma\omega + \kappa^2 + \|\boldsymbol{k}\|^2\right\} \widehat{u}(\boldsymbol{k}, \omega) = \widehat{\mathcal{E}}(\boldsymbol{k}, \omega),$$
$$(\kappa^2 + \|\boldsymbol{k}\|^2)^{\alpha/2} \widehat{\mathcal{E}}(\boldsymbol{k}, \omega) = \widehat{\mathcal{W}}(\boldsymbol{k}, \omega),$$

and

$$S_u(\boldsymbol{k}, \omega) = \frac{1}{(2\pi)^{d+1}(\gamma^2\omega^2 + (\kappa^2 + \|\boldsymbol{k}\|^2)^2)(\kappa^2 + \|\boldsymbol{k}\|^2)^\alpha}$$

How differentiable are the realisations?

Simple heat equation (cont)

Using that, in the standardised Whittle-Matérn SPDE, the variance is

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)\kappa^{2\nu}(4\pi)^{d/2}}, \quad \nu = \alpha - d/2,$$

the marginal spatial spectrum for the heat model is

$$S_u(\boldsymbol{k}) = \int_{\mathbb{R}} S_u(\boldsymbol{k}, \omega)\, \mathrm{d}\omega = \frac{1}{4\pi\gamma} \frac{1}{(2\pi)^d(\kappa^2 + \|\boldsymbol{k}\|^2)^{\alpha+1}},$$

which is a scaled Whittle spectrum for a Matérn covariance with smoothness $\nu = \alpha + 1 - d/2$.

A generalised generalised case

If $\alpha = 0$, $d = 2$, then $\nu = 0$, which is just outside of the allowed range of the Matérn family. However, for every t, $u(\cdot, t)$ is a generalised random field with singular kernel $K(\mathbf{s}, \mathbf{s}') = \frac{1}{4\pi\gamma}\frac{1}{2\pi}K_0(\kappa\|\mathbf{s}' - \mathbf{s}\|)$.

Simple heat equation (cont)

To help understand the temporal properties, take the Fourier transform in only the spatial directions:

$$\left\{\gamma\frac{\partial}{\partial t} + \kappa^2 + \|\boldsymbol{k}\|^2\right\}\widetilde{u}(\boldsymbol{k},t) = \frac{\widetilde{\mathcal{W}}(\boldsymbol{k},t)}{(\kappa^2+\|\boldsymbol{k}\|^2)^{\alpha/2}},$$

so for each spatial frequency $\boldsymbol{k}$, the temporal evolution of $\widetilde{u}(\boldsymbol{k},t)$ is an Ornstein-Uhlenbeck process with covariance

$$\frac{1}{4\pi\gamma(\kappa^2+\|\boldsymbol{k}\|^2)^{\alpha+1}}\exp\left(-|t|\frac{\kappa^2+\|\boldsymbol{k}\|^2}{\gamma}\right).$$

There is one more property we need to understand: Markov in space

First order Markov in time

Filtration σ-algebras:

$$a \in \mathcal{F}^{\sigma}_{(-\infty,t]} \equiv \sigma(u(s), s \leq t), \quad b \in \mathcal{F}^{\sigma}_{[t,\infty)} \equiv \sigma(u(s), s \geq t)$$
$$\mathsf{P}(a \cap b \mid u(t)) = \mathsf{P}(a \mid u(t))\mathsf{P}(b \mid u(t))$$

Higher order Markov on spatial and spatio-temporal domains

Let $A, B, S \subset D$, such that S separates A and B.

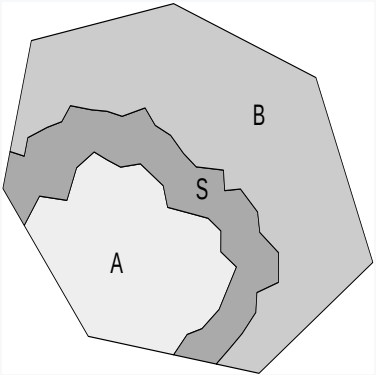
$$\mathcal{F}^{\sigma}_{S} \equiv \sigma(u(\mathbf{s}), \mathbf{s} \in S), \quad a \in \mathcal{F}^{\sigma}_{A}, \quad b \in \mathcal{F}^{\sigma}_{B},$$
$$\mathsf{P}(a \cap b \mid \mathcal{F}^{\sigma}_{S}) = \mathsf{P}(a \mid \mathcal{F}^{\sigma}_{S})\mathsf{P}(b \mid \mathcal{F}^{\sigma}_{S})$$

Markov for generalised random fields

$$\mathcal{F}^{\sigma}_{S} \equiv \sigma(\langle f, u \rangle_{S}, f \in H_{\mathcal{R}}(S)), \quad a \in \mathcal{F}^{\sigma}_{A}, \quad b \in \mathcal{F}^{\sigma}_{B},$$
$$\mathsf{P}(a \cap b \mid \mathcal{F}^{\sigma}_{S}) = \mathsf{P}(a \mid \mathcal{F}^{\sigma}_{S})\mathsf{P}(b \mid \mathcal{F}^{\sigma}_{S})$$

Markov in space

Markov properties



S is a separating set for A and B: $u(A) \perp u(B) \mid u(S)$

Solutions to

$$\left(\kappa^2 - \nabla \cdot \nabla\right)^{\alpha/2} u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s})$$

are Markov when α is an integer.

More generally, when the reciprocal of the spectral density is a polynomial, Rozanov, 1977

In graphs with no edges between A and B ($\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$):

$$\boldsymbol{Q}_{AB} = \boldsymbol{0}$$
$$\boldsymbol{Q}_{A|S,B} = \boldsymbol{Q}_{AA}$$
$$\boldsymbol{\mu}_{A|S,B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AS}(\boldsymbol{u}_S - \boldsymbol{\mu}_S)$$

Generally: Markov iff the precision *operator* $\mathcal{Q} = \mathcal{R}^{-1}$ is local.

Markov in space

Precision matrix block structure: $\begin{bmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AS} & \mathbf{0} \\ \boldsymbol{Q}_{SA} & \boldsymbol{Q}_{SS} & \boldsymbol{Q}_{SB} \\ \mathbf{0} & \boldsymbol{Q}_{BS} & \boldsymbol{Q}_{BB} \end{bmatrix}$

A partial history of Markov random fields

Rozanov (1977)

Generally: Markov iff the precision *operator* $\mathcal{Q} = \mathcal{R}^{-1}$ is local.

Stationary case:

$$u(\mathbf{s}) \text{ is stationary Markov} \iff S_u(\boldsymbol{k}) \propto P(\boldsymbol{k})^{-1}$$

where $P(\boldsymbol{k}) \geq 0$ is a symmetric polynomial

Matérn/Whittle is Markov for $\alpha = 1, 2, 3 \ldots$: $S_u(\boldsymbol{k}) \propto (\kappa^2 + \|\boldsymbol{k}\|^2)^{-\alpha}$

GMRF	Covariance on $\mathbb{R}^2$	
SAR(1) CAR(2)	$\propto \kappa\|\boldsymbol{u}\| K_1(\kappa\|\boldsymbol{u}\|)$	Whittle (1954)
CAR(1)	$\frac{1}{2\pi} K_0(\kappa\|\boldsymbol{u}\|)$	Besag (1981)
ICAR(1)	$-\frac{1}{2\pi}\log(\|\boldsymbol{u}\|)$	Besag & Mondal (2005)

On lattices, classical CAR $\rightarrow$ Matérn models (limits of).

Hilbert space approximation ("The SPDE approach" from Lindgren et al, 2011)

Can extend to (non-)stationary SPDE models on irregular triangulations.

From continuous to discrete

We want to construct finite dimensional approximations to the distribution of $u(\cdot)$, where

$$\left[\left\langle f_i, (\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} u(\cdot)\right\rangle_D, i = 1, \ldots, m\right] \stackrel{d}{=} \left[\left\langle f_i, \mathcal{W}(\cdot)\right\rangle_D, i = 1, \ldots, m\right]$$

for all finite collections of test functions $f_i \in H_{\mathcal{R}}(D)$.

A finite basis expansion

$$u(\mathbf{s}) = \sum_{j=1}^{n} \psi_j(\mathbf{s}) u_j$$

can only hope to achieve this for a subspace of size n.

Two main approaches:

- Galerkin: $\{f_i = \psi_i,\ i = 1, \ldots, n\}$
- Least squares: $\{f_i = (\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} \psi_i,\ i = 1, \ldots, n\}$

We use least squares for $\alpha = 1$, Galerkin for $\alpha = 2$, and a recursion for $\alpha \geq 3$.

Stochastic Green's first identity

On any sufficiently smooth manifold domain D,

$$\langle f, -\nabla \cdot \nabla g \rangle_D = \langle \nabla f, \nabla g \rangle_D - \langle f, \partial_n g \rangle_{\partial D}$$

holds, even if either ∇f or $-\nabla \cdot \nabla g$ are as generalised as white noise.

For now, we'll impose deterministic Neumann boundary conditions, informally $\partial_n u(\mathbf{s}) = 0$ for all $\mathbf{s} \in \partial D$. For $\alpha = 2$ and Galerkin,

$$\begin{aligned}\left\langle \psi_i, (\kappa^2 - \nabla \cdot \nabla) \sum_j \psi_j u_j \right\rangle_D &= \sum_j \left\{ \kappa^2 \langle \psi_i, \psi_j \rangle_D + \langle \nabla \psi_i, \nabla \psi_j \rangle_D \right\} u_j \\ &= (\kappa^2 \boldsymbol{C} + \boldsymbol{G}) \boldsymbol{u}\end{aligned}$$

The covariance for the RHS of the SPDE is

$$\left[\mathsf{Cov}(\langle \psi_i, \mathcal{W} \rangle_D, \langle \psi_j, \mathcal{W} \rangle_D \right] = \left[\langle \psi_i, \psi_j \rangle_D \right] = \boldsymbol{C}$$

by the definition of $\mathcal{W}$.

We seek $\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ such that $\mathsf{Var}\{(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{u}\} = \boldsymbol{C}$:

$$(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{\Sigma}(\kappa^2\boldsymbol{C} + \boldsymbol{G}) = \boldsymbol{C}$$
$$\boldsymbol{\Sigma} = (\kappa^2\boldsymbol{C} + \boldsymbol{G})^{-1}\boldsymbol{C}(\kappa^2\boldsymbol{C} + \boldsymbol{G})^{-1}$$

If ψ_i are piecewise linear on a triangulation of D, then $\boldsymbol{C}$ and $\boldsymbol{G}$ are both very sparse, and in addition, $\boldsymbol{C} = \mathrm{diag}(\langle \psi_i, 1 \rangle_D)$ is a valid approximation. Then, the *precision* matrix is also sparse,

$$\boldsymbol{Q} = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G})$$

and $\boldsymbol{u}$ is Markov on the adjacency graph given by the non-zero structure of $\boldsymbol{Q}$.

Least squares and Galerkin recursion gives precisions for all $\alpha = 1, 2, \ldots$:

- $\boldsymbol{Q}_1 = (\kappa^2\boldsymbol{C} + \boldsymbol{G})$
- $\boldsymbol{Q}_2 = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G}) = \kappa^4\boldsymbol{C} + 2\kappa^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}$
- $\boldsymbol{Q}_\alpha = (\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1}\boldsymbol{Q}_{\alpha-2}\boldsymbol{C}^{-1}(\kappa^2\boldsymbol{C} + \boldsymbol{G})$
- Any $\alpha \geq 0$: $\boldsymbol{Q}_\alpha = \boldsymbol{C}^{1/2}\left\{\boldsymbol{C}^{-1/2}(\kappa^2\boldsymbol{C} + \boldsymbol{G})\boldsymbol{C}^{-1/2}\right\}^\alpha \boldsymbol{C}^{1/2}$
  (non-sparse for non-integer α)

Basis function representations for Gaussian Matérn fields

Basis definitions

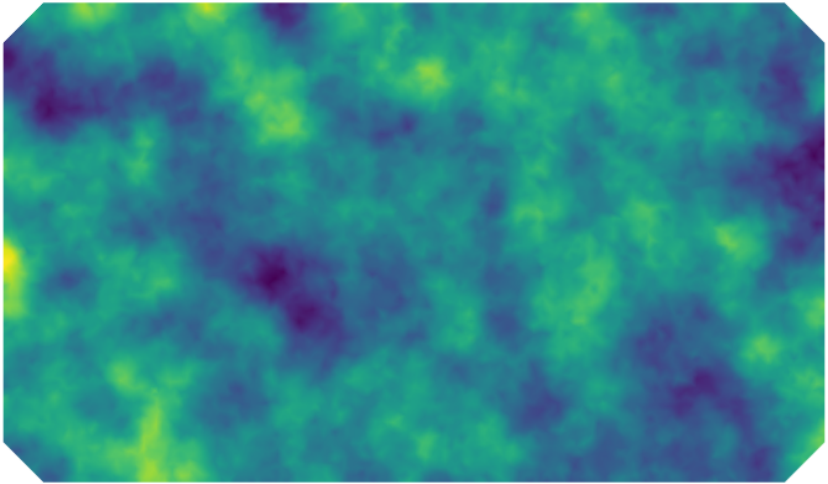
	Finite basis set ($k = 1, \ldots, n$)
Karhunen-Loève	$(\kappa^2 - \nabla \cdot \nabla)^{-\alpha} e_{\kappa,k}(\boldsymbol{s}) = \lambda_{\kappa,k} e_{\kappa,k}(\boldsymbol{s})$
Fourier	$-\nabla \cdot \nabla e_k(\boldsymbol{s}) = \lambda_k e_k(\boldsymbol{s})$
Convolution	$(\kappa^2 - \nabla \cdot \nabla)^{\alpha/2} g_\kappa(\boldsymbol{s}) = \delta(\boldsymbol{s})$
General	$\psi_k(\boldsymbol{s})$

Field representations

	Field $u(\boldsymbol{s})$	Weights
Karhunen-Loève	$\propto \sum_k e_{\kappa,k}(\boldsymbol{s}) z_k$	$z_k \sim \mathsf{N}(0, \lambda_{\kappa,k})$
Fourier	$\propto \sum_k e_k(\boldsymbol{s}) z_k$	$z_k \sim \mathsf{N}(0, (\kappa^2 + \lambda_k)^{-\alpha})$
Convolution	$\propto \sum_k g_\kappa(\boldsymbol{s} - \boldsymbol{s}_k) z_k$	$z_k \sim \mathsf{N}(0, \lvert\mathsf{cell}_k\rvert)$
General	$\propto \sum_k \psi_k(\boldsymbol{s}) u_k$	$\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}_\kappa^{-1})$

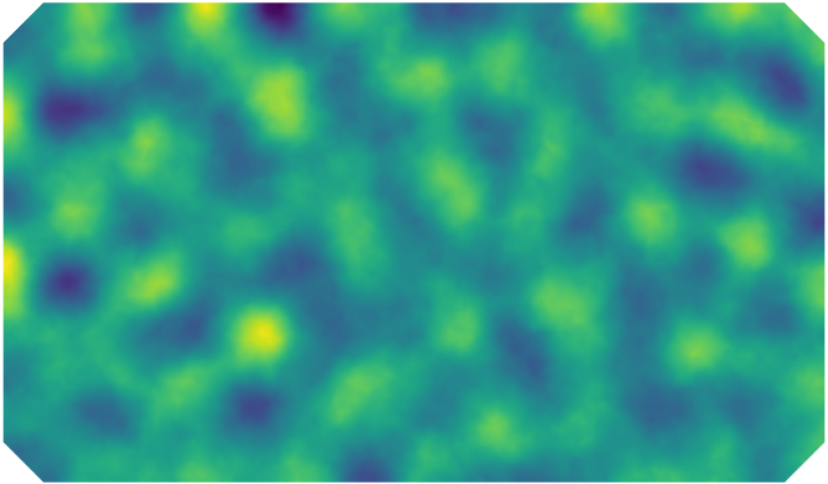
Note: Harmonic basis functions (as in the Fourier approach) give a diagonal $\boldsymbol{Q}_\kappa$, but lead to dense posterior precision matrices.

SPDE/GMRF realisations and non-stationary models



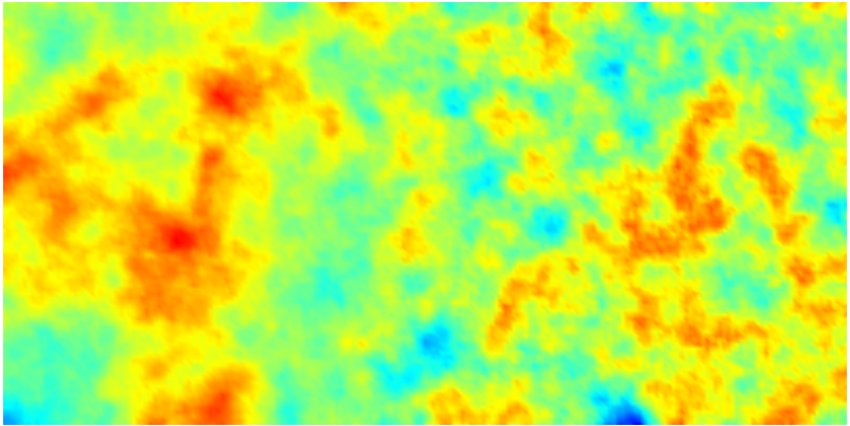
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



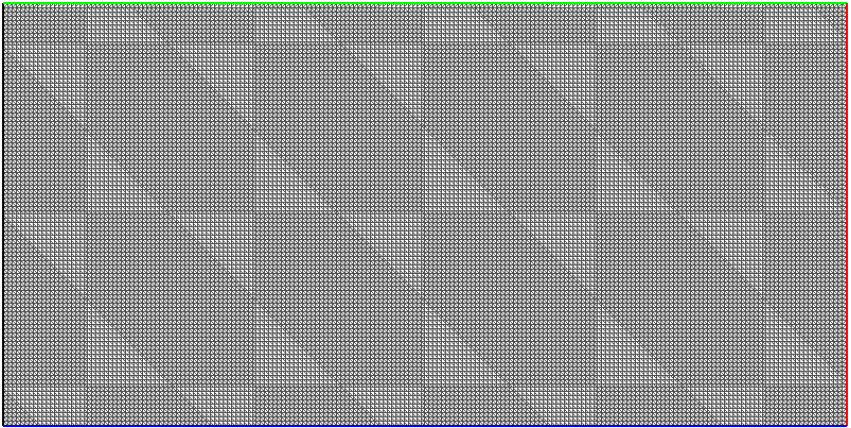
$(\kappa^2 \exp(i\theta) - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \boldsymbol{s} \in D$, $\mathsf{Re}(u)$ independent of $\mathsf{Im}(u)$

Link to Sampson&Guttorp (1992) deformation non-stationarity



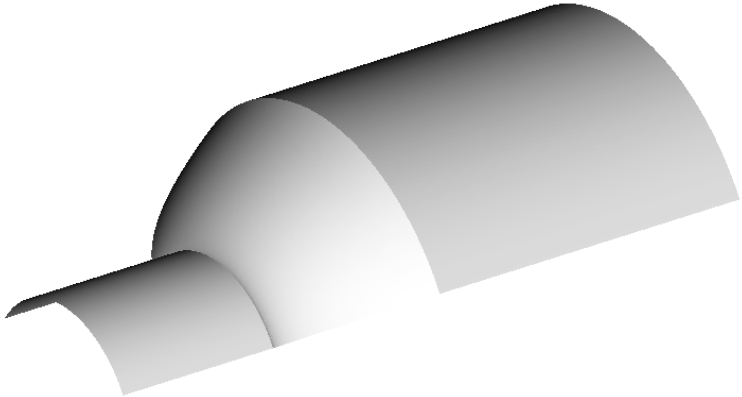
$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



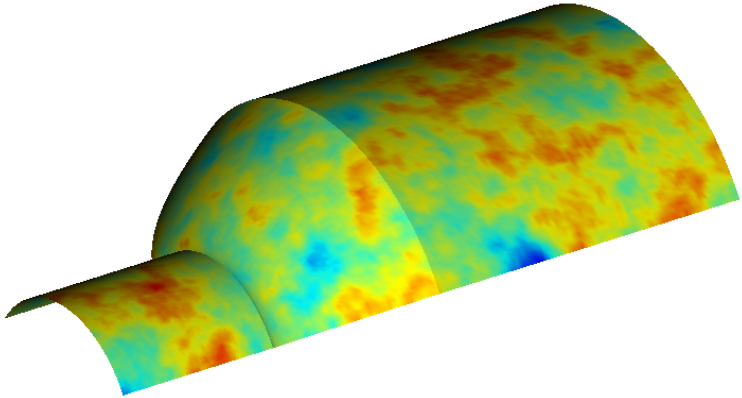
$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



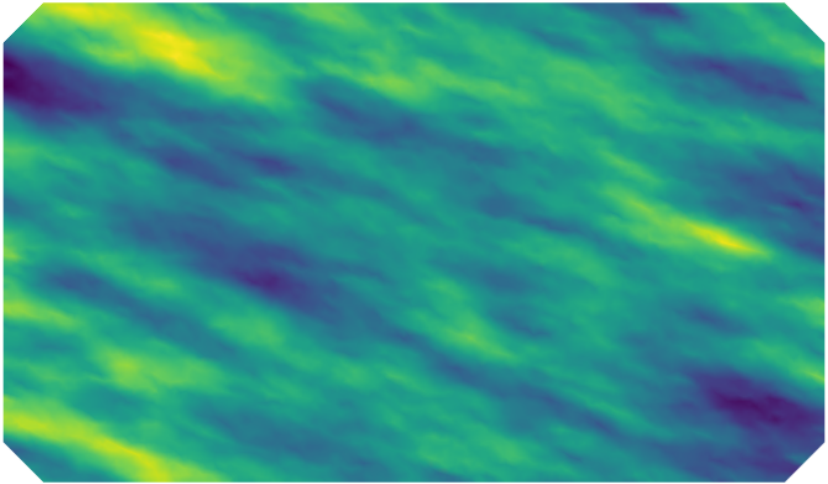
$$(\widetilde{\kappa}^2 - \nabla \cdot \nabla) u(\widetilde{\boldsymbol{s}}) = \widetilde{\kappa}\widetilde{\mathcal{W}}(\widetilde{\boldsymbol{s}}), \quad \widetilde{\boldsymbol{s}} \in \widetilde{\Omega}$$

Link to Sampson&Guttorp (1992) deformation non-stationarity



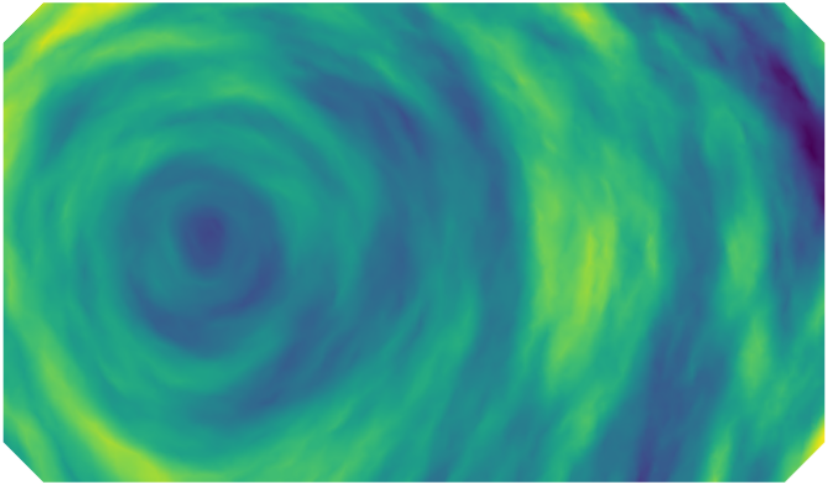
$$(\widetilde{\kappa}^2 - \nabla \cdot \nabla) u(\widetilde{\boldsymbol{s}}) = \widetilde{\kappa}\widetilde{\mathcal{W}}(\widetilde{\boldsymbol{s}}), \quad \widetilde{\boldsymbol{s}} \in \widetilde{\Omega}$$

SPDE/GMRF realisations and non-stationary models



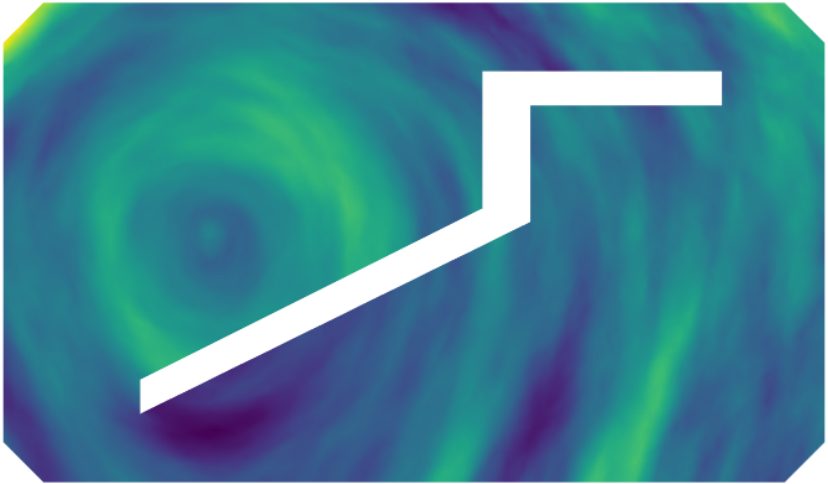
$$(\kappa^2 - \nabla \cdot \boldsymbol{H} \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



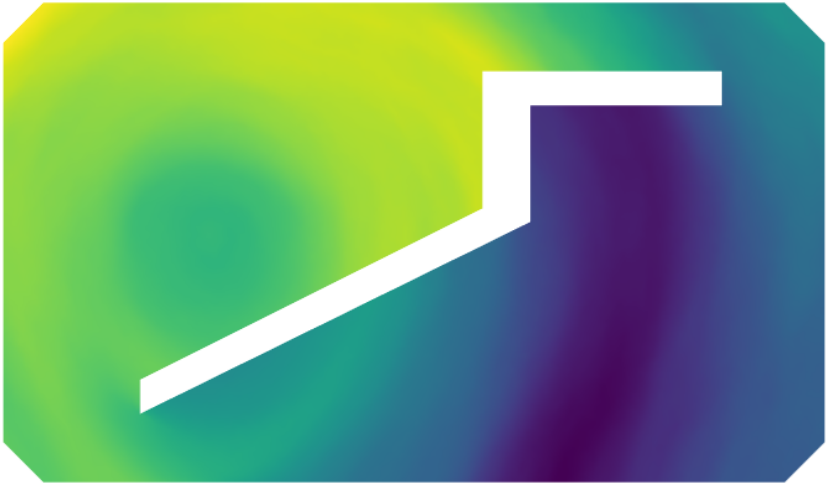
$$(\kappa^2 - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



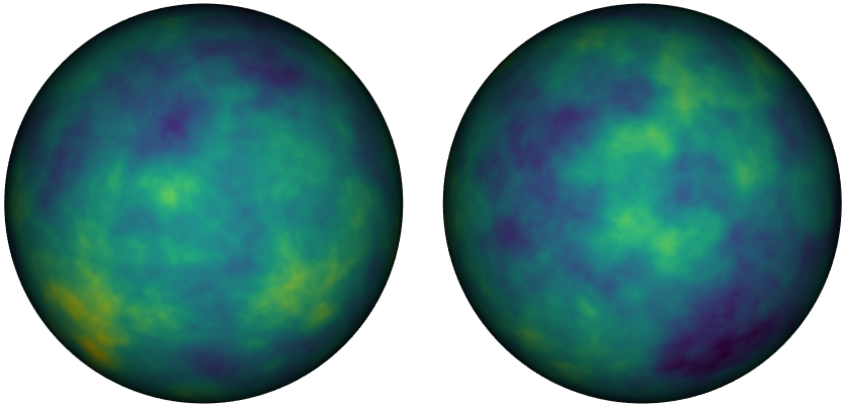
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



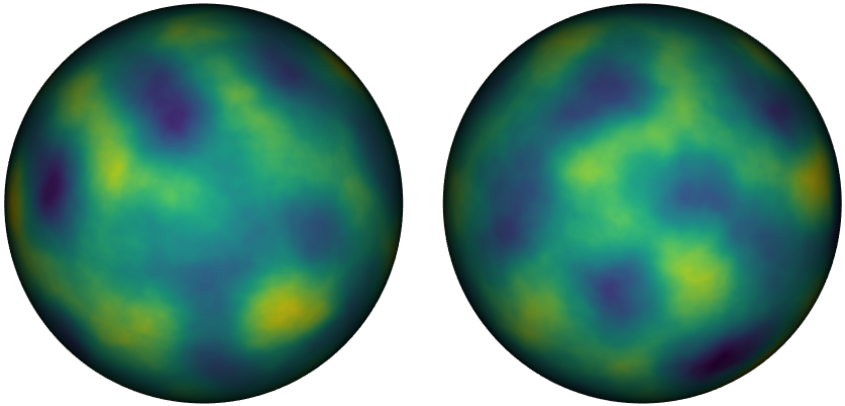
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

SPDE/GMRF realisations and non-stationary models



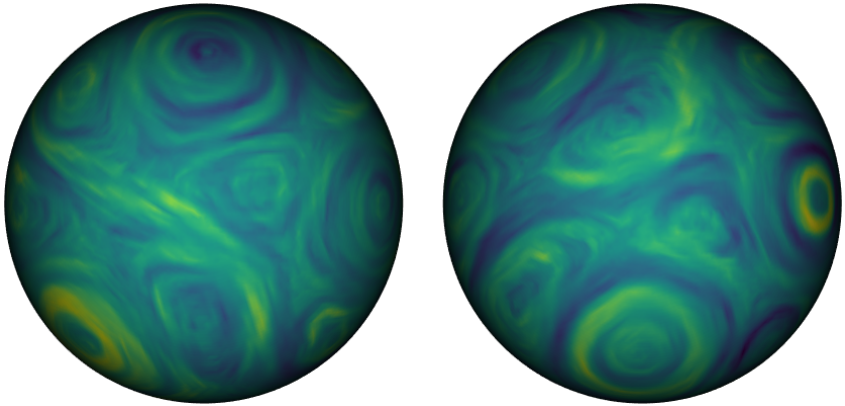
$$(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D = \mathbb{S}^2$$

SPDE/GMRF realisations and non-stationary models



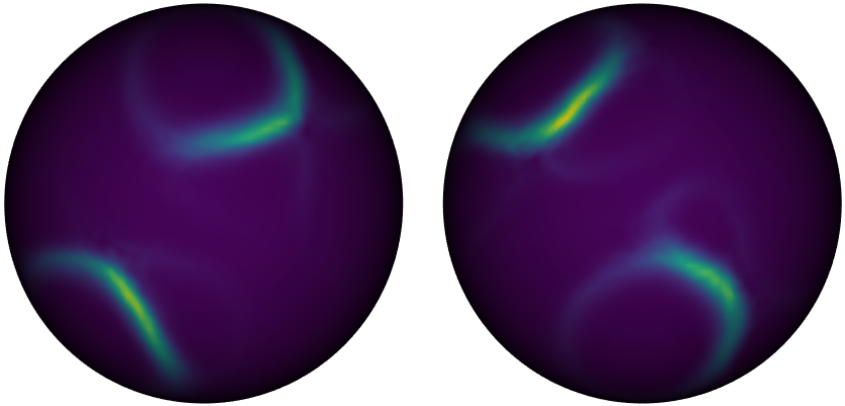
$$(\kappa^2 \exp(i\theta) - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in D = \mathbb{S}^2$$

Markov does *not* mean that dependence is only local



$$(\kappa(\boldsymbol{s})^2 - \nabla \cdot \boldsymbol{H}(\boldsymbol{s})\nabla)u(\boldsymbol{s}) = \kappa(\boldsymbol{s})\mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$$

Covariances for four reference points



Hierarchical models

Continuous Markovian spatial models (Lindgren et al, 2011)

Local basis: $u(\boldsymbol{s}) = \sum_k \psi_k(\boldsymbol{s}) u_k$, (compact, piecewise linear)

Basis weights: $\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, sparse $\boldsymbol{Q}$ based on an SPDE

Special case: $(\kappa^2 - \nabla \cdot \nabla) u(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega$

Precision: $\boldsymbol{Q} = \kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G} + \boldsymbol{G}_2$ $\quad (\kappa^4 + 2\kappa^2 |\boldsymbol{\omega}|^2 + |\boldsymbol{\omega}|^4)$

Conditional distribution in a jointly Gaussian model

$$\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}), \quad \boldsymbol{y}|\boldsymbol{u} \sim \mathsf{N}(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{Q}_{y|u}^{-1}) \qquad (A_{ij} = \psi_j(\boldsymbol{s}_i))$$

$$\boldsymbol{u}|\boldsymbol{y} \sim \mathsf{N}(\boldsymbol{\mu}_{u|y}, \boldsymbol{Q}_{u|y}^{-1})$$

$$\boldsymbol{Q}_{u|y} = \boldsymbol{Q}_u + \boldsymbol{A}^T \boldsymbol{Q}_{y|u} \boldsymbol{A} \quad (\sim\text{"Sparse iff } \psi_k \text{ have compact support"})$$

$$\boldsymbol{\mu}_{u|y} = \boldsymbol{\mu}_u + \boldsymbol{Q}_{u|y}^{-1} \boldsymbol{A}^T \boldsymbol{Q}_{y|u} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}_u)$$

The computational GMRF work-horse

Cholesky decomposition (Cholesky, 1924)

$$\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^\top, \quad \boldsymbol{L} \text{ lower triangular } (\sim \mathcal{O}(n^{(d+1)/2}) \text{ for } d = 1, 2, 3)$$

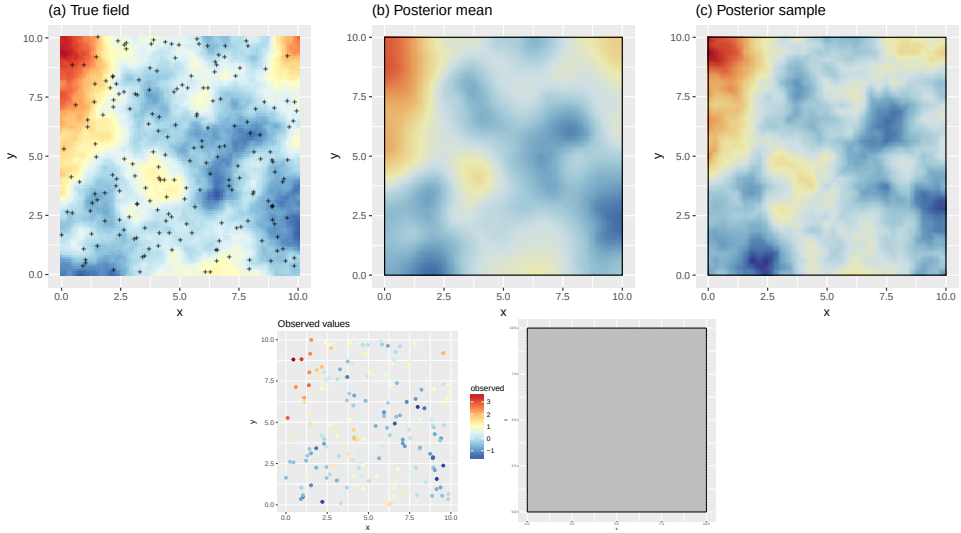
$$\boldsymbol{Q}^{-1}\boldsymbol{x} = \boldsymbol{L}^{-\top}\boldsymbol{L}^{-1}\boldsymbol{x}, \quad \text{via forward/backward substitution}$$

$$\log\det \boldsymbol{Q} = 2\log\det \boldsymbol{L} = 2\sum_i \log L_{ii}$$

André-Louis Cholesky (1875–1918)

"He invented, for the solution of the condition equations in the method of least squares, a very ingenious computational procedure which immediately proved extremely useful, and which most assuredly would have great benefits for all geodesists, if it were published some day." (Euology by Commandant Benoit, 1922)

Example: 2D georeferenced data



How to choose a triangulation mesh?

- SPDE solutions with Neumann boundary conditions are not stationary processes; there is a boundary effect on the covariance structure; visible as inflated variance (factor 2 for straight boundaries)
- Easy workaround: extend the domain boundary
- Small triangles lead to good continuous function approximation properties
- Small triangles lead to expensive calculations
- Resolve the tradeoff by choosing the triangles to be *small enough* in relation to the correlation length. Need intuition!
- Exercise: Given $\mathsf{E}(u_0) = \mathsf{E}(u_1) = 0$, $\mathsf{Var}(u_0) = \sigma_0^2$, $\mathsf{Var}(u_1) = \sigma_1^2$, and $\mathsf{Cov}(u_0, u_1) = \rho\sigma_0\sigma_1$, what is the variance of the linear interpolation $(1 - z)u_0 + zu_1$, $z \in [0, 1]$?
- When the triangle edge lengths decrease, the "ρ" values increase and the continuous/discrete model discrepancy decreases. This can be visualised:
  The interactive tool `INLA::meshbuilder()` can help build intuition

Part 2: Fast Bayesian inference & method and model assessment

Laplace approximations for non-Gaussian observations

Quadratic posterior log-likelihood approximation

$$
\begin{aligned}
p(\boldsymbol{u} \mid \boldsymbol{\theta}) &\sim \mathsf{N}(\boldsymbol{\mu}_u, \boldsymbol{Q}_u^{-1}), \quad \boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta} \sim p(\boldsymbol{y} \mid \boldsymbol{u}) \\
p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) &\sim \mathsf{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{Q}}^{-1}) \\
\mathbf{0} &= \nabla_{\boldsymbol{u}} \left\{ \ln p(\boldsymbol{u} \mid \boldsymbol{\theta}) + \ln p(\boldsymbol{y} \mid \boldsymbol{u}) \right\} \big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}} \\
\widetilde{\boldsymbol{Q}} &= \boldsymbol{Q}_u - \nabla_{\boldsymbol{u}}^2 \ln p(\boldsymbol{y} \mid \boldsymbol{u}) \big|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}}
\end{aligned}
$$

Direct Bayesian inference with INLA (`r-inla.org` & `inlabru.org`)

$$
\begin{aligned}
\widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) &\propto \left. \frac{p(\boldsymbol{\theta}) p(\boldsymbol{u} \mid \boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\theta})}{p_G(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{u}=\widetilde{\boldsymbol{\mu}}} \\
\widetilde{p}(\boldsymbol{u}_i \mid \boldsymbol{y}) &\propto \int p_{GG}(\boldsymbol{u}_i \mid \boldsymbol{y}, \boldsymbol{\theta}) \widetilde{p}(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}
\end{aligned}
$$

The main practical limiting factors for the INLA method are the number of latent variables and the number model parameters.

Example: Point process data

Log-Gaussian Cox processes

Point intensity:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + u(\boldsymbol{s})\right)$$

Inhomogeneous Poisson process log-likelihood:

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) = |D| - \int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} + \sum_{k=1}^{N} \ln \lambda(\boldsymbol{y}_k)$$

The likelihood can be approximated numerically, e.g.

$$\int_D \lambda(\boldsymbol{s})\, d\boldsymbol{s} \approx \sum_{j=1}^{n} \lambda(\boldsymbol{s}_j) w_j,$$

where s_j are mesh nodes, and $w_j = \langle \psi_j, 1 \rangle_D$

Example: Point process data (cont)

Discretised field and likelihood:

$$\lambda(\boldsymbol{s}) = \exp\left(\sum_i b_i(\boldsymbol{s})\beta_i + \sum_j \psi_j(\mathbf{s})u_j\right)$$

$$\ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx |D| - \sum_{j=1}^{n} \lambda(\boldsymbol{s}_j)w_j + \sum_{k=1}^{N} \ln \lambda(\boldsymbol{y}_k)$$

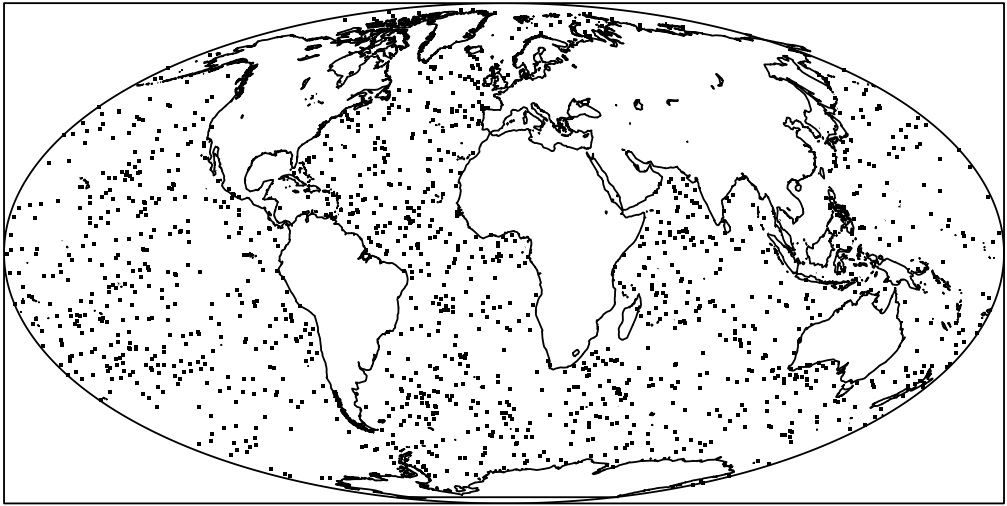
Then, with $\boldsymbol{\lambda_D} = \big[\lambda(s_i)\big]$, $\boldsymbol{A}_D = \big[\psi_j(s_i)\big]$, and $\boldsymbol{A}_y = \big[\psi_j(y_i)\big]$,

$$\nabla_{\boldsymbol{u}} \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w})\boldsymbol{\lambda}_D + \boldsymbol{A}_y^\top \mathbf{1}$$

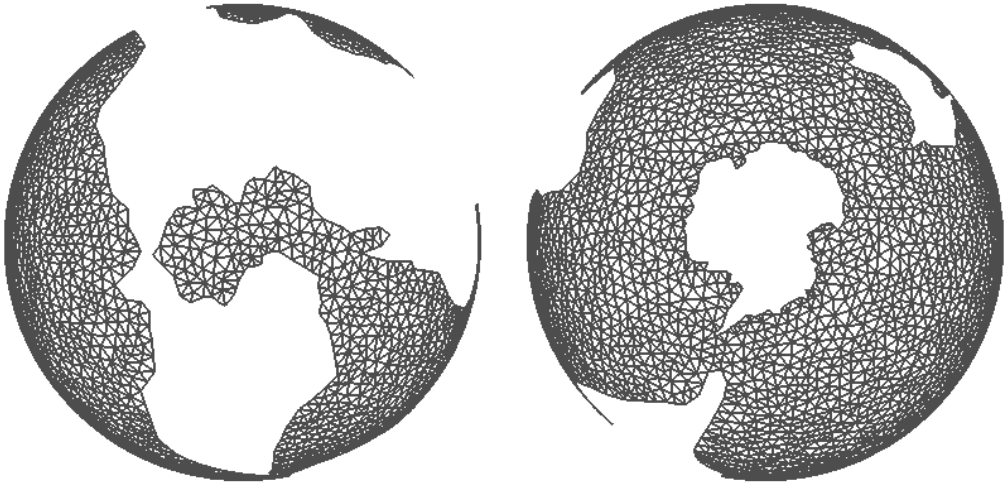
$$\nabla^2_{\boldsymbol{u}} \ln p(\{\boldsymbol{y}_k\} \mid \boldsymbol{\lambda}) \approx -\boldsymbol{A}_D^\top \operatorname{diag}(\boldsymbol{w}) \operatorname{diag}(\boldsymbol{\lambda}_D)\boldsymbol{A}_D$$

and similarly for ∇_β, ∇^2_β, and $\nabla_u \nabla_\beta$.

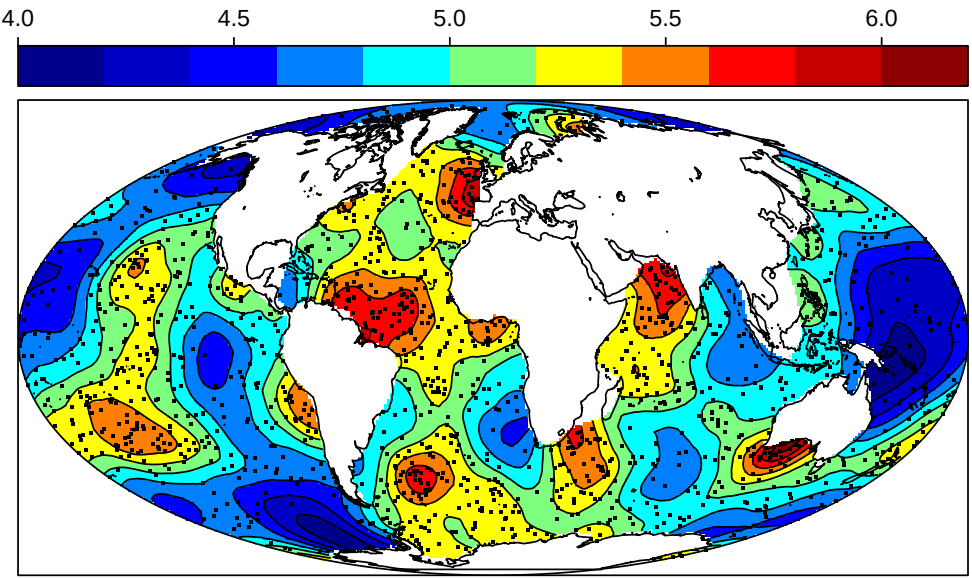
Concept illustration: rogue waves



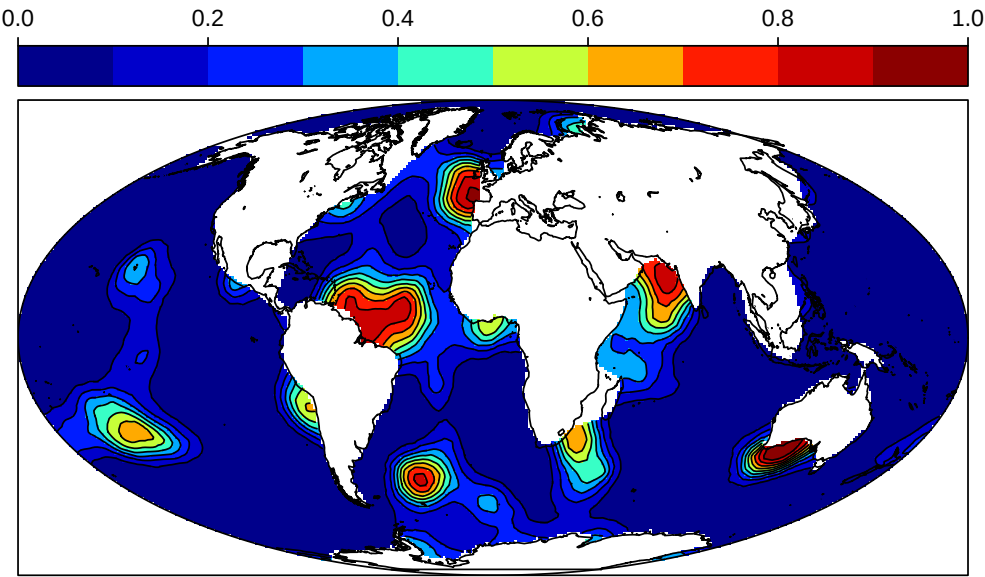
Mesh of the ocean subset of the globe



Posterior mean of the log-intensity



Marginal posterior probabilities for exceeding a threshold



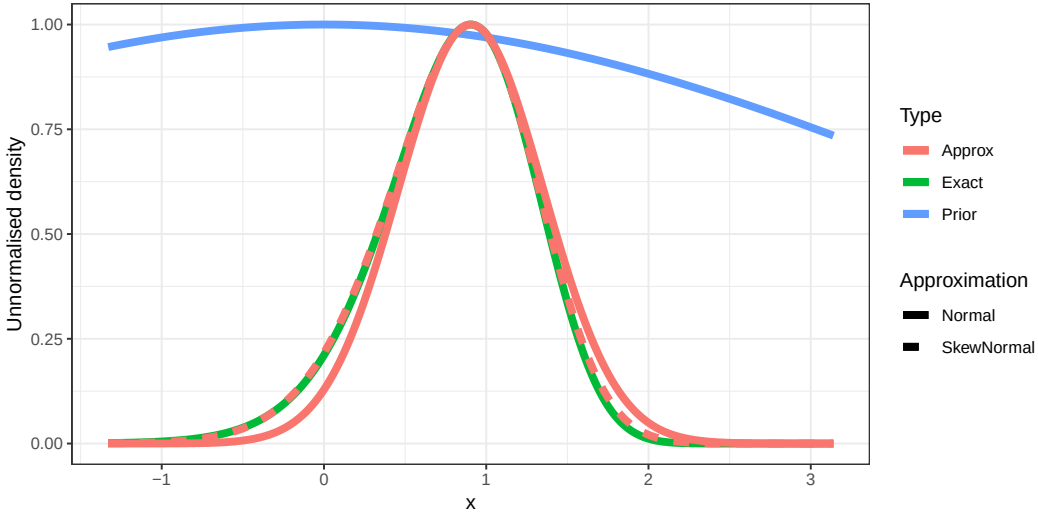
Bias and skewness improvement

- For skewed posteriors, the Normal approximation at the mode is biased
- Can use higher order derivatives at the mode to find better approximations
- Example: Match 2nd and 3rd order derivatives of the log-posterior density to a skew-Normal distribution, at the posterior mode.

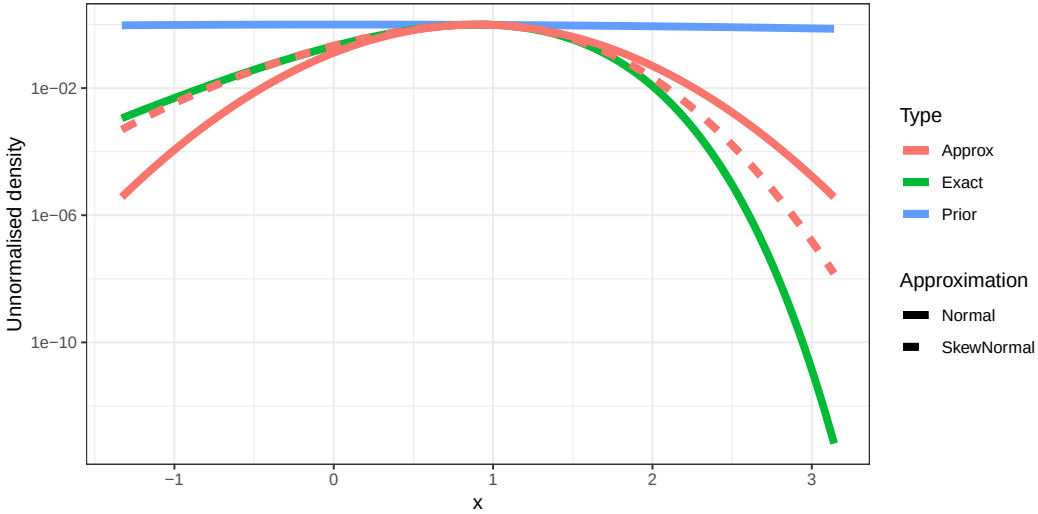
Skew-Normal distribution

Let $z = (x - m)/s$.
The skew-Normal density is defined by $p(x) = \frac{2}{s}\phi(z)\Phi(\alpha z)$, where $\alpha \in \mathbb{R}$ controls the skewness.
The first order derivative of the log-density is $-z + \frac{\alpha\phi(\alpha z)}{s\Phi(\alpha z)}$.
Higher order derivatives are straightforward (but tedious) to derive.

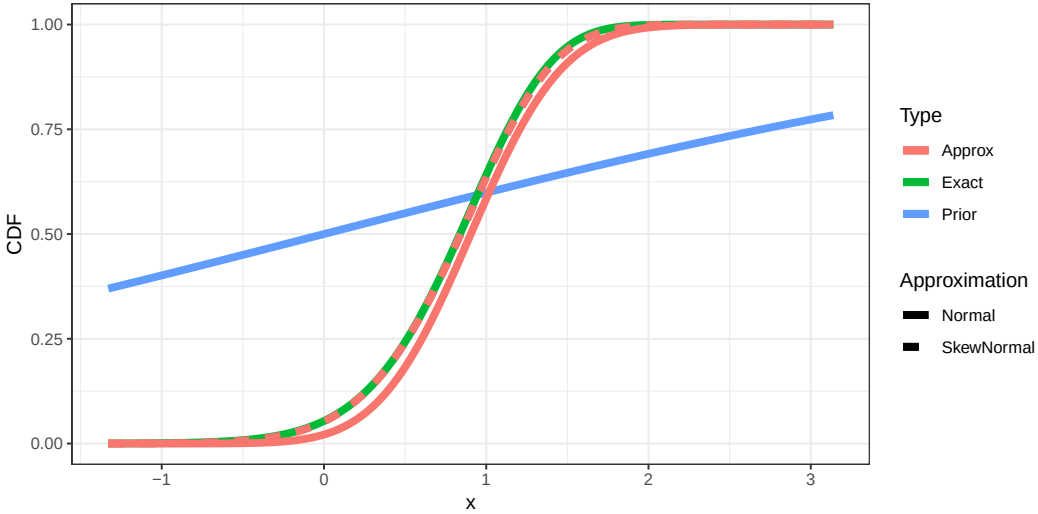
Densities


1.00
0.75
0.50
0.25
0.00
Unnormalised density
−1
0
1
2
3
x
Type
Approx
Exact
Prior
Approximation
Normal
SkewNormal

Log-densities



Cumulative distribution functions (CDF)



Bayesian method correctness assessment

For each $k = 1, \ldots, K$,

1. Sample

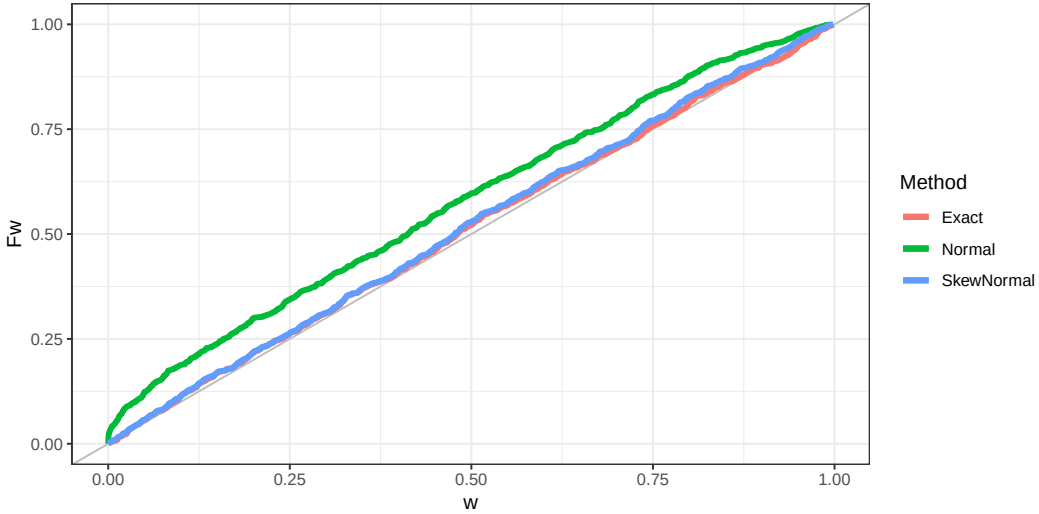
$$\begin{aligned}\boldsymbol{\theta}^{(k)} &\sim p(\boldsymbol{\theta}) \\ \boldsymbol{u}^{(k)} &\sim p(\boldsymbol{u} \mid \boldsymbol{\theta}^{(k)}) \\ \boldsymbol{y}^{(k)} &\sim p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)})\end{aligned}$$

2. The method being assessed has posterior density approximation $\widehat{p}(h(\boldsymbol{\theta}, \boldsymbol{u})|\boldsymbol{y}^{(k)})$
3. Compute The CDF value $w^{(k)} = F_{\widehat{p}(h(\boldsymbol{\theta},\boldsymbol{u})|\boldsymbol{y}^{(k)})}(h(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}))$

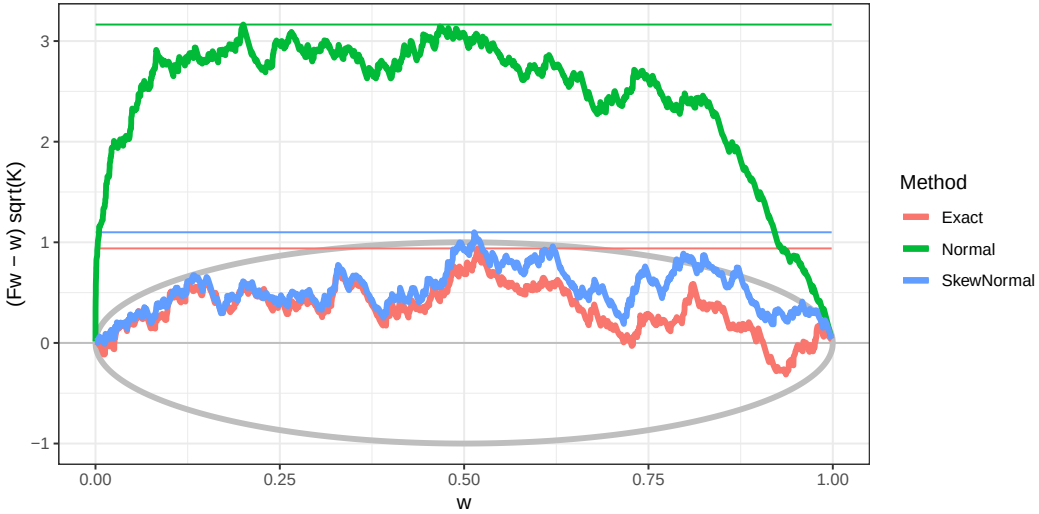
If the method recovers the correct posterior distributions, then $w^{(k)} \sim \mathsf{Unif}(0, 1)$, independent over $k = 1, \ldots, K$.

If the method does not recover the correct posterior distributions, then we expect to see some deviation from $\mathsf{Unif}(0, 1)$.

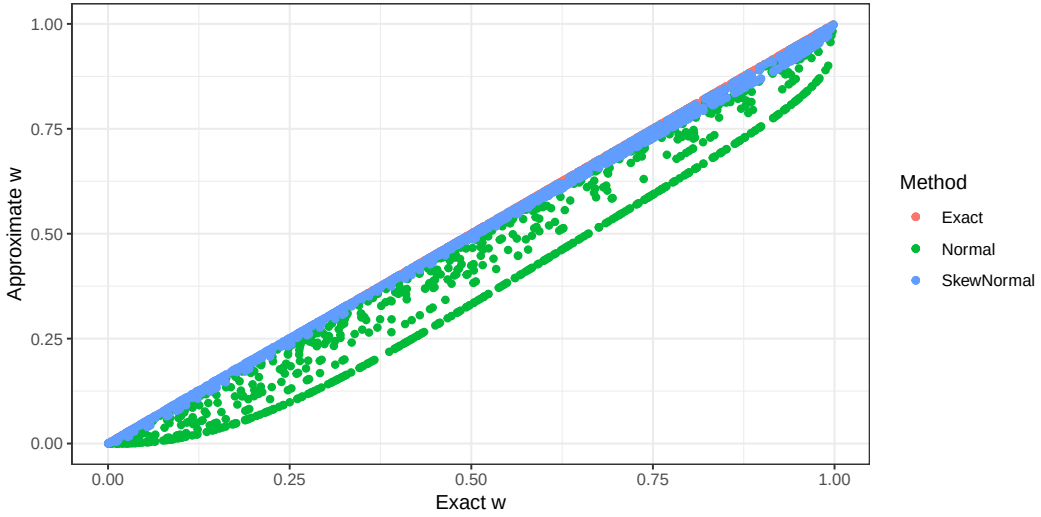
CDF comparison



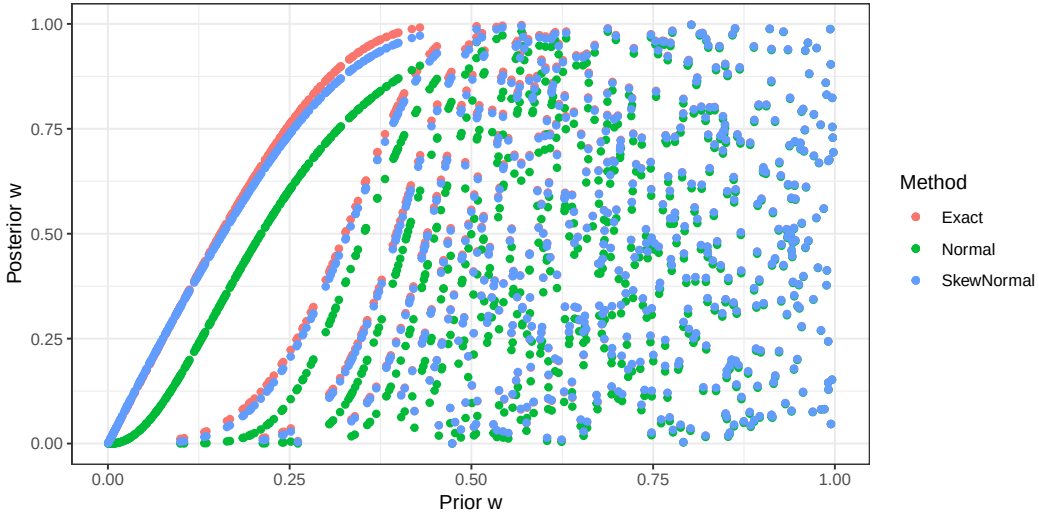
Kolmogorov-Smirnov diagnostic plots



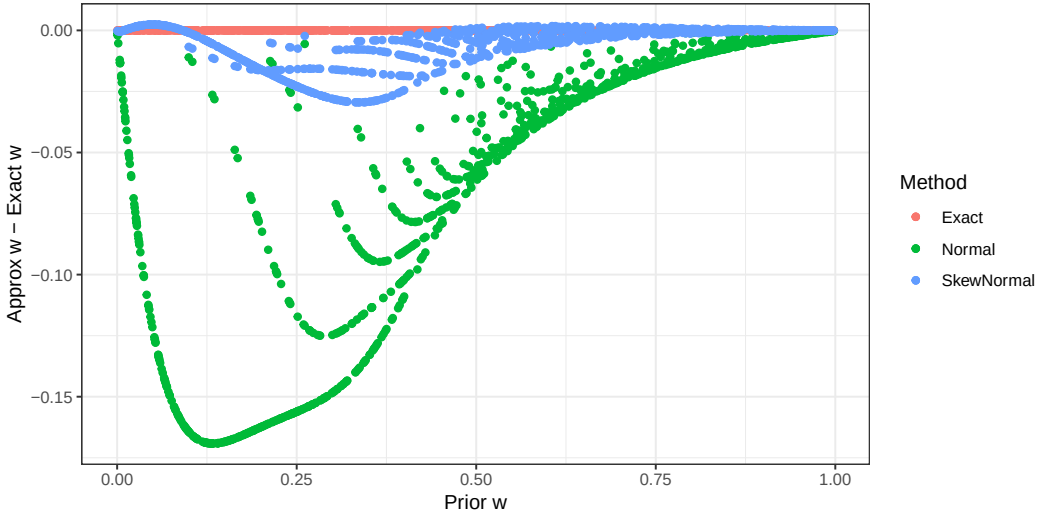
Exact vs Approximate w



Prior vs Posterior w



Prior vs Posterior w difference



Procedure for sampling based Bayesian methods

MCMC and other Monte Carlo methods do not provide CDF values, which then need to be estimated.

Posterior correctness assessment from samples

Generate samples $(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}, \boldsymbol{y}^{(k)})$ from the full model, as before.
For each $k = 1, \ldots, K$, generate J samples from $(\boldsymbol{\theta}^{(j|k)}, \boldsymbol{u}^{(j|k)}) \sim p(\boldsymbol{\theta}, \boldsymbol{u} \mid \boldsymbol{y}^{(k)})$.
Compute the approximate CDF value as an empirical CDF for the samples:

$$w^{(k)} = \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\left\{ h(\boldsymbol{\theta}^{(j|k)}, \boldsymbol{u}^{(j|k)}) \leq h(\boldsymbol{\theta}^{(k)}, \boldsymbol{u}^{(k)}) \right\} - \frac{1}{2J}$$

which is a normalised order statistic.

Notes:

- The assessment approach assumes we can sample *exactly* (and independently) from the prior model.
- MCMC methods capable of posterior samples are not necessarily good at generating from the prior.
- The null distribution for the K-S test now depends on both K and J, as well as the dependence between the Monte Carlo samples.

Model vs method vs implementation

- The Bayesian method correctness assessment method specifically targets the implementation of a method
- *Model assessment* based on output from a method implementation is meaningless if we we don't have some trust in the method and implementation
- Information criteria based on the full model likelihood are popular but are often hard to interpret
- Probabilistic predictions can be easier to interpret, and are often cheap to compute (in particular if one is already doing expensive Bayesian inference)
- We let F denote the CDF of a probabilistic prediction of an observation y
- The context can be cross-validation or estimation/validation/test data splits

Scores

- We want to quantify how well our predictions represent the test data.
- We define *scores* $S(F, y)$ that in some way measure how well the prediction F matched the actual value, y.
- The scores defined here are *negatively oriented*, meaning that the *lower the score, the better*.

Squared errors and log-likelihood scores

- Squared Error (SE): $S_{\text{SE}}(F, y) = (y - \widehat{y}_F)^2$,
  where $\widehat{y}_F$ is a point estimate under F, e.g. the expectation μ_F.
- Logarithmic/Ignorance score (LOG/IGN): $S_{\text{LOG}}(F, y) = -\log f(y)$,
  where $f(\cdot)$ is the predictive density.
- Dawid-Sebastiani (DS): $S_{\text{DS}}(F, y) = \frac{(y-\mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)$.

Score expectations and proper scoring rules

- What functions of the predictive distributions are useful scores?
- We want to reward accurate (unbiased) and precise (small variance) predictions, but not at the expense of understating true uncertainty.
- First, we define the expectation of a score under a true distribution G as

$$S(F, G) = \mathsf{E}_{y\sim G}[S(F, y)]$$

Proper scores/scoring rules

A negatively oriented score is *proper* if it fulfils

$$S(F, G) \geq S(G, G).$$

A proper score that has equality of the expectations *only* when F and G are the same, $F(\cdot) \equiv G(\cdot)$, is said to be *strictly proper*.

The practical interpretation of this is that a proper score does not reward cheating; stating a lower (or higher) forecast/prediction uncertainty will not, on average, give a better score than stating the truth.

Absolute error and CRPS

Absolute error and Continuous Ranked Probability Score

- Absolute Error (AE): $S_{\text{AE}}(F, y) = |y - \widehat{y}_F|$, where $\widehat{y}_F$ is a point estimate under F, e.g. the *median* $F^{-1}(1/2)$.
- CRPS: $S_{\text{CRPS}}(F, y) = \int_{-\infty}^{\infty} \left[\mathbb{I}(y \leq x) - F(x)\right]^2 \, \mathrm{d}x$

Average scores

Average score

Given a collection of prediction/truth pairs, $\{(F_i, y_i), i = 1, \ldots, n\}$, define the *average* or *mean* score:

$$\overline{S}(\{(F_i, y_i), i = 1, \ldots, n\}) = \frac{1}{n} \sum_{i=1}^{n} S(F_i, y_i)$$

- When comparing prediction quality, we often look at the difference in average scores across the test data set.
- For modern, complex models with explicit spatial and temporal model components, the *pairwise* differences may be useful: For two prediction methods, F and F',

$$S_i^{\Delta}(F_i, F_i', y_i) = S(F_i, y_i) - S(F_i', y_i)$$

  We can have $\overline{S}^{\Delta} \approx 0$ at the same time as all $|S_i^{\Delta}| \gg 0$, if the two models/methods are both good, but e.g. at different spatial locations.
- How can we assess whether the score differences are indistinguishable?

How good are confidence/prediction interval procedures?

Tradeoffs for CIs

Desired properties for methods generating CIs for a quantity Y:

1. Appropriate *coverage* under the true distribution, G: $\mathsf{P}_G(Y \in CI_F) \geq 1 - \alpha$
2. Narrow intervals

- A wide prediction F helps with 1 but makes 2 difficult
- A narrow prediction F helps with 2 but makes 1 difficult

A proper score for interval predictions

The *Interval Score* For a CI (L_F, U_F) is defined by

$$S_{\text{INT}}(F, y) = U_F - L_F + \frac{2}{\alpha}(L_F - y)\mathbb{I}(y < L_F) + \frac{2}{\alpha}(y - U_F)\mathbb{I}(y > U_F)$$

It is a proper scoring rule, consistent for equal-tail error probability intervals:
$S(F, G)$ is minimised for the narrowest CI that has expected coverage $1 - \alpha$.

Proper scores

$$
\begin{aligned}
S_{\mathsf{SE}}(F, G) &= \mathsf{E}_{y\sim G}[S_{\mathsf{SE}}(F, y)] = \mathsf{E}_{y\sim G}[(y - \mu_F)^2] = \mathsf{E}_{y\sim G}[(y - \mu_G + \mu_G - \mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y - \mu_G)^2 + 2(y - \mu_G)(\mu_G - \mu_F) + (\mu_G - \mu_F)^2] \\
&= \mathsf{E}_{y\sim G}[(y - \mu_G)^2] + 2(\mu_G - \mu_F)\mathsf{E}_{y\sim G}[y - \mu_G] + (\mu_G - \mu_F)^2 \\
&= \sigma_G^2 + (\mu_G - \mu_F)^2
\end{aligned}
$$

This is minimised when $\mu_F = \mu_G$. Therefore $S_{\mathsf{SE}}(F, G) \geq S_{\mathsf{SE}}(G, G) = \sigma_G^2$, so the score is proper. Is it strictly proper?

$$
\begin{aligned}
S_{\mathsf{DS}}(F, G) = \mathsf{E}_{y\sim G}[S_{\mathsf{DS}}(F, y)] &= \frac{\mathsf{E}_{y\sim G}[(y - \mu_F)^2]}{\sigma_F^2} + \log(\sigma_F^2) \\
&= \frac{\sigma_G^2 + (\mu_G - \mu_F)^2}{\sigma_F^2} + \log(\sigma_F^2)
\end{aligned}
$$

This is minimised when $\mu_F = \mu_G$ and $\sigma_F = \sigma_G$. Therefore $S_{\mathsf{DS}}(F, G) \geq S_{\mathsf{DS}}(G, G) = 1 + \log(\sigma_G^2)$, so the score is proper. Is it strictly proper?

Part 3: Lessons from the EUSTACE project