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

Garrett Dancik

October 6, 2007

1 *mlegp*: an overview

Gaussian processes (GPs) are commonly used as surrogate statistical models for predicting output of computer experiments (Santner *et al.*, 2003). Generlly, 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 product exponential correlation structures; constant or linear regression mean functions; no nugget term, constant nugget terms, or a nugget matrix that can be specified up to a multiplicative constant. The latter is an extension of previous Gaussian process models and provides some flexibility for using GPs to model heteroscedastic responses. 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 $z_{\text{known}} = \left[z(\theta^{(1)}), \ldots, z(\theta^{(m)})\right]$ be a vector of *observed* responses, where $z(\theta^{(i)})$ is the response observed at the design point $\theta^{(i)}$, the parameter 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 parameter setting $\theta^{(\text{new})}$. The correlation between any two responses (observed or unobserved) is assumed to have the (prodcut exponential) form

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

The correlation matrix $\text{C}(\beta) = [\text{C}(\beta)]_{i,j}$, 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

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

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

$$z_{\text{known}} \sim MVN_m(M, \sigma_{GP}^2 C(\beta) + \sigma_e^2 I), \quad (3)$$