Response to the discussion of “Gaussian Process Regression for Multivariate Spectroscopic Calibration”

Tao Chen, Julian Morris and Elaine Martin*

*School of Chemical Engineering and Advanced Materials,
Newcastle University, Newcastle upon Tyne, NE1 7RU, U.K.*

* E-mail: e.b.martin@ncl.ac.uk; Tel.: +44 191 222 6231; Fax: +44 191 222 5748.

We are grateful to the discussants for their comments which highlight some interesting points, and raise issues to be addressed in future work. Our response to the discussion is given below.

1. Response to Dr. Naijun Sha’s comments

Dr. Sha [1] has proposed an extension to our work which addresses the problem of the development of a Gaussian process regression model for the calibration of multiple response variables. A covariance matrix function was formulated for the problem of multi-response calibration. The approach proposed is analogous to that described in our paper. Prior to receiving the comments from Dr. Sha, we had previously derived a similar form of covariance function for multi-response calibration. The details are discussed below.

Let $\mathbf{y}_i$, a column vector, be the q -dimensional response with sample index i , and $\mathbf{x}_i$ be the corresponding covariate vector. The multi-response linear regression model is:

$$\mathbf{y}_i = \mathbf{K}_i \mathbf{b} + \boldsymbol{\varepsilon}_i \quad (1)$$

where $\mathbf{K}_i$ is the Kronecker product of the identity matrix, $\mathbf{I}(q \times q)$, and $\mathbf{x}_i^T$: $\mathbf{K}_i = \mathbf{I} \otimes \mathbf{x}_i^T$, and $\mathbf{b}$ is the concatenation of the regression vectors for the q response variables: $\mathbf{b} = [\mathbf{b}_1^T, \dots, \mathbf{b}_q^T]^T$. In a Bayesian framework, the prior for the q -dimensional vector, $\boldsymbol{\varepsilon}_i$, is given by a zero mean Gaussian distribution: $\boldsymbol{\varepsilon}_i \sim G(\mathbf{0}, \boldsymbol{\Sigma})$, where the full matrix $\boldsymbol{\Sigma}$ reflects the correlation between the multiple responses. The prior for the regression vectors is defined as follows:

$$p(\mathbf{b}_g) = G(\mathbf{0}, \sigma_b^2 \mathbf{I}), \quad g = 1, \dots, q \quad (2)$$

The prior distributions for $\boldsymbol{\varepsilon}_i$ and $\mathbf{b}_g$ can be regarded as a simplified form of those presented in Sha [1]. More specifically, the correlation between $\boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_j$, $i \neq j$, and between $\mathbf{b}_g$ and $\mathbf{b}_h$, $g \neq h$, is ignored. To consider this correlation issue, the introduction of matrix variates, and corresponding matrix normal distributions is necessary [1]. Following the approach described in Boyle & Freaun [2], the response variables are organized in a long vector as follows:

$$\mathbf{y} = [y_{11}, \dots, y_{1n}, \dots, y_{q1}, \dots, y_{qn}]^T \quad (3)$$

where n is the number of training samples. The Gaussian process covariance function between y_{gi} and y_{hj} can be obtained as follows:

$$\begin{aligned} C_{ij}^{gh} &= \text{Cov}(y_{gi}, y_{hj}) = E(y_{gi} y_{hj}^T) = E(\mathbf{x}_i^T \mathbf{b}_g \mathbf{b}_h^T \mathbf{x}_j + \boldsymbol{\varepsilon}_{gi} \boldsymbol{\varepsilon}_{hj}^T) = \mathbf{x}_i^T E(\mathbf{b}_g \mathbf{b}_h^T) \mathbf{x}_j + E(\boldsymbol{\varepsilon}_{gi} \boldsymbol{\varepsilon}_{hj}^T) \\ &= \sigma_b^2 \mathbf{x}_i^T \mathbf{x}_j \boldsymbol{\delta}_{gh} + \boldsymbol{\Sigma}_{gh} \boldsymbol{\delta}_{ij} \end{aligned} \quad (4)$$

where $\boldsymbol{\delta}_{ij} = 1$ if $i=j$, otherwise $\boldsymbol{\delta}_{ij} = 0$, and $\boldsymbol{\Sigma}_{gh}$ is the g -th row and h -th column of $\boldsymbol{\Sigma}$. Equation (4) implies that the covariance function comprises two parts, that part which is required to model the correlation between two data points, and that part that reflects the correlation between two response variables.

Furthermore, the covariance function can be extended to more complex forms as proposed in our paper:

$$C_{ij}^{gh} = \left[a_0 + a_1 \sum_{d=1}^p x_{id} x_{jd} + v_0 \exp \left(- \sum_{d=1}^p w_d (x_{id} - x_{jd})^2 \right) \right] \boldsymbol{\delta}_{gh} + \boldsymbol{\Sigma}_{gh} \boldsymbol{\delta}_{ij} \quad (5)$$

Hence the covariance matrix for $\mathbf{y}$ can be written as:

$$\mathbf{C} = \mathbf{I}_q \otimes \mathbf{Q} + \mathbf{\Sigma} \otimes \mathbf{I}_n \quad (6)$$

where the i -th row and j -th column of $\mathbf{Q}$ is given by:

$$Q_{ij} = a_0 + a_1 \sum_{d=1}^p x_{id} x_{jd} + v_0 \exp \left(- \sum_{d=1}^p w_d (x_{id} - x_{jd})^2 \right) \quad (7)$$

The covariance matrix in Eq. (6) is a simplified form of that given in Sha [1], but it still accounts for the effect of the multiple responses through the introduction of the hyper-parameter, $\mathbf{\Sigma}$.

However, the implementation of the multi-response Gaussian process using the covariance matrix defined in Eq. (6), or the more general form given in Sha [1], is not straightforward numerically. More specifically, it has been observed in our simulation studies that it is difficult to ensure a positive definite covariance matrix. This is partly due to the increase in the size of the covariance matrix. However it is hypothesized that there is a more serious reason. More specifically it is considered to be a consequence of the increase in the complexity of the covariance matrix, as a result of introducing additional hyper-parameters that account for the correlation between the multiple responses. It is not clear how this issue can be addressed. In the implementation using hybrid Monte Carlo [3], one simplified approach that we have considered has been the rejection of those Monte Carlo samples that result in a non-positive definite covariance matrix. This approach is equivalent to assigning a zero prior probability to these samples. However from the results of the simulations, it was observed that this approach dramatically reduces the acceptance rate of the hybrid Monte Carlo algorithm, and thus this approach is not efficient. Therefore alternative solutions require to be investigated to address this numerical issue.

2. Response to Prof. Philip J. Brown's comments

Professor Brown [4] has raised a number of important issues which are addressed under the following headings: 1) model interpretability, 2) covariance function and the prior assumptions, 3) computational cost, 4) hierarchical structure and 5) variable selection strategy. The last point has also been raised by Dr. Sha [1].

Model interpretability. Model interpretability, i.e. how the transmittance at different wavelengths are weighted in the Gaussian process model, is possible through the approach of “automatic relevance determination” (ARD) [3][5]. The Gaussian process with the covariance function as defined in Eq. (10) of our paper falls within the family of ARD models, therefore each predictor is associated with a hyper-parameter, w_d , thereby enabling the determination of the relevance of the corresponding variable in the predictive model. More specifically the w_d ’s serve as the “weights” for the predictor variables at different wavelengths.

Covariance function and the prior assumptions. The rationale for specifying the covariance function, and the corresponding prior distributions, has previously been discussed in Rasmussen’s thesis [5]. More specifically, the hyper-parameter, w_d , determines the relative relevance of each predictor for the prediction as discussed in the response to the model interpretability issue. Consequently the scaling of the mean for the inverse Gamma prior for w_d reflects that *a priori* the w_d ’s should decrease when p increases, thereby materializing in a small number of relevant predictors when p is large. Empirically, the application of the chosen covariance function and this specific prior been reported in a number of applications, for example Shi et al. [6]. Thus we would conclude that these choices are not specific to the present calibration problem, but are applicable to a wide variety of problems.

Computational cost. We have mentioned in our paper that the computational time scales cubically with sample size. For the two case studies presented in the paper, the CPU time for the training stage, where 1000 MCMC iterations were required, was 27.3 seconds and 558.5 seconds for the “Tablet” (60 training data points) and “Meat” (173 training data points) data sets, respectively. The Gaussian process was implemented using C++, and the program was executed on a Pentium-4 3.0 GHz computer running under Windows XP. For applications with more than 1000 training data points (a rare situation in spectroscopic calibration), sparse training strategies may be required to reduce the overall computational burden (see the references cited in our paper).

Hierarchical structure. To model observations that exhibit cluster structures, a mixture of Gaussian process models can be implemented [6][7]. However, there has been limited research undertaken into the block effects of covariates (e.g. one set of near infrared spectra plus one set of Raman spectra) in the context of Gaussian processes. One possible approach is to assign separate prior distributions to the ARD parameters of each block, and a hyper-prior across the blocks. This

Bayesian hierarchical approach appears attractive, but its effectiveness and applicability needs to be assessed through further case studies.

Variable selection strategy. We agree with Dr. Sha and Prof. Brown that variable selection is a powerful approach in regression problem, with regard to improving predictive performance, reducing future measurement costs and alleviating computational burden. We are currently implementing the Markov chain Monte Carlo approach with variable selection within a Gaussian process regression model.

Acknowledgments

We appreciate Professor Cliff Spiegelman's effort to make this discussion happen. T. Chen would like to acknowledge the financial support from the EPSRC KNOW-HOW (GR/R19366/01) and Chemicals Behaving Badly II (GR/R43853/01), and the UK ORS Award for his PhD study.

References

- [1] N. Sha. Discussion of "Gaussian process regression for multivariate spectroscopic calibration", Chemometrics and Intelligent Laboratory Systems, in press, 2006.
- [2] P. K. Boyle, M. R. Frea. Dependent Gaussian processes. In: Advances in Neural Information Processing Systems 17, MIT Press, 2005.
- [3] R. M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702, Department of Statistics, University of Toronto, Canada, 1997.
- [4] P. J. Brown. Discussion of the paper by Chen, Morris and Martin, Chemometrics and Intelligent Laboratory Systems, in press, 2006.
- [5] C. E. Rasmussen. Evaluation of Gaussian processes and other methods for non-linear regression. Ph. D. thesis, University of Toronto, Canada, 1996.
- [6] J. Q. Shi, R. Murray-Smith, D. M. Titterton. Hierarchical Gaussian process mixtures for regression, Statistics and Computing 15 (2005) 31-41.
- [7] C. E. Rasmussen, Z. Ghahramani. Infinite mixtures of Gaussian process experts. In: Advances in Neural Information Processing Systems 14, MIT Press, 2002.