A Bayesian Framework for Real-Time Identification of Locally Weighted Partial Least Squares

Ming Ma, Shima Khatibisepehr, and Biao Huang
Dept. of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2G6

DOI 10.1002/aic.14663
Published online November 7, 2014 in Wiley Online Library (wileyonlinelibrary.com)

Just-in-time (JIT) learning methods are widely used in dealing with nonlinear and multimode behavior of industrial processes. The locally weighted partial least squares (LW-PLS) method is among the most commonly used JIT methods. The performance of LW-PLS model depends on parameters of the similarity function as well as the structure and parameters of the local PLS model. However, the regular LW-PLS algorithm assumes that the parameters of the similarity function and structure of the local PLS model are known and do not fully utilize available knowledge to estimate the model parameters. A Bayesian framework is proposed to provide a systematic way for real-time parameterization of the similarity function, selection of the local PLS model structure, and estimation of the corresponding model parameters. By applying the Bayes' theorem, the proposed framework incorporates the prior knowledge into the identification process and takes into account the different contribution of measurement noises. Furthermore, Bayesian model structure selection can automatically deal with the model complexity problem to avoid the overfitting issue. The advantages of this new approach are highlighted through two case studies based on the real-world near infrared data. © 2014 American Institute of Chemical Engineers *AIChE J*, 61: 518–529, 2015
Keywords: Bayesian inference, just-in-time modeling, partial least squares, latent variable, near infrared

Introduction

Process modeling is one of the most important elements in development and implementation of advanced process monitoring and control techniques. The representativeness of process models has a significant effect on the performance of these techniques. Linear modeling techniques are commonly used to identify a model from the process variables. Ordinary least squares (OLS) regression is one of the most widely used classical modeling techniques due to its simplicity. The main assumption behind the OLS regression is that the process variables are not strongly dependent of each other. Principal component regression (PCR) and partial least squares (PLS) regression have noticeable advantages over the OLS regression in dealing with the collinearity issue.[1–4] The PCR first uses orthogonal transformation to convert correlated input variables into a set of uncorrelated, lower dimensional principal components. Next, the OLS is applied to reveal the parametric relationship between the principal components and the output variables. The orthogonal transformation used in the PCR only considers the relationships among input variables and fails to take into account any information about the output variables. Therefore, it may result in an ill-conditioned alignment.[5] The PLS regression overcomes this shortcoming by taking into account both input and output variables for finding the principal components.[6] The performance of these linear techniques will be satisfactory only if the underlying process can be assumed to be linear. To deal with the process which exhibits certain form of nonlinear behavior, several approaches have been proposed to integrate nonlinear features with the linear PLS framework, thus result in nonlinear PLS (NLPLS) algorithms such as quadratic PLS,[7] neural network PLS,[8] and fuzzy PLS.[9] These approaches retain the linear latent structure of PLS model. In the light of nonlinear principal component, Malthouse et al.[10] proposed a new approach named NLPLS to extract the nonlinear latent structures. However, these NLPLS approaches which provide global models to describe the data from different operation modes may not achieve satisfactory performance. Considering these issues, the locally weighted partial least squares (LW-PLS) regression can be used.[11] LW-PLS combines the nature of locally weighted regression and PLS so that it can deal with the nonlinearity, multimode behavior as well as the collinearity. In the LW-PLS method, local PLS models are built around each operating point through local calibration samples. To construct an LW-PLS model, the following aspects should be considered:

1. Selection of local calibration samples: Local calibration samples are often selected or prioritized using a certain similarity function. The similarity function takes into account the distance between a query sample and calibration ones. The similarity function is parameterized by a set of localization parameters which needs to be specified to control how steeply the similarity will decrease by increasing the distance. In this way, the localization parameters would greatly affect the selection or priorotization of local calibration samples.

2. Selection of model structure: After choosing or prioritizing proper local calibration samples, the next step is to

Correspondence concerning this article should be addressed to B. Huang at biao.huang@ualberta.ca.

choose a proper model structure. This could be equivalent to determining the dimensionality of the latent space that can best describe the underlying behavior of the process.

3. Estimation of model parameters: Having selected the local calibration samples and determined the model structure, model parameters can be identified via the LW-PLS algorithm.

Therefore, the problem of identification of an LW-PLS model boils down to obtaining the optimal combination of localization parameters, model structure, and model parameters. The common practice is to search for the globally optimal combination of localization parameters and model structure by minimizing the root mean square error of cross-validation (RMSECV).[12] This approach is often computationally inefficient for online identification of the LW-PLS models and may also result in the overfitting issue.[13,14] Khatibisepehr et al. (submitted) has developed an off-line identification method to find the locally optimal localization parameters and model structure within a known operating space using a hierarchical Bayesian optimization framework. The idea behind this method is to first partition the operating space into a finite number of subspaces and then find the optimal combination of localization parameters and model structure for each subspace. The application of Bayes' theorem makes it possible to incorporate the prior knowledge over the localization parameters and model structures. The proposed Bayesian framework can also deal with the model complexity control to avoid overfitting. However, this method has the following shortcoming: (1) it does not utilize the prior knowledge over the main parameters for modeling, (2) like all the other existing methods, uncertainties in the parameter estimates are not taken into account in selection of the model structure and tuning of the localization parameters, (3) due to the multimode behavior of industrial processes, a finite number of subspaces may not cover the entire operating space especially over a long period.

Therefore, it is desired to tune the localization parameters, select the model structure, and estimate the model parameters all in a real-time phase. The main contribution of this article is to develop a novel integrated identification method to find the locally optimal combination of the model parameters, localization parameters, and model structure in a real-time manner to take full advantage of Bayesian methods. The real-time identification problem of interest is formulated under a holistic Bayesian framework consisting of consecutive levels of optimization. The resulting optimization problem is hierarchically decomposed and a layered optimization strategy is implemented. To obtain explicit solutions, an iterative hierarchical Bayesian approach is adopted to coordinate the solutions obtained in subsequent layers of optimization. The proposed method has the following advantages over the existing ones: (1) the developed hierarchical Bayesian framework offers a systematic way to select the model structure, determine the localization parameters as well as estimate the model parameters. (2) External information over the main parameters, localization parameters, and model structure can be incorporated in the identification process. (3) Sparsity and heteroscedasticity of training samples can be effectively handled. (4) Bayesian inference at a particular level takes into account the uncertainty in the estimates of the previous level. (5) Bayesian model selection can automatically penalize the model complexity to avoid the overfitting issue.[15]

The remainder of this article is organized as follows: the following section introduces the basic formulation of the LW-PLS model and discusses the limitations of the regular LW-PLS modeling method which lead to consideration of an integrated Bayesian framework. Then, the motivation behind adopting a hierarchical approach is outlined and each level of inference is explained in detail. Next, an overall procedure to implement the proposed hierarchical framework is shown. Two industrial case studies are considered to demonstrate the effectiveness of the proposed method based on the real-world near infrared (NIR) spectroscopy data. Finally, the article is summarized by concluding remarks.

Problem Statement

Suppose, we have a training (calibration) dataset with N samples denoted by

$$\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_N]^T \tag{1}$$

$$\mathbf{y}=[y_1,y_2,\ldots y_N]^T \tag{2}$$

$\mathbf{X}\in\mathbb{R}^{N\times M}$ and $\mathbf{y}\in\mathbb{R}^{N\times 1}$ are the input and output matrices, respectively. The ith sample consists of a vector of inputs, $\mathbf{x}_i=[x_{i1},x_{i2}\ldots x_{iM}]^T$, and an output, y_i, where M is the number of input variables. The formulation of the PLS model is given by

$$\mathbf{X}=\mathbf{T}\mathbf{P}^T+\mathbf{E}_X \tag{3}$$

$$\mathbf{y}=\mathbf{T}\mathbf{q}^T+\mathbf{e}_y \tag{4}$$

where $\mathbf{T}\in\mathbb{R}^{N\times H}$ denotes a matrix of latent variables, $\mathbf{P}\in\mathbb{R}^{M\times H}$ is a matrix of loadings and $\mathbf{q}\in\mathbb{R}^{1\times H}$ is a vector of regression coefficients. $\mathbf{E}_X\in\mathbb{R}^{N\times M}$ and $\mathbf{e}_y\in\mathbb{R}^{N\times 1}$ denote the matrices of input and output residuals, respectively.

LW-PLS is an online identification method which builds a local PLS model for each query sample. Given a query sample $\mathbf{x}_q$, a similarity matrix is constructed to prioritize the calibration samples

$$\mathbf{S}_q=\mathrm{diag}(s_{1|q},s_{2|q},\ldots,s_{N|q}) \tag{5}$$

where $s_{i|q}(i=1,2\ldots N)$ is the similarity between $\mathbf{x}_q$ and $\mathbf{x}_i$.

Generally, a measurement of similarity is defined based on a notion of distance between $\mathbf{x}_q$ and $\mathbf{x}_i$. One of the widely used similarity functions is

$$s_{i|q}=\exp\left(-\frac{d_i}{\sigma_d\lambda}\right) \tag{6}$$

$$d_i=\sqrt{(\mathbf{x}_i-\mathbf{x}_q)^T(\mathbf{x}_i-\mathbf{x}_q)} \tag{7}$$

where d_i is the Euclidean distance between $\mathbf{x}_q$ and $\mathbf{x}_i$, σ_d is the standard deviation of $\mathbf{d}=\{d_1,d_2,\ldots d_N\}$ and λ is the localization parameter. Given a σ_d, the similarity decreases more steeply by increasing the distance for larger values of λ. So, λ can determine the acceptable region for selecting the local calibration samples together with σ_d.

LW-PLS models can be constructed by following Algorithm I in Appendix (Khatibisepehr et al., submitted). However, this regular LW-PLS algorithm implicitly assumes that the number of latent variables H, that is, model structure, and localization parameter λ are given. In reality, these parameters are often unknown and have critical effects on the estimation accuracy. Even though proper combination of the model structure and localization parameter can be found in advance using RMSECV, this method cannot maintain good estimation accuracy in a longer term. Multimode behavior of processes and nonlinearity of underlying

mechanisms affect not only the model parameters, but also the model structure and similarity function. Furthermore, the available prior knowledge cannot be incorporated in the identification process using the regular LW-PLS algorithm.

Considering these points, in this work, a new similarity function is defined as

$$s_{i|q}(\varphi)=\exp(-d_i\varphi) \tag{8}$$

where the localization parameter is denoted by φ and treated as a hyperparameter of similarity function to be tuned for each local model. Compared with the similarity function in the regular LW-PLS (Eq. 6), in the new similarity function the term $\frac{1}{\sigma_d/2}$ has been substituted by φ. In this way, the acceptable region of calibration samples can be directly controlled by tuning φ.

The formulation of the PLS model remains the same as given by Eqs. 3 and 4. The number of retained latent variables H is treated as an unknown variable to be estimated. Therefore, the problem of identifying an LW-PLS model consists of the following steps: (1) prioritizing calibration samples, that can be equally achieved by properly tuning the localization parameter φ; (2) choosing the model structure or number of retained latent variables H; and (3) estimating the main parameters $\Theta=\{\mathbf{P},\mathbf{T},\mathbf{q}\}$, that is, loading matrix $\mathbf{P}$, latent variable matrix $\mathbf{T}$, regression coefficient vector $\mathbf{q}$.

From a Bayesian perspective, the problem is converted to maximizing the joint posterior distribution of main parameters, localization parameter, and model structure that is defined as the conditional probability distribution of these variables given the training dataset and query sample, that is, $p(\Theta,\varphi,H|\mathbf{X},\mathbf{y},\mathbf{x}_q)$.

Hierarchical Bayesian Optimization Framework

A Bayesian approach to identify an LW-PLS model is to maximize the posterior probability density function of the main parameters, localization parameter, and model structure, $p(\Theta,\varphi,H|\mathbf{X},\mathbf{y},\mathbf{x}_q)$. Because of the difficulties associated with the direct maximization of $p(\Theta,\varphi,H|\mathbf{X},\mathbf{y},\mathbf{x}_q)$, the problem of interest can be formulated and solved under an iterative hierarchical Bayesian optimization framework.[16] First, the chain rule of probability theory is used to expand the joint posterior probability distribution as

$$p(\Theta,\varphi,H|\mathbf{X},\mathbf{y},\mathbf{x}_q)=p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)p(H|\mathbf{X},\mathbf{y},\mathbf{x}_q) \tag{9}$$

Next, the optimization problem is decomposed hierarchically into following three layers

$$\begin{aligned}&\max_{\Theta,\varphi,H}p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)p(H|\mathbf{X},\mathbf{y},\mathbf{x}_q)\\&=\max_H\{p(H|\mathbf{X},\mathbf{y},\mathbf{x}_q)\{\max_\varphi p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)\\&\times\max_\Theta\{p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)\}\}\}\end{aligned} \tag{10}$$

Inference of main parameters

Applying Bayes' rule, the posterior probability density function (PDF) of main parameters can be written as

$$\begin{aligned}p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)&=\frac{p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|\varphi,H,\mathbf{x}_q)}{p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)}\\&\propto p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|\varphi,H,\mathbf{x}_q)\end{aligned} \tag{11}$$

where $p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)$ is a normalizing constant.

As prior, it is reasonable to assume that the main parameters are independent of the localization parameter and query sample. The prior can be explicitly expressed as the conditional joint probability of the loading matrix, regression coefficient vector, and latent variable matrix given the model structure

$$\begin{aligned}p(\Theta|\varphi,H,\mathbf{x}_q)&=p(\Theta|H)\\&=p(\mathbf{P},\mathbf{T},\mathbf{q}|H)\\&=p(\mathbf{T}|\mathbf{P},\mathbf{q},H)p(\mathbf{q}|\mathbf{P},H)p(\mathbf{P}|H)\end{aligned} \tag{12}$$

Given the loading matrix $\mathbf{P}$, it is reasonable to assume that $\mathbf{T}$ and $\mathbf{q}$ are independent, that is, $p(\mathbf{T}|\mathbf{P},\mathbf{q},H)=p(\mathbf{T}|\mathbf{P},H)$. Thus, the posterior PDF of main parameters can be explicitly written as

$$\begin{aligned}p(\mathbf{P},\mathbf{T},\mathbf{q}|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)&\propto p(\mathbf{X},\mathbf{y}|\mathbf{P},\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)\\&\times p(\mathbf{T}|\mathbf{P},H)p(\mathbf{q}|\mathbf{P},H)p(\mathbf{P}|H)\end{aligned}$$

Following the approach of Ref. 17, a new Bayesian approach to solve the problem of LW-PLS modeling is proposed in this section.

For each calibration sample, the LW-PLS formulation is given by

$$\mathbf{x}_i=\mathbf{P}\mathbf{t}_i+\mathbf{e}_{xi} \tag{14}$$

$$y_i=\mathbf{q}\mathbf{t}_i+e_{yi} \tag{15}$$

The noise-free inputs and output are given by

$$\tilde{\mathbf{x}}_i=\mathbf{P}\mathbf{t}_i \tag{16}$$

$$\tilde{y}_i=\mathbf{q}\mathbf{t}_i \tag{17}$$

The loading matrix $\mathbf{P}$ has the following unit orthogonal constraint

$$\mathbf{P}^T\mathbf{P}=\mathbf{I} \tag{18}$$

A vector of model parameters, $\mathbf{b}\in\mathbb{R}^{M\times 1}$, representing the relationship between the input and output variables, is defined as

$$\mathbf{b}=\mathbf{P}\mathbf{q}^T \tag{19}$$

The likelihood function relies on the nature of noise. Assume that the input and output measurements are contaminated by mutually independent Gaussian noise, $\mathbf{e}_{xi}$ and e_{yi}, with known variance $\mathbf{Q}_{e_x}$ and Q_{e_y}.[17] The estimation of these unknown variances will be discussed shortly. Given a query sample $\mathbf{x}_q$, the importance weight assigned to the ith calibration sample is denoted by $s_{i|q}$. This is equivalent to saying that

$$\mathbf{Q}_{e_{xi}}=\frac{\mathbf{Q}_{e_x}}{s_{i|q}} \tag{20}$$

$$Q_{e_{yi}}=\frac{Q_{e_y}}{s_{i|q}} \tag{21}$$

Normally, a calibration sample with large weight is strongly revelent to the local PLS model. If a calibration sample is far away from the query one, a relatively small importance weight is assigned to it to reduce its contribution to the local PLS model. This would be equivalent to resulting in a large noise, meaning that this point contains more information about noise or, equivalently, less information about the main parameters. Note that if the weight is equal

to zero, that is, $s_{i|q}=0$, the variance of noise will approach infinity and the corresponding point will be completely excluded in identifying the local PLS regression model. It is assumed that the measurement noises of the observations are independent. It is also assumed that the measurement noises of inputs and output are mutually independent. Thus, the likelihood can be simplified as follows

$$p(\mathbf{X},\mathbf{y}|\mathbf{P},\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)=p(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q)p(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q) \tag{22}$$

$$p(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q)=\prod_{i=1}^{N}p(\mathbf{x}_i|\mathbf{P},\mathbf{t}_i,\varphi,H,\mathbf{x}_q) \tag{23}$$

$$p(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q)=\prod_{i=1}^{N}p(y_i|\mathbf{t}_i,\mathbf{q},\varphi,H,\mathbf{x}_q) \tag{24}$$

$$\mathbf{x}_i|\mathbf{P},\mathbf{t}_i,\mathbf{q},\varphi,H,\mathbf{x}_q\sim\mathcal{N}\left(\mathbf{P}\mathbf{t}_i,\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right) \tag{25}$$

$$y_i|\mathbf{P},\mathbf{t}_i,\mathbf{q},\varphi,H,\mathbf{x}_q\sim\mathcal{N}\left(\mathbf{q}\mathbf{t}_i,\frac{Q_{e_y}}{s_{i|q}}\right) \tag{26}$$

The priors over the main parameters depend on the nature of the noise-free data. The noise-free inputs are assumed to follow a multivariate Gaussian distribution, that is

$$\tilde{\mathbf{x}}_i\sim\mathcal{N}(\boldsymbol{\mu}_\mathbf{x},\mathbf{Q}_\mathbf{x}) \tag{27}$$

As a result, given the loading matrix $\mathbf{P}$, the latent variable $\mathbf{t}_i$ will also follow a conditional multivariate Gaussian distribution

$$\mathbf{t}_i=\mathbf{P}^T\tilde{\mathbf{x}}_i \tag{28}$$

$$\mathbf{t}_i|\mathbf{P},H\sim\mathcal{N}\left(\mathbf{P}^T\boldsymbol{\mu}_x,\mathbf{P}^T\mathbf{Q}_x\mathbf{P}\right) \tag{29}$$

It is also assumed that the model parameters $\mathbf{b}$ follow a multivariate Gaussian distribution

$$\mathbf{b}\sim\mathcal{N}(\boldsymbol{\mu}_b,\mathbf{Q}_b) \tag{30}$$

Given the loading matrix $\mathbf{P}$, and the vector of model parameters $\mathbf{b}$, the regression coefficient vector $\mathbf{q}^T$ also follow a conditional multivariate Gaussian distribution

$$\mathbf{q}^T=\mathbf{P}^T\mathbf{b} \tag{31}$$

$$\mathbf{q}^T|\mathbf{P},H\sim\mathcal{N}\left(\mathbf{P}^T\boldsymbol{\mu}_b,\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right) \tag{32}$$

In the absence of any external knowledge over the loading matrix $\mathbf{P}$, a uniform prior distribution can be specified over $\mathbf{P}$. Based on the likelihood and prior distributions, the posterior distribution can be determined as

$$\begin{aligned}&p\left(\mathbf{X},\mathbf{y}|\mathbf{P},\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q\right)\propto p\left(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q\right)\\&\times p\left(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q\right)p(\mathbf{T}|\mathbf{P},H)p(\mathbf{q}|\mathbf{P},H)\end{aligned} \tag{33}$$

The maximum *a posteriori* probability (MAP) estimates can be obtained by solving the following optimization problem

$$\begin{aligned}\{\mathbf{P},\mathbf{T},\mathbf{q}\}_{\mathrm{MAP}}=&\arg\max_{\mathbf{P},\mathbf{T},\mathbf{q}}\{p\left(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q\right)\\&\times p\left(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q\right)p(\mathbf{T}|\mathbf{P},H)p(\mathbf{q}|\mathbf{P},H)\}\\&\text{s.t. }\mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{34}$$

It is intractable to solve this optimization problem directly. The overall objective function can be decomposed into the following three simultaneous parameter-estimation and data-reconciliation optimization problem

$$\begin{aligned}&\{\mathbf{P}\}_{\mathrm{MAP}}=\arg\max_{\mathbf{P}}p\left(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q\right)p\left(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q\right)\\&\{\mathbf{q}\}_{\mathrm{MAP}}=\arg\max_{\mathbf{q}}p\left(\mathbf{y}|\mathbf{T},\mathbf{q},\varphi,H,\mathbf{x}_q\right)p(\mathbf{q}|\mathbf{P},H)\\&\text{s.t.}\\&\{\mathbf{T}\}_{\mathrm{MAP}}=\arg\max_{\mathbf{T}}p\left(\mathbf{X}|\mathbf{P},\mathbf{T},\varphi,H,\mathbf{x}_q\right)p(\mathbf{T}|\mathbf{P},H)\\&\mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{35}$$

Because likelihood and priors are all multivariate Gaussian, the MAP estimates can be equivalently obtained by solving the following minimization problems

$$\begin{aligned}\{\mathbf{P}\}_{\mathrm{MAP}}=&\arg\min_{\mathbf{P}}\{\sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)\\&+\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{e_y}}{s_{i|q}}\right)^{-1}(y_i-\mathbf{q}\mathbf{t}_i)\}\\\{\mathbf{q}\}_{\mathrm{MAP}}=&\arg\min_{\mathbf{q}}\{\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{e_y}}{s_{i|q}}\right)^{-1}(y_i-\mathbf{q}\mathbf{t}_i)\\&+\left(\mathbf{q}^T-\mathbf{P}^T\boldsymbol{\mu}_b\right)^T\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\left(\mathbf{q}^T-\mathbf{P}^T\boldsymbol{\mu}_b\right)\}\\\{\mathbf{t}_i\}_{\mathrm{MAP}}=&\arg\min_{\mathbf{t}_i}\{(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)\\&+\left(\mathbf{t}_i-\mathbf{P}^T\boldsymbol{\mu}_\mathbf{x}\right)^T\left(\mathbf{P}^T\mathbf{Q}_\mathbf{x}\mathbf{P}\right)^{-1}\left(\mathbf{t}_i-\mathbf{P}^T\boldsymbol{\mu}_\mathbf{x}\right)\\&\text{s.t. }\mathbf{P}^T\mathbf{P}=\mathbf{I}\end{aligned} \tag{36}$$

The first optimization function is intractable to solve because of the unit orthonormal constraint. We can first use optimization methods that have a closed form solution, to estimate $\mathbf{P}$. In this way, both of the following optimization problems can be solved analytically

$$\{\mathbf{t}_i\}_{\mathrm{MAP}}=\left[\mathbf{P}^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}\mathbf{P}+\left(\mathbf{P}^T\mathbf{Q}_\mathbf{x}\mathbf{P}\right)^{-1}\right]^{-1}\left[\mathbf{P}^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}\mathbf{x}_i+\left(\mathbf{P}^T\mathbf{Q}_\mathbf{x}\mathbf{P}\right)^{-1}\mathbf{P}^T\boldsymbol{\mu}_\mathbf{x}\right] \tag{37}$$

$$\{\mathbf{q}^T\}_{\mathrm{MAP}}=\left[\mathbf{T}^T\mathbf{S}_q\mathbf{T}Q_{e_y}+\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\right]^{-1}\left[\mathbf{T}^T\mathbf{S}_q\mathbf{Y}Q_{e_y}+\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\mathbf{P}^T\mu_b\right] \tag{38}$$

In this Bayesian modeling algorithms, $\mathbf{Q}_{e_x},Q_{e_y},\mu_b,\mathbf{Q}_b,\mu_x,\mathbf{Q}_x$ are assumed to be known. That means the prior density was assumed to be fully specified in advance. In the presence of limited prior knowledge over the noise variance and main parameters, a widely used alternative is the empirical Bayesian analysis which estimates the prior from the available data assuming data is representative.[18] In the empirical Bayesian analysis, there are two kinds of approaches to estimate the prior from data: parametric approach and nonparametric approach.[17] The parametric approach assuming the structures of the prior distribution are known and it only needs to estimate the hyperparameters of the prior density function. The nonparametric approach will estimate the entire prior from the data which is more complex and time-consuming. For computational convenience, the parametric approach is used to estimate the prior in the light of training data using Algorithm II in Appendix.

Inference of localization parameter

Applying Bayes' rule, the posterior PDF of localization parameter can be expressed as

$$p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q) \propto p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)p(\varphi|H,\mathbf{x}_q) \tag{39}$$

As priors, one can assume that the localization parameter φ is statistically independent of the model structure H and the query sample $\mathbf{x}_q$

$$p(\varphi|H,\mathbf{x}_q)=p(\varphi) \tag{40}$$

In the absence of any external knowledge, a noninformative prior can be specified in the form of a constrained uniform distribution. To incorporate the available prior knowledge, conjugate priors are normally used for which the resulting posterior distribution can be conveniently evaluated. To assure generality, a Gamma prior distribution is specified over the localization parameter

$$p(\varphi)=\frac{\varphi^{a-1}}{b^a\Gamma(a)}\exp\left(-\frac{\varphi}{b}\right) \tag{41}$$

where a is the shape parameter and b is the scale parameter. The likelihood in Eq. 39 can be evaluated by integrating out the main parameters

$$p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)=\int_\Theta p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|H)d\Theta \tag{42}$$

As the above problem is often intractable, the integral in Eq. 42 can be approximated by applying Laplace's method of approximation[19]

$$\begin{aligned}&\int_\Theta p(\mathbf{X},\mathbf{y}|\Theta,\varphi,H,\mathbf{x}_q)p(\Theta|H)d\Theta\\ &\quad\approx p(\mathbf{X},\mathbf{y}|\Theta^{\mathrm{MAP}},\varphi,H)p(\Theta^{\mathrm{MAP}}|H)\det\left(\frac{\mathbf{A}_\Theta}{2\pi}\right)^{-\frac{1}{2}}\end{aligned} \tag{43}$$

where $\mathbf{A}_\Theta=-\nabla\nabla\log p(\Theta|\varphi,H,\mathbf{X},\mathbf{y},\mathbf{x}_q)$. The inverse of Hessian matrix $\mathbf{A}_\Theta$ reflects the posterior uncertainty in Θ. Then, the MAP estimate of localization parameter can be shown as

$$\begin{aligned}\{\varphi\}_{\mathrm{MAP}}&=\arg\max_\varphi\{p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)\}\\ &=\arg\max_\varphi\left\{p(\mathbf{X},\mathbf{y}|\Theta^{\mathrm{MAP}},\varphi,H,\mathbf{x}_q)p(\Theta^{\mathrm{MAP}}|H)\det\left(\frac{\mathbf{A}_\Theta}{2\pi}\right)^{-\frac{1}{2}}p(\varphi)\right\}\end{aligned} \tag{44}$$

As both the likelihood and prior probability density functions belong to the family of exponential PDFs, the MAP solution can be obtained by solving the following minimization problem

$$\{\varphi\}_{\mathrm{MAP}}=\arg\min_\varphi\left\{\begin{aligned}&\frac{1}{2}\sum_{i=1}^N(\mathbf{x}_i-\hat{\mathbf{x}}_i)^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}(\mathbf{x}_i-\hat{\mathbf{x}}_i)\\ &+\frac{1}{2}\sum_{i=1}^N(y_i-\hat{y}_i)^T\left(\frac{Q_{e_y}}{s_{i|q}}\right)^{-1}(y_i-\hat{y}_i)\\ &+(1-a)\log\varphi+\frac{1}{b}\varphi\\ &-\log\left[\det\left(\tfrac{\mathbf{A}_\Theta}{2\pi}\right)^{-\frac{1}{2}}\right]-\frac{M+1}{2}\log\prod_{i=1}^N s_{i|q}\end{aligned}\right\} \tag{45}$$

This optimization problem can be solved by sampling method instead of deriving a closed form solution which cannot be obtained directly. For instance, the continuous localization parameter can be discretized into a finite set of reasonable values $\{\varphi_1,\varphi_2,\ldots\varphi_f\}$. We can next draw samples from the posterior distribution using these candidate values of the localization parameters to approximate the MAP solution.

Inference of model structure

Applying Bayes' rule, the posterior PDF of model structure can be expressed as

$$p(H|\mathbf{X},\mathbf{y},\mathbf{x}_q) \propto p(\mathbf{X},\mathbf{y}|H,\mathbf{x}_q)p(H|\mathbf{x}_q) \tag{46}$$

As priors, it is reasonable to assume that the model structure is statistically independent of the query sample $\mathbf{x}_q$. Given a set of candidate model structures, that is, $H\in\{H_1,H_2\ldots H_L\}$, the random variable H is a categorical variable and can be modeled by

$$p(H)=\prod_{l=1}^L p(H=H_l)^{[H=H_l]} \tag{47}$$

where $[H=H_l]$ equals 1 if $H=H_l$ and equals 0 otherwise. In the absence of any prior information, a uniform distribution can be used for the candidate model structures, that is, $p(H=H_1)=p(H=H_2)\ldots=p(H=H_L)$.

The likelihood function $p(\mathbf{X},\mathbf{y}|H,\mathbf{x}_q)$ can be obtained by integrating out the localization parameter

$$p(\mathbf{X},\mathbf{y}|H,\mathbf{x}_q)=\int_\varphi p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)p(\varphi)d\varphi \tag{48}$$

As it is intractable to solve the above integral directly, Laplace's method of approximation is applied again

$$\int_\varphi p(\mathbf{X},\mathbf{y}|\varphi,H,\mathbf{x}_q)p(\varphi)d\varphi\approx p(\mathbf{X},\mathbf{y}|\varphi^{\mathrm{MAP}},H,\mathbf{x}_q)p(\varphi^{\mathrm{MAP}})\det\left(\frac{A_\varphi}{2\pi}\right)^{-\frac{1}{2}} \tag{49}$$

where $A_\varphi=-\nabla\nabla\log p(\varphi|H,\mathbf{X},\mathbf{y},\mathbf{x}_q)$. The inverse of Hessian matrix A_φ reflects the posterior uncertainty in φ.

Finally, the MAP estimate of the model structure can be obtained as follows

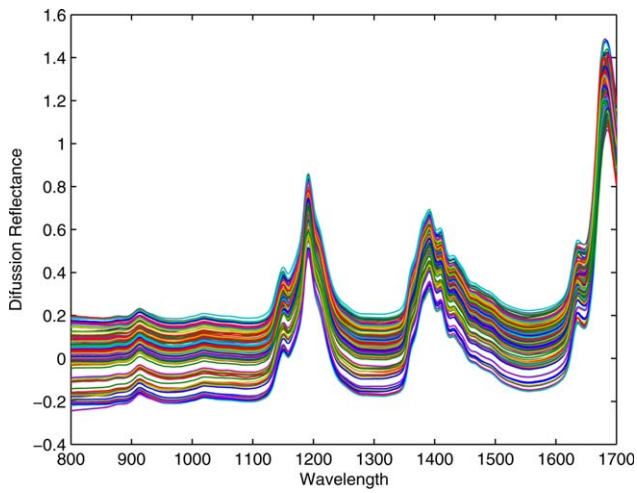

Figure 1. RVP of gasoline spectra data.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 1. Comparing Prediction Performance of the First Layer of Bayesian LW-PLS and Regular LW-PLS Using Dataset from Reid Vapor Pressure of Gasoline, Scenario I

	Bayesian LW-PLS	Regular LW-PLS
Localization parameter λ=0.5		
Selected number of retained LVs $H=30$		
MSE of cross-validation	15.4347	15.6006
Correlation of cross-validation	0.9732	0.9731

$$\{H\}_{\mathrm{MAP}}=\arg\max_{H}\left(p\left(H|\mathbf{X},\mathbf{y},\mathbf{x}_q\right)\right)$$

$$=\arg\max_{H}\left\{p\left(\mathbf{X},\mathbf{y}|\varphi^{\mathrm{MAP}},H,\mathbf{x}_q\right)p\left(\varphi^{\mathrm{MAP}}\right)\det\left(\frac{A_\varphi}{2\pi}\right)^{-\frac{1}{2}}p(H)\right\}$$

$$=\arg\max_{H}\left\{p\left(\mathbf{X},\mathbf{y}|\Theta^{\mathrm{MAP}},\varphi^{\mathrm{MAP}},H,\mathbf{x}_q\right)p\left(\Theta^{\mathrm{MAP}}|H\right)\right.$$

$$\left.p(\varphi^{\mathrm{MAP}})\det\left(\frac{\mathbf{A}_\Theta}{2\pi}\right)^{-\frac{1}{2}}\det\left(\frac{A_\varphi}{2\pi}\right)^{-\frac{1}{2}}p(H)\right\} \tag{50}$$

As both the likelihood and prior probability density functions belong to the family of exponential PDFs, the MAP solution can be obtained by solving the following minimization problem

$$\{H\}_{\mathrm{MAP}}=\arg\min_{H}$$

$$\left\{\begin{array}{l}
\frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\mathbf{x}}_i)^T\left(\frac{\mathbf{Q}_{e_x}}{s_{i|q}}\right)^{-1}(\mathbf{x}_i-\hat{\mathbf{x}}_i)\\
+\frac{1}{2}\sum_{i=1}^{N}(y_i-\hat{y}_i)^T\left(\frac{Q_{e_y}}{s_{i|q}}\right)^{-1}(y_i-\hat{y}_i)\\
+\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{t}_i-\mathbf{P}^T\mu_x\right)^T\left(\mathbf{P}^T\mathbf{Q}_x\mathbf{P}\right)^{-1}\left(\mathbf{t}_i-\mathbf{P}^T\mu_x\right)\\
+\frac{1}{2}\left(\mathbf{q}^T-\mathbf{P}^T\mu_b\right)^T\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\left(\mathbf{q}^T-\mathbf{P}^T\mu_b\right)\\
+(1-a)\log\varphi+\frac{1}{b}\varphi-\log\left[\det\left(\frac{\mathbf{A}_\Theta}{2\pi}\right)^{-\frac{1}{2}}\right]-\frac{M+1}{2}\log\prod_{i=1}^{N}s_{i|q}\\
+\frac{1}{2}(1+N)H\log 2\pi-\log\left[\det\left(\frac{A_\varphi}{2\pi}\right)^{-\frac{1}{2}}\right]
\end{array}\right\} \tag{51}$$

Hierarchical Bayesian Optimization Procedure

1. Choose the similarity function given in Eq. 8.
2. Select a proper set of candidate model structures $\{H_1,H_2...H_L\}$. If there is available prior information about the model structures, the candidates and their prior probabilities p(H) can be determine based on the prior knowledge. If there is no prior information, the candidate model structures can be selected based on empirical method: select several candidate model structures around the globally optimal one obtained from off-line leave-one-out cross-validation (LOOCV), and set a uniform prior distribution over this set of candidate model structures.

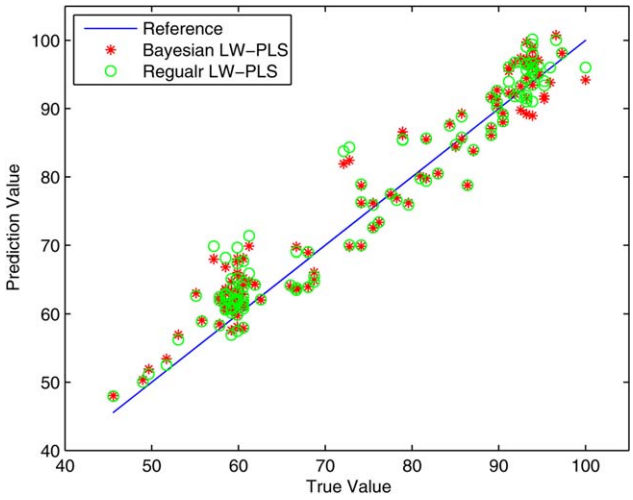

Figure 2. Cross-validation using dataset from RVP of gasoline, scenario I.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

3. Characterize the noise variances, $\mathbf{Q}_{e_x}$ and Q_{e_y}, and specify a prior distribution over the main parameters, $p(\Theta|H)$, using Algorithm II.
4. Characterize the prior distribution over localization parameter, $p(\varphi|H)$, using Algorithm III.
5. For $l=1:L$
   (1) Select H_l and choose an initial value for the localization parameter φ_l.
   (2) Although $\mathbf{P}_l,\mathbf{T}_l,\mathbf{q}_l$, and φ_l converge
      (2.1) calculate the similarity matrix, $\mathbf{S}_{q_l}$, using Eqs. 5, 7, and 8.
      (2.2) calculate the loading matrix, $\mathbf{P}_l$, by applying the LW-PLS algorithm to $\{\mathbf{X},\mathbf{y},\mathbf{x}_q\}$.
      (2.3) calculate the regression coefficient vector, $\mathbf{q}_l$, latent variable matrix, $\mathbf{T}_l$, using Eqs. 37 and 38.
      (2.4) calculate the localization parameter φ_l using Eq. 45.
   (3) Calculate the posterior probability of model structure, $p\left(H=H_l|\mathbf{X},\mathbf{y},\mathbf{x}_q\right)$, using Eq. 51.
6. Choose the model structure with the highest posterior probability as well as corresponding loading matrix, $\mathbf{P}$, and regression coefficient vector, $\mathbf{q}$.
7. Calculate output as $\hat{y}=\mathbf{x}_q\mathbf{P}\mathbf{q}^T$.

Table 2. Comparing Prediction Performance of the First and Second Layer of Bayesian LW-PLS and Regular LW-PLS Using Dataset from Reid Vapor Pressure of Gasoline, Scenario II

	Bayesian LW-PLS	Regular LW-PLS
Selected number of retained LVs $H=30$		
Localization parameter	$\varphi\in[0.1,2]$	λ=0.2
MSE of cross-validation	9.9807	57.5943
Correlation of cross-validation	0.9833	0.8846
Localization parameter	$\varphi\in[0.1,2]$	λ=0.8
MSE of cross-validation	9.9807	10.1754
Correlation of cross-validation	0.9833	0.9816
Localization parameter	$\varphi\in[0.1,2]$	λ=1.5
MSE of cross-validation	9.9807	14.1729
Correlation of cross-validation	0.9833	0.9743
Localization parameter	$\varphi\in[0.1,2]$	$\lambda=2$
MSE of cross-validation	9.9807	15.9304
Correlation of cross-validation	0.9833	0.9713

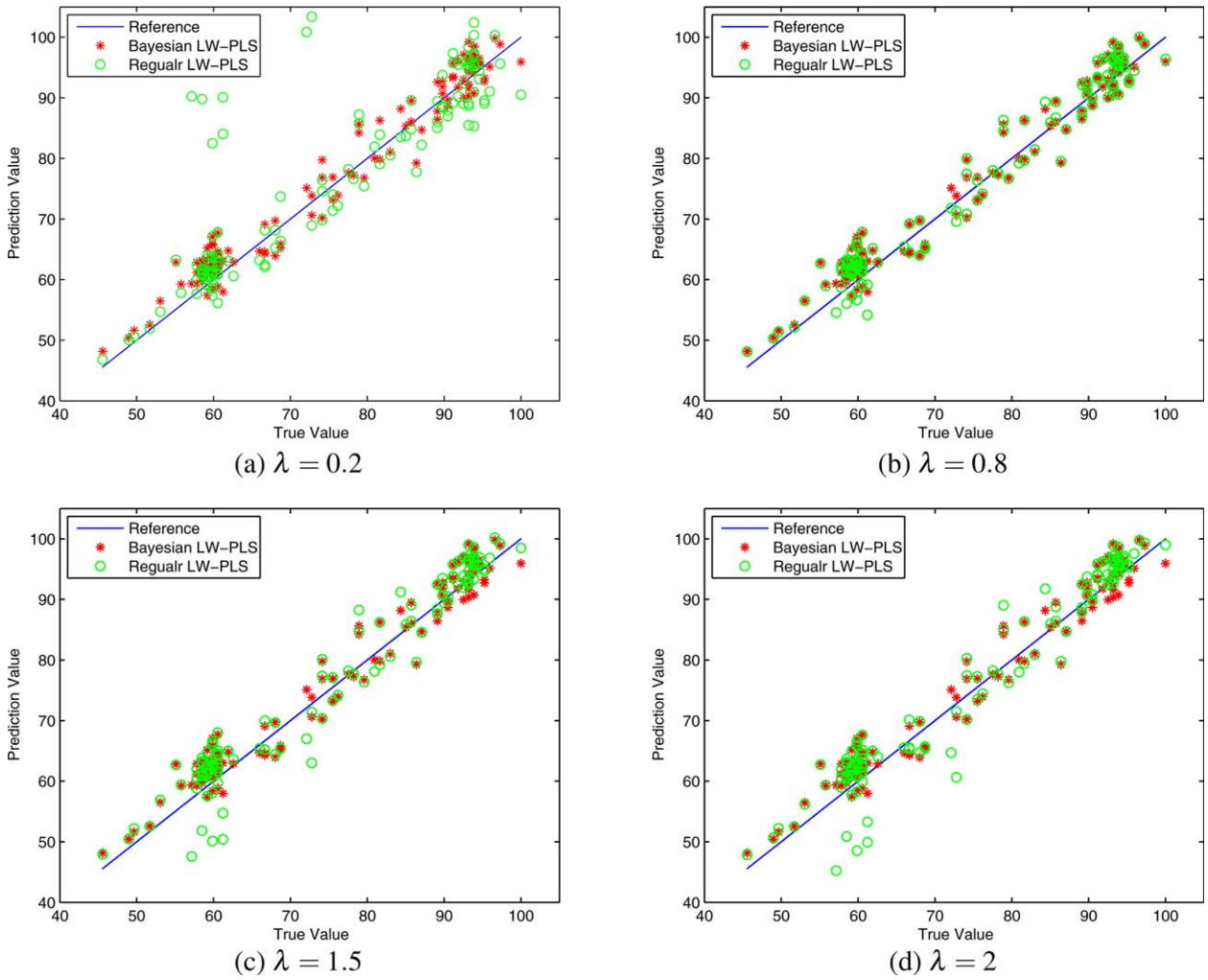

Figure 3. Cross-validation using dataset from RVP of gasoline, scenario II.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Case Studies

This section demonstrates the practical application of the Bayesian LW-PLS through case studies. To illustrate the advantages of hierarchical Bayesian optimization, two sets of near NIR data for real-time prediction of reid vapor pressure (RVP) of gasoline and wheat kernels are used. It is noteworthy that the NIR datasets have high dimension with strongly correlated spectra. All industrial data presented here have been normalized to protect proprietary information.

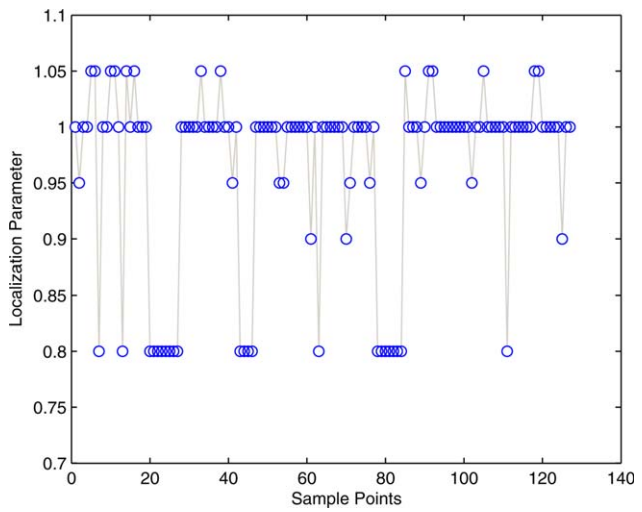

Figure 4. Localization parameter φ of Bayesian LW-PLS.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

RVP of gasoline

The objective of this study is to estimate RVP of gasoline from NIR spectra data. The set of data is taken from Khatibisepehr et al. (submitted). The dataset consists of NIR spectra for 423 gasoline samples. The diffusion reflectance

Table 3. Comparing Prediction Performance of Bayesian LW-PLS and RMSECV-Based LW-PLS Using Dataset from Reid Vapor Pressure of Gasoline, Scenario III

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[25,30]	30
Localization parameter	$\varphi \in [0.1, 2]$	$\lambda = 2$
MSE of cross-validation	9.9499	15.9304
Correlation of cross-validation	0.9835	0.9713

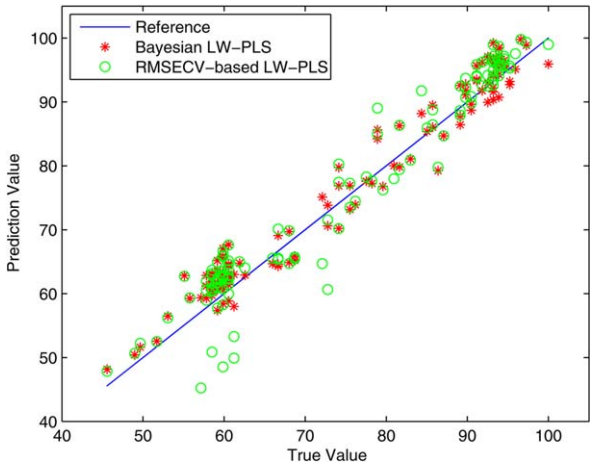

Figure 5. Cross-validation using dataset from RVP of gasoline, scenario III.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

spectra of samples are measured with wavelength range of 800–1700 nm in 1-nm intervals (Figure 1). The samples are divided into 296 calibration dataset and 127 test one. Standard ASTM testing methodologies have been used to obtain the reference measurements for RVP.

To show the features of the proposed method more clearly, the performance is evaluated in the following three scenarios:

Scenario I: Known localization parameter and model structure, but unknown main parameters.

The first layer of the proposed method, inference of main parameters, is applied to develop real-time LW-PLS models for the prediction of RVP. The prediction performance of the developed models is compared with that of the models identified using regular LW-PLS regression. The similarity functions are chosen as in Eq. 6 and the localization parameter λ and number of retained latent variables H are set as 0.5 and 30, respectively, and same for both methods. The prior distributions of main parameters are specified using Algorithm II. The comparison results are reported in Table 1 and Figure 2. It can be observed that the Bayesian parameter estimation is more accurate than the regular LW-PLS for some of the calibration samples. A slightly higher prediction performance has been achieved by incorporating the prior knowledge and taking into account the different contributions of noise in the measurements. The challenge in using this Bayesian approach for estimation exists not only in obtaining proper prior distribution but also in specifying

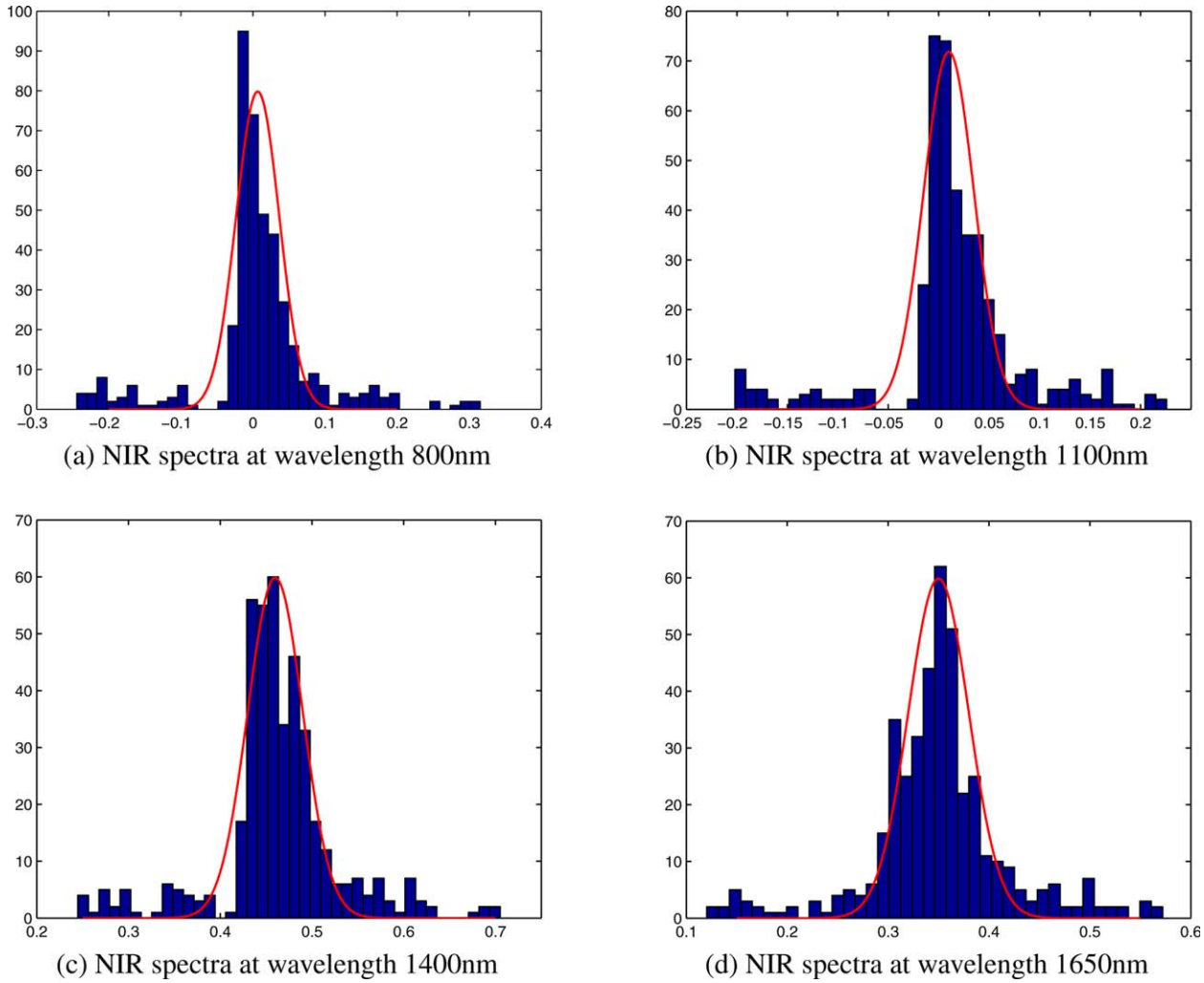

Figure 6. Distributions of selected inputs for RVP of gasoline example.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 4. Comparing Prediction Performance of the Bayesian LW-PLS and RMSECV-Based LW-PLS Using Dataset from Wheat Kernels

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[8,12]	9
Localization parameter	$\varphi \in [0.1, 2]$	λ=0.2
MSE of cross-validation	20.8174	48.3352
Correlation of cross-validation	0.9365	0.8517

Table 5. Comparing Prediction Performance of the Bayesian LW-PLS and RMSECV-Based LW-PLS Using Dataset from Wheat Kernels, Extrapolation Case

	Bayesian LW-PLS	RMSECV
Selected Number of retained LVs	[8,12]	9
Localization parameter	$\varphi \in [0.1, 2]$	λ=0.2
MSE of cross-validation	94.9904	380.0755
Correlation of cross-validation	0.8207	0.8163

appropriate noise variance. As no prior information is available, the variances of measurement noise can only be estimated from existing sources such as the calibration data. However, the main challenge in locally weighted methods is simultaneous estimation of localization parameter, model structure, and main parameters where the proposed Bayesian approach shows its great advantage, as demonstrated in the following scenarios.

Scenario II: Known model structure, but unknown localization parameter and main parameters.

The first and second layers of the proposed method, estimation of the main parameters, and selection of localization parameter are applied to identify the LW-PLS models. The number of retained latent variables is set as 30. For the regular LW-PLS, the classic similarity function (Eq. 6) is used and we consider four different values for localization parameter λ: 0.2, 0.8, 1.5, and 2. For the proposed method, the new similarity function (Eq. 8) is used. The prior distribution of main parameters is specified using Algorithm II. The prior distribution over the localization parameters φ is specified using Algorithm III within sampling range [0.1,2]. From the results shown in Table 2 and Figure 3, it can be observed that the performance of the regular LW-PLS method highly depends on the value of the localization parameter. Therefore, proper tuning of the localization parameters has a significant effect on the prediction performance of the LW-PLS models. As the Bayesian LW-PLS searches for the locally optimal value of the localization parameter within the developed hierarchical optimization framework, the prediction performance of the resulting LW-PLS models is superior. Figure 4 shows that for different local models, different optimal localization parameters have been obtained to achieve a better performance.

Scenario III: Unknown model structure, localization parameter and main parameters.

The proposed method, Bayesian LW-PLS and one widespread method, RMSECV-based LW-PLS are applied to develop the LW-PLS models for real-time prediction of RVP. The main idea behind RMSECV is to search for the globally optimal localization parameter and model structure by minimizing the RMSE of LOOCV in an off-line identification phase and then apply the LW-PLS to do online estimation of the main parameters. The candidate model structures are set as $[25, 30]$ for both methods. The result of RMSECV for optimal localization parameters and number of retained latent variables are 2 and 30, respectively. For Bayesian LW-PLS, first, the prior distributions over the main parameters are specified using Algorithm II. The prior distribution over the localization parameters φ is specified using Algorithm III within sampling range [0.1,2]. In the absence of the prior knowledge, a uniform distribution is used for the model structure. The comparison results are reported in Table 3 and illustrated in Figure 5. According to the results, Bayesian LW-PLS performs much better than the traditional method, RMSECV-based LW-PLS.

In all of these three scenarios, the priors over main parameter are obtained from estimation of empirical prior, that is, Algorithm II. The assumption behind this approach is Gaussian distributed inputs. As shown in Figure 6, the distribution of the input can be well approximated by Gaussian distribution.

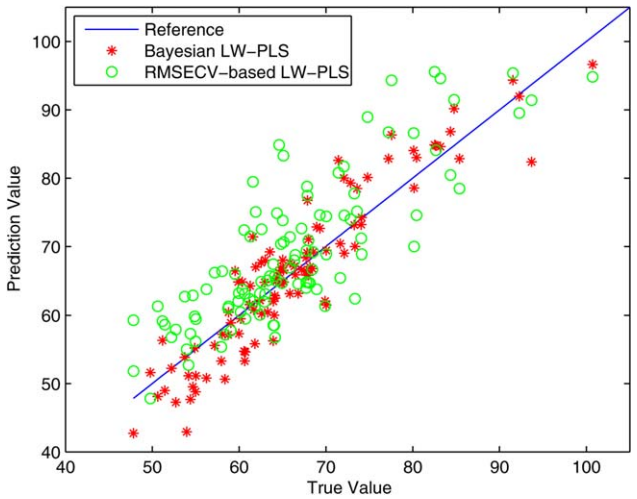

Figure 7. Cross-validation using dataset from protein content of wheat kernels.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

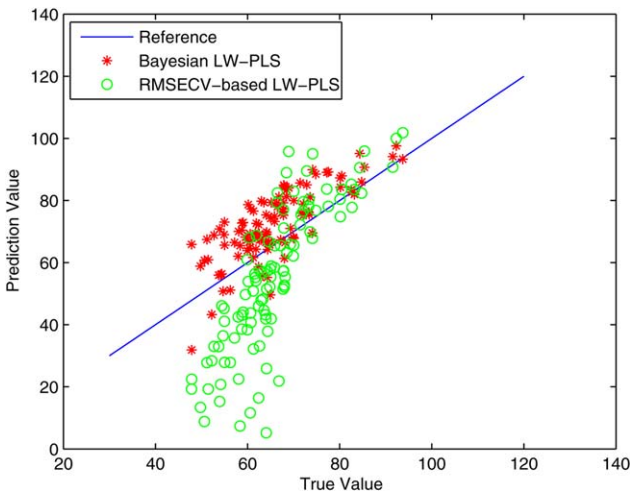

Figure 8. Cross-validation using dataset from protein content of wheat kernels, extrapolation case.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

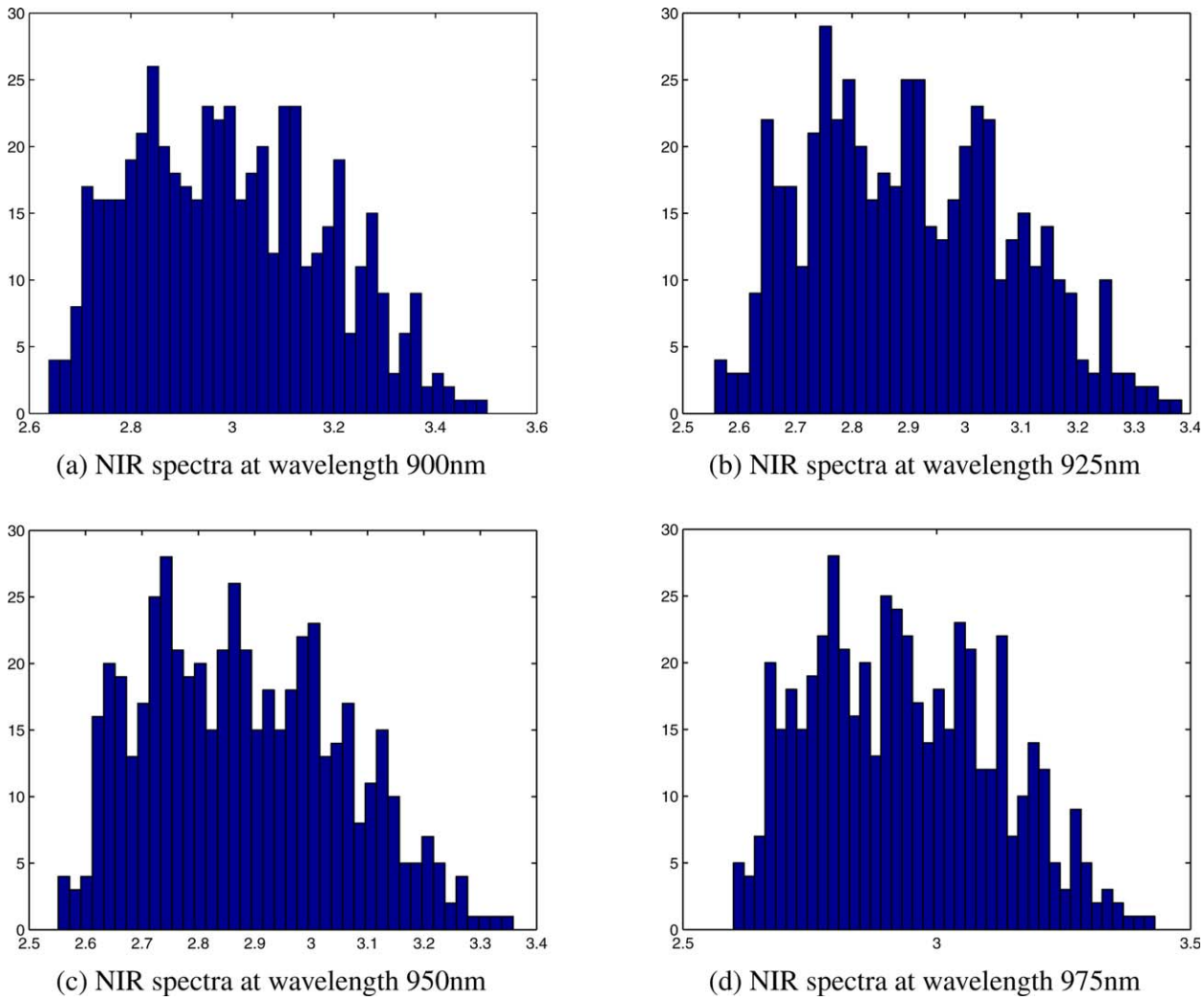

Figure 9. Distributions of selected inputs for protein content of wheat kernels example.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Protein content of wheat kernels

In this case study, the LW-PLS models are developed for online prediction of the protein content of wheat kernels from the NIR spectra. This dataset was used by Refs. 20 and [21] as a standard NIR dataset. The wheat kernels were randomly chosen from bulk samples representing different varieties or various mixtures from two different locations in Denmark.

The calibration and test datasets collected in this study consist of 100 and 105 samples with reference value ranges from 46.1 to 103.4 and 47.8 to 93.7, respectively. As stated by Refs. 20 and 21, the test samples were acquired with the calibration samples, but stored for about 2 additional months before measurement to provide a check for temporal drift in the samples and instrumentation.

The Bayesian LW-PLS and RMSECV-based LW-PLS are applied to develop the calibration models for protein content. The candidate model structures are set as [8,12] for both methods. The optimal localization parameter, λ, and number of retained latent variables, H, obtained via RMSECV, are 0.2 and 9, respectively. For the Bayesian LW-PLS, the prior distribution of main parameters is specified by following the procedure in Algorithm II. A Gamma prior distribution over the localization parameter φ is extracted from calibration data using Algorithm III and the corresponding sampling range is chosen as [0.1,2]. From comparison results reported in Table 4 and illustrated in Figure 7, it can, again, be observed that the Bayesian LW-PLS significantly outperforms the RMSECV-based LW-PLS.

To further evaluate the effectiveness of the proposed method, a case of extrapolation is performed on the same NIR dataset. The calibration samples which have output value in the range of 82.3–103.4 are selected to form the new calibration dataset. The test ones remain unchanged which have output value between 47.8 and 93.6 so that the calibration dataset does not overlap with all the operation region of test ones. It means some of the prediction can only be carried out by extrapolation. This situation can happen in real-world application if a process is shifted to a new operation mode.

From Table 5 and Figure 8, we can see that the performances of the proposed methods are again much better than the RMSECV-based LW-PLS. Especially in the extrapolated part where outputs range from 47.8 to 82.3, the predictions of RMSECV-based LW-PLS obviously deviate from the reference value, while the predictions of Bayesian one can still follow the reference.

In this case study, the priors over main parameters are also obtained from Algorithm II. As shown in Figure 9, even though the distribution of the input does not exactly follow Gaussian distribution, the proposed method can still outperform the compared one. Revisit the optimization problem in Eq. 34 which is equivalent to the following minimization problem

$$\{\mathbf{P},\mathbf{q},\mathbf{T}\}_{\mathrm{MAP}}=\arg\min_{\mathbf{P},\mathbf{q},\mathbf{T}} \left\{ \begin{array}{l} \sum_{i=1}^{N}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{ex}}{s_{i|q}}\right)^{-1}(\mathbf{x}_i-\mathbf{P}\mathbf{t}_i) \\ +\sum_{i=1}^{N}(y_i-\mathbf{q}\mathbf{t}_i)^T\left(\frac{\mathbf{Q}_{ey}}{s_{i|q}}\right)^{-1}(y_i-\mathbf{q}\mathbf{t}_i) \\ +\left(\mathbf{q}^T-\mathbf{P}^T\boldsymbol{\mu}_b\right)^T\left(\mathbf{P}^T\mathbf{Q}_b\mathbf{P}\right)^{-1}\left(\mathbf{q}^T-\mathbf{P}^T\boldsymbol{\mu}_b\right) \\ +\sum_{i=1}^{N}\left(\mathbf{t}_i-\mathbf{P}^T\boldsymbol{\mu}_{\mathbf{x}}\right)^T\left(\mathbf{P}^T\mathbf{Q}_{\mathbf{x}}\mathbf{P}\right)^{-1}\left(\mathbf{t}_i-\mathbf{P}^T\boldsymbol{\mu}_{\mathbf{x}}\right) \end{array} \right\} \quad (52)$$

$$\text{s.t. } \mathbf{P}^T\mathbf{P}=\mathbf{I}$$

The first two quadratic terms represent the information from historical data, last two quadratic terms contain the information from available prior knowledge over main parameters. If informative priors are obtained beforehand, the last two terms will make a contribution to a more accuracy estimation. If the prior contains little helpful information (situation in this case), a fairly good estimation of main parameters can be achieved by taking advantage of information contained in data. Moreover, the next two layers of Bayesian framework, that is, inference of localization parameter and model structure can further improve the performance.

Conclusion

This article proposed a holistic Bayesian framework for the LW-PLS regression. The proposed method has the following advantages over the regular LW-PLS regression: (1) by following a Bayesian approach to estimate the main parameters of the LW-PLS model, available prior knowledge can be incorporated into the identification process. (2) Different contributions of measurement noise can be taken into account. (3) Application of the hierarchical Bayesian optimization framework offers a systematic and tractable way to get the optimal combination of the model structure, localization parameters as well as main parameters for each operating point. (4) Bayesian model structure selection can automatically deals with the model complexity problem to avoid the overfitting issue. The attractive features of the proposed framework were illustrated through two industrial case studies in which NIR spectra were used to provide real-time estimates of RVP and wheat kernels using the LW-PLS models. In the first case study, different scenarios were investigated not only to illustrate the advantages of each layer of the proposed Bayesian formulation of the LW-PLS regression problem, but also to clearly demonstrate the integration mechanism adopted in the developed hierarchial Bayesian optimization framework.

Acknowledgment

This work is supported in part by Natural Sciences and Engineering Research Council of Canada (NSERC) Industrial Research Chair program and Alberta Innovates Technology Futures (AITF) Industry Chair Program.

Literature Cited

1. Abdi H. Partial least squares regression and projection on latent structure regression. *Wiley Interdiscip Rev Comput Stat*. 2010;2:97–106.
2. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Ser B Stat Methodol*. 2010;72:3–25.
3. Lin B, Jørgensen SB. Soft sensor design by multivariate fusion of image features and process measurements. *J Process Control*. 2011; 21:547–553.
4. Shao X, Xu F, Huang B, Espejo A. Estimation of bitumen froth quality using Bayesian information synthesis: an application to froth transportation process. *Can J Chem Eng*. 2012;90:1393–1399.
5. Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics*. 1993;35:109–135.
6. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal Chim Acta*. 1986;185:1–17.
7. Wold S, Kettaneh-wold N, Skagerberg B. Nonlinear PLS modeling. *Chemometr Intell Lab Syst*. 1989;7:53–65.
8. Qin SJ, McAvoy TJ. Nonlinear PLS modeling using neural networks. *Comput Chem Eng*. 1992;16:379–391.
9. Bang YH, Yoo CK, Lee IB. Nonlinear PLS modeling with fuzzy inference system. *Chemometri Intell Lab Syst*. 2002;64:137–155.
10. Malthouse EC, Tamhane AC, Mah RSH. Nonlinear partial least squares. *Comput Chem Eng*. 1997;21:875–890.
11. Kim S, Okajima R, Kano M, Hasebe S. Development of soft-sensor using locally weighted PLS with adaptive similarity measure. *Chemometr Intell Lab Syst*. 2013;124:43–49.
12. Perez-Guaita D, Kuligowski J, Quint G, Garrigues S. Modified locally weighted partial least squares regression improving clinical predictions from infrared spectra of human serum samples. *Talanta*. 2013;107:368–375.
13. Leardi R. *Nature-Inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*. Amsterdam: Elsevier, 2003.
14. Kim S, Kano M, Hasebe S, Takinami A, Seki T. Long-term industrial applications of inferential control based on just-in-time soft-sensors: economical impact and challenges. *Ind Eng Chem Res*. 2013; 52:12346–12356.
15. Guyon I, Saffari A, Dror G, Cawley G. Model selection: beyond the Bayesian/frequentist divide. *J Mach Learn Res*. 2010;11:61–87.
16. Mackay DJ. Comparison of approximate methods for handling hyperparameters. *Neural Comput*. 1999;11:1035–1068.
17. Nounou MN, Bakshi BR, Goel PK, Shen X. Process modeling by Bayesian latent variable regression. *AIChE J*. 2001;48:1775–1793.
18. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. CRC Press, 2003.
19. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90:773–795.
20. Pedersen DK, Martens H. Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): analysis of near-infrared transmittance spectra of single wheat seeds. *Appl Spectrosc*. 2002;56:1206–1214.
21. Nielsen JP, Pedersen DK, Munck L. Development of nondestructive screening methods for single kernel characterization of wheat. *Cereal Chem*. 2003;80:274–280.

Appendix

Algorithm I: Regular locally weighted partial least square

1. Determine the number of latent variables H, localization parameters λ.
2. When query sample $\mathbf{x}_q$ arrives, calculate the similarity matrix $\mathbf{S}_q$ using Eqs. 5, 6 and 7.
3. Calculate the weight matrix, loading matrix, and regression coefficient vector by

$$\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2\ldots\mathbf{w}_H] \quad (A1)$$

$$\mathbf{P}=[\mathbf{p}_1,\mathbf{p}_2\cdots\mathbf{p}_H] \quad (A2)$$

$$q=[q_1, q_2 \ldots q_H] \tag{A3}$$

$$\mathbf{w}_h=\frac{\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)^T\mathbf{S}_q\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)}{\left\|\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)^T\mathbf{S}_q\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)\right\|} \tag{A4}$$

$$\mathbf{p}_h=\frac{\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)^T\mathbf{S}_q\mathbf{t}_h}{\mathbf{t}_h{}^T\mathbf{S}_q\mathbf{t}_h} \tag{A5}$$

$$q_h=\frac{\left(y-\sum_{j=1}^{h-1}\mathbf{t}_jq_j{}^T\right)^T\mathbf{S}_q\mathbf{t}_h}{\mathbf{t}_h{}^T\mathbf{S}_q\mathbf{t}_h} \tag{A6}$$

where the columns of $\mathbf{W}\in\mathbb{R}^{M\times H}$ are orthonormal weight vectors and $\mathbf{t}_h$ denotes the hth column of $\mathbf{T}$ which is calculated by

$$\mathbf{t}_h=\left(\mathbf{X}-\sum_{j=1}^{h-1}\mathbf{t}_j\mathbf{p}_j{}^T\right)\mathbf{w}_h \tag{A7}$$

4. Calculate output of the local PLS model by

$$\hat{\mathbf{y}}=\mathbf{x}_q\mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{q}^T \tag{A8}$$

Algorithm II: Estimation of empirical prior over main parameters

1. For industry data, it is rational to assume in a short period of time (i.e., one sampling interval), the input and output are kept constant. The incremental input output measurements are resulted from the measurement noise. So, the noise variance $\mathbf{Q}_{e_x}, Q_{e_y}$ is calculated by the variances of the distribution of incremental input and output measurements

$$\mathbf{J}_x=[\mathbf{x}_1-\mathbf{x}_2, \mathbf{x}_2-\mathbf{x}_3, \ldots \mathbf{x}_{N-1}-\mathbf{x}_N]^T \tag{A9}$$

$$\mathbf{j}_y=[y_1-y_2, y_2-y_3, \ldots y_{N-1}-y_N]^T \tag{A10}$$

$$\mathbf{Q}_{e_x}=\frac{1}{2}\mathrm{var}(\mathbf{J}_x) \tag{A11}$$

$$Q_{e_y}=\frac{1}{2}\mathrm{var}\left(\mathbf{j}_y\right) \tag{A12}$$

2. Solve the Bayesian LW-PLS modeling problem with a uniform priori for all the main parameters.
3. Estimate the set of hyperparameters $\mu_b, \mathbf{Q}_b, \mu_x, \mathbf{Q}_x$ as follows

$$\mu_b=E\left[\mathbf{P}\mathbf{q}^T\right] \tag{A13}$$

$$\mathbf{Q}_b=c\left(\hat{\mathbf{X}}^T\mathbf{S}_q\hat{\mathbf{X}}\right)^{-1} \tag{A14}$$

$$\mu_x=E\left[\hat{\mathbf{X}}\right] \tag{A15}$$

$$\mathbf{Q}_x=\mathrm{Cov}\left[\hat{\mathbf{X}}\right] \tag{A16}$$

4. Solve the Bayesian LW-PLS modeling problem using the empirically estimate priori.

Algorithm III: Estimation of empirical prior over localization parameter

1. Determine proper model structure H.
2. Choose the similarity function given in Eq. 6 and determine a proper set of localization parameters $\left[\lambda_1, \lambda_2, \ldots \lambda_f\right]$.
3. For $f=1:F$
   (1) Choose λ_f as localization parameter.
   (2) For $n=1:N$

Let $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}\}$ denote calibration samples except $\{\mathbf{x}_n, y_n\}$. Choose $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}\}$ as calibaration samples and $\mathbf{x}_n$ as query sample, and apply the regular LW-PLS algorithm (Algorithm I) to $\{\mathbf{X}_{-n}, \mathbf{y}_{-n}, \mathbf{x}_n\}$ get the output prediction $\hat{y}_n$.

   (3) Calculate the prediction error

$$E_f=\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n-y_n)^2 \tag{A17}$$

4. Choose the localization parameter that results in the lowest prediction error, denote it as λ_k and record the value for each point in similarity function as

$$\varphi_i=\frac{1}{\lambda_k\sigma_{d_i}}(i=1, 2 \ldots N) \tag{A18}$$

5. Determine a Gamma prior distribution over φ based on $\{\varphi_1, \varphi_2 \ldots \varphi_N\}$.

Manuscript received Oct. 6, 2014.