Module 3: Gaussian Process Parameter Estimation, Prediction Uncertainty, and Diagnostics

Jerome Sacks and William J. Welch

National Institute of Statistical Sciences and University of British Columbia

Adapted from materials prepared by Jerry Sacks and Will Welch for various short courses

Acadia/SFU/UBC Course on Dynamic Computer Experiments
September–December 2014

Outline of Topics

Parameters of the Gaussian Process (GP) Model

Recall from Module 2 that the Gaussian process prior for $y(\mathbf{x}) = y(x_1, \ldots, x_d)$ has hyper-parameters:

- mean, μ,
- variance, σ^2
- correlation parameters, e.g., $\theta_1, \ldots, \theta_d$ and $p_1, \ldots, p_d$ for the power-exponential correlation function,

$$R(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} \exp(-\theta_j |x_j - x'_j|^{p_j}).$$

- Their values will be chosen to be consistent with the computer-model runs.

Maximum Likelihood

- Recall also that $y(\mathbf{x})$ is assumed to be Gaussian.
- Hence, $\mathbf{y} = [y(\mathbf{x}^{(1)}), \ldots, y(\mathbf{x}^{(n)})]^T$, the data from the computer model, are a sample from a multivariate-normal distribution.
- The likelihood, $L(\mathbf{y} \mid \mu, \sigma^2, \theta_1, \ldots, \theta_d, p_1, \ldots, p_d)$, is

$$\frac{1}{(2\pi\sigma^2)^{n/2} \det^{1/2}(\mathbf{R})} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mu\mathbf{1})^T \mathbf{R}^{-1}(\mathbf{y} - \mu\mathbf{1})\right).$$

- Maximum likelihood estimation (MLE) chooses the hyper-parameters to maximize this.
- Or use Bayes' rule to get a posterior distribution for the hyper-parameters and for predictions of $y(\mathbf{x})$ (see Appendix A).

Maximum Likelihood: Computation

For fixed correlation parameters,

$$\hat{\mu} = \frac{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{y}}{\mathbf{1}^T \mathbf{R}^{-1} \mathbf{1}}$$

and

$$\widehat{\sigma^2} = \frac{1}{n} (\mathbf{y} - \hat{\mu}\mathbf{1})^T \mathbf{R}^{-1} (\mathbf{y} - \hat{\mu}\mathbf{1})$$

The likelihood function (with $\hat{\mu}$ and $\widehat{\sigma^2}$ substituted) has to be numerically maximized w.r.t. the correlation parameters.

G-Protein Computer Model

Biosystems model for so-termed ligand activation of G-protein in yeast.

$d = 4$ input variables

- x is concentration of ligand
- $u_1, \ldots, u_8$ is a vector of 8 kinetic parameters (only u_1, u_6, and u_7 are varied)

Output variable

- y is the normalized concentration of part of the complex

G-Protein System Dynamics: Differential Equations

1. $\dot{\eta}_1 = -u_1 \eta_1 x + u_2 \eta_2 - u_3 \eta_1 + u_5$
2. $\dot{\eta}_2 = u_1 \eta_1 x - u_2 \eta_2 - u_4 \eta_2$
3. $\dot{\eta}_3 = -u_6 \eta_2 \eta_3 + u_8 (G_{\text{tot}} - \eta_3 - \eta_4)(G_{\text{tot}} - \eta_3)$
4. $\dot{\eta}_4 = u_6 \eta_2 \eta_3 - u_7 \eta_4$
5. $y = (G_{\text{tot}} - \eta_3)/G_{\text{tot}}$

where

- $\eta_1, \ldots, \eta_4$ are concentrations of 4 chemical species and $\dot{\eta}_1 \equiv \frac{\partial \eta_1}{\partial t}$, etc.
- $G_{\text{tot}} =$ (fixed) total concentration of G-protein complex after 30 seconds

Inputs and Code Runs

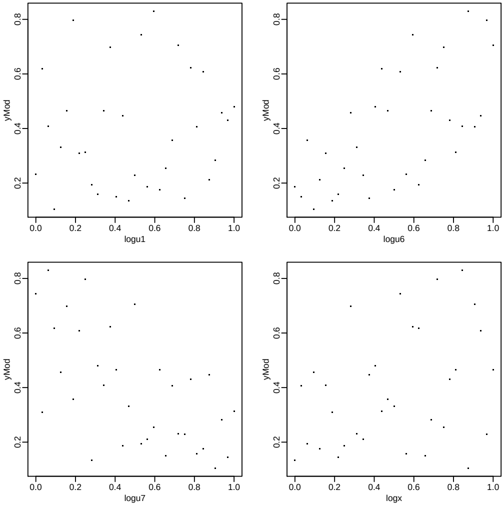
Input variables

- $d = 4$ variables
- Work with $\log(x)$, $\log(u_1)$, $\log(u_6)$, $\log(u_7)$.
- i.e., what we called the **x** vector before is $\log(x)$, $\log(u_1)$, $\log(u_6)$, and $\log(u_7)$ here
- All input variable ranges are normalized to $[0, 1]$ on the log scale

Number of runs

- $n = 33$
  (this choice and the design for the 33 runs is described in Module 4)

Computer Model Data



Gaussian Process (GP) Model

$y(\mathbf{x})$ is a realization of a Gaussian process with:

- mean μ
- variance σ^2
- correlations given by

$$\mathrm{Cor}(y(\mathbf{x}), y(\mathbf{x}')) \equiv R(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{4} e^{-\theta_j |x_j - x_j'|^{p_j}}.$$

The parameters in red need to be estimated.

Maximum Likelihood Estimates

- $\hat{\mu} = 0.36$
- $\hat{\sigma^2} = 0.51$
-

Variable	$\hat{\theta}$	$\hat{p}$
$\log(x)$	0.929	1.98
$\log(u_1)$	0.179	2
$\log(u_6)$	0.082	2
$\log(u_7)$	0.083	2

- It is difficult to interpret the magnitudes of the estimates. (we will revisit this example in Module 5 and do a sensitivity analysis).

"Plug-In" Prediction and Standard Error

Replace all hyper-parameters by their MLEs in the conditional mean and variance formulas:

$$\text{prediction of } y(\mathbf{x}) = \hat{y} = \hat{m}(\mathbf{x}) = \hat{\mu} + \mathbf{r}^T(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}).$$

and

$$\text{estimated variance of prediction} = \hat{v}(\mathbf{x}) = \widehat{\sigma^2}(1 - \mathbf{r}^T(\mathbf{x})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})).$$

($\mathbf{R}$ and $\mathbf{r}(\mathbf{x})$ are also estimates.)
The plug-in estimated variance ignores uncertainty in estimating the hyper-parameters. It can be adapted to include uncertainty from estimating μ:

$$\hat{v}(\mathbf{x}) = \widehat{\sigma^2}\left(1 - \mathbf{r}^T(\mathbf{x})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) + \frac{[1 - \mathbf{1}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x})]^2}{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}\right).$$

This plug-in formula is often used to give a standard error, i.e., $s(\mathbf{x}) = \sqrt{\hat{v}(\mathbf{x})}$.

Measures of Accuracy

- We could rely on the standard error, $\sqrt{\hat{v}(\mathbf{x})}$.
- If we have m test data observations, the root mean squared error (RMSE) of prediction is

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{\text{test pts}} (\hat{y} - y(\mathbf{x}))^2}.$$

  But rarely available.
- Cross validation (CV)

Cross Validation (CV)

Let $\mathbf{x}^{(i)}$ denote $\mathbf{x}$ for run i in the data $(i = 1, \ldots, n)$. For run i:

- The cross validated prediction of $y(\mathbf{x}^{(i)})$ is

$$\hat{y}_{-i}(\mathbf{x}^{(i)}),$$

  i.e., $\hat{y}(\mathbf{x}) = \hat{m}(\mathbf{x})$ computed from the $n-1$ runs excluding run i.

- The cross validated standard error of $\hat{y}_{-i}(\mathbf{x}^{(i)})$ is

$$s_{-i}(\mathbf{x}^{(i)}),$$

  i.e., $s(\mathbf{x}) = \sqrt{\hat{v}(\mathbf{x})}$ computed from the $n-1$ runs excluding run i.

- The cross-validated residual for run i is

$$y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)}).$$

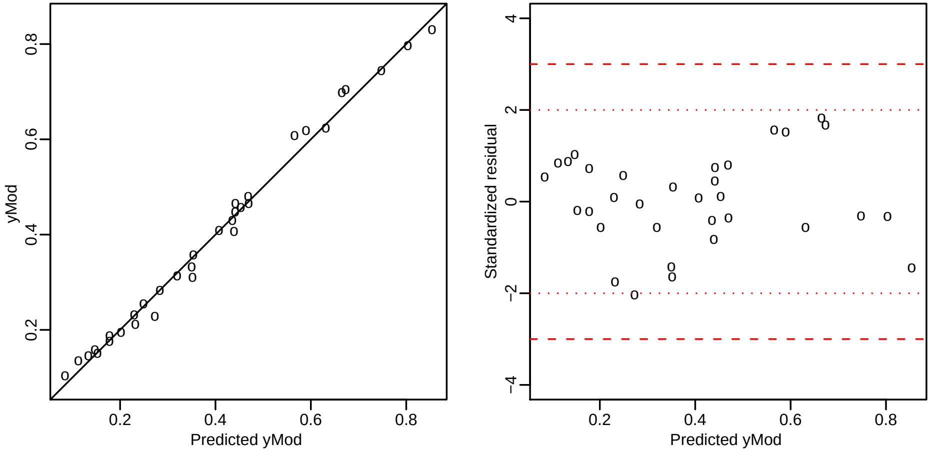
- The standardized cross-validated residual for run i is

$$\frac{y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)})}{s_{-i}(\mathbf{x}^{(i)})}.$$

Diagnostic Plots

- Plot the cross-validated residuals to assess the overall magnitude of error.
- Plot the standardized cross-validated residuals to assess the validity of the standard error for individual predictions.

G-Protein Diagnostic Plots



Cross Validation: Numerical Summaries

Magnitude of error

- The cross-validated root mean squared error is

$$CVRMSE \equiv \sqrt{\frac{1}{n}\sum(y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)}))^2} = .020.$$

- Maximum cross-validated residual is .044
- Fairly accurate relative to a range of about 0.7 in y

Standard errors?

- $\frac{y(\mathbf{x}^{(i)}) - \hat{y}_{-i}(\mathbf{x}^{(i)})}{s_{-i}(\mathbf{x}^{(i)})}$ for $i = 1, \ldots, n$ are roughly in $(-2, 2)$
- Standard errors look reliable.

Fast and Slow CV

- When run i is removed, the hyper-parameters should be re-estimated.
- For computational reasons the correlation parameters are often not updated (it is cheap to update the estimates of μ and σ^2), producing a "fast" CV.
- For "slow" CV, do say 10-fold cross-validation, re-estimating all hyper-parameters.
- The agreement between "fast" CVRMSE and "slow" CVRMSE is often good.
- The agreement between "fast" CVRMSE and the RMSE from test points has been good in examples.

Module Summary

- The GP model has to be "tuned" to data so that its properties match those of the computer model.
- Tuning (fitting) the GP by maximum likelihood is computationally feasible for up to about $n = 1000$ runs and $d = 50$ input variables.
- GP model gives an approximation and a measure of accuracy.
- The measure of accuracy (standard error) can be checked for validity by cross validation.

Appendix A: Bayesian Treatment of the Hyper-parameters

- Posterior distribution of the hyper-parameters ("hyper" below), μ, σ^2, $\theta_1, \ldots, \theta_d$, etc., of the GP
  - From Bayes rule, given the data $\mathbf{y}$

$$p(\text{hyper} \mid \mathbf{y}) \propto \pi(\text{hyper}) L(\mathbf{y} \mid \text{hyper}),$$

  - $\pi(\text{hyper})$ is the prior for hyper
  - $L(\mathbf{y} \mid \text{hyper})$ is the multivariate normal likelihood.
- Predictive distribution for $y(\mathbf{x})$ at a "new" $\mathbf{x}$
  - $p(y(\mathbf{x}) \mid \mathbf{y}) = \int p(y(\mathbf{x}) \mid \mathbf{y}, \text{hyper}) p(\text{hyper} \mid \mathbf{y})\, d\text{hyper}$
  - Usually, the integration is not carried out explicitly.
  - Rather, properties such as the posterior predictive mean and variance of $p(y(\mathbf{x}_0) \mid \mathbf{y})$ are obtained by MCMC sampling of the posterior distribution for the hyper-parameters, $p(\text{hyper} \mid \mathbf{y})$.