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

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

where V is the variance-covariance matrix defined as

$$V \equiv \sigma_{GP}^2 C(\beta) + N, \tag{4}$$

where σ_{GP}^2 is the unconditional variance of an expected response and N is a diagonal *nugget matrix* with the i^{th} diagonal element equal to $\sigma_{\mathrm{e}}^2(\theta^{(i)})$, which is variance due to the stochasticity of the response (e.g., random noise) that may depend on θ. If output is *deterministic*, the nugget is not present so that $\sigma_{\mathrm{e}}^2(\theta) \equiv 0$. For *stochastic* responses, variance is traditionally taken to be constant so that $\sigma_{\mathrm{e}}^2(\theta) \equiv \sigma_{\mathrm{e}}^2$ and $N = \sigma_{\mathrm{e}}^2 I$. The package *mlegp* extends the traditional GP model by allowing the user to specify N exactly or N up to a multiplicative constant.

Define $r_i = \mathrm{cor}(z(\theta^{(new)}), z(\theta^{(i)}))$, following equation (1), and $r = [r_1, \ldots, r_m]'$. Under the GP assumption, the predictive distribution of $z(\theta^{(new)})$ is normal with mean

$$\widehat{z}\left(\theta^{(i)}\right) = \mathrm{E}[z(\theta^{(new)})|z_{\mathrm{known}}] = \mu(\theta^{(new)}) + \sigma_{GP}^2 r' V^{-1}(z_{\mathrm{known}} - M) \tag{5}$$

and variance

$$\mathrm{Var}[z(\theta^{(new)})|z_{\mathrm{known}}] = \sigma_{\mathrm{GP}}^2 + \sigma_{\mathrm{e}}^2(\theta) - \sigma_{GP}^4 r' V^{-1} r. \tag{6}$$

For more details, see Santner *et al.* (2003).

2.2 Maximum likelihood estimation

We first need some additional notation. Mean functions that are constant or linear in design parameters have the form $\mu(\theta) = x(\theta)F$, where $x(\theta)$ is a row vector of regression parameters, and F is a column vector of regression coefficients. Note that for a constant mean function, $x(\cdot) \equiv 1$ and F is a single value corresponding to the constant mean. The mean matrix M defined in equation (2) has the form $M = XF$, where the i^{th} row of X is equal to $x\left(\theta^{(i)}\right)$.

Let us also rewrite the variance-covariance matrix V from equation (4) to be

$$V \equiv \sigma_{\mathrm{GP}}^2(C(\beta) + aN_s) \equiv \sigma_{\mathrm{GP}}^2 W(\beta, a), \tag{7}$$

where N_s is the nugget matrix specified up to a multiplicative constant, with $N = \sigma_{GP}^2 a N_s$ and the matrix W depends on the correlation parameters $\beta = [\beta_1, \ldots, \beta_p]$ and a proportionality constant a.

When the matrix W is fully specified, maximum likelihood estimates of the mean regression parameters and σ_{GP}^2 exist in closed form and are

$$\widehat{F} = (X^T W^{-1} X)^{-1} X^T W^{-1} z_{\mathrm{known}} \tag{8}$$

and

$$\widehat{\sigma}_{\mathrm{GP}}^2 = \frac{1}{m}(z_{\mathrm{known}} - \widehat{M})^T W^{-1}(z_{\mathrm{known}} - \widehat{M}), \tag{9}$$

where $\widehat{M} = X\widehat{F}$.

2.3 Diagnostics

The cross-validated prediction $\widehat{z}_{\text{-}i}(\theta^{(i)})$ is the predicted response obtained using equation (5) after removing all responses at input vector $\theta^{(i)}$ from z_{known} to produce $z_{\mathrm{known,\text{-}i}}$. Note that it is possible for multiple $\theta^{(i)}$'s, for various i's, to be identical, in which case all corresponding observations are removed. The cross-validated residual for this observations is

$$\frac{z(\theta^{(i)}) - z_{\text{-}i}(\theta^{(i)})}{\sqrt{\mathrm{Var}(z(\theta^{(i)})|z_{\mathrm{known,\text{-}i}})}}. \tag{10}$$

2.4 What does *mlegp* do?

The package *mlegp* extends the standard GP model of (3), which assumes that $N = \sigma_{\mathrm{e}}^2 I$, by allowing the user to specify the diagonal nugget matrix N exactly or up to a multiplicative constant (i.e., N_s). This extension provides some flexibility for modeling heteroscedastic responses. The user also has the option of fitting a GP with a constant mean (i.e., $\mu(\theta) \equiv \mu_0$) or mean functions that are linear regression functions in all elements of θ (plus an intercept term). For multi-dimensional output, the user has the option of fitting independent GPs to each dimension (i.e., each type of observation), or to the most important principle component weights following singular value decomposition. The latter is ideal for data rich situations, such as functional output, and is explained further in Section (5). GP accuracy is analyzed through diagnostic plots of cross-validated predictions and cross-validated residuals, which were described in Section (2.3). Sensitivity analysis tools including FANOVA decomposition, and plotting of main and two-way factor interactions are described in Section (4).

The package *mlegp* employs two general approaches to GP fitting. In the standard approach, *mlegp* uses numerical methods in conjunction with equations (8) and (9) to find maximum likelihood estimates (MLEs) of all GP parameters. However, when replicate runs are available, it is usually more accurate and computationallly more efficient to fit a GP to a collection of *sample means* while using a plug-in estimate for the nugget (matrix).

Let $z_{ij} \equiv z_j\left(\theta^{(i)}\right)$ be the j^{th} replicate output from the computer model evaluated at the input vector $\theta^{(i)}$, $i = 1, \ldots k, j = 1, \ldots n_i$, so that the computer model is evaluated n_i times at the input vector $\theta^{(i)}$. Let $\overline{z} = (\overline{z_{1.}}, \ldots \overline{z_{k.}})$ be a collection of k sample mean computer model outputs, where

$$\overline{z_{i.}} = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}$$

is the sample mean output when the computer model is evaluated at θ.

The GP model of $\overline{z}$ is similar to the GP model of z_{known} described above, with the $(i,t)^{th}$ element of the matrix $C(\beta)$ given by $\mathrm{cor}(\overline{z_{i.}}, \overline{z_{t.}})$, following Eq. (1). and the i^{th} element of the nugget matrix N given by $\frac{\sigma_{\mathrm{e}}^2(\theta)}{n_i}$. The covariance matrix V has the same form as Eq. (4). Predicted means and variances have the same form as Eqs. (5 - 6), but with the vector z_{known} replaced by $\overline{z}$. For a fixed nugget term or nugget matrix, the package *mlegp* can fit a GP to a set of sample means by using numerical methods in combination with Eq. (8) to find the MLE of all remaining GP parameters. The user may specify a value for the constant nugget or nugget matrix to use. Alternatively, if replicate runs are available and a nugget term is not specified, *mlegp* will automatically take $N = \sigma_{\mathrm{e}}^2 I$ and estimate the nugget as

$$\widehat{\sigma_e^2} = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) s_i^2,$$

where, s_i^2 is the sample variance for design point i and $N = \sum_{i=1}^{k} n_i$. This estimate is the best linear unbiased estimate (BLUE) of σ_e^2 (which is linear in s_i^2).

The above *means* approach is computationally more efficient when replicate runs are available. If the nugget term or nugget matrix is well known or can be accurately estimated, the *means* approach is also more accurate than the standard approach.

3 Examples: Gaussian process fitting and diagnostics

3.1 A simple example

The function *mlegp* is used to fit one or more Gaussian processes (GPs) to a vector or matrix of responses observed under the same set of inputs. Data can be input from within R or read from a text file using the command *read.table* (type '?read.table' from within R for more information).