Bayesian treed Gaussian process models

Robert B. Gramacy and Herbert K. H. Lee
{rbgramacy,herbie}@ams.ucsc.edu
Department of Applied Math & Statistics
University of California, Santa Cruz

Abstract

This paper explores nonparametric and semiparametric nonstationary modeling methodologies that couple stationary Gaussian processes and (limiting) linear models with treed partitioning. Partitioning is a simple but effective method for dealing with nonstationarity. Mixing between full Gaussian processes and simple linear models can yield a more parsimonious spatial model while significantly reducing computational effort. The methodological developments and statistical computing details which make this approach efficient are described in detail. Illustrations of our model are given for both synthetic and real datasets.

Key words: recursive partitioning, nonstationary spatial model, nonparametric regression, Bayesian model averaging

1 Introduction

The Gaussian process (GP) model is a well-established approach for spatial modeling (e.g., Cressie, 1993; Wackernagel, 2003; Banerjee et al., 2003) as well as for modeling other stochastic processes such as computer experiment output (e.g., Sacks et al., 1989; Santner et al., 2003) and unknown functions (O'Hagan, 1991). However, the standard GP model has a number of known drawbacks, as discussed in the next paragraph. Fully flexible nonstationary formulations can be difficult to work with, particularly for larger datasets. In this paper we introduce an expansion of Gaussian processes based on the idea of Bayesian partition models (Chipman et al., 2002; Denison et al., 2002) which is able to address many of these issues.

GPs are conceptually straightforward, can easily accommodate prior knowledge in the form of covariance functions, and can return estimates of predictive confidence. However, we highlight three potential disadvantages to the standard form of a GP. First, inference on the GP scales poorly with the number of data points, typically requiring computing time that grows with the cube of the sample size. Second, GP models are usually stationary in that the same covariance structure is used throughout the entire input space, which

may be too strong an assumption. Third, the estimated predictive error does not directly depend on the locally observed response values. Rather, it depends on them only indirectly through the distance to the spatial locations of the nearest observations and a global measure of error, which also stems from the stationarity assumption. In many real-world spatial and stochastic problems, uncertainty will not be uniform in this sense, but instead, some regions of the space will tend to exhibit larger variability than others. On the other hand, fully nonstationary Bayesian GP models (e.g., Higdon et al., 1999; Paciorek, 2003) can be difficult to fit, and not computationally tractable for more than a relatively small number of datapoints. Further discussion of nonstationary models is deferred until the end of Section 1.2.

All of these shortcomings can be addressed by partitioning the input space into regions, and fitting separate GP models within each region (e.g., Kim et al., 2005). Partitioning allows for the modeling of nonstationary behavior, and can ameliorate some of the computational demands by fitting models to less data. A Bayesian model averaging approach allows for the explicit estimation of predictive uncertainty, which can now vary in a nonstationary manner. Finally, an `R` package with implementations of all of the models discussed in this paper is available at `http://www.cran.r-project.org/src/contrib/Descriptions/tgp.html`.

This paper is in two parts and combines work from multiple research areas in statistics. Section 2 combines stationary Gaussian processes (GPs) and treed partitioning to create treed GPs, implementing a tractable nonstationary model for nonparametric regression. The methodology is illustrated and validated on synthetic data, as well as on a number of classic nonstationary data sets. Section 3 exploits a particular Gaussian process parameterization which implements a semiparametric model that treats some or all of the input dimensions as linear, decoupling them from the GP correlation function. The utility of the this model will be made apparent in its own right, however the greatest "bang for your buck" is obtained when combining it with treed partitioning. The result is a uniquely efficient nonstationary semiparametric regression tool. Section 4 concludes with some discussion.

1.1 Related work

Our approach for nonparametric and semiparametric nonstationary modeling combines standard GPs and treed partitioning within the context of Bayesian hierarchical modeling and model averaging. We assume that the reader is familiar with the basic concepts of Bayesian model averaging (e.g., Hoeting et al., 1999) and inference via Markov chain Monte Carlo (e.g., Gilks et al., 1996). An introduction to GPs and treed partition modeling follows.

1.1.1 Stationary Gaussian Processes

A common specification of stochastic processes for spatial data, of which the stationary Gaussian Process (GP) is a particular case, specifies that model outputs (responses) z depend on multivariate inputs (explanatory variables) $\mathbf{x}$ as $z(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + w(\mathbf{x})$ where $\boldsymbol{\beta}$ are linear trend coefficients, $w(\mathbf{x})$ is a zero mean random process with covariance $C(\mathbf{x}, \mathbf{x}') = \sigma^2 K(\mathbf{x}, \mathbf{x}')$, and $\mathbf{K}$ is a correlation matrix. Low-order polynomials are sometimes used instead of the simple linear mean $\boldsymbol{\beta}^\top \mathbf{x}$, or the mean process is specified generically, often as $m(\mathbf{x}, \boldsymbol{\beta})$ or $m(\mathbf{x})$ (Stein, 1999).

GPs are a popular method for nonparametric regression and classification, with a history going back to Kriging (Matheron, 1963). Consider a training set $D = \{\mathbf{x}_i, z_i\}_{i=1}^N$. The collection of inputs is indicated as the $N \times m_X$ matrix $\mathbf{X}$ whose i^{th} row is $\mathbf{x}_i^\top$. Formally (Stein, 1999), a Gaussian process is a collection of random variables $\mathbf{Z}(\mathbf{x})$ indexed by $\mathbf{x}$ having a jointly Gaussian distribution for any finite subset of indices. It is specified by a mean $\boldsymbol{\mu}(\mathbf{x}) = E\big(\mathbf{Z}(\mathbf{x})\big)$ and correlation function $K(\mathbf{x}, \mathbf{x}') = \frac{1}{\sigma^2} E\big([\mathbf{Z}(\mathbf{x}) - \boldsymbol{\mu}(\mathbf{x})][\mathbf{Z}(\mathbf{x}') - \boldsymbol{\mu}(\mathbf{x}')]^\top\big)$. With data D, the resulting predictive density for a new point $\mathbf{x}$, assuming (for now) that the $m(\mathbf{x}, \boldsymbol{\beta}) = 0$, has a Normal distribution, with mean $\hat{z}(\mathbf{x})$ and variance $\hat{\sigma}^2_{\hat{z}}(\mathbf{x})$:

$$\hat{z}(\mathbf{x}) = \mathbf{k}^\top(\mathbf{x}) \mathbf{K}^{-1} \mathbf{Z} \qquad \hat{\sigma}^2(\mathbf{x}) = \sigma^2 [K(\mathbf{x}, \mathbf{x}) - \mathbf{k}^\top(\mathbf{x}) \mathbf{K}_N^{-1} \mathbf{k}(\mathbf{x})], \tag{1}$$

where $\mathbf{k}^\top(\mathbf{x})$ is the N-vector whose ith component is $K(\mathbf{x}, \mathbf{x}_i)$, $\mathbf{K}$ is the $N \times N$ matrix with i, j element $K(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{Z}$ is the N-vector of observations with i^{th} component z_i. Notice how $\hat{z}(\mathbf{x})$ is linear in the responses ($\mathbf{Z}$), but $\hat{\sigma}^2(\mathbf{x})$ depends only indirectly on $\mathbf{Z}$ through inference on $\mathbf{K}$.

We assume that the GP correlation functions $K(\cdot, \cdot)$ can be written in the form of

$$K(\mathbf{x}_j, \mathbf{x}_k | g) = K^*(\mathbf{x}_j, \mathbf{x}_k) + g \delta_{j,k}. \tag{2}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker delta function, and K^* is the *true* underlying parametric correlation function. The g term in Eq. (2) is referred to as the *nugget* in the geostatistics literature (Matheron, 1963; Cressie, 1993) and sometimes as *jitter* in Machine Learning literature (Neal, 1997). It must always be positive ($g > 0$), and serves two purposes. Primarily, it provides a mechanism for introducing measurement error into the stochastic process. It arises when considering a model of the form: $Z(\mathbf{X}) = m(\mathbf{X}, \boldsymbol{\beta}) + \varepsilon(\mathbf{X}) + \eta(\mathbf{X})$, where $m(\cdot, \cdot)$ is the underlying (often linear) mean process, $\varepsilon(\cdot)$ is a process covariance whose underlying correlation is governed by K^*, and $\eta(\cdot)$ is simply Gaussian noise. Secondarily, though perhaps of equal practical importance, the nugget (or jitter) helps prevent $\mathbf{K}$ from becoming numerically singular. Notational convenience and conceptual congruence motivates referral to $\mathbf{K}$ as a correlation matrix, even though the nugget term (g) forces $K(\mathbf{x}_i, \mathbf{x}_i) > 1$. There is an isomorphic model specification wherein $\mathbf{K}$ depicts honest

correlations. Under both specifications K^* does indeed define a valid correlation matrix $\mathbf{K}^*$.

The correlation functions $K^*(\cdot,\cdot)$ are typically specified through a low dimensional parametric structure, and produce correlation matrices which are symmetric ($\mathbf{K}^* = (\mathbf{K}^*)^\top$) and positive semi-definite ($\mathbf{a}^\top \mathbf{K}^* \mathbf{a} \geq 0$, for any column-vector $\mathbf{a}$). A general reference for families of correlation functions K^* is provided by Abrahamsen (1997). Here we focus on the power family, although our methods are clearly extensible to other families, such as the Matérn class (Matérn, 1986). Further discussion of correlation structures can be found in Adler (1990), Abrahamsen (1997), or Stein (1999). The power family of correlation functions includes the simple isotropic parameterization

$$K^*(\mathbf{x}_j, \mathbf{x}_k | d) = \exp\left\{ -\frac{||\mathbf{x}_j - \mathbf{x}_k||^{p_0}}{d} \right\}, \tag{3}$$

where $d > 0$ is a single range parameter and $p_0 \in (0, 2]$ determines the smoothness of the process. Thus the correlation of two points depends only on the Euclidean distance $||\mathbf{x}_j - \mathbf{x}_k||$ between them. A straightforward enhancement to the isotropic power family is to employ a separate range parameter d_i in each dimension ($i = 1, \ldots, m_X$). The resulting correlation function is still stationary, but no longer isotropic:

$$K^*(\mathbf{x}_j, \mathbf{x}_k | \mathbf{d}) = \exp\left\{ -\sum_{i=1}^{m_X} \frac{|x_{ij} - x_{ik}|^{p_0}}{d_i} \right\}. \tag{4}$$

When the true underlying correlation structure is isotropic, the extra parameters of the separable model represent a sort of overkill, and in terms of efficiency of implementation, a hindrance.

Fitting parameter values for the correlation function can be done either by maximizing the likelihood, integrating over them, or by taking a Bayesian approach. The usual priors (Gelman et al., 1995) can be placed on the linear ($\boldsymbol{\beta}$) part of the model, including a conditionally conjugate inverse-gamma prior for σ^2, allowing Gibbs sampling for these parameters. Priors also need to be placed on the hyperparameters to the correlation structure $\mathbf{K}$. If little is known in advance about the process, then reference priors can be used (Berger et al., 2001). The posteriors can be sampled using the Metropolis-Hastings algorithm.

1.2 Treed Partitioning for Nonstationary Modeling

Many spatial modeling problems require more flexibility than is offered by a stationary GP. One way to achieve a more flexible, nonstationary, process is to use a partition model—a meta-model which somehow divides up the input space and fits different base models to data independently in the regions depicted by the partitions. Treed partitioning is one possible approach.

Treed partition models typically divide up the input space by making binary splits on the value of a single variable (e.g., $x_1 > 0.8$) so that partition boundaries are parallel to coordinate axes. Partitioning

is recursive, so each new partition is a sub-partition of a previous one. For example, a first partition may divide the space in half by whether the first variable is above or below its midpoint. The second partition will then divide only the space below (or above) the midpoint of the first variable, so that there are now three partitions (not four). Since variables may be revisited, there is no loss of generality by using binary splits, as multiple splits on the same variable will be equivalent to a non-binary split. In each partition (leaf of the tree), an independent model is applied. Classification and Regression Trees (CART) (Breiman et al., 1984) are an example of a treed partition model. CART, which fits a constant surface in each leaf, has become popular because of its ease of use, clear interpretation, and ability to provide a good fit in many cases.

The Bayesian approach is straightforward to apply to tree models (Chipman et al., 1998; Denison et al., 1998), provided that one can specify a meaningful prior for the size of the tree. We follow Chipman et al. (1998, 2002) who specify the prior through a tree-generating process. Starting with a null tree (all data in a single partition), a leaf node $\eta \in \mathcal{T}$, representing a region of the input space, splits with probability $a(1+q_\eta)^{-b}$, where q_η is the depth of $\eta \in \mathcal{T}$ and a and b are parameters chosen to give an appropriate size and spread to the distribution of trees. Further details are available in the Chipman et al. papers. A sensible prior might further require that each new region have at least a minimal number of data points to ensure that there is enough data to infer the parameters of the independent models, and we impose this constraint as well. The prior for the splitting process involves first choosing the splitting dimension u from a discrete uniform, and then the split location s is chosen uniformly from a subset of the locations $\mathbf{X}$ in the u^{th} dimension. Integrating out dependence on the tree structure $\mathcal{T}$ can be accomplished via Reversible-Jump MCMC as further described in Section 2.2.2.

Section 2 takes this approach to another level by fitting stationary GPs in each of the leaves of the tree. Our models bear some similarity to those of Kim et al. (2005), who fit separate GPs in each element of a Voronoi tessellation. The treed GP approach is better geared toward problems with a smaller number of distinct partitions, leading to a simpler overall model. Voronoi tessellations allow an intricate partitioning of the space, but have the trade-off of added complexity and can produce a final model that is difficult to interpret. A nice review of Bayesian partition modeling is provided by Denison et al. (2002).

Other approaches to nonstationary modeling include those which use spatial deformations and process convolutions. The idea behind the spatial deformation approach is to map nonstationary inputs in the original, geographical, space into a dispersion space wherein the process is stationary. The approach taken by Sampson and Guttorp (1992) uses thin-plate spline models and multidimensional scaling (MDS) to construct the mapping. Damian et al. (2001) explore a similar methodology from a Bayesian perspective. Schmidt and O'Hagan (2003) also take the Bayesian approach, but put a Gaussian process prior on the mapping.

The process convolution approach (Higdon et al., 1999; Fuentes and Smith, 2001; Paciorek, 2003) proceeds by allowing the convolution kernels $K_{\mathbf{s}}(\cdot)$ to vary in parameterization as a function of their location $\mathbf{s} \in \Re^d$. $K_{\mathbf{s}}$ is treated as an unknown, smooth function of $\mathbf{s}$.

A common theme among such nonstationary models is the introduction of meta-structure which ratchets up the flexibility of the model, ratcheting up the computational demands as well. This is in stark contrast to the treed approach which introduces a structural mechanism, the tree $\mathcal{T}$, that actually reduces the computational burden relative to the base model, e.g., a GP, because smaller matrices are inverted.

2 Treed Gaussian process models

Extending the partitioning ideas of Chipman et al. (1998, 2002) for simple Bayesian treed models, we fit stationary GP models with linear trends independently within each of R regions, $\{r_\nu\}_{\nu=1}^R$, depicted at the leaves of the tree $\mathcal{T}$, instead of constant (1998) or linear (2002) models. The tree is averaged out by integrating over possible trees, using reversible-jump Markov chain Monte Carlo (RJ-MCMC) (Richardson and Green, 1997), with the tree prior specified through a tree-generating process. Prediction is conditioned on the tree structure, and is averaged over in the posterior to get a full accounting of uncertainty.

2.1 Hierarchical Model

A tree $\mathcal{T}$ recursively partitions the input space into into R non-overlapping regions: $\{r_\nu\}_{\nu=1}^R$. Each region r_ν contains data $D_\nu = \{\mathbf{X}_\nu, \mathbf{Z}_\nu\}$, consisting of n_ν observations. Let $m \equiv m_X + 1$ be number of covariates in the design (input) matrix $\mathbf{X}$ plus an intercept. For each region r_ν, the hierarchical generative GP model is

$$\begin{aligned} \mathbf{Z}_\nu|\boldsymbol{\beta}_\nu, \sigma_\nu^2, \mathbf{K}_\nu &\sim N_{n_\nu}(\mathbf{F}_\nu\boldsymbol{\beta}_\nu, \sigma_\nu^2\mathbf{K}_\nu), & \boldsymbol{\beta}_0 &\sim N_m(\boldsymbol{\mu}, \mathbf{B}) \\ \boldsymbol{\beta}_\nu|\sigma_\nu^2, \tau_\nu^2, \mathbf{W}, \boldsymbol{\beta}_0 &\sim N_m(\boldsymbol{\beta}_0, \sigma_\nu^2\tau_\nu^2\mathbf{W}) & \tau_\nu^2 &\sim IG(\alpha_\tau/2, q_\tau/2), \\ \sigma_\nu^2 &\sim IG(\alpha_\sigma/2, q_\sigma/2), & \mathbf{W}^{-1} &\sim W((\rho\mathbf{V})^{-1}, \rho), \end{aligned} \tag{5}$$

with $\mathbf{F}_\nu = (\mathbf{1}, \mathbf{X}_\nu)$, and $\mathbf{W}$ is an $m \times m$ matrix. N, IG, and W are the (Multivariate) Normal, Inverse-Gamma, and Wishart distributions, respectively. Constants $\boldsymbol{\mu}, \mathbf{B}, \mathbf{V}, \rho, \alpha_\sigma, q_\sigma, \alpha_\tau, q_\tau$ are treated as known. The model (5) specifies a multivariate normal likelihood with linear trend coefficients $\boldsymbol{\beta}_\nu$, variance σ_ν^2 and $n_\nu \times n_\nu$ correlation matrix $\mathbf{K}_\nu$. The coefficients $\boldsymbol{\beta}_\nu$ are believed to have come from a common unknown mean $\boldsymbol{\beta}_0$ and region-specific variance $\sigma_\nu^2\tau_\nu^2$. There is no explicit mechanism in the model (5) to ensure that the process near the boundary of two adjacent regions is continuous across the partitions depicted by $\mathcal{T}$. However the model can capture smoothness through model averaging, as will be discussed in Section 2.3. In our work with models for physical processes, we frequently encounter problems with phase transitions where

the response surface is not smooth at the boundary between distinct physical regimes (such as sub-sonic vs. super-sonic flight), so we view the ability to fit a discontinuous surface as a feature of this model.

The GP correlation structure $K_\nu(\mathbf{x}_j, \mathbf{x}_k) = K^*_\nu(\mathbf{x}_j, \mathbf{x}_k) + g_\nu \delta_{j,k}$ generating $\mathbf{K}_\nu$ for each partition r_ν takes K^*_ν to be from the isotropic power family (3), or separable power family (4), with a fixed power p_0, but unknown (random) range and nugget parameters. However, since most of the following discussion holds for K^*_ν generated by other families, as well as for unknown p_0, we shall refer to the correlation parameters indirectly via the resulting correlation matrix $\mathbf{K}$, or function $K(\cdot,\cdot)$. For example, $p(\mathbf{K}_\nu)$ can represent either of $p(d_\nu, g_\nu)$ or $p(\mathbf{d}_\nu, g_\nu)$, etc. Priors which encode a belief that the global covariance structure is nonstationary are chosen for parameters to K^*_ν and g_ν, as further described in Section 3, Equation (20).

2.2 Estimation

The data $D_\nu = \{\mathbf{X}, \mathbf{Z}\}_\nu$ are used to estimate the GP parameters $\boldsymbol{\theta}_\nu \equiv \{\boldsymbol{\beta}, \sigma^2, \mathbf{K}\}_\nu$, for $\nu = 1, \ldots, R$. Conditional on the tree $\mathcal{T}$, the full set of parameters is denoted as $\boldsymbol{\theta} = \boldsymbol{\theta}_0 \cup \bigcup_{\nu=1}^R \boldsymbol{\theta}_\nu$, where $\boldsymbol{\theta}_0 = \{\mathbf{W}, \boldsymbol{\beta}_0, \boldsymbol{\tau}\}$ denotes hyperparameters that are also estimated. Samples from the posterior distribution of $\boldsymbol{\theta}$ are gathered using Markov chain Monte Carlo (MCMC) by first conditioning on the hierarchical priors $\boldsymbol{\theta}_0$ and drawing $\boldsymbol{\theta}_\nu|\boldsymbol{\theta}_0$ for $\nu_1, \ldots, \nu_R$, and then $\boldsymbol{\theta}_0$ is drawn as $\boldsymbol{\theta}_0|\bigcup_{\nu=1}^R \boldsymbol{\theta}_\nu$. Section 2.2.1 gives the details. All parameters can be sampled with Gibbs steps, except those which parameterize the covariance function $K(\cdot,\cdot)$, e.g., $\{d, g\}_\nu$, which require Metropolis-Hastings (MH) draws. Section 2.2.2 shows how RJ-MCMC is used to gather samples from the joint posterior of $(\boldsymbol{\theta}, \mathcal{T})$ by alternately drawing $\boldsymbol{\theta}|\mathcal{T}$ and $\mathcal{T}|\boldsymbol{\theta}$.

2.2.1 GP parameters given a tree ($\mathcal{T}$)

Finding full conditionals is the first step towards efficient sampling. Since parameters associated with the linear trend have conditionally conjugate priors, they can be sampled using Gibbs steps. Some parameters ($\{\mathbf{K}, \sigma^2\}_\nu$) are sampled more efficiently if their full conditionals can be marginalized by analytically integrating out dependence on other parameters. Full derivations are included in Appendix A.1.

The linear regression parameters $\boldsymbol{\beta}_\nu$ and prior mean $\boldsymbol{\beta}_0$ both have conditionally conjugate multivariate normal full conditionals: $\boldsymbol{\beta}_\nu|\text{rest} \sim N_m(\tilde{\boldsymbol{\beta}}_\nu, \sigma^2_\nu \mathbf{V}_{\tilde{\beta}_\nu})$, and $\boldsymbol{\beta}_0|\text{rest} \sim N_m(\tilde{\boldsymbol{\beta}}_0 \boldsymbol{V}_{\tilde{\beta}_0})$, where

$$\mathbf{V}_{\tilde{\beta}_\nu} = (\mathbf{F}_\nu^\top \mathbf{K}_\nu^{-1} \mathbf{F}_\nu + \mathbf{W}^{-1}/\tau^2_\nu)^{-1} \qquad \tilde{\boldsymbol{\beta}}_\nu = \mathbf{V}_{\tilde{\beta}_\nu}(\mathbf{F}_\nu^\top \mathbf{K}_\nu^{-1} \mathbf{Z}_\nu + \mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau^2_\nu). \tag{6}$$

$$\mathbf{V}_{\tilde{\beta}_0} = \left(\mathbf{B}^{-1} + \mathbf{W}^{-1} \textstyle\sum_{i=0}^r (\sigma_\nu \tau_\nu)^{-2}\right)^{-1} \qquad \tilde{\boldsymbol{\beta}}_0 = \mathbf{V}_{\tilde{\beta}_0}\left(\mathbf{B}^{-1}\mu + \mathbf{W}^{-1} \textstyle\sum_{i=1}^r \boldsymbol{\beta}_\nu (\sigma_\nu \tau_\nu)^{-2}\right). \tag{7}$$

The linear variance parameter τ^2 follows the conditionally conjugate inverse-gamma:

$$\tau^2_\nu|\text{rest} \sim IG((\alpha_\tau + m)/2, (q_\tau + b_\nu)/2) \qquad \text{where} \qquad b_\nu = (\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)^\top \mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)/\sigma^2_\nu. \tag{8}$$

The linear model covariance matrix $\mathbf{W}$ follows the conditionally conjugate inverse-Wishart:

$$\mathbf{W}^{-1}|\text{rest} \sim W_m\left(\rho\mathbf{V}+\mathbf{V}_{\hat{T}}, \rho + r\right) \qquad \text{where} \qquad \mathbf{V}_{\hat{T}} = \sum_{i=1}^{r} \frac{1}{(\sigma_\nu\tau_\nu)^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)^\top. \quad (9)$$

Analytically integrating out $\boldsymbol{\beta}_\nu$ and σ_ν^2 gives a marginal posterior for $\mathbf{K}_\nu$ and improves mixing of the Markov chain (Berger et al., 2001). Details are left to Appendix A.2.

$$p(\mathbf{K}_\nu|\mathbf{Z}_\nu, \boldsymbol{\beta}_0, \mathbf{W}, \tau^2) = \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}|\mathbf{K}_\nu||\mathbf{W}|\tau^{2m}}\right)^{\frac{1}{2}} \frac{(q_\sigma/2)^{\alpha_\sigma/2}}{[(q_\sigma+\psi_\nu)/2]^{(\alpha_\sigma+n_\nu)/2}} \frac{\Gamma\left[(\alpha_\sigma+n_\nu)/2\right]}{\Gamma\left[\alpha_\sigma/2\right]} p(\mathbf{K}_\nu), \quad (10)$$

$$\text{where} \qquad \psi_\nu = \mathbf{Z}_\nu^\top \mathbf{K}_\nu^{-1}\mathbf{Z}_\nu + \boldsymbol{\beta}_0^\top \mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau^2 - \tilde{\boldsymbol{\beta}}_\nu^\top \mathbf{V}_{\tilde{\beta}_\nu}^{-1}\tilde{\boldsymbol{\beta}}_\nu. \quad (11)$$

Eq. (10) can be used to iteratively obtain draws for the parameters of $K_\nu(\cdot,\cdot)$ via Metropolis-Hastings (MH), or as part of the acceptance ratio for proposed modifications to $\mathcal{T}$ [see Section 2.2.2]. Many terms in (10) cancel when examining the MH acceptance ratio for $\mathbf{K}_\nu$ in isolation. Dropping constants that would be common in the numerator and denominator results in the simplified expression

$$p(\mathbf{K}_\nu|\mathbf{Z}_\nu, \boldsymbol{\beta}_0, \tau_\nu^2, \mathbf{W}) \propto p(\mathbf{K}_\nu) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{|\mathbf{K}_\nu|}\right)^{\frac{1}{2}} \times \left(\frac{q_\sigma+\psi_\nu}{2}\right)^{-\frac{\alpha_\sigma+n_\nu}{2}}. \quad (12)$$

Any hyperparameters to $K_\nu(\cdot,\cdot)$, e.g., parameters to priors for $\{d, g\}_\nu$ of the isotropic power family, would also require MH draws. Dropping the prior $p(\mathbf{K}_\nu)$ gives an integrated likelihood (Berger et al., 2001).

The conditional distribution of σ_ν^2 with $\boldsymbol{\beta}_\nu$ integrated out is

$$\sigma_\nu^2|d_\nu, g, \boldsymbol{\beta}_0, \mathbf{W} \sim IG((\alpha_\sigma+n_\nu)/2, (q_\sigma+\psi_\nu)/2), \quad (13)$$

which allows Gibbs sampling. The full derivation of (13) is also included in Appendix A.2.

2.2.2 Tree ($\mathcal{T}$)

Integrating out dependence on the tree structure ($\mathcal{T}$) is accomplished by RJ-MCMC. We augment the tree operations of Chipman et al. (1998)—*grow, prune, change, swap*—with a rotate operation. Tree proposals can change the size of the parameter space ($\boldsymbol{\theta}$). Proposals for new parameters (via an increase in the number of partitions R) are drawn from their priors, thus eliminating the Jacobian term usually present in RJ-MCMC. New splits are chosen uniformly from the set of marginalized input locations ($\mathbf{X}$).

Swap and *change* tree operations are straightforward because the number of partitions, and thus parameters, stays the same. A *change* operation proposes moving an existing split-point $\{u, s\}$ to either the next greater or lesser value of s (s_+ or s_-) along the u^{th} column of $\mathbf{X}$. This is accomplished by sampling s' uniformly from the set $\{u_\nu, s_\nu\}_{\nu=1}^{\lceil R/2\rceil} \times \{+,-\}$ which causes the MH acceptance ratio for *change* to reduce

to a simple likelihood ratio since parameters $\boldsymbol{\theta}_r$ in regions r below the split-point $\{u, s'\}$ are held fixed.

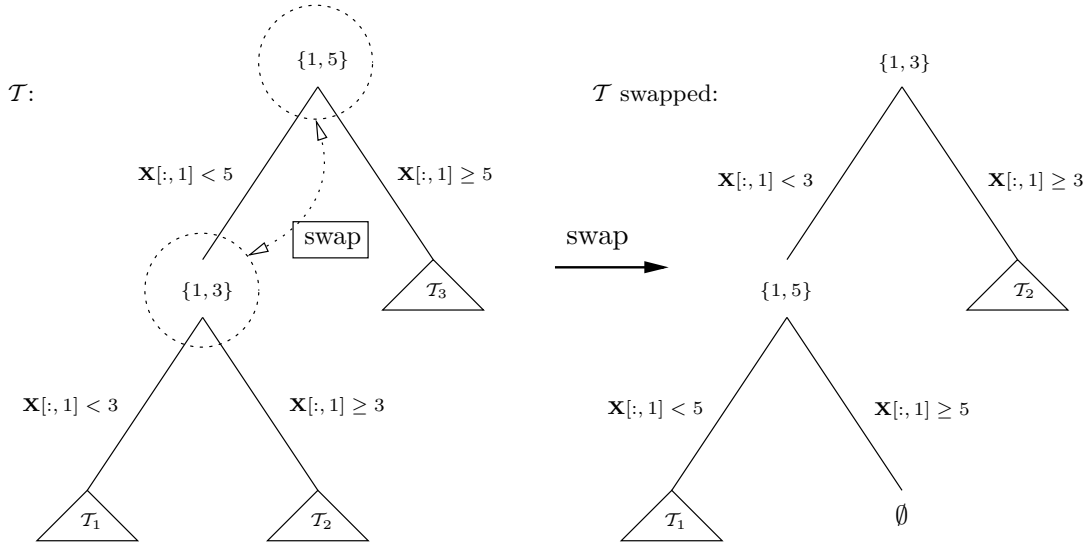

Figure 1: Swapping on the same variable is always rejected because one of the leaves corresponds to an empty region. $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ are arbitrary sub-trees (could be leaves).

A *swap* operation proposes changing the order in which two adjacent parent–child (internal) nodes split up the inputs. An internal parent–child node pair is picked at random from the tree and their splitting rules are swapped. When both child splitting rules are the same, Chipman et al. (1998) propose jointly swapping the parent with both of its children. We found that this situation is rare in practice, especially for continuous explanatory variables. So instead, we supplement *swap* with a modification for our more common situation. Swaps on parent-child internal nodes which split on the same variable cause problems because a child region below both parents becomes empty after the operation. Figure 1 gives an illustration. However, if instead a *rotate* operation from Binary Search Trees (BSTs) is performed, the proposal will almost always accept. Rotations are a way of adjusting the configuration and height of a BST without violating the BST property. *Red-Black Trees* make extensive use of *rotate* operations (Cormen et al., 1990).

In the context of a Bayesian MCMC tree proposal, rotations encourage better mixing of the Markov chain by providing a more dynamic set of candidate nodes for pruning, thereby helping escape local minima in the marginal posterior of $\mathcal{T}$. Figure 2 shows an example of a successful right-rotation where the swap of Figure 1 fails. Since the partitions at the leaves remain unchanged, the likelihood ratio of a proposed rotate is always 1. The only "active" part of the MH acceptance ratio is the prior on $\mathcal{T}$, preferring trees of minimal depth. Still, calculating the acceptance ratio for a *rotate* is non-trivial because the depth of *two* of its sub-trees change. Sub-trees $\mathcal{T}_1$ and $\mathcal{T}_3$ of Figure 2 change depth, either increasing or decreasing respectively, depending on the direction of the rotation. In a right-rotate, nodes in $\mathcal{T}_1$ decrease in depth,

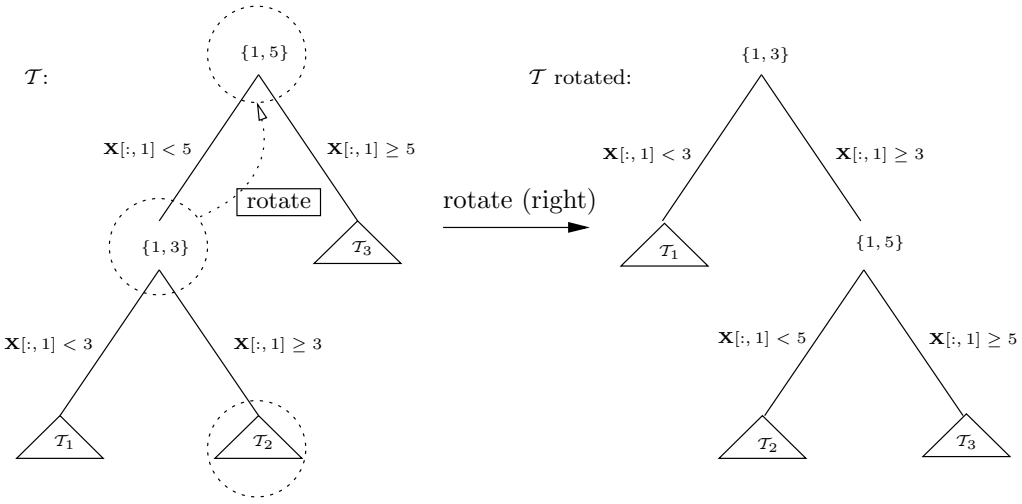

Figure 2: Rotating on the same variable is almost always accepted. $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$ are arbitrary sub-trees (could be leaves).

while those in $\mathcal{T}_3$ increase. The opposite is true for left-rotation. If $I = \{I_i, I_\ell\}$ is the set of nodes (internals and leaves) of $\mathcal{T}_1$ and $\mathcal{T}_3$, before rotation, which increase in depth after rotation, and $D = \{D_i, D_\ell\}$ are those which decrease in depth, then the MH acceptance ratio for a right-rotate is

$$\frac{p(\mathcal{T}^*)}{p(\mathcal{T})} = \frac{p(\mathcal{T}_1^*)p(\mathcal{T}_3^*)}{p(\mathcal{T}_1)p(\mathcal{T}_3)} = \frac{\prod_{\eta\in I_i} a(2+q_\eta)^{-b} \prod_{\eta\in I_\ell} [1-a(2+q_\eta)^{-b}]}{\prod_{\eta\in I_i} a(1+q_\eta)^{-b} \prod_{\eta\in I_\ell} [1-a(1+q_\eta)^{-b}]} \cdot \frac{\prod_{\eta\in D_i} aq_\eta^{-b} \prod_{\eta\in D_\ell} [1-aq_\eta^{-b}]}{\prod_{\eta\in D_i} a(1+q_\eta)^{-b} \prod_{\eta\in D_\ell} [1-a(1+q_\eta)^{-b}]}.$$

The MH acceptance ratio for a left-rotate is analogous.

Grow and *prune* operations are complex because they add or remove partitions, changing the dimension of the parameter space. The first step for either operation is to uniformly select a leaf node (for *grow*), or the parent of a pair of leaf nodes (for *prune*). When a new region r is added, new parameters $\{K(\cdot,\cdot), \tau^2\}_r$ must be proposed, and when a region is taken away the parameters must be absorbed by the parent region, or discarded. When evaluating the MH acceptance ratio the linear model parameters $\{\boldsymbol{\beta}, \sigma^2\}_r$ are integrated out (10). One of the newly grown children is uniformly chosen to receive the correlation function $K(\cdot,\cdot)$ of its parent, essentially inheriting a block from its parent's correlation matrix. To ensure that the resulting Markov chain is ergodic and reversible, the other new sibling draws its $K(\cdot,\cdot)$ from the prior. Symmetrically, *prune* operations randomly select parameters from $K(\cdot,\cdot)$ for the consolidated node from one of the children being absorbed. After accepting a *grow* or *prune*, σ_r^2 can be drawn from its marginal posterior, with $\boldsymbol{\beta}_r$ integrated out (13), followed by draws for $\boldsymbol{\beta}_r$ and the rest of the parameters in the r^{th} region.

Let $\{\mathbf{X}, \mathbf{Z}\}$ be the data at the new parent node η at depth q_η, and $\{\mathbf{X}_1, \mathbf{Z}_1\}$ and $\{\mathbf{X}_2, \mathbf{Z}_2\}$ be the partitioned child data at depth $q_\eta + 1$ created by a new split $\{u, s\}$. Also, let $\mathcal{P}$ be the set of pruneable nodes

of $\mathcal{T}$, $\mathcal{G}$ the set of growable nodes. If $\mathcal{P}'$ is the set of prunable nodes in $\mathcal{T}'$ after the (successful) grow at η and the parent of η was prunable in $\mathcal{T}$, then $|P'| = |P|$. Otherwise $|P'| = |P| + 1$. The MH acceptance ratio for *grow* is:

$$\frac{|\mathcal{P}'|}{|\mathcal{G}|} \times \frac{a(1+q_\eta)^{-b}(1-a(2+q_\eta)^{-b})^2}{1-a(1+q_\eta)^{-b}} \times \frac{p(\mathbf{K}_1, |\mathbf{Z}_1, \boldsymbol{\beta}_0, \tau_1^2, \mathbf{W}) p(\mathbf{K}_2|\mathbf{Z}_2, \boldsymbol{\beta}_0, \tau_2^2, \mathbf{W})}{p(\mathbf{K}|\mathbf{Z}, \boldsymbol{\beta}_0, \tau^2, \mathbf{W})}. \tag{14}$$

The *prune* operation is analogous. Note that in (14) the posteriors $p(\mathbf{K}|\mathbf{Z}, \boldsymbol{\beta}_0, \tau^2, \mathbf{W})$, $p(\mathbf{K}_1|\mathbf{Z}_1, \boldsymbol{\beta}_0, \tau_1^2, \mathbf{W})$ and $p(\mathbf{K}_2|\mathbf{Z}_2, \boldsymbol{\beta}_0, \tau_2^2, \mathbf{W})$ must be evaluated using the formula in (10), *not* the simplified one in (12), because the terms canceled from (10) do not occur the same number of times in the numerator and denominator.

2.3 Treed GP Prediction

Prediction under the above GP model, called Kriging (Matheron, 1963) in the geostatistics community, is straightforward (Hjort and Omre, 1994). The predicted value of $z(\mathbf{x} \in r_\nu)$ is normally distributed with

$$\begin{aligned} \text{mean} \qquad \hat{z}(\mathbf{x}) &= E(\mathbf{Z}(\mathbf{x})|\text{ data}, \mathbf{x} \in D_\nu) \\ &= \mathbf{f}^\top(\mathbf{x})\tilde{\boldsymbol{\beta}}_\nu + \mathbf{k}_\nu(\mathbf{x})^\top \mathbf{K}_\nu^{-1}(\mathbf{Z}_\nu - \mathbf{F}_\nu \tilde{\boldsymbol{\beta}}_\nu), \qquad (15) \\ \text{and variance} \qquad \hat{\sigma}(\mathbf{x})^2 &= \text{Var}(\mathbf{z}(\mathbf{x})|\text{ data}, \mathbf{x} \in D_\nu) \\ &= \sigma_\nu^2[\kappa(\mathbf{x},\mathbf{x}) - \mathbf{q}_\nu^\top(\mathbf{x})\mathbf{C}_\nu^{-1}\mathbf{q}_\nu(\mathbf{x})], \qquad (16) \end{aligned}$$

$$\text{where} \qquad \mathbf{C}_\nu^{-1} = (\mathbf{K}_\nu + \tau_\nu^2 \mathbf{F}_\nu \mathbf{W} \mathbf{F}_\nu^\top)^{-1} \qquad \mathbf{q}_\nu(\mathbf{x}) = \mathbf{k}_\nu(\mathbf{x}) + \tau_\nu^2 \mathbf{F}_\nu \mathbf{W}_\nu \mathbf{f}(\mathbf{x}) \tag{17}$$
$$\kappa(\mathbf{x},\mathbf{y}) = K_\nu(\mathbf{x},\mathbf{y}) + \tau_\nu^2 \mathbf{f}^\top(\mathbf{x}) \mathbf{W} \mathbf{f}(\mathbf{y})$$

with $\mathbf{f}^\top(\mathbf{x}) = (1, \mathbf{x}^\top)$, and $\mathbf{k}_\nu(\mathbf{x})$ is a $n_\nu-$vector with $\mathbf{k}_{\nu,j}(\mathbf{x}) = K_\nu(\mathbf{x}, \mathbf{x}_j)$, for all $\mathbf{x}_j \in \mathbf{X}_\nu$. Notice that the predictive mean equations use $\tilde{\boldsymbol{\beta}}_\nu$, the posterior mean estimate of $\boldsymbol{\beta}_\nu$. Using $\boldsymbol{\beta}_\nu$ instead requires the predictive variance relation in (1).

Conditional on a particular tree ($\mathcal{T}$), the posterior predictive surface described in Eqs. (15–16) is discontinuous across the partition boundaries of $\mathcal{T}$. However, in the aggregate of samples collected from the joint posterior distribution of $\{\mathcal{T}, \boldsymbol{\theta}\}$, the mean tends to smooth out near likely partition boundaries as the tree operations *grow, prune, change*, and *swap* integrate over trees and GPs with larger posterior probability. Uncertainty in the posterior for $\mathcal{T}$ translates into higher posterior predictive uncertainty near region boundaries. When the data actually indicate a non-smooth process, the treed GP retains the flexibility necessary to model discontinuities.

2.4 Implementation

The treed GP model is coded in a mixture of `C` and `C++`, using `C++` for the tree structure and `C` for the GP at each leaf of $\mathcal{T}$. The `C` code can interface with either standard platform-specific `Fortran BLAS/Lapack` libraries for the necessary linear algebra routines, or link to those automatically configured for fast execution on a variety of platforms via the `ATLAS` library (Whaley and Petitet, 2004). In most cases, the `ATLAS` implementation is significantly faster than standard `BLAS/Lapack`. The code has been tested on Unix (`Solaris, Linux, FreeBSD, OSX`) and Windows (2000, XP) platforms. To improve usability, the routines have been wrapped up in an intuitive `R` interface, and are available on CRAN (R Development Core Team, 2004) at `http://www.cran.r-project.org/src/contrib/Descriptions/tgp.html`, as a package called `tgp`.

It is useful to first translate and re-scale the input dataset ($\mathbf{X}$) so that it lies in an $\Re^{m_X}$ dimensional unit cube. This makes it easier to construct prior distributions for the width parameters to the correlation function $K(\cdot,\cdot)$. Conditioning on $\mathcal{T}$, proposals for all parameters which require MH sampling are taken from a uniform "sliding window" centered around the location of the last accepted setting. For example, a proposed a new nugget parameter g_ν to the correlation function $K(\cdot,\cdot)$ in region r_ν would go as $g_\nu^* \sim \mathrm{Unif}\,(3g_\nu/4, 4g_\nu/3)$. Calculating the forwards and backwards proposal probabilities for the MH acceptance ratio is straightforward.

After conditioning on $\{\mathcal{T}, \boldsymbol{\theta}\}$, prediction can be parallelized by using a producer/consumer model. This allows the use of `PThreads` in order to take advantage of multiple processors, and get speed-ups of at least a factor of two, which is helpful as multi-processor machines become commonplace. Parallel sampling of the posterior of $\boldsymbol{\theta}|\mathcal{T}$ for each of the $\{\theta_\nu\}_{\nu=1}^R$ is also possible.

2.5 Illustration & Experimentation

In this section the treed GP model is illustrated on synthetic and real world data. To keep things simple, for now, the isotropic power family (3) correlation function ($p_0 = 2$) is chosen for $K^*(\cdot,\cdot|d)$ in the following experiments, with range parameter d, combined with nugget g to form $K(\cdot,\cdot|d,g)$.

2.5.1 1-d Synthetic Sinusoidal data

Consider 1-dimensional simulated data on the input space $[0, 20]$. The true response comes partly from Higdon (2002), augmented to include a linear region. Eq. (18) gives a formula describing the data. This dataset typifies the type of nonstationary response surface that the treed GP model was designed to exploit. Zero mean Gaussian noise with sd = 0.1 is added to the response to keep things interesting.

$$z(x) = \begin{cases} \sin\left(\frac{\pi x}{5}\right) + \frac{1}{5}\cos\left(\frac{4\pi x}{5}\right) & x < 10 \\ x/10 - 1 & \text{otherwise} \end{cases} \tag{18}$$

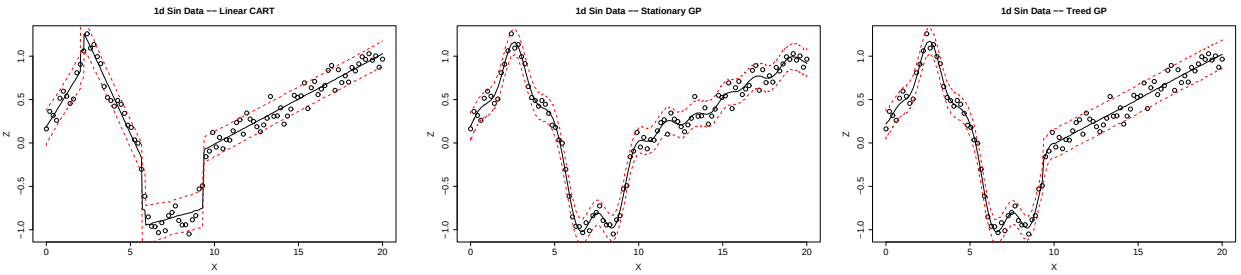

Figure 3: Comparison between Bayesian linear CART (*left*), stationary GP (*middle*) and the treed GP model (*right*), for the 1-d Sine data.

Figure 3 shows the posterior predictive surfaces of three regression models for comparison based on samples obtained at $N = 200$ evenly-spaced input locations—mean in solid black, and 95% intervals in dashed-red. The *left* panel is from a Bayesian Linear CART model (Chipman et al., 2002), which does well in the linear region, but comes up short in the sinusoidal region. The *middle* panel is from a stationary GP model which is heavily influenced by the sinusoidal region, and consequently fits it well, but is unable to model the more smooth linear process. This is because nonstationarity in the data cannot be captured by a stationary (or homogeneous) correlation structure. The *right* panel shows the best of both worlds: a treed GP where correlation is lower in the sinusoidal region and higher in the linear region.

2.5.2 2-d Synthetic Exponential data

Next, results are shown for a two-dimensional input space in $[-2, 6] \times [-2, 6]$. The true response is given by $z(\mathbf{x}) = x_1 \exp(-x_1^2 - x_2^2)$. A small amount of Gaussian noise (with sd = 0.001) is added. Besides its dimensionality, a key difference between this dataset and the last one is that it is not defined using step functions; this smooth function does not have any artificial breaks between regions.

Figure 4 shows plots comparing fits of Bayesian Linear CART (*left*) and the treed GP (*right*) which find an average of roughly five and three partitions, respectively. It is clear from the figure that the treed GP is better. The fit for a stationary GP is not shown because it looks very similar to that of the treed GP, since the data are indeed stationary. Much of the advantage of the treed GP in this situation, over a single stationary GP, is in speed of computation. Inverting three matrices, one of half and two of one quarter of the original size (N), is considerably faster than inverting a single $N \times N$ matrix.

2.5.3 Motorcycle data

The Motorcycle Accident Dataset (Silverman, 1985) is a classic dataset used in recent literature (Rasmussen and Ghahramani, 2002) to demonstrate the success of nonstationary models. The data set consists of measurements of acceleration of the head of a motorcycle rider as a function of time in the first moments

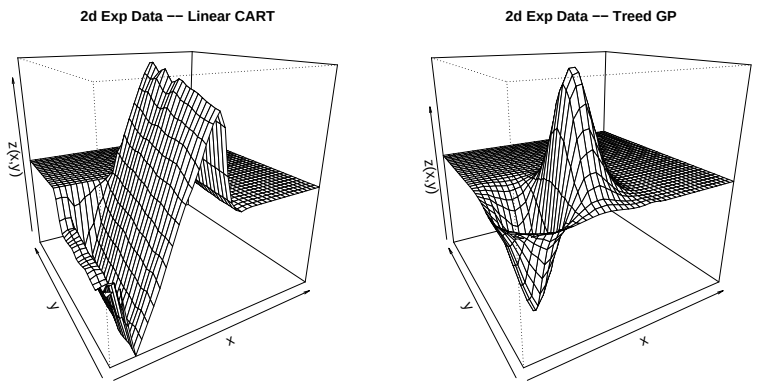

Figure 4: Comparison between Bayesian linear CART (*left*), and the treed GP model (*right*), for the 2-d Exponential data.

after an impact. In addition to being nonstationary, the data has input-dependent noise, an aspect overlooked by a number of nonparametric regression analyses of this dataset.

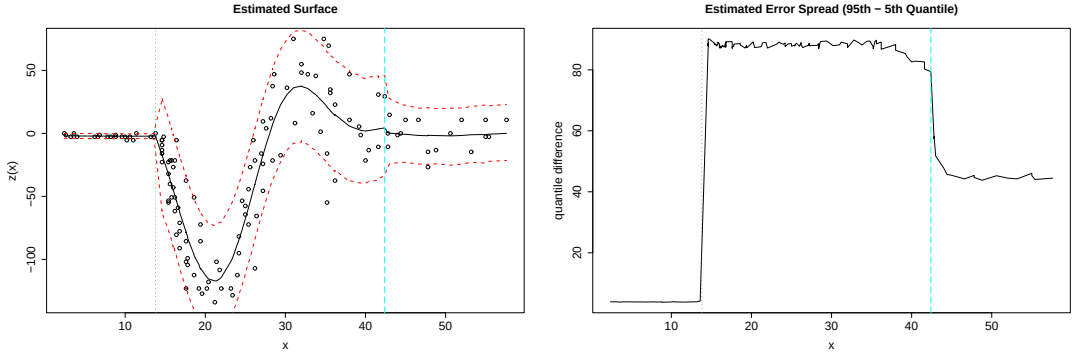

Figure 5: 1-d Motorcycle Dataset, fit by our nonstationary model.

Figure 5 shows the data and the fit given by the treed GP model. The *left* panel shows the estimate of the surface with 90%-quantile error bars; the *right* panel shows the difference in quantiles. Vertical lines on both panels illustrate a typical treed partition $\mathcal{T}$. The error bars, and estimated error spread, can give insight into the uncertainty in the posterior distribution for $\mathcal{T}$. Notice the sharp rise in estimated variance from the leftmost region to the center region. Contrast this with the more gradual descent in variance from the center to the rightmost region. 20,000 MCMC rounds yielded an average average of 3.11 partitions in $\mathcal{T}$.

These results are quite different from those reported by Rasmussen & Ghahramani (2002). In particular, the error-bars they report for the leftmost region seem too large relative to the center and rightmost regions. They use a what they call an "infinite" mixture of GP "experts" which is really a Dirichlet process mixture of GPs. They report that the posterior distribution uses between 3 and 10 experts to fit this dataset, which they admit has "roughly" three regions. In fact, in their histogram of the number of GP experts used

throughout the MCMC rounds, they show that between 3 and 10 experts are equally likely, and even 10-15 experts still have considerable posterior mass. Contrast this with the treed GP model which almost always partitions into three regions, occasionally four, rarely two. On speed grounds, the treed GP is also a winner. Rasmussen & Ghahramani (2002) report that they ran the mixture of GP experts model using a total of 11,000 MCMC rounds, discarding the first 1,000 and keeping every 100^{th} after that. This took roughly one hour on a 1 GHz Pentium. Allowing treed GP to use 25,000 MCMC rounds, discarding the first 5,000 and keeping every sample thereafter takes about 3 minutes on a 1.8 GHz Athalon.

3 Gaussian processes and limiting linear models

Gaussian processes (GPs) retain the linear model (LM) either as a special case, or in the limit. This section shows how the limiting parameterization can be exploited when the data are at least partially linear. However, from the prospective of the Bayesian posterior, the GPs which encode the LM either have probability of nearly zero or are otherwise unattainable without the explicit construction of a prior with the limiting linear model (LLM) in mind. Here, we develop such a prior and show how its practical benefits extend well beyond the computational and conceptual simplicity of the LM. For example, linearity can be extracted on a per-dimension basis yielding a semiparametric model, or can be combined with treed partitioning to yield a highly efficient nonstationary model.

The correlation function and its parameters are the focus of this section, so a particular parameterization is needed. Here we work with the power family (2), with power $p_0 = 2$, and nugget g, in the form: $K(\mathbf{x}_j, \mathbf{x}_k|g) = K^*(\mathbf{x}_j, \mathbf{x}_k) + g\delta_{j,k}$. Recall that the isotropic correlation function (3) is parameterized with a single range parameter, d: $K^*(\mathbf{x}_j, \mathbf{x}_k|d) = \exp\{-||\mathbf{x}_j - \mathbf{x}_k||^2/d\}$ and that the separable function (4) has m_X range parameters $\mathbf{d} = \{d_1, \ldots, d_{m_X}\}$: $K^*(\mathbf{x}_j, \mathbf{x}_k|\mathbf{d}) = \exp\{-\sum_{i=1}^{m_X} |x_{ij} - x_{ik}|^2/d_i\}$. When it applies to both separable and isotropic versions, d and $\mathbf{d}$ are used interchangeably since the isotropic version is a special case of the separable one. Subscripts (ν) are dropped as the discussion applies generally to any GP. However, when coupled with treed partitioning, it may be possible to treat formerly non-linear data as piecewise linear and gain a great advantage. This work is clearly extensible to the case of unknown power p_0 or to other families of correlation functions.

3.1 Limiting Linear Models

A special limiting case of the Gaussian process model is the standard linear model. Replacing the likelihood $\mathbf{Z}|\boldsymbol{\beta}, \sigma^2, \mathbf{K} \sim N_N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{K})$ in the hierarchical model given in Eq. (5) with $\mathbf{Z}|\boldsymbol{\beta}, \sigma^2 \sim N_N(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the $N \times N$ identity matrix, gives a parameterization of a linear model. From a phenomenological perspective,

GP regression is more flexible than standard linear regression in that it can capture nonlinearities in the interaction between covariates ($\mathbf{x}$) and responses (z). From a modeling perspective, the GP can be more than just overkill for linear data. Parsimony and over-fitting considerations are the tip of the iceberg. Inference for GPs on linear data unnecessarily requires the inversion of a large covariance matrix—an operation that can be costly as well as numerically unstable as the smooth/linear data support large finite range parameters (d) which can cause the off-diagonal elements of $\mathbf{K}$ to be nearly one.

So in other words, for some parameterizations, the GP is operationally equivalent to the limiting linear model (LLM), but comes with none of its benefits, e.g., speed and stability. Exploiting and/or manipulating such equivalence can be of great practical benefit. As Bayesians, this means constructing a prior distribution on $\mathbf{K}$ that makes it clear in which situations each model is preferred; i.e., when should $\mathbf{K} \rightarrow c\mathbf{I}$.

Theoretically, there are only two parameterizations to a GP correlation structure $K(\cdot, \cdot)$ which encode the LLM. Though they are well-known, without intervention they are quite unhelpful from the perspective of *practical* estimation and inference. The first one is when the range parameter d is set to zero. In this case $\mathbf{K} = (1 + g)\mathbf{I}$, and the result is clearly a linear model. The second is when the nugget goes to infinity. A third, hybrid, parameterization is alluded to by Cressie (1993, Section 3.2.1) in an analysis of the "effect of variogram parameters on kriging". Specifically, he remarks that a large nugget coupled with a large range drives the interpolator towards the linear mean. Thus an essentially linear model can be reached with nonzero d and finite g. This is refreshing since constructing a prior for the LLM by exploiting the former GP parameterization (range $d \rightarrow 0$) is difficult, and for the latter (nugget $g \rightarrow \infty$) near impossible.

3.2 Understanding the models

Before constructing a prior, it makes sense to study the kriging neighborhood. The following exploratory analysis focuses on studying likelihoods and posteriors for GPs fit to data generated from a linear model with evenly spaced x-values in the range $[0, 1]$:

$$z_i = 1 + 2x_i + \epsilon, \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, 1) \tag{19}$$

Illustrating the limiting linear model parameterization ($g \rightarrow \infty$), the *left*-hand side of Figure 6 shows how as the nugget g increases, the likelihood of the GP approaches that of the linear model for a sample of size $n = 100$ from (19) with the range parameter fixed at $d = 1$. The nugget must be quite large relative to the actual variability in the data be before the likelihoods of the GP and LLM become comparable. The *right*-hand side of Figure 6 summarizes the ratio of the ML GP parameterization over the ML linear model based on 1000 simulations of ten evenly spaced random draws from (19). A likelihood ratio of one means

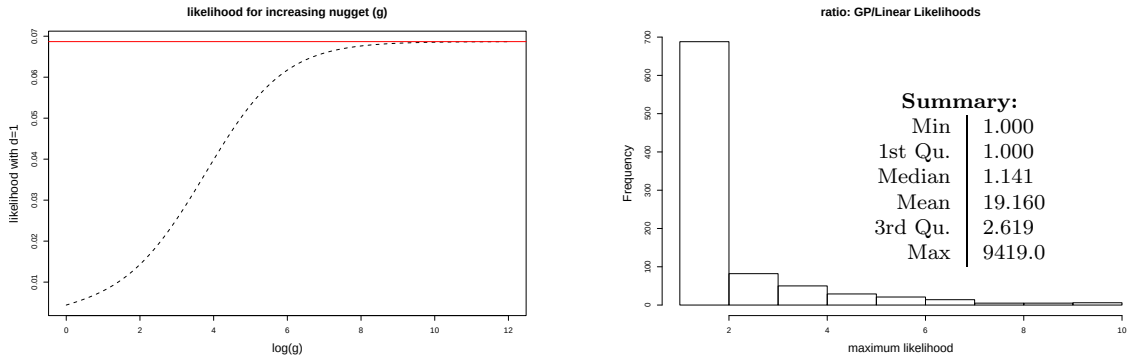

Figure 6: *Left:* Likelihoods as the nugget gets large for an $n = 100$ sample from Eq. (19). The x-axis is $(\log g)$, the range is fixed at $d = 1$. Likelihood of the LLM ($d = 0$) shown for comparison. *Right:* Histogram of the ratio of the ML GP parameterization over the likelihood of the limiting LM. For visibility, the horizontal axis is truncated at 10; full summary statistics are shown.

that the LLM was best for a particular sample. The histogram (with the x-axis truncated at 10 for visibility) and summary statistics in Figure 6 (*right*) show that the GP is seldom much better than the linear model. More than two-thirds of the ratios are close to one—approximately 1/3 (362) were exactly one but 2/3 (616) had ratios less than 1.5—which means that posterior inference for borderline linear data is likely to depend heavily the prior specification of $K(\cdot, \cdot)$.

If it is suspected that the data might be linear then this bias should be encoded in the prior somehow. This is a non-trivial task given the nature of the GP parameterizations which encode the LLM. The marginalized posterior $p(\mathbf{K}|\mathbf{Z}, \boldsymbol{\beta}_0, \tau^2, \mathbf{W})$ of Eq. (12) can be used, which integrates out $\boldsymbol{\beta}$ and σ^2, which for the power family means specifying $p(d, g)$. Alternatively, one could consider dropping the $p(\mathbf{K})$ term from (12) and look solely at the marginalized likelihood (Berger et al., 2001). Consider a mixture of gammas prior for d:

$$p(d, g) = p(d) \times p(g) = p(g) \times \frac{1}{2}[G(d|\alpha = 1, \beta = 20) + G(d|\alpha = 10, \beta = 10)]. \tag{20}$$

It gives roughly equal mass to small d representing a population of GP parameterizations for wavy surfaces, and a separate population for those which are quite smooth or approximately linear. Figure 8 depicts $p(d)$ via histogram, ignoring $p(g)$ which we take to be a simple exponential distribution. Alternatively, one could encode the prior as $p(d, g) = p(d|g)p(g)$ and then use a reference prior (Berger et al., 2001) for $p(d|g)$. We prefer the more deliberate, mixture, specification for reasons that will become apparent shortly.

Figure 7 (*left*) shows a representative MAP GP ($d \approx 1$) fit for a sample of size $n = 100$ from (19). The likelihood around $d = 0$, shown in the *middle* panel, is severely peaked, but small, nonzero, d-values have extremely low likelihood. So the large likelihood at $d = 0$ is effectively a point-mass which does not get picked up by the posterior because $p(d = 0) = 0$; instead the data give a continuum large d parameters,

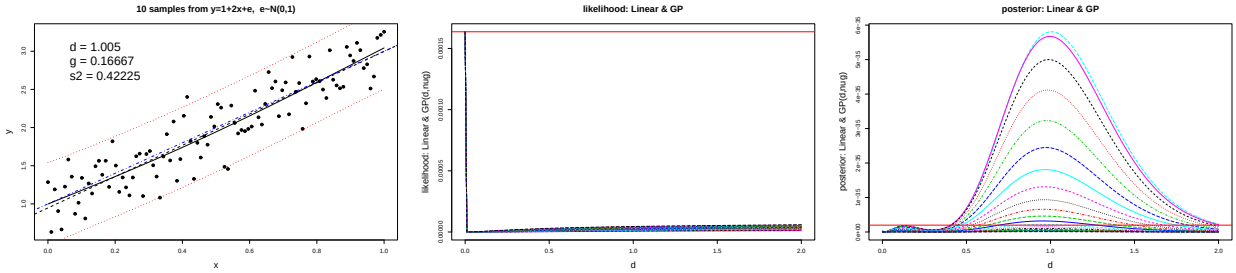

Figure 7: *Left* shows the GP(d, g) fit with a sample of size $n = 100$; *middle* shows the likelihood and *right* shows the integrated posterior distribution for range (d, x-axis) and nugget (g, lines) settings.

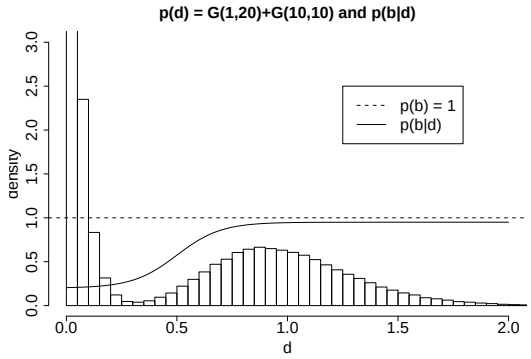

Figure 8: Prior distribution for the boolean (b) superimposed on $p(d)$.

and hence the GP, high posterior probability (*right*). Thus, the MAP estimate ($d \approx 1$) results in a linear "looking" predictive surface which bears the computational burden implied by full-fledged GPs.

3.3 Model Selection Priors

Motivated by the discussion above, we set out to construct a prior for the "mixture" of the GP with its LLM. The key idea is an augmentation of the parameter space by m_X indicators $\mathbf{b} = \{b\}_{i=1}^{m_X} \in \{0,1\}^{m_X}$. The boolean b_i is intended to select either the GP ($b_i = 1$) or its LLM for the i^{th} dimension. The actual range parameter used by the correlation function is multiplied by $\mathbf{b}$: e.g., $K^*(\cdot,\cdot|\mathbf{bd})$.[1] To encode the preference that GPs with larger range parameters be more likely to "jump" to the LLM, the prior on b_i is specified as a function of the range parameter d_i: $p(b_i, d_i) = p(b_i|d_i)p(d_i)$.

Probability mass functions which increase as a function of d_i, e.g.,

$$p_{\gamma,\theta_1,\theta_2}(b_i = 0|d_i) = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp\{-\gamma(d_i - 0.5)\}} \tag{21}$$

with $0 < \gamma$ and $0 \le \theta_1 \le \theta_2 < 1$, can encode such a preference by calling for the exclusion of dimensions i with large d_i when constructing $\mathbf{K}$. Thus b_i determines whether the GP or the LLM is in charge of the

[1] i.e. component-wise multiplication—like the "$\mathbf{b}.*\mathbf{d}$" operation in `Matlab`

marginal process in the i^{th} dimension. Accordingly, θ_1 and θ_2 represent minimum and maximum probabilities of jumping to the LLM, while γ governs the rate at which $p(b_i = 0|d_i)$ grows to θ_2 as d_i increases. Figure 8 plots $p(b|d)$ with $(\gamma, \theta_1, \theta_2) = (10, 0.2, 0.95)$ superimposed on the mixture of Gammas prior $p(d_i)$ from (20). The θ_2 parameter is taken to be strictly less than one so as not to preclude a GP which models a genuinely nonlinear surface using an uncommonly large range setting.

The implied prior probability of the full m_X-dimensional LLM is

$$p(\text{linear model}) = \prod_{i=1}^{m_X} p(b_i = 0|d_i) = \prod_{i=1}^{m_X} \left[\theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp\{-\gamma(d_i - 0.5)\}}\right]. \tag{22}$$

The resulting process is still a GP if any of the booleans b_i are one. The primary computational advantage associated with the LLM is foregone unless all of the b_i's are zero. However, the intermediate result offers an improvement in numerical stability in addition to describing a unique transitionary model lying somewhere between the GP and the LLM. Specifically, it allows for the implementation of semiparametric stochastic processes like $Z(\mathbf{x}) = \boldsymbol{\beta} f(\mathbf{x}) + \varepsilon(\tilde{\mathbf{x}})$ representing a piecemeal spatial extension of a simple linear model. The first part ($\boldsymbol{\beta} f(\mathbf{x})$) of the process is linear in some known function of the the full set of covariates $\mathbf{x} = \{x_i\}_{i=1}^{m_X}$, and $\varepsilon(\cdot)$ is a spatial random process, e.g., a GP, which acts on a subset of the covariates $\tilde{\mathbf{x}}$. Such models are commonplace in the statistics literature (e.g., Dey et al., 1998). Traditionally, $\tilde{\mathbf{x}}$ is determined and fixed *a priori*. The separable boolean prior in (21) implements an adaptively semiparametric process where the subset $\tilde{\mathbf{x}} = \{x_i : b_i = 1, i = 1, \ldots, m_X\}$ is given a prior distribution, instead of being fixed.

3.3.1 Prediction

Prediction under the limiting GP model is a simplification of Eqs. (15–17) since it is known that $\mathbf{K} = (1+g)\mathbf{I}$. A characteristic of the standard linear model is that all input configurations ($\mathbf{x}$) are treated as independent conditional on knowing $\boldsymbol{\beta}$. Moreover, the terms $k(\mathbf{x})$ and $K(\mathbf{x}, \mathbf{y})$ in (15–17) are zero for all $\mathbf{x}$, and $\mathbf{y} \neq \mathbf{x}$. Thus, the predicted value of z at $\mathbf{x}$ is normally distributed with mean $\hat{z}(\mathbf{x}) = \mathbf{f}^\top(\mathbf{x})\tilde{\boldsymbol{\beta}}$, and variance

$$\hat{\sigma}(\mathbf{x})^2 = \sigma^2[1 + \tau^2\mathbf{f}^\top(\mathbf{x})\mathbf{W}\mathbf{f}(\mathbf{x}) - \tau^2\mathbf{f}^\top(\mathbf{x})\mathbf{W}\mathbf{F}^\top((1+g)\mathbf{I} + \tau^2\mathbf{F}\mathbf{W}\mathbf{F}^\top)^{-1}\mathbf{F}\mathbf{W}\mathbf{f}(\mathbf{x})\tau^2]. \tag{23}$$

It is helpful to re-write the above expression for the variance as

$$\hat{\sigma}(\mathbf{x})^2 = \sigma^2[1 + \tau^2\mathbf{f}^\top(\mathbf{x})\mathbf{W}\mathbf{f}(\mathbf{x})] - \sigma^2\left[\frac{\tau^2}{1+g}\mathbf{f}^\top(\mathbf{x})\mathbf{W}\mathbf{F}^\top\left(\mathbf{I} + \frac{\tau^2}{1+g}\mathbf{F}\mathbf{W}\mathbf{F}^\top\right)^{-1}\mathbf{F}\mathbf{W}\mathbf{f}(\mathbf{x})\tau^2\right]. \tag{24}$$

A matrix inversion lemma called the Woodbury formula (Golub and Van Loan, 1996, p. 51) or the Sherman-Morrison-Woodbury formula (Bernstein, 2005, p. 67; best to see `Mathworld` for easy access to both formulas):

states that for $(\mathbf{I} + \mathbf{V}^\top \mathbf{A}\mathbf{V})$ non-singular, $(\mathbf{A}^{-1} + \mathbf{V}\mathbf{V}^\top)^{-1} = \mathbf{A} - (\mathbf{A}\mathbf{V})(\mathbf{I} + \mathbf{V}^\top \mathbf{A}\mathbf{V})^{-1}\mathbf{V}^\top \mathbf{A}$. Taking $\mathbf{V} \equiv \mathbf{F}^\top (1+g)^{-\frac{1}{2}}$ and $\mathbf{A} \equiv \tau^2 \mathbf{W}$ in (24) gives

$$\hat{\sigma}(\mathbf{x})^2 = \sigma^2 \left[1 + \mathbf{f}^\top(\mathbf{x}) \left(\frac{\mathbf{W}^{-1}}{\tau^2} + \frac{\mathbf{F}^\top \mathbf{F}}{1+g} \right)^{-1} \mathbf{f}(\mathbf{x}) \right]. \tag{25}$$

Not only is (25) a simplification of the predictive variance given in (23), but it should be familiar. Recall the expression for the posterior variance of the regression coefficients $\mathbf{V}_{\tilde{\beta}}$ given in (7). Writing $\mathbf{V}_{\tilde{\beta}}$ with $\mathbf{K}^{-1} = \mathbf{I}/(1+g)$ gives $\mathbf{V}_{\tilde{\beta}} = \left(\frac{\mathbf{W}^{-1}}{\tau^2} + \frac{\mathbf{F}^\top \mathbf{F}}{1+g} \right)^{-1}$. Thus the predictive variance for the LLM is actually

$$\hat{\sigma}(\mathbf{x})^2 = \sigma^2 \left[1 + \mathbf{f}^\top(\mathbf{x}) \mathbf{V}_{\tilde{\beta}} \mathbf{f}(\mathbf{x}) \right]. \tag{26}$$

But this is just the usual result for the predictive variance at $\mathbf{x}$ under the standard linear model. Therefore, the posterior predictive distribution under the LLM is simply

$$y(\mathbf{x}) = N[\mathbf{f}^\top(\mathbf{x})\tilde{\boldsymbol{\beta}}, \sigma^2(1 + \mathbf{f}^\top(\mathbf{x}) \mathbf{V}_{\tilde{\beta}} \mathbf{f}(\mathbf{x}))]. \tag{27}$$

This means we have a choice when it comes to obtaining samples from the posterior predictive distribution under the LLM. Eq. (26) is preferred over (23) because the latter involves inverting the $N \times N$ matrix, $\mathbf{I} + \tau^2 \mathbf{F}\mathbf{W}\mathbf{F}^\top/(1+g)$, whereas the former only requires the inversion of an $m \times m$ matrix.

3.4 Implementation, results, and comparisons

Here, the GP with jumps to the LLM (hereafter GP LLM) is illustrated on synthetic and real data. Most of the experiments are in the context of applying the GP LLM at the leaves of the tree, upgrading the treed GP model of Section 2 to a treed GP LLM model. Section 3.4.2 shows an example without treed partitioning. Partition models are an ideal setting for evaluating the utility of the GP LLM as linearity can be extracted in large areas of the input space. The result is a uniquely tractable nonstationary semiparametric spatial model. Treed and non-treed GP LLM models are implemented in the `tgp` package on CRAN.

A separable correlation function is used throughout this section for brevity and consistency, even though in some cases the process which generated the data is clearly isotropic. Proposals for the booleans $\mathbf{b}$ are drawn from the prior, conditional on $\mathbf{d}$, and accepted or rejected on the basis of the constructed covariance matrix $\mathbf{K}$. The same prior parameterizations are used for all experiments unless otherwise noted, the idea being to develop a method that works "right out of the box" as much as possible.

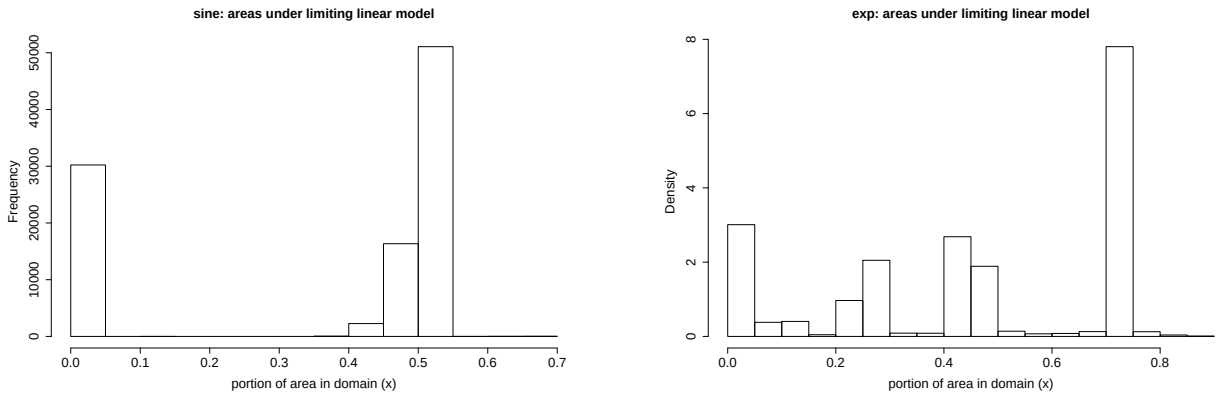

Figure 9: Histograms of the areas of the domain under the LLM spread over 20 repeated $n = 100$ samples from (*left*) the sine data and (*right*) exponential data.

3.4.1 Revisiting simple synthetic and real data with the treed GP LLM

Recall the synthetic sinusoidal data from Section 2.5.1, which we know from (18) is linear for exactly half of the domain. The *left* panel of Figure 9 shows a histogram of the areas under the LLM for each of 10,000 MCMC samples collected for the treed GP LLM model over 20 repeated experiments of size $n = 100$. An average of 42% of the input domain was under the LLM, and the mode can be seen to be near 0.5. A similar experiment which included predicting at $n' = 200$ new locations revealed that the treed GP LLM was 27% faster than treed GP alone.

Recall the 2-d synthetic exponential data from Section 2.5.2. On this dataset, the partitioning structure of the treed GP LLM first splits the region into two halves, one of which can be fit linearly. It then recursively partitions the half with the "action" into a piece which requires a GP and another piece which is also linear. The *right* panel of Figure 9 shows a histogram of the areas of the domain under the LLM over 20-fold repeated experiments. The four modes of the histogram clump around 0%, 25%, 50%, and 75% showing that most often the obvious three-quarters of the space are under the LLM, although sometimes one of the two partitions will use a very smooth GP. On average, 66% of the domain was under the LLM. The treed GP LLM was 40% faster than the treed GP alone when combining estimation and sampling from the posterior predictive distributions at the remaining $n' = 241$ points from the grid.

Recall the Motorcycle Accident Dataset from Section 2.5.3. In an experiment using the treed GP LLM, an average of 29% of the domain was under the LLM, split between the left low-noise region (before impact) and the noisier right region. The Rasmussen and Ghahramani (2002) analysis of this dataset with the DPGP reportedly took one hour on a 1 GHz Pentium. Such times are typical of nonstationary modeling because of the computational effort required to construct and invert large covariance matrices. In contrast, the treed GP LLM fits this dataset with comparable accuracy but in less than one minute on a 1.8 GHz Athalon.

In all three experiments, the predictive surfaces obtained are virtually identical to those shown in Section 2.5, so they are not shown here. The main advantage of the treed GP LLM for these data is speed. Three things make the treed GP LLM fast relative to most nonstationary spatial models. (1) Partitioning fits models to less data, yielding smaller matrices to invert. (2) Jumps to the LLM mean fewer inversions all together. (3) MCMC mixes better because under the LLM the parameters $\mathbf{d}$ and g are out of the picture and all sampling can be performed via Gibbs steps.

3.4.2 Friedman data

This Friedman data set is the first one of a suite that was used to illustrate MARS (Multivariate Adaptive Regression Splines) (Friedman, 1991). There are 10 covariates in the data ($\mathbf{x} = \{x_1, x_2, \ldots, x_{10}\}$), but the function that describes the responses (Z), observed with standard Normal noise,

$$E(Z|\mathbf{x}) = \mu = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5, \tag{28}$$

depends only on $\{x_1, \ldots, x_5\}$, thus combining nonlinear, linear, and irrelevant effects. Comparisons are made on this dataset to results provided for several other models in recent literature. The $\mathbf{x}$'s are taken to be randomly distributed on the unit interval. Chipman et al. (2002) used this dataset to compare their linear CART algorithm to four other methods of varying parameterization: linear regression, greedy tree, MARS, and neural networks. The statistic they use for comparison is root mean-square error (RMSE): $\text{MSE} = \sum_{i=1}^{n'}(\mu_i - \hat{z}_i)^2/n'$ and $\text{RMSE} = \sqrt{\text{MSE}}$ where $\hat{z}_i$ is the model-predicted response for input $\mathbf{x}_i$. RMSE's are gathered for fifty noisy simulations of size $n = 100$ from (28), and the posterior mean predictive responses at the training data (i.e., $n' = n$) are compared to the true expectation. Chipman et al. provide a nice collection of boxplots showing the results. However, they do not provide any numerical results, so we have extracted some key numbers from their plots and refer the reader to that paper for the full results.

We duplicated this experiment using the GP LLM. For this dataset, a single model was used, not a treed model, as the function is essentially stationary in the spatial statistical sense (so if we were to try to fit a treed GP, it would keep all of the data in a single partition). Linearizing boolean prior parameters $(\gamma, \theta_1, \theta_2) = (10, 0.2, 0.9)$ were used, which gave the LLM a relatively low prior probability of 0.35, for large range parameters d_i. The RMSEs obtained for the GP LLM are summarized in the table below.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
GP LLM	0.4341	0.5743	0.6233	0.6258	0.6707	0.7891
LM	1.710	2.165	2.291	2.325	2.500	2.794

Results on the linear model are reported for calibration purposes, and can be seen to be essentially the same as those reported by Chipman et al. RMSEs for the GP LLM are on average significantly better than *all* of

those reported for the above methods, with lower variance. For example, the best mean RMSE shown in the boxplot is about 0.9. That is 1.4 times higher than the worst one obtained for GP LLM. Further comparison to the boxplots provided by Chipman et al. shows that the GP LLM is the clear winner.

In fitting the model, the Markov chain quickly keyed in on the fact that only the first three covariates contribute nonlinearly. After burn-in, the booleans $\mathbf{b}$ almost never deviated from $(1, 1, 1, 0, 0, 0, 0, 0, 0, 0)$. From the following table summarizing the posterior for the linear regression coefficients $\boldsymbol{\beta}$ it can be seen that the coefficients for x_4 and x_5 (between double-bars) were estimated accurately, and that the model correctly determined that $\{x_6, \ldots x_{10}\}$ were irrelevant, i.e., not included in the GP, and had β's close to zero.

		x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
	5% Qu.	8.40	2.60	-1.23	-0.89	-1.82	-0.60	- 0.91
$\boldsymbol{\beta}$	Mean	9.75	4.59	-0.190	0.049	-0.612	0.326	0.066
	95% Qu.	10.99	9.98	0.92	1.00	0.68	1.21	1.02

For a final comparison, consider an SVM method (Drucker et al., 1996) illustrated on this dataset and compared to Bagging (Breiman, 1996) regression trees. Note that the SVM method required cross-validation (CV) to set some of its parameters. In the comparison, 100 randomized training sets of size $n = 200$ were used, and MSEs were collected for a (single) test set of size $n' = 1000$. An average MSE of 0.67 is reported, showing the SVM to be uniformly better than the Bagging method with an MSE of 2.26. We repeated the experiment for the GP LLM (which requires no CV), and obtained an average MSE of 0.293, which is 2.28 times better than the SVM, and 7.71 times better than Bagging.

3.4.3 Boston housing data

A commonly used data set for validating multivariate models is the Boston Housing Data (Harrison and Rubinfeld, 1978) available from the UCI Machine Learning repository (Newman et al., 1998), which contains 506 responses over 13 covariates. Chipman et al. (2002) showed that their (Bayesian) linear CART model gave lower RMSEs, on average, compared to a number of popular techniques (the same ones listed above). The treed GP LLM is a generalization of the linear CART model, retaining the original linear CART as an accessible special case. Though computationally more intensive than linear CART, the treed GP LLM gives impressive results. To mitigate some of the computational demands, the LLM can be used to initialize the Markov chain by breaking the larger data set into smaller partitions. Before treed GP burn-in begins, the model is fit using only the faster (limiting) linear CART model. Once the treed partitioning has stabilized, this fit is taken as the starting value for a full MCMC exploration of the posterior for the treed GP LLM. This initialization process allows fitting of GPs to smaller segments of the data, reducing the size of matrices that need to be inverted and greatly reducing computation time.

Experiments in the Bayesian linear CART paper (Chipman et al., 2002) consist of calculating RMSEs via 10-fold CV. The data are randomly partitioned into 10 groups, iteratively trained on 9/10 of the data, and tested on the remaining 1/10. This is repeated for 20 random partitions, and boxplots are shown. The logarithm of the response is used, and CV is only used to assess predictive error, not to tune parameters. Samples are gathered from the posterior predictive distribution of the linear CART model for six parameterizations using 20 restarts of 4000 iterations. In order to obtain a fair comparison, we followed suit for the treed GP LLM . Settings of $(\gamma, \theta_1, \theta_2) = (10, 0.2, 0.95)$ were used, which gives the LLM a prior probability of $0.95^{13} \approx 0.51$, when the d_i's are large. Our "boxplot" for training and testing RMSEs are summarized numerically in the table below. As before, linear regression (on the log responses) is used for calibration.

		Min	1st Qu.	Median	Mean	3rd Qu.	Max
train	GP LLM	0.0701	0.0716	0.0724	0.0728	0.0730	0.0818
	LM	0.1868	0.1869	0.1869	0.1869	0.1869	0.1870
test	GP LLM	0.1321	0.1327	0.1346	0.1346	0.1356	0.1389
	LM	0.1926	0.1945	0.1950	0.1950	0.1953	0.1982

The RMSEs for the linear model have extremely low variability. This is similar to the results provided by Chipman et al. and was a key factor in determining that the experiment was well-calibrated. Upon comparison of the above numbers with the boxplots in Chipman et al., it can readily be seen that the treed GP LLM is leaps and bounds better than linear CART, and *all* of the other methods in the study. The treed GP LLM's worst training RMSE is almost two times lower than the best ones from the boxplot. All testing RMSEs are lower than the lowest ones from the boxplot, and the median RMSE (0.1346) is 1.26 times lower than the lowest median RMSE (≈ 0.17) from the boxplot.

More recently, Chu et al. (2004, Table V) performed a similar experiment, but instead of 10-fold CV, they randomly partitioned the data 100 times into training/test sets of size 481/25 and reported average MSEs on the un-transformed responses. They compare their Bayesian SVM regression algorithm (BSVR) to other high-powered techniques like Ridge Regression, Relevance Vector Machine, GPs, etc., with and without ARD (automatic relevance determination). Repeating their experiment for the treed GP LLM gave an average MSE of 6.96 compared to that of 6.99 for the BSVR with ARD, making the two algorithms by far the best in the comparison. However, without ARD the MSE of BSVR was 12.34, 1.77 times higher than the treed GP LLM, and the worst in the comparison. The reported results for a GP with (8.32) and without (9.13) ARD showed the same effect, but to a lesser degree. Thus the GP LLM might similarly benefit from an ARD-like approach. Perhaps not surprisingly, the average MSEs do not tell the whole story. The 1st, median, and 3rd quantile MSEs obtained for the treed GP LLM were 3.72, 5.32 and 8.48 respectively, showing that its distribution had a heavy right-hand tail. This may be an indication that several responses in the data are either misleading, noisy, or otherwise very hard to predict.

4 Conclusion

In this paper, we introduced the treed Gaussian Process model as a simple and efficient method for non-stationary modeling, and validated it on synthetic and real data. A fully Bayesian treatment of the treed GP model was laid out, treating the hierarchical parameterization of the correlation function $K(\cdot, \cdot)$ as a modular component, easily replaced by a different family of correlations.

We also argued that Gaussian processes are a flexible modeling tool which can be overkill for many applications. We showed how the limiting linear model parameterization of the GP can be both useful and accessible in terms of Bayesian posterior estimation and prediction. The benefits include speed, parsimony, and a relatively straightforward implementation of an adaptively semiparametric model. Combined with treed partitioning, the GP LLM further extends the treed GP model, resulting in a uniquely nonstationary, semiparametric, tractable, and highly accurate model that contains linear CART as a special case.

We believe that a large contribution of the treed GP (and LLM) will be in the domain of sequential design of computer experiments (Santner et al., 2003; Gramacy et al., 2004) which was the inspiration for much of the work presented here. Empirical evidence suggests that many computer experiments are nearly linear. That is, either the response is linear in most of its input dimensions, or the process is entirely linear in a subset of the input domain. Supremely relevant, but receiving less emphasis in this paper, is that the Bayesian treed GP LLM provides a *full* posterior predictive distribution (particularly a nonstationary and thus region-specific estimate of predictive variance) which can be used towards active learning in the input domain. Exploitation of these characteristics should lead to a efficient framework for the adaptive exploration of computer experiment parameter spaces.

Acknowledgments

This work was partially supported by research subaward 08008-002-011-000 from the Universities Space Research Association and NASA, NASA/University Affiliated Research Center grant SC 2003028 NAS2-03144, Sandia National Laboratories grant 496420, and National Science Foundation grants DMS 0233710 and 0504851.

A Parameter Estimation Details

The following sub-sections show full derivations of conditional and marginalized posteriors of the parameters to the Gaussian processes at the leaves of the tree.

A.1 Full Conditionals

$\boxed{\boldsymbol{\beta}:}$ $p(\boldsymbol{\beta}_\nu|\text{rest}) \propto p(\mathbf{Z}_\nu|\boldsymbol{\beta}_\nu, \sigma_\nu^2, d_\nu, g_\nu)p(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2, \tau_\nu^2, \mathbf{W}) = N(\mathbf{Z}_\nu|\mathbf{F}_\nu\boldsymbol{\beta}_\nu, \sigma_\nu^2\mathbf{K}_\nu) \cdot N(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2\tau^2\mathbf{W})$

$$\begin{aligned}
&\propto \exp\left\{-(2\sigma_\nu)^{-2}\left[(\mathbf{Z}_\nu - \mathbf{F}_\nu\boldsymbol{\beta}_\nu)'\mathbf{K}_\nu^{-1}(\mathbf{Z}_\nu - \mathbf{F}_\nu\boldsymbol{\beta}_\nu) + (\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\left(\tau_\nu^2\mathbf{W}\right)^{-1}\tau_\nu^2\left(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0\right)\right]\right\} \\
&\propto \exp\left\{-(2\sigma_\nu)^{-2}\left[-2\mathbf{Z}_\nu'\mathbf{K}_\nu^{-1}\mathbf{F}_\nu\boldsymbol{\beta}_\nu + \boldsymbol{\beta}_\nu'\mathbf{F}_\nu'\mathbf{K}_\nu^{-1}\mathbf{F}_\nu\boldsymbol{\beta}_\nu + \boldsymbol{\beta}_\nu'(\tau_\nu\mathbf{W})^{-1}\boldsymbol{\beta}_\nu - 2\boldsymbol{\beta}_\nu'(\tau_\nu^2\mathbf{W})^{-1}\boldsymbol{\beta}_0\right]\right\} \\
&= \exp\left\{-(2\sigma_\nu)^{-2}\left[\boldsymbol{\beta}_\nu'(\mathbf{F}_\nu'\mathbf{K}_\nu^{-1}\mathbf{F}_\nu + \mathbf{W}^{-1}/\tau^2)\boldsymbol{\beta}_\nu - 2\boldsymbol{\beta}_\nu'(\mathbf{F}_\nu'\mathbf{K}_\nu^{-1}\mathbf{Z}_\nu + \mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau^2)\right]\right\}
\end{aligned}$$

$$\text{giving} \qquad \boldsymbol{\beta}_\nu|\text{rest} \sim N(\tilde{\boldsymbol{\beta}}_\nu, \sigma_\nu^2\mathbf{V}_{\tilde{\beta}_\nu}) \tag{29}$$

$$\text{where} \qquad \mathbf{V}_{\tilde{\beta}_\nu} = (\mathbf{F}_\nu'\mathbf{K}_\nu^{-1}\mathbf{F}_\nu + \mathbf{W}^{-1}/\tau^2)^{-1} \qquad \tilde{\boldsymbol{\beta}}_\nu = \mathbf{V}_{\tilde{\beta}_\nu}(\mathbf{F}_\nu'\mathbf{K}_\nu^{-1}\mathbf{Z}_\nu + \mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau^2).$$

$\boxed{\boldsymbol{\beta}_0:}$ $p(\boldsymbol{\beta}_0|\text{rest}) = p(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \sigma^2, \tau^2, \mathbf{W})p(\boldsymbol{\beta}_0)$

$$\begin{aligned}
&= p(\boldsymbol{\beta}_0)\prod_{i=1}^{r} p(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2, \tau_\nu^2, \mathbf{W}) = N(\boldsymbol{\beta}_0|\mu, \mathbf{B})\prod_{i=1}^{r} N(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2\tau_\nu^2\mathbf{W}) \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_0 - \mu)'\mathbf{B}^{-1}(\boldsymbol{\beta}_0 - \mu)\right\}\prod_{i=1}^{r}\exp\left\{-\frac{1}{2\sigma_\nu^2\tau_\nu^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}_0 - \mu)'\mathbf{B}^{-1}(\boldsymbol{\beta}_0 - \mu) + \sum_{i=1}^{r}\frac{1}{\sigma_\nu^2\tau_\nu^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)\right]\right\}
\end{aligned}$$

$$\begin{aligned}
p(\boldsymbol{\beta}_0|\text{rest}) &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_0'\mathbf{B}^{-1}\boldsymbol{\beta}_0 - 2\boldsymbol{\beta}_0'\mathbf{B}^{-1}\mu + \boldsymbol{\beta}_0'\mathbf{W}^{-1}\sum_{i=1}^{r}\frac{\boldsymbol{\beta}_0}{\sigma_\nu^2\tau_\nu^2} - 2\boldsymbol{\beta}_0'\mathbf{W}^{-1}\sum_{i=1}^{r}\frac{\boldsymbol{\beta}_\nu}{\sigma_\nu^2\tau_\nu^2}\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}_0'\left(\mathbf{B}^{-1} + \sum_{i=0}^{r}\frac{\mathbf{W}^{-1}}{\sigma_\nu^2\tau_\nu^2}\right)\boldsymbol{\beta}_0 - 2\boldsymbol{\beta}_0'\left(\mathbf{B}^{-1}\mu + \mathbf{W}^{-1}\sum_{i=1}^{r}\frac{\boldsymbol{\beta}_\nu}{\sigma_\nu^2\tau_\nu^2}\right)\right]\right\}
\end{aligned}$$

$$\text{giving} \qquad \boldsymbol{\beta}_0|\text{rest} \sim N(\tilde{\boldsymbol{\beta}}_0, \boldsymbol{V}_{\tilde{\beta}_0}) \tag{30}$$

$$\text{where} \qquad \mathbf{V}_{\tilde{\beta}_0} = \left(\mathbf{B}^{-1} + \mathbf{W}^{-1}\sum_{i=0}^{r}(\sigma_\nu\tau_\nu)^{-2}\right)^{-1} \qquad \tilde{\boldsymbol{\beta}}_0 = \mathbf{V}_{\tilde{\beta}_0}\left(\mathbf{B}^{-1}\mu + \mathbf{W}^{-1}\sum_{i=1}^{r}\boldsymbol{\beta}_\nu(\sigma_\nu\tau_\nu)^{-2}\right).$$

$\boxed{\tau^2:}$ $p(\tau_\nu^2|\text{rest}) = p(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2, \mathbf{W})p(\tau_\nu^2) = N(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma^2\tau_\nu^2\mathbf{W})IG(\tau_\nu^2|\alpha_\tau/2, q_\tau/2)$

$$\begin{aligned}
&\propto (\tau_\nu^2)^{-\frac{m}{2}}\exp\left\{-\frac{(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)^\top\mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)}{2\sigma_\nu^2\tau_\nu^2}\right\} \times (\tau_\nu^2)^{-(\alpha_\tau/2+1)}\exp\left\{-\frac{q_\tau}{2\tau^2}\right\} \\
&\propto (\tau_\nu^2)^{-(\frac{\alpha_\tau+m}{2}+1)}\exp\left\{-\frac{q_\tau + (\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)^\top\mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)/\sigma_\nu^2}{2\tau_\nu^2}\right\}
\end{aligned}$$

$$\text{giving} \qquad \tau_\nu^2 \sim IG\left(\frac{\alpha_\tau + m}{2}, \frac{q_\tau + (\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)^\top \mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)/\sigma_\nu^2}{2}\right). \tag{31}$$

$\boxed{\mathbf{W^{-1}}:}$

$$\begin{aligned} p(\mathbf{W}^{-1}|\text{rest}) &= p(\mathbf{W})p(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \sigma^2, \tau^2, \mathbf{W}) \\ &= W(\mathbf{W}^{-1}|(\rho\mathbf{V})^{-1}, \rho) \cdot \prod_{i=1}^{r} N(\boldsymbol{\beta}_\nu|\boldsymbol{\beta}_0, \sigma_\nu^2\tau_\nu^2\mathbf{W}) \\ &\propto |\mathbf{W}^{-1}|^{(\rho-m-1)/2} \exp\left\{-\frac{1}{2}\text{tr}((\rho\mathbf{V})\mathbf{W}^{-1})\right\} \times \\ &\qquad |\mathbf{W}^{-1}|^{r/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{r}\frac{1}{\sigma_\nu^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\mathbf{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)\}\right\} \end{aligned}$$

$$p(\mathbf{W}^{-1}|\text{rest}) \propto |\mathbf{W}^{-1}|^{(\rho+r-m-1)/2} \exp\left\{-\frac{1}{2}\left[\text{tr}((\rho\mathbf{V})\boldsymbol{W}^{-1}) + \text{tr}\left(\sum_{i=1}^{r}\frac{1}{\sigma_\nu^2\tau_\nu^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\boldsymbol{W}^{-1}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)\right)\right]\right\}$$

obtained because a scalar is equal to its trace. Applying more properties of the trace operation gives

$$p(\mathbf{W}^{-1}|\text{rest}) \propto |\mathbf{W}^{-1}|^{\frac{\rho+r-m-1}{2}} \exp\left\{-\frac{1}{2}\left[\text{tr}\left(\left(\rho\mathbf{V} + \sum_{i=1}^{r}\frac{(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'}{\sigma_\nu^2\tau_\nu^2}\right)\mathbf{W}^{-1}\right)\right]\right\}$$

$$\text{giving} \qquad \mathbf{W}^{-1}|\text{rest} \sim W\left(\left(\rho\mathbf{V} + \sum_{i=1}^{r}\frac{1}{\sigma_\nu^2\tau_\nu^2}(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)(\boldsymbol{\beta}_\nu - \boldsymbol{\beta}_0)'\right)^{-1}, \rho + r\right). \tag{32}$$

A.2 Marginalized Conditional Posteriors

Complete conditional posteriors for the parameters to the correlation function $K(\cdot,\cdot)$ can be obtained by analytically integrating out $\boldsymbol{\beta}$ and σ^2 to get a marginal posterior.

$$\begin{aligned} p(\mathbf{K}|\mathbf{Z}, \boldsymbol{\beta}_0, \mathbf{W}, \tau^2) &= \prod_\nu p(\mathbf{K}_\nu|\mathbf{Z}_\nu, \boldsymbol{\beta}_0, \tau^2, \mathbf{W}) \\ &\propto \prod_\nu p(\mathbf{K}_\nu) \int p(\sigma_\nu^2) \int p(\mathbf{Z}_\nu|d_\nu, g_\nu, \boldsymbol{\beta}_\nu, \sigma_\nu^2) p(\boldsymbol{\beta}_\nu|\sigma_\nu^2, \boldsymbol{\beta}_0, \tau_\nu^2, \mathbf{W})\, d\boldsymbol{\beta}_\nu\, d\sigma_\nu^2 \\ &= \prod_\nu p(\mathbf{K}_\nu) \int p(\sigma_\nu^2) \int N(\boldsymbol{\beta}_\nu|\tilde{\boldsymbol{\beta}}_\nu, \sigma_\nu^2\mathbf{V}_{\tilde{\beta}_\nu})\, d\boldsymbol{\beta}_{\boldsymbol{\nu}} \times (2\pi)^{-\frac{n_\nu}{2}}\sigma_\nu^{-n_\nu}|\mathbf{K}_\nu|^{-\frac{1}{2}}\tau_\nu^{-m}|\mathbf{W}|^{-\frac{1}{2}}|\mathbf{V}_{\tilde{\beta}_\nu}|^{\frac{1}{2}} \\ &\qquad \times \exp\left\{-\frac{1}{2\sigma_\nu^2}\left[\mathbf{Z}_\nu'\mathbf{K}^{-1}\mathbf{Z}_\nu + \boldsymbol{\beta}_0'\mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau^2 - \tilde{\boldsymbol{\beta}}_\nu'\mathbf{V}_{\tilde{\beta}_\nu}^{-1}\tilde{\boldsymbol{\beta}}_\nu)\right]\right\}\, d\sigma^2. \\ &= \prod_\nu p(\mathbf{K}_\nu) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}\tau_\nu^{2m}|\mathbf{K}_\nu||\mathbf{W}|}\right)^{\frac{1}{2}} \int \sigma_\nu^{-n_\nu} p(\sigma_\nu^2) \exp\left\{-\frac{\psi_\nu}{2\sigma_\nu^2}\right\}\, d\sigma_\nu^2, \end{aligned}$$

$$\text{where} \qquad \psi_\nu = \mathbf{Z}_\nu'\mathbf{K}^{-1}\mathbf{Z}_\nu + \boldsymbol{\beta}_0'\mathbf{W}^{-1}\boldsymbol{\beta}_0/\tau_\nu^2 - \tilde{\boldsymbol{\beta}}_\nu'\mathbf{V}_{\tilde{\beta}_\nu}^{-1}\tilde{\boldsymbol{\beta}}_\nu. \tag{33}$$

Expanding the prior for σ_ν^2 gives:

$$p(\mathbf{K}_\nu|\mathbf{Z},\boldsymbol{\beta}_0,\tau^2,\mathbf{W}) \propto \prod_\nu p(\mathbf{K}) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}\tau_\nu^{2m}|\mathbf{K}_\nu||\mathbf{W}|}\right)^{\frac{1}{2}} \times \int (\sigma_\nu^2)^{-\frac{n_\nu}{2}} \frac{\left(\frac{q_\sigma}{2}\right)^{\frac{\alpha_\sigma}{2}}}{\Gamma(\frac{\alpha_\sigma}{2})} (\sigma_\nu^2)^{-(\frac{\alpha_\sigma}{2}+1)} \exp\left\{-\frac{q_\sigma}{2\sigma_\nu^2}\right\} \exp\left\{-\frac{\psi_\nu}{2\sigma_\nu^2}\right\}\, d\sigma_\nu^2$$

$$p(\mathbf{K}_\nu|\mathbf{Z},\boldsymbol{\beta}_0,\tau^2,\mathbf{W}) \propto \prod_\nu p(\mathbf{K}_\nu) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}\tau_\nu^{2m}|\mathbf{K}_\nu||\mathbf{W}|}\right)^{\frac{1}{2}} \times \frac{\left(\frac{q_\sigma}{2}\right)^{\frac{\alpha_\sigma}{2}}}{\Gamma(\frac{\alpha_\sigma}{2})} \times \frac{\Gamma(\frac{\alpha_\sigma+n_\nu}{2})}{(\frac{q_\sigma+\psi_\nu}{2})^{\frac{\alpha_\sigma+n_\nu}{2}}} \times \int \frac{(\frac{q_\sigma+\psi_\nu}{2})^{\frac{\alpha_\sigma+n_\nu}{2}}}{\Gamma(\frac{\alpha_\sigma+n_\nu}{2})} (\sigma_\nu^2)^{-(\frac{\alpha_\sigma+n_\nu}{2}+1)} \times \exp\left\{-\frac{q_\sigma+\psi_\nu}{2\sigma_\nu^2}\right\}\, d\sigma_\nu^2,$$

since the integrand above is really $IG((\alpha_\sigma+n_\nu)/2,(q_\sigma+\psi_\nu)/2)$, the integral evaluates to 1, giving:

$$p(\mathbf{K}|\mathbf{Z},\boldsymbol{\beta}_0,\tau^2,\mathbf{W}) \propto \prod_\nu p(\mathbf{K}_\nu) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}\tau_\nu^{2m}|\mathbf{K}_\nu||\mathbf{W}|}\right)^{\frac{1}{2}} \times \frac{\left(\frac{q_\sigma}{2}\right)^{\frac{\alpha_\sigma}{2}}}{\left(\frac{q_\sigma+\psi_\nu}{2}\right)^{\frac{\alpha_\sigma+n_\nu}{2}}} \times \frac{\Gamma\left(\frac{\alpha_\sigma+n_\nu}{2}\right)}{\Gamma\left(\frac{\alpha_\sigma}{2}\right)}. \tag{34}$$

Eq. (34) can be used in place of the likelihood of the data conditional on all parameters. It can be thought of as a likelihood of the data, conditional on only the parameterization of $K(\cdot,\cdot)$. When computing a Metropolis-Hastings acceptance ratio for proposed $\mathbf{K}_\nu$ in a particular region r_ν, it suffices to use only the terms in (34) which contain some function of the imputed correlation matrix $\mathbf{K}_\nu$:

$$p(\mathbf{K}_\nu|\mathbf{Z}_\nu,\boldsymbol{\beta}_0,\tau_\nu^2,\mathbf{W}) \propto p(\mathbf{K}_\nu) \times \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{|\mathbf{K}_\nu|}\right)^{\frac{1}{2}} \times \left(\frac{q_\sigma+\psi_\nu}{2}\right)^{-\frac{\alpha_\sigma+n_\nu}{2}}. \tag{35}$$

Using the same ideas one can obtain the complete conditional of σ_ν^2 with $\boldsymbol{\beta}_\nu$ integrated out, which strangely enough involves the same ψ_ν quantity:

$$\begin{aligned}
p(\sigma_\nu^2|\mathbf{Z}_\nu,d_\nu,g_\nu,\boldsymbol{\beta}_0,\tau^2,\mathbf{W}) &= \int p(\boldsymbol{\beta}_\nu,\sigma_\nu^2|\mathbf{Z}_\nu,d_\nu,g_\nu,\boldsymbol{\beta}_0,\tau^2,\mathbf{W})\, d\boldsymbol{\beta}_\nu \\
&= p(\sigma_\nu^2) \int p(\mathbf{Z}_\nu|d_\nu,g_\nu,\boldsymbol{\beta}_\nu,\sigma_\nu^2) p(\boldsymbol{\beta}_\nu|\sigma_\nu^2,\boldsymbol{\beta}_0,\mathbf{W})\, d\boldsymbol{\beta}_\nu \\
&= \left(\frac{|\mathbf{V}_{\tilde{\beta}_\nu}|}{(2\pi)^{n_\nu}\tau_\nu^{2m}|\mathbf{K}_\nu||\mathbf{W}|}\right)^{\frac{1}{2}} \sigma_\nu^{-n_\nu} p(\sigma_\nu^2) \exp\left\{-\frac{\psi_\nu}{2\sigma_\nu^2}\right\} \\
&\propto \sigma_\nu^{-n_\nu} (\sigma_\nu^2)^{-(\alpha_\sigma/2+1)} \exp\left\{-\frac{q_\sigma}{2\sigma_\nu^2}\right\} \exp\left\{-\frac{\psi_\nu}{2\sigma_\nu^2}\right\} \\
&= (\sigma_\nu^2)^{-((\alpha_\sigma+n_\nu)/2+1)} \exp\left\{-\frac{q_\sigma+\psi_\nu}{2\sigma_\nu^2}\right\},
\end{aligned}$$

which means that

$$\sigma_\nu^2|d,g,\boldsymbol{\beta}_0,\mathbf{W} \sim IG((\alpha_\sigma+n_\nu)/2,(q_\sigma+\psi_\nu)/2). \tag{36}$$

References

Abrahamsen, P. (1997). "A Review of Gaussian Random Fields and Correlation Functions." Tech. Rep. 917, Norwegian Computing Center, Box 114 Blindern, N-0314 Oslo, Norway.

Adler, R. J. (1990). "An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes." Tech. rep., Institute of Mathematical Statistics, Hayward, CA.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman & Hall.

Berger, J. O., de Oliveira, V., and Sansó, B. (2001). "Objective Bayesian Analysis of Spatially Correlated Data." *Journal of the American Statistical Association*, 96, 456, 1361–1374.

Bernstein, D. (2005). *Matrix Mathematics*. Princeton, NJ: Princeton University Press.

Breiman, L. (1996). "Bagging Predictors." *Machine Learning*, 24, 2, 123–140.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART Model Search (with discussion)." *Journal of the American Statistical Association*, 93, 935–960.

— (2002). "Bayesian Treed Models." *Machine Learning*, 48, 303–324.

Chu, W., Keerthi, S. S., and Ong, C. J. (2004). "Bayesian Support Vector Regression using a Unified Loss Function." *IEEE Transactions on Neural Networks*, 15(1), 29–44.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1990). *Introduction to Algorithms*. The MIT Electrical Engineering and Computer Science Series. MIT Press/McGraw Hill.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, revised edition*. John Wiley and Sons.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). "Bayesian Estimation of Semiparametric Nonstationary Spatial Covariance Structure." *Environmetrics*, 12, 161–178.

Denison, D., Adams, N., Holmes, C., and Hand, D. (2002). "Bayesian Partition Modelling." *Computational Statistics and Data Analysis*, 38, 475–485.

Denison, D., Mallick, B., and Smith, A. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85, 363–377.

Dey, D., Müller, P., and Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York, NY, USA: Springer-Verlag New York, Inc.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. (1996). "Support Vector Regression Machines." In *Advances in Neural Information Processing Systems*, 155–161. MIT Press.

Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines." *Annals of Statistics*, 19, No. 1, 1–67.

Fuentes, M. and Smith, R. L. (2001). "A New Class of Nonstationary Spatial Models." Tech. rep., North Carolina State University, Raleigh, NC.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Baltimore, MD: Johns Hopkins.

Gramacy, R. B., Lee, H. K. H., and Macready, W. (2004). "Parameter Space Exploration With Gaussian Process Trees." In *ICML*, 353–360. Omnipress & ACM Digital Library.

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management*, 5, 81–102.

Higdon, D. (2002). "Space and Space-time Modeling Using Process Convolutions." In *Quantitative Methods for Current Environmental Issues*, eds. C. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, 37–56. London: Springer-Verlag.

Higdon, D., Swall, J., and Kern, J. (1999). "Non-Stationary Spatial Modeling." In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 761–768. Oxford University Press.

Hjort, N. L. and Omre, H. (1994). "Topics in Spatial Statistics." *Scandinavian Journal of Statistics*, 21, 289–357.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). "Bayesian Model Averaging: A Tutorial (with discussion)." *Statistical Science*, 14, 382–417.

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes." *Journal of the American Statistical Association*, 100, 653–668.

Matérn, B. (1986). *Spatial Variation*. 2nd ed. New York: Springer-Verlag.

Matheron, G. (1963). "Principles of Geostatistics." *Economic Geology*, 58, 1246–1266.

Neal, R. (1997). "Monte Carlo implementation of Gaussian process models for Bayesian regression and classification"." Tech. Rep. CRG–TR–97–2, Dept. of Computer Science, University of Toronto.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). "UCI Repository of Machine Learning Databases." *url:* http://www.ics.uci.edu/∼mlearn/MLRepository.html.

O'Hagan, A. (1991). "Bayes-Hermite quadrature." *Journal of Statistical Planning and Inference*, 29, 145–260.

Paciorek, C. (2003). "Nonstationary Gaussian Processes for Regression and Spatial Modelling." Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Rasmussen, C. and Ghahramani, Z. (2002). "Infinite Mixtures of Gaussian Process Experts." In *Advances in Neural Information Processing Systems*, vol. 14, 881–888. MIT Press.

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures With An Unknown Number of Components." *Journal of the Royal Statistical Society, Series B, Methodological*, 59, 731–758.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). "Design and Analysis of Computer Experiments." *Statistical Science*, 4, 409–435.

Sampson, P. D. and Guttorp, P. (1992). "Nonparametric Estimation of Nonstationary Spatial Covariance Structure." *Journal of the American Statistical Association*, 87(417), 108–119.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York, NY: Springer-Verlag.

Schmidt, A. M. and O'Hagan, A. (2003). "Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations." *Journal of the Royal Statistical Society, Series B*, 65, 745–758.

Silverman, B. W. (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Curve Fitting." *Journal of the Royal Statistical Society Series B*, 47, 1–52.

Stein, M. L. (1999). *Interpolation of Spatial Data*. New York, NY: Springer.

Wackernagel, H. (2003). *Multivariate Geostatistics*. Berlin: Springer-Verlag.

Whaley, R. C. and Petitet, A. (2004). "`ATLAS` (Automatically Tuned Linear Algebra Software)." Http://math-atlas.sourceforge.net/.