Nonstationary Covariance Functions for Gaussian Process Regression

Christopher J. Paciorek
Department of Biostatistics
Harvard School of Public Health

Mark J. Schervish
Department of Statistics
Carnegie Mellon University

Neural Information Processing Systems

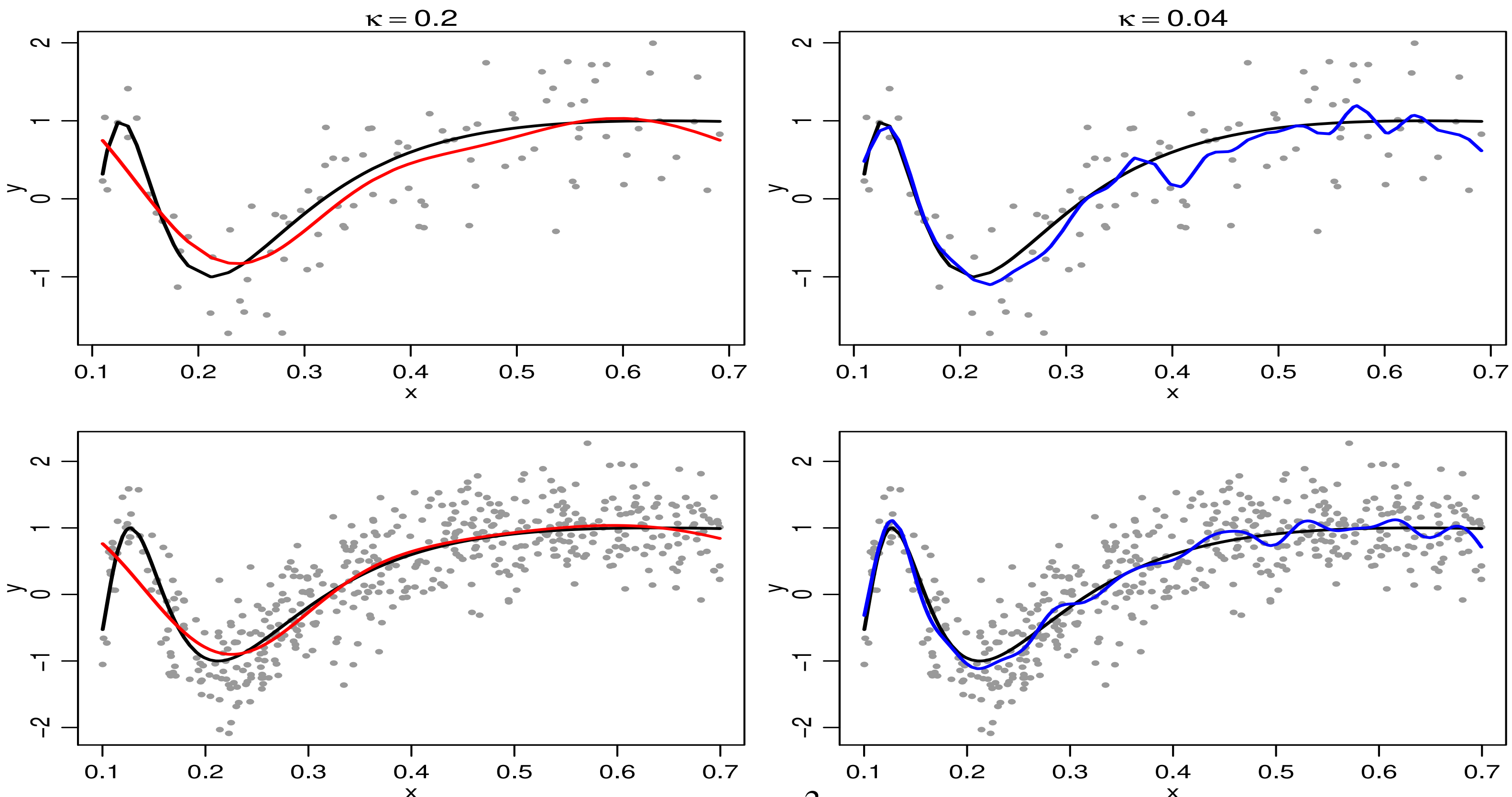
December 9, 2003

Abstract

We introduce a class of nonstationary covariance functions for Gaussian process (GP) regression. Nonstationary covariance functions allow the model to adapt to functions whose smoothness varies with the inputs. The class includes a nonstationary version of the Matérn stationary covariance, in which the differentiability of the regression function is controlled by a parameter, freeing one from fixing the differentiability in advance. In experiments, the nonstationary GP regression model performs well when the input space is two or three dimensions, performing comparably to a Bayesian neural network and outperforming an optimized neural network model and Bayesian free-knot spline models. In one dimension, it is outperformed by a state-of-the-art Bayesian free-knot spline model and by the Bayesian neural network model. The model readily generalizes to non-Gaussian data. Use of computational methods for speeding GP fitting allows for implementation of the method on somewhat larger datasets.

The General Problem of Variable Smoothness

For functions whose smoothness varies with the input values, most nonparametric methods, including Gaussian process (GP) regression with a stationary covariance function, will oversmooth in some regions and undersmooth in others. Below I show the posterior mean regression function from a GP regression implementation with two different fixed values (blue and red lines) of the correlation scale hyperparameter. The top panels are for 100 data points using Gaussian error around the true function (black line) while the bottom panels are for 500 data points.



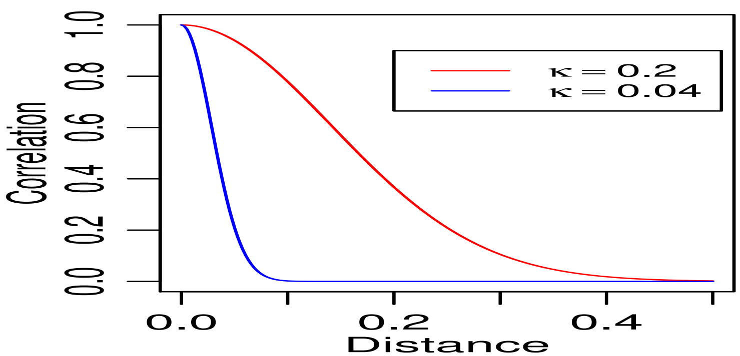
Current approaches

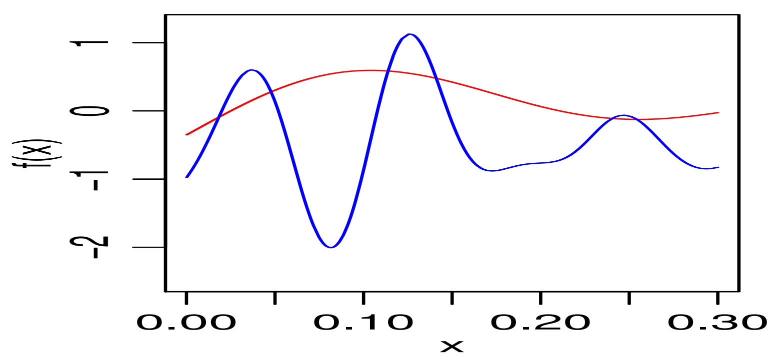
- Many other approaches to this problem have been presented, but good approaches for multivariate features and comparison amongst the approaches are still needed. Some current approaches are:
  - ❖ Free-knot regression spline models: number and location of knots optimized, with more knots in locations where function varies more quickly
    - ✦ Bayesian Adaptive Regression Splines (BARS) (DiMatteo, Genovese, and Kass 2002) (single feature only)
    - ✦ Bayesian Multivariate Adaptive Regression Splines (BMARS) (Denison, Mallick, and Smith 1998) and Bayesian Multivariate Linear Splines (BMLS) (Holmes and Mallick 2001) (multiple features)
  - ❖ Adaptive penalty smoothing spline models (MacKay and Takeuchi 1995)
  - ❖ Neural network models (Neal, etc.)
  - ❖ Mixtures of stationary Gaussian processes (Tresp 2001; Rasmussen and Ghahramani 2002)
  - ❖ Stationary Gaussian process regression in a deformed feature space (Damian, Sampson, and Guttorp 2001, Schmidt and O'Hagan 2000) (used for spatial features)
- In this poster and accompanying paper, we describe an approach to the variable smoothness problem using Gaussian process regression with nonstationary covariance functions.

Stationary Correlation Functions

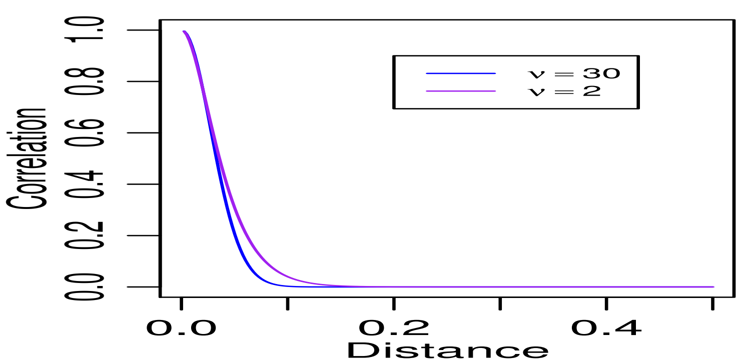
Here are two stationary correlation functions, with examples of the correlation function (left) and sample functions drawn from a Gaussian process parameterized by the correlation function (right). The squared exponential correlation function (top) gives sample functions with infinitely many derivatives, while the Matérn correlation function (bottom) gives sample functions whose number of derivatives varies with ν: $\lceil \nu - 1 \rceil$ derivatives. For $\nu \rightarrow \infty$, one recovers the squared exponential form.

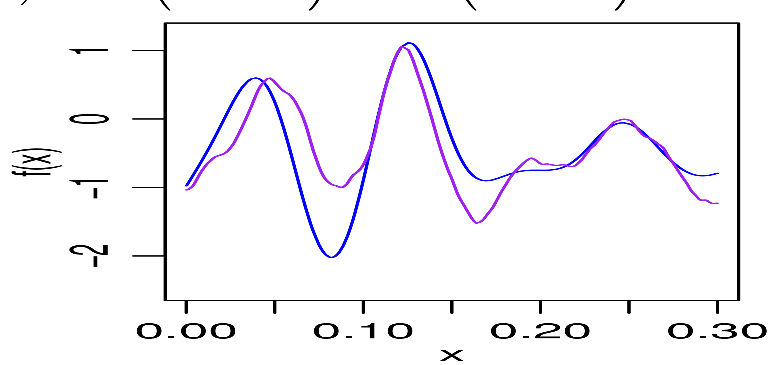
Squared exponential: $$R(\tau) = \exp\left(-\left(\frac{\tau}{\kappa}\right)^2\right)$$





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





Correlation function

Sample functions

A NONSTATIONARY COVARIANCE FUNCTION

- Higdon, Swall, and Kern (1999) (HSK) introduced the following nonstationary correlation function, where c_{ij} is a normalizing term and $k_i(\boldsymbol{u})$ is a function (called the kernel function) centered at $\boldsymbol{x_i}$.

  $R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = c_{ij} \int_{\Re^P} k_i(\boldsymbol{u}) k_j(\boldsymbol{u}) d\boldsymbol{u}$

- Guaranteed positive definite

- Using Gaussian kernels, one gets a closed form for the correlation:

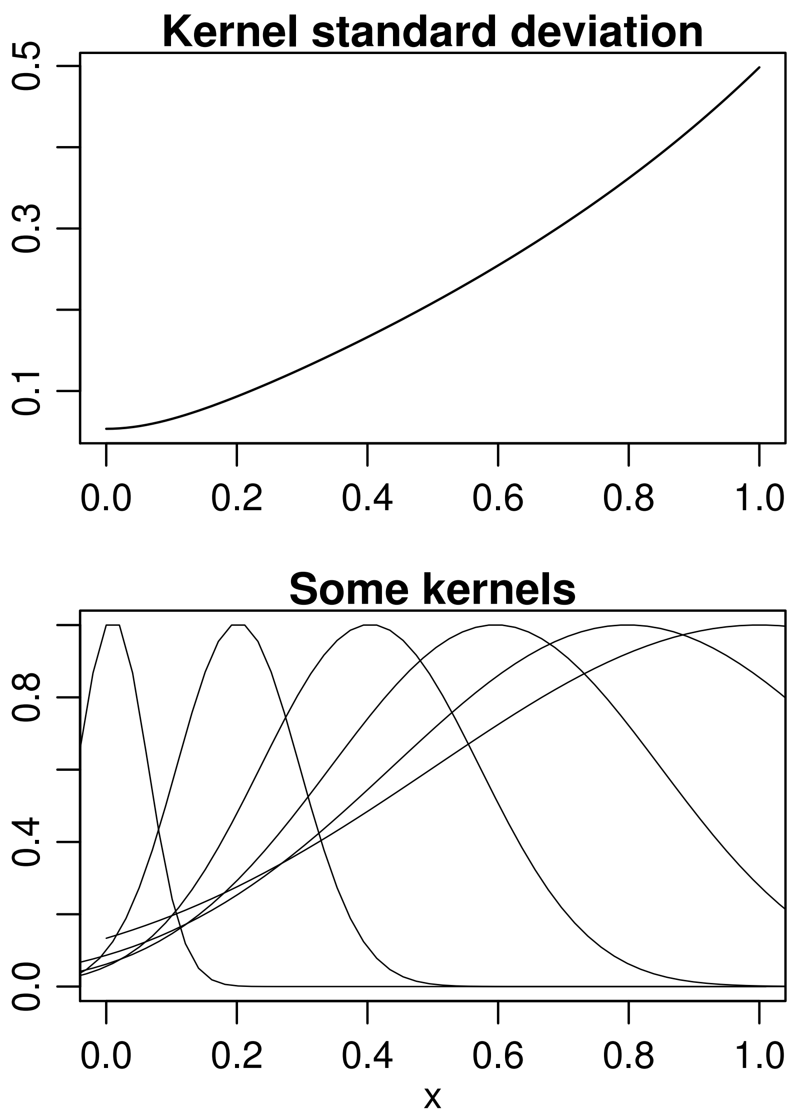
$$\begin{aligned} k_i(\boldsymbol{u}) &\propto \exp\left(-(\boldsymbol{u} - \boldsymbol{x_i})^T \Sigma_i^{-1} (\boldsymbol{u} - \boldsymbol{x_i})\right) \\ R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) &= c_{ij} \exp\left(-(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1}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

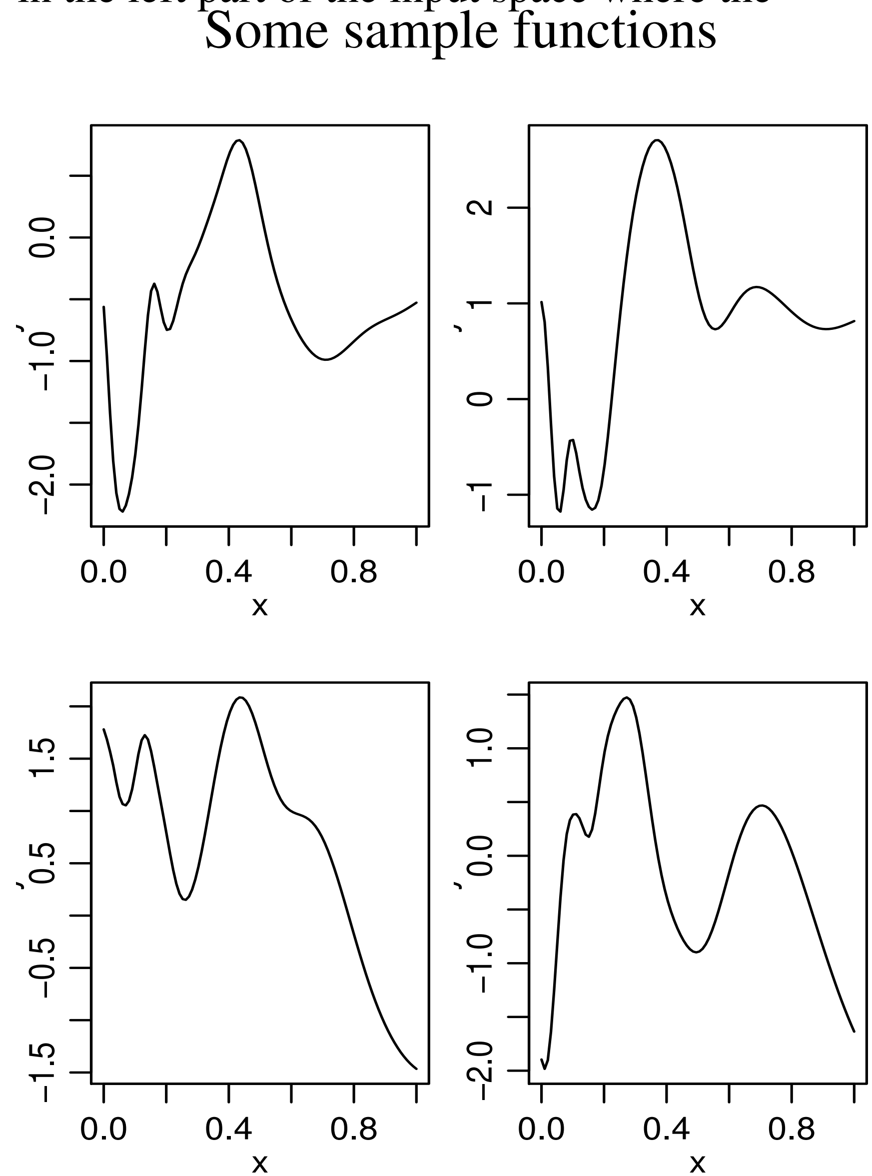
  Gibbs (1997) gave a special case with diagonal Σ_i, Σ_j

- $f(\cdot) \sim \text{GP}(\mu, \sigma^2 R^{NS}(\cdot, \cdot; \Sigma(\cdot)))$ is a nonstationary Gaussian process

Nonstationary GPs in 1D

Here are four sample functions (right) drawn from a nonstationary Gaussian process distribution whose Gaussian kernels are defined based on their standard deviation (left). Note that the sample functions are more wiggly in the left part of the input space where the kernels are less broad.

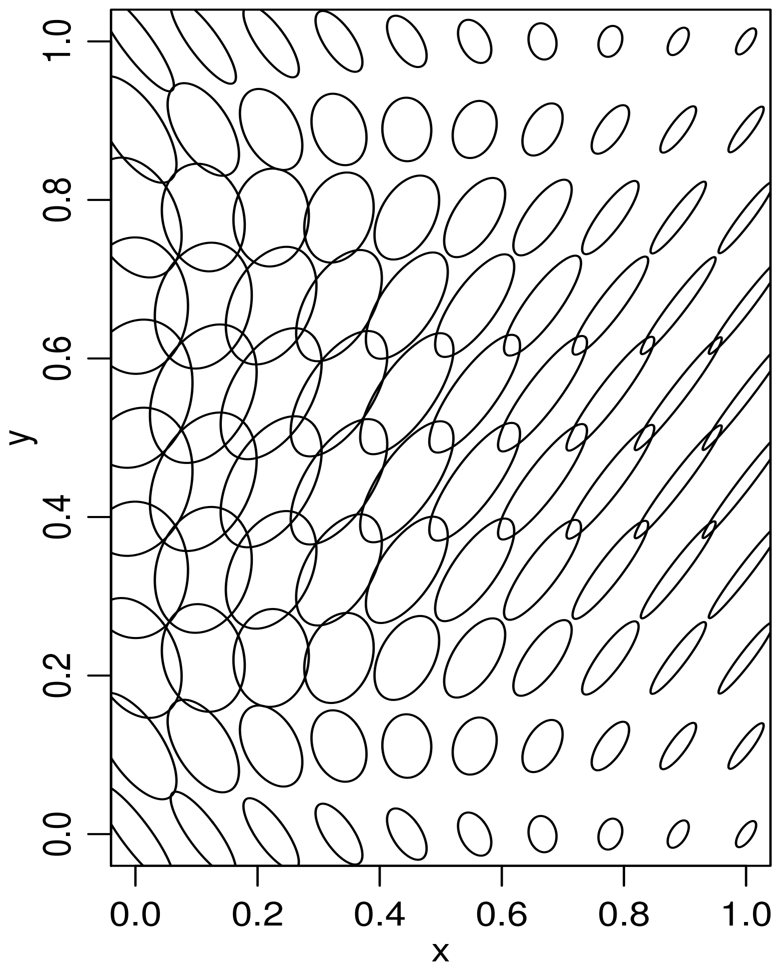




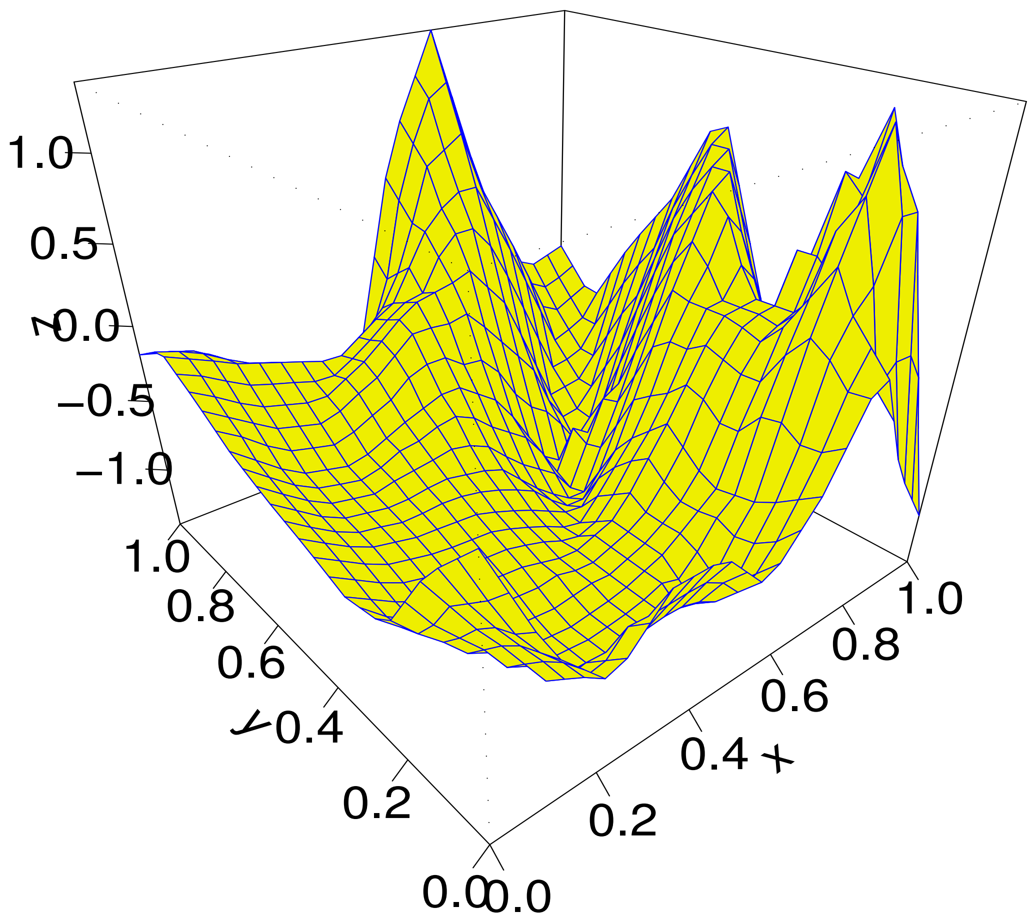
Nonstationary GPs in 2D

Here (right) is one sample function (of two inputs) from a nonstationary Gaussian process distribution whose Gaussian kernels are depicted using ellipses of constant density (left). Note the smoothness of the function where the kernels are large and the directionality of the smoothness where the kernels have strong directionality.

Kernel Structure



Sample Function



Generalizing the HSK Kernel Method

- The stationary squared exponential form (left) has the same form as the HSK nonstationary covariance (right):

$$\exp\left(-\left(\frac{\tau}{\kappa}\right)^2\right) \qquad c_{ij}\exp\left(-(\boldsymbol{x_i}-\boldsymbol{x_j})^T\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})\right)$$

  - ❖ The HSK nonstationary covariance merely replaces Euclidean distance with a quadratic form in the exponential.
  - ❖ Gaussian process distributions with this nonstationary covariance have infinitely-differentiable sample paths if $\Sigma(\cdot)$ vary smoothly.

- Consider the following ‘distance measures’ (the nonstationary one does not satisfy the triangle inequality):

isotropy
$$\tau_{ij}^2 = (\boldsymbol{x_i}-\boldsymbol{x_j})^T(\boldsymbol{x_i}-\boldsymbol{x_j})$$

anisotropy
$$\tau^{*2}_{ij} = (\boldsymbol{x_i}-\boldsymbol{x_j})^T\Sigma^{-1}(\boldsymbol{x_i}-\boldsymbol{x_j})$$

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 and still retain positive definiteness?

Generalized Nonstationary Covariance

- Theorem 1 (Paciorek 2003): if a stationary correlation function, $R(\tau)$, is positive definite for $\Re^P$, $P = 1, 2, \ldots$, then

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

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

- Theorem 2 (Paciorek 2003): Smoothness (differentiability) properties of original stationary correlation $R(\tau)$ are retained if elements of $\Sigma(\cdot)$ vary smoothly in the feature space.

PROOF OF THEOREM 1 (SKETCH)

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

$$R(\tau) = \int_0^\infty \exp(-\tau^2 w) h(w) dw \quad \text{(Schoenberg 1938)}$$

- Reexpress the proposed nonstationary correlation function:

$$\begin{aligned}
R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) &= \frac{2^{\frac{P}{2}} \left|\Sigma_i\right|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}}}{\left|\Sigma_i + \Sigma_j\right|^{\frac{1}{2}}} \int_0^\infty \exp(-Q_{ij} w) h(w) dw \\
&= \frac{2^{\frac{P}{2}} \left|\Sigma_i\right|^{\frac{1}{4}} \left|\Sigma_j\right|^{\frac{1}{4}}}{\left|\Sigma_i + \Sigma_j\right|^{\frac{1}{2}}} \cdot \\
& \quad \int_0^\infty \exp\left(-\frac{1}{2}(\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\frac{\Sigma_i + \Sigma_j}{2w}\right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})\right) h(w) dw \\
&= \int_0^\infty \int_{\Re^P} k_{i,w}(\boldsymbol{u}) k_{j,w}(\boldsymbol{u}) d\boldsymbol{u} h(w) dw
\end{aligned}$$

- Now check the definition of positive definiteness directly:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \int_0^\infty \int_{\Re^P} k_{i,w}(\boldsymbol{u}) k_{j,w}(\boldsymbol{u}) d\boldsymbol{u} h(w) dw \\
&= \int_0^\infty \int_{\Re^P} \sum_{i=1}^{n} a_i k_{i,w}(\boldsymbol{u}) \sum_{j=1}^{n} a_j k_{j,w}(\boldsymbol{u}) d\boldsymbol{u} h(w) dw \\
&= \int_0^\infty \int_{\Re^P} \left(\sum_{i=1}^{n} a_i k_{i,w}(\boldsymbol{u}) \right)^2 d\boldsymbol{u} h(w) dw \geq 0.
\end{aligned}$$

- The key is that the covariance must depend only on location-specific kernels

Nonstationary Matérn Covariance

- A particular nonstationary covariance of interest is a Matérn form, which satisfies the conditions of Theorem 1.

stationary form — nonstationary form

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

- Provided $\Sigma(\cdot)$ varies smoothly, by Theorem 2, this form will give sample functions whose differentiability varies with ν. By not constraining the differentiability, this gives a more flexible form of the correlation function than the original HSK nonstationary correlation, and has asymptotic advantages (Stein 1999).

A BAYESIAN NONSTATIONARY GP REGRESSION MODEL

- Bayesian model

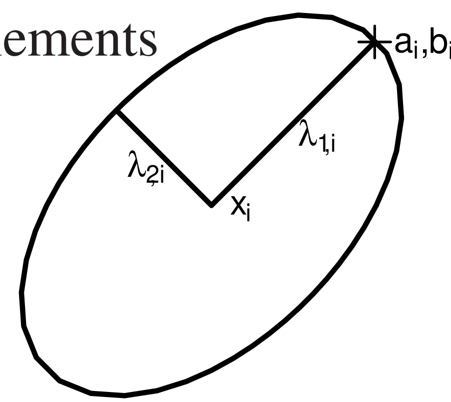
$$\begin{aligned} Y_i &\sim N(f(\boldsymbol{x_i}), \eta^2),\ \boldsymbol{x_i} \in \Re^P \\ f(\cdot) &\sim \mathrm{GP}(\mu, \sigma^2 R^{NS}(\cdot, \cdot; \Sigma(\cdot), \nu)) \end{aligned}$$

  ❖ Let R^{NS} be the nonstationary Matérn correlation

  ❖ Kernels, $\Sigma(\cdot)$, are constructed based on stationary GP priors, as described next.

SMOOTHLY-VARYING KERNEL MATRICES

- Goals:
  - ❖ Define multiple kernel matrices, Σ_i
    one for each training observation and test/prediction observation
  - ❖ Matrix elements should be smoothly-varying in input space
  - ❖ Matrices must be positive definite
- Use spectral decomposition ($\Sigma_i = \Gamma_i^T \Lambda_i \Gamma_i$)
  - ❖ Eigenvector matrix, Γ_i, is parameterized as first eigenvector plus successive orthogonal vectors in reduced-dimension subspaces
    - ✦ stationary GP priors on unnormalized eigenvector coordinates [(a_i, b_i) in 2-d cartoon]
  - ❖ Eigenvalue matrix, Λ_i, parameterized by diagonal elements
    - ✦ stationary GP priors on logarithms of diagonal elements (log $\lambda_{1,i}$, log $\lambda_{2,i}$ in 2-d cartoon)
  - ❖ gets unwieldy and highly-parameterized for large P



EFFICIENT REPRESENTATIONS OF STATIONARY GPS

- Let $\mathbf{\Phi}$ be a vector of process values

1. Matérn basis functions (Kammann and Wand 2003)
   - ❖ $\mathbf{\Phi} = \mu + \sigma Z \Omega^{-\frac{1}{2}} \boldsymbol{w}$
     - ✦ $Z = (C(\| x_i - \kappa_k \|))_{,1 \le i \le n, 1 \le k \le K}$
     - ✦ $\Omega = (C(\| \kappa_j - \kappa_k \|))_{,1 \le j \le K, 1 \le k \le K}$
     - ✦ $C(\cdot)$ a stationary covariance function
   - ❖ matrix operations based on K knots, $\{\kappa_k\}$, so more efficient
   - ❖ motivation: if $\{\kappa_k\} = \{x_i\}$, Cov$(\mathbf{\Phi}) = C(\cdot)$
   - ❖ When sample hyperparameters in MCMC, sample process as well:
     $\mathbf{\Phi} = \mu + \sigma Z \Omega^{-\frac{1}{2}} \boldsymbol{w}$

2. Fourier basis functions (Wikle 2002)

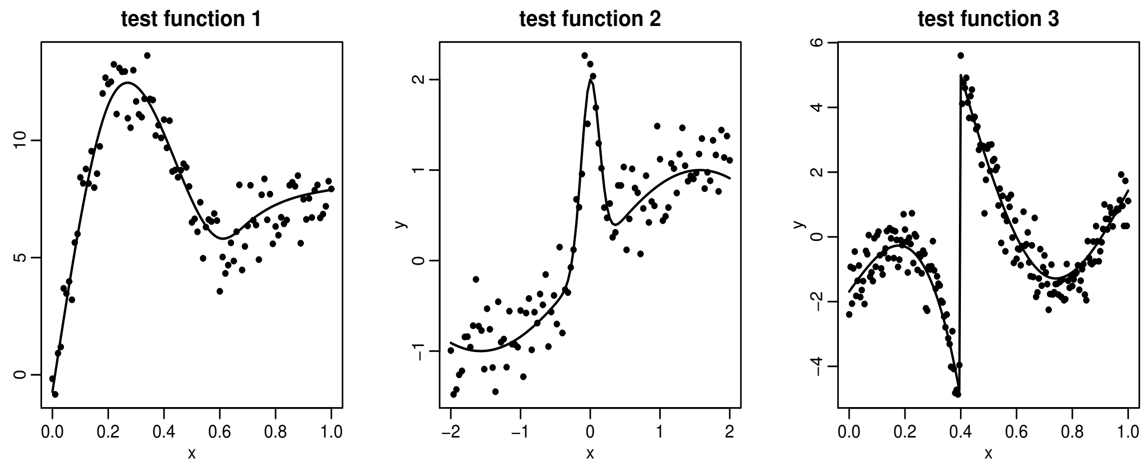
- $\boldsymbol{\Phi}_{\text{dat}} = \mu + \sigma A \boldsymbol{\Phi}_{\text{grid}}$ ($\boldsymbol{\Phi}_{\text{dat}}$ is process values at data locations)
- $\boldsymbol{\Phi}_{\text{grid}} = \Psi \boldsymbol{w}$ ($\boldsymbol{\Phi}_{\text{grid}}$ is discretized process on a grid)
- $\boldsymbol{w}$ elements are independent, complex-valued RVs
  - their variance is based on spectral density of stationary $C(\cdot)$
- $\Psi \boldsymbol{w}$ is the inverse FFT (Ψ are Fourier basis vectors)
- propose blocks of values of $\boldsymbol{w}$ with focus on low-frequency coefficients
- motivation: Cov$(\boldsymbol{\Phi}) = C(\cdot)$ asymptotically as grid gets finer
- When sample hyperparameters in MCMC, sample process as well:

  $\boldsymbol{\Phi}_{\text{dat}} = \mu + \sigma A \Psi \boldsymbol{w}$

Nonparametric Regression Comparison - 1D

- Compare to:
  - ❖ stationary GP
  - ❖ free-knot regression spline model (BARS) (DiMatteo et al. 2002)
  - ❖ neural network (with number of hidden units that give best result) (R nnet library)
  - ❖ Bayesian neural network (R. Neal software as implemented by A. Vehtari)
- Simulate 50 datasets and fit each model to each dataset
- Compare results based on mean squared error (relative to true function)

REGRESSION RESULTS - 1D



Method	Function 1	Function 2	Function 3
BARS	.0081 (.0071,.0092)	.012 (.011,.013)	.0050 (.0043,.0056)
Bayesian neural network	.0082 (.0072,.0093)	.011 (.010,.014)	.015 (.014,.016)
Nonstat. GP	.0083 (.0073,.0093)	.015 (.013,.016)	.026 (.021,.030)
Stat. GP	.0083 (.0073,.0093)	.026 (.024,.029)	.071 (.067,.074)
neural network	.0108 (.0095,.012)	.013 (.012,.015)	.0095 (.0086,.010)

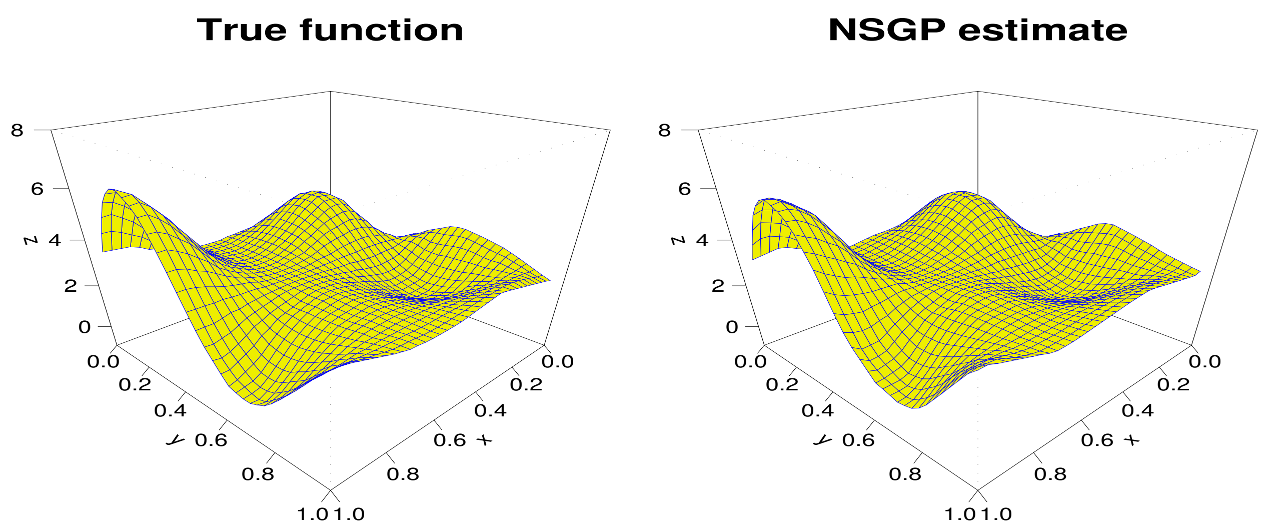
- Free-knot spline (BARS) performs best, followed by Bayesian neural network and nonstationary GP
- Most methods are similar on smoothly varying Function 1

NONPARAMETRIC REGRESSION COMPARISON - 2/3D

- Compare to:
  - stationary GP
  - free-knot regression spline with tensor products of univariate splines (BMARS) (Denison et al. 1998)
  - free-knot regression spline with multivariate linear splines (BMLS) (Holmes and Mallick 2001)
  - neural network (optimal number of hidden units)(R nnet library)
  - Bayesian neural network (R. Neal software as implemented by A. Vehtari)

EXAMPLES

- 1.) Simulated dataset with 2 inputs: $P = 2$, $n = 225$
  - ❖ simulate 50 datasets and compare using MSE (relative to true function)



- 2.) Real dataset of Dec. 1993 mean temperatures in Americas, $n = 109$
  - ❖ $P = 2$: longitude, latitude
  - ❖ MSE based on 50-fold cross-validation (MSE relative to test data)
- 3.) Real dataset of daily ozone in NY, $n = 111$
  - ❖ $P = 3$: radiation, temperature, wind speed
  - ❖ MSE based on 50-fold cross-validation (MSE relative to test data)

MEAN SQUARED ERROR

Method	Function with 2 inputs	Temp. data	Ozone data
Bayesian neural network	.020 (.019,.022)	.35	.32
Nonstat. GP	.023 (.020,.026)	.36	.29
Stat. GP	.024 (.021,.026)	.46	.33
BMARS	.076 (.065,.087)	.53	.33
BMLS	.033 (.029,.038)	.78	.33
neural network	.040* (.033,.047)	.60	.34

* Holmes and Mallick (2001) report a value of .07 for a neural network

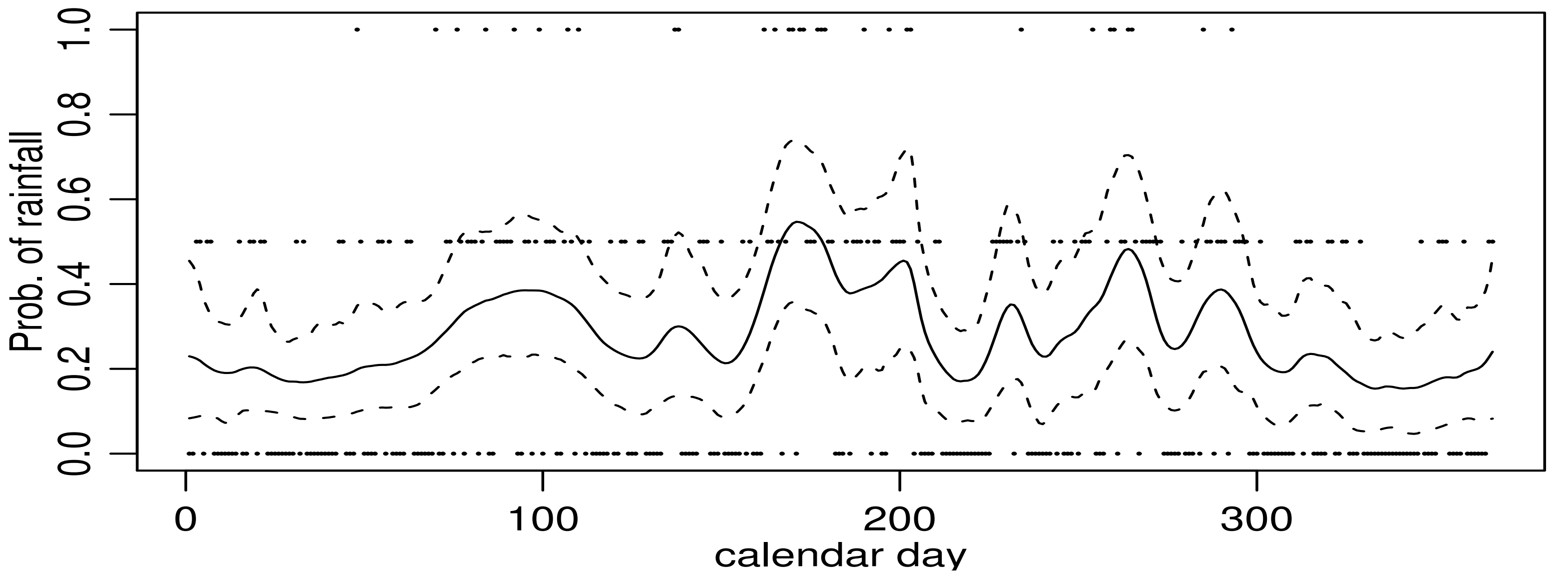
- Bayesian neural network and nonstationary Gaussian process appear to outperform other approaches.

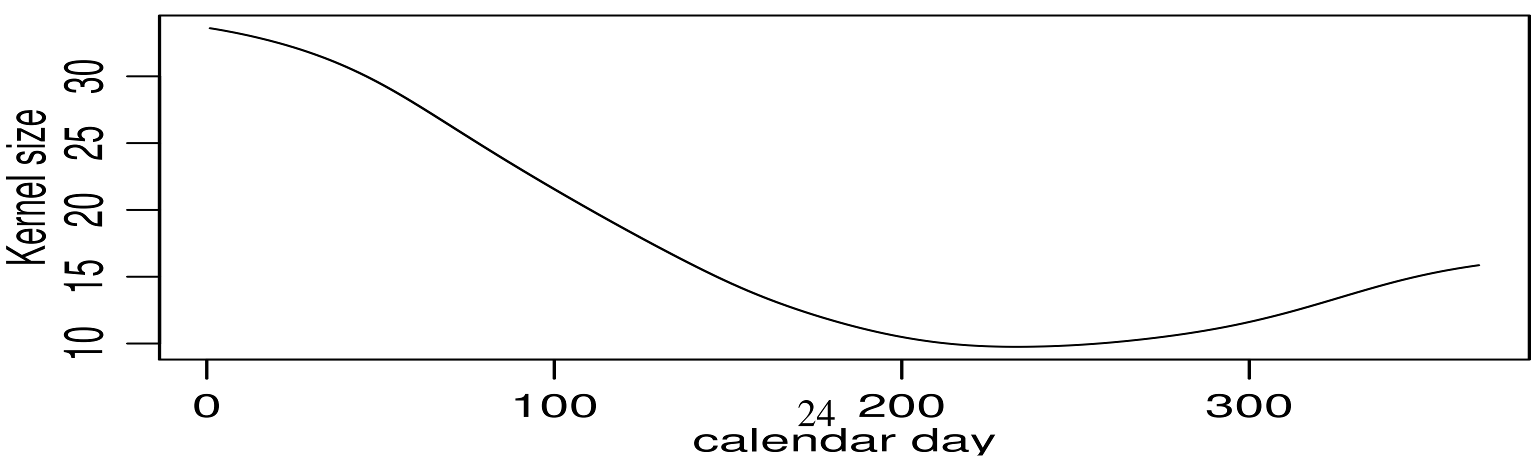
Generalized Nonparametric Regression

- Model:
  - ❖ $Y_i \sim D(g(f(x_i)))$
  - ❖ $f(\cdot) \sim \text{GP}(\mu, \sigma^2 R^{NS}(\cdot, \cdot; \Sigma(\cdot), \nu))$
  - ❖ D is a probability distribution and $g(\cdot)$ is a link function
- Examples:
  - ✥ count data (D = Poisson, $g^{-1} = \log$)
  - ✥ binary data (D = Bernoulli, $g^{-1} = \text{logit}$)

Tokyo Rainfall data example

Data are presence/absence of rainfall for the calendar days, 1983-1984 ($Y_i \in \{0, 1, 2\}$). Top panel shows posterior mean probability of rainfall as a function of calendar day, with pointwise 95% uncertainty intervals, and bottom panel shows standard deviation of the Gaussian kernels as a function of calendar day. The probability of rainfall appears to be more variable toward the end of the year.





CONCLUSIONS

We introduce a class of nonstationary covariance functions that can be used in Gaussian process regression (and classification) models and allow the model to adapt to variable smoothness in the unknown function. The nonstationary GPs improve on stationary GP models on several test datasets. In test functions on one-dimensional spaces, a state-of-the-art free-knot spline model and Bayesian neural network outperform the nonstationary GP, but in higher dimensions, the nonstationary GP outperforms two free-knot spline approaches and a non-Bayesian neural network while being comparable to a Bayesian neural network. The nonstationary GP may be of particular interest for data indexed by spatial coordinates.

Unfortunately, the nonstationary GP requires many more parameters than a stationary GP, particularly as the dimension grows, losing the attractive simplicity of the stationary GP model. Use of stationary GP priors in the hierarchy of the model to parameterize the nonstationary covariance results in slow computation. More efficient representations of these stationary GPs improves efficiency but still limits the model to approximately $n < 1000$. The slow part of the computation for Gaussian data is that calculating the marginal likelihood requires the Cholesky decomposition of an n by n matrix. Our approach provides a general modelling framework; other low-rank approximations to the covariance matrix (e.g., Smola and Bartlett 2001; Seeger and Williams 2003) may further speed fitting.

References

Damian, D., Sampson, P., and Guttorp, P. (2001), "Bayesian estimation of semi-parametric non-stationary spatial covariance structure," *Environmetrics*, 12, 161–178.

Denison, D., Mallick, B., and Smith, A. (1998), "Bayesian MARS," *Statistics and Computing*, 8, 337–346.

DiMatteo, I., Genovese, C., and Kass, R. (2002), "Bayesian curve-fitting with free-knot splines," *Biometrika*, 88, 1055–1071.

Gibbs, M. (1997), *Bayesian Gaussian Processes for Classification and Regression*, unpublished Ph.D. dissertation, University of Cambridge.

Higdon, D., Swall, J., and Kern, J. (1999), "Non-stationary spatial modeling," in *Bayesian Statistics 6*, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, Oxford, U.K.: Oxford University Press, pp. 761–768.

Holmes, C. and Mallick, B. (2001), "Bayesian regression with multivariate linear splines," *Journal of the Royal Statistical Society, Series B*, 63, 3–17.

Kammann, E. and Wand, M. (2003), "Geoadditive models," *Applied Statistics*, 52, 1–18.

MacKay, D. and Takeuchi, R. (1995), "Interpolation models with multiple hyperparameters,".

Paciorek, C. (2003), *Nonstationary Gaussian Processes for Regression and Spatial Modelling*, unpublished Ph.D. dissertation, Carnegie Mellon University, Department of Statistics.

Rasmussen, C. and Ghahramani, Z. (2002), "Infinite mixtures of Gaussian process experts," in *Advances* *in Neural Information Processing Systems 14*, eds. T. G. Dietterich, S. Becker, and Z. Ghahramani, Cambridge, Massachusetts: MIT Press.

Schmidt, A. and O'Hagan, A. (2000), "Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations," Technical Report 498/00, University of Sheffield, Department of Probability and Statistics.

Schoenberg, I. (1938), "Metric spaces and completely monotone functions," *Ann. of Math.*, 39, 811–841.

Seeger, M. and Williams, C. (2003), "Fast forward selection to speed up sparse Gaussian process regression," in *Workshop on AI and Statistics 9*.

Smola, A. and Bartlett, P. (2001), "Sparse greedy Gaussian process approximation," in *Advances in Neural Information Processing Systems 13*, eds. T. Leen, T. Dietterich, and V. Tresp, Cambridge, Massachusetts: MIT Press.

Tresp, V. (2001), "Mixtures of Gaussian Processes," in *Advances in Neural Information Processing Systems 13*, eds. T. K. Leen, T. G. Dietterich, and V. Tresp, MIT Press, pp. 654–660.

Wikle, C. (2002), "Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains," in *Spatial Cluster Modelling*, eds. A. Lawson and D. Denison, Chapman & Hall, pp. 199–209.