Accounting for space in regression models with binary outcomes: A Bayesian spectral basis approach outperforms other methods

Chris Paciorek

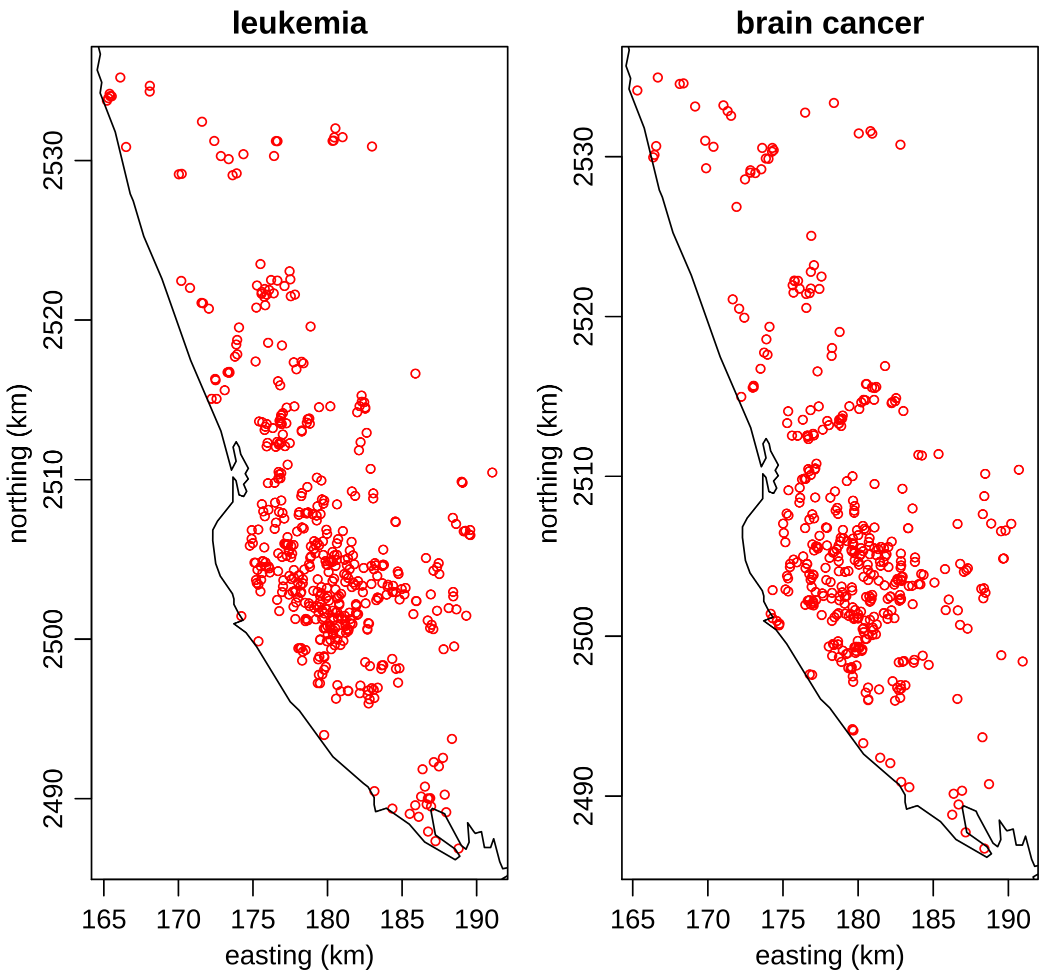
October 23, 2004

Department of Biostatistics

Harvard School of Public Health

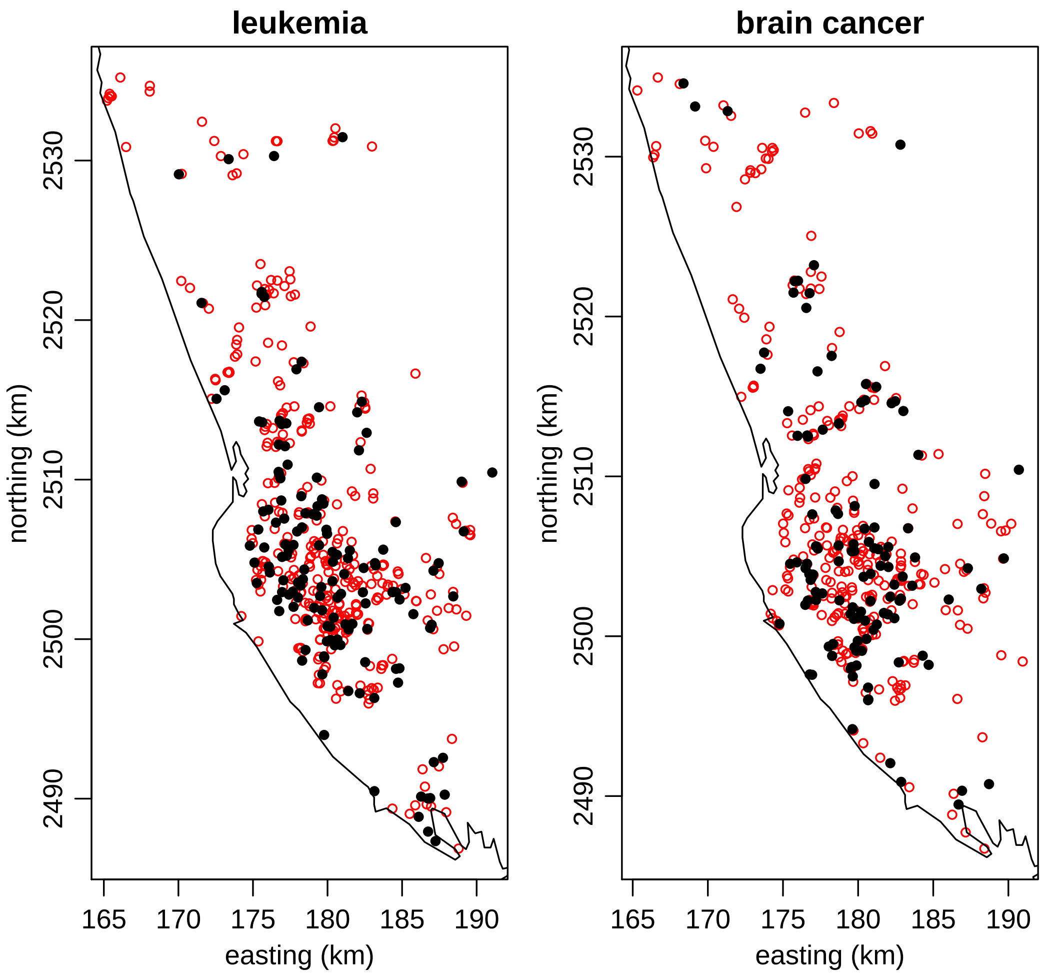
www.biostat.harvard.edu/~paciorek

Petrochemical exposure in Kaohsiung, Taiwan



$n = 495$ $n = 433$

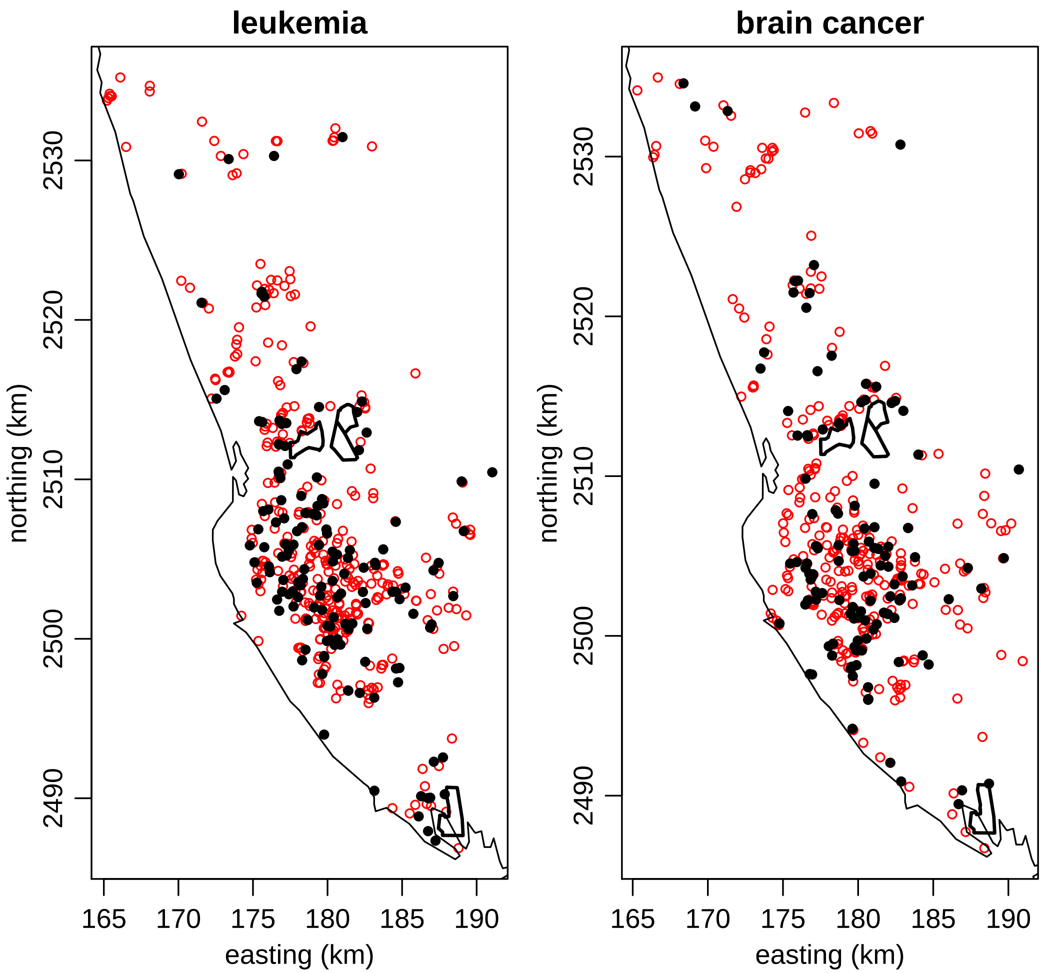
Petrochemical exposure in Kaohsiung, Taiwan



$$n = 495 \qquad n = 433$$

$$n_1 = 141 \qquad n_1 = 121$$

Petrochemical exposure in Kaohsiung, Taiwan



$n = 495$ $\qquad$ $n = 433$

$n_1 = 141$ $\qquad$ $n_1 = 121$

Possible approaches for health analysis

- Estimate exposure and use as covariate in health model
- Use distance to exposure source as covariate
- **Explicitly include space as a covariate**
  - **Map of risk - exploratory**
  - **Account for spatially-related unmeasured confounders**
  - **Test for spatial effect**

Outline

- Motivating example
- Generalized additive model and generalized mixed model approaches
- Difficulties in fitting regression for non-normal outcomes with 2-d smooth terms
- Parameterizations and fitting methods
- Simulations
  - binary responses
  - Poisson responses
- Revisit the example
- Goals for computational environmetrics

Goals for Computational Environmetrics

- reproducibility and ease of implementation, particularly for Bayesian methods
- modularity
- comparison and evaluation of models and fitting methods

GAM and GLMM frameworks

- basic model

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

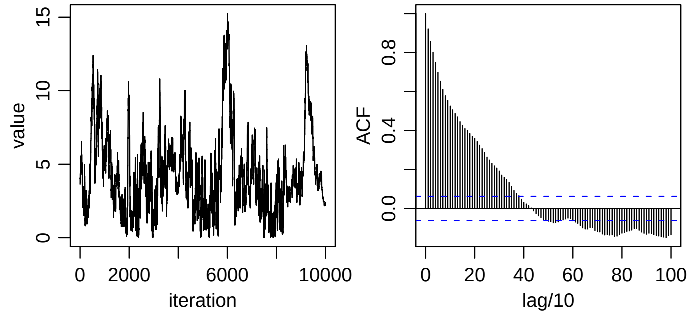
- basic spatial model for $\boldsymbol{g}_\theta^s = (g_\theta(\boldsymbol{s_1}), \ldots, g_\theta(\boldsymbol{s_n}))$
  - GAM: $g_\theta(\cdot)$ is a two-dimensional smooth term
    - basis representation, $\boldsymbol{g}_\theta^s = Z\boldsymbol{u}$
    - Gaussian process representation:
      $g(\cdot) \sim \text{GP}(\mu(\cdot), C_\theta(\cdot, \cdot)) \Rightarrow \boldsymbol{g}_\theta^s \sim N(\boldsymbol{\mu}, C_\theta)$
  - GLMM: $g_\theta(\boldsymbol{s_i}) = \boldsymbol{z_i}^T \boldsymbol{u}$
    - correlated random effects, $\boldsymbol{u} \sim N(0, \Sigma)$

Difficulties: speed and mixing

- Gaussian responses: closed form marginal likelihood
  – estimate $\boldsymbol{\beta}, \theta$

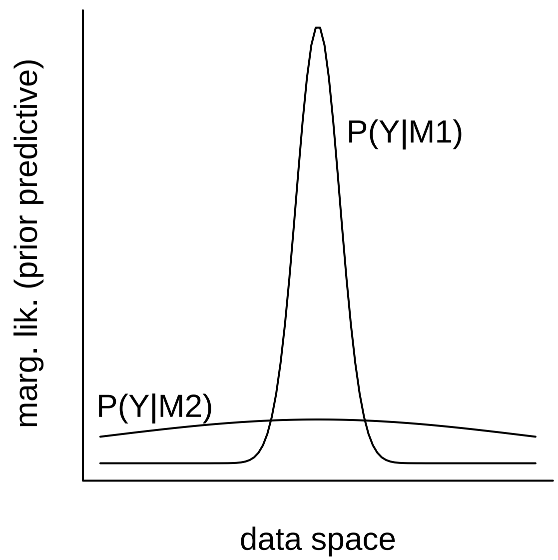
- non-Gaussian: no closed form $\Rightarrow$ high dimensional estimation
  – estimate $\boldsymbol{\beta}, \theta, \boldsymbol{u}$

- Challenges:
  - Classical mixed model: how approximate integral over random effects?
  - Bayesian methods: how perform large matrix calculations and avoid poor mixing?



Fitting approaches

- penalized likelihood, $l(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{g}^s_\theta) - \lambda J(\boldsymbol{u})$
  - fit by iterative weighted least squares
- Bayesian model for $(\boldsymbol{\beta}, \theta, \boldsymbol{u})$
  - fit by MCMC
  - implicit Bayesian penalty on complex spatial functions



Goals for implementations

- fast computations, avoiding large matrix calculations
- methods that scale reasonably with n
- reasonable fitting of simple risk surfaces we expect to model
- ease of implementation for applied work

Models and fitting methods considered

- penalized likelihood based on mixed model 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 geoadditive model-style radial basis functions fit by MCMC (Zhao and Wand 2004) [B-Geo]
- Bayesian spectral basis representation fit by MCMC using the FFT (Wikle 2002; Paciorek and Ryan, in prep.) [B-SB]
- Bayesian neural network model fit by MCMC (R. Neal) [B-NN]

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

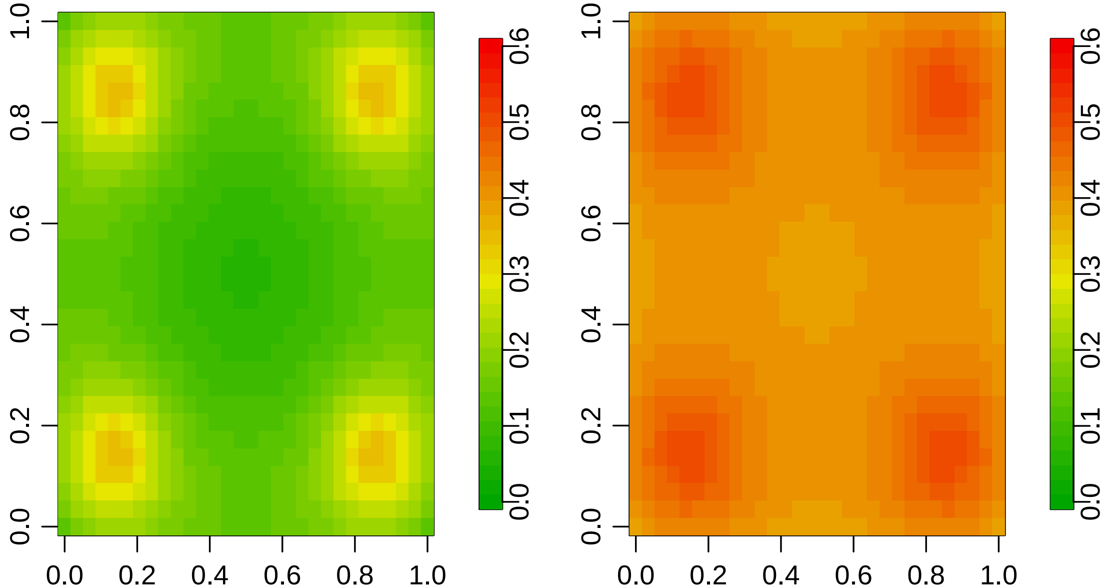
- potential covariance functions:
  - thin plate spline generalized covariance function, $C(\tau) = \tau^2 \log \tau$
  - Matérn correlation function, $R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho}\right) K_\nu \left(\frac{2\sqrt{\nu}\tau}{\rho}\right)$, with ρ and ν fixed

- 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 $\boldsymbol{u}$

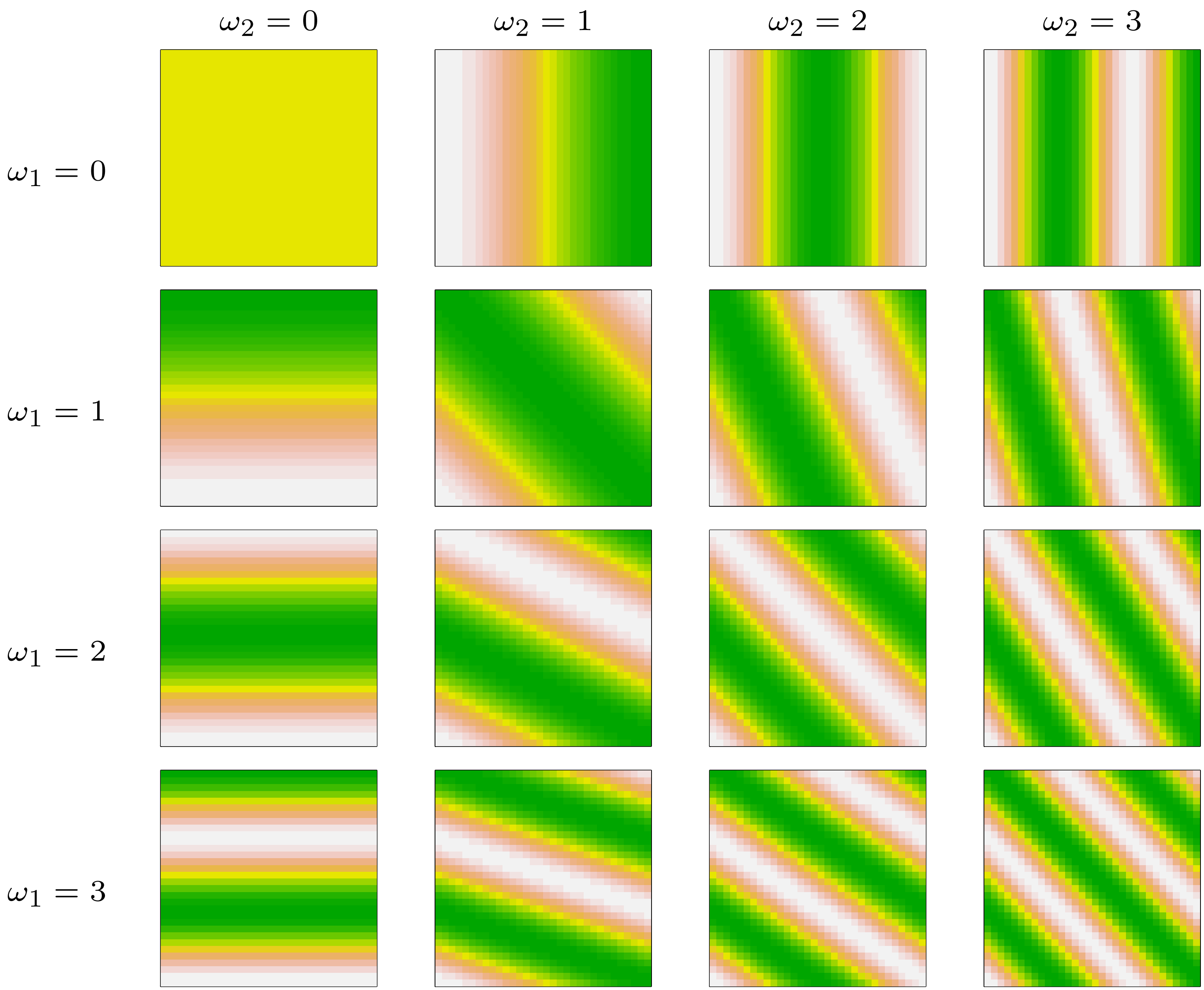
- thin plate spline covariance or Matérn correlation basis construction of Ψ and Ω

- 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

Bayesian spectral basis function model [B-SB]

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

Bayesian spectral basis functions

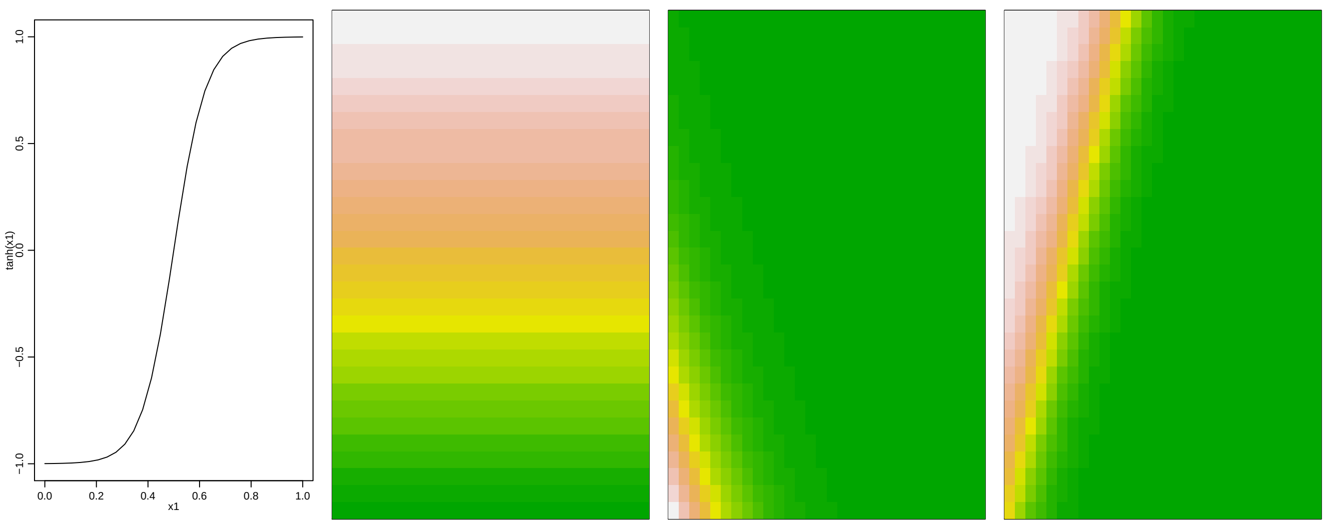


Comparison with usual GP specification

- usual GP model: $\boldsymbol{g}^s \sim N(\boldsymbol{\mu}, C_\theta)$
  - $O(n^3)$ fitting: $|C_\theta|$ and $C_\theta^{-1}\boldsymbol{g}$
- spectral basis uses FFT
  - $O(k^2)\log(k^2)$
  - additional observations are essentially free for a fixed grid
  - fast computation and prediction of surface given coefficients
  - a priori independent coefficients give fast computation of prior and help with mixing

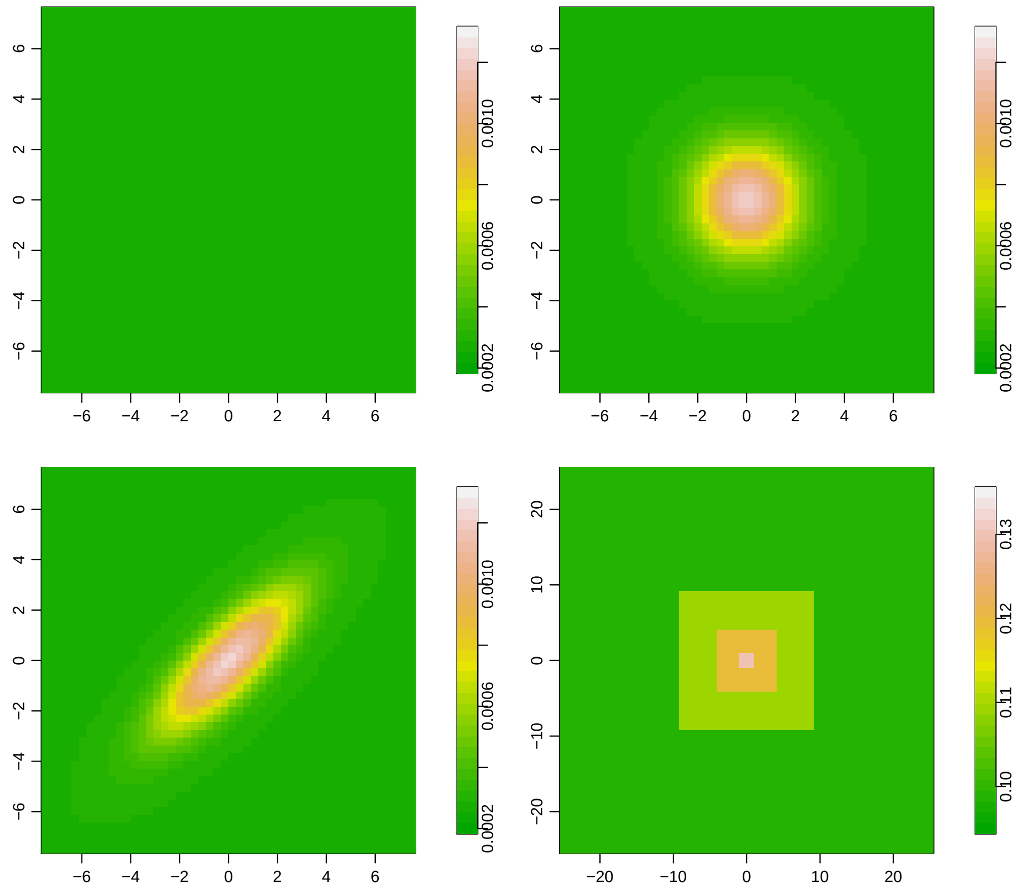
Bayesian neural network [B-NN]

- multilayer perceptron with one hidden layer gives basis representation:
  - $g(\boldsymbol{s_i}) = \sum_k \text{tanh}({\boldsymbol{\phi_k}}^T \boldsymbol{s_i}) u_k$
- position and orientation of basis functions change with $\boldsymbol{\phi_k}$
- implemented with software of R. Neal; somewhat complicated proposal scheme

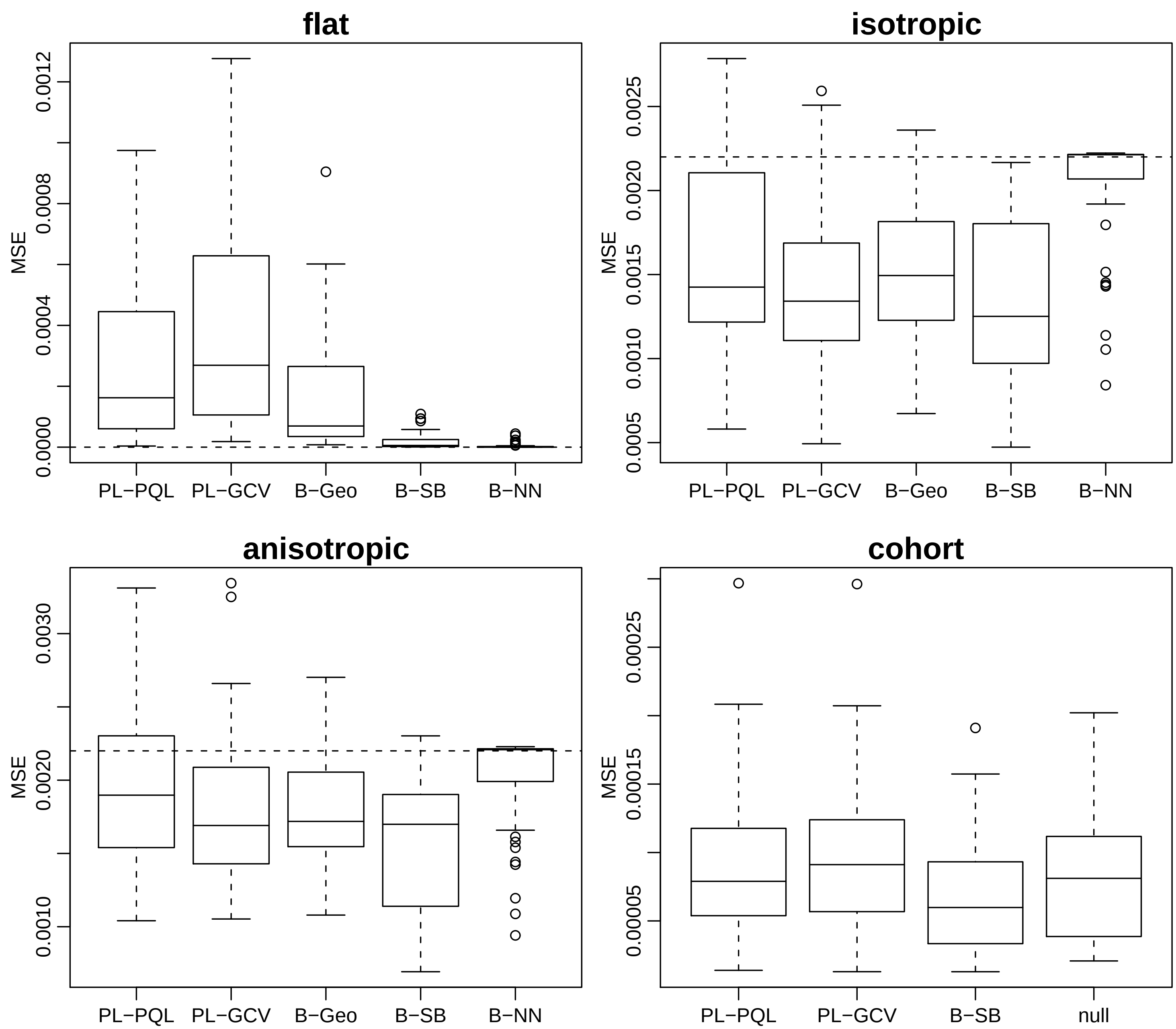


Simulated datasets

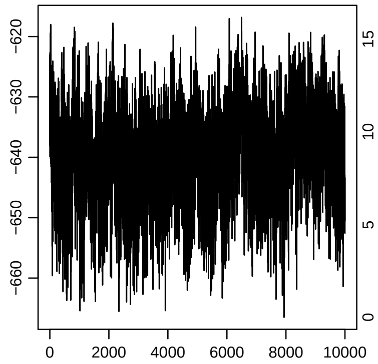
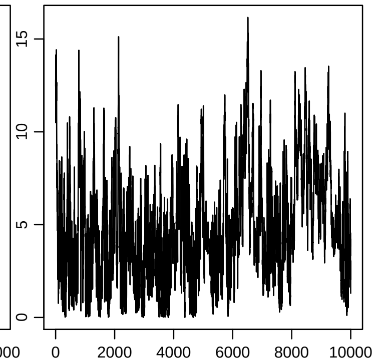
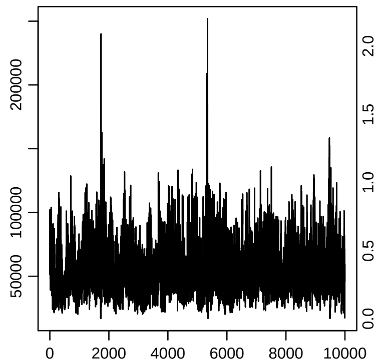
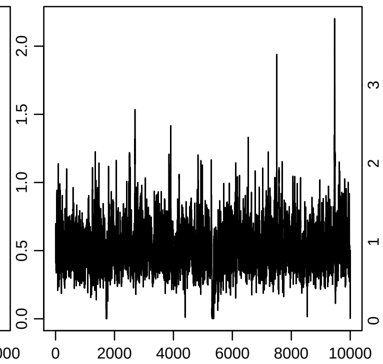
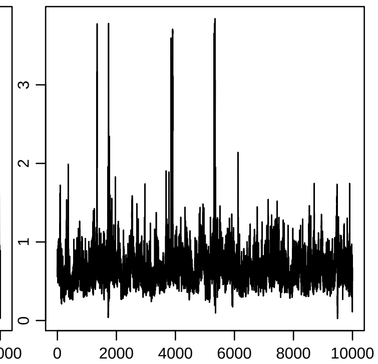
- 3 case-control scenarios: $n_0 = 1,000$; $n_1 = 200$; $n_{\text{test}} = 2500$ on 50 by 50 grid
- 1 cohort scenario: $n = 10,000$; $n_{\text{test}} = 2500$ on 50 by 50 grid



Assessment on 50 simulated datasets



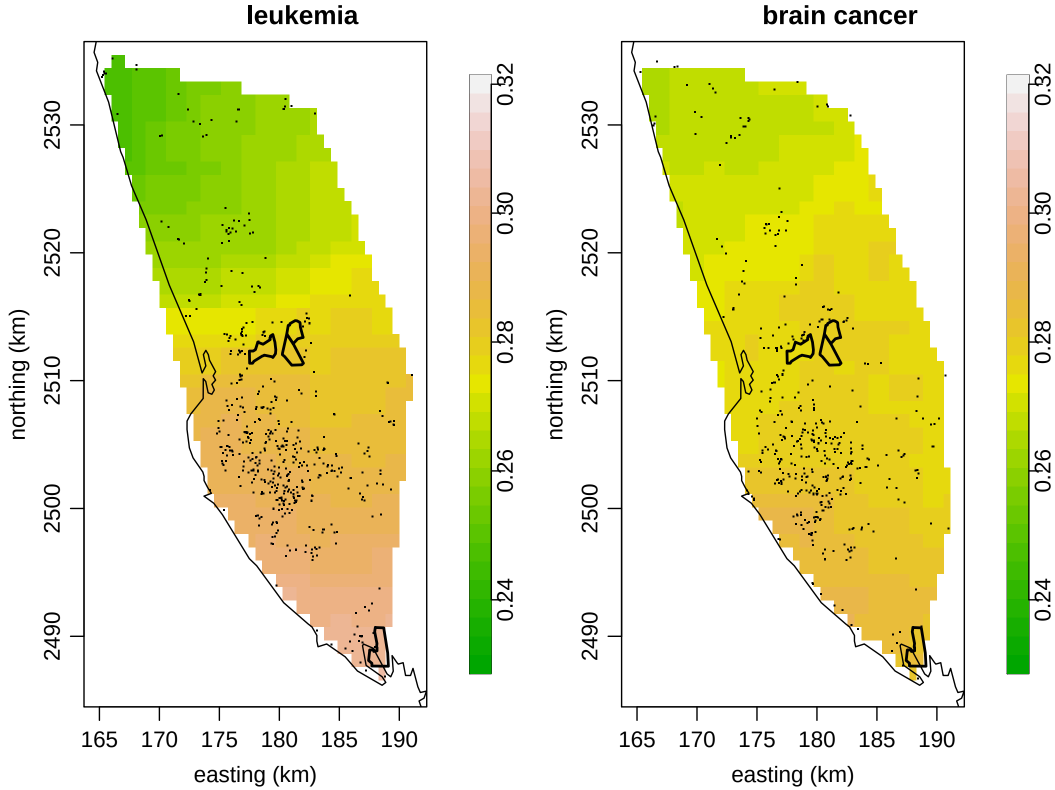
Mixing and speed of Bayesian methods

	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.			

Example revisited - assessment

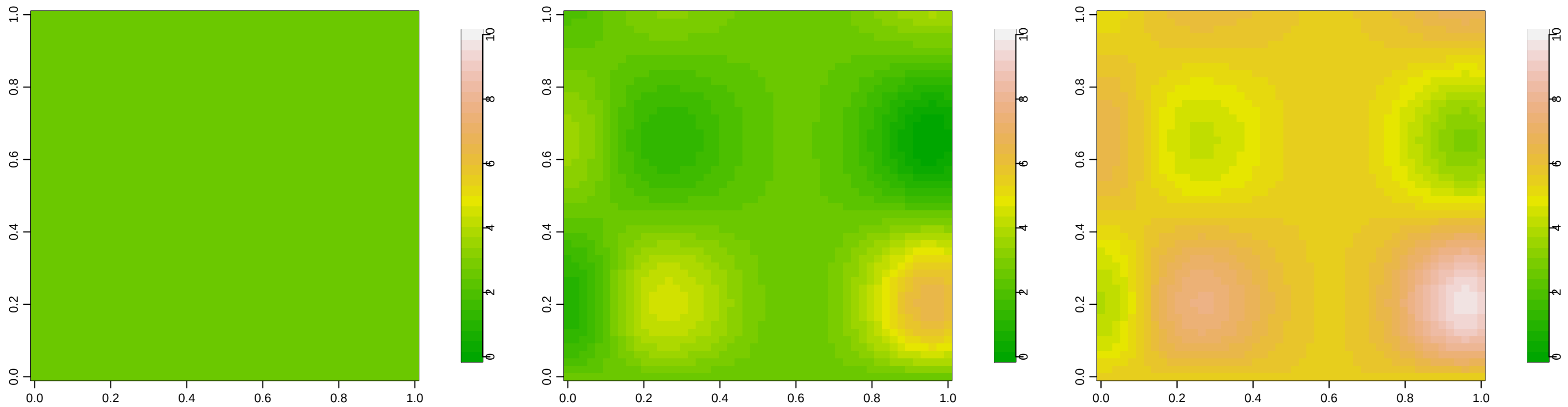
Summed test deviance over 10-fold C-V sets

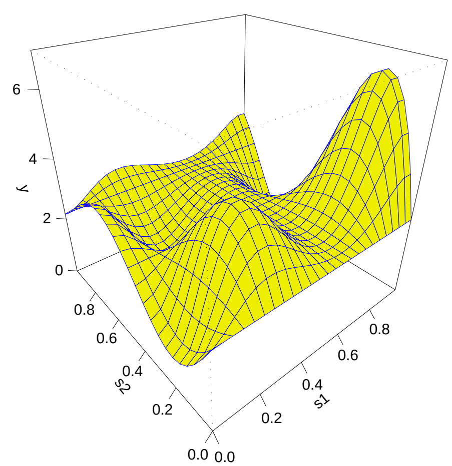
	leukemia	brain cancer
PL-GCV	590.1	529.8
PL-PQL	585.6	529.5
B-Geo	583.3	525.7
B-SB	582.1	525.1
null	581.6	525.5



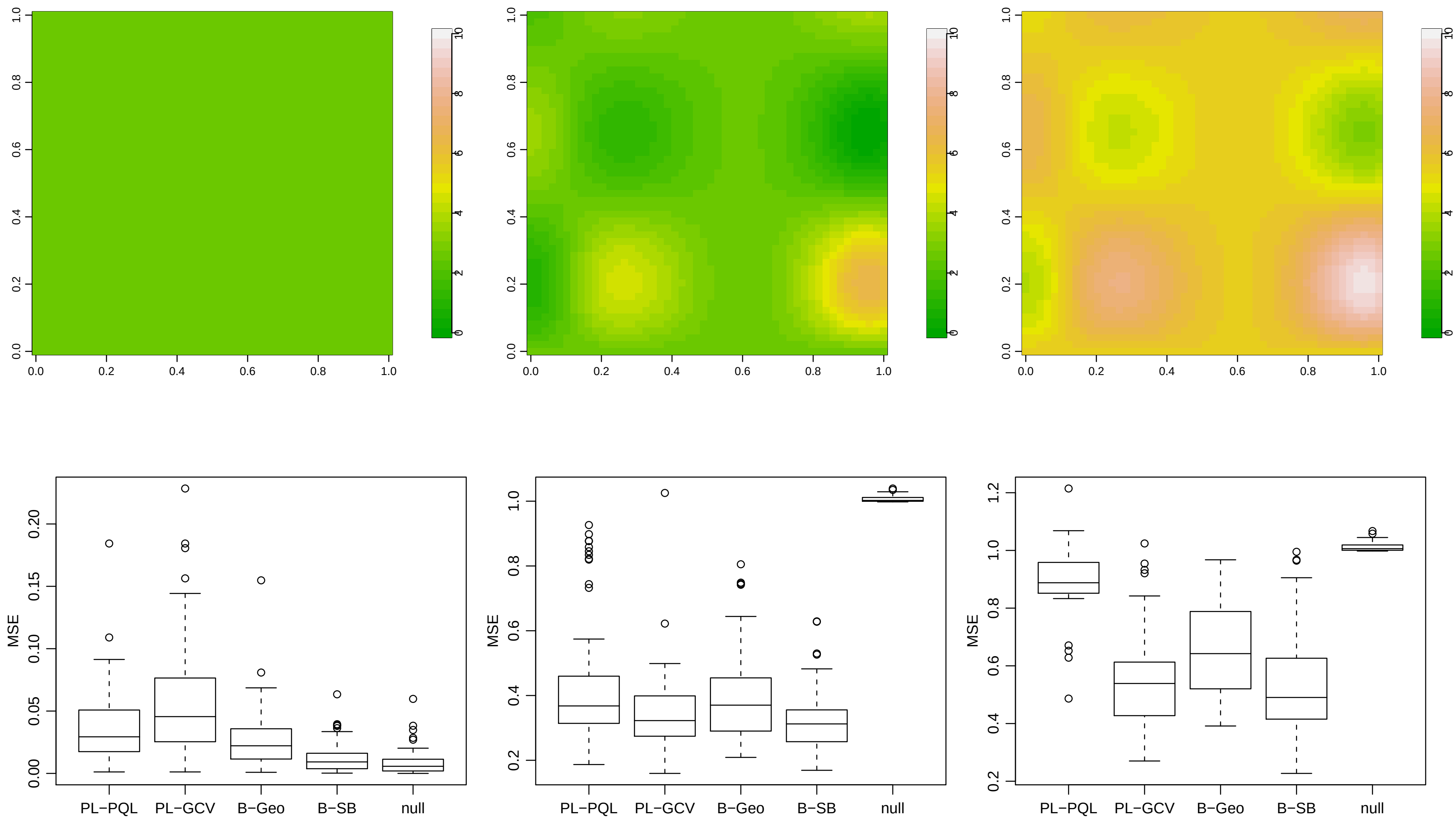
Simulated count data

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





Assessment on count simulations



Methodology conclusions

- Effective process parameterization allows for faster Bayesian estimation
  - effective for spatial models with thousands of observations
- Natural Bayesian complexity penalty works well; other automatic criteria appear to overfit
- R code for spectral basis model to ease implementation
- Power is an issue with binary observations
- Results hold for count data

Suggestions for computational environmetrics

- reproducibility
  - requires code and detailed description (supplemental material/web)
  - standard computing environment (R) helps
  - enabling reproducible MCMC (beyond BUGS)
    - class structures, templates, and proposal functions for R
- modularity
  - spectral basis as modular component in hierarchical models
- comparison of methods/models
  - rare
  - difficult without reproducibility, particularly with Bayesian methods

Future methodological work

- Importance of basis functions vs. speed/mixing in MCMC vs. penalty estimation method in determining fitting success
  - Why don't automatic criteria for penalized likelihood work as well?
  - Importance of fitting both variance and spatial range parameters - small-sample results (consider effective basis functions) vs. asymptotics (Zhang, 2004)
- Simple approaches for testing necessity of spatial term
- Other process parameterizations allowing fast Bayesian estimation:
  - Simple prior structures for wavelet basis coefficient (co)variances?