Nonstationary Covariance Functions for Gaussian Process Regression

Christopher J. Paciorek and Mark J. Schervish
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
paciorek@alumni.cmu.edu,mark@stat.cmu.edu

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, outperforming a neural network model and Bayesian free-knot spline models, and competitive with a Bayesian neural network, but is outperformed in one dimension by a state-of-the-art Bayesian free-knot spline model. The model readily generalizes to non-Gaussian data. Use of computational methods for speeding GP fitting may allow for implementation of the method on larger datasets.

1 Introduction

Gaussian processes (GPs) have been used successfully for regression and classification tasks. Standard GP models use a stationary covariance, in which the covariance between any two points is a function of Euclidean distance. However, stationary GPs fail to adapt to variable smoothness in the function of interest [1, 2]. This is of particular importance in geophysical and other spatial datasets, in which domain knowledge suggests that the function may vary more quickly in some parts of the input space than in others. For example, in mountainous areas, environmental variables are likely to be much less smooth than in flat regions. Spatial statistics researchers have made some progress in defining nonstationary covariance structures for kriging, a form of GP regression. We extend the nonstationary covariance structure of [3], of which [1] gives a special case, to a class of nonstationary covariance functions. The class includes a Matérn form, which in contrast to most covariance functions has the added flexibility of a parameter that controls the differentiability of sample functions drawn from the GP distribution. We use the nonstationary covariance structure for one, two, and three dimensional input spaces in a standard GP regression model, as done previously only for one-dimensional input spaces [1].

The problem of variable smoothness has been attacked in spatial statistics by mapping the original input space to a new space in which stationarity is assumed, but research has focused on multiple noisy replicates of the regression function with no development nor assessment of the method in the standard regression setting [4, 5]. The issue has been addressed in regression spline models by choosing the knot locations during the fitting [6] and in smoothing splines by choosing an adaptive penalizer on the integrated squared derivative [7]. The general approach in spline and other models involves learning the underlying basis functions, either explicitly or implicitly, rather than fixing the functions in advance. One alternative to a nonstationary GP model is mixtures of stationary GPs [8, 9]. Such methods adapt to variable smoothness by using different stationary GPs in different parts of the input space. The main difficulty is that the class membership is a function of the inputs; this involves additional unknown functions in the hierarchy of the model. One possibility is to use stationary GPs for these additional unknown functions [8], while [9] reduce computational complexity by using a local estimate of the class membership, but do not know if the resulting model is well-defined probabilistically. While the mixture approach is intriguing, neither of [8, 9] compare their model to other methods. In our model, there are unknown functions in the hierarchy of the model that determine the nonstationary covariance structure. We choose to fully model the functions as Gaussian processes themselves, but recognize the computational cost and suggest that simpler representations are worth investigating.

2 Covariance functions and sample function differentiability

The covariance function is crucial in GP regression because it controls how much the data are smoothed in estimating the unknown function. GP distributions are distributions over functions; the covariance function determines the properties of sample functions drawn from the distribution. The stochastic process literature gives conditions for determining sample function properties of GPs based on the covariance function of the process, summarized in [10] for several common covariance functions. Stationary, isotropic covariance functions are functions only of Euclidean distance, τ. Of particular note, the squared exponential (also called the Gaussian) covariance function, $C(\tau) = \sigma^2 \exp\left(-(\tau/\kappa)^2\right)$, where σ^2 is the variance and κ is a correlation scale parameter, has sample functions with infinitely many derivatives. In contrast, spline regression models have sample functions that are typically only twice differentiable. In addition to being of theoretical concern from an asymptotic perspective [11], other covariance forms might better fit real data for which it is unlikely that the unknown function is so highly differentiable. In spatial statistics, the exponential covariance, $C(\tau) = \sigma^2 \exp\left(-\tau/\kappa\right)$, is commonly used, but this form gives sample functions that, while continuous, are not differentiable. Recent work in spatial statistics has focused on the Matérn form, $C(\tau) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}\tau/\kappa\right)^{\nu} K_{\nu}\left(2\sqrt{\nu}\tau/\kappa\right)$, where $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind, whose order is the differentiability parameter, $\nu > 0$. This form has the desirable property that sample functions are $\lfloor \nu - 1 \rfloor$ times differentiable. As $\nu \rightarrow \infty$, the Matérn approaches the squared exponential form, while for $\nu = 0.5,$ the Matérn takes the exponential form. Standard covariance functions require one to place all of one's prior probability on a particular degree of differentiability; use of the Matérn allows one to more accurately, yet easily, express prior lack of knowledge about sample function differentiability. One application for which this may be of particular interest is geophysical data.

[12] suggest using the squared exponential covariance but with anisotropic distance, $\tau(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j})^T \Delta^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j})}$, where Δ is an arbitrary positive definite matrix, rather than the standard diagonal matrix. This allows the GP model to more easily model interactions between the inputs. The nonstationary covariance function we introduce next builds on this more general form.

3 Nonstationary covariance functions

One nonstationary covariance function, introduced by [3], is $C(\boldsymbol{x_i}, \boldsymbol{x_j}) = \int_{\Re^2} k_{\boldsymbol{x_i}}(\boldsymbol{u}) k_{\boldsymbol{x_j}}(\boldsymbol{u}) d\boldsymbol{u}$, where $\boldsymbol{x_i}$, $\boldsymbol{x_j}$, and $\boldsymbol{u}$ are locations in $\Re^2$, and $k_{\boldsymbol{x}}(\cdot)$ is a kernel function centered at $\boldsymbol{x}$. One can show directly that $C(\boldsymbol{x_i}, \boldsymbol{x_j})$ is positive definite in $\Re^p, p = 1, 2, \ldots$, [10]. For Gaussian kernels, the covariance takes the simple form,

$$C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sigma^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left| \left(\Sigma_i + \Sigma_j \right) / 2 \right|^{-\frac{1}{2}} \exp\left(-Q_{ij} \right), \tag{1}$$

with quadratic form

$$Q_{ij} = (\boldsymbol{x_i} - \boldsymbol{x_j})^T \left(\left(\Sigma_i + \Sigma_j \right) / 2 \right)^{-1} (\boldsymbol{x_i} - \boldsymbol{x_j}), \tag{2}$$

where Σ_i, which we call the kernel matrix, is the covariance matrix of the Gaussian kernel at $\boldsymbol{x_i}$. The form (1) is a squared exponential correlation function, but in place of a fixed matrix, Δ, in the quadratic form, we average the kernel matrices for the two locations. The evolution of the kernel matrices in space produces nonstationary covariance, with kernels that drop off quickly producing locally short correlation scales. Independently, [1] derived a special case in which the kernel matrices are diagonal. Unfortunately, so long as the kernel matrices vary smoothly in the input space, sample functions from GPs with the covariance (1) are infinitely differentiable [10], just as for the stationary squared exponential.

To generalize (1) and introduce functions for which sample path differentiability varies, we extend (1) as proven in [10]:

Theorem 1 *Let Q_{ij} be defined as in (2). If a stationary correlation function, $R^S(\tau)$, is positive definite on $\Re^p$ for every $p = 1, 2, \ldots$, then*

$$R^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}} \left| \left(\Sigma_i + \Sigma_j \right) / 2 \right|^{-\frac{1}{2}} R^S\left(\sqrt{Q_{ij}} \right) \tag{3}$$

is a nonstationary correlation function, positive definite on $\Re^p$, $p = 1, 2, \ldots$.

One example of nonstationary covariance functions constructed in this way is a nonstationary version of the Matérn covariance,

$$C^{NS}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{\sigma^2 |\Sigma_i|^{\frac{1}{4}} |\Sigma_j|^{\frac{1}{4}}}{\Gamma(\nu) 2^{\nu - 1}} \left| \frac{\Sigma_i + \Sigma_j}{2} \right|^{-\frac{1}{2}} \left(2\sqrt{\nu Q_{ij}} \right)^{\nu} K_{\nu} \left(2\sqrt{\nu Q_{ij}} \right). \tag{4}$$

Provided the kernel matrices vary smoothly in space, the sample function differentiability of the nonstationary form follows that of the stationary form, so for the nonstationary Matérn, the sample function differentiability increases with ν [10].

4 Bayesian regression model and implementation

Assume independent observations, $Y_1, \ldots, Y_n$, indexed by a vector of input or feature values, $\boldsymbol{x_i} \in \Re^P$, with $Y_i \sim \mathcal{N}(f(\boldsymbol{x_i}), \eta^2)$, where η^2 is the noise variance. Specify a Gaussian process prior, $f(\cdot) \sim \text{GP}\left(\mu_f, C_f^{NS}(\cdot, \cdot) \right)$, where $C_f^{NS}(\cdot, \cdot)$ is the nonstationary Matérn covariance function (4) constructed from a set of Gaussian kernels as described below. For the differentiability parameter, we use the prior, $\nu_f \sim \text{U}(0.5, 30)$, which varies between non-differentiability (0.5) and high differentiability. We use proper, but diffuse, priors for μ_f, σ_f^2, and η^2.The main challenge is to parameterize the kernel matrices, since their evolution determines how quickly the covariance structure changes in the input space and the degree to which the model adapts to variable smoothness in the unknown function. In many problems, it seems natural that the covariance structure would evolve smoothly; if so, the differentiability of the regression function will be determined by ν_f.

We put a prior distribution on the kernel matrices as follows. Any location in the input space, $\boldsymbol{x_i}$, has a Gaussian kernel with mean $\boldsymbol{x_i}$ and covariance (kernel) matrix, Σ_i. When the input space is one-dimensional, each kernel 'matrix' is just a scalar, the variance of the kernel, and we use a stationary Matérn GP prior on the log variance so that the variances evolve smoothly in the input space. Next consider multi-dimensional input spaces; since there are (implicitly) kernel matrices at each location in the input space, we have a multivariate process, the matrix-valued function, $\Sigma(\cdot)$. Parameterizing positive definite matrices as a function of the input space is a difficult problem; see [7]. We use the spectral decomposition of an individual covariance matrix, Σ_i,

$$\Sigma_i = \Gamma(\gamma_1(\boldsymbol{x_i}), \ldots, \gamma_Q(\boldsymbol{x_i}))D(\lambda_1(\boldsymbol{x_i}), \ldots, \lambda_P(\boldsymbol{x_i}))\Gamma(\gamma_1(\boldsymbol{x_i}), \ldots, \gamma_Q(\boldsymbol{x_i}))^T, \qquad (5)$$

where D is a diagonal matrix of eigenvalues and Γ is an eigenvector matrix constructed as described below. $\lambda_p(\cdot), p = 1, \ldots, P$, and $\gamma_q(\cdot), q = 1, \ldots, Q$, which are functions on the input space, construct $\Sigma(\cdot)$. We will refer to these as the eigenvalue and eigenvector processes, and to them collectively as the eigenprocesses. Let $\phi(\cdot) \in \{\log(\lambda_1(\cdot)), \ldots, \log(\lambda_P(\cdot)), \gamma_1(\cdot), \ldots, \gamma_Q(\cdot)\}$ denote any one of these eigenprocesses. To have the kernel matrices vary smoothly, we ensure that their eigenvalues and eigenvectors vary smoothly by taking each $\phi(\cdot)$ to have a GP prior with a single stationary, anisotropic Matérn correlation function, common to all the processes and described later. Using a shared correlation function gives us smoothly-varying kernels, while limiting the number of parameters. We force the eigenprocesses to be very smooth by fixing $\nu = 30$. We do not let ν vary, because it should have minimal impact on the regression estimate and is not well-informed by the data.

Parameterizing the eigenvectors of the kernel matrices using Givens angles, with each angle a function on $\Re^P$, the input space, is difficult, because the angle functions have range $[0, 2\pi) \equiv S^1$, which is not compatible with the range of a GP. To avoid this, we overparameterize the eigenvectors, using $Q = P(P-1)/2 + P - 1$ Gaussian processes, $\gamma_q(\cdot)$, that determine the directions of a set of orthogonal vectors. Here, we demonstrate the construction of the eigenvectors for $\boldsymbol{x_i} \in \Re^2$ and $\boldsymbol{x_i} \in \Re^3$; a similar approach, albeit with more parameters, applies to higher-dimensional spaces, but is probably infeasible in dimensions larger than five or so. In $\Re^3$, we construct an eigenvector matrix for an individual location as $\Gamma = \Gamma_3\Gamma_2$, where

$$\Gamma_3 = \begin{pmatrix} \frac{a}{l_{abc}} & \frac{-b}{l_{ab}} & \frac{-ac}{l_{ab}l_{abc}} \\ \frac{b}{l_{abc}} & \frac{a}{l_{ab}} & \frac{-bc}{l_{ab}l_{abc}} \\ \frac{c}{l_{abc}} & 0 & \frac{l_{ab}}{l_{abc}} \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{u}{l_{uv}} & \frac{-v}{l_{uv}} \\ 0 & \frac{v}{l_{uv}} & \frac{u}{l_{uv}} \end{pmatrix}.$$

The elements of Γ_3 are functions of three random variables, $\{A, B, C\}$, where $l_{abc} = \sqrt{a^2 + b^2 + c^2}$ and $l_{ab} = \sqrt{a^2 + b^2}$. $(\Gamma_3)_{32} = 0$ is a constraint that saves a degree of freedom for the two-dimensional subspace orthogonal to Γ_3. The elements of Γ_2 are based on two random variables, U and V. To have the matrices, $\Sigma(\cdot)$, vary smoothly in space, a, b, c, u and v, are the values of the processes, $\gamma_1(\cdot), \ldots, \gamma_5(\cdot)$ at the input of interest.

One can integrate $\boldsymbol{f}$, the function evaluated at the inputs, out of the GP model. In the stationary GP model, the marginal posterior contains a small number of hyperparameters to either optimize or sample via MCMC. In the nonstationary case, the presence of the additional GPs for the kernel matrices (5) precludes straightforward optimization, leaving MCMC. For each of the eigenprocesses, we reparameterize the vector, $\boldsymbol{\phi}$, of values of the process at the input locations, $\boldsymbol{\phi} = \mu_\phi + \sigma_\phi L(\Delta(\boldsymbol{\theta}))\boldsymbol{\omega}_\phi$, where $\boldsymbol{\omega}_\phi \sim \mathcal{N}(0, I)$ a priori and L is a matrix defined below. We sample μ_ϕ, σ_ϕ, and $\boldsymbol{\omega}_\phi$ via Metropolis-Hastings separately for each eigenprocess. The parameter vector $\boldsymbol{\theta}$, involving P correlation scale parameters and $P(P-1)/2$ Givens angles, is used to construct an anisotropic distance matrix, $\Delta(\boldsymbol{\theta})$, shared by the $\boldsymbol{\phi}$ vectors, creating a stationary, anisotropic correlation structure common to all the eigenprocesses. $\boldsymbol{\theta}$ is also sampled via Metropolis-Hastings. $L(\Delta(\boldsymbol{\theta}))$ is a generalized Cholesky decomposition of the correlation matrix shared by the $\boldsymbol{\phi}$ vectors that deals

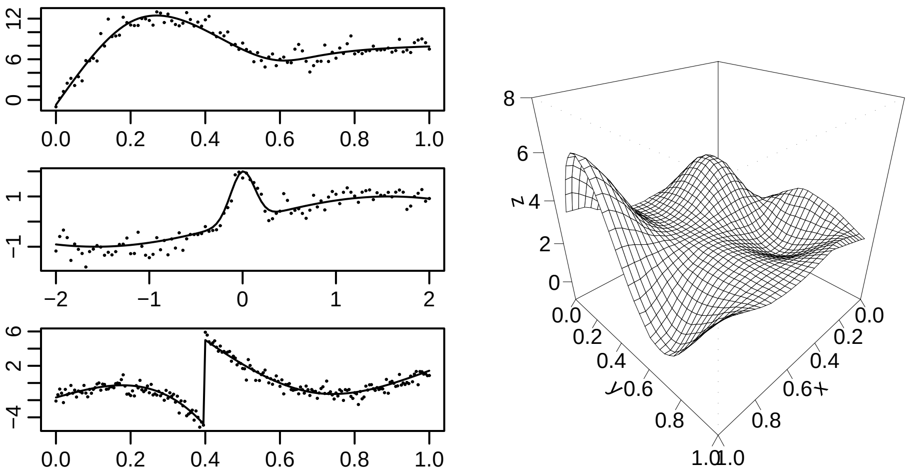

Figure 1: On the left are the three test functions in one dimension, with one simulated set of observations (of the 50 used in the evaluation), while the right shows the test function with two inputs.

with numerically singular correlation matrices by setting the ith column of the matrix to all zeroes when ϕ_i is numerically a linear combination of $\phi_1, \ldots, \phi_{i-1}$ [13]. One never calculates $L(\Delta(\boldsymbol{\theta}))^{-1}$ or $|L(\Delta(\boldsymbol{\theta}))|$, which are not defined, and does not need to introduce jitter, and therefore discontinuity in $\phi(\cdot)$, into the covariance structure.

5 Experiments

For one-dimensional functions, we compare the nonstationary GP method to a stationary GP model[1], two neural network implementations[2] , and Bayesian adaptive regression splines (BARS), a Bayesian free-knot spline model that has been very successful in comparisons in the statistical literature [6]. We use three test functions [6]: a smoothly-varying function, a spatially inhomogeneous function, and a function with a sharp jump (Figure 1a). For each, we generate 50 sets of noisy data and compare the models using the means, averaged over the 50 sets, of the standardized MSE, $\sum_i(\hat{f}_i - f_i)^2/\sum_i(f_i - \bar{f})^2$, where $\hat{f}_i$ is the posterior mean at x_i, and $\bar{f}$ is the mean of the true values. In the non-Bayesian neural network model, $\hat{f}_i$ is the fitted value and, as a simplification, we use a network with the optimal number of hidden units (3, 3, and 8 for the three functions), thereby giving an overly optimistic assessment of the performance. To avoid local minima, we used the network fit that minimized the MSE (relative to the data, with y_i in place of f_i in the expression for MSE) over five fits with different random seeds.

For higher-dimensional inputs, we compare the nonstationary GP to the stationary GP, the neural network models, and two free-knot spline methods, Bayesian multivariate linear splines (BMLS) [14] and Bayesian multivariate automatic regression splines (BMARS) [15], a Bayesian version of MARS [16]. We choose to compare to neural networks and

[1] We implement the stationary GP model by replacing $C_f^{NS}(\cdot,\cdot)$ with the Matérn stationary correlation, still using a differentiability parameter, ν_f, that is allowed to vary.

[2] For a non-Bayesian model, we use the implementation in the statistical software R, which fits a multilayer perceptron with one hidden layer. For a Bayesian version, results from R. Neal's FBM software were kindly provided by A. Vehtari.

Table 1: Mean (over 50 data samples) and 95% confidence interval for standardized MSE for the five methods on the three test functions with one-dimensional input.

Method	Function 1	Function 2	Function 3
Stat. GP	.0083 (.0073,.0093)	.026 (.024,.029)	.071 (.067,.074)
Nonstat. GP	.0083 (0.0073,.0093)	.015 (.013,.016)	.026 (.021,.030)
BARS	.0081 (.0071,.0092)	.012 (.011,.013)	.0050 (.0043,.0056)
Bayes. neural net.	.0082 (.0072,.0093)	.011 (.010,.014)	.015 (.014,.016)
neural network	.0108 (.0095,.012)	.013 (.012,.015)	.0095 (.0086,.010)

splines, because they are popular and these particular implementations have the ability to adapt to variable smoothness. BMLS uses piecewise, continuous linear splines, while BMARS uses tensor products of univariate splines; both are fit via reversible jump MCMC. We use three datasets, the first a function with two inputs [14] (Figure 1b), for which we use 225 training inputs and test on 225 inputs, for each of 50 simulated datasets. The second is a real dataset of air temperature as a function of latitude and longitude [17] that allows assessment on a spatial dataset with distinct variable smoothness. We use a 109 observation subset of the original data, focusing on the Western hemisphere, $222.5^\circ - 322.5^\circ$ E and 62.5°S-82.5°N and fit the models on 54 splits with 107 training examples and two test examples and one split with 108 training examples and one test example, thereby including each data point as a test point once. The third is a real dataset of 111 daily measurements of ozone [18] included in the S-plus statistical software. The goal is to predict the cube root of ozone based on three features: radiation, temperature, and wind speed. We do 55 splits with 109 training examples and two test examples and one split of 110 training examples and one test example. For the non-Bayesian neural network, 10, 50, and 3 hidden units were optimal for the three datasets, respectively.

Table 1 shows that the nonstationary GP does as well or better than the stationary GP, but that BARS does as well or better than the other methods on all three datasets with one input. Part of the difficulty for the nonstationary GP with the third function, which has the sharp jump, is that our parameterization forces smoothly-varying kernel matrices, which prevents our particular implementation from picking up sharp jumps. A potential improvement would be to parameterize kernel matrices that do not vary so smoothly. Table 2 shows that for the known function on two dimensions, the GP models outperform both the spline models and the non-Bayesian neural network, but not the Bayesian network. The stationary and nonstationary GPs are very similar, indicative of the relative homogeneity of the function. For the two real datasets, the nonstationary GP model outperforms the other methods, except the Bayesian network on the temperature dataset. Predictive density calculations that assess the fits of the functions drawn during the MCMC are similar to the point estimate MSE calculations in terms of model comparison, although we do not have predictive density values for the non-Bayesian neural network implementation.

6 Non-Gaussian data

We can model non-Gaussian data, using the usual extension from a linear model to a generalized linear model, for observations, $Y_i \sim \mathrm{D}\left(g\left(f\left(\boldsymbol{x_i}\right)\right)\right)$, where $\mathrm{D}(\cdot)$ $(g(\cdot))$ is an appropriate distribution (link) function, such as the Poisson (log) for count data or the binomial (logit) for binary data. Take $f(\cdot)$ to have a nonstationary GP prior; it cannot be integrated out of the model because of the lack of conjugacy, which causes slow MCMC mixing. [10] improves mixing, which remains slow, using a sampling scheme in which the hyperparameters (including the kernel structure for the nonstationarity) are sampled jointly with the function values, $\boldsymbol{f}$, in a way that makes use of information in the likelihood.

Table 2: For test function with two inputs, mean (over 50 data samples) and 95% confidence interval for standardized MSE at 225 test locations, and for the temperature and ozone datasets, cross-validated standardized MSE, for the six methods.

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

* [14] report a value of .07 for a neural network implementation

We fit the model to the Tokyo rainfall dataset [19]. The data are the presence of rainfall greater than 1 mm for every calendar day in 1983 and 1984. Assuming independence between years [19], conditional on $f(\cdot) = \text{logit}(p(\cdot))$, the likelihood for a given calendar day, x_i, is binomial with two trials and unknown probability of rainfall, $p(x_i)$. Figure 2a shows that the estimated function reasonably follows the data and is quite variable because the data in some areas are clustered. The model detects inhomogeneity in the function, with more smoothness in the first few months and less smoothness later (Figure 2b).

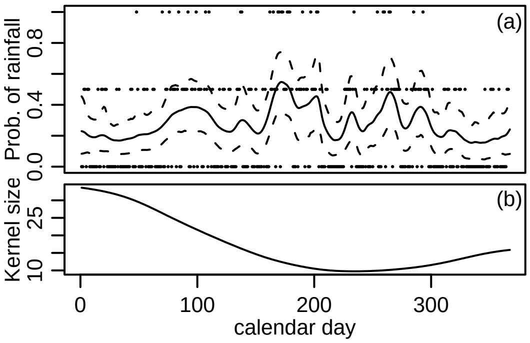

Figure 2. (a) Posterior mean estimate, from nonstationary GP model, of $p(\cdot)$, the probability of rainfall as a function of calendar day, with 95% pointwise credible intervals. Dots are empirical probabilities of rainfall based on the two binomial trials. (b) Posterior geometric mean kernel size (square root of geometric mean kernel eigenvalue).

7 Discussion

We introduce a class of nonstationary covariance functions that can be used in GP 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 outperforms the nonstationary GP, but in higher dimensions, the nonstationary GP outperforms two free-knot spline approaches and a non-Bayesian neural network, while being competitive with a Bayesian neural network. The nonstationary GP may be of particular interest for data indexed by spatial coordinates, where the low dimensionality keeps the parameter complexity manageable.

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 GP priors in the hierarchy of the model to parameterize the nonstationary covariance results in slow computation, limiting the feasibility of the model to approximately $n < 1000$, because the Cholesky decomposition is $O(n^3)$. Our approach provides a general framework; work is ongoing on simpler, more computationally efficient parameterizations of the kernel matrices. Also, approaches that use low-rank approximations to the covariance matrix [20, 21] may speed fitting.

References

[1] M.N. Gibbs. *Bayesian Gaussian Processes for Classification and Regression*. PhD thesis, Univ. of Cambridge, Cambridge, U.K., 1997.

[2] D.J.C. MacKay. Introduction to Gaussian processes. Technical report, Univ. of Cambridge, 1997.

[3] D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 761–768, Oxford, U.K., 1999. Oxford University Press.

[4] A.M. Schmidt and A. O'Hagan. Bayesian inference for nonstationary spatial covariance structure via spatial deformations. Technical Report 498/00, University of Sheffield, 2000.

[5] D. Damian, P.D. Sampson, and P. Guttorp. Bayesian estimation of semi-parametric non-stationary spatial covariance structure. *Environmetrics*, 12:161–178, 2001.

[6] I. DiMatteo, C.R. Genovese, and R.E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88:1055–1071, 2002.

[7] D. MacKay and R. Takeuchi. Interpolation models with multiple hyperparameters, 1995.

[8] Volker Tresp. Mixtures of Gaussian processes. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 654–660. MIT Press, 2001.

[9] C.E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, Massachusetts, 2002. MIT Press.

[10] C.J. Paciorek. *Nonstationary Gaussian Processes for Regression and Spatial Modelling*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2003.

[11] M.L. Stein. *Interpolation of Spatial Data : Some Theory for Kriging*. Springer, N.Y., 1999.

[12] F. Vivarelli and C.K.I. Williams. Discovering hidden features with Gaussian processes regression. In M.J. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, 1999.

[13] J.R. Lockwood, M.J. Schervish, P.L. Gurian, and M.J. Small. Characterization of arsenic occurrence in source waters of U.S. community water systems. *J. Am. Stat. Assoc.*, 96:1184–1193, 2001.

[14] C.C. Holmes and B.K. Mallick. Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society, Series B*, 63:3–17, 2001.

[15] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. Bayesian MARS. *Statistics and Computing*, 8:337–346, 1998.

[16] J.H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–141, 1991.

[17] S.A. Wood, W.X. Jiang, and M. Tanner. Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89:513–528, 2002.

[18] S.M. Bruntz, W.S. Cleveland, B. Kleiner, and J.L. Warner. The dependence of ambient ozone on solar radiation, temperature, and mixing height. In American Meteorological Society, editor, *Symposium on Atmospheric Diffusion and Air Pollution*, pages 125–128, 1974.

[19] C. Biller. Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9:122–140, 2000.

[20] A.J. Smola and P. Bartlett. Sparse greedy Gaussian process approximation. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, Cambridge, Massachusetts, 2001. MIT Press.

[21] M. Seeger and C. Williams. Fast forward selection to speed up sparse Gaussian process regression. In *Workshop on AI and Statistics 9*, 2003.