8th International IFAC Symposium on
Dynamics and Control of Process Systems
Preprints Vol.3, June 6-8, 2007, Cancún, Mexico

CALIBRATION OF SPECTROSCOPIC SENSORS WITH GAUSSIAN PROCESS AND VARIABLE SELECTION

Tao Chen, Xiaoling Ou and Elaine Martin

School of Chemical Engineering and Advanced Materials,
University of Newcastle, Newcastle upon Tyne, NE1 7RU, UK

Email: {tao.chen, xiaoling.ou, e.b.martin}@ncl.ac.uk

Abstract: Multivariate spectroscopic calibration models can be improved by selecting a subset of spectral variables that are identified as being the most informative in terms of inferring sample properties. This paper proposes the application of Gaussian processes for both variable selection and the development of calibration models. A Gaussian process is a Bayesian regression technique that assigns Gaussian priors over the regression functions. The covariance function of the Gaussian process is characterized by a number of hyper-parameters and by associating each spectral variable with a hyper-parameter, the relevance of the corresponding variable to the prediction can be automatically determined. Prior to the training of a Gaussian process using a Markov chain Monte Carlo approach, a pre-processing step is proposed based on a statistical significance test to reduce the computational time materialising from the large number of variables present within spectroscopic data. The methodology presented is applied to two sets of near infrared spectral data, and enhanced prediction performance is achieved when both the pre-processing step and a Gaussian process are implemented.

Keywords: Bayesian inference, Gaussian process, multivariate spectroscopic calibration, variable selection.

1. INTRODUCTION

Multivariate calibration techniques have been widely applied to infer sample properties through the extraction of relevant information from data generated from spectroscopic instruments. A calibration model is constructed from the available training data, using multivariate regression techniques, such as partial least squares (PLS) or principal component regression (PCR). The input variables for the regression model can be of high dimension, e.g. from several hundred to over one thousand wavelengths. Although PLS has been shown to be efficient when all the available wavelengths are included, both theoretical and experimental evidence exists that demonstrate that it is possible to enhance prediction performance through the implementation of variable selection (Bangalore *et al.*, 1996; McShane *et al.*, 1997). The assumption is that there may be parts of the spectra that contain little information about the chemical properties. When these spectra are included in the regression model, they may by chance appear, in a finite training set, to be closely associated with the chemical properties, especially when a large number of variables are considered. Predictive performance on unseen test cases will then be poor (Neal, 1996). Another issue, al-

though less well documented in the literature, is that there may exist information "redundancy" among wavelengths, i.e. similar relevant information for the inference of sample properties can be provided by a number of wavelengths. Therefore by selecting a subset of these wavelengths, efficiency in terms of measurement costs may be realised.

Traditionally relevant variables (wavelengths) have been selected by using a fundamental understanding of the spectroscopic properties of the test samples. More recently, researchers have focussed on automatic variable selection strategies. Most of these strategies are based on a PLS or MLR regression model whilst optimizing predictive performance by selecting/removing spectral variables. For example, iterative PLS (Osborne *et al.*, 1997) starts with the random selection of a small number of variables, with variables being added or removed based on the cross validation error. An alternative approach is that of uninformative variable elimination. This method of variable selection is based on an analysis of the PLS regression coefficients (Centner *et al.*, 1996).

A third method widely reported in the literature is that of genetic algorithms (GAs). Genetic algorithms were originally proposed as a family of stochastic optimization approaches that mimic the principles of genetics and natural selection. They have been successfully applied in spectroscopic applications for the selection of wavelengths (Bangalore *et al.*, 1996; Broadhurst *et al.*, 1997). A comparative study of a number of variable selection algorithms was reported (Abrahamsson *et al.*, 2003), and it was shown that the GA approach demonstrated improved prediction ability over conventional PLS and manual selection approaches.

In this paper a Gaussian process is applied for variable selection with the aim being that of the calibration of spectral data (Chen *et al.*, 2007). A Gaussian process is a non-parametric Bayesian regression model formulated from Gaussian prior distributions over the space of all possible regression functions (Neal, 1996). The use of prior distributions automatically addresses the issue of over-fitting, since Bayesian inference avoids the need for a separate validation data set or cross validation strategy. The behaviour of a Gaussian process is determined through the covariance function in terms of the underlying hyper-parameters. By associating each input variable with a hyper-parameter, the relevance of a variable to the prediction can be automatically determined from the training data. Consequently the relevant variables are selected according to the magnitude of their hyper-parameters. The training of a Gaussian process is performed through the Markov chain Monte Carlo (MCMC) sampling of the posterior probability of the hyper-parameters (Neal, 1997). Predictions can then be achieved by averaging over the Monte Carlo samples.

The rest of the paper is organised as follows. Section 2 introduces Gaussian process regression and its automatic relevance determination (ARD) property, prior to describing how Bayesian inference can be used to estimate the model parameters through MCMC simulations. One consequence of the large number of available spectral variables is that the Gaussian process may require substantial computational time for the training procedure. Thus a statistical significance test is proposed in Section 3 as a pre-processing step to select variables based on a correlation criterion. In Section 4, a Gaussian process is applied for variable selection prior to the development of a calibration model for two sets of near infrared spectral data. Compared with the widely used genetic algorithms, the Gaussian process can improve prediction performance through the selection of fewer spectral variables. Finally, conclusions and future work are given in Section 5.

2. GAUSSIAN PROCESSES

Gaussian processes emerged from the area of neural networks. In the Bayesian approach to neural networks, a prior distribution over the network weights materialises in a prior distribution over the functions. As discussed in Neal (1996), there is no rationale for restricting neural network models to a small number of hidden units to avoid the so-called over-fitting problem. As the number of hidden neurons tends to infinity, neural network models with one hidden layer converge to Gaussian processes, if standard "weight decay" priors are assumed (Neal, 1996; MacKay, 1998). A brief introduction to a Gaussian process based on regression analysis is given in the next section prior to describing the automatic relevance determination technique, which is the key to variable selection. Finally the training of Gaussian processes is briefly discussed.

2.1 Regression with Gaussian Process

From the perspective of a regression problem, a functional relationship is identified between the input variables, $\mathbf{x}$, and the output variable, y. Consider a training data set consisting of N data points, $\{\mathbf{x}_i, y_i, i = 1, \cdots, N\}$. A Gaussian process for regression is defined such that $y(\mathbf{x})$ has a Gaussian prior distribution with zero mean (assuming data are standardized to zero mean and unit standard deviation) and covariance function $C(\mathbf{x}_i, \mathbf{x}_j)$. An example of such a covariance function is:

$$C(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = v_0 \exp\left(-0.5 \sum_{d=1}^{D} w_d (x_{id} - x_{jd})^2\right) + a_0 + \delta_{ij}\sigma_v^2 \quad (1)$$

where $\boldsymbol{\theta} = [w_1, \cdots, w_D, v_0, a_0, \sigma_v^2]^T$, and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. $\boldsymbol{\theta}$ is a vector of hyper-parameters. The first term defines the correlation between the outputs with nearby inputs, and the remaining two terms define the bias and noise, respectively. This covariance function has been widely used for predicting stationary signals and is thus adopted in this paper. Other forms of covariance function are discussed in (Neal, 1997; MacKay, 1998). The predictive distribution of the output variable y^*, given its input $\mathbf{x}^*$, is Gaussian with mean and variance given by:

$$y^* = \mathbf{k}^T(\mathbf{x})\Sigma^{-1}\mathbf{y} \quad (2)$$

$$\sigma^{*2} = C(\mathbf{x}, \mathbf{x}) - \mathbf{k}^T(\mathbf{x})\Sigma^{-1}\mathbf{k}(\mathbf{x}) \quad (3)$$

where $\mathbf{k}(\mathbf{x}) = [C(\mathbf{x}^*, \mathbf{x}_1), \cdots, C(\mathbf{x}^*, \mathbf{x}_N)]^T$, Σ is the covariance matrix for the training cases: $\Sigma_{ij} = C(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{y} = [y_1, \cdots, y_N]^T$. The matrix inversion step in Eqs. (2) and (3) is computationally intensive, i.e. of the order $O(N^3)$. This is feasible for a moderate sized training data set (less than several thousand) on conventional computers. For larger problems sparse training strategies, such as projection techniques (Csató and Opper, 2002), are required.

2.2 Automatic Relevance Determination

An important feature of a Gaussian process is that it falls within the family of automatic relevance determination (ARD) models. The idea of ARD models was developed by MacKay (MacKay, 1998) and Neal (Neal, 1996). More recently the ARD approach has been used to prune irrelevant basis functions, resulting in the "relevance vector machine" (Tipping, 2001). Likewise for the calibration of spectral data, there is normally a large number of wavelengths which potentially contain information for predicting the material properties. The ARD approach assesses which measurements are more likely to be relevant to the prediction. In the Bayesian framework, ARD is implemented by associating each input variable with a hyper-parameter, thereby determining the magnitude of the relevance of the corresponding variable to the prediction. For example, if the covariance function in Eq. (1) is used, $\{w_1, \cdots, w_D\}$ serves as the relevant parameters. Fig. 1 illustrates how the hyper-parameters determine the relevance of specific variables. Here a simple covariance function with one input variable is considered:

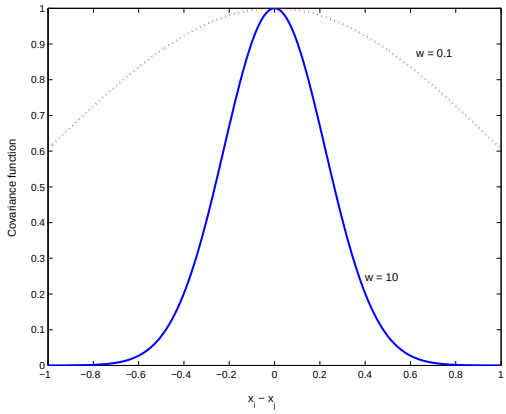

Fig. 1. Automatic relevance determination.

$$C(x_i, x_j) = \exp\left(-0.5w(x_i - x_j)^2\right) \quad (4)$$

According to Fig. 1, it is evident that for a large hyper-parameter w, the variation in the input data, $|x_i - x_j|$, will lead to large variation in the covariance function, indicating that this variable significantly affects the prediction. However, for smaller values of w, the covariance function is insensitive to the variation in the input data, implying lower relevance. For variable selection, a threshold is therefore needed to remove those variables with hyper-parameters close to zero. A value of 10^{-6} is adopted as the threshold since it has been observed to retain a small number of variables while giving good prediction performance in initial experiments,.

2.3 Training a Gaussian Process

Given a covariance function, the log likelihood of the training data is:

$$L = -\frac{1}{2}\log\det\Sigma - \frac{1}{2}\mathbf{y}^T\Sigma^{-1}\mathbf{y} - \frac{N}{2}\log 2\pi \quad (5)$$

Most training algorithms also require the derivative of L with respect to each hyper-parameter θ:

$$\frac{\partial L}{\partial \theta} = -\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta}\right) + \frac{1}{2}\mathbf{y}^T\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta}\Sigma^{-1}\mathbf{y} \quad (6)$$

The hyper-parameters can be obtained by maximizing the likelihood of the training data, using the conjugate gradient method. However this approach is sensitive to initializations and may converge to local minima (MacKay, 1998). Therefore a number of random initializations are needed to guarantee reliable results. A more elaborate approach is that of Bayesian inference. According to a Bayesian framework, a prior distribution $p(\boldsymbol{\theta})$ is defined over the hyper-parameters. With the available training data, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x}, y)$ is calculated by taking the product of the prior and the likelihood. Predictions are then made by integrating over the posterior:

$$y(\mathbf{x}^*) = \int p(y|\boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta}|\mathbf{x}, y) d\boldsymbol{\theta} \qquad (7)$$

Since it is not feasible to perform the integration analytically, the Markov chain Monte Carlo (MCMC) method is applied. However, due to the high dimensionality of the hyper-parameters, the random walk search strategy in conventional MCMC may converge very slowly. The hybrid MCMC approach can significantly increase the convergence speed using the gradient information. In the present study, the MCMC method described by Neal (1997) is implemented. After generating a sufficient number of samples for the hyper-parameters, the predictions can be made by taking an average of Eqs. (2)(3) over these samples.

3. SIGNIFICANCE TEST

The strategy of ARD is to assign one hyper-parameter to each input variable. When applying ARD to a data set with a large number of input variables, such as for the calibration of spectral data, Gaussian processes may incur substantial computational costs when attaining Monte Carlo samples from the posterior probability. One possible solution is to assign one hyper-parameter over multiple input variables. Therefore a criterion is required to determine how to cluster multiple input variables to a shared window, and possibly split the window at a later stage.

In this study, a simple statistical significance test is adopted as a "pre-processing" step to select an initial subset of variables (Neal and Zhang, 2003). The correlation between each input and output variable is used, where a low correlation coefficient implies that the corresponding variable is not informative for the prediction. The significance test starts by constructing a hypothesis test:

- H_0: No correlation exists between the input and output variables.
- H_1: Correlation exists between the input and output variables.

A significance level, α, is then chosen as a threshold to accept or reject the null hypothesis. Finally the p-value, which is the probability that the null hypothesis is true given the available data, can be calculated and compared with the significance level. If $p < \alpha$, the null hypothesis is rejected, and the corresponding variable is selected. α is normally selected to be 0.05. The significance test of correlation can be carried out in many software packages, including the statistical toolbox in Matlab, and S-Plus.

In summary, the proposed variable selection strategy is as follows:

Table 1. Summary of spectral data sets.

Property	Wavelengths	Training	Test
Viscosity	401	136	116
Fat	100	210	30

(1) Perform a significance test to pre-select potentially informative wavelengths.
(2) Based on the retained wavelengths from the significance test, a Gaussian process model is trained and the ARD strategy further selects a subset of the retained wavelengths.

4. EXAMPLES

4.1 Spectral Data Case Studies

Two sets of spectral data are considered in this study. The first relates to near infrared spectra of diesel fuels measured at Southwest Research Institute on a project sponsored by the US Army. This data set has been published by Eigenvector Research, Inc. (`http://software.eigenvector.com/Data/SWRI/index.html`) as a benchmark for variable selection and calibration. Of the six data sets relating to different properties, the data set relating to viscosity is used. Based on the recommendations by the authors, the high leverage samples and a random set are used to train the calibration model. A second random set is used for testing. According to the pre-analysis, the relationship between the spectral data and associated property is approximately linear.

The second data set was recorded using a Tecator near infrared spectrometer which measured the spectrum of light transmitted through a sample of pork meat (Borggaard and Thodberg, 1992). The spectrum consists of the absorbencies at 100 wavelengths in the region 850-1050 nm. The spectrometer is calibrated to determine the fat content from the spectrum. In previous studies (Borggaard and Thodberg, 1992; Thodberg, 1996), this data set was shown to be highly non-linear. The training and testing sets are randomly partitioned 100 times to evaluate the robustness of the proposed algorithm. The two data sets are summarized in Table 1. All the data are scaled to zero mean and unit standard deviation before the development of the calibration model.

4.2 Results

For comparison with the proposed variable selection strategy, the results for PLS models based on the full spectra and also following variable selection using GAs are reported. The PLS and GA analyses were performed using the PLS Toolbox from Eigenvector Research, Inc.. Ten-fold cross-validation was applied to determine the number of

Table 2. Parameters for GA.

Parameter	Value
Maximum number of generations	200
Chromosome population size	128
Number of splits for cross validation	10
Percentage of duplicates as convergence	80
Window width	1, 5, 10, 20

Table 3. Variable selection and calibration of viscosity. GA(i) represents genetic algorithm (window=i).

Model	Selected Wavelengths	%RMSE
PLS	401	3.96
GA(1) + PLS	73	4.17
GA(5) + PLS	85	3.74
GA(10) + PLS	90	3.75
GA(20) + PLS	100	3.70
GP(ARD)	70	3.24
Sig. Test + GP(ARD)	49	2.87

latent variables to retain in the PLS model. The tuning parameters for the GAs are summarised in Table 2 with a number of different values for the window width being considered. Since there are a significant number of tuning parameters for GAs, it is not feasible to determine them all through cross validation. Therefore the parameters given in Table 2 are taken from the literature and by using "trial and error".

The results for the first data set are given in Table 3. The predictive performance is evaluated using the percentage root mean square error (%RMSE) for the test data sets. For this case study, PLS performs well since the data is approximately linear. However using variable selection with GAs and then applying PLS to the reduced variable set, a slightly lower prediction error is achieved (except for the case where a window size of one was used). On the other hand, the Gaussian process gives further improvement in terms of the prediction error. Seventy relevant variables are selected using the ARD strategy. If the significance test preprocessing procedure is applied, 201 wavelengths are retained. Then, applying the Gaussian process, the number of wavelengths is reduced to 49 and the smallest %RMSE is achieved.

Fig. 2 illustrates the variables selected using GAs and the Gaussian process, following the application of the pre-processing significance test. Although the two approaches select quite different wavelengths, both achieve acceptable results, indicating that there exists information redundancy, along with uninformative variables. In addition, due to the effect of a fixed window, the GA tends to select a number of variables within a region. To address this effect, a window size of one can be used to eliminate this effect. Although this approach can select fewer variables, it may degrade the prediction performance, as reported for this data set.

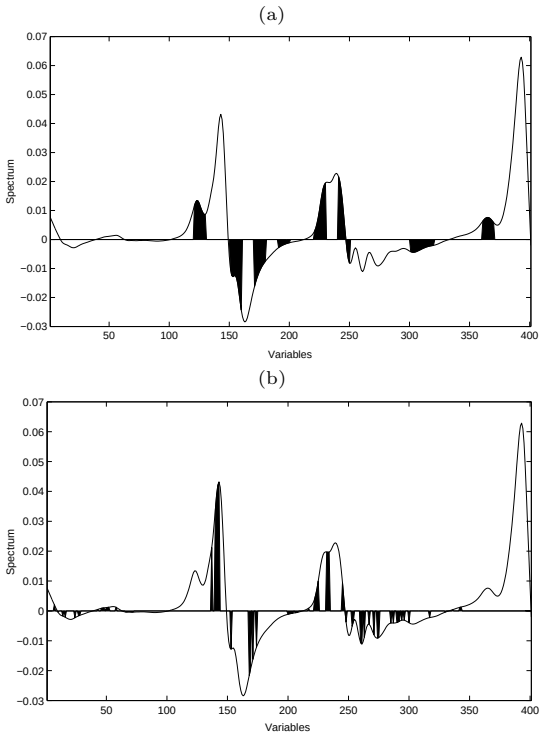

Fig. 2. Variable selection for viscosity. (a): GA (window = 10); (b): GP with significance test.

Table 4. Variable selection and calibration of fat in minced pork meat. Results are averaged over 100 random partitions of training and test data sets.

Model	Selected Wavelengths	RMSE
PLS	100	2.38
NN	100	0.72
GP	14.9	0.51

The second data set is non-linear, therefore linear calibration algorithms, such as PLS, are not expected to perform well. This is confirmed in Table 4. Therefore a neural network (NN) comprising one hidden layer with 15 neurons was trained using the Levenberg-Marquardt algorithm and early stopping, giving a RMSE of 0.72. It was suggested in (Thodberg, 1996) that the data was pre-processed using principal component analysis (PCA). However PCA does not reduce the number of wavelengths and thus the comparison with variable selection techniques is not appropriate.

An interesting finding was that the significance test showed that all the input variables are correlated to the output, with correlation coefficients lying in the range [0.4, 0.6]. This may be a good example of information redundancy, as all input variables are related to the output. For this data set, the Gaussian process is capable of selecting 14.9 wavelengths on average to achieve a RMSE of 0.51, for the 100 random experiments.

5. CONCLUSIONS AND DISCUSSIONS

This paper proposes using the Gaussian process for variable selection and the calibration of spectral data. Based on automatic relevance determination, variable selection and calibration are sequentially realised. In addition, as a result of Bayesian inference, a validation data set, which is essential for the tuning of PLS and genetic algorithm, is not required. One limitation of the proposed strategy is the difficulty of sampling the large number of relevance parameters in a Gaussian process, in terms of both computational cost and convergence rate of the MCMC. This issue can be partially alleviated by the implementation of significance test, which is introduced to eliminate apparently irrelevant variables prior to the application of the Gaussian process regression.

A general conclusion of the research undertaken is that for data with uninformative input variables, the significance test can remove apparently irrelevant variables, and the Gaussian process can then be applied to sequentially select the most relevant variables for building a calibration model. On the other hand, if high redundancy occurs in the data set, the Gaussian process is still capable of removing redundant variables, and providing satisfactory calibration performance.

The present research is focused on predicting a single chemical property. When extended to multiple output variables, the proposed approach may be applied separately to each output variable, and the significance test can consider the union of the relevant variables for each output. However a more appropriate way would be to also consider the relationship between the output variables. Ongoing work is focused on variable selection and calibration of multiple output variables by modelling the output covariance structure.

ACKNOWLEDGMENT

This work was supported by UK EPSRC grants KNOW-HOW (GR/R19366/01) and VERTIGO (GR/R64407/01).

REFERENCES

Abrahamsson, C., J. Johansson, A. Sparén and F. Lindgren (2003). Comparison of different variable selection methods conducted on NIR transmission measurements on intact tables. *Chemometrics and Intelligent Laboratory Systems* **69**, 3–12.

Bangalore, A.S., R.E. Shaffer, G.W. Small and M.A. Arnold (1996). Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: application to near-infrared spectroscopy. *Analytical Chemistry* **68**, 4200–4212.

Borggaard, C. and H. H. Thodberg (1992). Optimal minimal neural interpretation of spectra. *Analytical Chemistry* **64**, 545–551.

Broadhurst, D., R. Goodacre, A. Jones, J. J. Rowland and D. B. Kelp (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta* **348**, 71–86.

Centner, V., D.-L. Massart, O.E. deNoord, S. deJong, B. M. Vandeginste and C. Sterna (1996). Elimination of uninformative variables for multivariate calibration. *Anlytical Chemistry* **68**, 3851–3858.

Chen, T., J. Morris and E. Martin (2007). Gaussian process regression for multivariate spectroscopic calibration. *Chemometrics and Intelligent Laboratory Systems*. in press.

Csató, L. and M. Opper (2002). Sparse online Gaussian processes. *Neural Computation* **14**, 641–668.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. In: *Neural Networks and Machine Learning, volume 168 of F: Computer and Systems Sciences* (C. M. Bishop, Ed.). pp. 133–165. NATO Advanced Study Institute, Springer. Berlin, Heidelberg.

McShane, M.J., G.L. Cote and C.H. Spiegelman (1997). Variable selection in multivariate calibration of a spectroscopic glucose sensor. *Applied Spectroscopy* **51**, 1559–1564.

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer-Verlag. New York.

Neal, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702. Department of Statistics, University of Toronto. Canada. Available at `http://www.cs.utoronto.ca/~radford/ftp/mc-gp.pdf`.

Neal, R. M. and J. Zhang (2003). Classification for high dimensional problems using bayesian neural networks and dirichlet diffusion trees. In: *NIPS'2003 Feature Selection Workshop*. Whistler, British Columbia.

Osborne, S.D., R.B. Jordan and R. Kunnemeyer (1997). Method of wavelength selection for partial least squares. *Analyst* **122**, 1531–1537.

Thodberg, H. H. (1996). A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Networks* **7**, 56–72.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**, 211–244.