Bayesian variable selection for Gaussian process regression: application to chemometric calibration of spectrometers

Tao Chen*,[a], Bo Wang[b]

[a]School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore 637459
[b]Department of Mathematics, University of York, York, YO10 5DD, UK

Abstract

Gaussian processes have received significant interest for statistical data analysis as a result of the good predictive performance and attractive analytical properties. When developing a Gaussian process regression model with a large number of covariates, the selection of the most informative variables is desired in terms of improved interpretability and prediction accuracy. This paper proposes a Bayesian method, implemented through the Markov chain Monte Carlo sampling, for variable selection. The methodology presented is applied to the chemometric calibration of near infrared spectrometers, and enhanced predictive performance and model interpretation are achieved when compared with benchmark regression method of partial least squares.

Key words: Gaussian process regression, Markov chain Monte Carlo, model selection, multivariate calibration, variable selection

1. Introduction

Regression techniques are important data analysis tools in many scientific and engineering disciplines where empirical models are utilized to establish the relationship between two sets of variables. This paper is mainly concerned with the specific area of *chemometrics*, within which a major task is to calibrate various spectrometers such that the analyte properties (e.g. concentration) can be indirectly inferred from the measured spectra at hundreds or thousands of wavenumbers (or wavelengths) [1]. Fig. 1 gives the near infrared spectra of 120 samples of pharmaceutical tablets, where the objective is to predict the concentration of active substance from the absorbance recorded at 404 wavenumbers (i.e. 404 covariates). In this context, the conventional multiple linear regression is not applicable because of the high correlation between covariates, and the most widely used methods are principal component regression (PCR), partial least squares (PLS) and ridge regression [2, 3]. The wide acceptance of these linear methods is largely due to the inherent linearity between the analyte properties and spectral absorbance as stated in the *Beer-Lambert's law* [4]. However non-linear calibration methods have also been proposed for practical problems where the linearity does not hold, including neural networks [5], support vector machine and its variants [6, 7].

Recently, there has been significant interest in the Gaussian process regression model. Initially proposed by [8], Gaussian process was viewed as an alternative approach to neural networks since it was demonstrated that a large class of Bayesian regression models, based on neural networks, converged to a Gaussian process in the limit of an infinite network [9]. Gaussian processes can also be derived from the perspective of non-parametric Bayesian regression [10], by directly placing Gaussian prior distributions over the space of

*Corresponding author, Email: chentao@ntu.edu.sg; Tel.: +65 6513 8267; Fax: +65 6794 7553.
Email addresses: chentao@ntu.edu.sg (Tao Chen), bw527@york.ac.uk (Bo Wang)

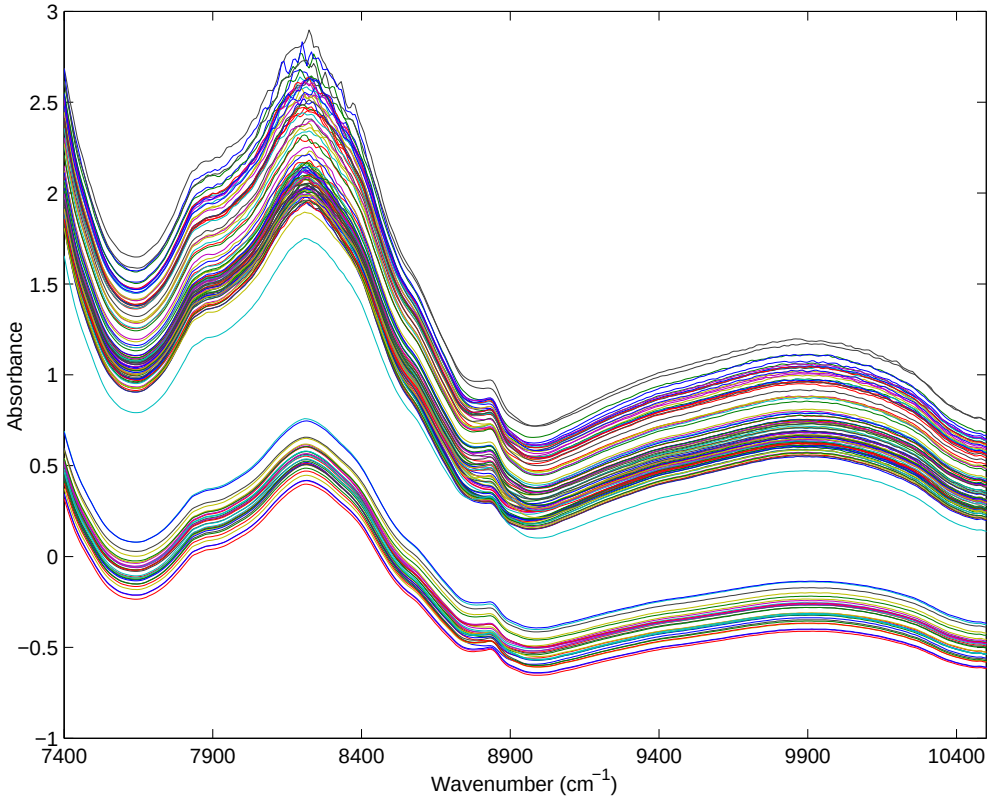

Figure 1: Near infrared spectra of 120 tablet samples.

regression functions. Empirical comparative studies have confirmed the outstanding performance of Gaussian process regression with respect to other non-linear models [11, 12, 13]. As a result, Gaussian process models have been widely applied to various problems in statistics and engineering [14, 11, 15, 16, 17, 18, 13].

In developing a Gaussian process model for regression on a large number of covariates, it is often desirable to reduce the dimension of the variables to alleviate the computational burden and to improve the prediction accuracy. This is typically realized by either projecting the original covariates onto lower-dimensional space, for example using principal component analysis (PCA) [11], or selecting a subset of these covariates. Compared with the projection techniques, variable selection attains the additional advantage of improved interpretability as to which covariates are the most informative to prediction. Therefore, the task of variable selection with respect to Gaussian process regression is considered in this paper. This is a special case of model selection problems, since a certain set of selected variables corresponds to a particular model.

In the field of chemometric calibration modelling, variable selection strategies are typically based on a PLS regression model whilst optimizing predictive performance by selecting/removing the covariates. For example, iterative PLS [19] 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 uninformative variable elimination based on an analysis of the PLS regression coefficients [20]. A third method widely reported in the literature is that of genetic algorithms (GAs) that were originally proposed as a family of stochastic optimization approaches that mimic the principles of genetics and natural selection [21]. GAs have been successfully applied for variable selection in spectroscopic calibration [22, 23].

More recently Bayesian approach to variable selection has received considerable attention, where the

main focus has been on linear regression using Markov chain Monte Carlo (MCMC) methods. Initially the Gibbs sampling algorithm was proposed to sample the posterior of the indicators [24, 25]. This was followed by its extension to multiple responses [26], and its implementation using the Metropolis-Hastings sampling algorithm [27]. [28] reviewed and compared several MCMC methods for variable selection. Basis function based regression models, for example splines [29, 30, 31] and wavelets [32], were proposed with the selection of the basis function being based on similar methodologies to those of variable selection.

The challenge with variable selection in a Gaussian process is that the posterior distribution of the "hyper-parameters" (to be defined subsequently) is not analytically available and needs to be sampled. As a consequence the tasks of model selection and hyper-parameter estimation are coupled, and this issue is addressed in this paper by sampling the model and hyper-parameters alternately. The rest of the paper is organized as follows. A brief description of Gaussian process regression is provided in Section 2. An overview of the Bayesian model selection techniques is described in Section 3, where the discussion focuses on MCMC based approaches. Section 4 presents the MCMC method for variable selection, and the proposed method is demonstrated through application study in Section 5. Finally conclusions are drawn in Section 6.

2. Gaussian process regression

Consider the case where a training set of N observations is available, $\{\mathbf{x}_i, y_i; i = 1, \ldots, N\} = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{x}_i$ is a vector of p covariates, and y_i is the scalar response. A Gaussian process prior for regression is then defined in such a way that the regression function has a Gaussian process prior with zero mean and covariance function $C(y_i, y_j) = C(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$. The covariance is a function of covariates given the hyper-parameters $\boldsymbol{\theta}$. An example of such a covariance function is:

$$C(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = a_0 + a_1 \sum_{d=1}^{p} x_{id} x_{jd} + v_0 \exp\left(-w \sum_{d=1}^{p} (x_{id} - x_{jd})^2\right) + \delta_{ij} \sigma^2 \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$; $\delta_{ij} = 1$ if $i = j$, otherwise it is equal to zero. We denote $\boldsymbol{\theta} = (a_0, a_1, v_0, w, \sigma^2)$ "hyper-parameters" to differentiate Gaussian processes from parametric regression. In this covariance function, the first two terms represent constant bias and linear correlation respectively. The exponential term is similar to the form of a radial basis function, and defines the correlation between the responses and nearby covariates. Finally σ^2 captures random noise. By combining linear and non-linear terms in the covariance function, the Gaussian process is capable of handling both linear and non-linear data structures [11]. Other forms of the covariance function are discussed in [10, 33].

The training of a Gaussian process, i.e., the inference of $\boldsymbol{\theta}$, can be carried out using maximum likelihood estimation, or using maximum a posterior estimation if prior distributions are given for the hyper-parameters. Since the posterior distribution of $\boldsymbol{\theta}$ is generally of complicated form, it is desirable to use MCMC methods to draw random samples from $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$. Particularly appropriate for this task is the *Hamiltonian Monte Carlo*[1] method [34, 35]. The detailed algorithm for Hamiltonian MC is given in the Appendix.

For a new data point with covariates $\mathbf{x}^*$, the predictive distribution of the response y^*, conditional on the hyper-parameters, is also Gaussian distributed with mean and variance calculated as follows:

$$\hat{y}^* = \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\, \mathbf{C}^{-1} \mathbf{y} \tag{2}$$

$$\sigma^2_{\hat{y}^*} = C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^{\mathrm{T}}(\mathbf{x}^*)\, \mathbf{C}^{-1}\, \mathbf{k}(\mathbf{x}^*) \tag{3}$$

[1] Hamiltonian Monte Carlo is usually termed *hybrid Monte Carlo* in the literature. We adopt the term Hamiltonian MC to avoid confusion with the hybrid MCMC method to be discussed in Section 4. In this paper hybrid MCMC is referred to a combination of the Gibbs, Metropolis-Hastings and Hamiltonian MC algorithms.

where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}^*, \mathbf{x}_1), \ldots, C(\mathbf{x}^*, \mathbf{x}_N)]^{\mathrm{T}}$, and $\mathbf{C}$ is the covariance matrix calculated from all the training data. If Monte Carlo samples have been obtained for the hyper-parameters, the above prediction can be modified by averaging over these samples [11, 35].

3. Bayesian model selection using MCMC

Consider the case of K candidate models, $\{\mathscr{M}_k, k = 1, \ldots, K\}$, each associated with a vector of unknown parameters, $\boldsymbol{\theta}_k$. The task of model selection is to identify which model is more probable given the data $\mathscr{D}$. Within a Bayesian framework, this is equivalent to assessing the following probability for each of the K models [26, 27, 36, 37]:

$$p(\mathscr{M}_k|\mathscr{D}) \propto p(\mathscr{D}|\mathscr{M}_k)p(\mathscr{M}_k) \tag{4}$$

where $p(\mathscr{D}|\mathscr{M}_k)$ is the likelihood, and $p(\mathscr{M}_k)$ is the prior of the k-th model. The likelihood function can be further expanded to include the effect of model parameters:

$$p(\mathscr{D}|\mathscr{M}_k) = \int p(\mathscr{D}|\boldsymbol{\theta}_k, \mathscr{M}_k)p(\boldsymbol{\theta}_k|\mathscr{M}_k)d\boldsymbol{\theta}_k \tag{5}$$

Only for some special cases can the calculation of the likelihood in (5) be achieved analytically. One example is in linear regression where $p(\boldsymbol{\theta}_k|\mathscr{M}_k)$ is a conjugate prior to the likelihood function [26]. In general this equation must be approximated. One of the approximation techniques is Bayesian information criterion (BIC) [38]. However, the BIC was not developed to address the model selection issue in the context of non-parametric models, since the number of (hyper-) parameters in a non-parametric model may not be directly associated with model complexity. Hence utilizing the BIC metric that penalizes model complexity based on the number of parameters becomes inappropriate. An alternative approach is Monte Carlo sampling as discussed subsequently.

3.1. Monte Carlo methods

With the rapid development of computing power, the use of Monte Carlo (MC) methods for model selection has become feasible and desirable, due to their ease of implementation and good performance in practice. The direct implementation of MC methods would be to draw J samples, $\{\boldsymbol{\theta}_k^j, j = 1, \ldots, J\}$ from its prior, $p(\boldsymbol{\theta}_k|\mathscr{M}_k)$, and approximate the likelihood in Eq. (5) as [37]:

$$p(\mathscr{D}|\mathscr{M}_k) \approx \sum_{j=1}^{J} p(\mathscr{D}|\boldsymbol{\theta}_k^j, \mathscr{M}_k) \tag{6}$$

However, this method is inefficient since it requires J samples to be drawn for each candidate model. When a large number of candidate models are considered, for example in the context of variable selection for regression modelling, this method is computationally infeasible.

Alternatively, the target distribution for MC sampling can be defined as the joint posterior distribution of the model and parameters, $p(\mathscr{M}_k, \boldsymbol{\theta}_k|\mathscr{D})$, from which samples are drawn: $\{\mathscr{M}_{k(j)}, \boldsymbol{\theta}_{k(j)}; j = 1, \ldots, J\}$. Here the index of the currently selected model $k(j)$ is dependent on the iteration index j, since for each sampling iteration, the selected model may differ. It should also be noted that the same model may be selected multiple times during the sampling iterations. Consequently the marginal posterior distribution of the k-th model can be approximated as follows:

$$
\begin{aligned}
p(\mathscr{M}_k|\mathscr{D}) &= \int p(\mathscr{M}_k, \boldsymbol{\theta}_k|\mathscr{D}) d\boldsymbol{\theta}_k \\
&\approx \frac{1}{J}\sum_{j=1}^{J} \delta\left(k(j)=k\right) \\
&= \frac{\text{Number of times the } k\text{-th model is selected}}{J}
\end{aligned} \tag{7}
$$

where $\delta\left(k(j)=k\right)=1$ if $k(j)=k$, otherwise it is equal to zero. The detailed sampling strategy is discussed in Section 4 with respect to variable selection for Gaussian process regression.

3.2. Prediction with Monte Carlo methods

Based on the Monte Carlo methods, prediction is made by calculating the following integral for new data $\mathscr{D}^*$:

$$
\begin{aligned}
p(\mathscr{D}^*|\mathscr{D}) &= \int\int p(\mathscr{D}^*|\mathscr{M}, \boldsymbol{\theta}) p(\mathscr{M}, \boldsymbol{\theta}|\mathscr{D}) d\mathscr{M}\, d\boldsymbol{\theta} \\
&\approx \frac{1}{J}\sum_{j=1}^{J} p\left(\mathscr{D}^*|\mathscr{M}_{k(j)}, \boldsymbol{\theta}_{k(j)}\right)
\end{aligned} \tag{8}
$$

However, from the perspective of model selection, the above model averaging may not be the most desired route. It may be more appropriate to select I models with the largest posterior probability, denoted by $\mathcal{M}_I = \{\mathscr{M}_i, i=1,\ldots,I\}$, and then the prediction is achieved by averaging over these selected models:

$$
\begin{aligned}
p(\mathscr{D}^*|\mathscr{D}, \mathcal{M}_I) &= \int p(\mathscr{D}^*|\mathcal{M}_I, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathscr{D}, \mathcal{M}_I) d\boldsymbol{\theta} \\
&\approx \frac{1}{J_I}\sum_{\mathscr{M}_{k(j)} \in \mathcal{M}_I} p\left(\mathscr{D}^*|\mathscr{M}_{k(j)}, \boldsymbol{\theta}_{k(j)}\right)
\end{aligned} \tag{9}
$$

where J_I is the total number of MC samples associated with the I selected models. The prediction equation in Eq. (9) is particularly relevant in certain situations of model selection. For example in the context of variable selection, it is desirable to retain those models with the largest probability, and thus only the selected variables need to be measured in the future.

4. Variable selection for Gaussian process regression

Variable selection is a special case of model selection, and it is especially appealing in spectroscopic calibration since typically only a small subset of the wavenumbers contain relevant information to the analyte properties to be predicted. In addition, variable selection is an effective way for model interpretation by identifying the most informative covariates. In industrial and laboratory practice, domain experts would be in general more convinced of the results from an interpretable model than a pure "black-box" model such as neural networks or Gaussian processes. Furthermore, variable selection potentially reduces the number of sensors, or covariates required to obtain the measurements, thereby reducing future costs. These considerations motivate the research in this paper.

Specifically let $\mathbf{c} = \{c_d, d=1,\ldots,p\}$ be a binary vector of which the d-th element indicates whether the d-th variable is included in the regression model, where p is the dimension of the covariates. The general

goal of variable selection is to maximize an objective function by searching over the space of $\mathbf{c}$, which comprises 2^p candidate models. Even when p is of moderate size, an exhaustive search over all candidates is computationally infeasible. In this section, an MCMC method is proposed for joint variable selection and hyper-parameter estimation within a Gaussian process regression model.

4.1. The proposed MCMC method

Let $\mathbf{X}$ and $\mathbf{y}$ be the covariates and responses of a Gaussian process regression model, respectively. In the context of variable selection, the covariance function in Eq. (1) is re-written as:

$$C(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = a_0 + a_1 \sum_{d:c_d=1} x_{id}x_{jd} + v_0 \exp\left(-w \sum_{d:c_d=1} (x_{id} - x_{jd})^2\right) + \delta_{ij}\sigma^2 \tag{10}$$

where $\boldsymbol{\theta} = (a_0, a_1, v_0, w, \sigma^2)$. In Gaussian process regression, variable selection necessitates the inference of the joint posterior distribution of $\boldsymbol{\theta}$ and the indicators $\mathbf{c}$: $p(\boldsymbol{\theta}, \mathbf{c}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{c}, \mathbf{X})p(\boldsymbol{\theta}|\mathbf{c})p(\mathbf{c})$. The likelihood, $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{c}, \mathbf{X})$, is Gaussian and the prior distributions for $p(\boldsymbol{\theta}|\mathbf{c})$ are similar to those used in the literature [35, 11]. Specifically the priors on all the hyper-parameters are defined as log-normal distributions, i.e. the logarithm of each element of $\boldsymbol{\theta}$ is normally distributed with mean -3 and standard deviation 3, corresponding to fairly vague priors. In addition, we follow [29] to define $p(\mathbf{c})$ indirectly by using a truncated geometric distribution based on the number of selected variables (i.e., the number of 1's in the vector $\mathbf{c}$ and is denoted by q):

$$p(q|\lambda) = \frac{\lambda(1-\lambda)^q}{1-(1-\lambda)^{p+1}}, \quad q = 0, 1, \ldots, p \tag{11}$$

The selection of $\lambda \in (0, 1)$ in the above prior distribution requires tuning, and this is discussed in the application studies subsequently.

To sample from the posterior distribution $p(\boldsymbol{\theta}, \mathbf{c}|\mathbf{X}, \mathbf{y})$, we adopt a hybrid MCMC method where Gibbs sampling [37] is used to sample alternately the two conditional posteriors: $p(\mathbf{c}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$ and $p(\boldsymbol{\theta}|\mathbf{c}, \mathbf{X}, \mathbf{y})$. Within each iteration of the Gibbs algorithm, $p(\boldsymbol{\theta}|\mathbf{c}, \mathbf{X}, \mathbf{y})$ can be sampled by using the Hamiltonian MC as for conventional inference of Gaussian process models; see Appendix. The sampling for $p(\mathbf{c}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})$ is implemented through the Metropolis-Hastings (M-H) algorithm [37]. The central component of the M-H algorithm is to propose a random move from the current sample, and the proposal is accepted with a certain probability such that the algorithm converges to the target distribution. Following [27, 26], the random move, $\mathbf{c}^*$, is proposed using the "birth" and "death" steps from the current model, $\mathbf{c}$:

- Birth: propose to add a new variable, uniformly drawn from those variables not selected currently.
- Death: uniformly select one of the included variables and propose to remove it.

The probabilities of selecting between the two steps in each MCMC iteration, i.e. the proposal distribution, are denoted by b_q and d_q for birth and death respectively. According to [29], the two terms are defined as:

$$\begin{cases} b_q = 1, d_q = 0 & \text{if } q = 0; \\ b_q = 0, d_q = 1 & \text{if } q = p; \\ b_q = d_q = 0.5 & \text{otherwise.} \end{cases} \tag{12}$$

Therefore the acceptance ratio for this M-H algorithm is:

$$r = \frac{p(\mathbf{c}^*|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})}{p(\mathbf{c}|\boldsymbol{\theta}, \mathbf{X}, \mathbf{y})} \times R \tag{13}$$

where R is the ratio of proposal probabilities given by d_{q+1}/b_q for a birth step, and b_{q-1}/d_q for a death step.

In summary, one iteration of the proposed hybrid MCMC method is as follows:

1. Generate a random move, $\mathbf{c}^*$, by proposing a birth or death step according to the probabilities b_q and d_q.
2. Calculate the acceptance ratio r as in Eq. (13).
3. Accept the random move with probability $\min(r, 1)$.
4. Run the Hamiltonian MC algorithm to sample for the hyper-parameters.

The initial sample of $\boldsymbol{\theta}$ and $\mathbf{c}$ is typically generated from the prior distribution. The convergence of the MCMC method can be assessed empirically through the monitoring of the posterior probability over the iterations [37]. Once the posterior probability reaches a relatively stable region, the sampling algorithm can be identified being convergent. Clearly, the computational cost of the algorithm is dominated by the large number of MCMC iterations needed to reach convergence, since the task is to jointly estimate the hyper-parameters and search through the space of binary vector $\mathbf{c}$ that has 2^p possible combinations. The actual computational time will be given for the case study in Section 5.

The marginal distribution of the indicators, $p(\mathbf{c}|\mathbf{X}, \mathbf{y})$, may be approximated by the frequency of a certain indicator $\mathbf{c}$ being selected as in Eq. (7). It may also be interesting to evaluate $p(c_d = 1|\mathbf{X}, \mathbf{y})$, the marginal probability of selecting the d-th variable, which is the sum of $p(\mathbf{c}|\mathbf{X}, \mathbf{y})$ over all distinct $\mathbf{c}$'s, subject to $c_d = 1$:

$$p(c_d = 1|\mathbf{X}, \mathbf{y}) = \sum_{\mathbf{c}:c_d=1} p(\mathbf{c}|\mathbf{X}, \mathbf{y}) \tag{14}$$

This marginal probability is useful to assess the relevance of individual covariates with respect to the prediction of the response variable.

Although the hybrid MCMC method is developed based on a specific form of covariance function in Eq. (10), it is applicable to a wide range of covariance functions as long as the dimension of the hyper-parameters does not change with the variable selection procedure. The following sub-section considers alternative forms of covariance function where special sampling techniques are required.

4.2. Alternative covariance function

One well known alternative form is to associate each covariate with an individual hyper-parameter in the exponential component of the covariance function in Eq. (10):

$$C(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = a_0 + a_1 \sum_{d:c_d=1} x_{id}x_{jd} + v_0 \exp\left(-\sum_{d:c_d=1} w_d(x_{id} - x_{jd})^2\right) + \delta_{ij}\sigma^2 \tag{15}$$

where the hyper-parameters are denoted by $\boldsymbol{\theta} = (a_0, a_1, v_0, \sigma^2, w_{d:c_d=1})$. Gaussian processes with such a covariance function fall within the family of automatic relevance determination (ARD) models [10, 9]. This ARD model, through assigning w_d's for each covariate, is capable of assessing the magnitude of relevance of the corresponding variable to the prediction, and it naturally provides a mechanism for variable selection. A straightforward approach is to define a threshold such that the covariates whose w_d's are lower than the threshold will be eliminated from the regression model. However, this approach requires to estimate a large number of hyper-parameters (e.g. 408 hyper-parameters for the spectral data considered in this paper, including a_0, a_1, v_0, σ^2 and 404 w_d's), and thus is computationally demanding. In addition, the high dimension of the problem typically results in a complex posterior distribution with many local maxima, further complicating the inference task. One possibility is to employ principal component analysis (PCA) to reduce the dimension of the covariates prior to the development of Gaussian process model [11]; however the capability to facilitate model interpretation using variable selection is compromised.

It is also possible to modify the hybrid MCMC algorithm developed in the previous sub-section for variable selection with the ARD model. The major difficulty is that the dimension of the hyper-parameters can change with variable selection, and thus conventional MCMC methods are not directly applicable. To address this issue, the reversible jump MCMC (RJMCMC) algorithm [36], is required to undertake the

inference of the hyper-parameters with varying dimensionality. RJMCMC is known to converge slowly and typically gives low acceptance rate. For example, when the d-th variable is included, the corresponding hyper-parameter w_d must also be drawn from a proposal distribution, which is typically defined as its prior. However, without information from the training data, the prior distribution is often specified as relatively vague compared with the likelihood function. Hence the random sample drawn from the prior distribution normally has a very small posterior probability. Therefore this strategy will give a very low acceptance rate, which is undesirable since a large amount of computational effect is wasted, and only a small number of candidate models can be explored.

In view of these practical difficulties, the ARD-type covariance functions are not explored further in this paper. By replacing the w_d's with a single hyper-parameter w, the Gaussian process may lose flexibility and accuracy in terms of modelling the regression function. However, in the situation of high dimensional covariates, the challenge of effective inference and computation may well offset the advantage of ARD models; see [39] for more detailed discussion on this matter.

5. Application study

The proposed variable selection method for Gaussian process regression is applied to the near infrared (NIR) spectroscopic data briefly mentioned in Section 1, where the spectra was depicted in Fig. 1. The transmittance spectra were generated from the analysis of Escitalopram® tablets, manufactured by a pharmaceutical company. The aim of the study was to determine the weight percentage of the active substance in the tablets, based on the NIR spectra over the wavenumber range of $7400 - 10500$ cm^{-1} (corresponding to 404 covariates) [40]. The tablet samples corresponding to the "preliminary calibration set" defined in the original paper [40] were used, where the data set comprised 120 samples. The data is randomly partitioned into training and testing sets, each having 60 samples. The original data set is available at `http://www.models.kvl.dk/research/data/Tablets/`. Following the common practice, the data is pre-processed to have zero mean and unit standard deviation at each variable before it is used for training a Gaussian process. The focus of the case study is to examine the effectiveness of the variable selection procedure, its computational efficiency, and the resulting predictive performance in terms of the root mean squared error for prediction (RMSEP). Note that although RMSEP is among the most widely used measure to assess the prediction performance of regression models, it may not be the most meaningful criterion in practice. For example, if the spectrometer is calibrated only for plant monitoring purpose, then a certain improvement in RMSEP may not have practical difference. In contrast, if the sensor is being used in an on-line feedback control loop to regulate the product quality, then a seemingly insignificant improvement may greatly enhance the controller performance in terms of stability and responsiveness [41]. The choice of an appropriate performance measure in practice depends on the intended applications, and it is outwith the scope of this paper.

Initially a number of preliminary MCMC runs were undertaken to determine $\lambda \in (0, 1)$ in the truncated geometric prior distribution for the number of selected variables (Eq. (11)). Selecting a λ close to unity encourages to select a few covariates, but it typically leads to a low acceptance rate. On the other hand, a small value of λ does not reflect the prior belief that a small subset of covariates are sufficient for predicting the response. In the current study, a subjective prior distribution is adopted which ensures that the MCMC algorithms converge to the region where approximately fewer than 10 covariates are selected. This reflects the belief that *a priori* the number of the most relevant variables to the prediction should not significantly exceed 10, because the spectrum appears to have only one "visually" identifiable band (one band can contain several variables) that is affected by the active substance [40]. Using this criterion, λ was selected and resulted in reasonable acceptance rate of approximately 0.25. More elaborate methods, such as the hierarchical Bayesian approach [42], can be used to facilitate the specification of the prior distribution.

Based on the selected prior, the proposed MCMC algorithm was then executed for the Gaussian process regression model, with 50 covariates being randomly selected from the 404 variables as the initial setting. Due

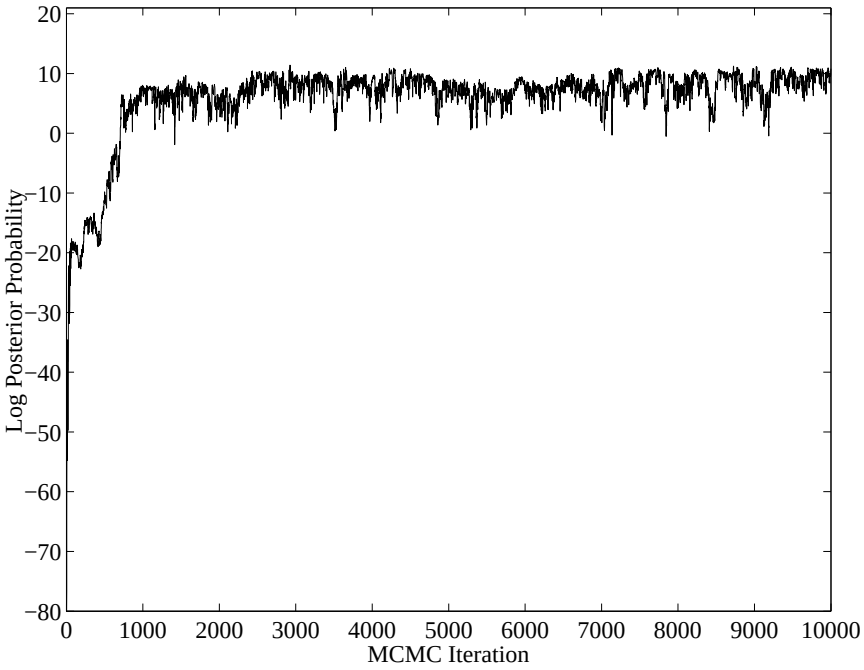

Figure 2: Log posterior probability during the MCMC iterations.

to the large number of candidate models, the algorithms were run for 10000 iterations to enable adequate exploration of the model space. Although this number of iterations is only capable of assessing a small proportion of all the $2^{404} \approx 4.13 \times 10^{121}$ candidate models, the MCMC methods tend to concentrate on regions of high posterior probability, and thus satisfactory results are expected [26].

Fig. 2 and 3 display the log posterior probability and the number of selected variables respectively during the MCMC iterations. Based on the plot of log posterior probability, the MCMC algorithm tends to converge after approximately 1000 iterations. Fig. 3 shows that the number of selected variables moves from the initial value 50 to between 2 and 10 when the algorithm appears to have converged.

To assess the consistency of the variable selection method, the MCMC algorithm was executed for 10000 iterations for other initial settings of the selected variables, i.e. the first 50 variables were selected; the last 50 variables were selected; the middle 50 variables were selected; and 50 of the 404 variables were randomly selected. In each run, the relative marginal probabilities of the indicators, $p(\mathbf{c}|\mathbf{X}, \mathbf{y})$, were calculated based on the last 9000 MC samples when the algorithm was judged to be convergent, and the marginal probabilities were used to obtain $p(c_d|\mathbf{X}, \mathbf{y})$. The results are shown in Fig. 4. The location of the spikes corresponds to those variables selected with high probability. Although there is some variation, the four plots are largely similar, and the location of the spikes are approximately aligned. This indicates that the proposed MCMC sampling strategy can consistently identify a similar set of relevant covariates for prediction, even when the initial settings of the selected variables are significantly different.

The most notable finding from Fig. 4 is that those variables close to wavenumber 8830 cm^{-1} have a relatively high probability of being selected. According to the fundamental spectroscopic analysis, in the NIR spectrum for the tablets studied, this is the only visually identifiable band, and is believed to be informative with regard to the prediction of the active substance in the tablets [40]. This confirms that the proposed strategies are capable of selecting informative covariates for the calibration model.

The predictive performance of the various calibration methods is summarized in Table 1. For a reliable

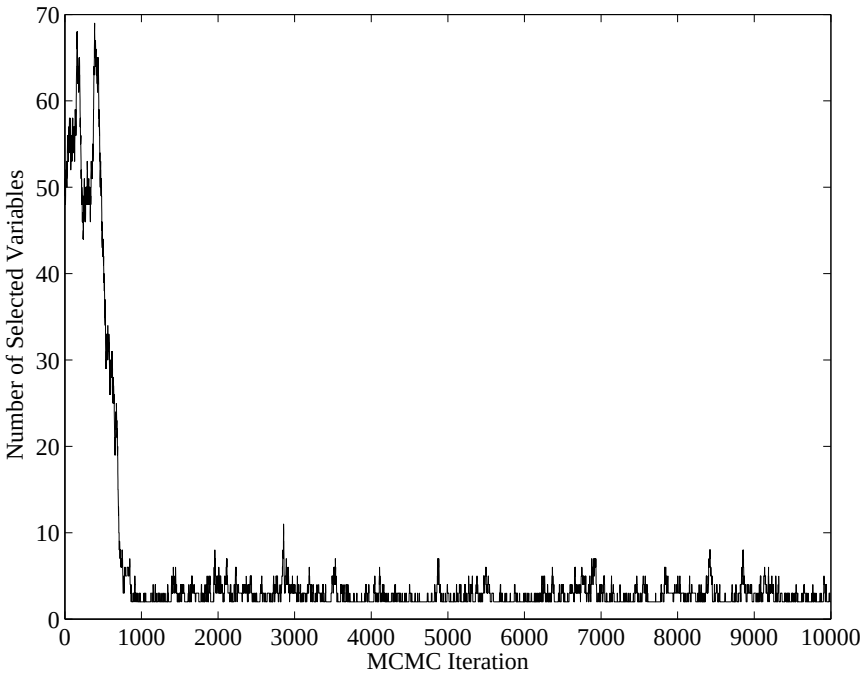

Figure 3: Number of selected variables during the MCMC iterations.

evaluation of these techniques, a common strategy is adopted where the random partitioning of training and test sets is repeated 50 times, and the average RMSEP is reported along with its standard error. We also report the number of selected covariates and CPU time for training, all averaged over 50 repeats. The PLS model is regarded as the benchmark method, and the number of latent variables to be retained is determined through five-fold cross-validation. The method "GP+PCA", as adopted in our earlier work [11], is to apply principal component analysis (PCA) to the original data for dimension reduction, and then to develop a Gaussian process calibration model. Five principal components that explain in excess of 99.99% of the total variance were retained, and this low dimensional data makes possible the use of an ARD-type covariance function (i.e. through replacing a single w in Eq. (1) by five w_d's), which is also the effective method reported in [11]. The inference of the hyper-parameters was conducted using the Hamiltonian MC as given in Appendix.

Table 1 clearly indicates the superior predictive performance of GP based models to the benchmark

Table 1: Calibration results and CPU time (of training) for different methods. All the values are averaged over 50 repeats and the standard error of RMSEP is also shown. The relative improvement (Rel. Imp.) in prediction is with respect to PLS.

Method	RMSEP	Rel. Imp.	No. of Covariates	CPU (s)
PLS	0.260 ± 0.004	–	404	3.2
GP + PCA	0.194 ± 0.004	25.4%	404	26.8
GP + Variable Selection				83.2
One Best	0.193 ± 0.006	25.8%	3	
Five Best	0.180 ± 0.005	30.8%	8	
Twenty Best	0.160 ± 0.005	38.5%	40	

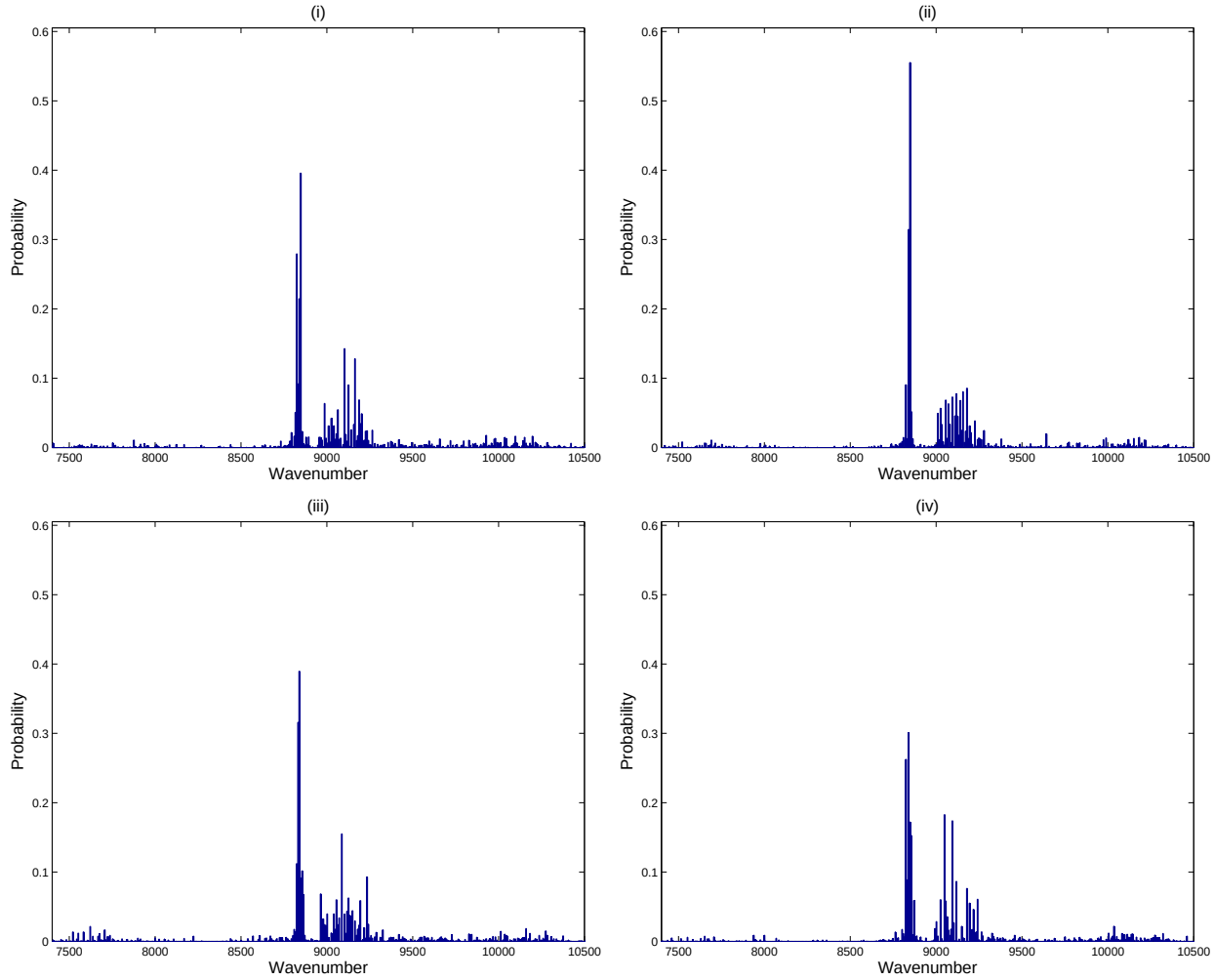

Figure 4: Marginal probability of the individual variables for the four runs starting with different initial settings: (i) $\{c_d = 1, d = 1, \ldots, 50\}$; (ii) $\{c_d = 1, d = 355, \ldots, 404\}$; (iii) $\{c_d = 1, d = 178, \ldots, 227\}$; (iv) randomly select 50 variables and set the corresponding c_d's to one.

PLS, as a result of more flexible modelling capability of Gaussian process than the linear PLS method. However, the "GP+PCA" model has the difficulty of interpretability with regard to which wavenumbers (covariates) contribute the most to the prediction, and this issue can be addressed by using variable selection method. The prediction with variable selection was made based on Eq. (9), where only the predictions from the most probable models, in terms of the highest posterior probability of $\mathbf{c}$, are averaged. The model with the highest posterior probability has only 3 selected variables, and gives an RMSEP of 0.193 that is statistically indistinguishable from "GP+PCA". If the best five models are used with eight variables being selected, a lower RMSEP of 0.180 can be achieved. Furthermore, even lower RMSEP is attained if the best twenty models are included, corresponding to 40 selected variables. Clearly, averaging over more models can result in more reliable predictions at the cost of more variables being selected, and this trade-off should be determined with regard to practical requirement. In addition, the computational cost is another factor to consider in practice. The flexible GP model requires much more time for training than PLS, and even

more time if variable selection is considered. Nevertheless, we believe that the CPU time given in Table 1 is manageable for many application scenarios.

6. Conclusions

This paper proposed the application of MCMC method to the Bayesian variable selection problem for Gaussian process regression. The general Bayesian model selection framework was first discussed, and then a hybrid MCMC algorithm was proposed to alternately sample for variables and hyper-parameters of a Gaussian process regression model. The proposed strategy was evaluated through its application to chemometric calibration of spectrometers, where typically several hundred covariates are considered. The results indicated that the variable selection method can attain enhanced prediction accuracy and model interpretability in comparison with conventional PLS and PCA-based Gaussian process regression models. In principle, the proposed MCMC strategy is equally applicable to variable selection with other non-linear regression models, and this will be further investigated.

In view of the computational complexity of MCMC methods, we are studying more efficient methods for variable selection. Among various alternatives, the approach of LASSO (least absolute shrinking and selection operator) [43] appears to be attractive in a number model selection problems, and it may be further extended to variable selection with Gaussian process regression.

Appendix: The Hamiltonian Monte Carlo algorithm

The Hamiltonian Monte Carlo (HMC) is a family of MCMC methods based on the concept of dynamic systems in physics [34]. Recently HMC has been shown in various applications to converge significantly faster than the Metropolis-Hastings algorithm, since the dynamic method avoids the random walk behavior inherent in conventional approaches [34, 35, 33]. This section gives a brief description of HMC based on the implementation of [12].

The method of HMC creates a virtual dynamic system by augmenting hyper-parameters $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]$ with *momentum* variables $\mathbf{p} = [p_1, \ldots, p_d]$. The objective is thus to sample from the distribution of the combined system $p(\boldsymbol{\theta}, \mathbf{p}) \propto \exp(-E - K)$, where E is the "potential" energy of the parameters defined in terms of the posterior distribution as $\exp(-E) \propto p(\boldsymbol{\theta}|\mathscr{D})$, and K is the "kinetic" energy of momenta, $K = 0.5 \sum_j p_j^2$. The total energy of the system is defined as $\mathcal{H} = E + K$. The dynamic system is simulated through a discretization approach, known as the leapfrog steps:

$$p_j(t + \frac{\epsilon}{2}) = p_j(t) - \frac{\epsilon}{2} \frac{\partial E}{\partial \theta_j} (\boldsymbol{\theta}(t)) \tag{16}$$

$$\theta_j(t + \epsilon) = \theta_j(t) + \epsilon p_j \left(t + \frac{\epsilon}{2}\right) \tag{17}$$

$$p_j(t + \epsilon) = p_j(t + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \frac{\partial E}{\partial \theta_j} (\boldsymbol{\theta}(t + \epsilon)) \tag{18}$$

where t is the virtual time for the dynamic system, ϵ is the step size for discretizing the dynamic system. Note that the definition of E results in the following expression for the above derivative terms:

$$\frac{\partial E}{\partial \theta_j} (\boldsymbol{\theta}(t)) = -\frac{\partial}{\partial \theta_j} \log \left[p \left(\boldsymbol{\theta}(t) | \mathscr{D} \right) \right]$$

The exploration of the hyper-parameter space is thus facilitated by using the gradient information, and the random walk search is avoided.

The step size, ϵ, can be heuristically determined by executing some initial runs of the HMC algorithm, and it was set to 0.1 in this study. In summary, one iteration of the HMC sampling is as follows:

1. Starting from previous sample $(\boldsymbol{\theta}^{i-1}, \mathbf{p}^{i-1})$ at iteration $i-1$, perform one leapfrog step as in Eqs. (16)(17)(18) with step size ϵ, resulting in the proposed value $(\boldsymbol{\theta}^*, \mathbf{p}^*)$.
2. Compute the acceptance rate:

$$r = \exp\left[\mathcal{H}(\boldsymbol{\theta}^{i-1}, \mathbf{p}^{i-1}) - \mathcal{H}(\boldsymbol{\theta}^*, \mathbf{p}^*)\right]$$

3. Accept the proposal with probability, $\min(1, r)$; otherwise retain the old values with negative momenta as $(\boldsymbol{\theta}^i, \mathbf{p}^i) = (\boldsymbol{\theta}^{i-1}, -\mathbf{p}^{i-1})$.
4. Update the total energy of the system by perturbing the momenta according to $p_j = \alpha\, p_j + \sqrt{1-\alpha^2}\, v_j$ for all j, where v_j are drawn from a zero-mean unit-variance Gaussian. α is normally set to 0.95 to ensure a reasonable level of perturbation [12].

As $\boldsymbol{\theta}$ and $\mathbf{p}$ are independent, the last step is Gibbs sampling of the momenta, which allows the HMC to explore regions with different values of $\mathcal{H}$. The sequence, $\{\boldsymbol{\theta}^i, i = 1, 2, \ldots\}$, defines the samples generated from the posterior distribution $p(\boldsymbol{\theta}|\mathscr{D})$.

References

[1] P. J. Brown, Measurement, regression, and calibration, Oxford University Press, 1993.
[2] P. Geladi, B. R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17.
[3] E. Vigneau, M. F. Devaux, E. M. Qannari, P. Robert, Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration, Journal of Chemometrics 11 (1997) 239–249.
[4] B. G. Osborne, T. Fearn, P. H. Hindle, Practical NIR Spectroscopy, Longman, Harlow, U.K., 1993.
[5] H. H. Thodberg, A review of Bayesian neural networks with an application to near infrared spectroscopy, IEEE Transactions on Neural Networks 7 (1996) 56–72.
[6] N. Hernández, I. Talavera, A. Dago, R. J. Biscay, M. M. C. Ferreira, D. Porro, Relevance vector machines for multivariate calibration purposes, Journal of Chemometrics (2008) 686–694.
[7] U. Thissen, B. Ustun, W. J. Melssen, L. M. C. Buydens, Multivariate calibration with least-squares support vector machines, Analytical Chemistry 76 (2004) 3099–3105.
[8] A. O'Hagan, Curve fitting and optimal design for prediction (with discussion), Journal of the Royal Statistical Society B 40 (1978) 1–42.
[9] R. M. Neal, Bayesian learning for neural networks, Springer-Verlag, New York, 1996.
[10] D. J. C. MacKay, Introduction to Gaussian processes, in: C. M. Bishop (Ed.), Neural Networks and Machine Learning, volume 168 of F: Computer and Systems Sciences, NATO Advanced Study Institute, Springer, Berlin, Heidelberg, 1998, pp. 133–165.
[11] T. Chen, J. Morris, E. Martin, Gaussian process regression for multivariate spectroscopic calibration, Chemometrics and Intelligent Laboratory Systems 87 (2007) 59–67.
[12] C. E. Rasmussen, Evaluation of Gaussian processes and other methods for non-linear regression, Ph.D. thesis, University of Toronto, Canada (1996).
[13] J. Yuan, K. Wang, T. Yu, M. Fang, Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression, International Journal of Machine Tools and Manufacture 48 (2008) 47–60.
[14] S. Brahim-Belhouari, A. Bermak, Gaussian process for nonstationary time series prediction, Computational Statistics and Data Analysis 47 (2004) 705–712.
[15] T. Chen, J. Ren, Bagging for Gaussian process regression, Neurocomputing 72 (2009) 1605–1610.
[16] B. Likar, J. Kocijan, Predictive control of a gas-liquid separation plant based on a Gaussian process model, Computers and Chemical Engineering 31 (2007) 142–152.
[17] J. Q. Shi, B. Wang, R. Murray-Smith, D. M. Titterington, Gaussian process functional regression modeling for batch data, Biometrics 63 (2007) 714–723.
[18] Q. Tang, Y. Lau, S. Hu, W. Yan, Y. Yang, T. Chen, Response surface methodology using Gaussian processes: towards optimizing the trans-stilbene epoxidation over Co^{2+}-NaX catalysts, Chemical Engineering Journal 156 (2010) 423–431.
[19] S. Osborne, R. Jordan, R. Kunnemeyer, Method of wavelength selection for partial least squares, Analyst 122 (1997) 1531–1537.
[20] V. Centner, D.-L. Massart, O. de Noord, S. de Jong, B. M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anlytical Chemistry 68 (1996) 3851–3858.
[21] D. Whitley, A genetic algorithm tutorial, Statistics and Computing 4 (1994) 65–85.
[22] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tables, Chemometrics and Intelligent Laboratory Systems 69 (2003) 3–12.
[23] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, D. B. Kelp, 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 (1997) 71–86.

[24] H. Chipman, Bayesian variable selection with related predictors, Canadian Journal of Statistics 24 (1996) 17–36.
[25] E. I. George, R. E. McCulloch, Variable selection via gibbs sampling, Journal of the American Statistical Association 88 (1993) 881–889.
[26] P. J. Brown, M. Vannucci, T. Fearn, Multivariate Bayesian variable selection and prediction, Journal of the Royal Statistical Society B 60 (1998) 627–641.
[27] P. J. Brown, M. Vannucci, T. Fearn, Bayesian wavelength selection in multicomponent analysis, Journal of Chemometrics 12 (1998) 173–182.
[28] P. Dellaportas, J. Forster, I. Ntzoufras, On bayesian model and variable selection using MCMC, Statistics and Computing 12 (2002) 27–36.
[29] D. G. T. Denison, B. K. Mallick, A. F. M. Smith, Bayesian MARS, Statistics and Computing 8 (1998) 337–346.
[30] R. Kohn, M. Smith, D. Chan, Nonparametric regression using linear combinations of basis functions, Statistics and Computing 11 (2001) 313–322.
[31] M. Smith, R. Kohn, Nonparametric regression using bayesian variable selection, Journal of Econometrics 75 (1996) 317–343.
[32] P. J. Brown, T. Fearn, M. Vannucci, Bayesian wavelet regression on curves with application to a spectroscopic calibration problem, Journal of the American Statistical Association 96 (2001) 398–408.
[33] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[34] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid Monte Carlo, Physics Letters B 195 (1987) 216–222.
[35] R. M. Neal, Monte Carlo implementation of Gaussian process models for Bayesian regression and classification, Tech. Rep. No. 9702, Department of Statistics, University of Toronto, Canada, available at `http://www.cs.utoronto.ca/~radford/ftp/mc-gp.pdf` (1997).
[36] P. J. Green, Reversible jump markov chain monte carlo computation and bayesian model determination, Biometrika 82 (1995) 711–732.
[37] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer, New York, 1999.
[38] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[39] T. Chen, Bayesian data analysis via Monte Carlo computation – application studies in process systems engineering, Ph.D. thesis, University of Newcastle upon Tyne, UK (2006).
[40] M. Dyrby, S. B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric quantitation of the active substance in a pharmaceutical tablet using near infrared (NIR) transmittance and NIR FT raman spectra, Applied Spectroscopy 56 (2002) 579–585.
[41] A.-J. Su, C.-C. Yu, B. A. Ogunnaike, On the interaction between measurement strategy and control performance in semiconductor manufacturing, Journal of Process Control 18 (3-4) (2008) 266–276.
[42] D. G. T. Denison, C. C. Holmes, B. K. Mallick, A. F. M. Smith, Bayesian Methods for Nonlinear Classification and Regression, John Wiley & Sons, 2002.
[43] T. Park, G. Casella, The Bayesian LASSO, Journal of the American Statistical Association 103 (2008) 681–686.