Gaussian processes for spatial modelling in environmental health: parameterizing for flexibility vs. computational efficiency

Chris Paciorek

March 11, 2005

Department of Biostatistics

Harvard School of Public Health

www.biostat.harvard.edu/~paciorek

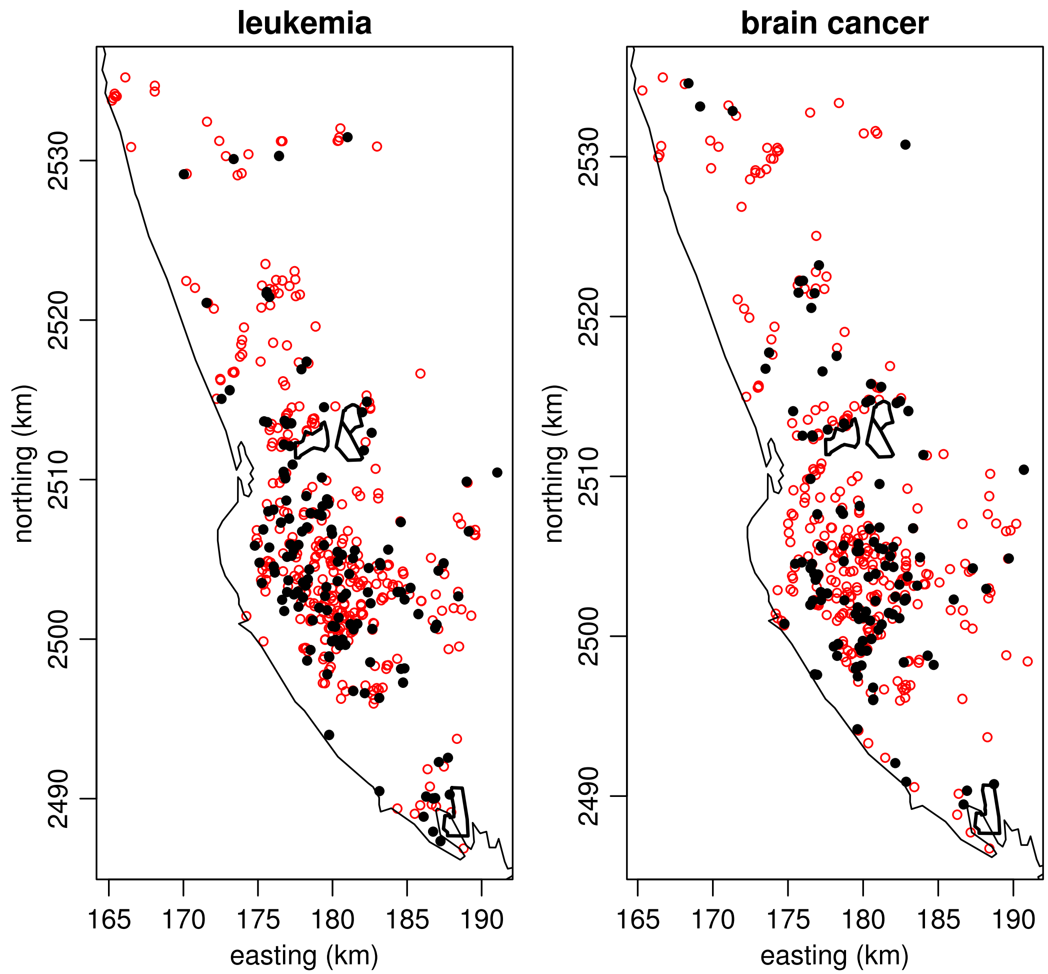
Increased attention to spatial analysis in public health

- data availability: geocoding and GPS for assigning point locations to individuals and monitors

- GIS software:
  - easy data management and manipulation
  - graphical presentation
  - spatially-varying covariate generation

- interest amongst researchers:
  - strong applied interest in kriging and related smoothing methods
  - opportunities for more sophisticated spatio-temporal modelling, particularly Bayesian hierarchical modelling

Petrochemical exposure in Kaohsiung, Taiwan



$n = 495$ $\quad$ $n = 433$

$n_1 = 141$ $\quad$ $n_1 = 121$

Possible approaches for health analysis

- Explicitly estimate pollutant exposure - difficult retrospectively
- Use distance to exposure source as covariate
- Use a moving window/multiple testing to detect clusters of cases
  - default approach - software available
- **Include space as a covariate to provide a map of risk**

$$\begin{aligned} H_i &\sim \text{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) \\ \text{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) &= \boldsymbol{x_i}^T \boldsymbol{\beta} + g_{\boldsymbol{\theta}}(\boldsymbol{s_i}) \end{aligned}$$

Particulate matter exposure in the Nurses' Health Study

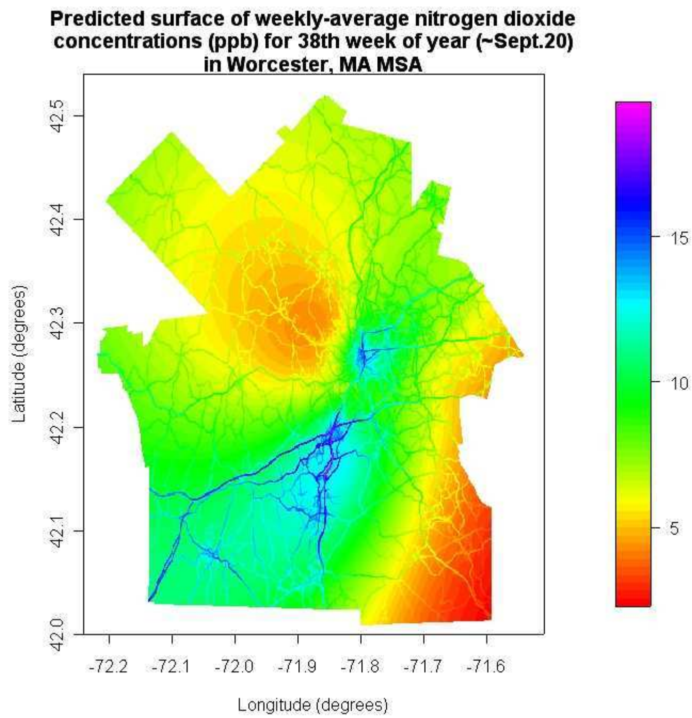
- estimate individual exposure, 1985-2003
  - EPA monitoring for large-scale spatio-temporal heterogeneity
  - spatially-varying covariates for local heterogenity
    * distance to roads, climate variables, local land use, ...
    * generated using GIS

- basic additive exposure model:

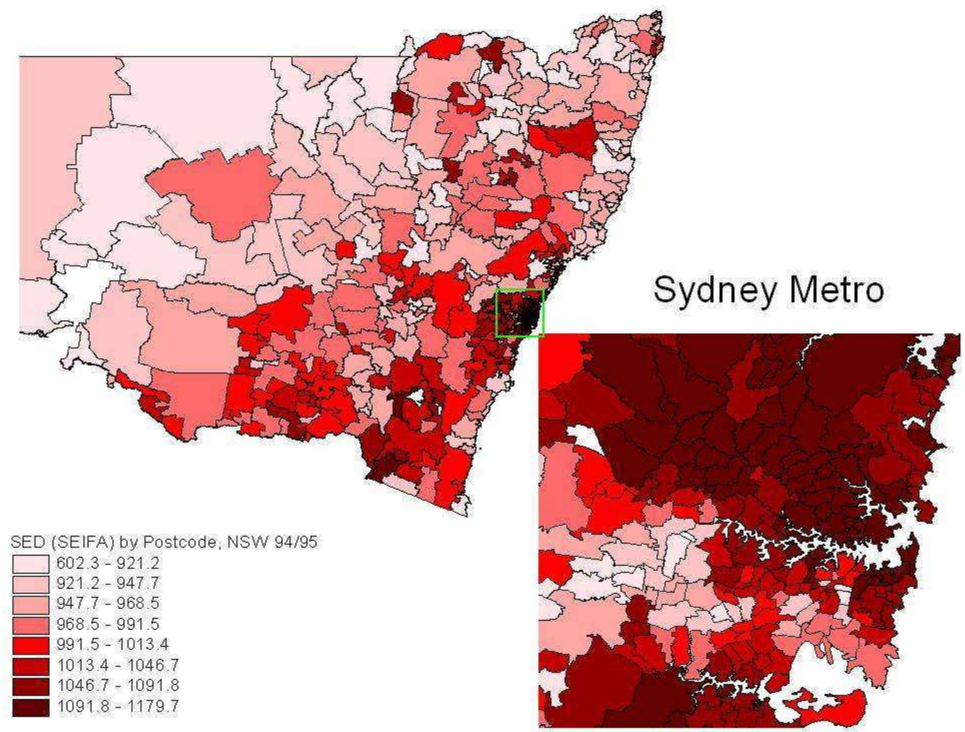
$$\begin{aligned} \log E_i &\sim \mathsf{N}(f(\boldsymbol{x_i}, \boldsymbol{s_i}), \eta^2) \\ f(\boldsymbol{x_i}, \boldsymbol{s_i}) &= \sum_p h_p(x_i) + g_{\boldsymbol{\theta}}(\boldsymbol{s_i}) \end{aligned}$$

- geocoding of individual residences every two years
  - relate estimated exposure to health outcomes (chronic heart disease)

- geocoding and GIS make this possible; spatial statistics provides a rigorous framework



Health outcomes by postcode in NSW, Australia



- methodological challenges
  - areal (postcode) units vary drastically in size
  - data misalignment
- relate areal data to a latent smooth process, $g_{\boldsymbol{\theta}}(\cdot)$ (Kelsall & Wakefield, Rathouz)
- computational challenges: 650 units, 5 years daily data, 2 sexes, 9 age groups

Outline

- Motivating examples
- Introduction to Gaussian processes (GPs)
- Fast Gaussian process modelling
- Flexible Gaussian process modelling
- Bayes and overfitting
- The future: flexibility + efficiency + hierarchical modelling

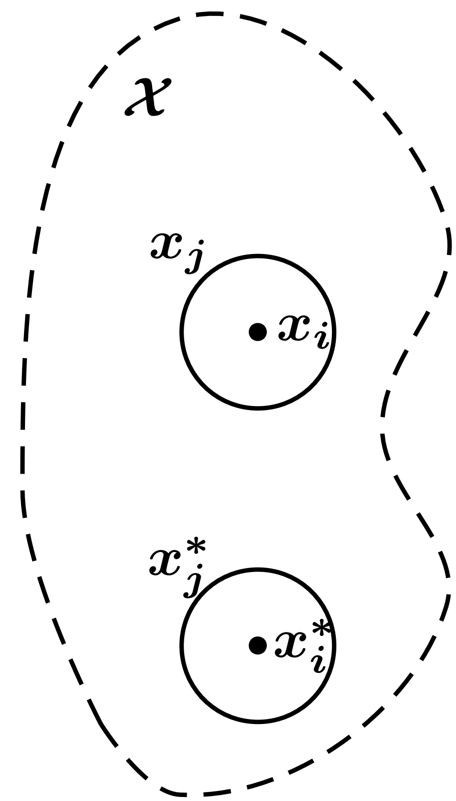
Kriging as a GP model

$$\begin{aligned} Y_i &\sim \mathsf{N}(g(\boldsymbol{s_i}), \eta^2) \\ g(\cdot) &\sim \mathrm{GP}(\mu, C(\cdot; \boldsymbol{\theta})) \end{aligned}$$

- Bayesian model specifies prior distributions for $\boldsymbol{\theta}$ (Bayesian kriging)
- Empirical Bayes/marginal likelihood (i.e., kriging)
  - integrate $g_{\text{train}} = (g(s_1), \ldots, g(s_n))$ out of model
  - estimate $\boldsymbol{\theta}$
    * maximizing marginal posterior
    * maximizing marginal likelihood
    * fitting variogram model for $C(\cdot; \boldsymbol{\theta})$
  - point estimate for spatial process:
    $\mathrm{E}(\boldsymbol{g}_{\text{test}} | \boldsymbol{Y}, \tilde{\boldsymbol{\theta}})$ based conditional normal calculations

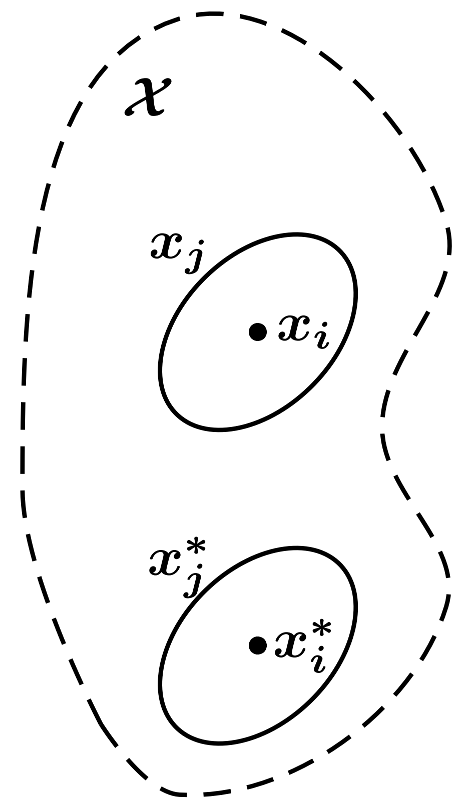
GAUSSIAN PROCESS DISTRIBUTION

- Infinite-dimensional joint distribution for $g(x), \; x \in \mathcal{X}$:
  - ❖ Example: $g(\cdot)$ a spatial process, $\mathcal{X} = \Re^2$
  - ❖ $g(\cdot) \sim \mathbf{GP}(\mu(\cdot), C(\cdot, \cdot))$
- Finite-dimensional marginals are normal
- Types of covariance functions, $C(x_i, x_j)$:
  - ❖ stationary, isotropic
  - ❖ stationary, anisotropic
  - ❖ nonstationary



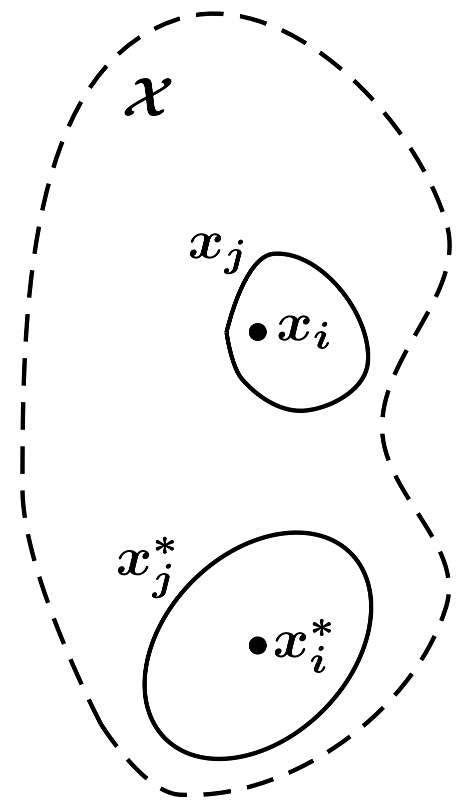
GAUSSIAN PROCESS DISTRIBUTION

- Infinite-dimensional joint distribution for $g(x),\ x \in \mathcal{X}$:
  - ❖ Example: $g(\cdot)$ a spatial process, $\mathcal{X} = \Re^2$
  - ❖ $g(\cdot) \sim \mathbf{GP}(\mu(\cdot), C(\cdot, \cdot))$
- Finite-dimensional marginals are normal
- Types of covariance functions, $C(x_i, x_j)$:
  - ❖ stationary, isotropic
  - ❖ stationary, anisotropic
  - ❖ nonstationary

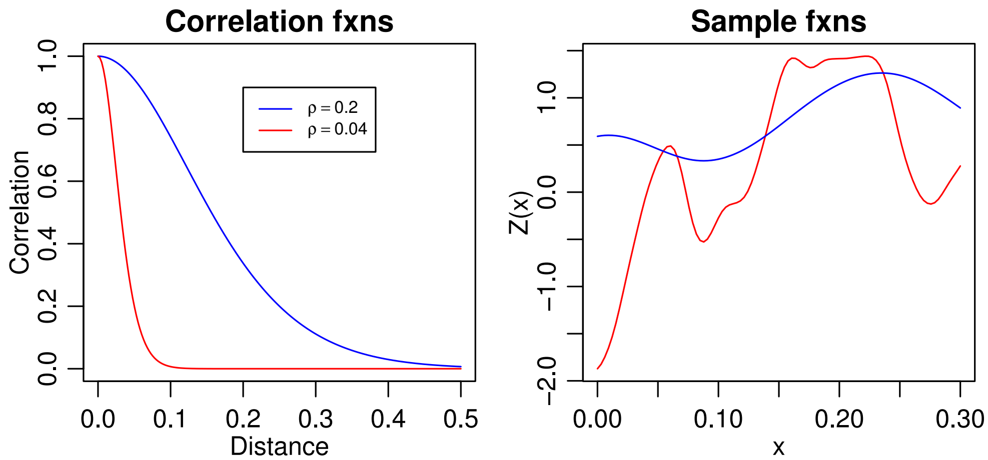


GAUSSIAN PROCESS DISTRIBUTION

- Infinite-dimensional joint distribution for $g(x),\ x \in \mathcal{X}$:
  - ❖ Example: $g(\cdot)$ a spatial process, $\mathcal{X} = \Re^2$
  - ❖ $g(\cdot) \sim \mathbf{GP}(\mu(\cdot), C(\cdot,\cdot))$
- Finite-dimensional marginals are normal
- Types of covariance functions, $C(x_i, x_j)$:
  - ❖ stationary, isotropic
  - ❖ stationary, anisotropic
  - ❖ nonstationary



Stationary Correlation Functions



Matérn form

- $R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\tau}{\rho}\right); \nu > 0,\ \rho > 0$
- Differentiability controlled by ν, asymptotic advantages (Stein)
- Familiar exponential ($\nu = 0.5$) and squared exponential (Gaussian) ($\nu \to \infty$) correlations as special and limiting cases

Computational challenges of GPs

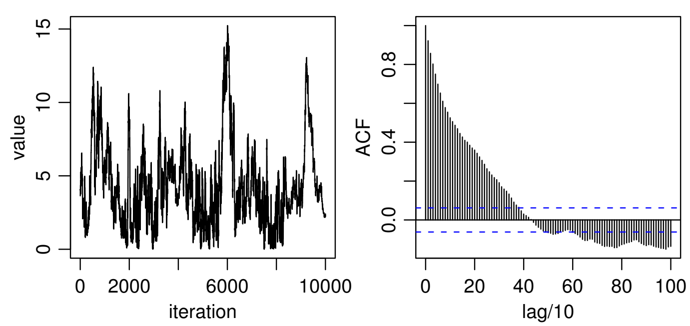
- even marginal likelihood in normal error model is intensive:

$$g(\cdot) \sim \text{GP}(\mu(\cdot), C_{\boldsymbol{\theta}}(\cdot,\cdot)) \Rightarrow \boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, C_{\boldsymbol{\theta}} + \eta^2 I)$$

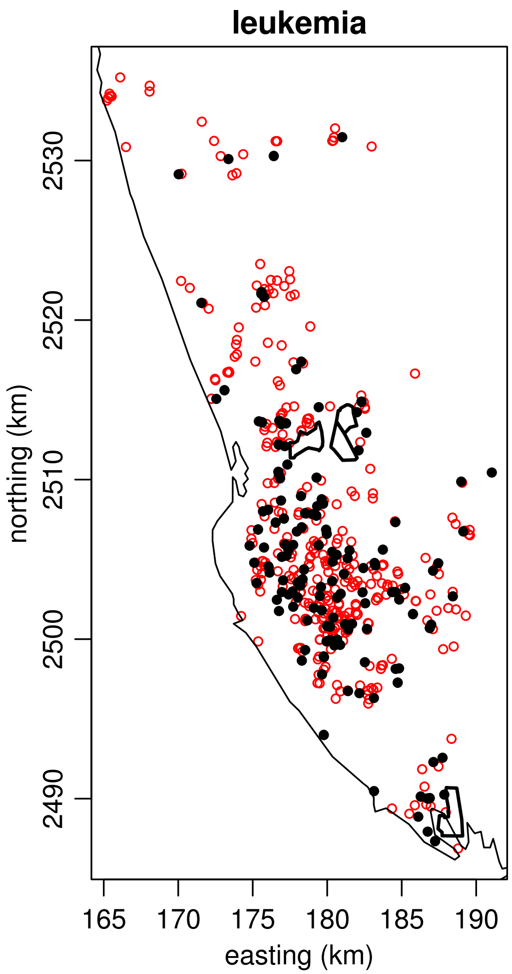
  – $O(n^3)$ fitting: $|C_{\boldsymbol{\theta}} + \eta^2 I|$ and $(C_{\boldsymbol{\theta}} + \eta^2 I)^{-1}(\boldsymbol{Y} - \mu\mathbf{1})$

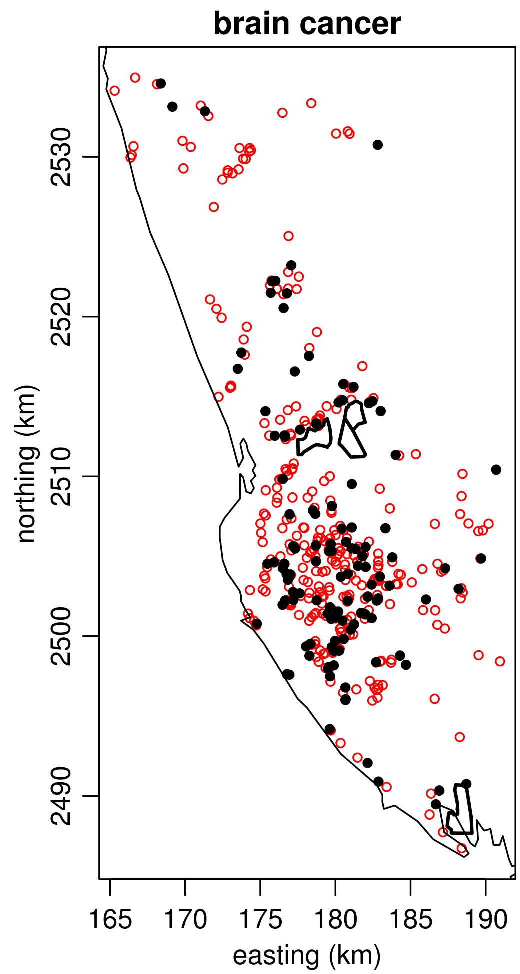
- non-Gaussian spatial models particularly difficult

  – spatial process can't be integrated out
  – MCMC mixing is very slow because of high-level structure
    * correlation amongst process values and between process values and process hyperparameters



Petrochemical exposure in Kaohsiung, Taiwan





$n = 495$ $\quad$ $n = 433$

$n_1 = 141$ $\quad$ $n_1 = 121$

Modelling Framework

$$\begin{aligned} H_i &\sim \text{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) \\ \text{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) &= \boldsymbol{x_i}^T \boldsymbol{\beta} + g_{\boldsymbol{\theta}}(\boldsymbol{s_i}) \end{aligned}$$

- basic spatial model for $\boldsymbol{g}_{\boldsymbol{\theta}}^s = (g_{\boldsymbol{\theta}}(\boldsymbol{s_1}), \ldots, g_{\boldsymbol{\theta}}(\boldsymbol{s_n}))$
  - GAM: $g_{\boldsymbol{\theta}}(\cdot)$ is a two-dimensional smooth term
    - basis representation

$$\boldsymbol{g}_{\boldsymbol{\theta}}^s = Z\boldsymbol{u}$$

    - Gaussian process representation:

$$g(\cdot) \sim \text{GP}(\mu(\cdot), C_{\boldsymbol{\theta}}(\cdot, \cdot)) \Rightarrow \boldsymbol{g}_{\boldsymbol{\theta}}^s \sim N(\boldsymbol{\mu}, C_{\boldsymbol{\theta}})$$

  - GLMM: $\boldsymbol{g}_{\boldsymbol{\theta}}^s = Z\boldsymbol{u}$
    - correlated random effects, $\boldsymbol{u} \sim N(0, \Sigma)$

Approaches

- Bayesian spectral basis model fit by MCMC (Wikle, 2002) [B-SB]

- penalized likelihood based on mixed model (radial basis functions) with REML smoothing
  (Kammann and Wand, 2003; Ngo and Wand, 2004) [PL-PQL]

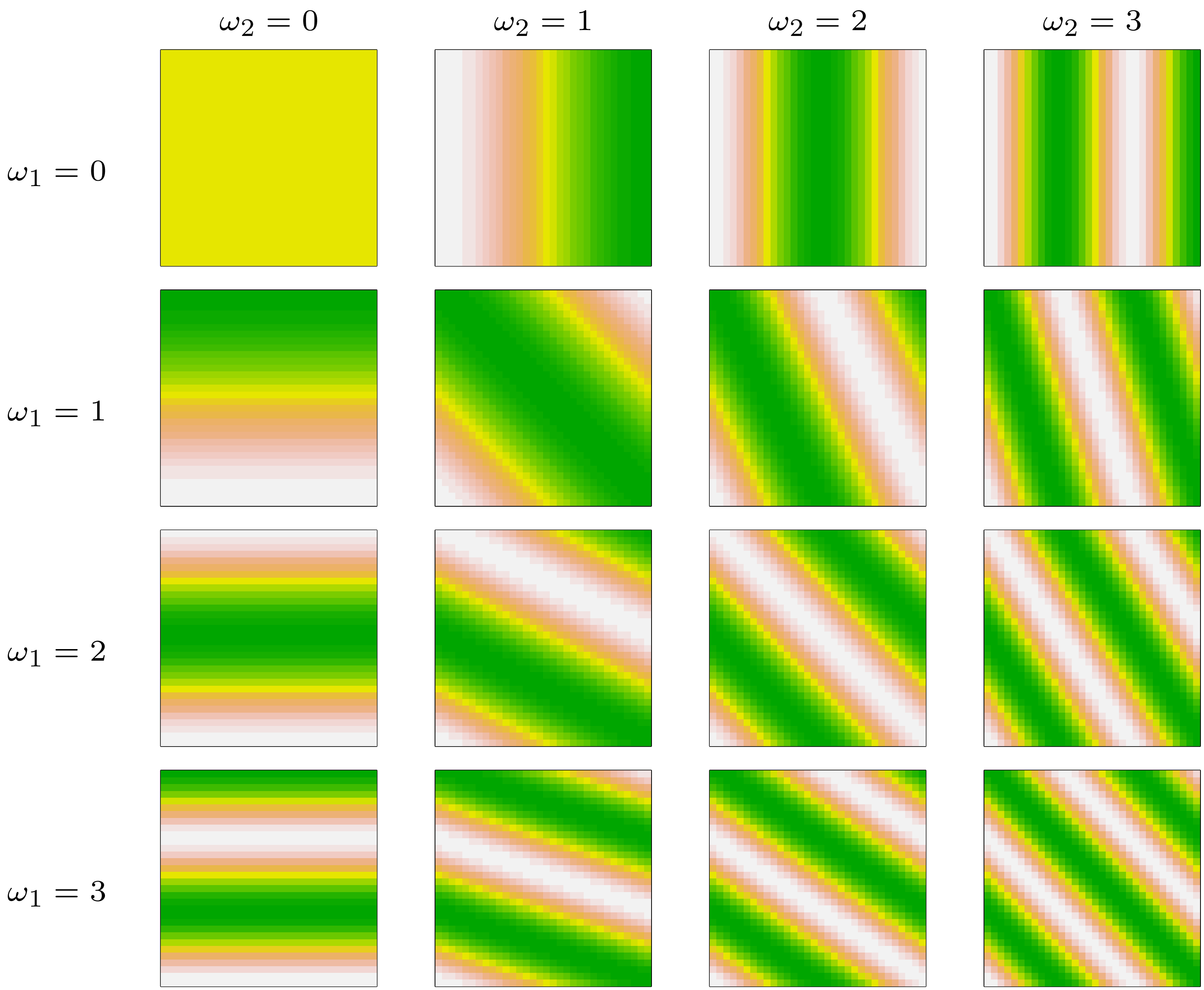
- penalized likelihood with GCV smoothing
  (Wood, 2001, 2003, 2004) [PL-GCV]

- Bayesian mixed model/radial basis functions fit by MCMC
  (Zhao and Wand 2004) [B-Geo]

Bayesian spectral basis function model

- computationally efficient basis function construction (Wikle 2002)
- $\boldsymbol{g}^{\#} = Z\boldsymbol{u}$ and $\boldsymbol{g}^{s} = \sigma P \boldsymbol{g}^{\#}$
  - piecewise constant gridded surface on k by k grid
  - P maps observation locations to nearest grid point
- Z is the Fourier (spectral) basis and $Z\boldsymbol{u}$ is the inverse FFT
- $Z\boldsymbol{u}$ is approximately a Gaussian process (GP) when...
  - $\boldsymbol{u} \sim N(0, \text{diag}(\pi_\theta(\boldsymbol{\omega})))$ for Fourier frequencies, $\boldsymbol{\omega}$
  - spectral density, $\pi_\theta(\cdot)$, of GP covariance function defines V$(\boldsymbol{u})$

Bayesian spectral basis functions



Comparison with usual GP specification

- spectral basis uses FFT
  - $O\left((k^2)\log(k^2)\right)$
  - additional observations are essentially free for fixed grid
  - fast computation and prediction of surface given coefficients
  - a priori independent coefficients give fast computation of prior and help with mixing

Penalized likelihood using GLMM framework with REML [PL-PQL]

- $\boldsymbol{g}^s = Z\boldsymbol{u}$, $Z = \Psi_{nk}\Omega_{kk}^{-\frac{1}{2}}$, $\boldsymbol{u} \sim N(0, \sigma_u^2)$ - variance component provides complexity penalty

- Ω contains pairwise spatial covariances between k knot locations and Ψ between n data locations and k knot locations

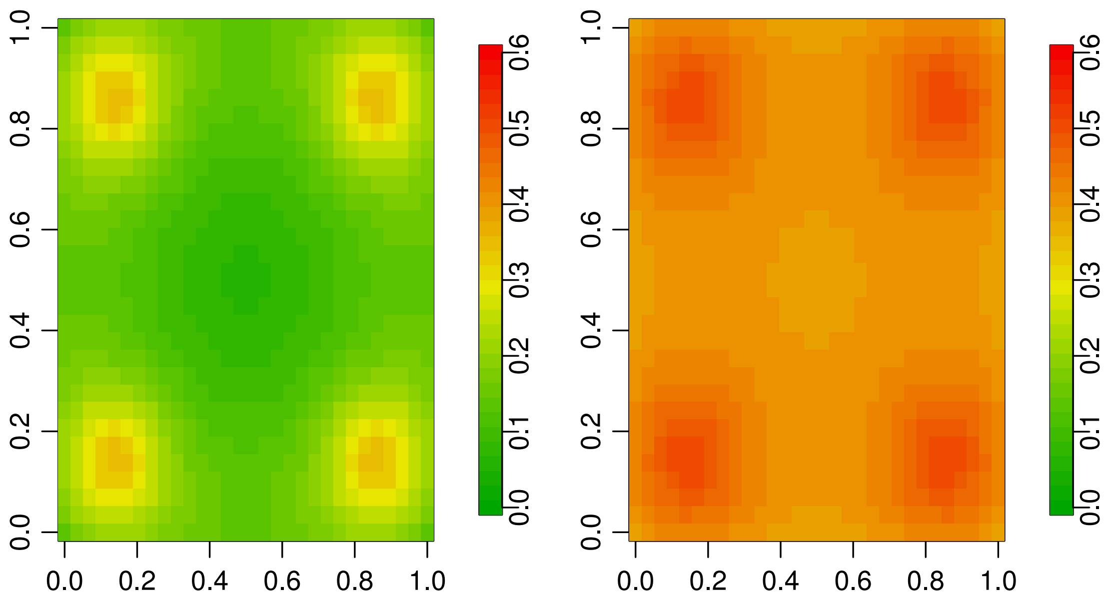
- potential covariance functions:
  - thin plate spline generalized covariance function, $C(\tau) = \tau^2 \log \tau$
  - Matérn correlation function, $R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\tau}{\rho}\right)$, with ρ and ν fixed

- computationally efficient approximation of a Gaussian process representation for $\boldsymbol{g}^s$

- PQL approach - IWLS fitting of $(\boldsymbol{\beta}, \boldsymbol{u})$ with REML estimation of σ_u^2 within the iterations using MM software

GLMM basis functions

- radial basis functions centered at the knots
- 4 of 64 functions displayed:



TPS Matérn

Penalized likelihood using GCV [PL-GCV]

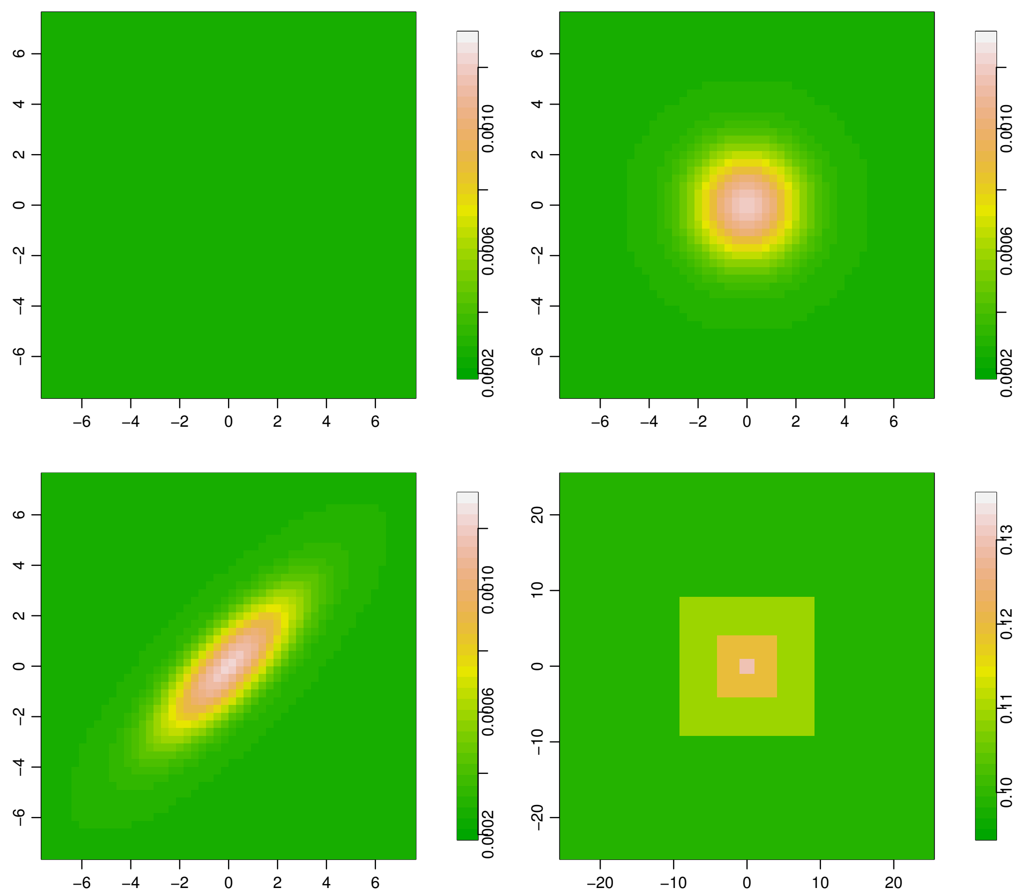
- thin plate spline basis for $g(\cdot)$
- truncated eigendecomposition of basis matrix increases computational efficiency
- IWLS fitting of $(\boldsymbol{\beta}, \boldsymbol{u})$ with GCV estimation of penalty
- easy implementation using the R mgcv library – gam()

Bayesian geoadditive model [B-Geo]

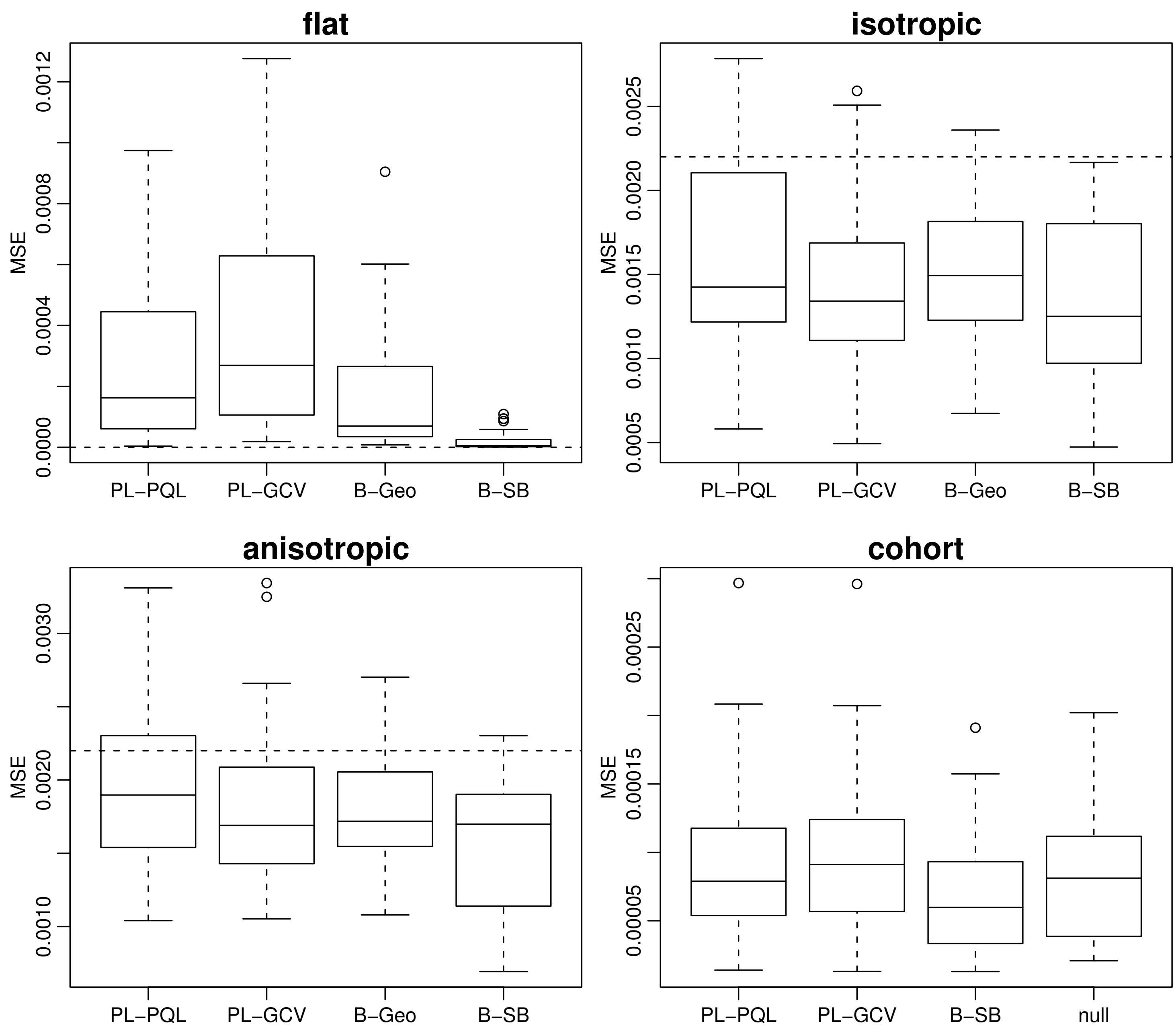
- Bayesian version of GLMM framework already described
  - $\boldsymbol{g}^s = Z\boldsymbol{u}$, $Z = \Psi_{nk}\Omega_{kk}^{-\frac{1}{2}}$, $\boldsymbol{u} \sim N(0, \sigma_u^2)$
  - natural Bayesian complexity penalty through prior on $\boldsymbol{u}$
- thin plate spline covariance or Matérn correlation basis construction of Ψ and Ω
- MCMC implementation - ensuring mixing is not simple
  - Metropolis-Hastings for $\boldsymbol{u}$ using conditional posterior mean and variance based on linearized observations
  - joint proposals for σ_u^2 and $\boldsymbol{u}$ to ensure that $\boldsymbol{u}$ remains compatible with its variance component

Simulated datasets

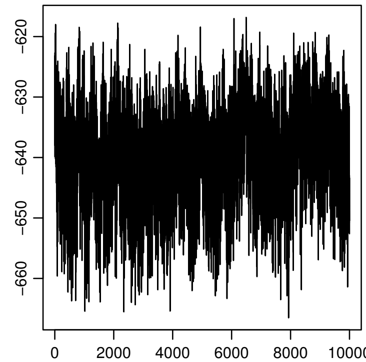
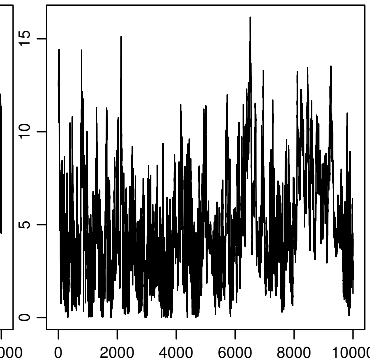
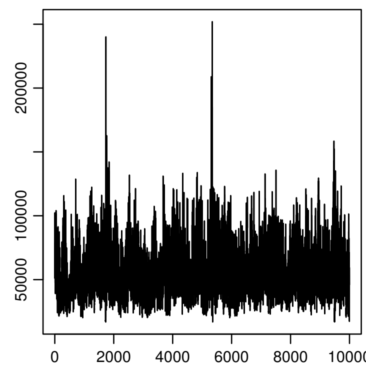
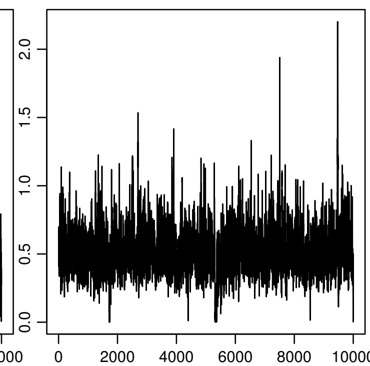
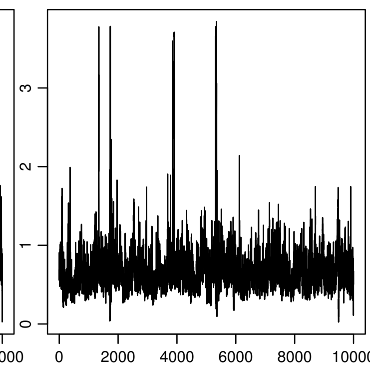
- 3 case-control scenarios: $n_0 = 1,000$; $n_1 = 200$; $n_{\text{test}} = 2500$ on 50 by 50 grid
- 1 cohort scenario: $n = 10,000$; $n_{\text{test}} = 2500$ on 50 by 50 grid



Assessment on 50 simulated datasets



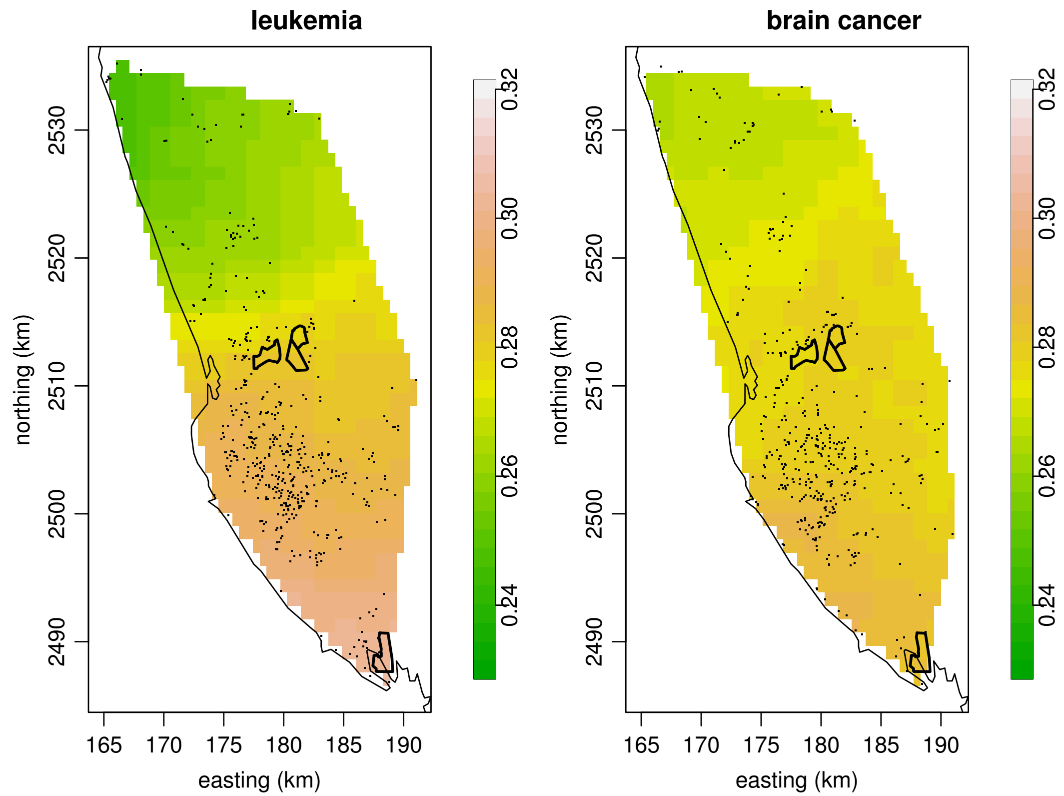
Mixing and speed of Bayesian methods

	speed (1000 its)	speed (1000 eff. its)	log posterior trace	σ trace	ρ trace
B-Geo	15 min.	104 hr.			
B-SB	1.3 min.	6 hr.			

Taiwan revisited - assessment

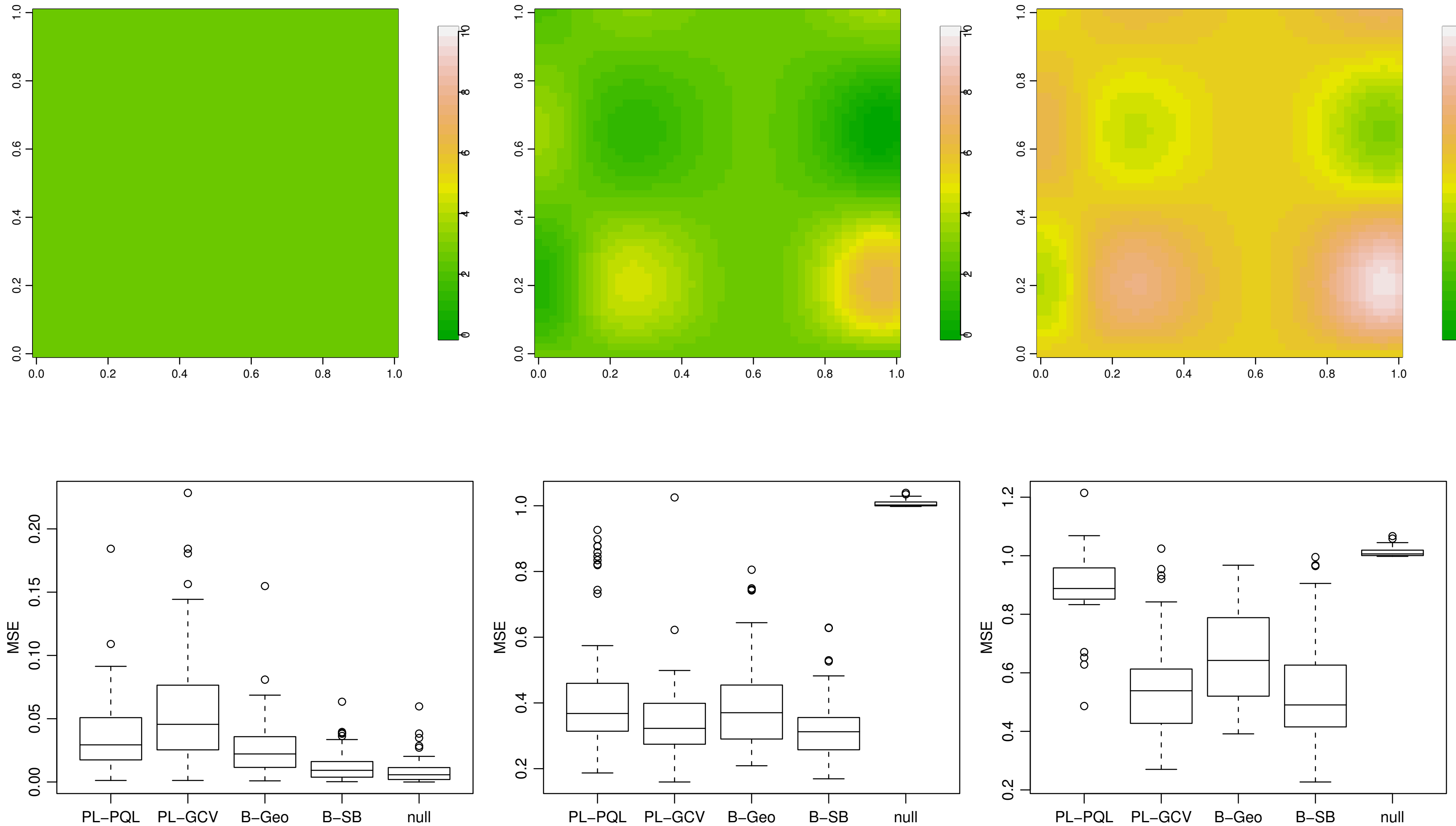
Summed test deviance over 10-fold C-V sets

	leukemia	brain cancer
PL-GCV	590.1	529.8
PL-PQL	585.6	529.5
B-Geo	583.3	525.7
B-SB	582.1	525.1
null	581.6	525.5



Assessment on count simulations

$n = 225$, $n_{\text{test}} = 2500$ on 50 by 50 grid

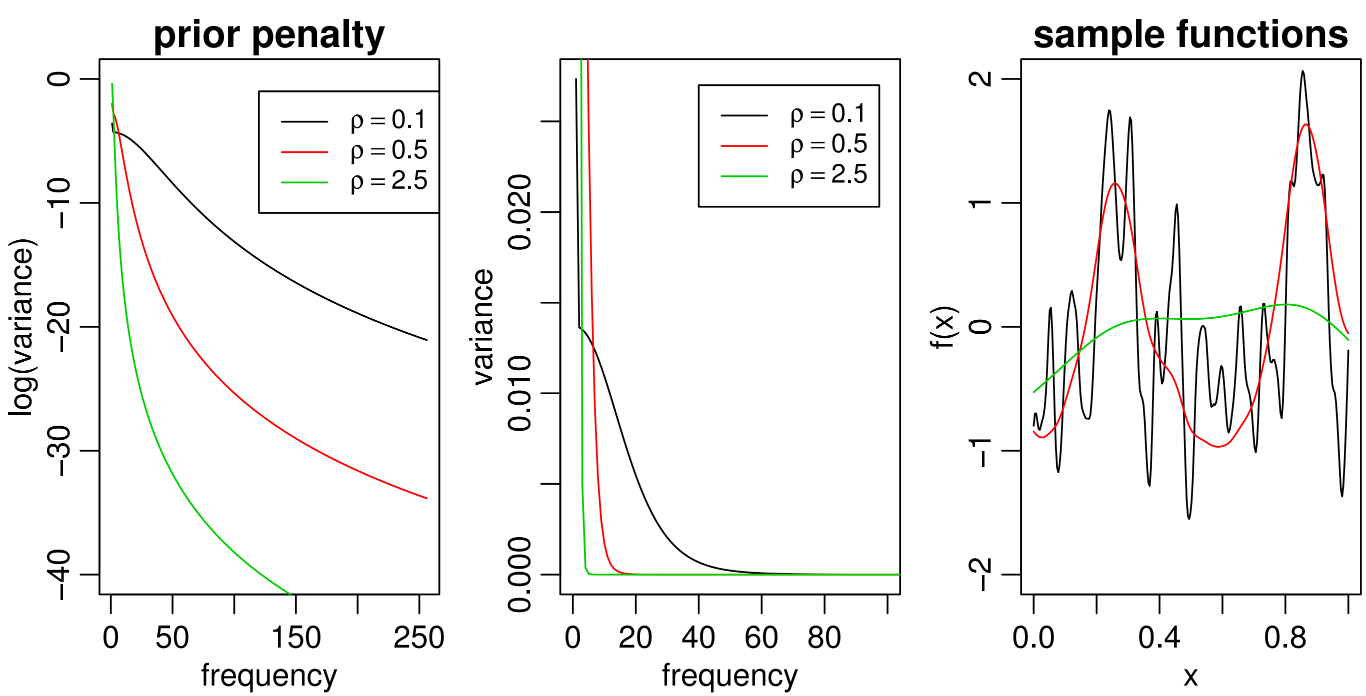


Penalization in the spectral approach

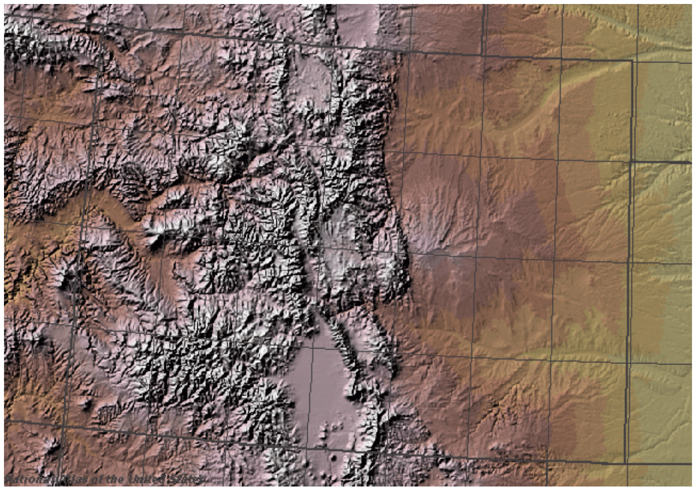
- GP representation zeroes out high-frequency coefficients as appropriate

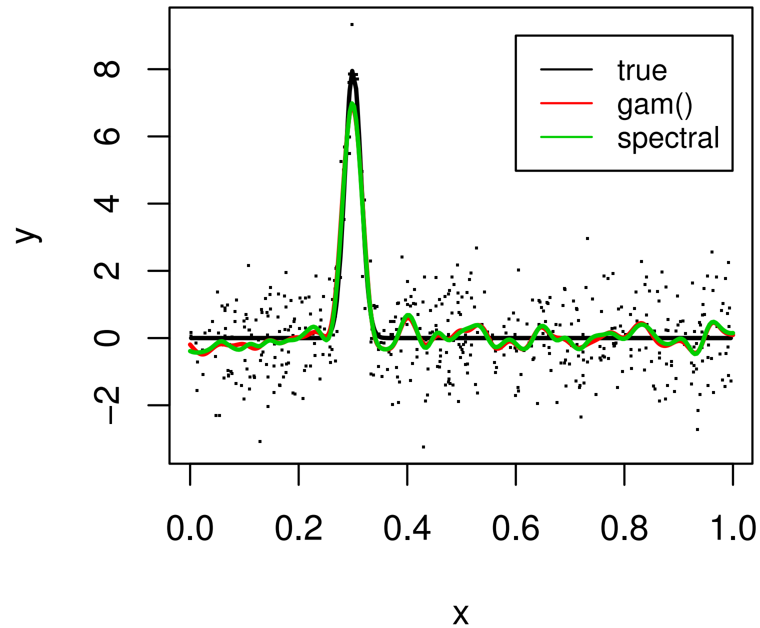
- Spatial hyperparameter controls coefficient variances

$$g(\cdot) \sim \text{GP}(\mu(\cdot), \sigma^2 R(\cdot, \cdot; \rho, \nu))$$



Heterogeneous penalties





- spatially-varying penalties are one option (e.g., Lang & Brezger 2004; Crainiceanu et al. 2004)
- spatially-varying ρ in a GP context is another

Outline

- Motivating examples
- Introduction to Gaussian processes (GPs)
- Fast Gaussian process modelling
- Flexible Gaussian process modelling
- Why Bayes works for smoothing
- The future: flexibility + efficiency + hierarchical modelling

A nonstationary covariance

- Higdon, Swall, and Kern (1999) model:

$$C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^p} k_{\boldsymbol{x_i}}(\boldsymbol{u}) k_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u}$$

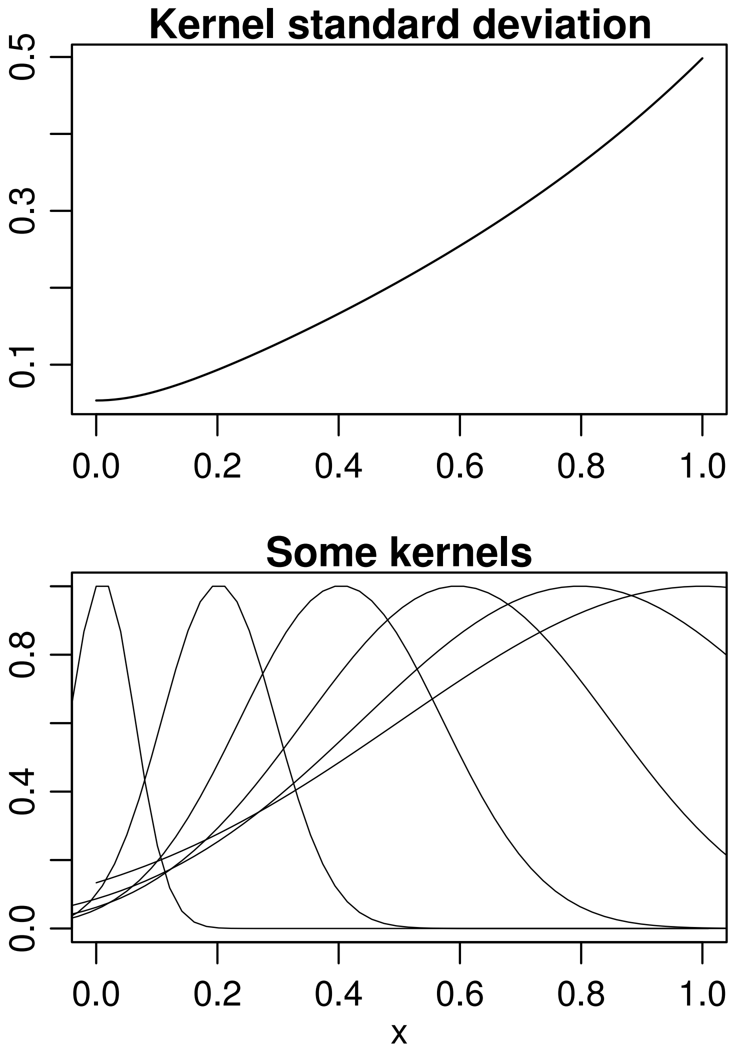
- guaranteed positive definite

- Gaussian kernels give closed form:

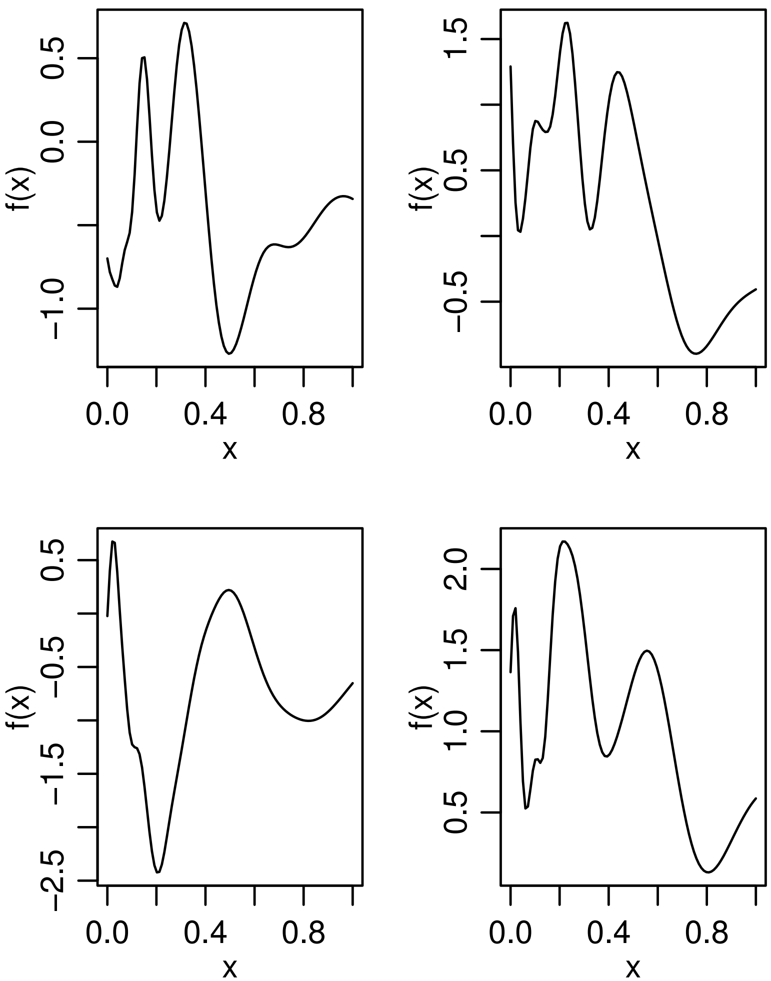
$$\begin{aligned} k_{\boldsymbol{x_i}}(\boldsymbol{u}) &\propto \exp\left(-(\boldsymbol{u} - \boldsymbol{x_i})^T \Sigma_i^{-1} (\boldsymbol{u} - \boldsymbol{x_i})\right) \\ R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) &= c_{ij} \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right) \end{aligned}$$

- $g(\cdot) \sim \mathrm{GP}(\mu, \sigma^2 R^{NS}(\cdot, \cdot; \Sigma(\cdot)))$

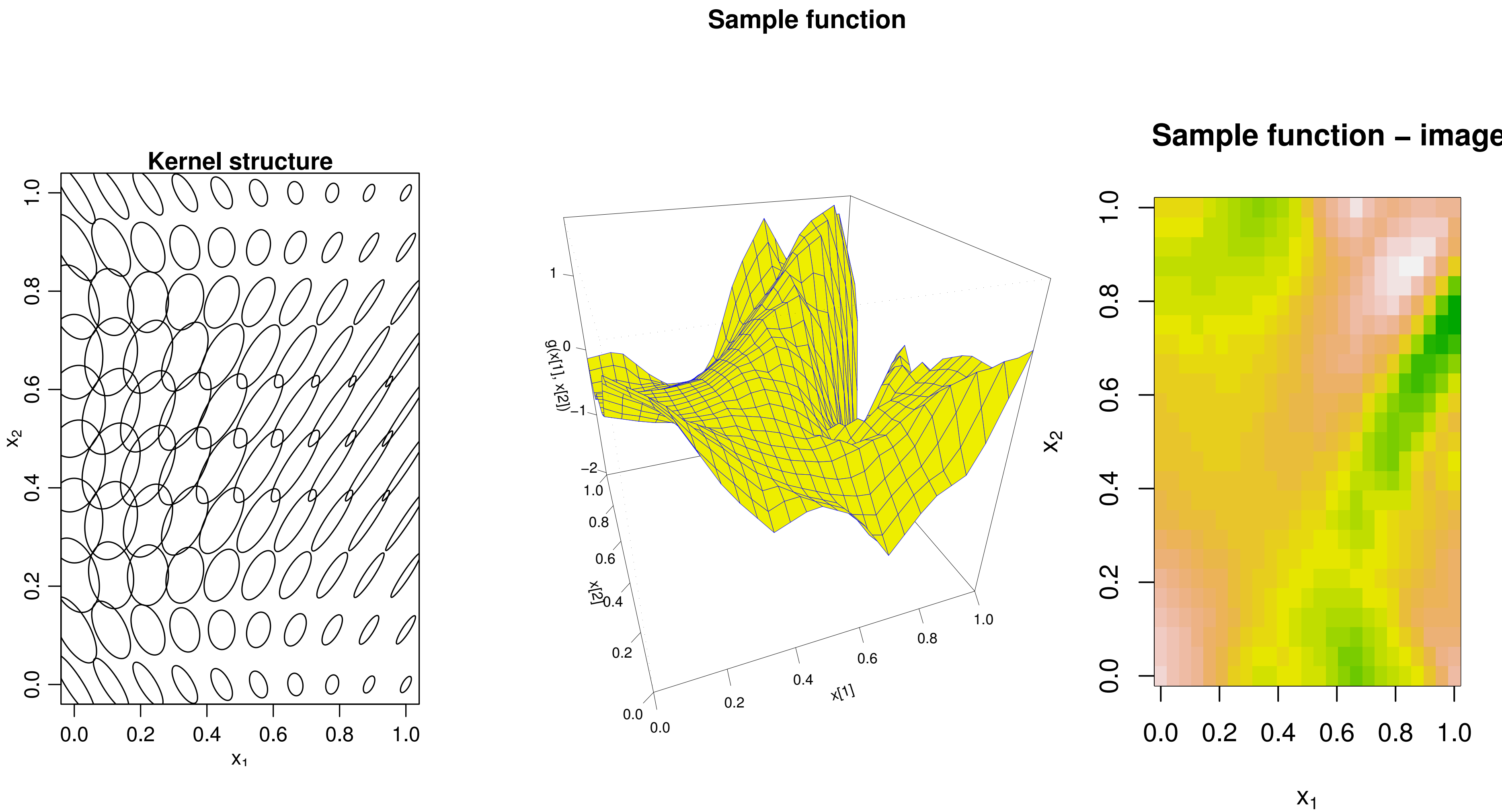
Nonstationary GPs in 1-D



Some sample functions



Nonstationary GPs in 2-D



Generalizing the kernel convolution approach

- Squared exponential form:

$$\exp\left(-\left(\frac{\tau_{ij}}{\rho}\right)^2\right) \Rightarrow c_{ij}\exp\left(-(\boldsymbol{x_i}-\boldsymbol{x_j})^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})\right)$$

  infinitely-differentiable sample paths

- 'Distance measures':

isotropy	$\tau_{ij}^2 = (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j})$
anisotropy	$\tau_{ij}^{*2} = (\boldsymbol{x_i}-\boldsymbol{x_j})^T\Sigma^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})$
nonstationarity	$Q_{ij} = (\boldsymbol{x_i}-\boldsymbol{x_j})^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})$

- Can we replace τ_{ij}^2 with Q_{ij} in other stationary correlation functions?

A class of nonstationary covariance functions

- Theorem 1: if $R(\tau)$ is positive definite for $\Re^P,\ P = 1, 2, \ldots,$ then

$$R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = c_{ij} R(\sqrt{Q_{ij}})$$

  is positive definite for $\Re^P,\ P = 1, 2, \ldots$

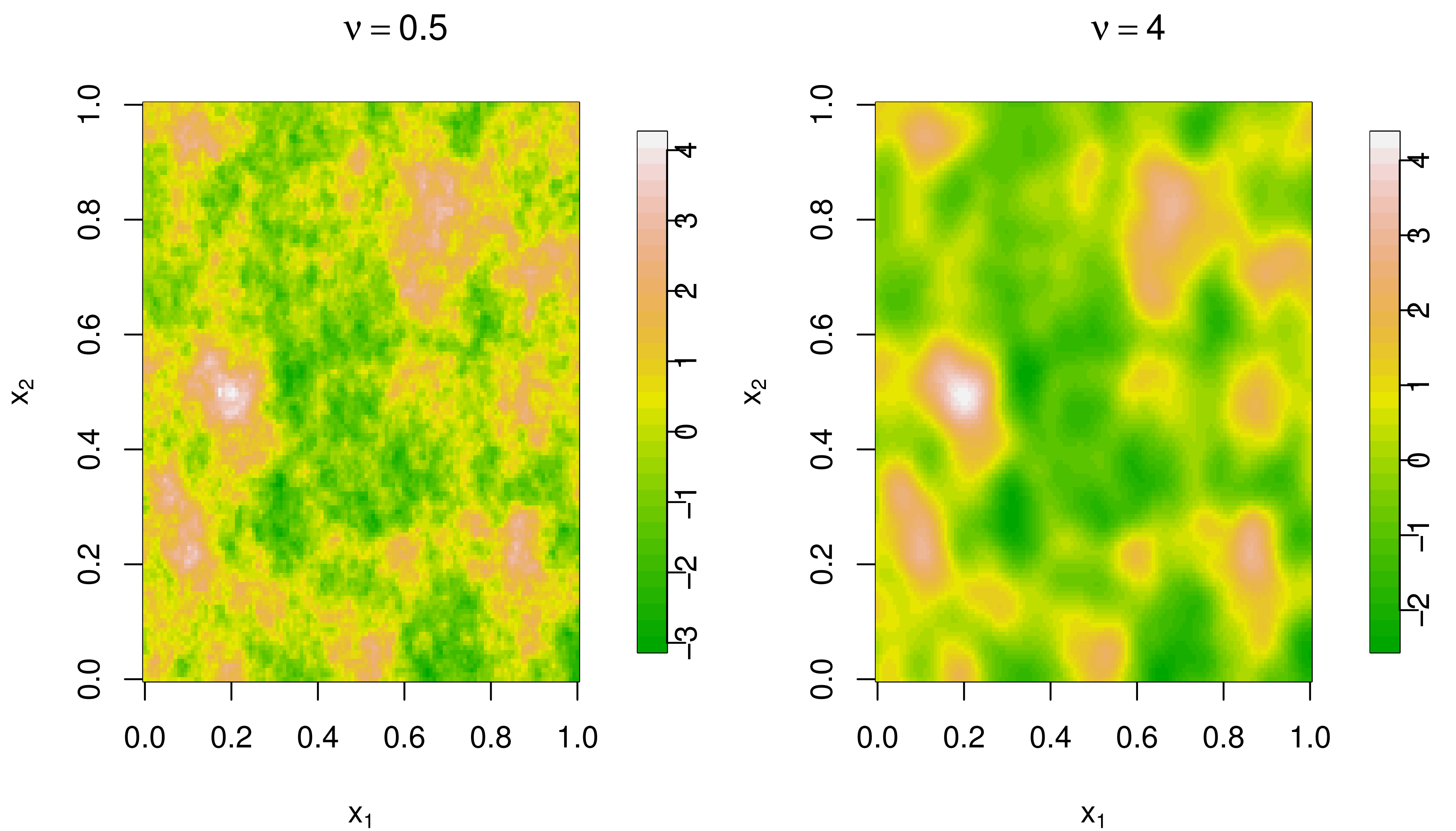
- Theorem 2: smoothness (differentiability) properties of the original stationary correlation retained

- Specific case of Matérn nonstationary covariance:

$$\frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}\tau}{\rho}\right) \Rightarrow \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu Q_{ij}}\right)^{\nu} K_{\nu}\left(2\sqrt{\nu Q_{ij}}\right)$$

  – advantages: more flexible form, differentiability not constrained, possible asymptotic advantages

Exponential and Matérn sample functions (stationary)



A basic Bayesian nonstationary spatial model

- Bayesian nonstationary kriging model

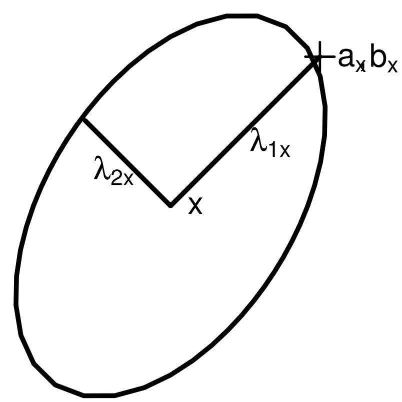
$$\begin{aligned} Y_i &\sim \mathrm{N}(g(\boldsymbol{x_i}), \eta^2),\ \boldsymbol{x_i} \in \Re^2 \\ g(\cdot) &\sim \mathrm{GP}(\mu, \sigma^2 R^{NS}(\cdot,\cdot;\nu,\Sigma(\cdot)) \end{aligned}$$

- Let R^{NS} be the nonstationary Matérn correlation
- Kernels ($\Sigma_{\boldsymbol{x}}$) constructed based on stationary GP priors
  - define multiple kernel matrices, $\Sigma_{\boldsymbol{x}}$, $\boldsymbol{x} \in \mathcal{X}$
  - smoothly-varying (element-wise) in domain
  - positive definite
- Fit via MCMC, including parameters determining $\Sigma(\cdot)$

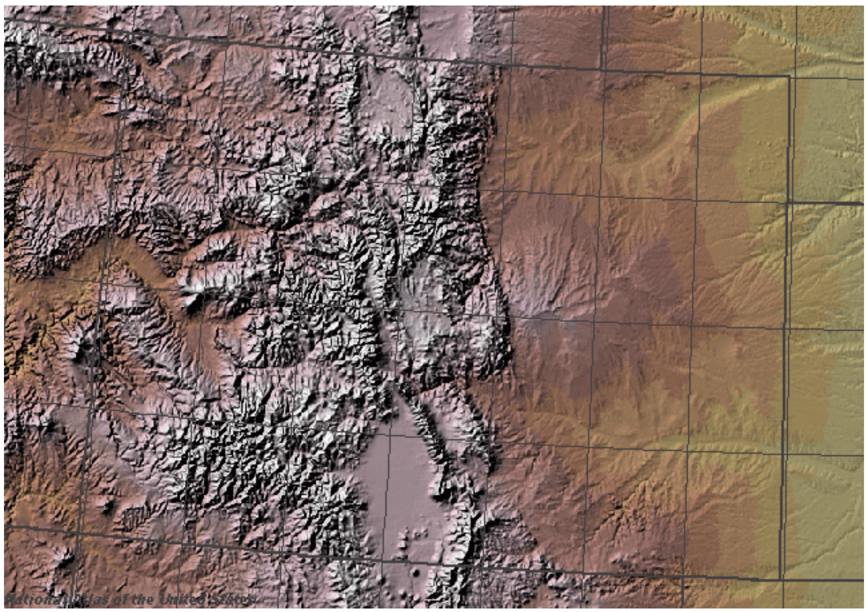
Smoothly-varying kernel matrices

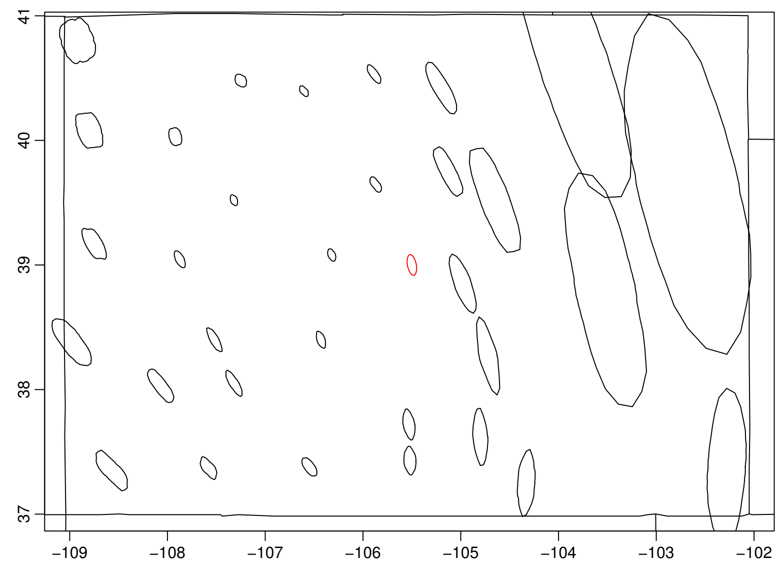
Spectral decomposition for each $\Sigma_{\boldsymbol{x}} = \Gamma_{\boldsymbol{x}}^T \Lambda_{\boldsymbol{x}} \Gamma_{\boldsymbol{x}}$

- in $\Re^2$, parameterize each kernel using unnormalized eigenvector coordinates $(a_{\boldsymbol{x}}, b_{\boldsymbol{x}})$ and the second eigenvalue $(\log \lambda_{2,\boldsymbol{x}})$
- define stationary GP priors for $\Phi(\cdot) \in \{(a(\cdot), b(\cdot), \log(\lambda_2(\cdot))\}$
- efficiently parameterize each GP using basis function approximation (Zhao & Wand, 2004)

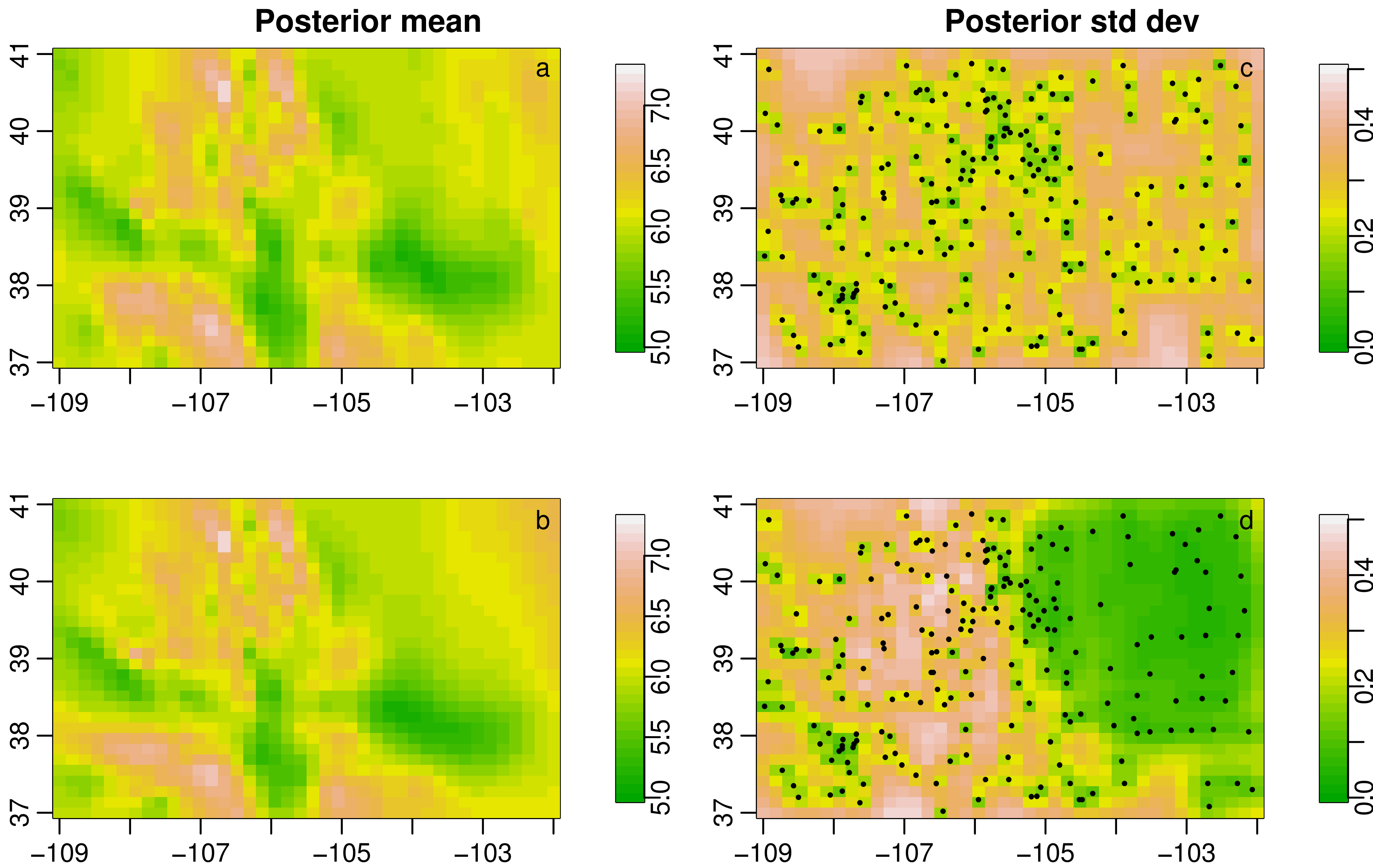


Colorado precipitation characterization



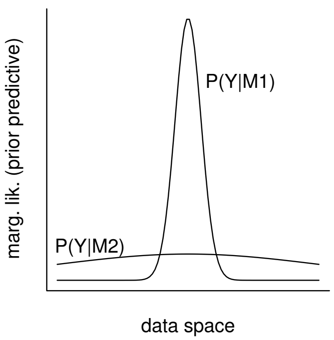


Estimating Colorado precipitation

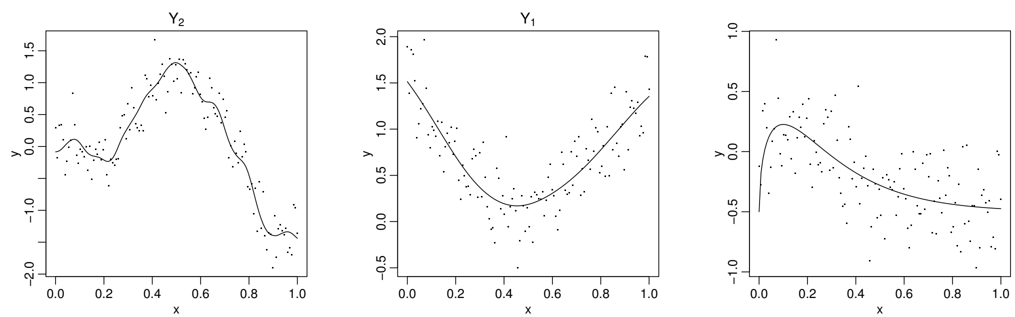


Why doesn't Bayes overfit?

- Fourier basis involves k^2 (=4096, e.g.) coefficients
- Nonstationary covariance involves very highly-parameterized covariance structure
- No direct penalty on complicated spatial functions



$P(y|M_i)$



Model 1 $\rho = 2.5$	-40.3	-18.9	-12.7
Model 2 $\rho = 0.5$	-27.5	-21.1	-16.4

What does a Bayesian approach give us?

- ability to create rich hierarchical models that reflect our understanding of the system
- in environmental health applications
  - the ability to incorporate time, latent variables, misaligned data
- a natural penalty on overfitting
- a recipe (perhaps slow) for estimation
- proper characterization of the uncertainty
- challenge lies less on the modelling side than with computations, model comparison and evaluation, and reproducibility

Future methodological work

- collaborative work on spatio-temporal modelling
  - computational approaches for applying existing methodological ideas to real health data

- GP computations and parameterization: flexibility + efficiency + hierarchical modelling
  - computational tricks for the nonstationary covariance; e.g., knot-based approaches for faster computation
  - use of a wavelet basis with irregular spatial data in a similar framework as the spectral basis

- combining deterministic and stochastic models, e.g. for air pollution

- useful, practical methods for designing spatial monitoring networks and determining power