BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 17 2008, pages 1966–1967
doi:10.1093/bioinformatics/btn329

Systems biology

mlegp: statistical analysis for computer models of biological systems using R

Garrett M. Dancik [1,2] and Karin S. Dorman [1,2,3,*]

[1]Program in Bioinformatics & Computational Biology, [2]Department of Statistics and [3]Department of Genetics, Development & Cell Biology, Iowa State University, Ames, IA 50010, USA

Received on October 7, 2007; revised on May 2, 2008; accepted on June 23, 2008

Advance Access publication July 17, 2008

Associate Editor: John Quackenbush

ABSTRACT

Summary: Gaussian processes (GPs) are flexible statistical models commonly used for predicting output from complex computer codes. As such, GPs are well suited for the analysis of computer models of biological systems, which have been traditionally difficult to analyze due to their high-dimensional, non-linear and resource-intensive nature. We describe an R package, *mlegp*, that fits GPs to computer model outputs and performs sensitivity analysis to identify and characterize the effects of important model inputs.

Availability: http://www.biomath.org/mlegp

Contact: kdorman@iastate.edu

Supplementary information: See http://www.biomath.org/mlegp for a user manual and examples.

1 INTRODUCTION

Gaussian processes (GPs) commonly facilitate analysis of computer models that are high-dimensional, non-linear and resource-intensive (Santner *et al*., 2003) by serving as fast and accurate emulators of these models. GPs play a prominent role in computer model calibration and validation (Bayarri *et al*., 2007; Kennedy and O'Hagan, 2001), as well as sensitivity analysis (SA) to rank inputs in order of importance [based on functional analysis of variance (FANOVA) decomposition] and to characterize their effects (through visual plots of main and two-way factor interactions) (Schonlau and Welch, 2006).

We describe an R package, *mlegp*, that implements GP modeling with power exponential correlation structure (Santner *et al*., 2003), the SA methods described in Schonlau and Welch (2006) and the modeling of functional computer model output described in Heitmann *et al*. (2006). In addition, *mlegp* extends previous GP models to handle stochastic computer output with non-constant (heteroscedastic) variance by no longer requiring a constant nugget term across observations.

The package is appropriate for what Kitano (2002) describes as simulation-based research in systems biology. In this context, computer models have been used to simulate gene expression and signal transduction pathways, e.g. in *Escherichia coli* (Dobrzyński *et al*., 2007); and infectious disease at the cellular level, e.g. *Mycobacterium tuberculosis* infection (Segovia-Juarez *et al*., 2004).

*To whom correspondence should be addressed.

We demonstrate the capabilities of *mlegp* by analyzing a computer model of parasitic infection.

2 STATISTICAL METHODS

2.1 The Gaussian process

Let $z_{\text{obs}} = [z(x^{(1)}), \ldots, z(x^{(m)})]$ be a vector of observed computer model outputs for m choices of the input vector $x^{(i)} = [x_1^{(i)}, \ldots, x_p^{(i)}]$. We are interested in predicting output $z(t)$ at untried input t. The correlation between any two computer model outputs is assumed to have the form

$$C_{ij} \equiv \text{cor}(z(x^{(i)}), z(x^{(j)})) = e^{-\sum_{k=1}^{p} \beta_k (x_k^{(i)} - x_k^{(j)})^2} \quad (1)$$

Let $\mu(x)$ be the unconditional mean $\text{E}[z(x)]$. Define the mean matrix

$$M \equiv \text{E}[z_{\text{obs}}] = [\mu(x^{(1)}), \ldots, \mu(x^{(m)})].$$

Under the GP model, computer output follows a multivariate normal distribution

$$z_{\text{obs}} \sim MVN_m(M, \sigma_{GP}^2 C + \sigma_e^2 I), \quad (2)$$

where I is the $m \times m$ identity matrix, $C \equiv \{C_{ij}\}$ from Equation (1), σ_{GP}^2 is the unconditional variance of mean computer model output and the nugget σ_e^2 accounts for computer model stochasticity. For convenience, denote the variance–covariance matrix as V. Then, the GP predictive distribution of $z(t)$ is normal with mean and variance

$$\text{E}[z(t)|z_{\text{obs}}] = \mu(t) + \sigma_{GP}^2 r' V^{-1} (z_{\text{obs}} - M)$$

$$\text{Var}[z(t)|z_{\text{obs}}] = \sigma_{GP}^2 + \sigma_e^2 - \sigma_{GP}^4 r' V^{-1} r,$$

where $\text{r} = [r_1, \ldots, r_m]'$, with $r_i = \text{cor}(z(t), z(x^{(i)}))$ following Equation (1). For more details, see Santner *et al*. (2003).

2.2 Sensitivity analysis

Schonlau and Welch (2006) describe SA of computer models using GPs. For independent marginal priors on the components of x, the total variance of the GP predictor can be decomposed into contributions from single and interacting inputs, a technique called FANOVA decomposition. The percentage of total functional variance attributed to the main effect of an input or the interaction effect among inputs provides a measure of the importance of that effect. The main effect $\text{E}[z(x)|z_{\text{obs}}, x_k]$ of the k-th input variable predicts output, given x_k and known results z_{obs}, by integrating against a prior $\pi(x_{-k})$ on all remaining variables in x. The two-way interaction effect $\text{E}[z(x)|z_{\text{obs}}, x_k, x_l]$ is similarly defined. Main effects plots and contour plots conveniently illustrate main effects and two-factor interactions as functions of the model inputs x_k and (x_k, x_l).

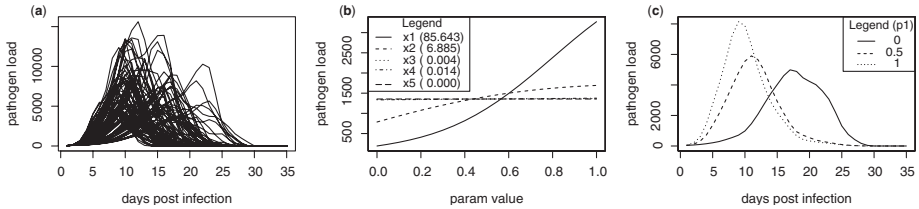

Fig. 1. Sensitivity analysis of a computer model of parasitic infection using the R package *mlegp*. (**a**) Computer model output, consisting of pathogen load over time, for 100 simulations obtained by varying inputs x_1 through x_5, (**b**) main effects for all inputs on pathogen load at 5 dpi, along with the percentage contribution of each effect to the total functional variance of the GP predictor and (**c**) main effect of x_1 on pathogen load over time.

3 SOFTWARE

The package *mlegp*, available in R (R Development Core Team, 2007), finds maximum likelihood estimates of Gaussian process parameters using the likelihood that follows from Equation (2). The package extends previous GP models by allowing the user to replace the identity matrix I in (2) with a user-defined diagonal matrix N. This extension leads to more accurate GP emulators of heteroscedastic computer models when the variance is known or well estimated. Another approach to this problem is implemented in the R package *tgp*, which fits separate GPs to a partitioned input space using a fully Bayesian approach (Gramacy, 2007). Not all non-constant variance can be partitioned in this way. On the other hand, our model requires knowledge of the nugget matrix up to a multiplicative constant.

GPs with constant mean functions (i.e. $\mu(x) \equiv \mu_0$) or linear functions in x are supported. For high-dimensional or functional output such as time-series data, the user can opt to fit independent GPs to individual outputs or, instead, to the most important principle component weights following singular value decomposition of the output (Heitmann *et al*., 2006). For each GP, the R package provides cross-validated diagnostics, performs FANOVA decomposition, and produces plots for all main and two-way factor interaction effects. Main effects for functional output can also be produced.

4 EXAMPLE APPLICATION: ANALYSIS OF A COMPUTER MODEL OF DISEASE

The SA methods we describe have been used to analyze computer models with up to 40 input variables (Schonlau and Welch, 2006). For demonstration purposes, we use *mlegp* to analyze the effects of five input variables ($x = [x_1, \ldots, x_5]$) in a computer model of *Leishmania major* infection (Dancik *et al*., 2006). Inputs are scaled between 0 and 1 and are described in Supplementary Material. Computer model output consists of pathogen load over time (Fig. 1a). Using *mlegp*, we fit a GP to 100 observations of pathogen load at 5 days post infection (dpi). Main effects for all inputs and their FANOVA contributions are provided in Figure 1b. Lastly, we use *mlegp* to calculate the main effect of x_1 (pathogen growth rate) on the temporal evolution of pathogen load by fitting independent GPs to the six most important principle component weights (Fig. 1c). The SA shows that the input variable x_1 is the most important input for determining pathogen load at 5 dpi and has a positive relationship with this response (Fig. 1b). Low values of x_1 result in a gradual increase in pathogen load and a relatively longer infection, whereas high values of x_1 result in a sharp increase in pathogen load and a higher peak, but a fast resolution of the infection (Fig. 1c).

In Supplementary Material, we report additional output from *mlegp*, including GP diagnostic plots, the FANOVA decomposition, and two-way interaction contour plots, for pathogen load at both 5 and 18 dpi. We also illustrate the advantage of using a non-constant nugget term for heteroscedastic computer model output.

ACKNOWLEDGEMENTS

Funding: This work was funded by National Institutes of Health (GM068955) and United States Department of Agriculture (2001-52100-11506).

Conflict of Interest: none declared.

REFERENCES

Bayarri,M. *et al*. (2007) A framework for validation of computer models. *Technometrics*, **49**, 138–154.

Dancik,G.M. *et al*. (2006) An agent-based model for *Leishmania* infection. *Interj. Complex Sys.*, 1853.

Dobrzyński,M. *et al*. (2007). Computational methods for diffusion-influenced biochemical reactions. *Bioinformatics*, **23**, 1969–1977.

Gramacy,R.B. (2007) tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *J. Stat. Soft.*, 19.

Heitmann,K. *et al*. (2006) Cosmic Calibration, *Astrophys. J.*, **646**, L1–L4.

Kennedy,M.C., O'Hagan,A. (2001) Bayesian calibration of computer models. *J.R. Stat. Soc. B*, **63**, 425–464.

Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.

R Development Core Team. (2007) *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at URL http://www.R-project.org/. (last accessed date July 15 2008).

Santner,T.J. *et al*.(2003) *The Design and Analysis of Computer Experiments*. Springer, New York.

Schonlau,M. and Welch,W. (2006). Screening the input variables to a computer model via analysis of variance and visualization. In Dean,A. and Lewis,S. (eds), *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*. Springer, New York, pp. 308–327.

Segovia-Juarez,J.L. *et al*. (2004) Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *J. Theor. Biol.*, **231**, 357–376.