Bayesian linear regression and variable selection for spectroscopic calibration

Tao Chen [a,*], Elaine Martin [b]

[a] School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore
[b] School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

ARTICLE INFO

ABSTRACT

This paper presents a Bayesian approach to the development of spectroscopic calibration models. By formulating the linear regression in a probabilistic framework, a Bayesian linear regression model is derived, and a specific optimization method, i.e. Bayesian evidence approximation, is utilized to estimate the model "hyper-parameters". The relation of the proposed approach to the calibration models in the literature is discussed, including ridge regression and Gaussian process model. The Bayesian model may be modified for the calibration of multivariate response variables. Furthermore, a variable selection strategy is implemented within the Bayesian framework, the motivation being that the predictive performance may be improved by selecting a subset of the most informative spectral variables. The Bayesian calibration models are applied to two spectroscopic data sets, and they demonstrate improved prediction results in comparison with the benchmark method of partial least squares.

1. Introduction

Multivariate calibration techniques have been widely applied to spectroscopic measurements for the extraction of chemical and/or physical information about the analytes [1–4]. For the development of calibration models, a variety of techniques have been proposed to address the collinearity issue as a result of the large number of spectral wavelengths, including partial least squares (PLS) [2], principal component regression (PCR) [3], and ridge regression [5]. These models have become standard approaches to spectroscopic calibration, and they are available in most chemometric software packages.

During the last decade, when facing increasing requirement for more accurate and reliable calibration models from industry and laboratory analysis, the chemometric community has put significant effort in the development of advanced calibration techniques. These research have materialized in both enhanced pre-processing algorithms (such as extended multiplicative signal correction [6], extended inverted signal correction [7] and direct orthogonal signal correction [8]) and improved regression models (including locally weighted regression [9], neural networks [9,10], support vector machine [11] and Gaussian process [12]). This paper is mainly focused on the development of advanced regression models for calibration, and thus the pre-processing techniques are not discussed further.

Although calibration models are typically developed by including all the available wavelengths, both theoretical and experimental evidence exists to demonstrate that it is possible to enhance prediction performance through the implementation of variable selection [1,13,14], also termed wavelength selection. The assumption is that there may be parts of the wavelengths that contain little information about the analyte properties. When these wavelengths are included in the regression model, predictive performance on unseen test data will then be poor.

To develop calibration techniques with variable selection, most strategies are based on a regression model whilst optimizing calibration performance by selecting/removing spectral variables. For example, iterative PLS [15] 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 based on analysis of the PLS regression coefficients [16]. Despagne and Massart investigated several variable selection methods for neural networks [17], including Hinton diagrams, magnitude approach, determination of saliency, variance propagation and partial modeling. The other 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 for variable selection in spectroscopic applications [13,18]. A comparative study of a number of variable selection algorithms was reported [19], and it was shown that the GA approach demonstrated improved prediction ability over conventional PLS models.

* Corresponding author. Tel.: +65 6513 8267; fax: +65 6794 7553.
E-mail address: chentao@ntu.edu.sg (T. Chen).

More recently, there has been a significant interest in Bayesian statistical approach to the development of calibration models. Nounou et al. [20] presented a Bayesian latent variable regression model for the analysis of process data, and the technique can be applied for calibration purpose. The advantage of introducing Bayesian methodology into neural networks was reviewed [10] through the application to the calibration of near infrared spectroscopy. As early advocates of Bayesian approach, Brown and co-workers proposed a number of calibration models that demonstrated promising results, such as Bayesian variable selection methods for calibration [21,22] and wavelet regression [23]. One salient feature of Brown's work is that Markov chain Monte Carlo (MCMC) simulation is employed for the inference of the model parameters. MCMC method provides the possibility of model averaging that has been shown to attain more robust predictions than a single model [24]; however, the high computational cost of MCMC simulation may restrict its wide acceptance in practice.

This paper presents a Bayesian linear regression approach to the development of spectroscopic calibration models. A specific optimization method, Bayesian evidence approximation, is utilized to estimate the model "hyper-parameters" (to be defined in the subsequent section). The proposed method attains the general advantages of Bayesian models, including: (a) maintaining a balance between model accuracy and complexity (and thus a way to address the "over-fitting" issue); (b) providing a predictive distribution that conveniently provides the prediction intervals; and (c) removing the computationally intensive cross-validation step (such as in PLS and ridge regression). Furthermore, a variable selection strategy is developed within the Bayesian framework. The variable selection algorithm is also efficiently implemented using the Bayesian evidence approximation that materializes in substantially lower computation than the MCMC simulation adopted by Brown et al. [21,22]. The Bayesian calibration models are applied to two spectroscopic data sets, and they demonstrate improved prediction when compared with the standard PLS technique.

2. Bayesian linear regression

The focus of this paper is to develop a regression model for spectroscopic calibration given a training data set comprising N observations $\mathcal{D} = \{\mathbf{x}_n, y_n\}$, $n = 1, \ldots, N$, where $\mathbf{x}_n$ and y_n are M-dimensional predictor and scalar response variables respectively. The Bayesian approach to statistical modeling typically consists of the following two stages: (1) the inference of the posterior distribution of model parameters $\mathbf{w}$, that is the proportional to the product of the prior distribution and likelihood function: $p(\mathbf{w}|\mathcal{D}) \propto p(\mathbf{w})p(\mathcal{D}|\mathbf{w})$; and (2) the calculation of the predictive distribution of y^* given any new predictors $\mathbf{x}^*$. Formally the prediction is obtained by integrating over the posterior distribution of $\mathbf{w}$ as follows:

$$p(y^*|\mathbf{x}^*, \mathcal{D}) = \int p(y^*|\mathbf{x}^*, \mathbf{w})p(\mathbf{w}|\mathcal{D})\,d\mathbf{w} \tag{1}$$

Through the integration in Eq. (1), the predictive distribution explicitly quantifies the uncertainty associated with the model parameters, and it provides a natural way to construct the prediction intervals. This is in contrast to the non-probabilistic regression models (e.g. PLS), where additional strategies, typically involving Taylor expansion, must be employed to determine the prediction intervals [25].

Within this section, the development of Bayesian inference and prediction method for linear regression models will be presented based on the discussions in [1,26]. For ease of derivation, univariate response is initially considered. The development of a calibration model for multivariate responses is deferred to Section 2.4.

2.1. The model

The linear regression model can be written as $y = \mathbf{w}^T\mathbf{x} + \epsilon$, where ϵ is a zero mean Gaussian noise term with variance σ^2. Thus the conditional distribution of y is $p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = G(y; \mathbf{w}^T\mathbf{x}, \sigma^2)$. For the rest of this paper the dependency of the distributions on $\mathbf{x}$ will be omitted to simplify the notations. In the framework of maximum likelihood estimation (MLE), $\mathbf{w}$ is treated as fixed parameter and is estimated as

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$, and $\mathbf{y} = (y_1, \ldots, y_N)^T$. This is equivalent to the result obtained using traditional least squares. The noise variance may also be determined using MLE:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{\mathbf{w}}^T\mathbf{x}_n)^2 \tag{3}$$

One of the major issues with the MLE (and least squares) is that $\mathbf{X}^T\mathbf{X}$ may not be invertible when the predictor variables are highly correlated or the number of predictors is larger than the number of available training data points, i.e. $M > N$. There are a number of techniques to address this issue in the chemometric literature, including PCR [3], PLS [2], ridge regression [5] and variable selection [1].

Alternatively the Bayesian treatment of linear regression model introduces a prior probability distribution over the model parameters $\mathbf{w}$. Specifically a zero-mean isotropic Gaussian prior is usually adopted such that $p(\mathbf{w}|\alpha) = G(\mathbf{w}; \mathbf{0}, \alpha^{-1}\mathbf{I})$. Since this prior distribution is conjugate to the likelihood function, the posterior distribution is also Gaussian [26]:

$$p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) \propto p(\mathbf{y}|\mathbf{w}, \alpha, \sigma^2)p(\mathbf{w}|\alpha) = G(\mathbf{w}; \mathbf{m}, \mathbf{S}) \tag{4}$$

where

$$\mathbf{m} = \sigma^{-2}\mathbf{S}\mathbf{X}^T\mathbf{y} \tag{5}$$

$$\mathbf{S} = (\alpha\mathbf{I} + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1} \tag{6}$$

If a non-informative prior is adopted, i.e. $\alpha \to 0$, the posterior mean in Eq. (5) reduces to the MLE result given by Eq. (2). In the development of spectroscopic calibration models, the number of data points is typically smaller than the number of predictor variables (wavelengths), i.e. $N < M$. In this case the Woodbury inversion identity (see e.g. [27], §A.2.4) can be applied to calculate $\mathbf{S}$, reducing the computation to $O(N^3)$.

After the posterior distribution of $\mathbf{w}$ is obtained, the prediction y^* can be calculated for a new point $\mathbf{x}^*$ through the integration as in Eq. (1). It is well known that the predictive distribution is also Gaussian:

$$p(y^*|\mathbf{y}, \alpha, \sigma^2) = \int p(y^*|\mathbf{w}, \sigma^2)p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2)\,d\mathbf{w}$$
$$= G(y^*; \mathbf{m}^T\mathbf{x}^*, \sigma^2(\mathbf{x}^*)) \tag{7}$$

where $\sigma^2(\mathbf{x}^*) = \sigma^2 + \mathbf{x}^{*T}\mathbf{S}\mathbf{x}^*$. It can be seen that the predictive variance consists of two parts: the noise on the data and the uncertainty about the parameters $\mathbf{w}$.

The appropriate value for σ^2 and α (termed hyper-parameters since they differ from model parameters $\mathbf{w}$) will be determined through the method of Bayesian evidence approximation as discussed subsequently.

2.2. Bayesian evidence approximation

In a formal Bayesian treatment of the linear regression model, higher-level prior distributions (i.e. hyper-prior) can be introduced over σ^2 and α, and the prediction can be made by integrating over σ^2, α, and the regression parameters $\mathbf{w}$. This methodology is known in the statistics literature as Bayesian hierarchical approach [28]. However, it is not possible to integrate over all these parameters analytically, and thus approximate methods must be adopted, such as Monte Carlo sampling approaches [28]. In this paper an efficient approximation is utilized where the hyper-parameters are set to specific values such that the following marginal likelihood function is maximized:

$$p(\mathbf{y}|\alpha, \sigma^2) = \int p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)\,d\mathbf{w}$$

$$= (2\pi)^{-(N/2)}|\mathbf{C}|^{-(1/2)}\exp\left(-\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y}\right) \tag{8}$$

where $\mathbf{C} = \sigma^2\mathbf{I} + \alpha^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$. This approach is known as empirical Bayes, type-II maximum likelihood, and evidence approximation in the literature [26,29]. To find the optimal value for α and σ^2, the logarithm of the marginal likelihood in Eq. (8) can be written as [26,30]

$$\ln p(\mathbf{y}|\alpha, \beta) = \frac{M}{2}\ln\alpha - \frac{N}{2}\ln\sigma^2 - \frac{1}{2}\ln|\mathbf{S}^{-1}| - \frac{\sum_{n=1}^{N}(y_n - \mathbf{m}^{\mathrm{T}}\mathbf{x}_n)^2}{2\sigma^2}$$
$$- \frac{\alpha\mathbf{m}^{\mathrm{T}}\mathbf{m}}{2} - \frac{N}{2}\ln(2\pi) \tag{9}$$

The derivative of $\ln p(\mathbf{y}|\alpha, \beta)$ with respect to α can then be derived:

$$\frac{d\ln p(\mathbf{y}|\alpha, \beta)}{d\alpha} = \frac{M}{2\alpha} - \frac{1}{2}\mathrm{Tr}(\mathbf{S}) - \frac{\mathbf{m}^{\mathrm{T}}\mathbf{m}}{2} \tag{10}$$

By setting the above derivative to zero, α can be obtained by

$$\alpha = \frac{\gamma}{\mathbf{m}^{\mathrm{T}}\mathbf{m}} \tag{11}$$

where γ is defined as

$$\gamma = M - \alpha\,\mathrm{Tr}(\mathbf{S}) \tag{12}$$

Similarly setting $d\ln p(\mathbf{y}|\alpha, \beta)/d\sigma^2 = 0$ gives

$$\sigma^2 = \frac{\sum_{n=1}^{N}(y_n - \mathbf{m}^{\mathrm{T}}\mathbf{x}_n)^2}{N - \gamma} \tag{13}$$

It should be noted that Eqs. (11)–(13) are implicit solutions for the hyper-parameters because $\mathbf{m}$, $\mathbf{S}$ and γ are dependent on the value of α and σ^2. Therefore α and σ^2 are estimated iteratively using Eqs. (11)–(13), where $\mathbf{m}$ and $\mathbf{S}$ must be updated using Eqs. (5) and (6) in each iteration. The convergence of the algorithm can be determined if the difference in the log-likelihood (Eq. (9)) between two successive iterations is sufficiently small.

However, preliminary study has confirmed that the log likelihood function may have multiple local maxima, and thus the iterative procedure does not guarantee to find the global maximum. The issue of multi-modes is typically more serious for variable selection to be discussed in Section 3, where much more hyper-parameters need to be estimated. To alleviate the effect of local maxima, we adopt the common practice to try a number of random starting values for the hyper-parameters, and then select the model with the largest log likelihood value.

2.3. Relation to ridge regression

Ridge regression [5] addresses the collinearity issue in ordinary least squares by introducing a regularization term $\lambda \geq 0$ to penalize the magnitude of the regression parameters:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\sum_{n=1}^{N}(y_n - \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)^2 + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{14}$$

Compared with ordinary least squares and maximum likelihood estimation (Eq. (2)), the parameter estimation in Eq. (14) requires the calculation of $(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}$. This can improve the numerical stability if the predictors are highly correlated and thus $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is close to singular. Furthermore, the introduction of the regularization term is helpful to control the model complexity and to alleviate the "overfitting" problem [26]. A number of techniques have been proposed to determine the value of λ, including cross-validation [5] and more recently harmonious approach [31].

The Bayesian linear regression model proposed in this paper is closely related to the ridge regression. Through the comparison between Eq. (14) and Eqs. (5), (6), it is clear that the posterior mean of $\mathbf{w}$ is equivalent to the estimated parameter in ridge regression with $\lambda = \alpha\sigma^2$. However, the advantage of the Bayesian framework is two-fold. Firstly, it provides a probabilistic model where the uncertainty associated with both the model parameters and the prediction can be quantified. By utilizing the probabilistic model, the regularization term λ can be interpreted as the ratio between the data variance (σ^2) and the prior variance associated with the model parameters (α^{-1}). Secondly, the value for σ^2 and α can be determined through the direct maximization of the marginal likelihood as presented in Section 2.2, as opposed to the cross-validation where multiple models are developed based on multiple partitions of the training data. The cross-validation procedure is especially undesirable for the determination of continuous terms such as λ, since the range and step-size must be specified to obtain the candidate values for validation. The proper range and step-size are typically identified through trial-and-error, which incurs additional computational cost.

2.4. Regression with multiple responses

This subsection extends the Bayesian linear regression models discussed previously to the case of multivariate regression, where Q response variables are considered. The multivariate linear regression model can be written as

$$\mathbf{y}_n = \mathbf{W}^{\mathrm{T}}\mathbf{x}_n + \boldsymbol{\epsilon}_n \tag{15}$$

where $\mathbf{y}_n = (y_{n1}, \ldots, y_{nQ})^{\mathrm{T}}$, $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_q)$, and $\boldsymbol{\epsilon}_n$ is Q-dimensional zero-mean random noise with Gaussian distribution: $G(\boldsymbol{\epsilon}_n; \mathbf{0}, \boldsymbol{\Sigma})$. Similar to the case of single response variable, $\mathbf{y}_n$ is conditionally Gaussian distributed: $p(\mathbf{y}_n|\mathbf{W}, \boldsymbol{\Sigma}) = G(\mathbf{y}_n; \mathbf{W}^{\mathrm{T}}\mathbf{x}_n, \boldsymbol{\Sigma})$. For the purpose of Bayesian inference, it may be convenient to introduce the matrix Gaussian distribution as prior for $\mathbf{W}$[1,21], and then to integrate out the model parameters to obtain the marginal distribution [21,22].

In this paper a simplified approach is adopted where the response variables are assumed to be independent given the predictor variables, and thus the covariance matrix is diagonal: $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_Q^2)$. This is a commonly utilized assumption to simplify the development of multivariate regression models [26]. Based on this covariance matrix, the likelihood of the data set is

$$p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N}\prod_{q=1}^{Q}G(y_{nq}; \mathbf{w}_q^{\mathrm{T}}\mathbf{x}_n, \sigma_q^2) \tag{16}$$

where the response matrix is $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)^{\mathrm{T}}$. Independent prior distributions are assigned for $\mathbf{w}_q$'s as[1]

$$p(\mathbf{W}|\alpha) = \prod_{q=1}^{Q} G(\mathbf{w}_q; \mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{17}$$

Then the posterior distribution can be derived as follows:

$$p(\mathbf{W}|\mathbf{Y}, \alpha, \boldsymbol{\Sigma}) \propto \left[\prod_{q=1}^{Q} G(\mathbf{w}_q; \mathbf{0}, \alpha^{-1}\mathbf{I}) \right] \left[\prod_{n=1}^{N} \prod_{q=1}^{Q} G(y_{nq}; \mathbf{w}_q^{\mathrm{T}}\mathbf{x}_n, \sigma_q^2) \right]$$

$$= \prod_{q=1}^{Q} \left[G(\mathbf{w}_q; \mathbf{0}, \alpha^{-1}\mathbf{I}) \prod_{n=1}^{N} G(y_{nq}; \mathbf{w}_q^{\mathrm{T}}\mathbf{x}_n, \sigma_q^2) \right] \tag{18}$$

In analogous to Eq. (4) the following result is obtained:

$$p(\mathbf{W}|\mathbf{Y}, \alpha, \boldsymbol{\Sigma}) = \prod_{q=1}^{Q} G(\mathbf{w}_q; \mathbf{m}_q, \mathbf{S}_q) \tag{19}$$

where

$$\mathbf{m}_q = \sigma_q^{-2} \mathbf{S}_q \mathbf{X}^{\mathrm{T}} \mathbf{Y}_q \tag{20}$$

$$\mathbf{S}_q = (\alpha\mathbf{I} + \sigma_q^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{21}$$

and $\mathbf{Y}_q$ is the q-th column vector of $\mathbf{Y}$. It can be seen that the utilization of a diagonal covariance matrix ($\boldsymbol{\Sigma}$) for regression noise leads to a decoupling of the multivariate regression into Q regression problems. However, this decoupling method is not equivalent to developing Q separate models by considering each response variable independently, because the prior distributions of the $\mathbf{w}_q$'s are dependent on the same hyper-parameter α in Eq. (17). Therefore the model presented in this subsection is still referred to as a multivariate regression model. Whether this multivariate regression model is more appropriate than a set of separate models is dependent on specific applications, and it will be discussed with the case studies in Section 5.

Finally, the predictive distribution for each response variable has the same form as Eq. (7) with $\mathbf{m}$, $\mathbf{S}$ and σ^2 being replaced by $\mathbf{m}_q$, $\mathbf{S}_q$ and σ_q^2, respectively.

To determine the values for the hyper-parameters (α and σ_q^2, $q = 1, \ldots, Q$), the Bayesian evidence approximation approach is applied to maximize the following marginal likelihood:

$$p(\mathbf{Y}|\alpha, \boldsymbol{\Sigma}) = \int p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\Sigma})p(\mathbf{W}|\alpha)\,d\mathbf{W}$$

$$= \int \cdots \int \prod_{q=1}^{Q} \left[G(\mathbf{w}_q; \mathbf{0}, \alpha^{-1}\mathbf{I}) \prod_{n=1}^{N} G(y_{nq}; \mathbf{w}_q^{\mathrm{T}}\mathbf{x}_n, \sigma_q^2) \right] d\mathbf{w}_1 \cdots d\mathbf{w}_Q = \prod_{q=1}^{Q} (2\pi)^{-(N/2)} |\mathbf{C}_q|^{-(1/2)} \exp\left(-\frac{1}{2} \mathbf{Y}_q^{\mathrm{T}} \mathbf{C}_q^{-1} \mathbf{Y}_q \right) \tag{22}$$

where $\mathbf{C}_q = \sigma_q^2\mathbf{I} + \alpha^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Similar to the univariate regression case, by maximizing the marginal likelihood, the estimates for the hyper-parameters can be obtained:

$$\alpha = \frac{\sum\limits_{q=1}^{Q} \gamma_q}{\sum\limits_{q=1}^{Q} \mathbf{m}_q^{\mathrm{T}}\mathbf{m}_q} \tag{23}$$

$$\sigma_q^2 = \frac{\sum\limits_{n=1}^{N} (y_{nq} - \mathbf{m}_q^{\mathrm{T}}\mathbf{x}_n)^2}{N - \gamma_q} \tag{24}$$

where γ_q is defined by

$$\gamma_q = M - \alpha\,\mathrm{Tr}(\mathbf{S}_q) \tag{25}$$

The hyper-parameters are estimated iteratively using Eqs. (23)–(25) with the update of $\mathbf{m}_q$ and $\mathbf{S}_q$ as in Eqs. (20) and (21) at each iteration.

3. Bayesian variable selection

As introduced in Section 1, the selection of a subset of predictor variables (wavelengths) may improve the prediction accuracy of the calibration models. Variable selection strategy can be implemented within the Bayesian linear regression model through the modification of the prior distributions. Firstly the situation of univariate regression is considered.

The approach of variable selection is essentially to determine the relative importance of each predictor variable for the prediction. Within the Bayesian framework, this goal can be achieved by assigning individual Gaussian prior distribution for each regression coefficients w_i:

$$p(w_i|\alpha_i) = G(w_i; 0, \alpha_i^{-1}), \qquad i = 1, \ldots, M \tag{26}$$

By assuming the prior distributions for w_i's are independent, the prior for the parameter vector $\mathbf{w}$ can be written as

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} G(w_i; 0, \alpha_i^{-1}) \tag{27}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_M]^{\mathrm{T}}$. The posterior distribution of $\mathbf{w}$ is still Gaussian as in Eq. (4), with the mean being defined in Eq. (5). However, since each regression parameter w_i is assigned a zero-mean Gaussian prior with variance α_i^{-1}, the posterior variance of $\mathbf{w}$ is given by

$$\mathbf{S} = (\mathbf{A} + \sigma^{-2}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \tag{28}$$

where $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$.

Based on the evidence approximation strategy, the hyper-parameters $\boldsymbol{\alpha}$ and β can be obtained through the maximization of the marginal likelihood. Similar to Eq. (8) the marginal likelihood is

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \int p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})\,d\mathbf{w}$$

$$= (2\pi)^{-(N/2)} |\mathbf{C}|^{-(1/2)} \exp\left(-\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{y} \right) \tag{29}$$

where $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathrm{T}}$. Note this marginal likelihood is in similar form as Eq. (8), the difference being the definition of $\mathbf{C}$ as the result of individual prior distributions for w_i's. The detailed derivation to maximize the likelihood is similar to that presented in Section 2.2, thus it will not be repeated here. The resultant iterative algorithm

---

[1] Alternative forms of the hyper-parameter are possible, for example $\alpha_q = k/\sigma_q^2$ and thus the prior for each $\mathbf{w}_q$ is $p(\mathbf{w}_q|\sigma_q^2, k) = G(\mathbf{w}_q; \mathbf{0}, \sigma_q^2\mathbf{I}/k)$. To keep the paper concise, alternative parameterizations are not explored further.

is as follows:

$$\alpha_i = \frac{\gamma_i}{m_i^2} \tag{30}$$

$$\sigma^2 = \frac{\sum_{n=1}^{N}(y_n - \mathbf{m}^T\mathbf{x}_n)^2}{N - \sum_{i=1}^{M}\gamma_i} \tag{31}$$

where m_i is the i-th component of the posterior mean $\mathbf{m}$ (Eq. (5)), and $\gamma_i \in [0, 1]$ is defined by

$$\gamma_i = 1 - \alpha_i S_{ii} \tag{32}$$

where S_{ii} is the i-th diagonal component of the matrix $\mathbf{S}$. The introduction of γ_i's may appear confusing with the notations in Section 2.4, where the subscript q in γ_q's was utilized to denote the q-th response variable. However this should not be a source of confusion if the context is consulted.

The optimal value for $\boldsymbol{\alpha}$ and σ^2 is determined by alternating Eqs. (30)–(32) until convergence. Variable selection can be realized by noting that the marginal posterior distribution of regression coefficient w_i is of the form: $p(w_i|\mathbf{y}, \alpha_i, \sigma^2) = G(w_i; m_i, s_i)$, where

$$m_i = s_i\sigma^{-2}\sum_{n=1}^{N}x_{ni}y_n \tag{33}$$

$$s_i = \left(\alpha_i + \sigma^{-2}\sum_{n=1}^{N}x_{ni}^2\right)^{-1} \tag{34}$$

and x_{ni} is the i-th element of the predictor vector $\mathbf{x}_n$. This result can be obtained based on Eqs. (4), (5) and (28). Therefore if during the training iterations, α_i tends to become a large value, the corresponding w_i can only be close to zero (i.e. $m_i \approx 0$ and $s_i \approx 0$), and thus the corresponding predictor has little impact on the prediction. Practically the i-th variable can be removed from the regression model if α_i's is larger than a threshold. Preliminary studies in this work have found that neither the selected variables, nor the predictive performance of the final model, is sensitive to this threshold, as long as it is sufficiently large. The appropriate value for the threshold also depends on the specific scale of the predictors $\mathbf{x}$. We suggest to follow the standard method by scaling the data to have zero mean and unit standard deviation at each variable, and then to empirically select a threshold such that the selected variables and log likelihood value are not susceptible to small change in this threshold. Based on this rationale, we found in the application study (to be presented subsequently) that a threshold between 10^5 and 10^{15} gives very similar results, and we used the value of 10^6 to produce the results in this paper.

The Bayesian linear regression model with the individual prior being defined for each parameter falls within the family of automatic relevance determination (ARD) models. The idea of ARD models was developed by MacKay [32] and Neal [33]. By associating each variable with a hyper-parameter α_i, the magnitude of the relevance of the corresponding variable to the prediction can be determined. Recently the ARD approach has been used to select relevant basis functions in regression models, resulting in the "relevance vector machine" [30].

The variable selection strategy may be extended for regression with multiple response variables. The detailed algorithm is straightforward to derive by following the discussion in Section 2.4, and thus it is neglected here. However care must be taken when variable selection is implemented in the case of multivariate responses. For different response variables, the corresponding informative predictors are typically different. Hence by considering all response variables simultaneously, the selected variables may not be optimal in terms of predicting each analyte property. Therefore for the purpose of variable selection, it may be more desirable to develop separate calibration models for each response variable. This issue will be discussed in the application study in Section 5.

4. A Gaussian process view

The form of the marginal likelihood given in Eq. (8) implies that the Bayesian linear regression falls into the family of Gaussian process models [33,32,34], since $p(\mathbf{y}|\alpha, \sigma^2) = G(\mathbf{y}; \mathbf{0}, \mathbf{C})$. Similarly the model with variable selection in Eq. (29) is also a Gaussian process. As the regression model can be summarized by integrating over parameters $\mathbf{w}$, the Gaussian process is a non-parametric regression technique.

One of the major advantages of Gaussian process models is their flexibility in handling different regression problems. The previous sections have shown that by modifying the prior distribution of regression parameters, the resultant model can achieve the goal of variable selection. Note the conventional Bayesian linear model in Eq. (8) differs from that for variable selection in Eq. (29) only in terms of the covariance matrix $\mathbf{C}$. In fact the Gaussian process model can take various forms by directly manipulating the covariance matrix. The element of $\mathbf{C}$, C_{uv}, is the covariance between the u-th and v-th data points, and it can be defined in terms of a "covariance function" as $C_{uv} = C(\mathbf{x}_u, \mathbf{x}_v)$. For example the following covariance function has been widely reported in the literature with promising modeling capability [12,35,36]:

$$C(\mathbf{x}_u, \mathbf{x}_v) = a_0 + a_1\sum_{m=1}^{M}x_{um}x_{vm}$$
$$+ v_0\exp\left(-\sum_{m=1}^{M}\eta_m(x_{um} - x_{vm})^2\right) + \sigma^2\delta_{uv} \tag{35}$$

where $\delta_{uv} = 1$ if $u = v$, otherwise $\delta_{uv} = 0$. This covariance function is flexible in accounting for different aspects of the data [12]. The first two terms represent constant bias (offset) and linear correlation respectively; the exponential term accounts for the non-linear relation, and $\sigma^2\delta_{uv}$ captures the random noise effect.

Let $\boldsymbol{\theta} = (a_0, a_1, v_0, \eta_1, \ldots, \eta_M, \sigma^2)$ denote the hyper-parameters of the Gaussian process model with covariance function being defined in Eq. (35). Due to the complex form of the covariance function, the task of maximizing the marginal likelihood is not straightforward, and in practice the likelihood function has been observed to exhibit many local optimum values [34]. Typically a conjugate gradient approach is used for the optimization, and the algorithm requires to be run multiple times with different initial hyper-parameter values to alleviate the local optima issue [34]. Alternatively Markov chain Monte Carlo (MCMC) simulation can be utilized for the inference of the hyper-parameters. Carefully designed MCMC approach can overcome the local optima problem; however, its computation is significantly higher than the Bayesian evidence approximation method.

This paper is focused on the application of Bayesian linear regression models developed in Sections 2 and 3. For a more detailed discussion of Gaussian process and its application in spectroscopic calibration, the readers are referred to [12].

5. Application to spectroscopic calibration

In this section, two application studies are presented to evaluate the predictive performance of the Bayesian calibration models. For
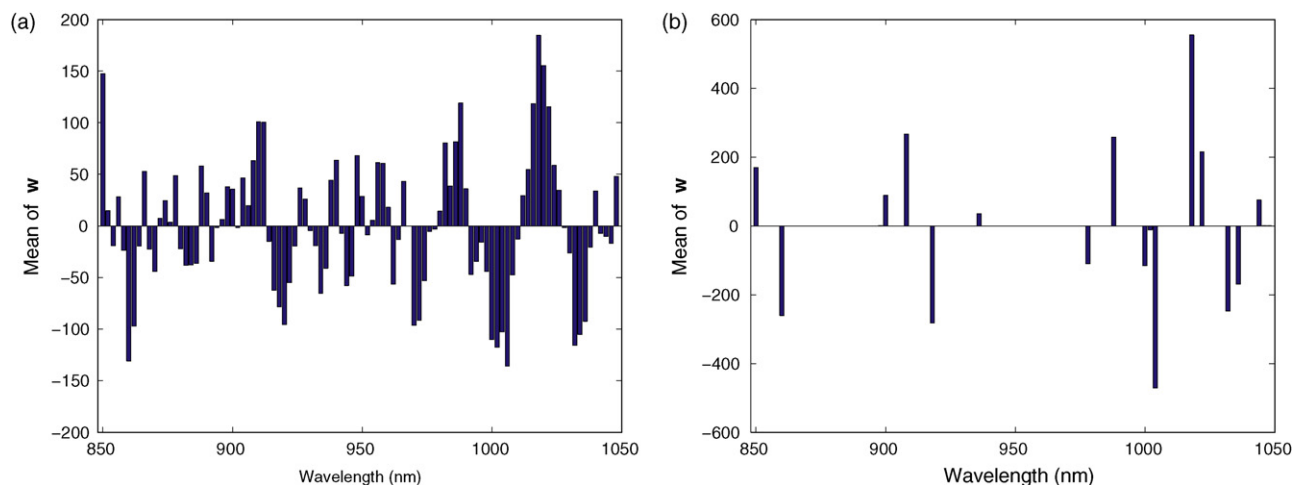
Fig. 1. Posterior mean of regression parameter **w** for the wheat data set: (a) Bayesian linear regression model; (b) Bayesian linear regression model with variable selection, where 16 predictors (wavelengths) were retained.

brevity the Bayesian linear regression model discussed in Section 2 is denoted by "BLR", whilst the BLR model with variable selection presented in Section 3 is termed "BLR-VS". The results from PLS are also quoted for the purpose of comparison. In both examples the data is pre-processed to have zero mean and unit standard deviation at each variable.

To alleviate the effect of multiple local optima, both BLR and BLR-VS models were trained from 10 random starting points for all the case studies reported in this paper. The random initial values were generated from a log-normal distribution (to ensure positive α and σ^2) with mean -3 and standard deviation 3, where the specific values of mean and standard deviation are selected based on related research in Gaussian processes [12,36] to have a fairly wide range for starting points. The 10 random starting points often give several very different estimates of α (or α_i's for variable selection) and σ^2. In terms of resulting likelihood, the best local optimum is typically several times or even orders of magnitude more probably than other local optima. Selecting the best local optimum as the final model for future prediction purpose is widely accepted in the literature [26,30,34,36] for these multi-modal optimization tasks.

5.1. Prediction of protein in wheat kernels

The first application is to develop calibration models for near-infrared (NIR) transmittance spectra recorded for the analysis of wheat kernels [7]. The objective of the study was to determine the percentage protein concentration in the wheat kernels, based on the NIR spectra recorded at 100 wavelengths across the region 850–1050 nm. The data set was divided into a training set of 415 samples and a test set of 108 samples. This is a challenging data set in terms of the development of calibration models, because there are significant variations within the data, including different sample varieties and different locations where the samples were collected. Furthermore, the test samples were stored for additional two months prior to analysis to evaluate the temporal effect on the samples and instruments. Further details about the data set can be found in [7]. The data set is publicly available from http://www.models.kvl.dk/research/data/wheat_kernels/index.asp.

To develop the calibration models, the CPU time was 0.8 s and 1.2 s for BLR and BLR-VS respectively, whilst a single PLS training process took 0.2 s. (The algorithms were implemented in Matlab and were executed on a Pentium-4 3.0 GHz computer running under Windows XP.) Note the above computational cost does not include the effect of multiple random starts for BLR/BLR-VS or

cross-validation procedure for PLS. Although the CPU time for the Bayesian models is longer than that for PLS, the difference is not critical for many application scenarios.

Fig. 1 gives the posterior mean of the regression parameter **w** (see Eq. (4)) for the BLR and BLR-VS models. In the BLR-VS model, 16 out of the total 100 wavelengths were selected, and thus the model is of less complexity. It is interesting to see that the locations of the selected variables (Fig. 1(b)) roughly match the locations of the regression coefficients with large magnitude in Fig. 1(a). This may be an indication that the BLR-VS model is capable of summarizing the information contained at the whole spectral region through the selection of a subset of representative and informative variables.

The predictive performance is evaluated using root mean squares error for prediction (RMSEP). The baseline RMSEP from PLS is 0.70. The BLR model gives a lower RMSEP of 0.62; if variable selection is applied, the best result of 0.55 is obtained by the BLR-VS model with only 16 predictors being selected. The prediction versus reference plot is shown in Fig. 2. For clarity only the results of PLS and the BLR-VS model are shown. Fig. 2 clearly illustrates that the Bayesian model with variable selection has achieved superior predictive performance to that of PLS on this data set.
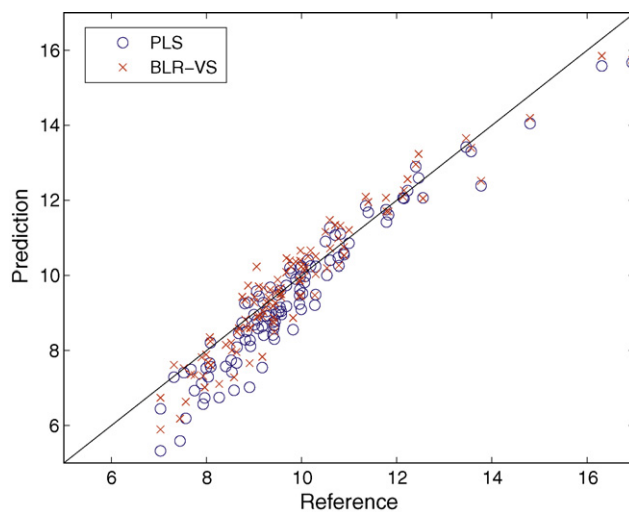

Fig. 2. Predictions using PLS (RMSE = 0.70) and BLR-VS (RMSE = 0.55) for the wheat data.

5.2. Prediction of properties of corn samples

The second data set consists of 80 corn samples measured on three different NIR spectrometers denoted as "m5", "mp5" and "mp6" respectively. The wavelength range is 1100–2498 nm at 2 nm intervals, resulting in 700 predictor variables in the spectra. The objective of this analysis is to predict the moisture, oil, protein and starch content of the corn samples. In this study the data was randomly divided into a training and a test set, each having 40 samples. This data set is available from the Eigenvector Research (http://software.eigenvector.com/Data/Corn/index.html), and it has been used in [8,37] for the development of calibration models.

In the presence of four response variables, there are two options for the development of calibration models: a single multivariate model or a set of separate models for each response variable. Table 1 gives the prediction results of the two methods, where the data was analyzed using the spectrometer "m5". It can be seen that for the PLS models, there is no consensus on whether multi-response model (PLS2) can achieve lower RMSEP than separate models (PLS1). (This phenomenon has been noted in the literature, e.g. [12].) In addition separate BLR models appear to outperform the single multivariate BLR model in the prediction of all four analyte properties. It could be argued that by adopting the diagonal covariance matrix Σ as in Section 2.4, significant information contained in the correlation structure between the response variables is ignored. However, utilizing a full covariance matrix would

Table 1
Prediction results (RMSEP) for the corn data set using the spectrometer "m5": (a) separate models for each response variable; (b) a single model for multivariate responses. The value inside the brackets for the model BLR-VS denotes the number of selected variables.

Model	Moisture	Oil	Protein	Starch
(a)				
PLS	0.036	0.154	0.317	0.261
BLR	0.018	0.140	0.127	0.230
BLR-VS	0.004 (2)	0.101 (23)	0.072 (27)	0.211 (21)
(b)				
PLS	0.046	0.206	0.144	0.408
BLR	0.038	0.142	0.143	0.330
BLR-VS	0.005 (10)	0.127 (10)	0.130 (10)	0.311 (10)

make the Bayesian inference significantly more difficult, and a computational intensive Monte Carlo simulation approach would be required [12]. Furthermore, adopting separate modeling approach may be conceptually appealing if variable selection is applied, since the informative predictors (the region of sensitive wavelengths in the context of spectroscopy) are typically different for different analyte properties. By developing separate BLR-VS models, the number of selected variables (Table 1(a)) and the corresponding regression parameters (Fig. 3) are distinctive across the four response variables. The RMSEPs of separate BLR-VS models are significantly lower than those of a single multivariate BLR-VS model with the same 10 variables being selected for the prediction of all responses.
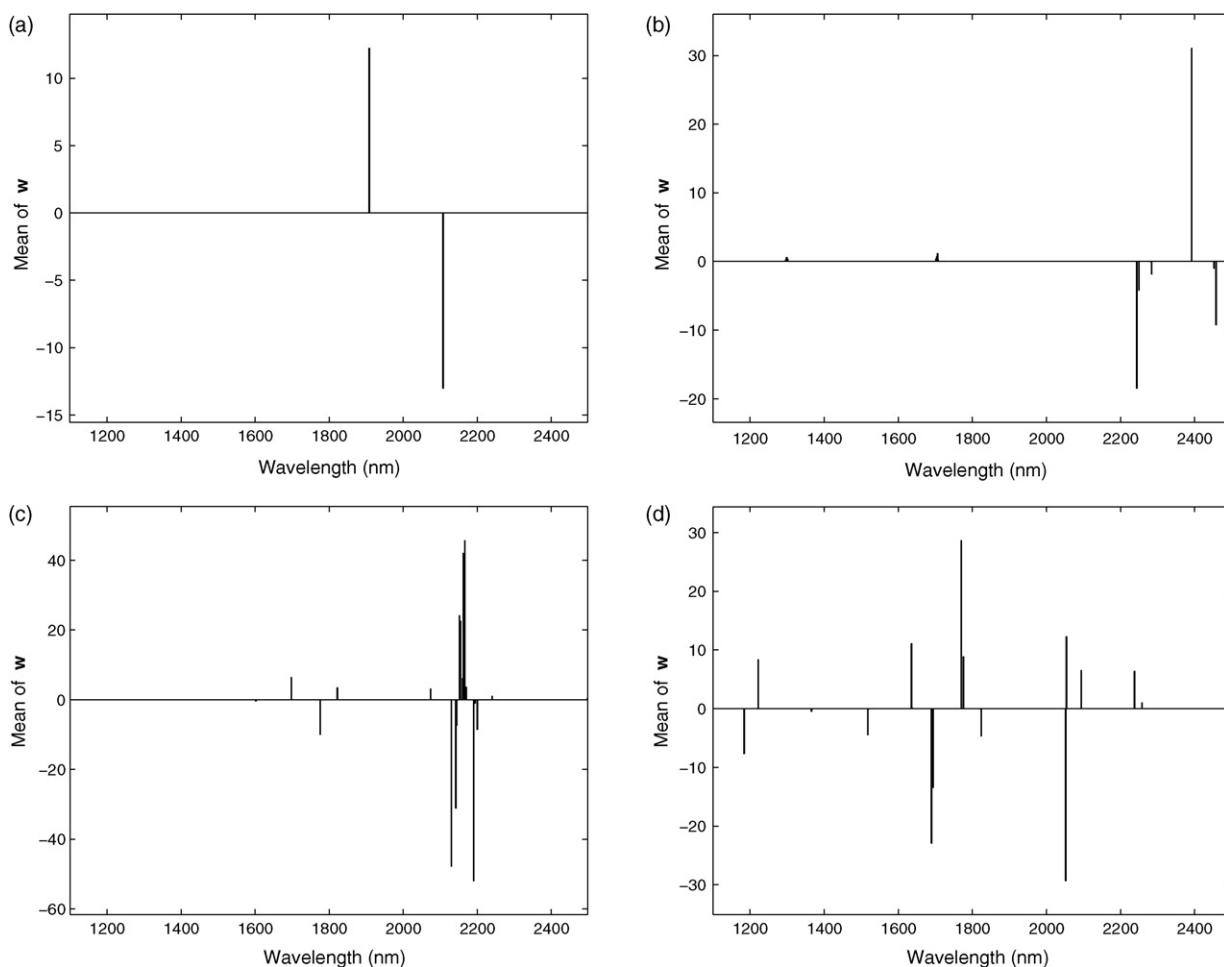

Fig. 3. Posterior mean of regression parameter **w** obtained by variable selection for the corn samples analyzed by the spectrometer "m5". Separate BLR-VS model is developed for each response variable: (a) moisture, (b) oil, (c) protein and (d) starch, where the number of selected variables are 2, 23, 27 and 21, respectively.
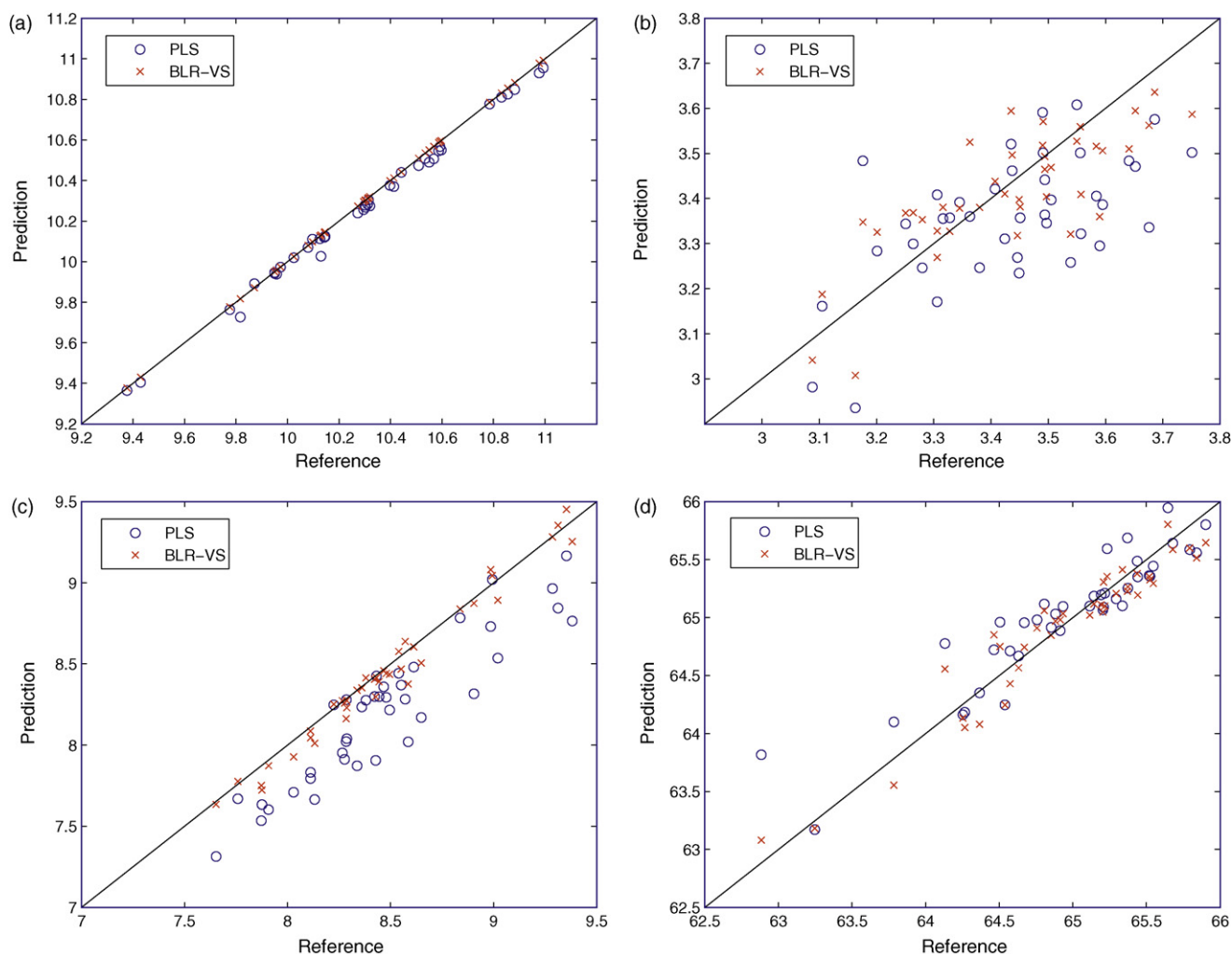
Fig. 4. Prediction vs. reference plot using PLS and Bayesian linear regression model with variable selection (BLR-VS) for the corn data analyzed by the spectrometer "m5". Separate models are developed for each response variable: (a) moisture, (b) oil, (c) protein, (d) starch.

Based on the above discussions, the separate modeling approach is utilized in the rest of this paper.

Table 1(a) also demonstrates the improved prediction performance of Bayesian regression models, both BLR and BLR-VS, in comparison with PLS. The results of BLR-VS model is especially promising, since it attained the lowest RMSEP for all analyte properties with only a small subset of predictors being selected. Fig. 4 illustrates the prediction versus reference plot for PLS and the BLR-VS model for the prediction of all four components, and it confirms the superior predictive performance of the BLR-VS model.

Finally Tables 2 and 3 summarize the prediction results for the corn samples analyzed using the other two spectrometers ("mp5" and "mp6" respectively). It appears that these two instruments are generally less accurate than the spectrometer "m5" demonstrated previously: the RMSEPs of all models are considerably higher than

Table 2
Prediction results (RMSEP) for the corn data set using the spectrometer "mp5". Separate calibration models are developed for each response variable. The value inside the brackets for the model BLR-VS denotes the number of selected variables.

Model	Moisture	Oil	Protein	Starch
PLS	0.184	0.207	0.365	0.698
BLR	0.190	0.172	0.297	0.556
BLR-VS	0.164 (20)	0.153 (32)	0.237 (20)	0.531 (22)

Table 3
Prediction results (RMSEP) for the corn data set using the spectrometer "mp6". Separate calibration models are developed for each response variable. The value inside the brackets for the model BLR-VS denotes the number of selected variables.

Model	Moisture	Oil	Protein	Starch
PLS	0.280	0.196	0.262	0.836
BLR	0.211	0.192	0.218	0.559
BLR-VS	0.194 (18)	0.180 (17)	0.198 (30)	0.508 (23)

those obtained from "m5". Again, the improvement of the Bayesian calibration models over PLS is manifest, and the variable selection strategy achieved the best prediction results on all four properties of interest.

6. Conclusions

This paper presents a Bayesian linear regression approach to the development of spectroscopic calibration models, and further introduces a variable selection strategy in conjunction of the proposed calibration models. The Bayesian models attain both theoretical advantages and practical improvement over the standard PLS method in terms of lower prediction errors. The variable selection strategy is particularly promising, since it not only improves the prediction accuracy but also provides clear interpretation as to

which wavelengths are the most informative to infer the analyte properties of interest.

A Matlab implementation of the Bayesian linear regression models discussed in this paper is available from http://www.ntu.edu.sg/home/chentao/.

References

[1] P.J. Brown, Measurement, Regression, and Calibration, Oxford University Press, 1993.
[2] P. Geladi, B.R. Kowalski, Analytica Chimica Acta 185 (1986) 1.
[3] T. Næs, H. Martens, Journal of Chemometrics 2 (1988) 155.
[4] S. Wold, M. Sjöström, L. Eriksson, Chemometrics and Intelligent Laboratory Systems 58 (2001) 109.
[5] E. Vigneau, M.F. Devaux, E.M. Qannari, P. Robert, Journal of Chemometrics 11 (1997) 239.
[6] H. Martens, J.P. Nielsen, S.B. Engelsen, Analytical Chemistry 75 (2003) 394.
[7] D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen, Applied Spectroscopy 56 (2002) 1206.
[8] J.A. Westerhuis, S. de Jong, A.K. Smilde, Chemometrics and Intelligent Laboratory Systems 56 (2001) 13.
[9] F. Despagne, D.-L. Massart, P. Chabot, Analytical Chemistry 72 (2000) 1657.
[10] H.H. Thodberg, IEEE Transactions on Neural Networks 7 (1996) 56.
[11] U. Thissen, B. Ustun, W.J. Melssen, L.M.C. Buydens, Analytical Chemistry 76 (2004) 3099.
[12] T. Chen, J. Morris, E. Martin, Chemometrics and Intelligent Laboratory Systems 87 (2007) 59.
[13] A. Bangalore, R. Shaffer, G. Small, M. Arnold, Analytical Chemistry 68 (1996) 4200.
[14] M. McShane, G. Cote, C. Spiegelman, Applied Spectroscopy 51 (1997) 1559.
[15] S. Osborne, R. Jordan, R. Kunnemeyer, Analyst 122 (1997) 1531.
[16] V. Centner, D.-L. Massart, O. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anlytical Chemistry 68 (1996) 3851.
[17] F. Despagne, D.-L. Massart, Chemometrics and Intelligent Laboratory Systems 40 (1998) 145.
[18] D. Broadhurst, R. Goodacre, A. Jones, J.J. Rowland, D.B. Kelp, Analytica Chimica Acta 348 (1997) 71.
[19] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Chemometrics and Intelligent Laboratory Systems 69 (2003) 3.
[20] M.N. Nounou, B.R. Bakshi, P.K. Goel, X. Shen, AIChE Journal 48 (2002) 1775.
[21] P.J. Brown, M. Vannucci, T. Fearn, Journal of the Royal Statistical Society B 60 (1998) 627.
[22] P.J. Brown, M. Vannucci, T. Fearn, Journal of Chemometrics 12 (1998) 173.
[23] P.J. Brown, T. Fearn, M. Vannucci, Journal of the American Statistical Association 96 (2001) 398.
[24] P.J. Brown, M. Vannucci, T. Fearn, Journal of the Royal Statistical Society B 64 (2002) 519.
[25] M.C. Denham, Journal of Chemometrics 11 (1997) 39.
[26] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[27] K. Mardia, J. Kent, J. Bibby, Multivariate Analysis, Academic Press, London, 1979.
[28] A.B. Gelman, J.S. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall/CRC, 1995.
[29] D.J.C. MacKay, Neural Computation 4 (1992) 415.
[30] M.E. Tipping, Journal of Machine Learning Research 1 (2001) 211.
[31] J.B. Forrester, J.H. Kalivas, Journal of Chemometrics 18 (2004) 372.
[32] D.J.C. MacKay, C.M. Bishop, in: Neural Networks and Machine Learning, volume 168 of F: Computer and Systems Sciences, NATO Advanced Study Institute, Springer, Berlin, Heidelberg, 1998, p. 133.
[33] R.M. Neal, Bayesian Learning for Neural Networks, Springer-Verlag, New York, 1996.
[34] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.
[35] X. Ou, E. Martin, Neural Computing and Applications 17 (2008) 471.
[36] J.Q. Shi, R. Murray-Smith, D.M. Titterington, International Journal of Adaptive Control and Signal Processing 17 (2003) 1.
[37] H. Tan, S.D. Brown, Journal of Chemometrics 17 (2003) 111.