laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R

Robert B. Gramacy
Virginia Tech
Department of Statistics

Abstract

Gaussian process (GP) regression models make for powerful predictors in out of sample exercises, but cubic runtimes for dense matrix decompositions severely limit the size of data—training and testing—on which they can be deployed. That means that in computer experiment, spatial/geo-physical, and machine learning contexts, GPs no longer enjoy privileged status as data sets continue to balloon in size. We discuss an implementation of local approximate Gaussian process models, in the **laGP** package for R, that offers a particular sparse-matrix remedy uniquely positioned to leverage modern parallel computing architectures. The **laGP** approach can be seen as an update on the spatial statistical method of local *kriging* neighborhoods. We briefly review the method, and provide extensive illustrations of the features in the package through worked-code examples. The appendix covers custom building options for symmetric multi-processor and graphical processing units, and built-in wrapper routines that automate distribution over a simple network of workstations.

Keywords: sequential design, active learning, surrogate/emulator, calibration, local kriging, symmetric multi-processor, graphical processing unit, cluster computing, big data.

1. Introduction

The **laGP** package (Gramacy 2014) for R (R Core Team 2014) provides functions for (local) approximate Gaussian process modeling and prediction for large spatial data and the emulation of large computer experiments. This document provides a review of the underlying methodology, with background on conventional Gaussian process modeling, and worked code examples demonstrating most of the features of the package. There are several packages on the Comprehensive R Archive Network (CRAN, `cran.R-project.org`) which implement full (i.e., not approximated) Gaussian process regression. These include **mleGP** (Dancik 2013), **GPfit** (MacDoanld, Chipman, and Ranjan 2014), **spatial** (Venables and Ripley 2002), and **fields** (Nychka, Furrer, and Sain 2014)—all performing maximum likelihood (or *a' posteriori*) inference; and **tgp** (Gramacy 2007; Gramacy and Taddy 2010) and **spBayes** (Finley and Banerjee 2013)—performing fully Bayesian inference. Approximate methods for large-scale inference include **tgp** and **sparseEM** (Kaufman, Bingham, Habib, Heitmann, and Frieman 2012, which is not on CRAN). In what follows we motivate **laGP** by, in part, arguing that none of these methods (or their accompanying software) cope well with the modern scale of data collection/generation for spatial, computer experiment, or machine learning applications.

The **laGP** package also provides hooks that allow limited non-approximate inference. These subroutines have been carefully engineered to support the the package's main approximation features, and in their own right largely out-perform conventional alternatives in terms of data size capability. An important exception is the distributed computations offered by the **bigGP** package (Paciorek, Lipshitz, Prabhat, Kaufman, Zhuo, and Thomas 2014). As we discuss, an attractive feature of the nature of the approximations implemented by **laGP** is that they too can be parallelized in several ways.

1.1. Gaussian process regression and sparse approximation

The Gaussian process (GP) regression model, sometimes called a Gaussian spatial processes (GaSP), has been popular for decades in spatial data contexts like geostatistics (e.g., Cressie 1993) where they are known as *kriging* (Matheron 1963), and in computer experiments where they are deployed as *surrogate models* or *emulators* (Sacks, Welch, Mitchell, and Wynn 1989; Santner, Williams, and Notz 2003). More recently, they have become a popular prediction engine in the machine learning literature (Rasmussen and Williams 2006). The reasons are many, but the most important are probably that: the Gaussian structure affords a large degree of analytic capability not enjoyed by other general-purpose approaches to nonparametric nonlinear modeling; and because they perform well in out-of-sample tests. They are not, however, without their drawbacks. Two important ones are computational tractability and nonstationary flexibility, which we shall return to shortly.

A GP is technically a prior over functions (Stein 1999), with finite dimensional distributions defined by a mean $\mu(x)$ and positive definite covariance $\Sigma(x, x')$, for p-dimensional input(s) x and x'. For N input x-values this defines a μ_N N-vector and Σ_N positive definite $N \times N$ matrix whereby the output is a random N-vector $Y_N \sim \mathcal{N}_N(\mu_N, \Sigma_N)$. However, for regression applications a likelihood perspective provides a more direct view of the relevant quantities for inference and prediction. In that setup, N data (training) pairs $D_N = (X_N, Y_N)$ define a multivariate normal (MVN) likelihood for an N-vector of scalar responses Y_N through a small number of parameters θ that describe how X_N, an $(N \times p)$-dimensional design matrix, is related to μ_N and Σ_N. Linear regression is a special case where $\theta = (\beta, \tau^2)$ and $\mu_N = X_N\beta$ and $\Sigma_N = \tau^2 I_N$.

Whereas the linear case puts most of the "modeling" structure in the mean, GP regression focuses more squarely on the covariance structure. In many computer experiments contexts the mean is taken to be zero (e.g., Santner *et al.* 2003). This is a simplifying assumption we shall make throughout, although it is easy to generalize to a mean described by a polynomial basis. Let $K_\theta(x, x')$ be a correlation function so that $Y_N \sim \mathcal{N}_N(0, \tau^2 K_N)$ where K_N is a $N \times N$ positive definite matrix comprised of entries $K_\theta(x_i, x_j)$ from the rows of X_N. Here we are changing the notation slightly so that θ is reserved explicitly for K_θ, isolating τ^2 as a separate scale parameter. Choices of $K_\theta(\cdot, \cdot)$ determine stationarity, smoothness, differentiability, etc., but most importantly they determine the decay of spatial correlation.

A common first choice is the so-called *isotropic Gaussian*: $K_\theta(x, x') = \exp\{-\sum_{k=1}^{p}(x_k - x'_k)^2/\theta\}$, where correlation decays exponentially fast at rate θ. Since $K_\theta(x, x) = 1$ the resulting regression function is an interpolator, which is appropriate for many deterministic computer experiments. For smoothing noisy data, or for a more robust approach to modeling computer experiments (Gramacy and Lee 2011), a *nugget* can be added to $K_{\theta,\eta}(x, x') = K_\theta(x, x') + \eta^{\mathbb{I}_{\{x=x'\}}}$. Much of the technical work described below, and particularly in Section 2, is generic

to the particular choice of $K(\cdot,\cdot)$, excepting that it be differentiable in all parameters. The **laGP** package favors the isotropic Gaussian case. Many of the drawbacks of that overly simplistic choice, which leads theorists and practitioners alike to prefer other choices like the Matérn (Stein 1999), are less of a concern in our particular *local* approach to sparse approximation. The package also provides a limited set of routines that can accommodate a separable Gaussian correlation function; more details are provided in Section 3.2. Our empirical work will contain examples where correlation parameters (θ, η) are *both* estimated from data, however we emphasize cases where η is fixed at a small value which is typical for numerically robust near-interpolation of computer experiments.

1.2. Inference and prediction

GP regression is popular because inference (for all parameters but particularly for θ) is easy, and (out-of-sample) prediction is highly accurate and conditionally (on θ and η) analytic. It the spatial and computer experiments literatures it has become convention to deploy a reference $\pi(\tau^2) \propto 1/\tau^2$ prior (Berger, De Oliveira, and Sanso 2001) and obtain a marginal likelihood for the remaining unknowns:

$$p(Y_N|K_\theta(\cdot,\cdot)) = \frac{\Gamma[N/2]}{(2\pi)^{N/2}|K_N|^{1/2}} \times \left(\frac{\psi_N}{2}\right)^{-\frac{N}{2}} \quad \text{where} \quad \psi_N = Y_N^\top K_N^{-1} Y_N. \tag{1}$$

Derivatives are available analytically, leading to fast Newton-like schemes for maximizing. Some complications can arise when the likelihood is multi-modal for θ, however, where fully Bayesian inference may be preferred (e.g., Rasmussen and Williams 2006, Chapter 5).[1]

The predictive distribution $p(y(x)|D_N, K_\theta(\cdot,\cdot))$, is Student-$t$ with degrees of freedom N,

$$\text{mean} \qquad \mu(x|D_N, K_\theta(\cdot,\cdot)) = k_N^\top(x) K_N^{-1} Y_N, \tag{2}$$

$$\text{and scale} \qquad \sigma^2(x|D_N, K(\cdot,\cdot)) = \frac{\psi_N[K_\theta(x,x) - k_N^\top(x) K_N^{-1} k_N(x)]}{N}, \tag{3}$$

where $k_N^\top(x)$ is the N-vector whose i^{th} component is $K_\theta(x, x_i)$. Using properties of the Student-t, the variance of $Y(x)$ is $V_N(x) \equiv \mathbb{V}\text{ar}[Y(x)|D_N, K_\theta(\cdot,\cdot)] = \sigma^2(x|D_N, K_\theta(\cdot,\cdot)) \times N/(N-2)$.

As an example illustrating both inference and prediction, consider a simple sinusoidal "data set" treated as a deterministic computer simulation, i.e., modeled without noise.

```
R> X <- matrix(seq(0, 2 * pi,length = 6), ncol = 1)
R> Z <- sin(X)
```

The code below uses some low-level routines in the package to initialize a full GP representation with $\theta = 2$ and $\eta = 10^{-6}$. Then, a derivative-based MLE sub-routine is used to find $\hat{\theta}_{N=6}$, maximizing the expression in Equation 1.

```
R> gp <- newGP(X, Z, 2, 1e-6, dK = TRUE)
R> mleGP(gp, tmax=20)
```

[1] Equation 1 emphasizes $K_\theta(\cdot,\cdot)$, dropping η to streamline the notation in the following discussion. Everything applies to $K_{\theta,\eta}(\cdot,\cdot)$ as well.

```
$d
[1] 4.386202

$its
[1] 6
```

The output printed to the screen shows the inferred $\hat{\theta}_N$ value, called `d` in the package, and the number of Newton iterations required. The `mleGP` command alters the stored GP object (`gp`) to contain the new representation of the GP using $\hat{\theta}_{N=6}$. By default, `mleGP` maximizes over the lengthscale, however by specifying `param = "g"` it can maximize over the nugget η instead. The function `jmleGP` automates a profile-likelihood approach to "joint" optimization over lengthscale (θ/`d`) and nugget (η/`g`) values.

The code below obtains the parameters of the predictive equations on a grid of new x-values `XX`, following Equations 2–3.

```
R> XX <- matrix(seq(-1, 2 * pi + 1, length = 499), ncol = ncol(X))
R> p <- predGP(gp, XX)
R> deleteGP(gp)
```

The last line, above, frees the internal representation of the GP object, as we no longer need it to complete this example. The moments stored in `p` can be used to plot mean predictions and generate sample predictive paths via multivariate Student-t draws using the **mvtnorm** package (Genz, Bretz, Miwa, Mi, Leisch, Scheipl, and Hothorn 2014; Genz and Bretz 2009).

```
R> if(require("mvtnorm")) {
+ N <- 100
+ ZZ <- rmvt(N, p$Sigma, p$df)
+ ZZ <- ZZ + t(matrix(rep(p$mean, N), ncol = N))
+ } else { cat("this example is not doable without mvtnorm") }
```

Figure 1 provides a visualization of those sample paths on a scatter plot of the data. Each gray line, plotted by `matplot`, is a single random realization of $Y(x)|D_N, \hat{\theta}_N$. Observe how the predictive variance narrows for x nearby elements of X_N and expands out in a "football shape" away from them. This feature has attractive uses in design: high variance inputs represent sensible choices for new simulations (Gramacy and Lee 2009).

1.3. Supercomputing and sparse approximation for big data

Despite its many attractive features, GP regression implementations are buckling under the weight of the growing size of data sets in many modern applications. For example, supercomputers make submitting one job as easy as thousands, leading to ever larger computer simulation data. The problem is the $O(N^3)$ matrix decompositions required to calculate K_N^{-1} and $|K_N|$ in Equations 1–3. In practice that limits N to the mid-upper thousands for point inference, and lower thousands for sampling-based inference like the bootstrap or Bayesian MCMC. This has pushed some practitioners towards wholly new modeling apparatuses, say via trees (Pratola, Chipman, Gattiker, Higdon, McCulloch, and Rust 2014; Gramacy, Taddy, and Wild 2013; Chipman, Ranjan, and Wang 2012). Although trees offer an appealing divide-and-conquer approach, their obvious drawback is that they struggle to capture the smoothness known, in many cases, to exist in the physical and mathematical quantities being modeled.

```
R> if(require("mvtnorm")) {
+ matplot(XX, t(ZZ), col = "gray", lwd = 0.5, lty = 1, type = "l",
+         bty = "n", main = "simple sinusoidal example", xlab = "x",
+         ylab = "Y(x) | thetahat")
+ points(X, Z, pch = 19)
+ } else { plot(X, Z, pch = 19, main = "install mvtnorm!") }
```

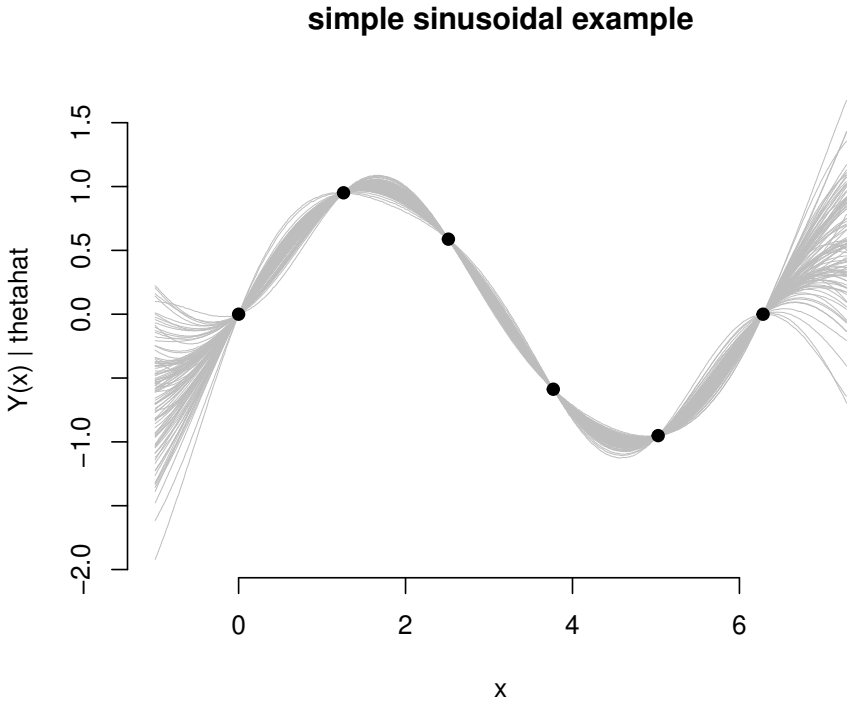

Figure 1: Predictions from fitted GP regression model on simple sinusoidal data.

One approach to salvaging GP inference for use in larger contexts has been to allocate supercomputer resources. Franey, Ranjan, and Chipman (2012) were the first to use graphical processing unit (GPU) linear algebra subroutines, extending the N by an order of magnitude. Paciorek, Lipshitz, Zhuo, Prabhat, Kaufman, and Thomas (2015) developed a package for R called **bigGP** (Paciorek *et al.* 2014) that combined symmetric-multiprocessor, cluster, and GPU facilities to gain yet another order of magnitude. Paciorek *et al.* were able to handle $N = 67275$. To go too far down that road, however, may miss the point in certain contexts. Computer model emulation is meant to *avoid* expensive computer simulation, not be a primary consumer of it.

An orthogonal approach is to perform approximate GP regression, and a common theme in that literature is sparsity, leading to fast matrix decompositions (e.g., Kaufman *et al.* 2012; Sang and Huang 2012). Again, the expansion of capability is one-to-two orders of magnitude, albeit without tapping supercomputer resources which is more practical for most applications. For example, Kaufman *et al.* reported on an experiment with $N = 20000$. Some approaches in a similar vein include fixed rank kriging (Cressie and Johannesson 2008) and

using ‘’‘pseudo-inputs” (Snelson and Ghahramani 2006).

Hybrid approximate GP regression and big-computer resources have been combined to push the boundary even farther. Eidsvik, Shaby, Reich, Wheeler, and Niemi (2014) suggest composite likelihood approach, rather than directly leveraging a sparse matrix library, and when combined with a GPU implementation their method is able to cope with $N = 173405$. This represents a substantial inroad into retaining many of the attractive features of GP regression in larger data applications. However, a larger (and thriftier) capability would certainly be welcome. Pratola *et al.* (2014) found it necessary to modify a tree-based approach for distribution over the nodes of a supercomputer in order to handle an $N = 7$M sized design.

The remainder of the paper is outlined as follows. In Section 2 we discuss the local approximate Gaussian process method for large scale inference and prediction. Several variations are discussed, including parallelized and GPU versions for combining with supercomputing resources in order to handle large-N problems in reasonable computation times (e.g., under an hour). Live-code examples, demonstrating the features of the **laGP** package for R, are peppered throughout paper, however Sections 3 and 4 are devoted to larger scale and more exhaustive illustration: first demonstrating local emulation/regression/smoothing and then with application to large scale computer model calibration. Section 5 discusses extra features, and the potential for end-user customization. Appendix A discusses default priors, and Appendix B describes how the package can be compiled to enable SMP and GPU support, as well as a variation a the key wrapper function `aGP` enabling distribution of predictions across the nodes of a cluster.

2. Local approximate Gaussian process models

The methods in the **laGP** package take a two-pronged approach to large data GP regression. They (1) leverage sparsity, but in fact only work with small dense matrices. And (2) the many-independent nature of calculations facilitates massive parallelization. The result is an approximate GP regression capability that can accommodate orders of magnitude larger training and testing sets than ever before. The method can be seen as a modernization of *local kriging* from the spatial statistics literature (Cressie 1993, pp,131–134). It involves approximating the predictive equations at a particular generic location, x, via a subset of the data $D_n(x) \subseteq D_N$, where the sub-design $X_n(x)$ is (primarily) comprised of X_N close to x. The thinking is that, with the typical choices of $K_\theta(x, x')$, where correlation between elements $x' \in X_N$ decays quickly for x' far from x, remote x's have vanishingly small influence on prediction. Ignoring them in order to work with much smaller, n-sized, matrices will bring a big computational savings with little impact on accuracy.

This is a sensible idea: It can be shown to induce a valid stochastic process (Datta, Banerjee, Finley, and Gelfand 2015); when $n \ll 1000$ the method is fast and accurate, and as n grows the predictions increasingly resemble their full N-data counterparts; and, for smaller n, $V_n(x)$ is organically inflated relative to $V_N(x)$, acknowledging greater uncertainty in approximation. The simplest version of such a scheme would be via nearest neighbors (NN): $X_n(x)$ comprised of closest elements of X_N to x. Emory (2009) showed that this works well for many common choices of K_θ. However, NN designs are known to be sub-optimal (Vecchia 1988; Stein, Chi, and Welty 2004) as it pays to have some spread in $X_n(x)$ in order to obtain good estimates of correlation hyperparameters like θ. Still, searching for the optimal sub-design, which involves

choosing n from N things, is a combinatorially huge undertaking.

Gramacy and Apley (2015) showed how a greedy search could provide designs $X_n(x)$ where predictors based on $D_n(x)$ out-performed the NN alternative out-of-sample, yet required no more computational effort than NN, i.e., they worked in $O(n^3)$ time. The idea is to search iteratively, starting with a small NN set $D_{n_0}(x)$, and choosing x_{j+1} to augment $X_j(x)$ to form $D_{j+1}(x)$ according to one of several simple objective criteria. Importantly, they showed that the criteria they chose, on which we elaborate below, along with the the other relevant GP quantities for inference and prediction (Equations 1–3) can be calculated, or updated as $j \to j+1$, in $O(j^2)$ time as long as the parameters, θ, remain constant across iterations. Therefore over the entirety of $j = n_0, \ldots, n$ iterations the scheme is in $O(n^3)$. The idea of sequential updating for GP inference is not new (Gramacy and Polson 2011; Haaland and Qian 2011), however the focus of previous approaches has been global. Working local to particular x brings both computational and modeling/accuracy advantages.

2.1. Criterion for local design

Gramacy and Apley considered two criteria in addition to NN, one being a special case of the other. The first is to minimize the empirical Bayes mean-square prediction error (MSPE)

$$J(x_{j+1}, x) = \mathbb{E}\{[Y(x) - \mu_{j+1}(x|D_{j+1}, \hat{\theta}_{j+1})]^2 | D_j(x)\}$$

where $\hat{\theta}_{j+1}$ is the estimate for θ based on D_{j+1}. The predictive mean $\mu_{j+1}(x|D_{j+1}, \hat{\theta}_{j+1})$ follows Equation 2, except that a $j+1$ subscript has been added to indicate dependence on x_{j+1} and the future, unknown y_{j+1}. They then derive the approximation

$$J(x_{j+1}, x) \approx V_j(x|x_{j+1}; \hat{\theta}_j) + \left(\frac{\partial \mu_j(x;\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}_j} \right)^2 \Big/ \mathcal{G}_{j+1}(\hat{\theta}_j). \tag{4}$$

The first term in Equation 4 estimates variance at x after x_{j+1} is added into the design,

$$V_j(x|x_{j+1};\theta) = \frac{\psi_j v_{j+1}(x;\theta)}{j-2}, \quad \text{where} \quad v_{j+1}(x;\theta) = \left[K_{j+1}(x,x) - k_{j+1}^\top(x) K_{j+1}^{-1} k_{j+1}(x)\right]. \tag{5}$$

Minimizing predictive variance at x is a sensible goal. The second term in Equation 4 estimates the rate of change of the predictive mean at x, weighted by the expected *future* inverse information, $\mathcal{G}_{j+1}(\hat{\theta}_j)$, after x_{j+1} and the corresponding y_{j+1} are added into the design. The weight, which is constant in x comments on the value of x_{j+1} for estimating the parameter of the correlation function, θ, by controlling the influence of the rate of change (derivative) of the predictive mean at x on the overall criteria.

The influence of that extra term beyond the reduced variance is small. The full MSPE criteria tends to yield qualitatively similar local designs $X_n(x)$ as ones obtained using just $V_j(x|x_{j+1};\hat{\theta}_j)$, which incurs a fraction of the computational cost (since no derivative calculations are necessary). This simplified criteria is equivalent to choosing x_{j+1} to maximize *reduction* in variance:

$$\begin{aligned} &v_j(x;\theta) - v_{j+1}(x;\theta), \quad \text{(dropping } \theta \text{ below for compactness)} \\ &\quad = k_j^\top(x) G_j(x_{j+1}) v_j(x_{j+1}) k_j(x) + 2 k_j^\top(x) g_j(x_{j+1}) K(x_{j+1}, x) + K(x_{j+1}, x)^2 / v_j(x_{j+1}), \end{aligned} \tag{6}$$

where $G_j(x') \equiv g_j(x')g_j^\top(x')$, and $g_j(x') = -K_j^{-1}k_j(x')/v_j(x')$. Observe that only $O(j^2)$ calculations are required above. Although known for some time in other contexts, Gramacy and Apley chose the acronym ALC to denote use of that decomposition in local design, recognizing similar approach to *global* design via a method called *active learning Cohn* (1996).

To illustrate local designs derived under greedy application of both criteria, consider the following gridded global design in $[-2, 2]^2$.

```
R> x <- seq(-2, 2, by = 0.02)
R> X <- as.matrix(expand.grid(x, x))
R> N <- nrow(X)
```

Here we have $N = 40401$, a very large design by traditional GP standards. You cannot invert an $N \times N$ matrix for N that big on even the best modern workstation. As a point of reference, it takes about seven seconds to perform a single decomposition of an 4000×4000 matrix using hyperthreaded libraries on a 2010 iMac.

The `laGP` function requires a vector of responses to perform local design, even though the design itself doesn't directly depend on the responses—a point which we will discuss at greater length shortly. The synthetic response Gramacy and Apley used for illustrations is coded below, and we shall elucidate the nature of input/output relationships therein in due course.

```
R> f2d <- function(x)
+   {
+     g <- function(z)
+       return(exp( - (z - 1)^2) + exp( -0.8 * (z + 1)^2)
+         - 0.05 * sin(8 * (z + 0.1)))
+     -g(x[,1]) * g(x[,2])
+   }
R> Y <- f2d(X)
```

Now, consider a prediction location x, denoted by `Xref` in the code below, and local designs for prediction at that x based on MSPE and ALC criteria.

```
R> Xref <- matrix(c(-1.725, 1.725), nrow = 1)
R> p.mspe <- laGP(Xref, 6, 50, X, Y, d = 0.1, method="mspe")
R> p.alc <- laGP(Xref, 6, 50, X, Y, d = 0.1, method="alc")
```

Both designs use $n_0 = 6$ nearest neighbors to start, make greedy selections until $n = 50$ locations are chosen, and use $\theta = 0.1$. The output object from `laGP` contains indices into the original design. Those locations, and the order in which they were chosen, are plotted in Figure 2. They are not identical under the two criteria, but any qualitative differences are subtle. Both contain a clump of nearby points with satellite points emanating along rays from x, the green dot. The satellite points are still relatively close to x considering the full scope of locations in X_N—all locations chosen are in the upper-left quadrant of the space.

It is perhaps intriguing that the greedy local designs differ from NN ones. An exponentially decaying $K_\theta(\cdot, \cdot)$, like our isotropic Gaussian choice, should substantially devalue locations far from x. Gramacy and Haaland (2015) offer an explanation, which surprisingly has little to do with the particular choice of K_θ. The explanation lies the form of Equation 6. Although

```
R> Xi <- rbind(X[p.mspe$Xi, ], X[p.alc$Xi, ])
R> plot(X[p.mspe$Xi, ], xlab = "x1", ylab = "x2", type = "n",
+   main = "comparing local designs", xlim = range(Xi[ ,1]),
+   ylim = range(Xi[ ,2]))
R> text(X[p.mspe$Xi, ], labels = 1:length(p.mspe$Xi), cex = 0.7)
R> text(X[p.alc$Xi, ], labels = 1:length(p.alc$Xi), cex = 0.7, col = 2)
R> points(Xref[1], Xref[2], pch=19, col=3)
R> legend("topright", c("mspe", "alc"), text.col = c(1, 2), bty="n")
```

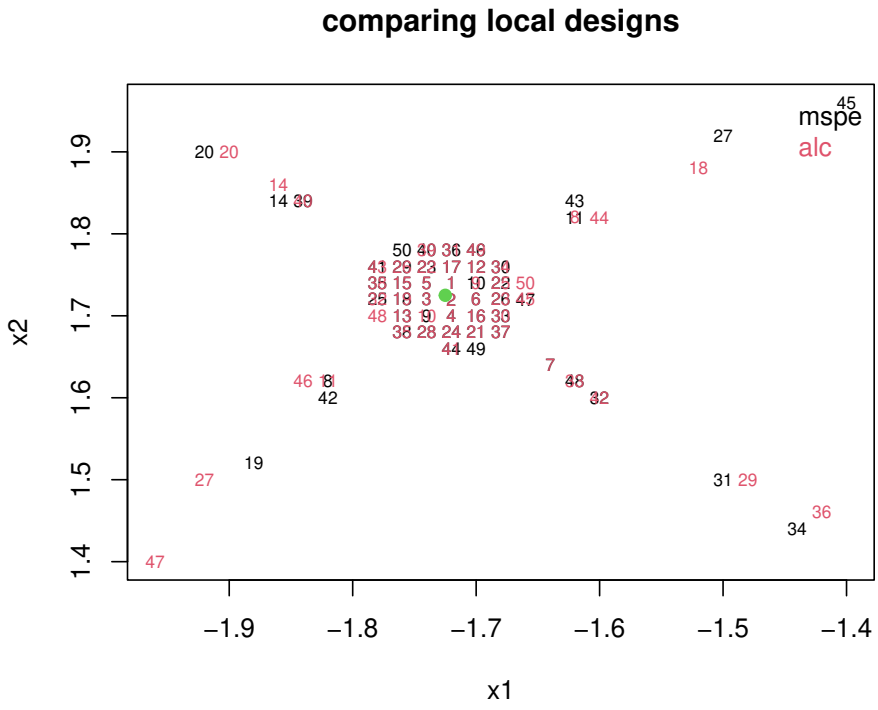

Figure 2: Local designs at x (green dot), derived under MSPE and ALC criteria.

quadratic in $K_\theta(x_{j+1}, x)$, the "distance" between the x and the potential new local design location x_{j+1}, it is also quadratic in $g_j(x_{j+1})$, a vector measuring "inverse distance", via K_j^{-1}, between x_{j+1} and the current local design $X_j(x)$. So the criteria makes a tradeoff: minimize "distance" to x while maximizing "distance" (or minimizing "inverse distance") to the existing design. Or in other words, the potential value of new design element (x_{j+1}, y_{j+1}) depends not just on its proximity to x, but also on how potentially different that information is to where we already have (lots of) it, at $X_j(x)$.

Returning to the code example, we see below that the predictive equations are also very similar under both local designs.

```
R> p <- rbind(c(p.mspe$mean, p.mspe$s2, p.mspe$df),
+   c(p.alc$mean, p.alc$s2, p.alc$df))
R> colnames(p) <- c("mean", "s2", "df")
```

```
R> rownames(p) <- c("mspe", "alc")
R> p
```

```
           mean           s2 df
mspe -0.3725312 2.518566e-06 50
alc  -0.3724820 2.445078e-06 50
```

Although the designs are built using a fixed $\theta = 0.1$, the predictive equations output at the end use local MLE calculation given the data $D_n(x)$.

```
R> p.mspe$mle
```

```
          d dits
1 0.3588186    7
```

```
R> p.alc$mle
```

```
          d dits
1 0.3378369    7
```

MLE calculations can be turned off by adjusting the `laGP` call to include `d=list(start=0.1, mle=FALSE)` as an argument. More about local inference for θ is deferred until Section 2.2. For now we note that the implementation is same as the one behind the `mleGP` routine described earlier in Section 1.2, under modest regularization [see Appendix A].

Finally, both local design methods are fast,

```
R> c(p.mspe$time, p.alc$time)
```

```
elapsed elapsed
  0.165   0.053
```

though ALC is about 3.1 times faster since it doesn't require evaluation of derivatives. Although a more thorough out-of-sample comparison on both time and accuracy fronts is left to Section 3, the factor of (at least) two speedup in execution time, together with the simpler implementation, led Gramacy and Apley to prefer ALC in most cases.

2.2. Global inference, prediction and parallelization

The simplest way to extend the analysis to cover a dense design of predictive locations $x \in \mathcal{X}$ is to serialize: loop over each x collecting approximate predictive equations, each in $O(n^3)$ time. For $T = |\mathcal{X}|$ the total computational time is in $O(Tn^3)$. Obtaining each of the full GP sets of predictive equations, by contrast, would require computational time in $O(TN^2 + N^3)$, where the latter N^3 is attributable to obtaining K^{-1}.[2] One of the nice features of standard GP emulation is that once K^{-1} has been obtained the computations are fast $O(N^2)$ operations

[2] If only the predictive mean is needed, and not the variance, then the time reduces to $O(TN + N^3)$.

for each location x. However, as long as $n \ll N$ our approximate method is even faster despite having to rebuild and re-decompose $K_j(x)$'s for each x.

The approximation at x is built up sequentially, but completely independently of other predictive locations. Since a high degree of local spatial correlation is a key modeling assumption this may seem like an inefficient use of computational resources, and indeed it would be in serial computation for each x. However, independence allows trivial parallelization requiring token programmer effort. When compiled correctly [see Appendix B.1] the **laGP** package can exploit symmetric multiprocessor (SMP) parallelization via `OpenMP` pragmas in its underlying `C` implementation. The simplest way this is accomplished is via a "parallel-for" pragma.

```
#ifdef _OPENMP
  #pragma omp parallel for private(i)
#endif
  for(i = 0; i < npred; i++) { ...
```

That is actual code from an early implementation, where `npred` $= |\mathcal{X}|$, leading to a nearly linear speedup: runtimes for P processors scale roughly as $1/P$. Later versions of the package use the "`parallel`" pragma which involves more code but incurs slightly less overhead.

To illustrate, consider the following predictive grid in $[-2, 2]^2$ spaced to avoid the original $N = 40$K design.

```
R> xx <- seq(-1.97, 1.95, by = 0.04)
R> XX <- as.matrix(expand.grid(xx, xx))
R> YY <- f2d(XX)
```

The `aGP` function iterates over the elements of $\tilde{X} \equiv$ `XX`. The package used in this illustration is compiled for `OpenMP` support, and the `omp.threads` argument controls the number of threads used by `aGP`, divvying up `XX`. You can specify any positive integer for `omp.threads`, however a good rule-of-thumb is to match the number of cores. Here we set the default to two, since nearly all machines these days have at least one hyperthreaded core (meaning it behaves like two cores). However, this can be overwritten by the `OMP_NUM_THREADS` environment variable.

```
R> nth <- as.numeric(Sys.getenv("OMP_NUM_THREADS"))
R> if(is.na(nth)) nth <- 2
R> print(nth)

[1] 8
```

If your machine has fewer cores, if your **laGP** is not compiled with `OpenMP` or if your operating system caps the number of `OpenMP` threads to a lower value (see Appendix B.1), then it will take longer to run the examples here.

```
R> P.alc <- aGP(X, Y, XX, omp.threads = nth, verb = 0)
```

Note that the default method is ALC. The results obtained with `method = "mspe"` are similar, but require more computation time. Further comparison is delayed until Section 3. The `verb = 0` argument suppresses a progress meter that is otherwise printed to the screen.

```
R> persp(xx, xx, -matrix(P.alc$mean, ncol = length(xx)), phi=45, theta=45,
+       main = "", xlab = "x1", ylab = "x2", zlab = "yhat(x)")
```

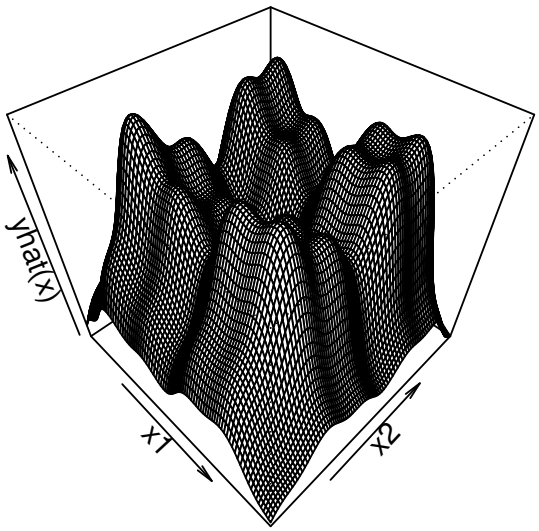

Figure 3: Emulated surface based on $N = 40\text{K}$ and $|\mathcal{X}| = 10\text{K}$ gridded predictive locations.

Figure 3 shows the resulting (predictive mean) emulation surface.[3] Although the input dimension is low, the input-output relationship is nuanced and merits a dense design in the input space to fully map.

For a closer look, Figure 4 shows a slice through that predictive surface at $x_2 = 0.51$ along with the true responses (completely covered by the prediction) and error-bars. Observe that the error bars are very tight on the scale of the response, and that although no continuity is enforced—calculations at nearby locations are independent and potentially occur in parallel—the resulting surface looks smooth to the eye. This is not always the case, as we illustrate in Section 3.3.

Accuracy, however, is not uniform. Figure 5 shows that predictive bias oscillates across the same slice of the input space shown in Figure 4. Crucially, however, notice that the magnitude of the bias is small: one-hundredth of a tick on the scale of the response. Still, given the density of the input design one could easily guess that the model may not be flexible enough to characterize the fast-moving changes in the input-output relationships.

Although an approximation, the local nature of modeling means that, from a global perspective, the predictor is *more* flexible that the full-N stationary Gaussian process predictor. Here, *stationary* loosely means that the covariance structure is modeled uniformly across the input space. Most choices of $K_\theta(\cdot, \cdot)$, like the isotropic Gaussian we use, induce stationarity in the spatial field. Inferring separate independent predictors across the elements of a vast predictive grid lends `aGP` a degree of nonstationarity. In fact, by default, `aGP` goes beyond that by learning separate $\hat{\theta}_n(x)$ local to each $x \in \mathcal{X}$ by maximizing the local likelihoods (or posterior probabilities). Figure 6 shows that, indeed, the estimated lengthscales vary spatially. So even though the spatial field may be *locally* restricted to isotropy, and therefore assumes

[3] The negative is shown for better visibility.

```
R> med <- 0.51
R> zs <- XX[, 2] == med
R> sv <- sqrt(P.alc$var[zs])
R> r <- range(c(-P.alc$mean[zs] + 2 * sv, -P.alc$mean[zs] - 2 * sv))
R> plot(XX[zs,1], -P.alc$mean[zs], type="l", lwd = 2, ylim = r, xlab = "x1",
+      ylab = "predicted & true response", bty = "n",
+      main = "slice through surface")
R> lines(XX[zs, 1], -P.alc$mean[zs] + 2 * sv, col = 2, lty = 2, lwd = 2)
R> lines(XX[zs, 1], -P.alc$mean[zs] - 2 * sv, col = 2, lty = 2, lwd = 2)
R> lines(XX[zs, 1], YY[zs], col = 3, lwd = 2, lty = 3)
```

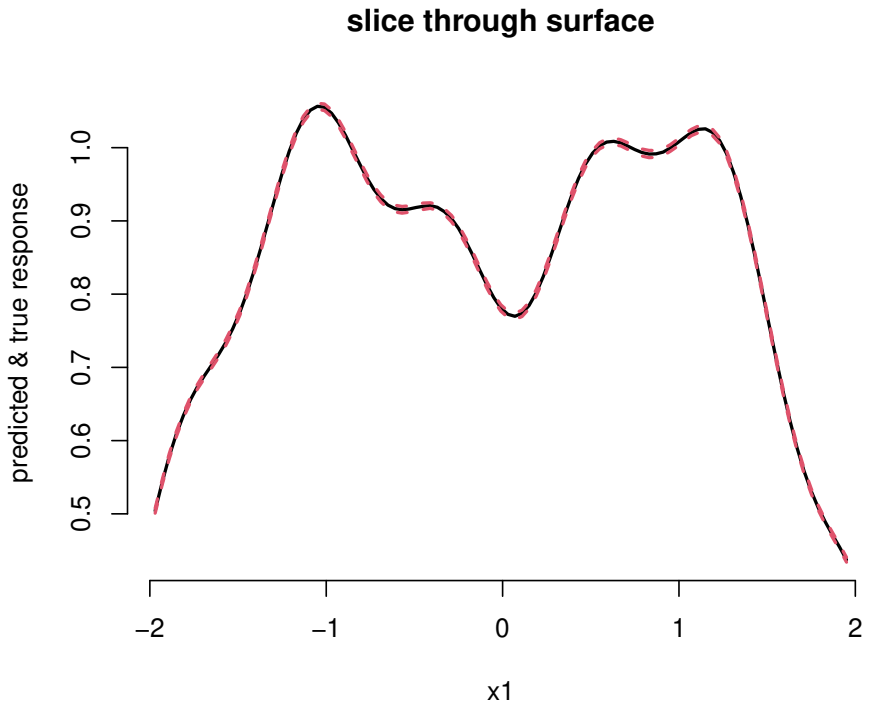

Figure 4: Slice of the predictive surface shown in Figure 3 including the true surface [covered by the mean] and predictive interval.

stationarity to a certain extent, *globally* the characteristics of the field are less constrained. Nevertheless, even the extra degree of flexibility afforded by spatially varying $\hat{\theta}_n(x)$ is not enough to entirely mitigate the small amount of bias shown in Figure 5.

Several enhancements offer scope for improvement. One is to explicitly accommodate global anisotropy with a separable correlation structure. A simple way to do that is discussed in Section 3.2. Another is to refine the local analysis, enhancing the degree of nonstationarity. Gramacy and Apley recommend a two-stage scheme wherein the above process is repeated and new $X_n(x)$ are chosen conditional upon $\hat{\theta}_n(x)$ values from the first stage. i.e., so that the second iteration's local designs use locally estimated parameters. This leads to a globally nonstationary model and generally more accurate predictions than the single-stage scheme.

```
R> diff <- P.alc$mean - YY
R> plot(XX[zs,1], diff[zs], type = "l", lwd = 2,
+      main = "systematic bias in prediction",
+      xlab = "x1", ylab = "y(x) - yhat(x)", bty = "n")
```

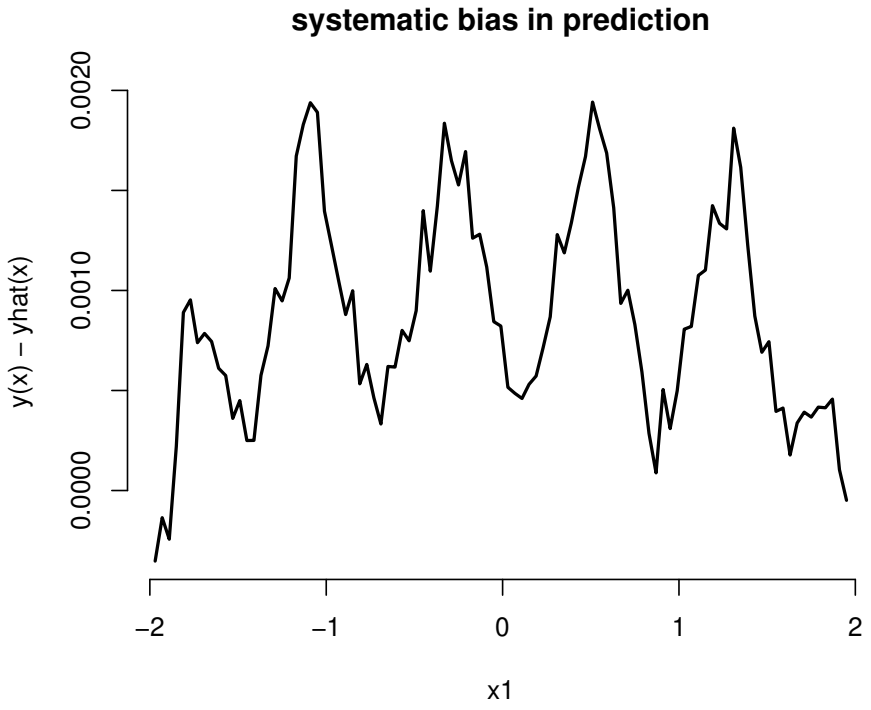

Figure 5: Bias in the predictive mean surface shown the same slice as in Figure 4.

The full scheme is outlined algorithmically in Figure 7. Step 2(b) of the algorithm implements the ALC reduction in variance scheme, via Equation 6, although MSPE (Equation 4) or any other criteria could be deployed there, at each greedy stage of local design. Of course, more than two repetitions of the global search scheme can be performed, but in many examples two has been sufficient to achieve rough convergence of the overall iterative scheme. Optionally, the $\hat{\theta}_n(x)$-values can be smoothed (e.g., by `loess`, as illustrated in Figure 4) before they are fed back into the local design schemes. Smoothing can guard against extreme and abrupt changes in lengthscale from one stage to the next. Considering other popular approaches to adapting a stationary model to achieve nonstationary surfaces—usually involving orders of magnitude more computation (e.g., Schmidt and O'Hagan 2003, and references therein)—this small adaptation is a thrifty alternative that does not change the overall computational order of the scheme.

Consider the following illustration continuing on from our example above.

```
R> P.alc2 <- aGP(X, Y, XX, d = exp(lo$fitted), omp.threads = nth, verb = 0)
```

This causes the design, for each element of `XX`, to initialize search based on the smoothed `d`-values output from the previous `aGP` run. Comparing the predictions from the first iteration

```
R> plot(XX[zs,1], P.alc$mle$d[zs], type = "l", lwd=2,
+      main = "spatially varying lengthscale",
+      xlab = "x1", ylab = "thetahat(x)", bty = "n")
R> df <- data.frame(y = log(P.alc$mle$d), XX)
R> lo <- loess(y ~ ., data = df, span = 0.01)
R> lines(XX[zs,1], exp(lo$fitted)[zs], col=2, lty=2, lwd=2)
R> legend("topright", "loess smoothed", col=2, lty=2, lwd=2, bty="n")
```

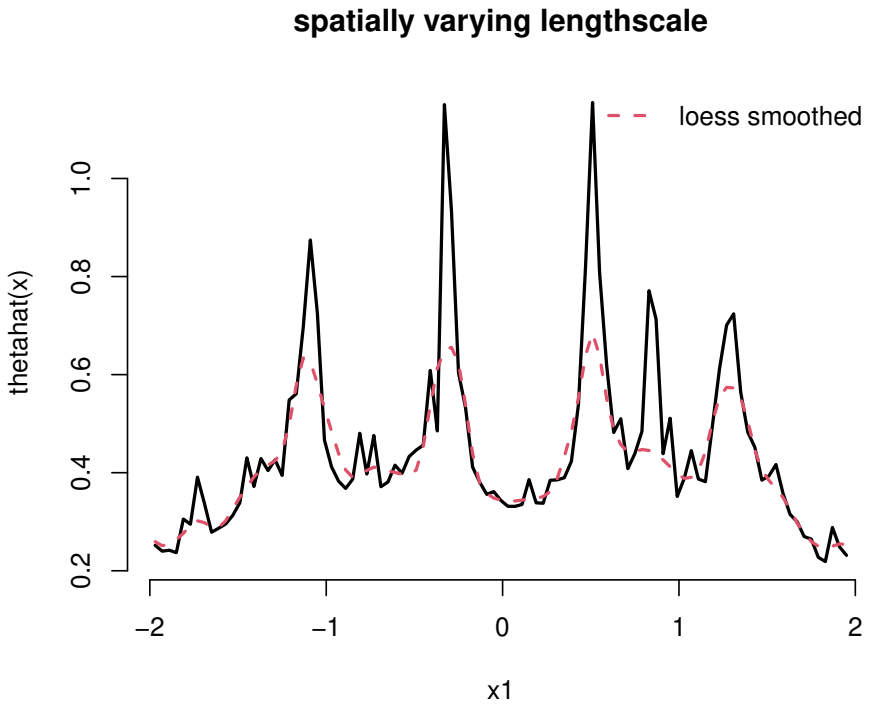

Figure 6: Spatially varying lengthscale estimated along the slice shown in Figure 4.

to those from the second, we can see that the latter has lower RMSE.

```
R> rmse <- data.frame(alc = sqrt(mean((P.alc$mean - YY)^2)),
+   alc2 = sqrt(mean((P.alc2$mean - YY)^2)))
R> rmse

           alc         alc2
1 0.0006227472 0.0003031463
```

This result is not impressive, but it is statistically significant across a wide range of examples. For example Gramacy and Apley (2015) provided an experiment based on the borehole data [more in Section 3] showing that the second iteration consistently improves upon predictions from the first. Although explicitly facilitating a limited degree of nonstationarity, second stage local designs do not solve the bias problem completely. The method is still locally

1. Choose a sensible starting global $\theta_x = \theta_0$ for all x.
2. Calculate local designs $X_n(x, \theta_x)$ based on ALC, independently for each x:
   (a) Choose a NN design $X_{n_0}(x)$ of size n_0.
   (b) For $j = n_0, \ldots, n-1$, set
   $$x_{j+1} = \arg \max_{x_{j+1} \in X_N \setminus X_j(x)} v_j(x; \theta_x) - v_{j+1}(x; \theta_x),$$
   and then update $D_{j+1}(x, \theta_x) = D_j(x, \theta_x) \cup (x_{j+1}, y(x_{j+1}))$.
3. Also independently, calculate the MLE $\hat{\theta}_n(x)|D_n(x, \theta_x)$ thereby explicitly obtaining a globally nonstationary predictive surface. Set $\theta_x = \hat{\theta}_n(x)$.
4. Repeat steps 2–3 as desired.
5. Output predictions $Y(x)|D_n(x, \theta_x)$ for each x.

Figure 7: Multi-stage approximate local GP modeling algorithm.

stationary, and indeed locally isotropic in its **laGP** implementation. Finally, we note that subsequent stages of design tend to be slightly faster than earlier stages since the number of Newton iterations required for $\hat{\theta}_n(x)$ is reduced given refined starting values for search.

2.3. Computational techniques for speeding up local search

The most expensive step in Algorithm 7 is the inner-loop of Step 2(b), iterating over all $N - j$ remaining candidates in $X_N \setminus X_j(x)$ in search of X_{j+1}. Assuming the criteria involves predictive variance (Equation 3) in some way, every candidate entertained involves an $O(j^2)$ calculation. Viewed pessimistically, one could argue the scheme actually requires computation in $O(Nn^3)$ not $O(n^3)$. However, there are several reasons to remain optimistic about computational aspects. One is that $O(Nn^3)$ is not $O(N^3)$. The others require more explanation, and potentially slight adjustments in implementation.

Not all $N - j$ candidates need be entertained for the method to work well. For the same reason prediction is localized to x in the first place, that correlation decays quickly away from x, we can usually afford to limit search to $N' \ll N - j$ candidates near to x. By default, `laGP` and `aGP` limit search to the nearest $N' = 1000$ locations, although this can be adjusted with the `close` argument. One can check [not shown here] that increasing `close` by an order of magnitude, to 2000 or 10,000 uses more compute cycles but yields identical results in the applications described in this document.

But it is risky to reduce `close` too much, as doing so will negate the benefits of search, eventually yielding the NN GP predictor. Another option, allowing N' to be greatly increased if desired, is to deploy further parallelization. Gramacy, Niemi, and Weiss (2014) showed that ALC-based greedy search is perfect for GPU parallelization. Each potential candidate, up to 65K candidates, can be entertained on a separate GPU block, and threads within that block can be used to perform many of the required dense linear algebra operations in Equation 6

in parallel. In practice they illustrate that this can result in speedups of between twenty and seventy times, with greater efficiencies for large n and N'. Enabling GPU subroutines requires custom compilation of CUDA source code via the NVIDIA compiler `nvcc` and re-compilation of the C code in the **laGP** package. For more details see Appendix B.2. For best results, enabling `OpenMP` support [Appendix B.1] is also recommended.

Finally, Gramacy and Haaland (2015) suggested that the discrete and exhaustive nature of search could be bypassed all together. They studied the topology of the reduction in variance landscape—the spatial surface searched in Step 2(b) via Equation 6—and observed that many regularities persist over choices of $K_\theta(\cdot,\cdot)$ and its parameterization. As long as X_N is reasonably space-filling, local designs predictably exhibit the features observed in Figure 2: a substantial proportion of NNs accompanied by farther out satellite points positioned roughly along rays emanating from the reference predictive location, x. To mimic that behavior without exhaustive search they proposed a continuous one-dimensional line search along rays emanating from x. Optimizing along the ray is fast and can be implemented with library routines, like `Brent_fmin` (Brent 1973), the workhorse behind R's `optimize` function.

The code below calculates such an ALC-ray based design, augmenting our example from Section 2.

```
R> p.alcray <- laGP(Xref, 6, 50, X, Y, d = 0.1, method = "alcray")
```

Although a similar idea could be deployed for finding MSPE-based designs based on rays, this is not implemented in the **laGP** package at the present time. Figure 8 compares local designs based on ray and exhaustive search. The exhaustive search design is identical to the ALC one shown in Figure 2, and just like in that example the ray-based version is not identical to the others but clearly exhibits similar qualitative features. The time required to derive the ALC-ray local design is:

```
R> p.alcray$time

elapsed
  0.012
```

and this is 4.4 times better than the exhaustive alternative. The predictive equations are nearly identical.

```
R> p <- rbind(p, c(p.alcray$mean, p.alcray$s2, p.alcray$df))
R> rownames(p)[3] <- c("alcray")
R> p

             mean           s2 df
mspe   -0.3725312 2.518566e-06 50
alc    -0.3724820 2.445078e-06 50
alcray -0.3723245 1.840874e-06 50
```

Gramacy and Haaland recommend using p rays per greedy search iteration, where p is the dimension of the input space. However this can be adjusted with the `numrays` argument, fine-tuning the exhaustiveness of search relative to the computational expense.

To complete the picture, the code below performs two stage global/local design based on ALC-ray searches.

```
R> plot(X[p.alc$Xi,], xlab = "x1", ylab = "x2", type = "n",
+   main="comparing local designs", xlim = range(Xi[ ,1]),
+   ylim = range(Xi[ ,2]))
R> text(X[p.alc$Xi,], labels = 1:length(p.alc$Xi), cex = 0.7, col = 2)
R> text(X[p.alcray$Xi,], labels=1:length(p.mspe$Xi), cex=0.7, col = 3)
R> points(Xref[1], Xref[2], pch = 19, col = 3)
R> legend("topright", c("alc", "alcray"), text.col = c(2,3), bty = "n")
```

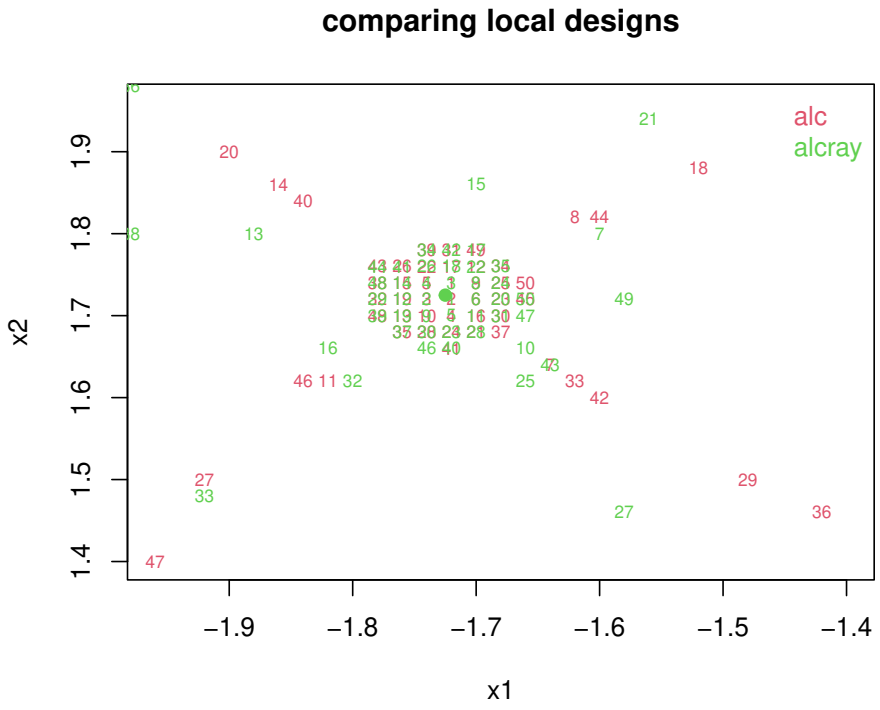

Figure 8: Local designs at x (green dot), derived under ALC and ALC-ray search criteria.

```
R> P.alcray <- aGP(X, Y, XX, method = "alcray", omp.threads = nth, verb = 0)
R> dfray <- data.frame(y = log(P.alcray$mle$d), XX)
R> loray <- loess(y ~ ., data = dfray, span = 0.01)
R> P.alcray2 <- aGP(X, Y, XX, method = "alcray", d = exp(loray$fitted),
+   omp.threads = nth, verb = 0)
```

The result is a global predictor that is 5.7 times faster than the non-raw version, echoing the single-x results from `laGP` above

```
R> c(P.alcray$time, P.alcray2$time)
```

```
elapsed elapsed
 13.722  10.782
```

and provides nearly identical out-of-sample accuracy via RMSE:

```
R> rmse <- cbind(rmse,
+   data.frame(alcray=sqrt(mean((P.alcray$mean - YY)^2)),
+     alcray2=sqrt(mean((P.alcray2$mean - YY)^2))))
R> rmse

           alc          alc2       alcray      alcray2
1 0.0006227472 0.0003031463 0.0004478262 0.0002044841
```

25 seconds on a 2010 desktop to accurately emulate at 10K locations from an input design of $N = 40$K is an unmatched capability in the recent computer experiment literature.

3. Examples

The 2-d example above, while illustrative, was somewhat simplistic. Below we present three further examples that offer a more convincing demonstration of the merits of local GP prediction and expand its feature set to accommodate a wider range of application. After exploring its performance on the "borehole" data, a classic computer experiment benchmark, we illustrate how noisy data can be accommodated by estimating local nuggets. Section 4 provides a further example of how it can be deployed for computer model calibration.

3.1. Borehole data

The borehole experiment (Worley 1987; Morris, Mitchell, and Ylvisaker 1993) involves an 8-dimensional input space, and our use of it here follows the setup of Kaufman *et al.* (2012); more details can be found therein. The response y is given by

$$y = \frac{2\pi T_u[H_u - H_l]}{\log\left(\frac{r}{r_w}\right)\left[1 + \frac{2LT_u}{\log(r/r_w)r_w^2 K_w} + \frac{T_u}{T_l}\right]}. \tag{7}$$

The eight inputs are constrained to lie in a rectangular domain:

$$\begin{array}{llll} r_w \in [0.05, 0.15] & r \in [100, 5000] & T_u \in [63070, 115600] & T_l \in [63.1, 116] \\ H_u \in [990, 1110] & H_l \in [700, 820] & L \in [1120, 1680] & K_w \in [9855, 12045]. \end{array}$$

We use the following implementation in R which accepts inputs in the unit 8-cube.

```
R> borehole <- function(x){
+   rw <- x[1] * (0.15 - 0.05) + 0.05
+   r <-  x[2] * (50000 - 100) + 100
+   Tu <- x[3] * (115600 - 63070) + 63070
+   Hu <- x[4] * (1110 - 990) + 990
+   Tl <- x[5] * (116 - 63.1) + 63.1
+   Hl <- x[6] * (820 - 700) + 700
+   L <-  x[7] * (1680 - 1120) + 1120
+   Kw <- x[8] * (12045 - 9855) + 9855
+   m1 <- 2 * pi * Tu * (Hu - Hl)
+   m2 <- log(r / rw)
```

```
+   m3 <- 1 + 2 * L * Tu / (m2 * rw^2 * Kw) + Tu / Tl
+   return(m1/m2/m3)
+ }
```

We consider a modestly big training set ($N = 100000$), to illustrate how large emulations can proceed with relatively little computational effort. However, we keep the testing set somewhat smaller so that we can so that we can duplicate part of a Monte Carlo experiment (i.e., multiple repeats of random training and testing sets) from Gramacy and Apley (2015) without requiring too many compute cycles.

```
R> N <- 100000
R> Npred <- 1000
R> dim <- 8
```

The experiment involves ten repetitions wherein a Latin hypercube sample (LHS; McKay, Conover, and Beckman 1979) defines random training data and testing sets, with responses from `borehole`. In each repetition a sequence of (local GP) estimators is fit to the training sets followed by out-of-sample RMSE calculations on the testing sets. Storage for those RMSEs, along with timing info, is allocated as follows

```
R> T <- 10
R> nas <- rep(NA, T)
R> times <- rmse <- data.frame(mspe = nas, mspe2 = nas,
+   alc.nomle = nas, alc = nas, alc2 = nas,
+   nn.nomle = nas, nn=nas, big.nn.nomle = nas, big.nn = nas,
+   big.alcray = nas, big.alcray2 = nas)
```

The names of the columns of the data frame are indicative of the corresponding estimator. For example, `big.nn.nomle` indicates a nearest neighbor (NN) estimator fit to with a larger local neighborhood ($n = 200$) using a sensible, but not likelihood maximizing, global value of θ. The other estimators describe variations either via a smaller local neighborhood ($n = 50$), greedy search, and local calculation of $\hat{\theta}_n(x)$.

The `for` loop below iterates over each Monte Carlo repetition. The first chunk in the loop generates the data via the **lhs** package (Carnell 2012); the second chunk assigns arguments common to all comparators; the remaining lines gather predictions and measure performance.

```
R> for(t in 1:T) {
+
+   if(require("lhs")) { x <- randomLHS(N + Npred, dim) }
+   else { x <- matrix(runif((N + Npred)*dim), ncol=dim) }
+   y <- apply(x, 1, borehole)
+   ypred.0 <- y[-(1:N)]; y <- y[1:N]
+   xpred <- x[-(1:N),]; x <- x[1:N,]
+
+   formals(aGP)[c("omp.threads", "verb")] <- c(nth, 0)
+   formals(aGP)[c("X", "Z", "XX")] <- list(x, y, xpred)
+
```

```
+   out1<- aGP(d=list(mle = FALSE, start = 0.7))
+   rmse$alc.nomle[t] <- sqrt(mean((out1$mean - ypred.0)^2))
+   times$alc.nomle[t] <- out1$time
+
+   out2 <- aGP(d = list(max = 20))
+   rmse$alc[t] <- sqrt(mean((out2$mean - ypred.0)^2))
+   times$alc[t] <- out2$time
+
+   out3 <- aGP(d = list(start = out2$mle$d, max = 20))
+   rmse$alc2[t] <- sqrt(mean((out3$mean - ypred.0)^2))
+   times$alc2[t] <- out3$time
+
+   out4 <- aGP(d = list(max = 20), method="alcray")
+   rmse$alcray[t] <- sqrt(mean((out4$mean - ypred.0)^2))
+   times$alcray[t] <- out4$time
+
+   out5 <- aGP(d = list(start = out4$mle$d, max = 20), method="alcray")
+   rmse$alcray2[t] <- sqrt(mean((out5$mean - ypred.0)^2))
+   times$alcray2[t] <- out5$time
+
+   out6<- aGP(d = list(max = 20), method="mspe")
+   rmse$mspe[t] <- sqrt(mean((out6$mean - ypred.0)^2))
+   times$mspe[t] <- out6$time
+
+   out7 <- aGP(d = list(start = out6$mle$d, max = 20), method="mspe")
+   rmse$mspe2[t] <- sqrt(mean((out7$mean - ypred.0)^2))
+   times$mspe2[t] <- out7$time
+
+   out8 <- aGP(d = list(mle = FALSE, start = 0.7), method = "nn")
+   rmse$nn.nomle[t] <- sqrt(mean((out8$mean - ypred.0)^2))
+   times$nn.nomle[t] <- out8$time
+
+   out9 <- aGP(end = 200, d = list(mle = FALSE), method = "nn")
+   rmse$big.nn.nomle[t] <- sqrt(mean((out9$mean - ypred.0)^2))
+   times$big.nn.nomle[t] <- out9$time
+
+   out10 <- aGP(d = list(max = 20), method = "nn")
+   rmse$nn[t] <- sqrt(mean((out10$mean - ypred.0)^2))
+   times$nn[t] <- out10$time
+
+   out11 <- aGP(end = 200, d = list(max = 20), method="nn")
+   rmse$big.nn[t] <- sqrt(mean((out11$mean - ypred.0)^2))
+   times$big.nn[t] <- out11$time
+
+   out12 <- aGP(end = 200, d = list(max = 20), method="alcray")
+   rmse$big.alcray[t] <- sqrt(mean((out12$mean - ypred.0)^2))
+   times$big.alcray[t] <- out12$time
```

```
+
+   out13 <- aGP(end = 200, d = list(start = out12$mle$d, max = 20),
+     method="alcray")
+   rmse$big.alcray2[t] <- sqrt(mean((out13$mean - ypred.0)^2))
+   times$big.alcray2[t] <- out13 $time
+ }
```

The code below collects summary information into a table, whose rows are ordered by average RMSE value. The final column of the table shows the p-value of a one-sided t-test for differences between adjacent rows in the table—indicating if the RMSE in the row is statistically distinguishable from the one below it.

```
R> timev <- apply(times, 2, mean, na.rm = TRUE)
R> rmsev <- apply(rmse, 2, mean)
R> tab <- cbind(timev, rmsev)
R> o <- order(rmsev, decreasing = FALSE)
R> tt <- rep(NA, length(rmsev))
R> for(i in 1:(length(o)-1)) {
+   tto <- t.test(rmse[ ,o[i]], rmse[ ,o[i+1]], alternative = "less",
+     paired = TRUE)
+   tt[o[i]] <- tto$p.value
+ }
R> tab <- cbind(tab, data.frame(tt))
R> tab[o, ]
```

```
               timev     rmsev           tt
big.alcray2  88.0849 0.1896538 5.463523e-08
big.alcray   98.2095 0.2095144 4.568460e-11
alc2          9.9326 0.2646101 1.443727e-01
mspe2        29.2107 0.2655748 2.144497e-09
big.nn       28.6229 0.2971100 9.275897e-07
alc          10.1819 0.3215997 3.758175e-01
mspe         29.4330 0.3225452 5.136214e-09
alcray2       4.2947 0.3962093 2.153505e-04
alcray        4.4383 0.4218566 5.695679e-15
big.nn.nomle  1.7696 0.8798026 5.857637e-08
alc.nomle     9.8037 1.0079727 6.777112e-08
nn            0.9762 1.1801521 2.842154e-15
nn.nomle      0.3890 3.0325168           NA
```

The two biggest takeaways from the table are that (1) everything is fast on a data set of this size by comparison to the state of the art in GP emulation, approximately or otherwise; (2) local inference of the lengthscale parameter, $\hat{\theta}_n(x)$ leads to substantial improvements in accuracy. Gramacy and Apley's similar experiments included variations on the method of compactly supported covariances (CSC) (Kaufman *et al.* 2012) yielding estimators with similar accuracies, but requiring at least an order magnitude more compute time. In fact, they commented that $N = 10000$ was the limit that CSC could accommodate on their machine due

to memory swapping issues. Moreover, the `laGP` method, despite restrictions to local isotropy, is competitive with, and often outperforms, comparators which model spatial correlations separably. CSC is one example. Gramacy and Haaland (2015) provide a detailed case study along these lines, including hybrid global/local approaches like those described in the following subsection.

The best methods, based on a larger local neighborhood and ray-based search, point to an impressive emulation capability. In a time that is comparable to a plain NN-based emulation strategy (with local inference for $\hat{\theta}_n(x)$; i.e., `nn` in the table), a greedy design is three times more accurate out-of-sample. Gramacy and Haaland (2015) show that the trend continues as N is increased, indicating the potential for extremely accurate emulation on testing and training sets of size $N > 1$M in a few hours. Pairing with cluster-style distribution, across 96 16-CPU nodes, that can be reduced to 188 seconds, or extended to $N > 8M$ in just over an hour. They show that for smaller (yet still large) designs $N < 100000$, searching exhaustively rather than by rays leads to more accurate predictors. In those cases, massive parallelization over a cluster and/or with GPUs (Gramacy *et al.* 2014) can provide (more) accurate predictions on a commensurately sized testing set (N) in about a minute.

3.2. Challenging global/local isotropy

Our choice of isotropic correlation function was primarily one of convenience. It is a common first choice for computer experiments, and since it has just one parameter, θ, inferential schemes like maximum likelihood via Newton methods are vastly simplified. When deployed for local inference over thousands of elements of a vast predictive grid, that simplicity is a near necessity from an engineering perspective. However, the local GP methodology is not limited to this choice. Indeed Gramacy and Apley (2015) developed all of the relevant equations for a generic choice of separable correlation function. Here, separable means the joint correlation over all input directions factors as a product of a simpler one in each direction, independently. The simplest example is a separable Gaussian form, $K_\theta(x, x') = \exp\{-\sum_{k=1}^{p}(x_k - x'_k)^2/\theta_k\}$. It is easy to imagine, as in our eight-dimensional borehole example above, that the spatial model could benefit for allowing differential rate of decay θ_k in each input direction.

The **laGP** package contains limited support for a separable correlation function. Functions like `laGPsep` and `aGPsep` perform the analog of `laGP` and `aGP`, and are currently considered to be *beta* functionality. Release-quality subroutines are provided for separable modeling in the context of *global*, that is canonical, GP inference. On a data set of size $N = 100$K like the one we entertain above, this is not a reasonable undertaking. But we have found it useful on subsets of the data for the purpose of obtaining a rough re-scaling of the inputs so that a (local) isotropic analysis is less objectionable. For example, the code below, after allocating space and setting reasonable starting values and ranges, considers ten random subsets of size $n = 1$K from the full $N = 100$K design, and collects $\hat{\theta}$ vectors under the separable Gaussian formulation.

```
R> thats <- matrix(NA, nrow = T, ncol = dim)
R> its <- rep(NA, T)
R> n <- 1000
R> g2 <- garg(list(mle = TRUE), y)
R> d2 <- darg(list(mle = TRUE, max = 100), x)
R> for(t in 1:T) {
```

```
+
+   subs <- sample(1:N, n, replace = FALSE)
+
+   gpsepi <- newGPsep(x[subs, ], y[subs], rep(d2$start, dim), g = 1/1000,
+     dK = TRUE)
+   that <- mleGPsep(gpsepi, param = "d", tmin = d2$min, tmax = d2$max,
+     ab = d2$ab, maxit = 200)
+   thats[t,] <- that$d
+   its[t] <- that$its
+
+   deleteGPsep(gpsepi)
+ }
```

The `mleGPsep` function uses `optim` with `method="L-BFGS-B"` together with analytic derivatives of the log likelihood; the function `mleGP` offers a similar feature for the isotropic Gaussian correlation, except that it uses a Newton-like method with analytic first and second derivatives. For details on `darg` and `garg`, which lightly regularize and determine initial values for the MLE calculations, see Appendix A.

The package also offers `jmleGPsep`, an analog of `jmleGP`, automating a profile approach to iterating over $\theta|\eta$ and $\eta|\theta$ where the latter is performed with a Newton-like scheme leveraging first and second derivatives.[4] We do not demonstrate `jmleGPsep` on this example since the large data subset ($n = 1000$) combined with very smooth deterministic outputs from moderately size (8-dim) inputs, via `borehole`, leads to estimating near-zero nuggets and ill-conditioning in the matrix decompositions owing to our choice of Gaussian decay. For estimating nuggets in this setup, where the response is both deterministic and extremely smooth (and stationary), we recommend **GPfit** (MacDoanld *et al.* 2014) based on the methods of Ranjan, Haynes, and Karsten (2011). However, we caution that in our experience **GPfit** can be slow on data sets as large as $N = 1000$.

Figure 9 shows the distribution of estimated lengthscales obtained by randomizing over subsets of size $n = 1000$. We see that some lengthscales are orders of magnitude smaller than others, suggesting that some inputs may be more important than others. Input one (r_w) has a distribution that is highly concentrated near small values, so it may be the most important. Perhaps treating all inputs equally when performing a global/local approximation, as in Section 3.1, is leaving some predictability on the table. The **laGP** package does not support using a separable correlation function for local analysis, however we can pre-scale the data globally to explore whether there is any benefit from a differential treatment of inputs.

```
R> scales <- sqrt(apply(thats, 2, median))
R> xs <- x; xpreds <- xpred
R> for(j in 1:ncol(xs)) {
+   xs[,j] <- xs[,j] / scales[j]
```

[4]Newer versions of the package also provide a `param = "both"` option to `mleGPsep`, leveraging a gradient over both separable lengthscale and nugget parameters for joint inference; i.e., not taking a profile approach. This setup usually requires fewer total iterations to converge unless one of the two parameters sets is already very close to the MLE setting. See the package documentation for more details. Unfortunately an analog in `mleGP` is not available at this time.

```
R> boxplot(thats, main = "distribution of thetas", xlab = "input",
+   ylab = "theta")
```

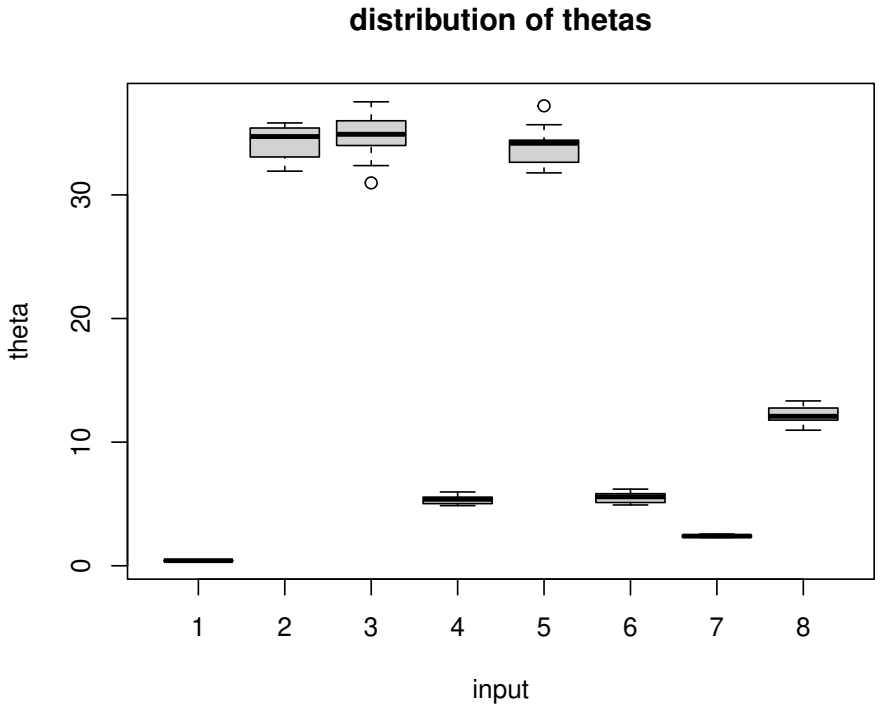

Figure 9: Distribution of maximum *a' posteriori* lengthscales over random subsets of the borehole data.

```
+   xpreds[,j] <- xpreds[,j] / scales[j]
+ }
```

Using the new inputs, consider the following global approximation for the final iteration in the Monte Carlo experiment from Section 3.1.

```
R> out14 <- aGP(xs, y, xpreds, d=list(start=1, max=20), method="alcray")
```

Since the imputs have been pre-scaled by an estimate of (square-root) lengthscale(s), it makes sense to initialize with a local lengthscale of one. The RMSE obtained,

```
R> sqrt(mean((out14$mean - ypred.0)^2))

[1] 0.1619715
```

is competitive with the best methods in the study above—those are based on $n = 200$ whereas only the default $n = 50$ was used here. Also observe that the RMSE we just obtained is better than half of the one we reported for "`alcray`" in the Monte Carlo experiment.

Determining if this reduction is statistically significant would require incorporating it into the Monte Carlo. We encourage the reader to test that off-line, if so inclined. We conclude here that it can be beneficial to perform a cursory global analysis with a separable correlation function to determine if the inputs should be scaled before performing a local (isotropic) analysis on the full data set.

3.3. Motorcycle data

For a simple illustration of heteroskedastic local GP modeling, consider the motorcycle accident data (Silverman 1985), simulating the acceleration of the head of a motorcycle rider as a function of time in the first moments after an impact. It can be found in the **MASS** package (Venables and Ripley 2002) for R. For comparison, we first fit a simple GP model to the full data set ($N = 133$), estimating both lengthscale θ and nugget η.

```
R> if(require("MASS")) {
+ d <- darg(NULL, mcycle[, 1, drop = FALSE])
+ g <- garg(list(mle = TRUE), mcycle[,2])
+ motogp <- newGP(mcycle[ , 1, drop=FALSE], mcycle[ ,2], d = d$start,
+   g = g$start, dK = TRUE)
+ jmleGP(motogp, drange = c(d$min, d$max), grange = c(d$min, d$max),
+   dab = d$ab, gab = g$ab)
+ } else { cat("this example is not doable without MASS") }

         d         g tot.its dits gits
1 54.28291 0.2771448      95   26   69
```

Now consider the predictive equations derived from that full-data, alongisde a local approximate alternative (via ALC) with a local neighborhood size of $n = 30$.

```
R> if(require("MASS")) {
+ XX <- matrix(seq(min(mcycle[ ,1]), max(mcycle[ ,1]), length = 100),
+   ncol = 1)
+ motogp.p <- predGP(motogp, XX = XX, lite = TRUE)
+ motoagp <- aGP(mcycle[ , 1, drop=FALSE], mcycle[,2], XX, end = 30,
+   d = d, g = g, verb = 0)
+ } else { cat("this example is not doable without MASS") }
```

Figure 10 shows the predictive surfaces obtained for the two predictors in terms of means and 90% credible intervals. The (full) GP mean surface, shown as solid-black, is smooth and tracks the center of the data nicely from left to right over the range of x-values. However, it is poor at capturing the heteroskedastic nature of the noise (dashed-black). The local GP mean is similar, except near $x = 35$ where it is not smooth. This is due to the small design. With only $N = 132$ there isn't much opportunity for smooth transition as the local predictor tracks across the input space, leaving little wiggle room to make a trade-off between smoothness ($n = 132$, reproducing the full GP results exactly) and adaptivity ($n \ll 132$). Although the mean of the local GP may disappoint, the variance offers an improvement over the full GP. It is conservative where the response is wiggly, being similar to the full GP but slightly wider, and narrower where the response is flat.

```
R> if(require("MASS")) {
+ plot(mcycle, cex = 0.5, main = "motorcycle data")
+ lines(XX, motogp.p$mean, lwd = 2)
+ q1 <- qnorm(0.05, mean = motogp.p$mean, sd = sqrt(motogp.p$s2))
+ q2 <- qnorm(0.95, mean = motogp.p$mean, sd = sqrt(motogp.p$s2))
+ lines(XX, q1, lty = 2, lwd = 2)
+ lines(XX, q2, lty = 2, lwd = 2)
+ lines(XX, motoagp$mean, col = 2, lwd = 2)
+ q1 <- qnorm(0.05, mean = motoagp$mean, sd = sqrt(motoagp$var))
+ q2 <- qnorm(0.95, mean = motoagp$mean, sd = sqrt(motoagp$var))
+ lines(XX, q1, lty = 2, col = 2, lwd = 2)
+ lines(XX, q2, lty = 2, col = 2, lwd = 2)
+ } else { plot(0, 0, main = "install MASS package!") }
```

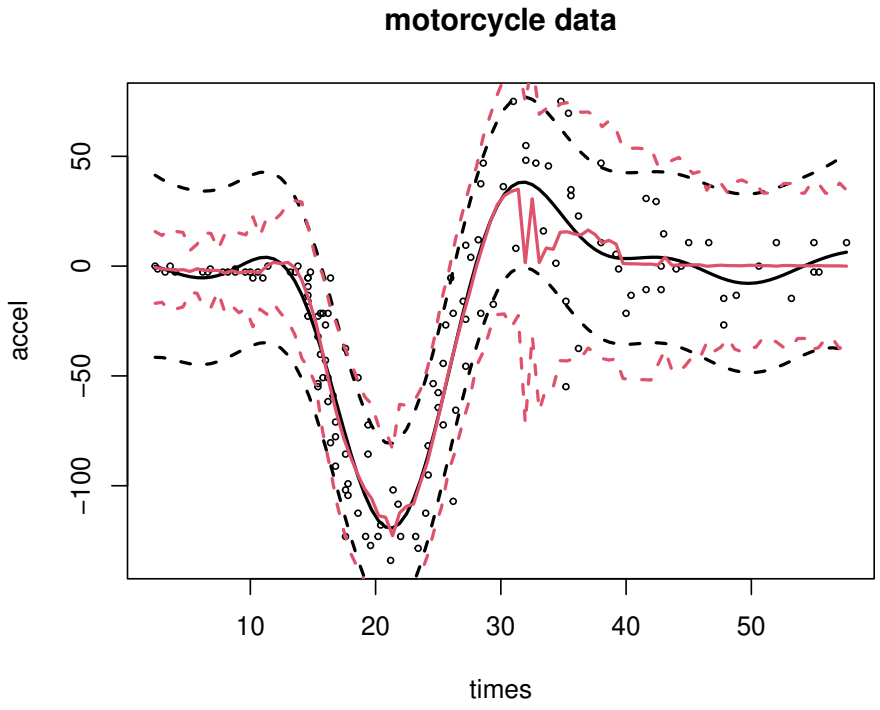

Figure 10: Comparison of a global GP predictive surface (black) with a local one (red). Predictive means (solid) and 90% interval (dashed) shown.

It is interesting to explore how the local GP approximation would fare on a larger version of the same problem, where otherwise a local approach is not only essential for computational reasons, but also potentially more appropriate from a nonstationary modeling perspective on this data. For a crude simulation of a larger data setup we replicated the data ten times with a little bit of noise on the inputs.

```
R> if(require("MASS")) {
```

```
+ X <- matrix(rep(mcycle[ ,1], 10), ncol = 1)
+ X <- X + rnorm(nrow(X), sd = 1)
+ Z <- rep(mcycle[ ,2], 10)
+ motoagp2 <- aGP(X, Z, XX, end = 30, d = d, g = g, verb = 0)
+ } else { cat("this example is not doable without MASS") }
```

Figure 11 shows the resulting predictive surface. Notice how it does a much better job of tracing predictive uncertainty across the input space. The predictive mean is still overly

```
R> if(require("MASS")) {
+ plot(X, Z, main = "simulating a larger data setup", xlab = "times",
+   ylab = "accel")
+ lines(XX, motoagp2$mean, col = 2, lwd = 2)
+ q1 <- qnorm(0.05, mean = motoagp2$mean, sd = sqrt(motoagp2$var))
+ q2 <- qnorm(0.95, mean = motoagp2$mean, sd = sqrt(motoagp2$var))
+ lines(XX, q1, col = 2, lty = 2, lwd = 2)
+ lines(XX, q2, col = 2, lty = 2, lwd = 2)
+ } else { plot(0, 0, main = "install MASS!") }
```

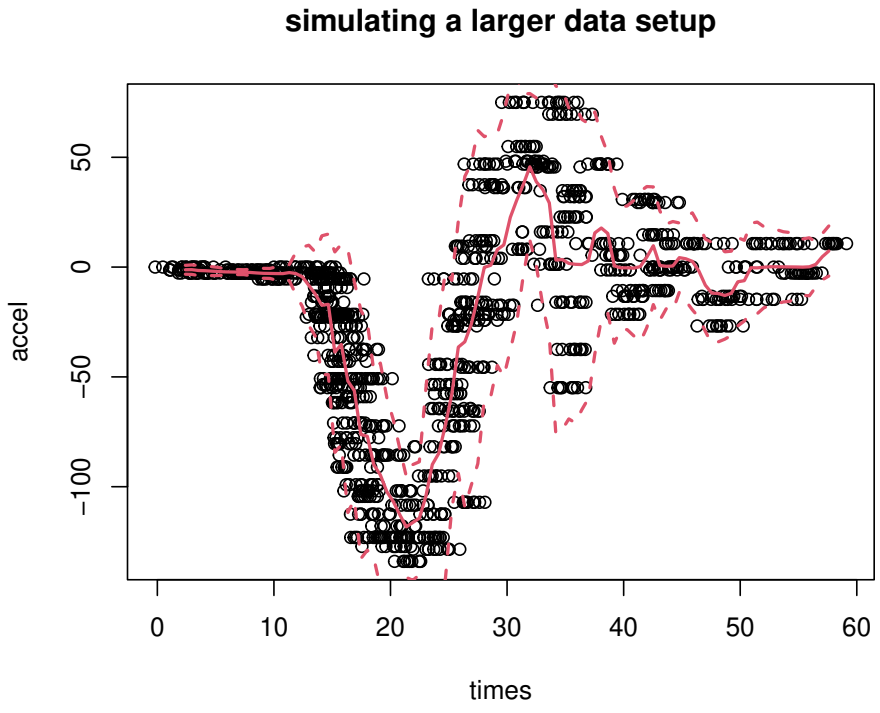

Figure 11: Predictive surface obtained after combining ten replications of the data with jittered x-values.

wiggly, but also reveals structure in the data that may not have been evident from the scatterplot alone, and likewise is disguised (or overly smoothed) by the full GP fit. The local GP is picking up oscillations for larger input values which makes sense considering the output is

measuring a whiplash effect. However, that may simply be wishful thinking; the replicated response values paired with the jittered predictors may not be representative of what would have been observed in a larger simulation.

4. Calibration

Computer model *calibration* is the enterprise of matching a simulation engine with real, or field, data to ultimately build an accurate predictor for the real process at novel inputs. In the case of large computer simulations, calibration represents a capstone application uniquely blending (and allowing review of) features, for both large and small-scale spatial modeling via GPs, provided by the **laGP** package.

Kennedy and O'Hagan (2001) described a statistical framework for combining potentially biased simulation output and noisy field observations for model calibration, via a hierarchical model. They proposed a Bayesian inferential framework for jointly estimating, using data from both processes, the bias, noise level, and any parameters required to run the computer simulation—so-called *calibration parameter(s)*—but which cannot be controlled or observed in the field. The setup, which we review below, has many attractive features, however it scales poorly when simulations get large. We explain how Gramacy, Bingham, Holloway, Grosskopf, Kuranz, Rutter, Trantham, and Drake (2015) modified that setup using **laGP** and provide a live demonstration via an example extracted from that paper.

4.1. A hierarchical model for Bayesian inference

Consider data comprised of runs of a computer model M at a large space-filling design, and a much smaller number observations from a physical or field experiment F following a design that respects limitations of the experimental apparatus. It is typical to assume that the runs of M are deterministic, and that its input space fully contains that of F. Use x to denote *design variables* that can be adjusted, or at leased measured, in the physical system; and let u to denote *calibration* or *tuning parameters*, whose values are required to simulate the system, but are unknown in the field. The primary goal is to predict the result of new field data experiments, via M, which in turn means finding a good u.

Toward that goal, Kennedy and O'Hagan (2001, hereafter KOH) proposed the following coupling of M and F. Let $y^F(x)$ denote a field observation at x, and $y^M(x,u)$ denote the (deterministic) output of a computer model run. KOH represent the *real* mean process R as the computer model output at the best setting of the tuning parameters, u^*, plus a bias term acknowledging potential for systematic discrepancies between the computer model and the underlying mean of the physical process. In symbols, the mean of the physical process is $y^R(x) = y^M(x,u^*) + b(x)$. The field observations connect reality with data:

$$y^F(x) = y^R(x) + \varepsilon = y^M(x,u^*) + b(x) + \varepsilon, \quad \text{where} \quad \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_\varepsilon^2). \tag{8}$$

The unknown parameters are u^*, σ_ε^2, and the discrepancy or bias $b(\cdot)$.

If evaluating the computer model is fast, then inference can proceed via residuals $y^F(x) - y^M(x,u)$, which can be computed at will for any (x,u) (Higdon, Kennedy, Cavendish, Cafeo, and Ryne 2004). However, y^M simulations are usually time consuming, in which case it helps to build an emulator $\hat{y}^M(\cdot,\cdot)$ fit to code outputs obtained on a computer experiment design

of N_M locations (x, u). KOH recommend a GP prior for y^M, however rather than learn $\hat{y}^M$ in isolation, using just the N_M runs, as we have been doing throughout this document, they recommend inference joint with $b(\cdot)$, u, and σ^2_ε using both field observations and runs of the computer model. From a Bayesian perspective this is the coherent thing to do: infer all unknowns jointly given all data.

This is a practical approach when the computer model is *very* slow, giving small N_M. In that setup, the field data can be informative for emulation of $y^M(\cdot, \cdot)$, especially when the bias $b(\cdot)$ is very small or easy to estimate. Generally however, the computation required for inference in this setup is fraught with challenges, especially in the fully Bayesian formulation recommended by KOH. The coupled $b(\cdot)$ and $y^M(\cdot, \cdot)$ lead to parameter identification and MCMC mixing issues. And GP regression, taking a substantial computational toll when deployed in isolation, faces a compounded burden when coupled with other processes.

4.2. Calibration as optimization

Gramacy *et al.* (2015) proposed a thriftier approach pairing local approximate GP models for emulation with a modularized calibration framework (Liu, Bayarri, and Berger 2009) and derivative free optimization (Conn, Scheinberg, and Vicente. 2009). *Modularized* calibration sounds fancy, but it really represents a reduction rather than expansion of ideas: fitting the emulator $\hat{y}^M(\cdot, \cdot)$ separately or independently from the bias, using only the outputs of runs at a design of N_M inputs (x, u). Liu *et al.*'s justification for modularization stemmed from a "contamination" concern echoed by other researchers (e.g., Joseph 2006; Santner *et al.* 2003) where, in the fully Bayesian scheme, joint inference allows "pristine" field observations to be contaminated by an imperfect computer model.

Gramacy *et al.* motivate modularization from a more practical perspective, that of decoupling inference for computational tractability in large N_M settings. They argue that there is little harm in doing so for most modern calibration applications, in terms of the quality of estimates obtained irrespective of computational considerations. Due to the relative costs, the number of computer model runs involved increasingly dwarfs the data available from the field, i.e., $N_M \gg N_F$, making it unlikely that field data would substantively enhance the quality of the emulator, leaving only risk that joint inference with the bias will obfuscate traditional computer model diagnostics, and possibly stunt their subsequent re-development or refinement.

Combining modularization with local approximate GPs for emulation, and full GP regressions (with nugget η) for estimating bias-plus-noise from a relatively small number of field data observations, N_F, Gramacy *et al.* recommend viewing calibration as an optimization, acting as the glue that "sticks it all together". Algorithm 1 provides pseudocode comprised of library functions describing the objective function. In **laGP**, this objective is implemented as `fcalib`, comprising of first (steps 1–5) a call to `aGP.seq` to emulate on a schedule of sequential stages of local refinements [Figure 7]; and then (6–8) a call to `discrep.est` which estimates the GP discrepancy or bias term. The notation used in the psuedo-code, and further explanation, is provided below.

Let the field data be denoted as $D^F_{N_F} = (X^F_{N_F}, Y^F_{N_F})$ where $X^F_{N_F}$ is the design matrix of N_F field data inputs, paired with an N_F vector of y^F observations $Y^F_{N_F}$. Similarly, let $D^M_{N_M} = ([X^M_{N_M}, U_{N_M}], Y^M_{N_M})$ be the N_M computer model input-output combinations with column-combined x- and u-design(s) and y^M-outputs. Then, with an emulator $\hat{y}^M(\cdot, u)$ trained on

Require: Calibration parameter u, fidelity parameter n_M, computer data $D^M_{N_M}$, and field data $D^F_{N_F}$.

1: **for** $j = 1, \ldots, N_F$ **do**
2: $I \leftarrow \texttt{laGP}(x^F_j, u \mid n_M, D^M_{N_M})$ {get indicies of local design}
3: $\hat{\theta}_j \leftarrow \texttt{mleGP}(D^M_{N_M}[I])$ {local MLE of correlation parameter(s)}
4: $\hat{y}^{M|u}_j \leftarrow \texttt{muGP}(x^F_j \mid D^M_{N_M}[I], \hat{\theta}_j)$ {predictive mean emulation following Eq. (3)}
5: **end for**
6: $\hat{Y}^{B|u}_{N_F} \leftarrow Y^F_{N_F} - \hat{Y}^{M|u}$ {vectorized bias calculation}
7: $D^B_{N_F}(u) \leftarrow (\hat{Y}^{B|u}_{N_F}, X^F_{N_F})$ {create data for estimating $\hat{b}(\cdot)|u$}
8: $\hat{\theta}_b \leftarrow \texttt{mleGP}(D^B_{N_F}(u))$ {full GP estimate of $\hat{b}(\cdot)|u$}
9: **return** $\texttt{llikGP}(\hat{\theta}_n, D^B_{N_F}(u))$ {the objective value of the `mleGP` call above}

Algorithm 1: Objective function evaluation for modularized local GP calibration.

$D^M_{N_M}$, let $\hat{Y}^{M|u}_{N_F} = \hat{y}^M(X^F_{N_F}, u)$ denote a vector of N_F emulated output y-values at the X_F locations obtained under a setting, u, of the calibration parameter. With local approximate GP modeling, each $\hat{y}^{M|u}_j$-value therein, for $j = 1, \ldots, N_F$, can be obtained independently (and in parallel) with the others via local sub-design $X_{n_M}(x^F_j, u) \subset [X^M_{N_M}, U_{N_M}]$ and local inference for the correlation structure. A key advantage of this approach, which makes **laGP** methods well-suited to the task, is that emulation is performed only where it is needed, at a small number N_F of locations $X^F_{N_F}$, regardless of the size N_M of the computer model data. The size of the local sub-design, n_M, is a fidelity parameter, meaning that larger values provide more accurate emulation at greater computational expense. Finally, denote the N_F-vector of fitted discrepancies as $\hat{Y}^{B|u}_{N_F} = Y^F_{N_F} - \hat{Y}^{M|u}_{N_F}$. Given these quantities, the objective function for calibration of u, coded in Algorithm 1, is the (log) joint probability density of observing $Y^F_{N_F}$ at inputs $X^F_{N_F}$. Since N_F is small, this can be obtained from a best-fitting GP regression model trained on data $D^B_{N_F}(u) = (\hat{Y}^{B|u}_{N_F}, X^F_{N_F})$, representing the bias estimate $\hat{b}(\cdot)$.

Objective function in hand, we turn to optimizing. The discrete nature of independent local design searches for $\hat{y}^M(x^F_j, u)$ ensures that the objective is not continuous in u. It can look 'noisy', although it is in fact deterministic. This means that optimization with derivatives—even numerically approximated ones—is fraught with challenges. Gramacy *et al.* suggest a derivative-free approach via the mesh adaptive direct search (MADS) algorithm (Audet and Dennis, Jr. 2006) implemented as **NOMAD** (Le Digabel 2011). The authors of the **crs** package (Racine and Nie 2012) provide `snomadr`, an R wrapper to the underlying C++. MADS/**NOMAD** proceeds by successive pairs of *search* and *poll* steps, trying inputs to the objective function on a sequence of meshes that are refined in such a way as to guarantee convergence to a local optima under very weak regularity conditions; for more details see Audet and Dennis, Jr. (2006).

As MADS is a local solver, **NOMAD** requires initialization. Gramacy *et al.* recommend choosing starting u-values from the best value(s) of the objective found on a small random space-filling design. We note here that although **laGP** provides functions like `fcalib`, `aGP.seq` and `discrep.est` to facilitate calibration via optimization, there is no single subroutine automating the combination of all elements: selection of initial search point, executing search, and finally utilizing the solution to make novel predictions in the field. The illustrative exam-

ple below in Section 4.3 is intended to double as a skeleton for novel application. It involves a `snomadr` call with objective `fcalib`, after pre-processing to find an initial u-value via simple iterative search over `fcalib` calls. Then, after optimization returns an optimal u^* value, the example demonstrates how estimates of $\hat{b}(x)$ and $\hat{y}^M(x, u^*)$ can be obtained by retracing steps in Algorithm 1 to extract a local design and correlation parameter (via `aGP.seq`), parallelized for many x. Finally, using saved $D^B_{N_F}(u)$ and $\hat{\theta}$ from the optimization, or quickly re-computing them via `discrep.est`, it builds a predictor for the field at new x locations. Emulations and biases are thus combined to form a distribution for $y^F(x)|u^*$, a sum of Student-t's for $\hat{y}^M(x, u)$ and $\hat{b}(x)$ comprising $y^F(x)|u^*$. However, if $N_F, n_M \geq 30$ summing normals suffices.

4.3. An illustrative example

Consider the following computer model test function used by Goh, Bingham, Holloway, Grosskopf, Kuranz, and Rutter (2013), which is an elaboration of one first described by Bastos and O'Hagan (2009).

```
R> M <- function(x,u)
+   {
+     x <- as.matrix(x)
+     u <- as.matrix(u)
+     out <- (1 - exp(-1 / (2 * x[,2])))
+     out <- out * (1000 * u[,1] * x[,1]^3 + 1900 * x[ ,1]^2
+       + 2092 * x[ ,1] + 60)
+     out <- out / (100 * u[,2] * x[,1]^3 + 500 * x[ ,1]^2 + 4 * x[ ,1] + 20)
+     return(out)
+   }
```

Goh *et al.* paired this with the following discrepancy function to simulate real data under a process like in Equation 8.

```
R> bias <- function(x)
+   {
+     x <- as.matrix(x)
+     out <- 2 * (10 * x[ ,1]^2 + 4 * x[ ,2]^2) / (50 * x[ ,1] * x[ ,2] + 10)
+     return(out)
+   }
```

Data coming from the "real" process is simulated under a true (but unknown) u-value, and then augmented with bias and noise.

```
R> library("tgp")
R> rect <- matrix(rep(0:1, 4), ncol = 2, byrow = 2)
R> ny <- 50
R> X <- lhs(ny, rect[1:2,] )
R> u <- c(0.2, 0.1)
R> Zu <- M(X, matrix(u, nrow = 1))
R> sd <- 0.5
R> reps <- 2
```

```
R> Y <- rep(Zu, reps) + rep(bias(X), reps) +
+   rnorm(reps * length(Zu), sd = sd)
```

The code uses `Y` denote field data observations $Y^F_{N_F}$ with N_F = `2*ny` = 100, storing two replicates at each $X^F_{N_F}$ = `X` location. Gramacy *et al.* (2015) illustrated this example with ten replicates. We keep it smaller here for faster execution in live demonstration. Observe that the code uses uses `lhs` from the **tgp** library (Gramacy 2007; Gramacy and Taddy 2010), rather than from **lhs**, because the **tgp** version allows a non-unit rectangle, which is required for our second use of `lhs` below.

The computer model runs are generated as follows

```
R> nz <- 10000
R> XU <- lhs(nz, rect)
R> XU2 <- matrix(NA, nrow=10 * ny, ncol = 4)
R> for(i in 1:10) {
+   I <- ((i - 1) * ny + 1):(ny * i)
+   XU2[I, 1:2] <- X
+ }
R> XU2[ ,3:4] <- lhs(10 * ny, rect[3:4, ])
R> XU <- rbind(XU, XU2)
R> Z <- M(XU[ ,1:2], XU[ ,3:4])
```

Observe that the design $X^M_{N_M}$ = `XU` is a large LHS in four dimensions, i.e., over design and calibration parameters jointly, augmented with ten-fold replicated field design inputs paired with LHS u-values. This recognizes that it is sensible to run the computer model at inputs where field runs have been observed. `Z` is used to denote $Y^M_{N_M}$.

The following block sets default priors, initial values and specifies details of the model(s) to be estimated. For more details on `darg` and `garg`, see Appendix A.

```
R> bias.est <- TRUE
R> methods <- rep("alc", 2)
R> da <- d <- darg(NULL, XU)
R> g <- garg(list(mle = TRUE), Y)
```

Changing `bias.est = FALSE` will cause estimation of bias $\hat{b}(\cdot)$ to be skipped, and instead only the level of noise between computer model and field data is estimated. The `methods` vector specifies the nature of search and number of passes through the data for local design and inference. Finally `da`, `d` and `g` contain default priors for the lengthscale of the computer model emulator, and the bias parameters respectively. The prior is completed with a (log) prior density on the calibration parameter, u, which we choose to be an independent Beta with a mode in the middle of the space.

```
R> beta.prior <- function(u, a = 2, b = 2, log = TRUE)
+ {
+   if(length(a) == 1) a <- rep(a, length(u))
+   else if(length(a) != length(u)) stop("length(a) must be 1 or length(u)")
+   if(length(b) == 1) b <- rep(b, length(u))
```

```
+   else if(length(b) != length(u)) stop("length(b) must be 1 or length(u)")
+   if(log) return(sum(dbeta(u, a, b, log=TRUE)))
+   else return(prod(dbeta(u, a, b, log=FALSE)))
+ }
```

Now we are ready to evaluate the objective function on a "grid" to search for a starting value for **NOMAD**. The "grid" is comprised of a space-filling design on a slightly smaller domain than the input space allows. Experience suggests that initializing too close to the boundary of the input space leads to poor performance in **NOMAD** searches.

```
R> initsize <- 10*ncol(X)
R> imesh <- 0.1
R> irect <- rect[1:2,]
R> irect[,1] <- irect[,1] + imesh/2
R> irect[,2] <- irect[,2] - imesh/2
R> uinit.cand <- lhs(10 * initsize, irect)
R> uinit <- dopt.gp(initsize, Xcand = lhs(10 * initsize, irect))$XX
R> llinit <- rep(NA, nrow(uinit))
R> for(i in 1:nrow(uinit)) {
+   llinit[i] <- fcalib(uinit[i,], XU, Z, X, Y, da, d, g, beta.prior,
+                       methods, M, bias.est, nth, verb = 0)
+ }
```

By default, `fcalib` echoes the input and calculated objective value (log likelihood or posterior probability) to the screen. This can be useful for tracking progress for an optimization, say via **NOMAD**, however we suppress this here to eliminate clutter. The `fcalib` function has an argument called `save.global` that (when not `FALSE`) causes the information that would otherwise be printed to the screen to be saved in a global variable called `fcalib.save` in the environment indicated (e.g., `save.global = .GlobalEnv`). Those prints can be handy for inspection once the optimization has completed. That flag isn't engaged above, since the required quantities, `uinit` and `llinit` respectively, are already in hand. We will, however, utilize this feature below as `snomadr` does not provide an alternative mechanism for saving progress information for later inspection.

The next code chunk loads the **crs** library containing `snomadr`, the R interface to **NOMAD**, and then creates a list of options that are passed to **NOMAD** via `snomadr`.

```
R> if(require("crs")) {
+ opts <- list("MAX_BB_EVAL" = 1000, "INITIAL_MESH_SIZE" = imesh,
+   "MIN_POLL_SIZE" = "r0.001", "DISPLAY_DEGREE" = 0)
+ } else { cat("this example is not doable without crs") }
```

We have found that these options work well when the input space is scaled to the unit cube. They are derived from defaults recommended in the **NOMAD** documentation.

Now we are ready to invoke `snomadr` on the best input(s) found on grid established above. The code below orders those inputs by their objective value, and then loops over them until a minimum number of **NOMAD** iterations has been reached. Usually, this threshold results in just one pass through the `while` loop, however it offers some robustness in the face of occasional pre-mature convergence. In practice it may be sensible to perform a more exhaustive search if computational resources are abundant.

```
R> its <- 0
R> o <- order(llinit)
R> i <- 1
R> out <- NULL
R> while(its < 10) {
+ if(require("crs")) {
+   outi <- snomadr(fcalib, 2, c(0,0), 0, x0 = uinit[o[i],],
+             lb = c(0,0), ub = c(1,1), opts = opts, XU = XU,
+             Z = Z, X = X, Y = Y, da = da, d = d, g = g,
+             methods = methods, M = M, bias = bias.est,
+             omp.threads = nth, uprior = beta.prior,
+             save.global = .GlobalEnv, verb = 0)
+   its <- its + outi$iterations
+ } else {
+   outi <- fcalib(uinit[o[i],], XU = XU,
+             Z = Z, X = X, Y = Y, da = da, d = d, g = g,
+             methods = methods, M = M, bias = bias.est,
+             omp.threads = nth, uprior = beta.prior,
+             save.global = .GlobalEnv, verb = 0)
+   outi <- list(objective=outi, solution=uinit[o[i],])
+   its <- its + 1
+ }
+   if(is.null(out) || outi$objective < out$objective) out <- outi
+   i <- i + 1;
+ }

iterations: 18
time:       102
```

From the two major chunks of code above, we collect evaluations of `fcalib`, combining a space-filling set of `u`-values and ones placed along stencils in search of the `u`-value maximizing the likelihood (or posterior probability). In this 2-d problem, that's enough to get good resolution on the log likelihood/posterior surface in `u`. The code below discards any input pairs that are not finite. Infinite values result when **NOMAD** tries input settings that lie exactly on the bounding box.

```
R> Xp <- rbind(uinit, as.matrix(fcalib.save[ ,1:2]))
R> Zp <- c(-llinit, fcalib.save[ ,3])
R> wi <- which(!is.finite(Zp))
R> if(length(wi) > 0) { Xp <- Xp[-wi, ]; Zp <- Zp[-wi]}
R> if(require("interp")) {
+ surf <- interp(Xp[ ,1], Xp[ ,2], Zp, duplicate = "mean")
+ } else { cat("visual not available without interp") }
```

Figure 12 shows an image plot of the surface, projected to a mesh via `interp` in the **akima** package (Akima, Gebhardt, Petzoldt, and Maechler 2013),[5] with lighter-colored values indicating a larger value of likelihood/posterior probability. The initialization points (open

[5]Recent versions have used the **interp** package instead.

```
R> if(require("interp")) {
+ image(surf, xlab = "u1", ylab = "u2", main = "posterior surface",
+   col = heat.colors(128), xlim = c(0,1), ylim = c(0,1))
+ } else {
+   plot(0, 0, xlim = c(0,1), ylim = c(0,1), type="n",
+     main = "install interp!")
+ }
R> points(uinit)
R> points(fcalib.save[,1:2], col = 3, pch = 18)
R> u.hat <- out$solution
R> points(u.hat[1], u.hat[2], col = 4, pch = 18)
R> abline(v = u[1], lty = 2)
R> abline(h = u[2], lty = 2)
```

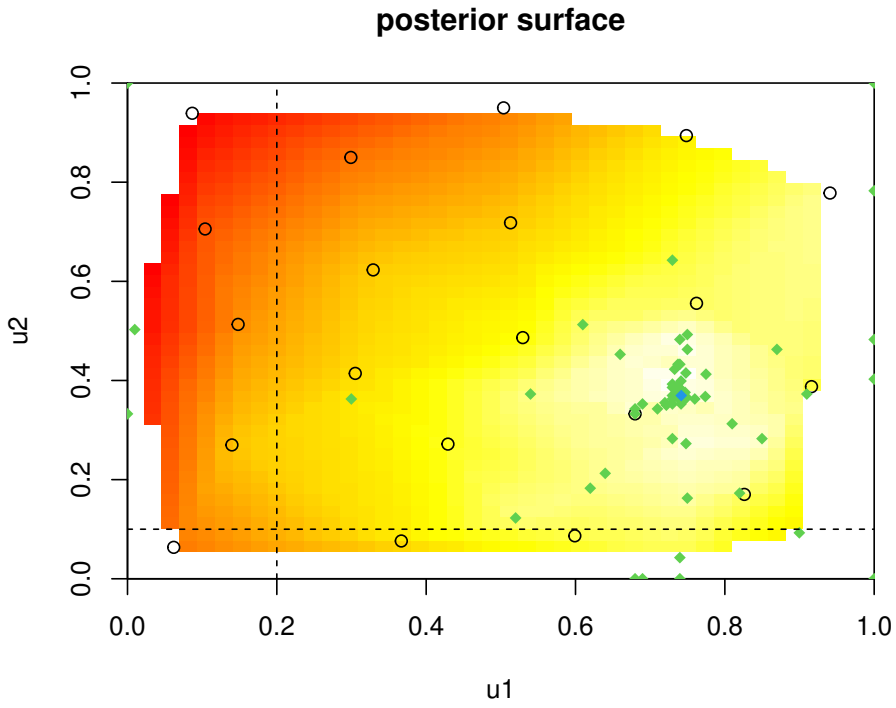

Figure 12: A view of the log likelihood/posterior surface as a function of the calibration inputs, with the optimal `u.hat` value (green dot), the initial grid (open circles) and points of evaluation along the **NOMAD** search (black dots), and the true `u`-value (cross-hairs) shown.

circles), evaluations along the **NOMAD** search (black dots), and the ultimate value found in optimization (green dot) are also shown.

Observe, by comparing to the true `u`-value (cross-hairs), that the `u.hat` value we found is far from the value that generated the data. In fact, while the surface is fairly peaked around the best `u.hat`-value that we found, it gives very little support to the true value. Since there are

were far fewer evaluations made near the true value, it is worth checking if the solver missed an area of high likelihood/probability.

```
R> Xu <- cbind(X, matrix(rep(u, ny), ncol = 2, byrow = TRUE))
R> Mhat.u <- aGP.seq(XU, Z, Xu, da, methods, ncalib = 2, omp.threads = nth,
+   verb = 0)
R> cmle.u <- discrep.est(X, Y, Mhat.u$mean, d, g, bias.est, FALSE)
R> cmle.u$ll <- cmle.u$ll + beta.prior(u)
```

Comparing log likelihood/posterior probabilities yields:

```
R> data.frame(u.hat = -out$objective, u = cmle.u$ll)

     u.hat         u
1 -128.479 -132.7951
```

Well that's reassuring in some ways—the optimization part is performing well—but not in others. Perhaps modeling apparatus introduces some identification issues that prevent recovering the data-generating `u`-value by maximizing likelihood/posterior probability.

Before searching for an explanation, lets check predictive accuracy in the field on a holdout set, again pitting the true `u`-value against our `u.hat`. We first create a random testing design and set aside the true predicted values on those inputs for later comparison.

```
R> nny <- 1000
R> XX <- lhs(nny, rect[1:2,])
R> ZZu <- M(XX, matrix(u, nrow = 1))
R> YYtrue <- ZZu + bias(XX)
```

Now we can calculate an out-of-sample RMSE value, first based on the true `u`-value.

```
R> XXu <- cbind(XX, matrix(rep(u, nny), ncol = 2, byrow = TRUE))
R> Mhat.oos.u <- aGP.seq(XU, Z, XXu, da, methods, ncalib = 2,
+   omp.threads = nth, verb = 0)
R> YYm.pred.u <- predGP(cmle.u$gp, XX)
R> YY.pred.u <- YYm.pred.u$mean + Mhat.oos.u$mean
R> rmse.u <- sqrt(mean((YY.pred.u - YYtrue)^2))
R> deleteGP(cmle.u$gp)
```

Turning to an RMSE calculation using the estimated `u.hat` value, we must re-build some key objects under that value as those objects are not returned to us via either `fcalib` or `snomadr`.

```
R> Xu <- cbind(X, matrix(rep(u.hat, ny), ncol = 2, byrow = TRUE))
R> Mhat <- aGP.seq(XU, Z, Xu, da, methods, ncalib = 2, omp.threads = nth,
+   verb = 0)
R> cmle <- discrep.est(X, Y, Mhat$mean, d, g, bias.est, FALSE)
R> cmle$ll <- cmle$ll + beta.prior(u.hat)
```

As a sanity check, it is nice to see that the value of the log likelihood/posterior probability matches with the one we obtained from `snomadr`:

```
R> print(c(cmle$ll, -out$objective))

[1] -128.479 -128.479
```

Now we can repeat what we did with the true `u`-value with our estimated one `u.hat`.

```
R> XXu <- cbind(XX, matrix(rep(u.hat, nny), ncol = 2, byrow = TRUE))
R> Mhat.oos <- aGP.seq(XU, Z, XXu, da, methods, ncalib = 2,
+   omp.threads = nth, verb = 0)
R> YYm.pred <- predGP(cmle$gp, XX)
R> YY.pred <- YYm.pred$mean + Mhat.oos$mean
R> rmse <- sqrt(mean((YY.pred - YYtrue)^2))
```

Wrapping up the comparison, we obtain the following:

```
R> data.frame(u.hat = rmse, u = rmse.u)

      u.hat         u
1 0.1849574 0.2250721
```

Indeed, our estimated `u.hat`-value leads to better predictions of the field data out-of-sample. Gramacy *et al.* (2015) offer an explanation. The KOH model is, with GPs for emulation and bias, overly flexible and consequently challenges identification of the unknown parameters. Authors have commented on this before, including KOH to a limited extent. Interlocking GP predictors (Ba and Joseph 2012) and the introduction of auxiliary inputs (Bornn, Shaddick, and Zidek 2012), of which the u-values are an example, have recently been proposed as deliberate mechanisms for handling nonstationary features in response surface models, particularly for computer experiments. The KOH framework combines both, and predates those works by more than a decade, so in some sense the model being fit is leveraging tools designed for flexibility in response surface modeling, possibly at the expense of being faithful to the underlying meanings of parameters like u and bias processes $b(\cdot)$. In any event, we draw comfort from evidence that the method yields accurate predictions, which in most calibration applications is the primary aim.

5. Ongoing development and extensions

The **laGP** package is under active development, and the corpus of code was developed with ease of extension in mind. The calibration application from Section 4 is a perfect example: simple functions tap into local GP emulators and full GP discrepancies alike, and are paired with existing direct optimizing subroutines from other packages for a powerful solution to large scale calibration problems that are becoming commonplace in the recent literature. As mentioned in Section 3.2, the implementation of separable modeling for local analysis is under active development and testing. Many of the associated subroutines (e.g., `laGPsep` and `aGPsep`) are available for use in the latest version of the package.

The library comprises roughly fifty R functions, although barely a fraction of those are elevated to the user's namespace for use in a typical R session. Many of the inaccesible/undocumented

functions have a purpose which, at this time, seem less directly useful outside their calling environment, but may eventually be promoted. Many higher level functions, like `laGP` and `aGP` which access `C` subroutines, have a development-analog (`laGP.R` and `aGP.R`) implementing similar (usually with identical output, our a superset of output) subroutines entirely in `R`. These were used as stepping stones in the development of the `C` versions; however they remain relevant as a window into the inner-workings of the package and as a skeleton for curious users' ambitions for new extensions. The local approximate GP methodology is, in a nutshell, just a judicious combination of established subroutines from the recent spatial statistics and computer experiments literature. We hope that exposing those combinations in well-organized code will spur others to take a similar tack in developing their own solutions in novel contexts.

One example involves deploying basic package functionality—only utilizing full (non local) GP subroutines—for solving blackbox optimization problems under constraints. Gramacy, Gray, Le Digabel, Lee, Ranjan, Wells, and Wild (2014) showed how the augmented Lagrangian (AL), an apparatus popular for solving similar constrained optimization problems in the recent literature (see, e.g., Kannan and Wild 2012), could be combined with the method of expected improvement (EI; Jones, Schonlau, and Welch 1998) to solve a particular type of optimization where the objective was known (and in particular was linear), but where the constraints required (potentially expensive) simulation. Searching for an optimal valid setting of the inputs to the blackbox function could be substantially complicated by a difficult-to-map constraint satisfaction boundary. The package includes a demo [see `demo("ALfhat")`] showcasing a variation on one of the examples from Gramacy *et al.* (2014). The problem therein involves modeling an objective and two constraints with GP predictors, together with an EI calculation on an AL predictive composite. The demo shows how the new, statistical, AL method outperforms the non-statistical analog.

Acknowledgments

Most of the work for this article was completed while the author was in the Booth School of Business at The University of Chicago. The author is grateful for partial support from National Science Foundation grant DMS-1521702.

A. Default regularization (priors) and initial values

In the bulk of this document, and in the core package routines (e.g., `laGP`, and `aGP`) the treatment and default generation of initial values, regularization (priors), and bounding boxes, is largely hidden from the user. Some exceptions include places where it is desirable to have each instance of a repeated call, e.g., in a Monte Carlo experiment, share identical inferential conditions across subtly varying (randomly generated) data sets. In those cases, `darg` and `garg` generate values that control and limit the behaviors of the estimating algorithms for the lengthscale (θ/`d`) and nugget (η/`g`), respectively. Although the package allows inference to proceed without regularization (true MLEs), and arbitrary starting values to be provided, generating sensible ones automatically is a key component in guaranteeing stable behavior out-of-the-box. In settings where potentially thousands of such calculations occur in parallel and without opportunity for individual scrutiny or intervention, such as via `aGP` [Section 2.2],

sensible defaults are essential.

The two methods `darg` and `garg`, which are invoked by `aGP` and `laGP` unless overrides are provided, leverage crude input summary statistics. For example, `darg` calculates squared distances between elements of the design matrix `X` to determine appropriate regularization. A bounding box for `d` is derived from the min and max distances, and a diffuse Gamma prior prescribed with `shape = 3/2` and `scale` set so that the maximum squared distance lies at the position of the 95% quantile. Together these define the regularization of MLE estimates for `d`, or equivalently depict (a search for) the maximum *a posteriori* (MAP) value. We prefer the term MLE as the purpose of the prior is to guard against pathologies, rather than to interject information. The starting `d`-value is chosen the 10% quantile of the calculated distances.

The `garg` function makes similar calculations on the sum of squared residuals in `y` from `mean(y)`, an exception being that by default the minimum nugget value is taken to be `sqrt(.Machine$double.eps)`. When invoked by a higher level routine such as `aGP` or `laGP`, the output values of `darg` and `garg` can be overridden via the `d` and `g` arguments by specifying list elements of the same names as the output values they are overriding. The outputs can also be fed to other, lower level routines such as `mleGP`.

B. Custom compilation

Here we provide hints for enabling the parallelization hooks, via **OpenMP** for multi-core machines and CUDA for graphics cards. The package also includes some wrapper functions, like `aGP.parallel`, which allow a large predictive set to be divvied up amongst multiple nodes in a cluster established via the **parallel** or **snow** (Tierney, Rossini, Li, and Sevcikova 2013) packages.

B.1. With OpenMP for SMP parallelization

Several routines in the **laGP** package include support for parallelization on multi-core machines. The most important one is `aGP`, allowing large prediction problems to be divvied up and distributed across multiple threads to be run in parallel. The speedups are roughly linear as long as the numbers of threads is less than or equal to the number of cores. This is controlled through the `omp.threads` argument.

If R is compiled with **OpenMP** support enabled—which at the time of writing is standard in most builds—then no special action is needed in order to extend that functionality to **laGP**. It will just work. One way to check if this is the case on your machine is to provide an `omp.threads` argument, say to `aGP`, that is bigger than one. If **OpenMP** support is not enabled then you will get a warning. If you are working within a well-managed supercomputing facility, with a custom R compilation, it is likely that R has been properly compiled with **OpenMP** support. If not, perhaps it is worth requesting that it be re-compiled as there are many benefits to doing so, beyond those that extend to the **laGP** package. For example, many linear algebra intensive packages, of which **laGP** is one, benefit from linking to MKL libraries from Intel, in addition to **OpenMP**. Note, however, that some customized libraries (e.g., **OpenBLAS**) are not compatible with **OpenMP** because they are not (at the time of writing) thread safe.

At the time of writing, some incompatibilities between multi-threaded BLAS (e.g., Intel MKL) and OpenMP (e.g., non-Intel, like with GCC) are still in the process of being resolved. In

some builds and instantiations **laGP** can create nested **OpenMP** threads of different types (Intel for linear algebra, and GCC for parallel local design). Problematic behavior has been observed when using `aGPsep` with GCC OpenMP and MKL multi-threaded linear algebra. Generally speaking, since **laGP** uses threads to divvy up local design tasks, a threaded linear algebra subroutine library is not recommended in combination with these routines.

In the case where you are using a standard R binary, it is still possible to compile **laGP** from source with **OpenMP** features assuming your compiler (e.g., GCC) supports them. This is a worthwhile step if you are working on a multi-core machine, which is rapidly becoming the standard setup. For those with experience compiling R packages from source, the procedure is straightforward and does not require installing a bespoke version of R. Obtain the package source (e.g., from CRAN) and, before compiling, open up the package and make two small edits to laGP/src/Makevars. These instructions assume a GCC compiler. For other compilers, please consult documentation for appropriate flags.

1. Replace `$(SHLIB_OPENMP_CFLAGS)` in the `PKG_CFLAGS` line with `-fopenmp`.
2. Replace `$(SHLIB_OPENMP_CFLAGS)` in the `PKG_LIBS` line with `-lgomp`

The laGP/src/Makevars file contains commented out lines which implement these changes. Once made, simply install the package as usual, either doing "R CMD INSTALL" on the modified directory, or after re-tarring it up. Note that for Apple machines as of Xcode v5, with OSX Mavericks, the `Clang` compiler provided by Apple does not include OpenMP support. We suggest downloading GCC v9 or later, for example from `http://hpc.sourceforge.net`, and following the instructions therein.

If hyperthreading is enabled, then a good default for `omp.threads` is two-times the number of cores. Choosing an `omp.threads` value which is greater than the max allowed by the **OpenMP** configuration on your machine leads to a notice being printed indicating that the max-value will be used instead.

B.2. With NVIDIA CUDA GPU support

The package supports graphics card acceleration of a key subroutine: searching for the next local design sight x_{j+1} over a potentially vast number of candidates $X_N \setminus X_n(x)$—Step 2(b) in Figure 7. Custom complication is required to enable this feature, the details of which are described here, and also requires a properly configured Nvidia Graphics card, drivers, and compilation programs (e.g., the Nvidia CUDA compiler `nvcc`). Compiling and linking to CUDA libraries can be highly architecture and operating system specific, therefore the very basic instructions here may not work widely. They have been tested on a variety of Unix-alikes including Intel-based Ubuntu Linux and OSX systems.

First compile the `alc_gpu.cu` file into an object using the Nvidia CUDA complier. E.g., after untarring the package change into `laGP/src` and do

```
% nvcc -arch=sm_20 -c -Xcompiler -fPIC alc_gpu.cu -o alc_gpu.o
```

Alternatively, you can use/edit the "`alc_gpu.o:`" definition in the `Makefile` provided.

Then, make the following changes to `laGP/src/Makevars`, possibly augmenting changes made above to accommodate **OpenMP** support. **OpenMP** (i.e., using multiple CPU threads) brings out the best in our GPU implementation.

1. Add `-D_GPU` to the `PKG_FLAGS`

2. Add `alc_gpu.o -L /software/cuda-5.0-el6-x86_64/lib64 -lcudart` to `PKG_LIBS`. Please replace "`/software/cuda-5.0-el6-x86_64/lib64`" with the path to the CUDA libs on your machine. CUDA 4.x has also been tested.

The `laGP/src/Makvars` file contains commented out lines which implement these changes. Once made, simply install the package as usual. Alternatively, use `make allgpu` via the definitions in the `Makefile` to compile a standalone shared object.

The four functions in the package with GPU support are `alcGP`, `laGP`, `aGP`, and `aGP.parallel`. The first two have a simple switch which allows a single search (Step 2(b)) to be off-loaded to a single GPU. Both also support off-loading the same calculations to multiple cores in a CPU, via `OpenMP` if enabled. The latter `aGP` variations control the GPU interface via two arguments: `num.gpus` and `gpu.threads`. The former specifies how many GPUs you wish to use, and indicating more than you actually have will trip an error. The latter, which defaults to `gpu.threads = num.gpus`, specifies how many CPU threads should be used to queue GPU jobs. Having `gpu.threads < num.gpus` is an inefficient use of resources, whereas `gpu.threads > num.gpus`, up to `2*num.gpus` will give modest speedups. Having multiple threads queue onto the same GPU reduces the amount of time the GPU is idle. `OpenMP` support must be included in the package to have more than one GPU thread.

By default, `omp.threads` is set to zero when `num.gpus > 1` since divvying the work amongst GPU and CPU threads can present load balancing challenges. However, if you get the load balancing right you can observe substantial speedups. Gramacy *et al.* (2014) saw up to 50% speedups, and recommend a scheme for allocating `omp.threads=10` with a setting of `nn.gpu` that allocates about 90% of the work to GPUs (`nn.gpu = floor(0.9*nrow(XX))`) and 10% to the ten **OpenMP** threads. As with `omp.threads`, `gpu.threads` maxes out at the maximum number of threads indicated by your **OpenMP** configuration. Moreover, `omp.threads + gpu.threads` must not exceed that value. When that happens both are first thresholded independently, then `omp.threads` may be further reduced to stay within the limit.

References

Akima H, Gebhardt A, Petzoldt T, Maechler M (2013). *akima: Interpolation Of Irregularly Spaced Data*. R package version 0.5-11, URL http://CRAN.R-project.org/package=akima.

Audet C, Dennis, Jr J (2006). "Mesh Adaptive Direct Search Algorithms For Constrained Optimization." *SIAM Journal on Optimization*, **17**(1), 188–217. doi:doi:10.1137/040603371. URL http://dx.doi.org/doi:10.1137/040603371.

Ba S, Joseph V (2012). "Composite Gaussian Process Models For Emulating Expensive Functions." *Annals of Applied Statistics*, **6**(4), 1838–1860.

Bastos L, O'Hagan A (2009). "Diagnostics For Gaussian Process Emulators." *Technometrics*, **51**(4), 425–438.

Berger J, De Oliveira V, Sanso B (2001). "Objective Bayesian Analysis Of Spatially Correlated Data." *Journal of the American Statistical Association*, **96**, 1361–1374.

Bornn L, Shaddick G, Zidek J (2012). "Modelling Nonstationary Processes Through Dimension Expansion." *Journal of the American Statistical Association*, **107**(497), 281–289.

Brent R (1973). *Algorithm For Minimization Without Derivatives*. Prentice–Hall, Englewood Cliffs, N.J.

Carnell R (2012). ***lhs**: Latin Hypercube Samples*. R package version 0.10, URL http://CRAN.R-project.org/package=lhs.

Chipman H, Ranjan P, Wang W (2012). "Sequential Design for Computer Experiments With A Flexible Bayesian Additive Model." *Canadian Journal of Statistics*, **40**(4), 663–678. ArXiv:1203.1078.

Cohn DA (1996). "Neural Network Exploration Using Optimal Experimental Design." In *Advances in Neural Information Processing Systems*, volume 6(9), pp. 679–686. Morgan Kaufmann Publishers.

Conn A, Scheinberg K, Vicente L (2009). *Introduction To Derivative-Free Optimization*. SIAM, Philadelphia.

Cressie N (1993). *Statistics For Spatial Data,* revised edition. John Wiley & Sons.

Cressie N, Johannesson G (2008). "Fixed Rank Kriging For Very Large Data Sets." *Journal of the Royal Statistical Soceity, Series B*, **70**(1), 209–226.

Dancik GM (2013). ***mlegp**: Maximum Likelihood Estimates Of Gaussian Processes*. R package version 3.1.4, URL http://CRAN.R-project.org/package=mlegp.

Datta A, Banerjee S, Finley AO, Gelfand AE (2015). "Hierarchical Nearest-Neighbor Gaussian Process Models For Large Geostatistical Datasets." *Journal of the American Statistical Association. to appear*; arXiv:1406.7343.

Eidsvik J, Shaby BA, Reich BJ, Wheeler M, Niemi J (2014). "Estimation And Prediction In Spatial Models With Block Composite Likelihoods." *Journal of Computational and Graphical Statistics*, **23**(2), 295–315. doi:10.1080/10618600.2012.760460.

Emory X (2009). "The Kriging Update Equations And Their Application To The Selection Of Neighboring Data." *Computational Geosciences*, **13**(3), 269–280.

Finley AO, Banerjee S (2013). ***spBayes**: Univariate And Multivariate Spatial-Temporal Modeling*. R package version 0.3-8, URL http://CRAN.R-project.org/package=spBayes.

Franey M, Ranjan P, Chipman H (2012). "A Short NoteOon Gaussian Process Modeling For Large Datasets Using Graphics Processing Units." *Technical report*, Acadia University.

Genz A, Bretz F (2009). *Computation Of Multivariate Normal And t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg. ISBN 978-3-642-01688-2.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2014). ***mvtnorm**: Multivariate Normal And t Distributions*. R package version 1.0-1, URL http://CRAN.R-project.org/package=mvtnorm.

Goh J, Bingham D, Holloway JP, Grosskopf MJ, Kuranz CC, Rutter E (2013). "Prediction And Computer Model Calibration Using Outputs From Multi-fidelity Simulators." *Technometrics*, **55**(4), 501–512.

Gramacy R, Bingham D, Holloway JP, Grosskopf MJ, Kuranz CC, Rutter E, Trantham M, Drake PR (2015). "Calibrating a large computer experiment simulating radiative shock hydrodynamics." *Annals of Applied Statistics*, **9**(3), 1141–1168. Also available as arXiv:1410.3293.

Gramacy R, Haaland B (2015). "Speeding Up neighborhood search In Local Gaussian Process Prediction." *Technometrics. to appear*; arXiv:1409.0074.

Gramacy R, Lee H (2011). "Cases For The Nugget In Modeling Computer Experiments." *Statistics and Computing*, **22**(3).

Gramacy R, Niemi J, Weiss R (2014). "Massively Parallel Approximate Gaussian Process Regression." *Journal of Uncertainty Quantification*, **2**(1), 564–584.

Gramacy R, Polson N (2011). "Particle Learning Of Gaussian Process Models For Sequential Design And Optimization." *Journal of Computational and Graphical Statistics*, **20**(1), 102–118. doi:10.1198/jcgs.2010.09171.

Gramacy R, Taddy M, Wild S (2013). "Variable Selection And Sensitivity Analysis Via Dynamic Trees With An Application To Computer Code Performance Tuning." *Annals of Applied Statistics*, **7**, 51–80. doi:10.1214/12-AOAS590.

Gramacy RB (2007). "**tgp**: An R Package For Bayesian Nonstationary, Semiparametric Nonlinear Regression And Design By Treed Gaussian Process Models." *Journal of Statistical Software*, **19**(9), 1–46. URL http://www.jstatsoft.org/v19/i09/.

Gramacy RB (2014). ***laGP**: Local Approximate Gaussian Process Regression*. R package version 1.1-4, URL http://bobby.gramacy.com/r_packages/laGP.

Gramacy RB, Apley DW (2015). "Local Gaussian Process Approximation For Large Computer Experiments." *Journal of Computational and Graphical Statistics*, **24**(2), 561–578. See arXiv:1303.0383.

Gramacy RB, Gray G, Le Digabel S, Lee H, Ranjan P, Wells G, Wild S (2014). "Modeling an Augmented Lagrangian for Blackbox Constrained Optimization." *Technometrics. to appear*; arXiv:1403.4890.

Gramacy RB, Lee HKH (2009). "Adaptive Design And Analysis Of Supercomputer Experiments." *Technometrics*, **51**(2), 130–145.

Gramacy RB, Taddy M (2010). "Categorical Inputs, Sensitivity Analysis, Optimization and Importance Tempering with **tgp** Version 2, an R Package for Treed Gaussian Process Models." *Journal of Statistical Software*, **33**(6), 1–48. URL http://www.jstatsoft.org/v33/i06/.

Haaland B, Qian P (2011). "Accurate Emulators for Large-Scale Computer Experiments." *Annals of Statistics*, **39**(6), 2974–3002. doi:10.1214/11-AOS929.

Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD (2004). "Combining Field Data And Computer Simulations For Calibration And Prediction." *SIAM Journal on Scientific Computing*, **26**(2), 448–466.

Jones DR, Schonlau M, Welch WJ (1998). "Efficient Global Optimization Of Expensive Black Box Functions." *J. of Global Optimization*, **13**, 455–492.

Joseph V (2006). "Limit Kriging." *Technometrics*, **48**(4), 548–466.

Kannan A, Wild SM (2012). "Benefits Of Deeper Analysis In Simulation-Based Groundwater Optimization Problems." In *Proceedings of the XIX International Conference on Computational Methods in Water Resources (CMWR 2012)*.

Kaufman C, Bingham D, Habib S, Heitmann K, Frieman J (2012). "Efficient Emulators Of Computer Experiments Using Compactly Supported Correlation Functions, With An Application to Cosmology." *Annals of Applied Statistics*, **5**(4), 2470–2492.

Kennedy M, O'Hagan A (2001). "Bayesian Calibration Of Computer Models (with discussion)." *Journal of the Royal Statistical Society B*, **63**, 425–464.

Le Digabel S (2011). "Algorithm 909: **NOMAD**: Nonlinear Optimization With The MADS Algorithm." *ACM Transactions on Mathematical Software*, **37**(4), 44:1–44:15. doi:10.1145/1916461.1916468. URL http://dx.doi.org/10.1145/1916461.1916468.

Liu F, Bayarri M, Berger J (2009). "Modularization In Bayesian Analysis, With Emphasis On Analysis Of Computer Models." *Bayesian Analysis*, **4**(1), 119–150.

MacDoanld B, Chipman H, Ranjan P (2014). ***GPfit**: Gaussian Processes Modeling*. R package version 0.2-0, URL http://CRAN.R-project.org/package=GPfit.

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

McKay M, Conover W, Beckman R (1979). "A Comparison Of Three Methods For Selecting Values Of Input Variables In The Analysis Of Output From A Computer Code." *Technometrics*, **21**(2), 239–245. URL http://www.jstor.org/stable/1268522.

Morris D, Mitchell T, Ylvisaker D (1993). "Bayesian Design And Analysis Of Computer Experiments: Use Of Derivatives In Surface Prediction." *Technometrics*, **35**, 243–255.

Nychka D, Furrer R, Sain S (2014). ***fields**: Tools For Spatial Data*. R package version 7.1, URL http://CRAN.R-project.org/package=fields.

Paciorek C, Lipshitz B, Prabhat, Kaufman C, Zhuo T, Thomas R (2014). ***bigGP**: Distributed Gaussian Process Calculations*. R package version 0.1-3, URL http://CRAN.R-project.org/package=bigGP.

Paciorek CJ, Lipshitz B, Zhuo W, Prabhat, Kaufman CGG, Thomas RC (2015). "Parallelizing Gaussian Process Calculations In R." *Journal of Statistical Software*, **63**(10), ??–?? ISSN 1548-7660. URL http://www.jstatsoft.org/v63/i10.

Pratola MT, Chipman H, Gattiker J, Higdon D, McCulloch R, Rust W (2014). "Parallel Bayesian Additive Regression Trees." *Journal of Computational and Graphical Statistics*, **23**(3), 830–852.

R Core Team (2014). *R: A Language And Environment For Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Racine JS, Nie Z (2012). ***crs**: Categorical Regression Splines*. R package version 0.15-18, URL https://github.com/JeffreyRacine/R-Package-crs/.

Ranjan P, Haynes R, Karsten R (2011). "A Computationally Stable Approach To Gaussian Process Interpolation Of Deterministic Computer Simulation Data." *Technometrics*, **53**(4), 363–378.

Rasmussen CE, Williams CKI (2006). *Gaussian Processes For Machine Learning*. The MIT Press.

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989). "Design And Analysis Of Computer Experiments." *Statistical Science*, **4**, 409–435.

Sang H, Huang JZ (2012). "A Full Scale Approximation Of Covariance Functions For Large Spatial Data Sets." *Journal of the Royal Statistical Soceity, Series B*, **74**(1), 111–132. ISSN 1467-9868. doi:10.1111/j.1467-9868.2011.01007.x.

Santner TJ, Williams BJ, Notz WI (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, NY.

Schmidt AM, 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 BW (1985). "Some Aspects Of The Spline Smoothing Approach To Non-Parametric Curve Fitting." *Journal of the Royal Statistical Society, Series B*, **47**, 1–52.

Snelson E, Ghahramani Z (2006). "Sparse Gaussian Processes Using Pseudo-inputs." In *Advances in Neural Information Processing Systems*, pp. 1257–1264. MIT press.

Stein ML (1999). *Interpolation Of Spatial Data*. Springer-Verlag, New York, NY.

Stein ML, Chi Z, Welty LJ (2004). "Approximating Likelihoods For Large Spatial Data Sets." *Journal of the Royal Statistical Society, Series B*, **66**(2), 275–296.

Tierney L, Rossini AJ, Li N, Sevcikova H (2013). ***snow**: Simple Network of Workstations*. R package version 0.3-12, URL http://CRAN.R-project.org/package=snow.

Vecchia A (1988). "Estimation And Model Identification For Continuous Spatial Processes." *Journal of the Royal Statistical Soceity, Series B*, **50**, 297–312.

Venables WN, Ripley BD (2002). *Modern Applied Statistics With S*. Fourth edition. Springer-Verlag, New York. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

Worley B (1987). "Deterministic Uncertainty Analysis." *Technical Report ORN-0628*, National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, USA.

Affiliation:

Robert B. Gramacy
Department of Statistics
Virginia Tech
250 Drillfield Drive
Blacksburg, VA 24061 USA
E-mail: rbg@vt.edu
URL: http://bobby.gramacy.com