mlegp: an R package for Gaussian process modeling and sensitivity analysis

Garrett Dancik

November 4, 2009

1 *mlegp*: an overview

Gaussian processes (GPs) are commonly used as surrogate statistical models for predicting output of computer experiments (Santner *et al.*, 2003). Generally, GPs are both interpolators and smoothers of data and are effective predictors when the response surface of interest is a smooth function of the parameter space. The package *mlegp* finds *m*aximum *l*ikelihood *e*stimates of *G*aussian *p*rocesses for univariate and multi-dimensional responses, for Gaussian processes with Gaussian correlation structures; constant or linear regression mean functions; and for responses with either constant or non-constant variance that can be specified exactly or up to a multiplicative constant. Unlike traditional GP models, GP models implemented in *mlegp* are appropriate for modelling heteroscedastic responses where variance is known or accurately estimated. Diagnostic plotting functions, and the sensitivity analysis tools of Functional Analysis of Variance (FANOVA) decomposition, and plotting of main and two-way factor interaction effects are implemented. Multi-dimensional output can be modelled by fitting independent GPs to each dimension of output, or to the most important principle component weights following singular value decomposition of the output. Plotting of main effects for functional output is also implemented. From within R, a complete list of functions and vignettes can be obtained by calling 'library(help = "mlegp")'.

2 Gaussian process modeling and diagnostics

2.1 Gaussian processes

Let $z_{\text{known}} = \left[\mathrm{z}(\theta^{(1)}), \ldots, \mathrm{z}(\theta^{(m)})\right]$ be a vector of *observed* responses, where $z(\theta^{(i)})$ is the response at the input vector $\theta^{(i)} = \left[\theta_1^{(i)}, \ldots, \theta_p^{(i)}\right]$, and we are interested in predicting output $z(\theta^{(\text{new})})$ at the untried input $\theta^{(\text{new})}$. The correlation between any two *unobserved* responses is assumed to have the form

$$C(\beta)_{i,t} \equiv \operatorname{cor}\left(z(\theta^{(i)}), z(\theta^{(t)})\right) = \exp\left\{\sum_{k=1}^{p}\left(-\beta_k\left(\theta_k^{(i)} - \theta_k^{(t)}\right)^2\right)\right\}. \tag{1}$$

The correlation matrix $\mathrm{C}(\beta) = [\mathrm{C}(\beta)]_{i,t}$, and depends on the correlation parameters $\beta = [\beta_1, \ldots, \beta_p]$

Let $\mu(\cdot)$ be the mean function for the unconditional mean of any observation, and the mean matrix of z_{known} be

$$M \equiv \left[\mu\left(\theta^{(1)}\right), \ldots, \mu\left(\theta^{(m)}\right)\right]. \tag{2}$$

The vector of observed responses, z_{known}, is distributed according to

$$z_{\text{known}} \sim MVN_m(M, V), \tag{3}$$