Adaptive Design for Non-Stationary Surfaces using Changes in Slope

Marian Frazier
Department of Statistics, The Ohio State University

Goal

To develop a sequential design method for computer experiments that is efficient at fitting the entire response surface. We desire this method to be especially effective for non-stationary responses.

Gaussian Processes

In computer experiments [4], we model a single output, $y(\cdot)$, evaluated at a set of inputs $\boldsymbol{x} \in \boldsymbol{X} \subset \mathbb{R}^{m_X}$. Assume that the output $y(\cdot)$ is a deterministic realization of a stochastic process, $Y(\boldsymbol{x})$.

$$\textbf{Model: } Y(\boldsymbol{x}) = \sum_{j=1}^{p} \beta_j f_j(\boldsymbol{x}) + Z(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\,\boldsymbol{\beta} + Z(\boldsymbol{x}), \quad (1)$$

where $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x})), f_2(\boldsymbol{x})), \ldots, f_p(\boldsymbol{x}))^T$ is a $p \times 1$ vector of known regression functions and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a $p \times 1$ vector of regression coefficients. $Z(\boldsymbol{x})$ is a Gaussian Random Function, with mean 0, variance σ_Z^2, and correlation function $R(\boldsymbol{x}_1, \boldsymbol{x}_2)$. The process $Z(\cdot)$ is completely determined by its covariance function

$$\text{Cov}[Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)] = \sigma_Z^2 R(\boldsymbol{x}_1, \boldsymbol{x}_2). \quad (2)$$

Bayesian Treed Gaussian Processes (BTGPs)

Classification and regression tree model (CART) [1]: Partition the space into two or more (non-overlapping) regions in which the distribution of the response variable Y is more homogeneous. In this way, build a "tree" of successive partitions that places each observation into some terminal node, each defining a model for Y that is appropriate to all of the observations in that node.
Bayesian Treed Gaussian Process model (TGP) [2]: Fit independent stationary Gaussian process models to each of the non-overlapping $\mathcal{N}$ terminal nodes of a tree $\mathcal{T}$. The correlation matrix $\boldsymbol{R}_\nu$ is specific to each region, but it is assumed that all $\boldsymbol{R}_\nu\,(\nu = 1, \ldots, \mathcal{N})$ come from the power exponential family, with range parameters $\boldsymbol{d}$ and the addition of a nugget parameter g_ν.

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth.

[2] Gramacy, R. B. (2005), "Bayesian Treed Gaussian Process Models," Ph.D. thesis, University of California, Santa Cruz.

[3] Gramacy, R. B. and Taddy, M. A. (2008), "tgp: Bayesian treed Gaussian process models," R package version 2.1-2.

[4] Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag.

Expected Difference of Slopes (E∆M)

This sequential design method focuses on the search for areas with large changes in slope, with the idea that sudden changes in slope are an indication of non-stationary "breaks" in the response.
Consider three points $(\boldsymbol{x}_1, Y(\boldsymbol{x}_1)), (\boldsymbol{x}_2, Y(\boldsymbol{x}_3)), (\boldsymbol{x}_3, y_3)$, such that $\boldsymbol{x}_3$ has already been sampled ($\boldsymbol{x}_3 \in \boldsymbol{X}$) but $\boldsymbol{x}_1, \boldsymbol{x}_2$ have not. We choose a new point $\boldsymbol{x}_0$ that resides in the region where the difference in slopes between these three points is expected to be the largest.

$$\begin{aligned} E[(\text{slope diff})^2] = \frac{1}{\Delta x_{32}^2 \Delta x_{21}^2} \times \{ & \Delta x_{21}^2 y_3^2 \\ & + \Delta x_{31}^2[\hat{\sigma}^2(\boldsymbol{x}_2) + \hat{Y}(\boldsymbol{x}_2)^2] + \Delta x_{32}^2[\hat{\sigma}^2(\boldsymbol{x}_1) + \hat{Y}(\boldsymbol{x}_1)^2] \\ & - 2\Delta x_{21}\Delta x_{31}\hat{Y}(\boldsymbol{x}_2) y_3 + 2\Delta x_{21}\Delta x_{32}\hat{Y}(\boldsymbol{x}_1) y_3 \\ & -2\Delta x_{31}\Delta x_{32}\left[\text{Cov}(\boldsymbol{x}_2, \boldsymbol{x}_1) + \hat{Y}(\boldsymbol{x}_2)\hat{Y}(\boldsymbol{x}_1)\right]\}. \end{aligned} \quad (3)$$

The prediction equations for $\hat{Y}(\boldsymbol{x}_i)$ and $\hat{\sigma}^2(\boldsymbol{x}_i)$ are using the BTGP model. $\text{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is estimated using the correlation matrix $\boldsymbol{R}$ and its range and nugget parameters $(\boldsymbol{d}, g)$. All of these estimates are calculated as part of the R package `tgp` [3].

Implementation: Choose a relatively dense grid of N' candidate points XX and an initial set of N_0 training points from which a predicted surface is estimated (using BTGP). To find the next point satisfying the E∆M criterion:

1. Pick a candidate point, $\boldsymbol{x}_0$, from XX.
2. Determine which c points in $\boldsymbol{X}$ are closest to $\boldsymbol{x}_0$; call these values $\boldsymbol{x}_*$.
3. For each of the c $\boldsymbol{x}_*$ points, find the point in XX that is halfway between this point and $\boldsymbol{x}_0$. These are called the "midpoints."
4. We now have a set of three points associated with each closest point: $\boldsymbol{x}_0$ (in XX), the closest point $\boldsymbol{x}_*$ (from $\boldsymbol{X}$), and the midpoint between them (in XX).
5. For each of the c sets, calculate $E[(\text{slope diff})^2]$ for that set.
6. Determine which $E[(\text{slope diff})^2]$ value is largest among the c $\boldsymbol{x}_*$ points; this tells us which *direction* has largest expected numerical second derivative for that $\boldsymbol{x}_0$.
7. Return to Step 1 and repeat for all $\boldsymbol{x}_0$ in XX that are not already part of the sample $\boldsymbol{X}$.
8. We now have $N' - N$ values of $\max\{E[(\text{slope diff})^2]\}$. The new sampled point is in the set with the largest $\max\{E[(\text{slope diff})^2]\}$. Specifically, the new point is the midpoint of that set.

Example: 2d Exponential Function

For $x_1, x_2 \in [-2, 6]$,

$$y(\boldsymbol{x}) = x_1 \exp(-x_1^2 - x_2^2) \quad (4)$$

with added $N(0, \sigma = 0.001)$ random noise. This surface is obviously non-stationary, with a steep peak and trough in $[-2, 2] \times [-2, 2]$ and a plane elsewhere.
On the left side of Figure 1 is the ERMSPE of the fit achieved as N increases for E∆M and Gramacy's Bayesian Adaptive Sampling (BAS) method. It is clear that E∆M achieves a good global fit much quicker than BAS.

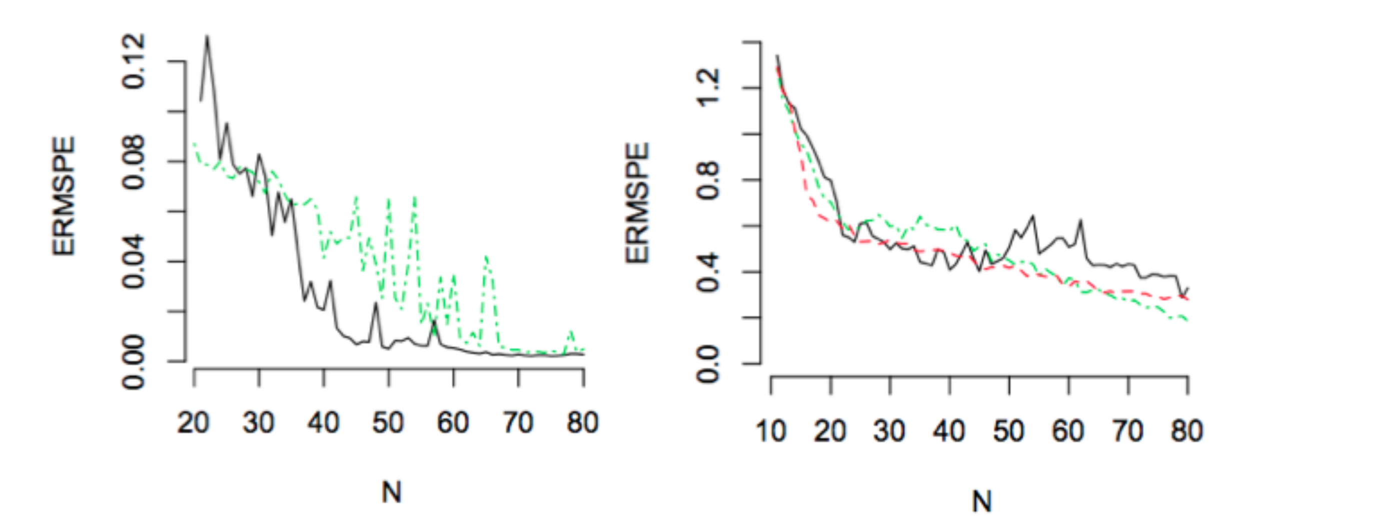

Figure 1: ERMSPE on the 2-d exponential data (left) and 6-hump camel-back function (right), using the E∆M (black), BAS (green), and modifications to the E∆M (red).

Next Steps

Although E∆M is more efficient than BAS in the example above, that is not always the case. However, an alteration of Equation (3) makes the E∆M method more robust on a variety of problems. We can see the effectiveness of this alternate version on the right side of Figure 1, comparing ERMSPE on the 2-dimensional six-hump camel-back function. While the original E∆M does not perform very well, the modified version is at least as efficient as BAS.

Conclusion

Tests on several one- and two-dimensional synthetic examples indicate that E∆M is successful at global fit, especially of non-stationary response surfaces. It performs at least as well as competing sequential design methods, and is a significant improvement over these methods on many examples. This is especially true for the modified version of E∆M, which outperforms the original E∆M in many cases.