Nonstationary Gaussian Processes for Regression and Spatial Modelling

Committee:

Mark Schervish

Doug Nychka

(Geophysical Statistics Project, National Center for Atmospheric Research)

James Risbey

(Department of Mathematical Sciences, Monash University)

Valérie Ventura

Larry Wasserman

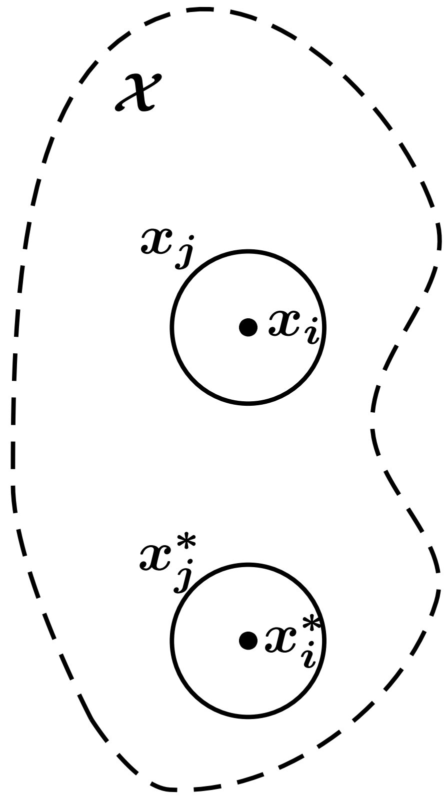
April 28, 2003

OUTLINE

- Gaussian process (GP) distribution
  - ❖ Stationary and nonstationary covariance models
- A Bayesian nonparametric regression model
  - ❖ Comparison with other adaptive smoothing methods
- Issues in fitting GP models
- A nonstationary model for replicated spatial data
- Future work

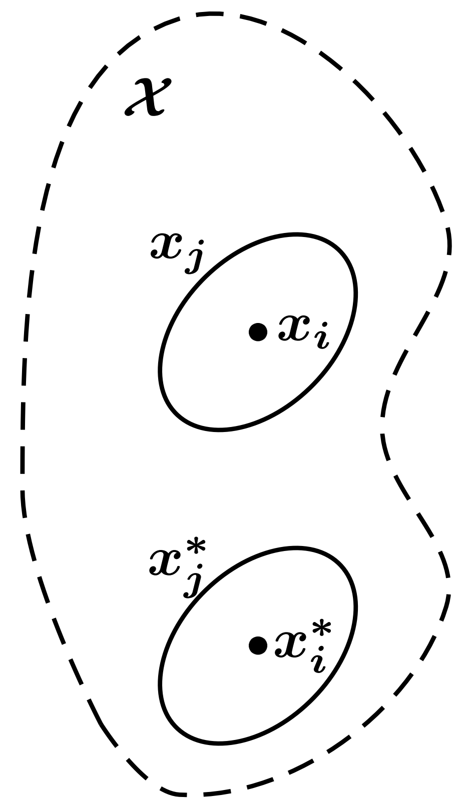
GAUSSIAN PROCESS DISTRIBUTION

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



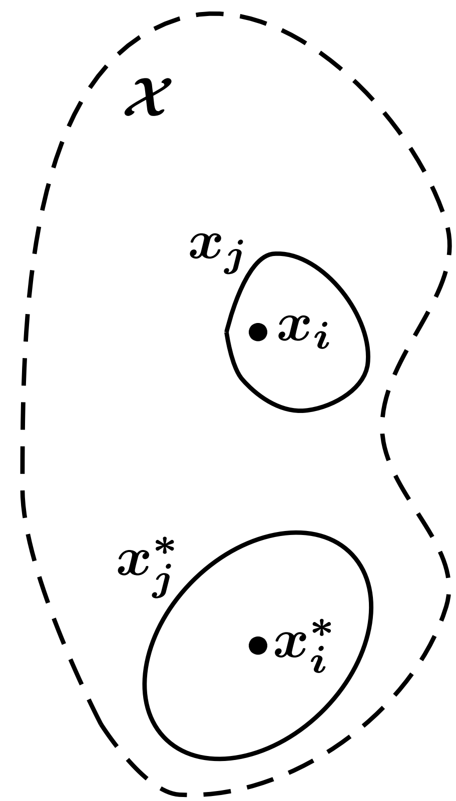
GAUSSIAN PROCESS DISTRIBUTION

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



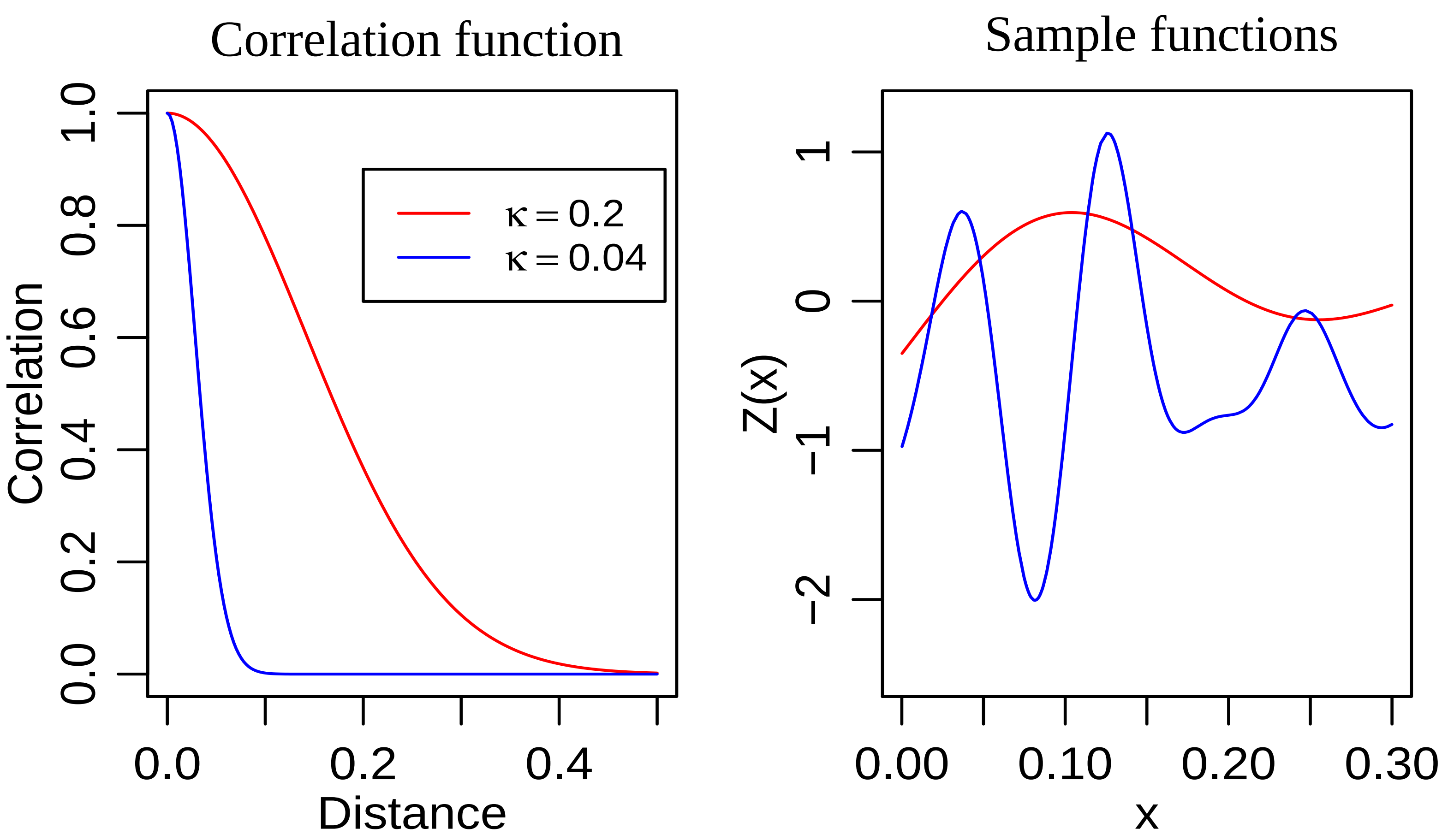
Gaussian process distribution

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

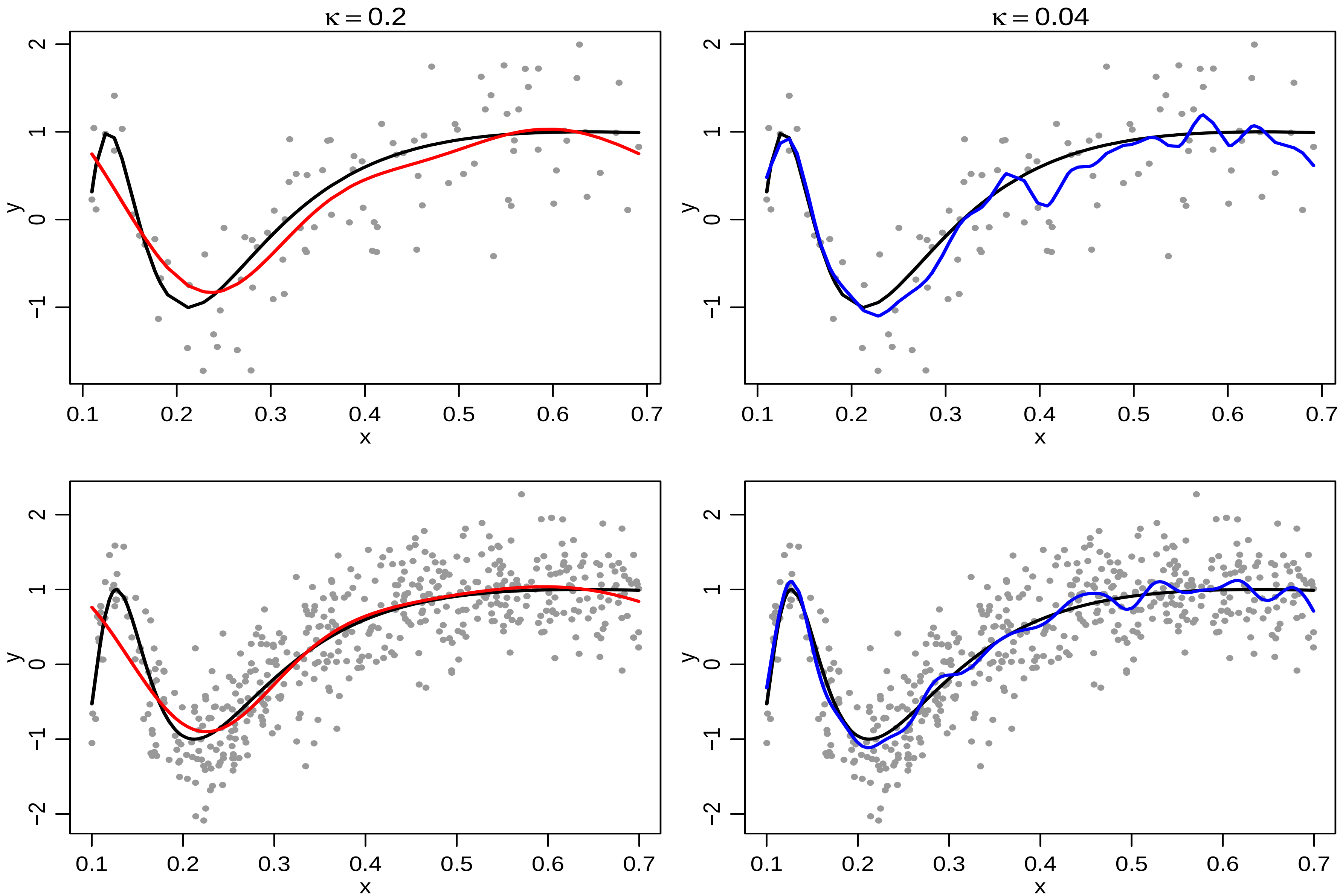


STATIONARY CORRELATION FUNCTIONS

Squared exponential: $\boldsymbol{R(\tau) = \exp\left(-\left(\frac{\tau}{\kappa}\right)^2\right)}$



Degree of Smoothing



NONSTATIONARY COVARIANCE

- Higdon, Swall, and Kern (1999)

  $$R^{NS}(x_i, x_j) = c \int_{\Re^P} k_{x_i}(u) k_{x_j}(u) du$$

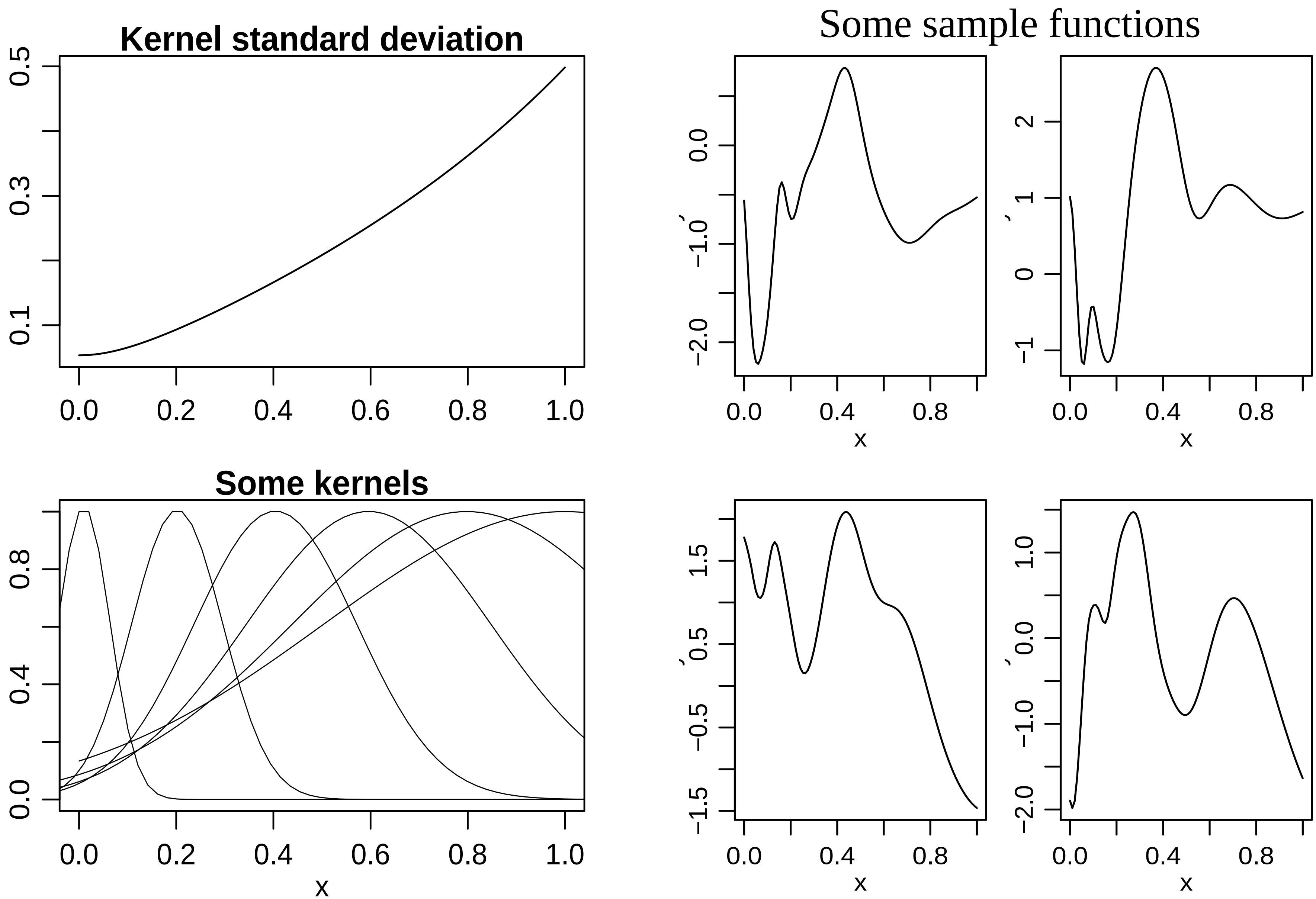
- Guaranteed positive definite

- Normal kernels:

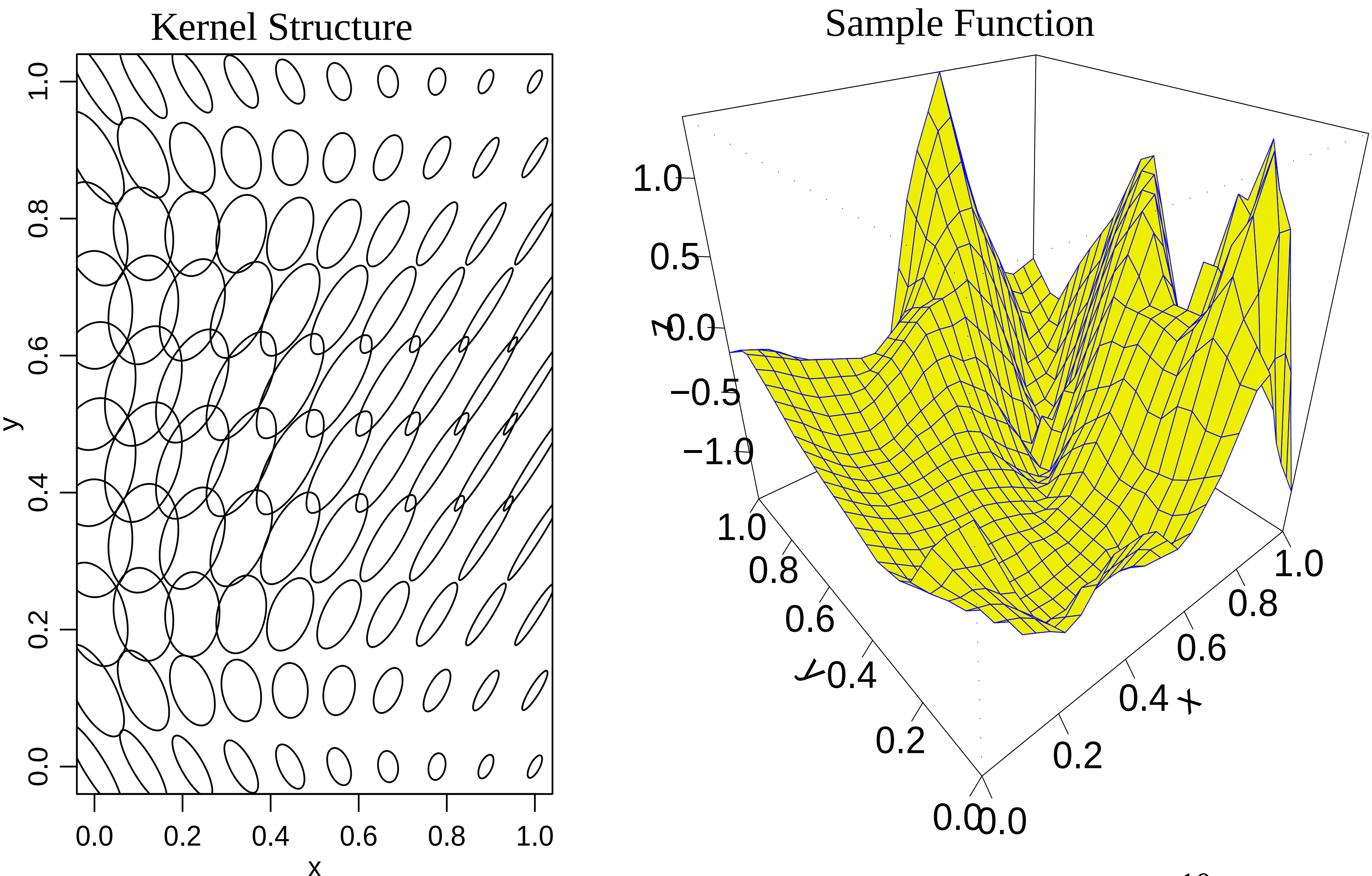
$$\begin{aligned} k_{x_i}(u) &\propto \exp\left(-(u - x_i)^T \Sigma_i^{-1} (u - x_i)\right) \\ R^{NS}(x_i, x_j) &= c \exp\left(-(x_i - x_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (x_i - x_j)\right) \end{aligned}$$

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

NONSTATIONARY GPS IN 1D



NONSTATIONARY GPS IN 2D



GENERALIZING THE KERNEL METHOD

- Squared exponential form:

$$\exp\left(-\left(\frac{\tau}{\kappa}\right)^2\right) \Rightarrow c\exp\left(-(x_i - x_j)^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(x_i - x_j)\right)$$

Infinitely-differentiable sample paths

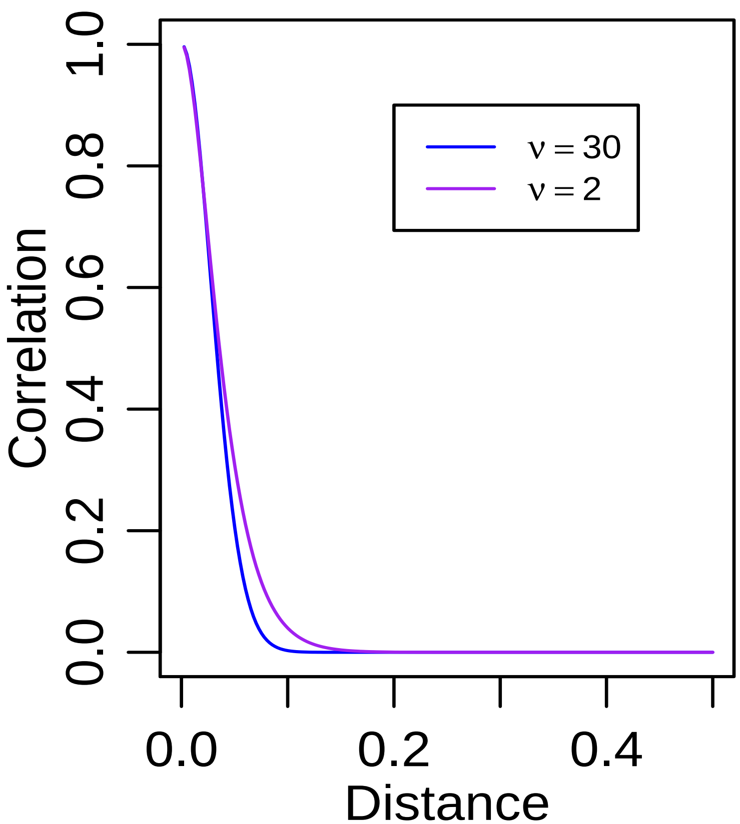
- 'Distance measures'

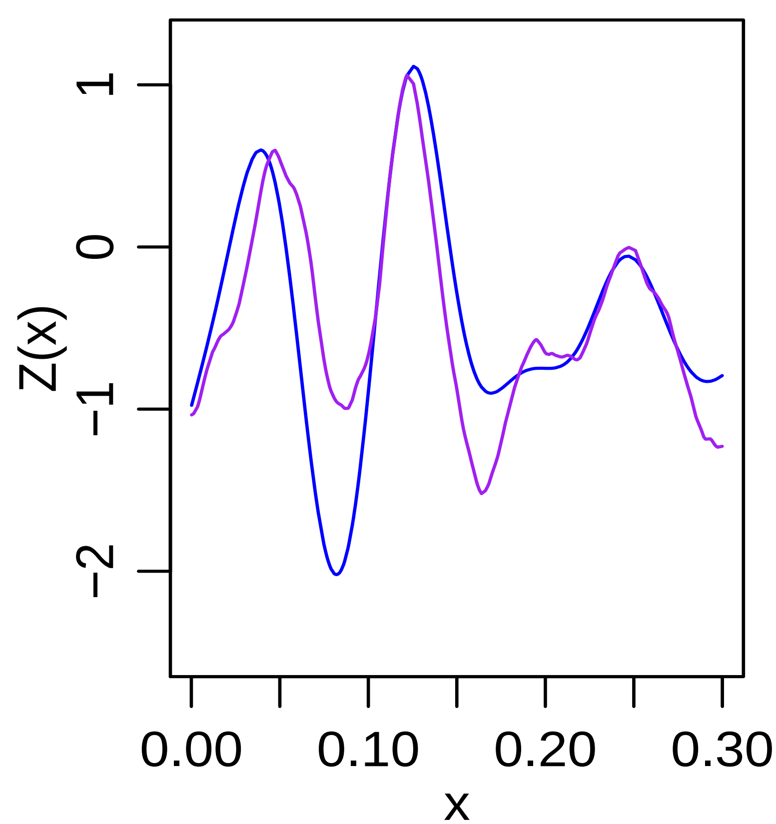
$$\begin{aligned} \tau^2_{x_i,x_j} &= (x_i - x_j)^T(x_i - x_j) \\ {\tau^*}^2_{x_i,x_j} &= (x_i - x_j)^T\Sigma^{-1}(x_i - x_j) \\ Q_{x_i,x_j} &= (x_i - x_j)^T\left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}(x_i - x_j) \end{aligned}$$

- Can we replace τ^2 with Q_{x_i,x_j} in other stationary correlation functions?

STATIONARY CORRELATION FUNCTIONS

Matérn form: $R(\tau) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu\tau^2}}{\kappa}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu\tau^2}}{\kappa}\right)$





- Differentiability controlled by ν, asymptotic advantages (Stein)
- Nonstationary form is positive definite

GENERALIZED KERNEL METHOD

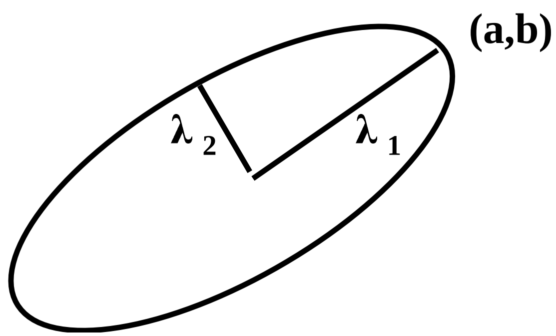
- Theorem : if $\boldsymbol{R(\tau)}$ is positive definite for $\Re^p, p = 1, 2, \ldots$, then

$$R^{NS}(x_i, x_j) = \frac{|\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|^{\frac{1}{2}}} R\left(\sqrt{Q_{x_i, x_j}}\right)$$

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

- Summary of theorems on smoothness properties of sample paths:
  - ❖ Based on original stationary correlation function
  - ❖ Provided kernel matrices vary sufficiently smoothly in covariate space

SMOOTHLY-VARYING KERNEL MATRICES



- Goals:
  - ❖ Define multiple kernel matrices, $\boldsymbol{\Sigma_x}$
  - ❖ Smoothly-varying in covariate space
  - ❖ Positive definite
- Use spectral decomposition ($\boldsymbol{\Sigma_x = \Gamma_x^T \Lambda_x \Gamma_x}$)
  - ❖ $\boldsymbol{\Gamma_x}$ parameterized as first eigenvector plus successive orthogonal vectors in reduced-dimension subspaces
  - ❖ stationary GP priors on unnormalized eigenvector coordinates $(\boldsymbol{a_x}, \boldsymbol{b_x})$ and on logarithm of eigenvalues $(\boldsymbol{\lambda_{x,1}}, \boldsymbol{\lambda_{x,2}})$
  - ❖ gets unwieldy and highly-parameterized for large $\boldsymbol{P}$

(MULTIVARIATE) NONPARAMETRIC REGRESSION MODEL

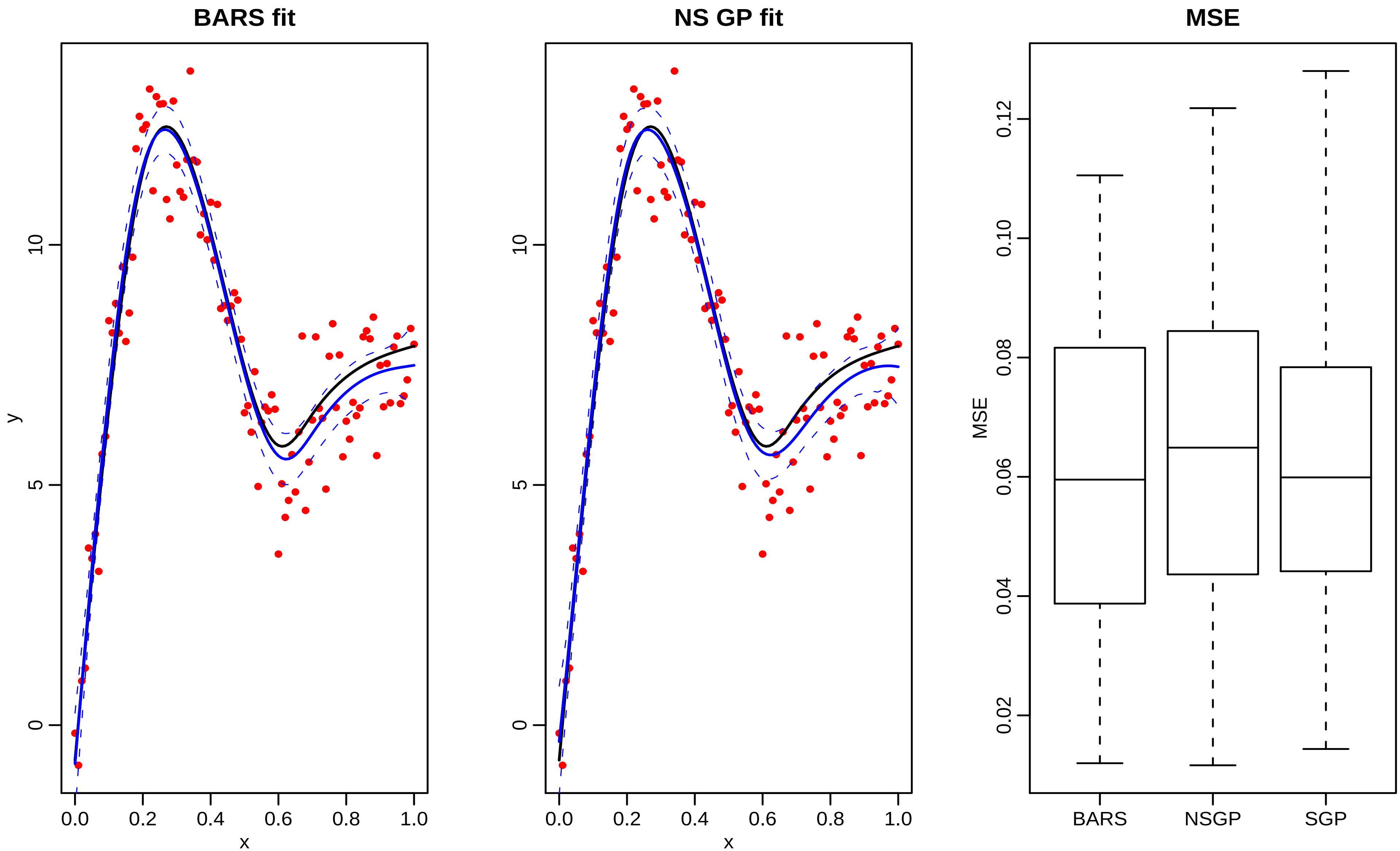
- Bayesian model

$$\begin{aligned} Y_i &\sim N(f(x_i), \eta^2),\ x_i \in \Re^P \\ f(\cdot) &\sim \text{GP}(\mu, \sigma^2 R^{NS}(\cdot, \cdot; \nu, \theta)) \end{aligned}$$

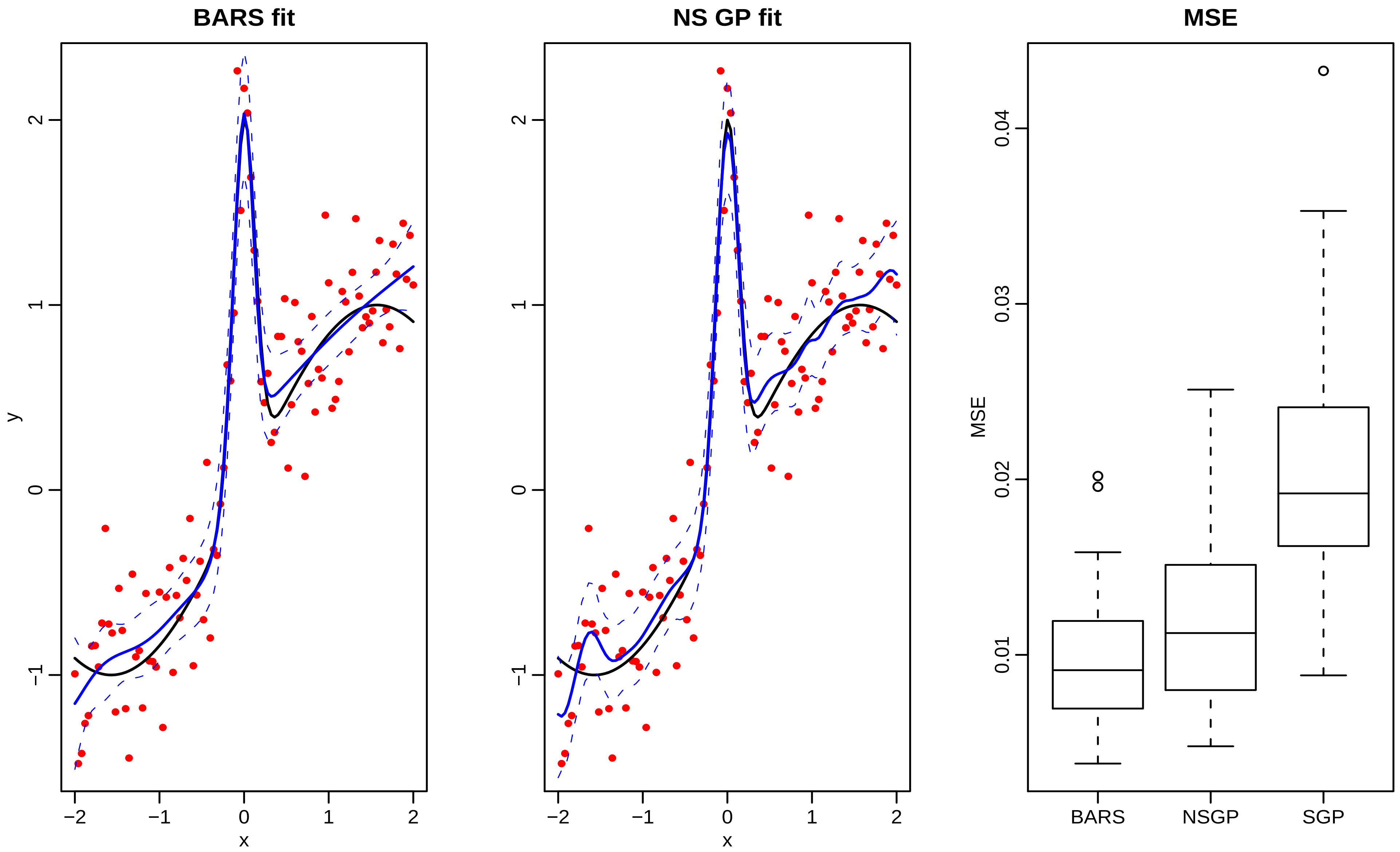
  - ❖ Let R^{NS} be the nonstationary version of the Matérn
  - ❖ Kernel parameters (θ) with stationary GP priors

- Compare performance to Bayesian models in which f is a spline
  - ❖ $x_i \in \Re^1$: BARS (DiMatteo, Genovese & Kass 2002)
  - ❖ $x_i \in \Re^P, P > 1$:
    - ✥ BMARS (Denison, Mallick & Smith 1998) - tensor products of univariate splines
    - ✥ BMLS (Holmes & Mallick 2001) - multivariate linear splines

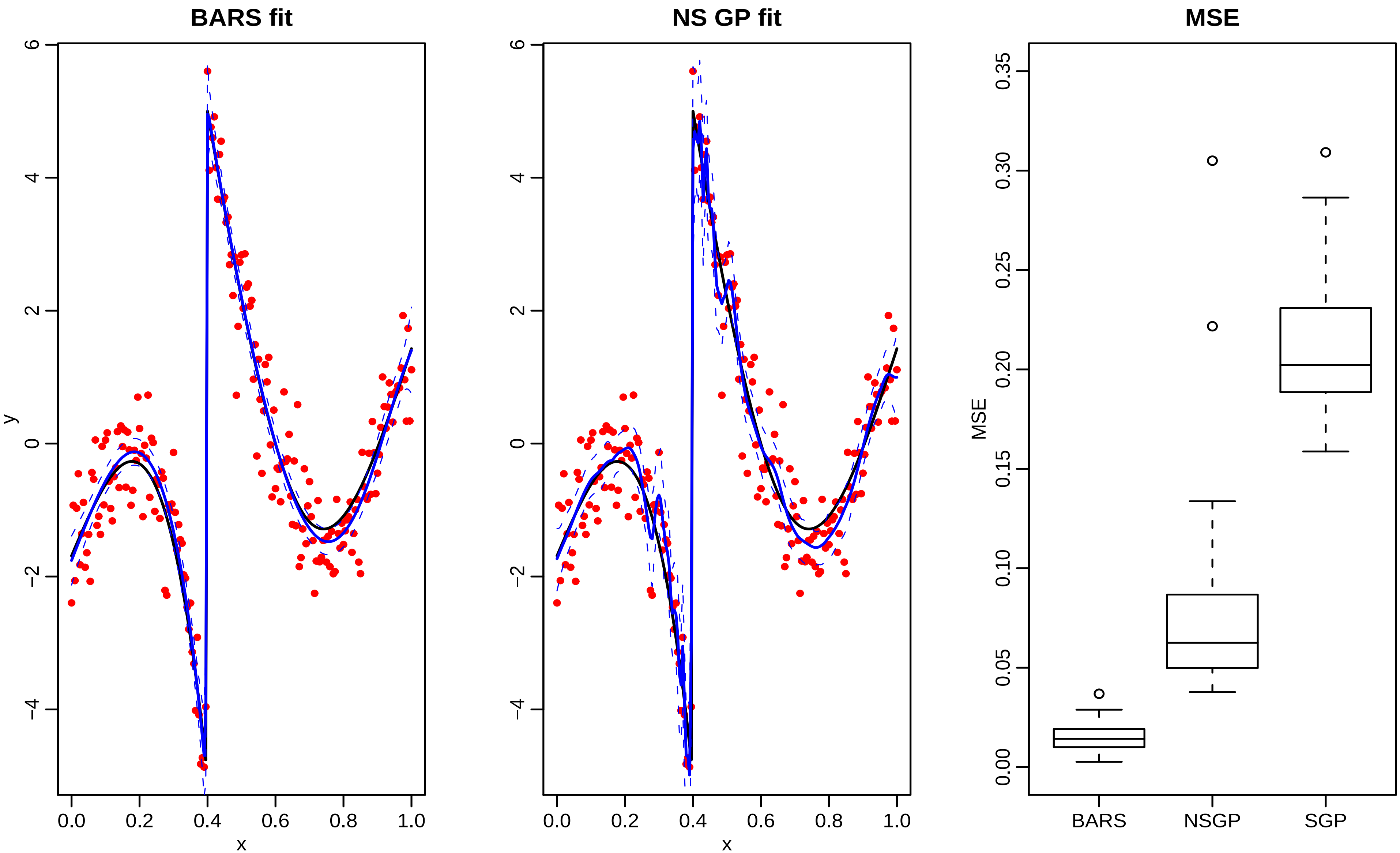
REGRESSION RESULTS - 1D



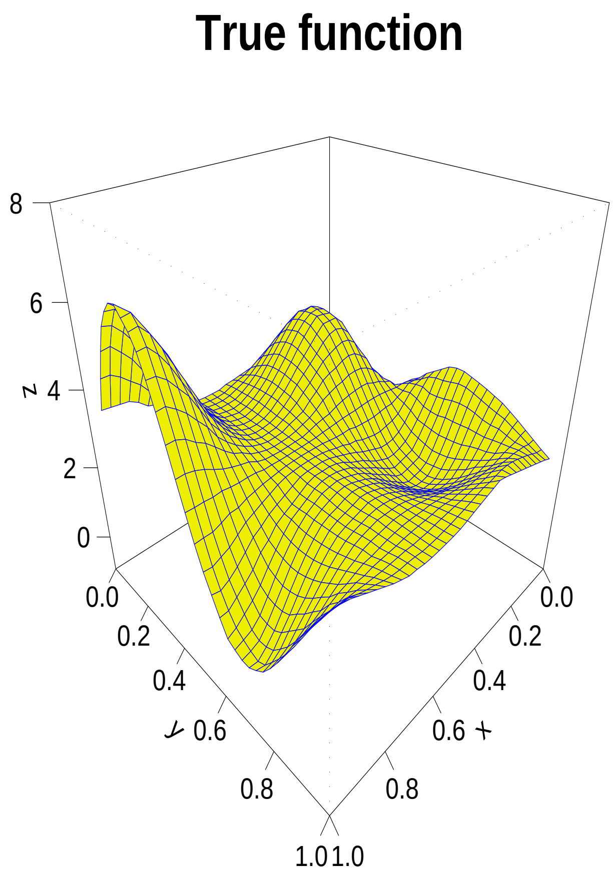
Regression results - 1D

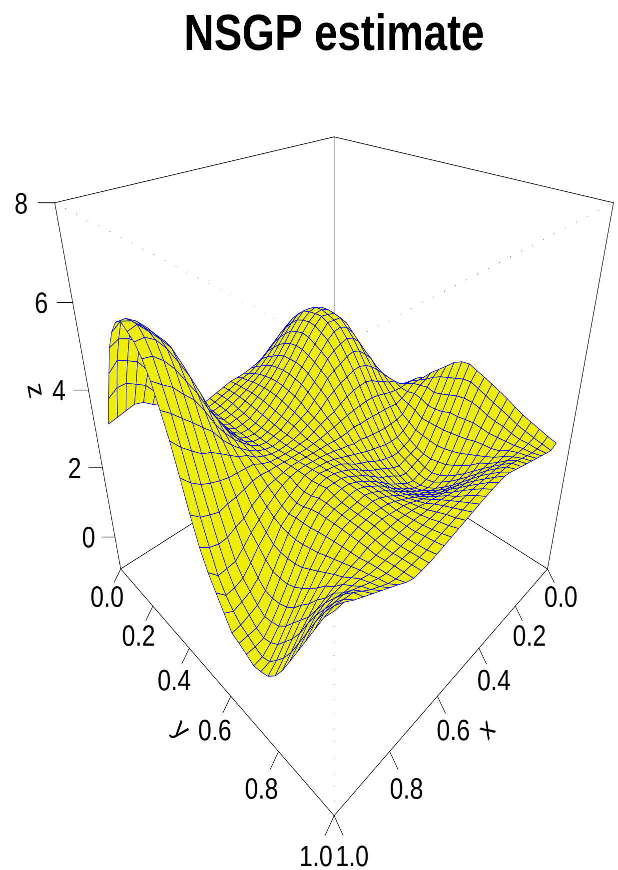


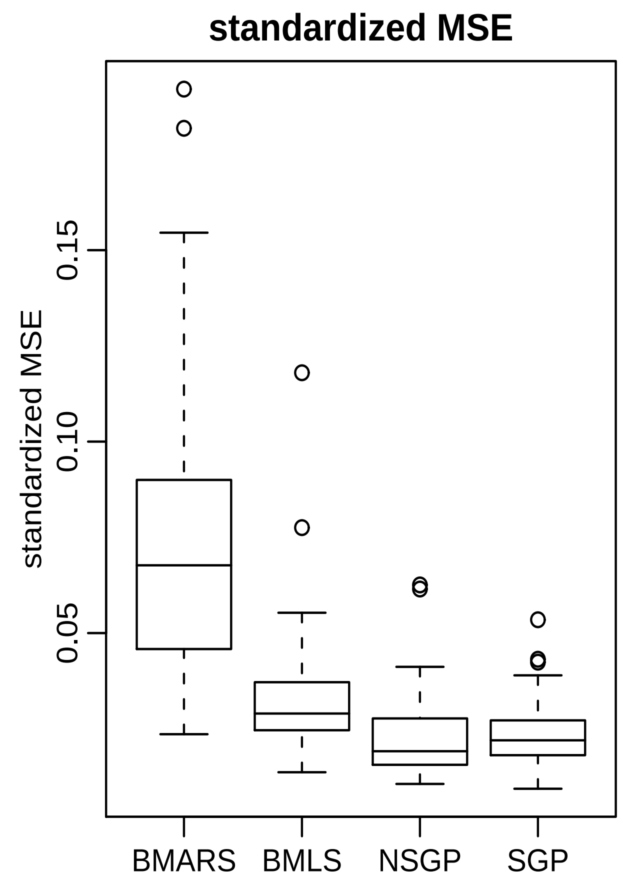
Regression results - 1D



REGRESSION RESULTS - 2D







test function: $P = 2, n = 225$

Regression results - real data

- Dec. 1993 mean temperatures in Americas, $n = 109$

  $P = 2$: longitude, latitude

- daily ozone in NY, $n = 111$

  $P = 3$: radiation, temperature, wind speed

- cross-validated MSE

model	temperature	ozone
Lin Regr	–	0.021
GAM	–	0.020
BMARS	1.74	0.0062
BMLS	2.40	0.0062
SGP	1.40	0.0062
NSGP	1.10	0.0054

RECOMMENDATIONS

- 1D:
  use BARS
- $>$1D: if response likely additive,
  use BARS
- 2-3D: if response likely relatively homogeneous,
  use stationary GP (for non-normal data, $\boldsymbol{n < 500}$)
- 2-3D: if response likely heterogeneous, $\boldsymbol{n < 250}$,
  use nonstationary GP (surface-fitting scenario)
- $\boldsymbol{P}$ or $\boldsymbol{n}$ large:
  use multivariate spline methods or another approach

Generalized Nonparametric Regression

- Model:

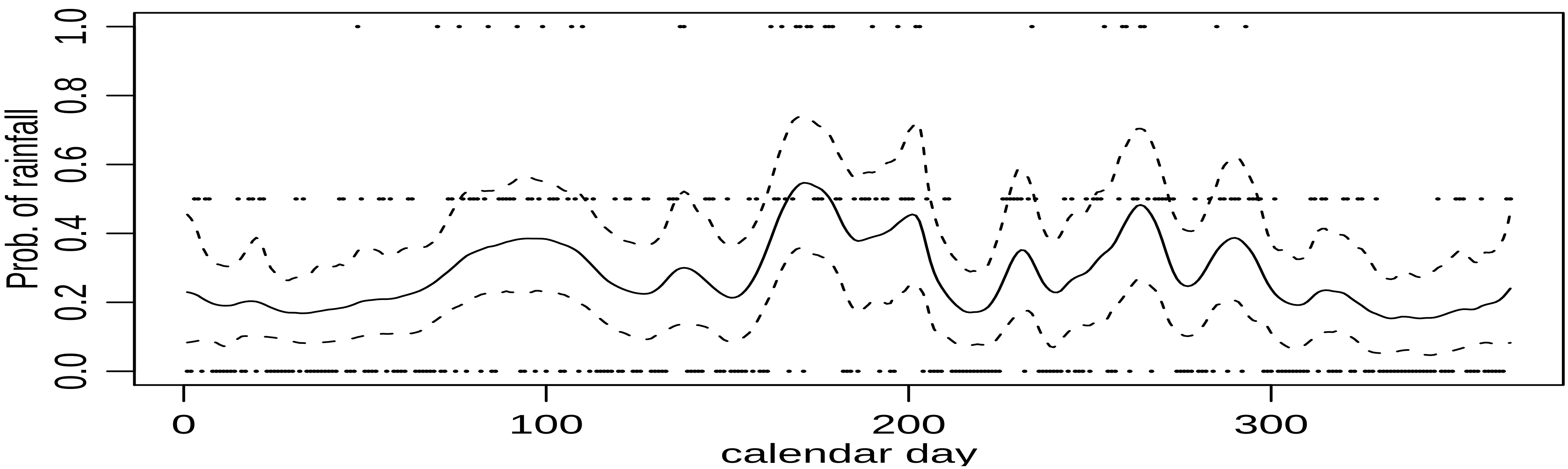
$$\boldsymbol{Y_i} \sim \boldsymbol{D}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x_i})))$$

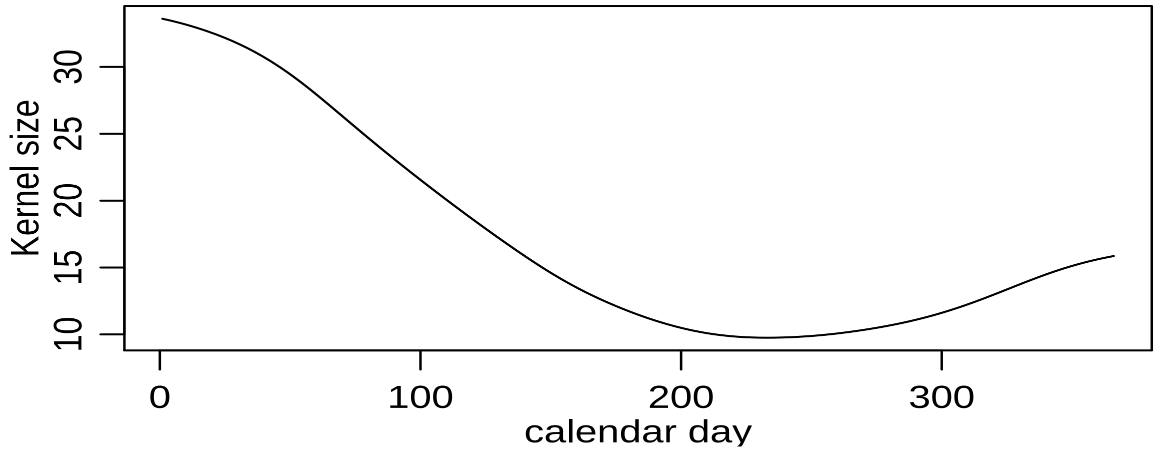
$$\boldsymbol{f}(\cdot) \sim \text{GP}(\boldsymbol{\mu}, \boldsymbol{\sigma^2} \boldsymbol{R^{NS}}(\cdot,\cdot))$$

- Examples:
  - ❖ count data ($\boldsymbol{D}$ = Poisson, $\boldsymbol{g^{-1}} = \log$)
  - ❖ binary data ($\boldsymbol{D}$ = Bernoulli, $\boldsymbol{g^{-1}} = \text{logit}$)

TOKYO RAINFALL DATA

Presence/absence of rainfall, calendar days 1983-1984

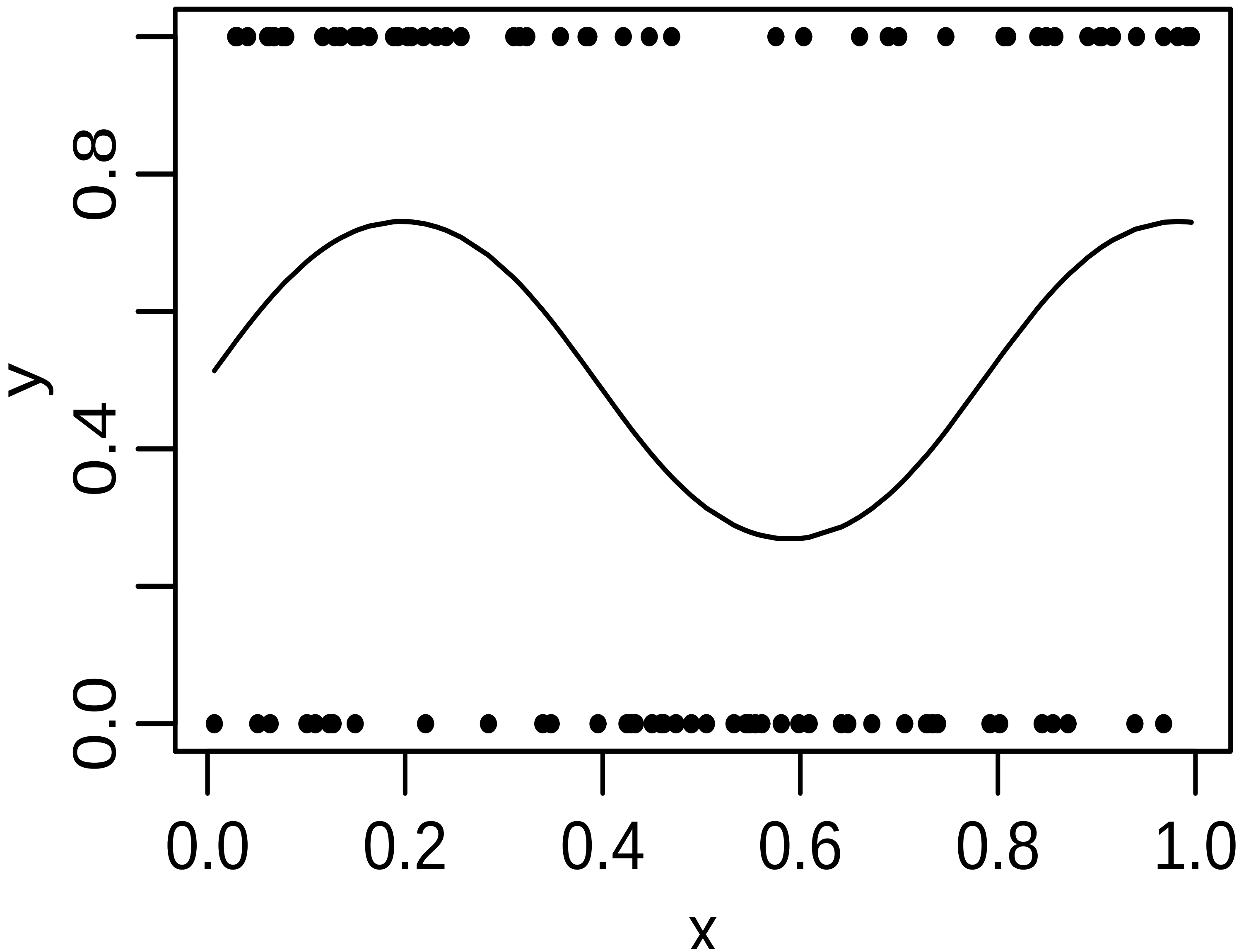




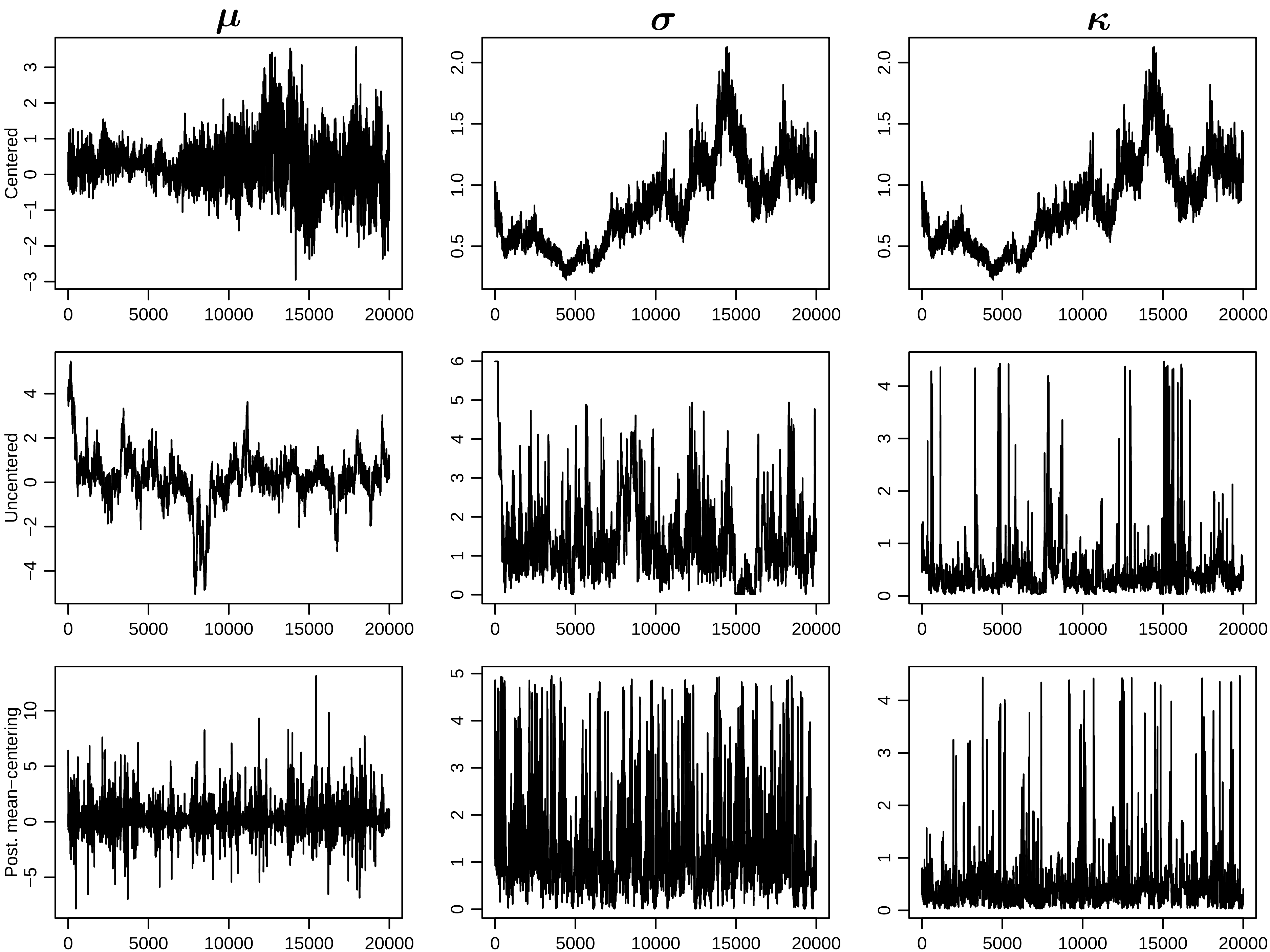
Issues in Fitting the GP Model

- Parameterizations
- Slow mixing
  - ❖ Posterior mean-centering for the generalized model:
    joint proposal for hyperparameter(s),
    f conditional on hyperparameter proposal
- Numerical sensitivity
- Parameter identifiability
- Computational speed

A SIMPLE BERNOULLI EXAMPLE



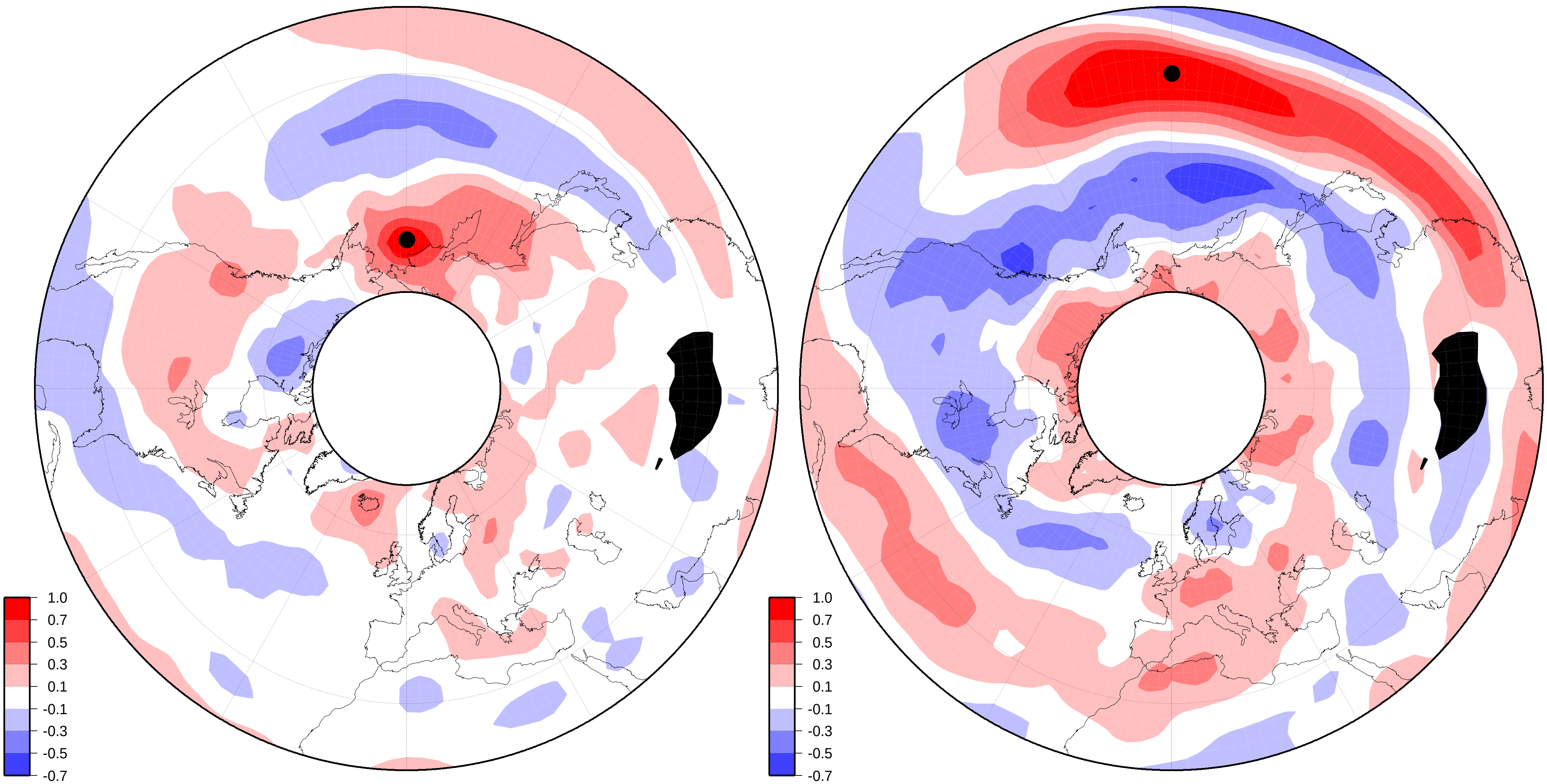
Fit a stationary GP model, $\boldsymbol{f}(\cdot) \sim \text{GP}(\boldsymbol{\mu}, \boldsymbol{\sigma^2} \boldsymbol{R}(\boldsymbol{\kappa}))$



μ
σ
κ
Centered
Uncentered
Post. mean−centering
0
5000
10000
15000
20000

NONSTATIONARY CORRELATION IN CLIMATOLOGY

$\hat{R}((60°, 180°), x)$ $\hat{R}((30°, 180°), x)$



BLACK= NO DATA

APPLICATION TO A CLIMATOLOGICAL DATASET

- Data: index of storm activity:
  - ❖ grid of 288 locations in Northern Hemisphere
  - ❖ 51 years of data (replicated data)
- Goal: analyze location-specific time trends simultaneously in space, accounting for residual spatial correlation
- Bayesian model:
  - ❖ $Y_{it} \sim N(Z_t(x_i), \delta^2)$
  - ❖ $Z_t(x_i) = \alpha(x_i) + \beta(x_i) \cdot t + \epsilon_t(x_i)$
  - ❖ $\epsilon_t(\cdot) \sim GP(0, C^{NS}(\cdot, \cdot))$
  - ❖ Stationary GP priors for the other processes

LESSONS FROM THE SPATIAL MODELLING

- Modelling complicated covariance structure is hard
  - ❖ Nonstationary model fits residual structure better than stationary model, but still seems to miss structure in the data
  - ❖ Lack of fit in stationary model drives up residual variances
  - ❖ Nychka, Wikle & Royle (2001) method for wavelet smoothing of empirical residual covariance fits poorly
- GP models shrink slope point estimates and standard errors
- For one dataset, simultaneous testing results give many more locations with significant trends than FDR, but not in a 2nd dataset

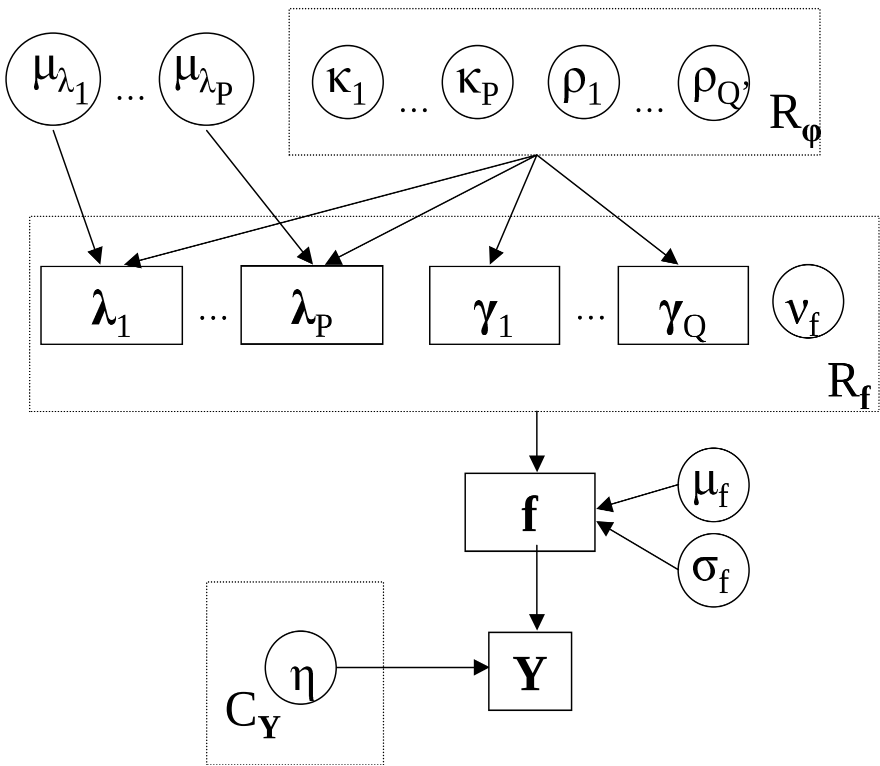
FUTURE WORK

- MCMC fitting: approaches to improve mixing and speed fitting
  - ❖ Simplified parameterization of the NS GP model
  - ❖ Computational efficiency
- Covariate selection in the NS GP model

FUTURE WORK – SPATIAL MODEL

- Further investigation of methods for flexibly fitting covariance structure of replicated data
  - ❖ Improved fitting criteria for wavelet smoothing of empirical covariance
- Incorporation of nonlinear time models

FITTING THE GP REGRESSION MODEL



$$Z \in \{\boldsymbol{f}, \boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_P}, \boldsymbol{\gamma_1}, \dots, \boldsymbol{\gamma_Q}\}$$

Some approaches

- Integrate $\boldsymbol{f}$ out of the model (normal likelihood)
- Fix the hyperparameters
- Full MCMC (non-conjugate processes)

PARAMETERIZATIONS OF GAUSSIAN PROCESSES

- (non-centered) $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \boldsymbol{L}(\boldsymbol{\kappa})\boldsymbol{\omega}$
  - ❖ directly sample $\boldsymbol{\omega}$ not $\boldsymbol{Z}$
- (centered) $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \boldsymbol{L}(\boldsymbol{\kappa})\boldsymbol{L}(\boldsymbol{\kappa})^T)$
  - ❖ straightforward sampling
    - ✥ $\boldsymbol{Z}$ not consistent with proposed hyperparameters
  - ❖ joint sampling of centered parameterization
    - ✥ $\boldsymbol{Z}, (\boldsymbol{\mu}, \boldsymbol{Z}), (\boldsymbol{\kappa}, \boldsymbol{Z}), (\boldsymbol{\sigma}, \boldsymbol{Z})$
    - ✥ $\boldsymbol{Z}^* = \boldsymbol{\mu}^* + \boldsymbol{\sigma}^* \boldsymbol{L}(\boldsymbol{\kappa}^*)(\boldsymbol{\sigma} \boldsymbol{L}(\boldsymbol{\kappa}))^{-1}(\boldsymbol{Z} - \boldsymbol{\mu})$
    - ✥ avoids numerical issues with $(\boldsymbol{\sigma} \boldsymbol{L}(\boldsymbol{\kappa}))^{-1}(\boldsymbol{Z} - \boldsymbol{\mu})$
    - ✥ $\boldsymbol{Z}^*$ consistent with proposed hyperparameters, but not likelihood
  - ❖ joint sampling but make use of approximate posterior mean, $\widetilde{\boldsymbol{Z}}$
    - ✥ Not feasible for processes involved in nonstationary correlation function

Posterior mean-centering

- Joint sampling centers around μ:
  $Z^* = \mu^* + \sigma^* L(\kappa^*)(\sigma L(\kappa))^{-1}(Z - \mu)$

  Proposal doesn't take account of likelihood

- PMC centers around $\widetilde{Z}$:
  $Z^* = \widetilde{Z}^* + \sigma^* L(\kappa^*)(\sigma L(\kappa))^{-1}(Z - \widetilde{Z})$

- conjugate case:

$$\begin{aligned}\widetilde{Z} &= C_Y(C_Z + C_Y)^{-1}\mu + C_Z(C_Z + C_Y)^{-1}y \\ &= \mu + C_Z(C_Z + C_Y)^{-1}(y - \mu)\end{aligned}$$

- generalized case: $\widetilde{Z} \approx \mu + C_f(C_f + C_{Y'})^{-1}(y' - \mu)$

- use IRLS approach

  $y'_i = g^{-1}(y_i) \approx f(x_i) + \frac{\partial g(x_i)}{\partial f(x_i)}(y_i - g(x_i))$

  $(C'_Y)_{ii} \approx \left(\frac{\partial g(x_i)}{\partial f(x_i)}\right)^2 (C_Y)_{ii}$