

Process Modeling by Bayesian Latent Variable Regression

Mohamed N. Nounou, Bhavik R. Bakshi*

Prem K. Goel, Xiaotong Shen

Department of Chemical Engineering

Department of Statistics

The Ohio State University, Columbus, OH 43210, USA

Abstract

Large quantities of measured data are being routinely collected in a variety of industries and used for extracting linear models for tasks such as, process control, fault diagnosis and process monitoring. However, existing linear modeling methods do not fully utilize all the information contained in the measurements. This paper presents a new approach for linear process modeling that makes *maximum* use of available process data and process knowledge. This approach, called Bayesian Latent Variable Regression (BLVR), permits extraction and incorporation of knowledge about the statistical behavior of measurements in developing linear process models. Furthermore, unlike existing methods, BLVR is able to handle noise in inputs and outputs, collinear variables, and incorporate prior knowledge about the regression parameters and measured variables. The resulting model is usually more accurate than that obtained by existing methods including, OLS, PCR and PLS. In this paper, BLVR considers a univariate output, and assumes the underlying variables and noise to be Gaussian, but the approach may be easily used for multivariate outputs and other distributions. An empirical Bayes approach is developed to extract the prior information from historical data or from the maximum likelihood solution of available data. Illustrative examples of steady state, dynamic and inferential modeling demonstrate the superior accuracy of BLVR over existing methods even when the assumptions of Gaussian distributions are violated. The relationship between BLVR and existing methods and opportunities for future work based on the proposed framework are also discussed.

1. Introduction

Process models are a core element in many process operation tasks including, model-based control, data reconciliation, and fault detection and diagnosis. Extracting accurate models is important because the performance of these tasks is directly tied with the accuracy of the models used. Linear process models are very commonly used, as they are simple and relatively accurate

* Correspondence should be addressed to Bhavik R. Bakshi. Fax: 1-614-292-3769; Email: bakshi.2@osu.edu

in the operating regions of interest. These models are usually empirical and are estimated from measurements of the process variables. Significant research has focused on developing modeling methods and on their application to process operation tasks. Common areas of application include statistical process monitoring (Kresta et al., 1991; Ku et al., 1995; Nomikos and MacGregor, 1994; Negiz and Cinar, 1997), system identification (Ljung, 1999; Box et al., 1994; Kaspar and Ray, 1993; Wise and Ricker, 1992; MacGregor et al., 1991; Lakshminarayanan et al., 1997; Shi and MacGregor, 2000), inferential modeling (Mejdell and Skogestad, 1991; Kresta et al., 1994; Kano et al., 2000), and many others.

When constructing process models from data, many challenges need to be met. The modeling technique should be able to,

- Account for the presence of measurement noise in all the variables. This noise may have different contributions across the variables and in time.
- Handle collinearity or redundancy in the data, since collinearity tends to increase variability of the estimated model parameters, which can deteriorate the model quality.
- Incorporate external information about variables *and* parameters when available, since such information can greatly enhance the accuracy of the estimated model.

Many techniques have been developed for meeting these challenges. Preprocessing the data by filtering or feature extraction has been popular for decreasing the contribution of noise (Whiteley and Davis, 1992; Bakshi and Stephanopoulos, 1994; Rengaswamy and Venkatasubramanian, 1995; Wong et al., 1998). However, processing the data without accounting for the relationship among the variables may not be very effective as important features may get distorted.

Furthermore, modeling without accounting for the noise can affect the accuracy of model parameters and prediction. Therefore, empirical modeling should be integrated with feature extraction or filtering.

Among existing linear regression methods, Ordinary Least Squares (OLS) regression is the simplest and most widely used. It estimates the model parameters by minimizing the mean square prediction error of the outputs. OLS implicitly assumes noise-free inputs. When this assumption is violated, the accuracy of the estimated model may deteriorate. Rigorous techniques that account for noise in all the variables include, Total Least Squares (TLS) regression (Van Huffel and Vandewalle, 1991; Van Huffel, 1997) and Error-In-Variables (EIV) regression (Kim et. al., 1990 and 1997; Valko and Vajda, 1987). These techniques

simultaneously estimate the model parameters and the underlying noise-free process variables. TLS assumes equal error contributions in all variables, while EIV accounts for different noise content by estimating the model that minimizes the mean square errors of all variables normalized by the noise covariance matrix. TLS and EIV show a noticeable advantage over OLS in estimating the model parameters for full rank data, however, they do not perform well in the presence of collinearity. Various techniques have been developed to account for collinear variables. For example, Ridge Regression (RR) (Hoerl and Kennard, 1970) extends OLS to handle collinear data by penalizing the magnitude of the model parameters. RR introduces a bias in the OLS solution to reduce the variance of the estimated model parameters. However, as in OLS, RR does not account for noise in all variables. Techniques which account for noise in all variables as well as collinearity include various latent variable or reduced rank regression methods such as, Principal Component Regression (PCR) and Partial Least Squares (PLS) regression (Frank and Friedman, 1993; Wold, 1992; and Lorber et. al., 1987). These techniques are very widely used in the chemical industry. PCR first applies PCA to the input data to eliminate collinearity and to capture the data in orthogonal principal component scores and loadings. It then applies OLS to relate the output to the scores. Since PCR transforms the input variables without considering their relationship with the output, the estimated principal components may span a subspace that is not properly aligned with the output space and may not result in the best model. This drawback is overcome by Partial Least Squares (PLS) regression which transforms the input variables to align them with the output space so as to improve model accuracy. Continuum regression unifies OLS, PCR and PLS, and can specialize to methods that lie between these methods, often resulting in improved model accuracy (Stone and Brooks, 1990).

The properties of these commonly used process modeling methods, listed in Table 1, indicate that none of them satisfy all the challenges listed above. In particular, none of these methods uses prior knowledge about the variables and parameters. For example, only EIV accounts for errors in all variables and in varying contributions, but does not account for collinearity. Among methods that do account for collinearity, RR does not account for errors in all variables, while PCR and PLS do not consider different contribution of error in each variable. Furthermore, most commonly used methods in chemical process modeling implicitly assume that no information about the underlying measurements and the model parameters is available, that is, the

measurements and parameters are uniformly distributed. Thus, none of the techniques described above can accommodate knowledge about the underlying data and model.

Advances in sensors, measurement technology, computing and networking have made large quantities of process data readily available in most industries. Knowledge about the underlying variables such as their range of variation and distribution may be extracted from these historical databases. Even if historical data are not available, knowledge about the underlying variables and parameters may be extracted from the data being used for the modeling. If this knowledge can be used in process modeling, it can improve the quality and accuracy of the models, and lead to more efficient processes.

Bayesian statistics provides a formal framework for using such prior knowledge about the variables and parameters to be estimated, as well as techniques for extracting prior information from available measured data. It considers all observable as well as unobservable quantities to be random. This general treatment permits incorporation of external knowledge through a density function called a prior. Bayesian estimation also satisfies the likelihood principle, that is, it uses all the information contained in the measured data about the quantities that need to be estimated. Consequently, a Bayesian approach can account for different extents of noise in all the variables. This likelihood principle is also satisfied by the EIV method. Bayesian statistics also provides a general framework through which existing modeling techniques may be understood better, related to other methods, and improved. Practical situations such as gross errors, bias, and missing data may also be handled in a rigorous manner. These attractive advantages of Bayesian estimation motivate the work described in this paper.

Bayesian methods have received limited attention in process engineering, with Kalman filtering being the most popular Bayesian approach. Bayesian methods have also been developed for data rectification (Tamhane et al., 1988; Johnston and Kramer, 1995; Albuquerque and Biegler, 1996; Bakshi et al., 2001). Statisticians have been working on Bayesian methods for several decades, and have developed many Bayesian linear regression methods. Good descriptions of Bayesian simple and multiple linear regression models are provided by Leamer (1978), Pliz (1983), Press (1989), Gelman et al. (1995), Congdon (2001) and many others. Most of these techniques only focus on using prior knowledge about the model parameters, and do not use prior knowledge about the variables. These methods also do not extract latent variables, which have been used extensively in many process operation tasks such as, monitoring and

diagnosis. Zellner (1971) used Bayesian methods to solve econometric problems, but did not account for noise in all variables or for collinearity. Many Bayesian time series and dynamic modeling techniques have also been developed. Zellner (1971) presented some of the earliest contributions in time series modeling. Later, West and Harrison (1989) showed how Bayesian estimation can be used in forecasting using dynamic models. De Alba et. al. (1995) studied Bayesian inference in ARMA forecasting models and McCulloch et. al. (1994) studied Bayesian analysis of autoregressive time series models. Bayesian methods have been developed for nonlinear regression by neural networks and related methods (Neal, 1996; de Freitas et al., 2000), but these methods also consider prior information only for the model parameters, and do not account for errors in all the variables.

Despite these developments, none of the current Bayesian methods satisfy all the needs of process modeling methods listed earlier, and are often not practical to use. Existing Bayesian regression methods typically require a priori knowledge about the distributions. Such information may not be readily available or may not be practically feasible to obtain due to a lack of familiarity with Bayesian statistics. Recent theoretical advances, and faster computing are making Bayesian methods much more practical, as indicated by their increasing popularity (Malakoff, 1999).

This paper develops a new approach for linear process modeling called Bayesian Latent Variable Regression (BLVR). This approach possesses all the desirable features listed earlier in this section, and in Table 1. Practical methods are developed for estimating the prior from historical data or only from the data available for modeling. This work focuses on variables and noise that are distributed as Gaussian. However, the proposed approach is general, and may be easily used to deal with non-Gaussian errors or variables. Illustrative examples of dynamic and inferential modeling demonstrate the superior performance of the proposed approach when the prior is obtained from different types of measured data, and when the assumption of Gaussian distributions is violated. The relationship between existing methods and BLVR is also discussed, and a general framework from which existing methods may be obtained is suggested.

The rest of this paper is organized as follows. The next section briefly introduces some of the existing linear modeling techniques. This is followed by an introduction to Bayesian estimation and its features. Subsequently, the BLVR methodology is developed. Simplifying assumptions to make the approach practical are discussed. Two variations of the proposed Bayesian Linear

Regression algorithms are presented according to how much the input variables affect the regression parameters. The Bayesian algorithm for prediction from new data and practical methods for estimating the prior parameters are also presented in this section. Finally, illustrative examples demonstrate the benefits of the proposed approach.

2. Existing Linear Modeling Techniques

Given noisy measurements of the input and output data, X (of size $n \times p$) and Y (of size $n \times 1$), such that noise-free variables are related as, $\tilde{Y} = \tilde{X}\tilde{b}$, it is desired to estimate the underlying model parameters and data. The variables are usually assumed to be contaminated with zero-mean additive Gaussian noise, ε_X and ε_Y . Many techniques have been developed to solve this modeling problem, some of which are briefly described below. For notational clarity, the superscripts (\sim and $\hat{\cdot}$) indicate a noise-free variable and an estimated variable, respectively, while no superscript indicates a measured variable. The vector, x_i^T represents the i -th row (measurement) of matrix, X .

2.1 Ordinary Least Squares (OLS)

OLS estimates the model parameters vector, \tilde{b} , by minimizing the sum of squared output prediction error as,

$$\begin{aligned} \{\hat{b}\}_{OLS} &= \underset{\tilde{b}}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \hat{y}_i)^T (y_i - \hat{y}_i) \\ \text{s.t.} \quad \hat{y}_i &= x_i^T \hat{b} \end{aligned} \quad (1)$$

where n is the number of observations, y_i and \hat{y}_i are the i -th measured and predicted output data points, respectively. This minimization problem has the following well-known closed form solution,

$$\{\hat{b}\}_{OLS} = (X^T X)^{-1} X^T Y. \quad (2)$$

Since OLS only minimizes the output prediction error, it implicitly assumes noise-free inputs. When the input variables are contaminated with measurement noise, however, the accuracy of estimated model parameters and prediction may deteriorate. Also, OLS does not account for the presence of collinearity in the input variables, which can result in unreliable estimates of the model parameters since the $(X^T X)$ matrix may be close to singular.

2.2 Total Least Squares (TLS)

TLS improves upon OLS by accounting for measurement errors in all variables. It estimates the model parameters and the underlying noise-free data such that the sum of square errors of all input and output variables is minimized (Van Huffel and Vandewalle, 1991; Van Huffel, 1997) as,

$$\begin{aligned} \{\hat{\mathbf{b}}, \hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i\}_{\text{TLS}} &= \underset{\hat{\mathbf{b}}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T (\mathbf{y}_i - \hat{\mathbf{y}}_i) \right\} \\ \text{s.t.} \quad \hat{\mathbf{y}}_i &= \hat{\mathbf{x}}_i^T \hat{\mathbf{b}} \end{aligned} \quad (3)$$

where \mathbf{x}_i and \mathbf{y}_i are the i -th input and output observations, respectively. This minimization problem has the following closed form solution

$$\left\{ \hat{\mathbf{b}} \right\}_{\text{TLS}} = \frac{-1}{v_{22}} v_{12} \quad (4)$$

where,

$$[\mathbf{X} \ \mathbf{Y}] = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (5)$$

is the singular value decomposition of the augmented matrix $[\mathbf{X} \ \mathbf{Y}]$,

$$\text{and} \quad \mathbf{V}^T = \begin{bmatrix} v_{11} & v_{12} \\ \underbrace{v_{21}}_r & \underbrace{v_{22}}_{\text{one}} \end{bmatrix} \quad (6)$$

where, p is the number of input variables. TLS is an improvement over OLS but it assumes equal noise contribution in all variables. When this assumption is violated, variables with larger noise contributions are given more importance in estimating the model parameters than those with lower error contributions as the noise is misinterpreted as variations in the noise-free data. Also, TLS does not account for collinearity in the input variables.

2.3 Error-In-Variables (EIV) Modeling

EIV modeling is a maximum likelihood estimator for Gaussian errors and has been extensively used for parameter estimation, data reconciliation, and gross error detection of noisy measurements (Kim et. al., 1990; 1997). It simultaneously estimates the model parameters and reconciles the data by maximizing the likelihood or probability that the estimated model fits the data as

$$\begin{aligned} \{\hat{\mathbf{b}}, \hat{\mathbf{X}}, \hat{\mathbf{Y}}\}_{\text{EIV}} &= \arg \max_{\hat{\mathbf{b}}, \hat{\mathbf{X}}, \hat{\mathbf{Y}}} L(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}; \mathbf{X}, \mathbf{Y}) = \arg \max_{\hat{\mathbf{b}}, \hat{\mathbf{X}}, \hat{\mathbf{Y}}} P(\mathbf{X}, \mathbf{Y} | \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \\ \text{s.t.} \quad \tilde{\mathbf{Y}} &= \tilde{\mathbf{X}}\hat{\mathbf{b}}. \end{aligned} \quad (7)$$

If the distribution of the errors is assumed to be Gaussian and the input and output noise to be independent, maximizing the likelihood is equivalent to minimizing the sum of squared input and output errors normalized by their error covariance matrices as,

$$\{\hat{\mathbf{b}}, \hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i\}_{\text{EIV}} = \arg \min_{\hat{\mathbf{b}}, \hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i} \left\{ \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{Q}_{\varepsilon_Y}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \right\} \quad (8)$$

where $\mathbf{Q}_{\varepsilon_X}$ and $\mathbf{Q}_{\varepsilon_Y}$ are assumed to be known and subject to the same constraint shown in Equation (7). This EIV optimization problem can be solved as two nested optimization problems, one solving for the model parameters and the other for the reconciled data as,

$$\begin{aligned} \{\hat{\mathbf{b}}\}_{\text{EIV}} &= \arg \min_{\hat{\mathbf{b}}} \left\{ (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{Q}_{\varepsilon_Y}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \right\} \\ \text{s.t.} \\ \{\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i\}_{\text{EIV}} &= \arg \min_{\hat{\mathbf{y}}_i, \hat{\mathbf{x}}_i} \left\{ (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{Q}_{\varepsilon_Y}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \right\} \end{aligned} \quad (9)$$

subject to the model constraint given by Equation (7). The EIV data reconciliation problem has the following closed form solution as shown in Appendix I,

$$\begin{aligned} \{\hat{\mathbf{x}}_i\}_{\text{EIV}} &= \left[\mathbf{Q}_{\varepsilon_X}^{-1} + \hat{\mathbf{b}} \mathbf{Q}_{\varepsilon_Y}^{-1} \hat{\mathbf{b}}^T \right]^{-1} \left(\mathbf{Q}_{\varepsilon_X}^{-1} \mathbf{x}_i + \hat{\mathbf{b}} \mathbf{Q}_{\varepsilon_Y}^{-1} \mathbf{y}_i \right) \\ \{\hat{\mathbf{y}}_i\}_{\text{EIV}} &= \hat{\mathbf{x}}_i^T \hat{\mathbf{b}} \end{aligned} \quad (10)$$

The EIV solution reduces to the TLS solution when $\mathbf{Q}_{\varepsilon_X}$ and $\mathbf{Q}_{\varepsilon_Y}$ are multiples of the identity matrix. The EIV method is a powerful technique, but it does not account for collinearity in the variables, making it unreliable when collinearity exists.

2.4 Ridge Regression (RR)

RR was introduced as a method for stabilizing the OLS regression estimates in the presence of extreme collinearity, i.e., the input covariance matrix $(\mathbf{X}^T \mathbf{X})$ being singular or nearly so (Hoerl and Kennard, 1970; Frank et. al., 1993). It decreases the variance of the estimated regression coefficients by imposing a penalty on their magnitude as,

$$\{\hat{\mathbf{b}}\}_{\text{RR}} = \arg \min_{\hat{\mathbf{b}}} \left\{ \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T (\mathbf{y}_i - \hat{\mathbf{y}}_i) + \lambda \hat{\mathbf{b}}^T \hat{\mathbf{b}} \right\}, \quad (11)$$

where, λ is a positive number. This minimization problem has the following closed form solution,

$$\{\hat{\mathbf{b}}\}_{\text{RR}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (12)$$

It can be seen from Equation (12) that the parameter λ stabilizes the estimated model parameters by increasing the rank of the input data matrix to the actual number of inputs to avoid inversion problems. The value of λ can be estimated using a model selection procedure, such as cross validation (Golub et. al., 1979). Like OLS, RR does not account for noise in the input variables.

2.5 Principal Component Regression (PCR)

PCR accounts for collinearity in the data by reducing the rank of the input data matrix. It combines PCA and OLS to handle collinearity in the input variables (Massy, 1965). First, it reduces the dimension of the input variables using PCA, and then it applies OLS to project the output on the retained principal components. Since the noise-free PCA model has the form

$$\tilde{\mathbf{X}} = \tilde{\mathbf{Z}} \tilde{\boldsymbol{\alpha}}^T, \quad (13)$$

The PCR model can be written as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}} \tilde{\mathbf{b}} = \tilde{\mathbf{Z}} \tilde{\boldsymbol{\alpha}}^T \tilde{\mathbf{b}} = \tilde{\mathbf{Z}} \tilde{\boldsymbol{\beta}}$$

where, $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\alpha}}^T \tilde{\mathbf{b}}.$ (14)

PCR can be formulated as two consecutive optimization problems,

$$\begin{aligned} \text{I.} \quad & \{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{z}}_i\}_{\text{PCA}} = \arg \min_{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{z}}_i} \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T (\mathbf{x}_i - \hat{\mathbf{x}}_i) \\ \text{s.t.} \quad & \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i, \text{ and} \quad \hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{\alpha}} = \mathbf{I}. \end{aligned} \quad (15a)$$

$$\begin{aligned} \text{II.} \quad & \{\hat{\boldsymbol{\beta}}\}_{\text{OLS}} = \arg \min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^n (y_i - \hat{y}_i)^T (y_i - \hat{y}_i) \\ \text{s.t.} \quad & \hat{y}_i = \hat{\mathbf{z}}_i^T \hat{\boldsymbol{\beta}}. \end{aligned} \quad (15b)$$

These optimization problems show that PCR eliminates some of the noise by reducing the dimension of the input variables. However, it does not account for possible varying error contributions in different variables. When all principal components are retained, PCR reduces to OLS.

2.6 Partial Least Squares (PLS)

PLS regression uses the same model structure used in PCR, but extends PCR to consider the output variables in computing the principal components. It determines the projection directions that capture the variations in the input variables and which are closest to the output by maximizing the following objective function (Wold, 1982),

$$\{\hat{\alpha}\}_{\text{PLS}} = \arg \max_{\hat{\alpha}} \{ \text{corr}^2(y, X\hat{\alpha}) \text{var}(X\hat{\alpha}) \}. \quad (16)$$

A similar formulation of PLS has also been used to extend PLS to deal with nonlinear problems (Malthouse et. al., 1997). In their formulation, the projection directions are estimated by minimizing the sum of input and output errors as,

$$\{\hat{\alpha}\}_{\text{PLS}} = \arg \min_{\hat{\alpha}} \left\{ \sum_{i=1}^n (x_i - \hat{x}_i)^T (x_i - \hat{x}_i) + \sum_{i=1}^n (y_i - \hat{y}_i)^T (y_i - \hat{y}_i) \right\} \quad (17)$$

subject to the constraints shown in Equations 15a and 15b. Like PCR, PLS also does not account for varying noise contributions in different variables.

3. Introduction to Bayesian Estimation

3.1. Basic Principles

A distinctive feature of Bayesian estimation is its assumption that all quantities, observable and unobservable, are random with a joint probability density function that describes their behavior (Kadane, 1995; Gelman et. al., 1995). This perspective is different from that adopted by non-Bayesian methods, which consider the quantities of interest as fixed unknown quantities to be determined by minimizing some objective function of the estimation errors. The methods commonly used for process modeling including those discussed in Section 2 are of this type. This assumption of all quantities being random allows Bayesian methods to incorporate external prior knowledge about the quantities of interest into the estimation problem. To estimate the quantity $\tilde{\theta}$, from a set of measurements of the quantity, y , Bayesian estimation starts by defining the conditional density of the variable to be estimated given the measurements, $P(\tilde{\theta} | y)$, which is called the posterior. The *posterior* is a density function that describes the behavior of the quantity, $\tilde{\theta}$, after observing the measurements. Using Bayes rule, the posterior can be written as follows,

$$P(\tilde{\theta} | y) = \frac{P(y | \tilde{\theta})P(\tilde{\theta})}{P(y)}. \quad (18)$$

The first term in the numerator of Equation (18) denotes the *likelihood* function, which is the conditional density of the observations given the true value of $\tilde{\theta}$. According to the Likelihood Principle (LP), the likelihood function contains all information brought by the observations, y , about the quantity, $\tilde{\theta}$. The second term in the numerator is the *prior*, which is the density function of the quantity $\tilde{\theta}$. It is called a prior since it quantifies our belief or knowledge about $\tilde{\theta}$ *before* observing the measurements. Through the prior, external knowledge about the quantity $\tilde{\theta}$ can be incorporated into the estimation problem. Finally, the denominator term is the density function of the observation, which can be assumed constant after observing the data. The posterior density can be written as,

$$P(\tilde{\theta} | y) \propto P(y | \tilde{\theta})P(\tilde{\theta})$$

or,

$$Posterior \propto Likelihood \times Prior, \quad (19)$$

which is sometimes referred to as the unnormalized posterior. Thus, the posterior combines the data information and any external information. Having constructed the posterior, a sample from it is selected as the final Bayesian estimate of the quantity $\tilde{\theta}$. In contrast to non-Bayesian or frequentist approaches, which rely only on the data for inference, Bayesian approaches combine the information brought by the data and any external knowledge represented by the prior to provide improved estimates.

3.2. General Methodology

The main steps in Bayesian estimation can be outlined as follows (Gelman et. al., 1995):

- i. Set up a full probability model (a joint probability density function) of all observable and unobservable quantities. This is possible since all variables are considered to be random.
- ii. Calculate the conditional density of the variables to be estimated given the observed data (posterior).
- iii. Evaluate the implication of the posterior and check the accuracy of the estimated quantities.

The second step is a mathematical one, and involves computing the posterior density function. When the likelihood and the prior densities are mathematically simple, such computation can be done analytically. However, for more complicated problems, it is usually done empirically by some sampling algorithm, such as Markov Chain Monte Carlo (MCMC), (Gilks et. al., 1996). The third step is more judgmental, since it requires a decision about the sample to be selected from the distribution of the posterior as the final Bayesian estimate. The first step, however, is usually the hardest since it involves defining the likelihood and prior density functions to be used in estimation, which usually are not completely defined. These steps of the Bayesian approach are schematically illustrated in Figure 1, which shows that the posterior density combines data and external information in one density function, from which a sample is chosen as the Bayesian estimate such that a predefined loss function is minimized.

3.3. Loss Function

The loss function, $L(\tilde{\theta}; \hat{\theta})$, is related to a utility or an objective function that decides which sample from the posterior may be selected as the Bayesian estimate. Here, $\hat{\theta}$ denotes the Bayesian estimate of the quantity θ . There are many loss functions that can be used such as, a quadratic loss function, a zero-one loss function, and others (Robert, 1994). A quadratic loss function defines a penalty of the squared error between the estimated and the true quantity, and corresponds to selecting the posterior mean as the Bayesian estimate. A zero-one loss function imposes a penalty of zero when the selected sample is the true one and a penalty of unity otherwise, i.e.,

$$L(\hat{\theta}; \tilde{\theta}) = \begin{cases} 0 & \text{when } \{\hat{\theta}\}_{\text{Bayesian}} = \tilde{\theta} \\ 1 & \text{otherwise} \end{cases}. \quad (20)$$

The use of a zero-one loss function corresponds to choosing the posterior mode or maximum as the Bayesian estimate, which is usually referred to as the maximum a posteriori (MAP) estimate. Thus,

$$\{\hat{\theta}\}_{\text{MAP}} = \arg \max_{\tilde{\theta}} P(y | \tilde{\theta})P(\tilde{\theta}). \quad (21)$$

One advantage of using a zero-one loss function is that it reduces Bayesian modeling into a minimization problem, which facilitates the comparison between BPCA and other existing methods, in which other objective functions are minimized. Also, a zero-one loss function can

be more computationally efficient as the Bayesian estimate of the data often has a closed form solution.

4. Bayesian Latent Variable Regression

The Bayesian Latent Variable Regression (BLVR) model may be represented by the following equations.

$$\tilde{Y} = \tilde{X}\tilde{b} = \tilde{Z}\tilde{\alpha}^T\tilde{b} = \tilde{Z}\tilde{\beta}, \text{ such that, } \tilde{\alpha}^T\tilde{\alpha} = I$$

This model is similar in form to that of PCR and PLS, but unlike existing methods, the approach for estimating the model parameters is Bayesian, as developed in the rest of this section.

4.1 Basic Formulation

BLVR modeling involves estimation of four parameters: projection directions, $\tilde{\alpha}$, principal components, \tilde{Z} , regression parameters, $\tilde{\beta}$, and model rank, \tilde{r} . Thus, from a Bayesian perspective, the posterior should be defined as the conditional density of these quantities given the measurements, X and Y . Using Bayes rule, the posterior may be written as

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{r}, \tilde{\beta} | X, Y) = \frac{P(X, Y | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r})P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r})}{P(X, Y)}. \quad (22)$$

The first term in the numerator is the likelihood function, which is the conditional density of the measured inputs and output given the noise-free model and data, and the second term is the prior density function. The denominator is the density function of the measurements, which is a normalization constant for a given set of measurements. Therefore, the unnormalized posterior can be written as

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{r}, \tilde{\beta} | X, Y) \propto P(X, Y | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r})P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}). \quad (23)$$

This formulation is similar to the BPCA formulation derived by Nounou et. al. (2001) except that in BPCA, it is only desired to estimate the projection directions, α , principal components, Z , and the model rank, r . Thus, the relationship between BPCA and BLVR is analogous to that between PCA and PCR or PCA and PLS.

The prior can be a complicated multivariate function since it represents the joint density function of the noise-free latent variables, projection directions of loadings, model rank, and regression parameters. Since the model depends on the assumed rank, the prior can be written as,

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) = P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta} | \tilde{r})P(\tilde{r}). \quad (24)$$

Note that $P(\tilde{r})$ is a discrete density function, which can be defined as

$$P(\tilde{r} = j) = k_j, \text{ such that, } \sum_j k_j = 1. \quad (25)$$

Also, the conditional density function of the loadings, latent variables, and regression parameters given the number of principal components can be expressed using the multiplication rule for probabilities as,

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta} | \tilde{r}) = P(\tilde{Z}, \tilde{\beta} | \tilde{\alpha}, \tilde{r})P(\tilde{\alpha} | \tilde{r}). \quad (26)$$

Assuming that \tilde{Z} and $\tilde{\beta}$ are independent, the prior becomes,

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta} | \tilde{r}) = P(\tilde{Z} | \tilde{\alpha}, \tilde{r})P(\tilde{\beta} | \tilde{\alpha}, \tilde{r})P(\tilde{\alpha} | \tilde{r}). \quad (27)$$

This assumption is usually valid since the regression parameters, $\tilde{\beta}$, represent the relationship between the inputs and outputs, which does not depend on the behavior of either set alone. Thus, the unnormalized posterior can be written as,

$$P(\tilde{Z}, \tilde{\alpha}, \tilde{r}, \tilde{\beta} | X, Y) \propto P(X, Y | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r})P(\tilde{Z} | \tilde{\alpha}, \tilde{r})P(\tilde{\beta} | \tilde{\alpha}, \tilde{r})P(\tilde{\alpha} | \tilde{r})P(\tilde{r}). \quad (28)$$

4.2 Simplifying Assumptions

Computing the posterior requires information about the structures of the likelihood and prior density functions, which depend on the nature of the noise and noise-free data, respectively. These density functions may be determined by parametric or non-parametric approaches. The parametric approach assumes a form of the distribution and estimates the parameters based on the data. The non-parametric approach uses numerical methods such as, Markov Chain Monte Carlo simulation to generate samples that represent the density functions. The parametric approach is usually simpler and less computationally expensive. Additional information about the trade-offs between these approaches is available in references such as, Johnston and Kramer (1998) and Silverman (1986). The approach developed in this paper is parametric, and is based on the following simplifying assumptions.

Known Model Rank. Estimating a reduced-rank model requires specifying the model rank. Assuming that the true model rank, \tilde{r} , is known implies that

$$P(\tilde{r}) = 1. \quad (29)$$

This assumption reduces Equation 27 to,

$$P(\tilde{\mathbf{Z}}|\tilde{\alpha})P(\tilde{\beta}|\tilde{\alpha})P(\tilde{\alpha}), \quad (30)$$

and the posterior in Equation 28 to,

$$P(\mathbf{X}, \mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})P(\tilde{\mathbf{Z}}|\tilde{\alpha})P(\tilde{\beta}|\tilde{\alpha})P(\tilde{\alpha}). \quad (31)$$

In practice, however, the true model rank is unknown and needs to be estimated. An overview of rank estimation techniques is presented in Section 4.6.

Gaussian Likelihood Function. For linear models, the structure of the likelihood function depends on the nature of the noise. If it is assumed that the measured inputs and outputs are contaminated with additive zero mean Gaussian noise, that is, $\varepsilon_X \sim N(0, \mathbf{Q}_{\varepsilon_X})$ and $\varepsilon_Y \sim N(0, \mathbf{Q}_{\varepsilon_Y})$, and that the input and output noise are independent, that is, $E[\varepsilon_X^T \varepsilon_Y] = 0$, then the likelihood function becomes the product of the following two density functions,

$$P(\mathbf{X}, \mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) = P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}), \quad (32)$$

Under the normality and independence assumptions of the noise, both density functions are normal with the following moments,

$$E[\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}] = E[\tilde{\mathbf{X}} + \varepsilon_X] = \mu_{\tilde{\mathbf{X}}}, \quad (33)$$

$$\text{Cov}[\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}] = E[(\mathbf{X} - \mu_{\tilde{\mathbf{X}}})^T (\mathbf{X} - \mu_{\tilde{\mathbf{X}}})] = \mathbf{Q}_{\varepsilon_X}, \quad (34)$$

$$E[\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}] = E[\tilde{\mathbf{Y}} + \varepsilon_Y] = \mu_{\tilde{\mathbf{Y}}}, \quad (35)$$

and

$$\text{Cov}[\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}] = E[(\mathbf{Y} - \mu_{\tilde{\mathbf{Y}}})^T (\mathbf{Y} - \mu_{\tilde{\mathbf{Y}}})] = \mathbf{Q}_{\varepsilon_Y}. \quad (36)$$

Therefore,

$$P(\mathbf{X}, \mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) = N(\mu_{\tilde{\mathbf{X}}}, \mathbf{Q}_{\varepsilon_X})N(\mu_{\tilde{\mathbf{Y}}}, \mathbf{Q}_{\varepsilon_Y}) \quad (37)$$

Zero-One Loss Function. This work uses a zero-one loss function of the form,

$$L(\hat{\mathbf{Z}}, \hat{\alpha}, \hat{\beta}; \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}) = \begin{cases} 0 & \{\hat{\mathbf{Z}}, \hat{\alpha}, \hat{\beta}\}_{\text{Bayes}} = \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta} \\ 1 & \text{otherwise} \end{cases} \quad (38)$$

This type of loss function defines the posterior mode as the Bayesian model estimate, which is usually called the Maximum a Posteriori (MAP) estimate. Thus, the Bayesian estimate of the reduced rank model can be obtained by the following objective function,

$$\{\hat{Z}, \hat{\alpha}, \hat{\beta}\}_{\text{MAP}} = \underset{\tilde{Z}, \tilde{\alpha}, \tilde{\beta}}{\text{argmax}} P(\mathbf{X} | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(\mathbf{Y} | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(\tilde{Z} | \tilde{\alpha}) P(\tilde{\beta} | \tilde{\alpha}) P(\tilde{\alpha}). \quad (39)$$

Uniform $P(\tilde{\alpha})$. Since the noise-free projection directions matrix has to satisfy the orthonormality constraint, $\tilde{\alpha}^T \tilde{\alpha} = \mathbf{I}$, each of its elements has to be bounded between -1 and 1 . Assuming that no external knowledge is available about the projection directions, the following uniform prior can be used to describe their behavior,

$$\tilde{\alpha}_j \sim \text{U}(-1, 1). \quad (40)$$

This means that $P(\tilde{\alpha})$ is constant over the interval $[-1, 1]$, and thus drops from the objective function (39), reducing the BLVR objective function to,

$$\{\hat{Z}, \hat{\alpha}, \hat{\beta}\}_{\text{MAP}} = \underset{\tilde{Z}, \tilde{\alpha}, \tilde{\beta}}{\text{argmax}} P(\mathbf{X} | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(\mathbf{Y} | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(\tilde{Z} | \tilde{\alpha}) P(\tilde{\beta} | \tilde{\alpha}). \quad (41)$$

The BPCA formulation developed by Nounou et. al. (2001) uses a Gaussian prior for the projection directions instead of the uniform prior suggested above. The Gaussian prior does improve the accuracy of estimated projection directions, but in this BLVR problem, a uniform prior for the projection directions is used to reduce the computational complexity of the modeling problem.

Gaussian Underlying Inputs. Defining the structure of the conditional density, $P(\tilde{Z} | \tilde{\alpha})$, requires an assumption about the nature of the noise-free data. In this work, the underlying noise-free input variables are assumed to follow a normal distribution, that is, $\tilde{X} \sim N(\mu_{\tilde{X}}, Q_{\tilde{X}})$. This assumption, as verified in Section 5 through illustrative examples, is not a bad one, since the distributions of many types of data, particularly from linear systems, can be approximated quite well by a Gaussian distribution. Such assumptions are commonly made in popular Bayesian methods such as Kalman Filtering. As discussed in more detail in Section 5, even when this assumption is not satisfied, the results can still be better than those obtained by existing methods. A multiscale formulation of BLVR (Nounou, 2000) makes the assumption of Gaussian distributions even more accurate due to the decorrelation ability of wavelets (Bakshi et al., 2001). According to this normality assumption, the conditional density of the principal component scores given the projection directions will also be normal with the following moments,

$$E[\tilde{Z} | \tilde{\alpha}] = E[\tilde{X}] \tilde{\alpha} = \mu_{\tilde{X}} \tilde{\alpha}, \quad (42)$$

$$\text{Cov}[\tilde{Z} | \tilde{\alpha}] = \tilde{\alpha}^T E[(\tilde{X} - \mu_{\tilde{X}})^T (\tilde{X} - \mu_{\tilde{X}})] \tilde{\alpha} = \tilde{\alpha}^T Q_{\tilde{X}} \tilde{\alpha}. \quad (43)$$

Therefore,

$$\tilde{Z} | \tilde{\alpha} \sim N(\mu_{\tilde{Z}|\tilde{\alpha}}, Q_{\tilde{Z}|\tilde{\alpha}}) = N(\mu_{\tilde{X}} \tilde{\alpha}, \tilde{\alpha}^T Q_{\tilde{X}} \tilde{\alpha}). \quad (44)$$

Gaussian Model Parameters. Finally, for computational simplification, the regression vector, $\tilde{\mathbf{b}}$, is assumed to follow a normal distribution, i.e.,

$$\tilde{\mathbf{b}} \sim N(\mu_{\tilde{\mathbf{b}}}, Q_{\tilde{\mathbf{b}}}). \quad (45)$$

Under this assumption and using the relationship, $\tilde{\beta} = \tilde{\alpha}^T \tilde{\mathbf{b}}$, the conditional distribution of the reduced-rank model parameters given the projection directions follows the following normal distribution,

$$\tilde{\beta} | \tilde{\alpha} \sim N(\mu_{\tilde{\beta}|\tilde{\alpha}}, Q_{\tilde{\beta}|\tilde{\alpha}}) = N(\tilde{\alpha}^T \mu_{\tilde{\mathbf{b}}}, \tilde{\alpha}^T Q_{\tilde{\mathbf{b}}} \tilde{\alpha}). \quad (46)$$

4.3 BLVR Algorithms

This section presents two variations of BLVR depending on the nature of the posterior used for estimating the latent variables and regression parameters. Algorithm I uses the posterior defined by Equation (41). In contrast, Algorithm II follows the spirit of many existing methods such as PCR and OLS, and estimates the latent variables without including prior knowledge about the outputs, and the regression parameters without including the prior for the inputs. As illustrated in Section 5, BLVR-II can do better prediction if all the prior distributions are assumed to be non-informative or uniform. With more accurate prior distributions, BLVR-I performs slightly better.

Bayesian Latent Variable Regression Algorithm I. The first BLVR algorithm (BLVR-I) is the MAP estimator of the reduced-rank model, which considers all parts of the posterior to be equally important. The MAP solution can be obtained by solving the following two simultaneous parameter estimation and data reconciliation optimization problems. The outer optimization problem solves for the model parameters and inner optimization problem solves for the data given the parameters as,

$$\{\hat{\alpha}, \hat{\beta}\}_{\text{MAP}} = \underset{\tilde{\alpha}, \tilde{\beta}}{\text{argmax}} P(X | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(Y | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r}) P(\tilde{\beta} | \tilde{\alpha})$$

s.t.

$$\{\hat{Z}\}_{\text{MAP}} = \underset{\tilde{Z}, \tilde{\alpha}}{\operatorname{argmax}} P(X | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) P(Y | \tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}) P(\tilde{Z} | \tilde{\alpha}).$$

s.t.

$$\hat{Y} = \hat{Z}\hat{\beta}, \hat{X} = \hat{Z}\hat{\alpha}^T, \text{ and } \hat{\alpha}^T \hat{\alpha} = I. \quad (47)$$

Under the simplifying assumption made in Section 4.2, all densities in the posterior are defined as multivariate normal, and thus the MAP solution of the Bayesian model can be equivalently obtained by solving the following simultaneous minimization problems for the model parameters and the reconciled data as follows,

$$\{\hat{\alpha}, \hat{\beta}\}_{\text{MAP}} = \underset{\hat{\alpha}, \hat{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + \sum_{i=1}^n (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + (\hat{\beta} - \mu_{\tilde{\beta}\tilde{\alpha}})^T Q_{\tilde{\beta}\tilde{\alpha}}^{-1} (\hat{\beta} - \mu_{\tilde{\beta}\tilde{\alpha}}) \right\} \quad (48a)$$

s.t.

$$\{\hat{z}_i\}_{\text{MAP}} = \underset{\hat{z}_i}{\operatorname{argmin}} \left\{ (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + (\hat{z}_i - \mu_{\tilde{z}\tilde{\alpha}})^T Q_{\tilde{z}\tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{\tilde{z}\tilde{\alpha}}) \right\} \quad (48b)$$

$$\hat{y}_i = \hat{z}_i^T \hat{\beta}, \hat{x}_i = \hat{\alpha}^T \hat{z}_i, \text{ and } \hat{\alpha}^T \hat{\alpha} = I \quad (48c)$$

where the moments of the densities, $\tilde{Z} | \tilde{\alpha}$ and $\tilde{\beta} | \tilde{\alpha}$, are given in Equations (44) and (46), respectively. The data reconciliation problem defined by Equations 48b and 48c has the following closed form solution as shown in Appendix II:

$$\{\hat{z}_i\}_{\text{MAP}} = \left[Q_{\varepsilon_z}^{-1} + Q_{\tilde{z}\tilde{\alpha}}^{-1} + \hat{\beta} Q_{\varepsilon_y}^{-1} \hat{\beta}^T \right]^{-1} \left(Q_{\varepsilon_z}^{-1} z_i + Q_{\tilde{z}\tilde{\alpha}}^{-1} \mu_{\tilde{z}\tilde{\alpha}} + \hat{\beta} Q_{\varepsilon_y}^{-1} y_i \right)$$

and

$$\{\hat{y}_i\}_{\text{MAP}} = \hat{\beta}^T \{\hat{z}_i\}_{\text{MAP}}. \quad (49)$$

Algorithm I, described above, is a Bayesian and reduced-rank version of the EIV method since it estimates all model parameters and the noise-free data by minimizing the sum of squared input and output errors. In fact, Algorithm I reduces exactly to the EIV method discussed in Section 2.3 when a uniform prior is used with a full rank model. Therefore, Algorithm I extends EIV modeling to handle collinearity and to incorporate external prior information about the true model and underlying data.

However, as will be shown through illustrative examples in Section 5, the EIV method is a good estimator of the available data and model parameters for full rank models, but is not always a good predictor. In other words, EIV estimates the model parameters that provide good estimates of the available data. These parameters, however, do not always provide good

predictions for unseen data. The reason behind this observation is that the output is not given much consideration when estimating the model parameters by EIV or Algorithm I. In fact, the output is considered as important as any other input. This can be seen clearly from Equation 48a, which minimizes the sum of squared errors of all inputs and outputs. Therefore, even though Algorithm I is expected to outperform EIV in estimating the noise-free model parameters and data, its prediction ability using unseen data for a uniform prior is not always superior to other more specialized predictive modeling methods, such as PCR and PLS.

Bayesian Latent Variable Regression Algorithm II. This variation of BLVR-I can provide more accurate prediction of output variables when an accurate prior is not available. BLVR-II eliminates the effect of the input part of the likelihood density when estimating the model parameter vector, $\tilde{\beta}$. Thus, the following algorithm is obtained:

$$\begin{aligned} \{\hat{\alpha}\} &= \underset{\tilde{\alpha}}{\operatorname{argmax}} P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) \\ \text{s.t.} \\ \{\hat{\mathbf{Z}}\} &= \underset{\tilde{\mathbf{Z}}}{\operatorname{argmax}} P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) P(\tilde{\mathbf{Z}} | \tilde{\alpha}) \\ \{\hat{\beta}\} &= \underset{\tilde{\beta}}{\operatorname{argmax}} P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}}) P(\tilde{\beta} | \tilde{\alpha}) \end{aligned} \quad (50)$$

subject to the same constraints shown in Equation (47). Again, assuming that all the simplifying assumptions described in Section 4.2 still hold, Algorithm II becomes,

$$\begin{aligned} \{\hat{\alpha}\} &= \underset{\hat{\alpha}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{Q}_{\varepsilon_y}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) \right\} \\ \text{s.t.} \\ \{\hat{\mathbf{z}}_i\} &= \underset{\hat{\mathbf{z}}_i}{\operatorname{argmin}} \left\{ (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\alpha}})^T \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\alpha}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\alpha}}) \right\} \\ \{\hat{\beta}\} &= \underset{\hat{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^T \mathbf{Q}_{\varepsilon_y}^{-1} (\mathbf{y}_i - \hat{\mathbf{y}}_i) + (\hat{\beta} - \boldsymbol{\mu}_{\tilde{\beta}|\tilde{\alpha}})^T \mathbf{Q}_{\tilde{\beta}|\tilde{\alpha}}^{-1} (\hat{\beta} - \boldsymbol{\mu}_{\tilde{\beta}|\tilde{\alpha}}) \right\} \\ \hat{\mathbf{y}}_i &= \hat{\mathbf{z}}_i^T \hat{\beta}, \quad \hat{\mathbf{x}}_i = \hat{\alpha} \hat{\mathbf{z}}_i, \quad \text{and} \quad \hat{\alpha}^T \hat{\alpha} = \mathbf{I} \end{aligned} \quad (51)$$

in which the outer optimization function solves for the projection directions, and the inner optimization functions solve for the input-output parameters and the estimated data. Both of the

inner optimization problems can be solved analytically, and have the following closed form solutions as proved in Appendices III and IV,

$$\{\hat{z}_i\} = (\hat{\alpha}^T Q_{\varepsilon_x}^{-1} \hat{\alpha} + Q_{\tilde{z}}^{-1})^{-1} (\hat{\alpha}^T Q_{\varepsilon_x}^{-1} x_i + Q_{\tilde{z}|\hat{\alpha}}^{-1} \mu_{\tilde{z}|\hat{\alpha}}) \quad (52)$$

and,

$$\{\hat{\beta}\} = (\hat{Z}^T \hat{Z} + Q_h^{-1})^{-1} (\hat{Z}^T Y + Q_h^{-1} \mu_{\tilde{\beta}|\hat{\alpha}}) \quad (53)$$

where $Q_h = Q_{\tilde{\beta}|\hat{\alpha}} Q_{\varepsilon_y}^{-1}$, assuming a single-output model.

The maximum likelihood version of this algorithm has a better prediction ability for unseen data than Algorithm I since it gives special consideration to the output when estimating the model parameters. In fact, like PCR and PLS, Algorithm II accounts for collinearity and focuses on the model's prediction when estimating its parameters. However, it has advantages over PCR and PLS since it can account for different noise contents in different variables and it can incorporate external knowledge about the model parameters and data.

4.4 Bayesian Prediction

A common use of the model obtained from the BLVR algorithms is to predict the output for new input measurements. This section formulates and solves the Bayesian prediction problem so that the BLVR model can be used to predict the output from new data. In prediction, it is desired to estimate the noise-free output given the noisy input data, the estimated model projection directions, $\hat{\alpha}$, and regression parameters, $\hat{\beta}$. Since the estimated output is linearly related to the estimated inputs, prediction can be performed by estimating the noise-free input data or principal component scores from the measured inputs, and then using the estimated model parameters to estimate the output. Therefore, the posterior density for Bayesian prediction can be defined as the conditional density of the noise-free principal components given the measured inputs and estimated projection directions, which can be written using Bayes rule as,

$$P(\tilde{Z} | X, \hat{\alpha}) \propto P(X | \tilde{Z}, \hat{\alpha}) P(\tilde{Z} | \hat{\alpha}) \quad (54)$$

Again, assuming a zero-one loss function, the problem reduces to,

$$\{\hat{z}\}_{\text{Pred}} = \underset{\tilde{z}}{\operatorname{argmax}} P(X | \tilde{z}, \hat{\alpha}) P(\tilde{z} | \hat{\alpha}). \quad (55)$$

Based on the assumptions made in Section 4.2, Bayesian prediction reduces to the following minimization problem,

$$\begin{aligned} \{\hat{z}_i\}_{\text{Pred}} &= \underset{\hat{z}_i}{\text{argmin}} \left\{ (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (\hat{z}_i - \mu_{z|\tilde{\alpha}})^T Q_{z|\tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{z|\tilde{\alpha}}) \right\} \\ \text{s.t.}, \quad \hat{z}_i &= \hat{\alpha}^T \hat{x}_i. \end{aligned} \quad (56)$$

which has the following closed form solution,

$$\{\hat{z}_i\}_{\text{Pred}} = \left(\hat{\alpha}^T Q_{\varepsilon_x}^{-1} \hat{\alpha} + Q_z^{-1} \right)^{-1} \left(\hat{\alpha}^T Q_{\varepsilon_x}^{-1} x_i + Q_{z|\tilde{\alpha}}^{-1} \mu_{z|\tilde{\alpha}} \right). \quad (57)$$

Once the latent variables are estimated, the predicted output can simply be computed as follows,

$$\{\hat{y}_i\}_{\text{Pred}} = \hat{\beta}^T \{\hat{z}_i\}_{\text{Pred}}. \quad (58)$$

4.5 Estimating the Prior Density

In the Bayesian modeling algorithms developed in Section 4.3, the structures of the densities, $P(\tilde{\mathbf{b}})$, and $P(\tilde{Z}|\tilde{\alpha})$, were assumed to be multivariate normal, and their parameters, $\mu_{\tilde{x}}$, $Q_{\tilde{x}}$, $\mu_{\tilde{b}}$, and $Q_{\tilde{b}}$ were also assumed to be known. In other words, the prior density was assumed to be defined a priori. Estimating the prior distribution is a critical part of any Bayesian approach. This section describes practical methods for estimating the prior for the BLVR algorithms based on the assumptions presented in previous sections.

Traditional Bayesian analysis assumes a fully predefined prior density. In practice, however, the necessary prior knowledge may not be available, and parts or the entire prior distribution might be unspecified. Empirical Bayesian analysis (Gelman et. al., 1995; Maritz, 1970) is an approach for estimating the prior from the available data. There are two general approaches for estimating the prior empirically: a parametric approach and a non-parametric approach. In the parametric approach, the structure of the prior distribution is defined first, and then the data are used to estimate its hyperparameters. In the non-parametric approach, the entire prior distribution is estimated from the data, making it more challenging and computationally more demanding than the parametric approach. This paper uses the parametric approach to reduce computational cost, and since under the simplifying assumptions described earlier, the structure of all parts of the prior distribution, except the parameters, is known.

If the set of hyperparameters to be estimated, $\{\mu_{\tilde{x}}, Q_{\tilde{x}}, \mu_{\tilde{b}}, Q_{\tilde{b}}\}$, is denoted by ξ , the prior for the empirical Bayesian modeling problem becomes $P(\tilde{Z}, \tilde{\alpha}, \tilde{r}, \tilde{\beta} | \xi)$. Now, the prior is dependent on the set of hyperparameters, ξ . When these hyperparameters are known, ξ drops from the prior as there is no need to express conditioning on a constant. The approach used in

this work estimates the set of the hyperparameters, ξ , from the maximum likelihood solution, and then using the empirically estimated prior to solve for the BLVR model. The maximum likelihood solution may be obtained easily as the BLVR solution with uniform priors for all the parameters. Therefore, solving the empirical BLVR problem involves the following three steps:

- I. Solve the BLVR modeling problem using any of the algorithms described in Section 4.3 with a uniform prior for all parameters.
- II. Estimate the set of hyperparameters, $\hat{\xi}$, as follows:
 1. Estimate $\mu_{\hat{X}}$ as $E[\hat{X}]$,
 2. Estimate $Q_{\hat{X}}$ as $Cov[\hat{X}]$,
 3. Set $\mu_{\tilde{b}} = E[\hat{\alpha}\hat{\beta}]$,
 4. Set $Q_{\tilde{b}} = c(\hat{X}^T \hat{X})^{-1}$,
- III. Solve the BLVR modeling problem using the following prior, $P(\tilde{Z}, \tilde{\alpha}, \tilde{\beta}, \tilde{r} | \hat{\xi})$.

Step II(4) represents the covariance of the posterior of the coefficients (Congdon, 2001, Page 95). The coefficient, c , represents the inverse of the covariance matrix or precision of the prior for b . Thus, c is a small positive number that allows changing the level of confidence about $\mu_{\tilde{b}}$. In this paper, c is selected to minimize the output prediction error.

4.6 Estimating the Number of Retained Principal Components

One of the challenges in applying reduced-rank models is determining the number of principal components. Geladi and Kowalski (1986) described some of the techniques used in rank estimation of PLS models. These approaches include checking the norm of the residual output until it falls below a certain threshold, comparing the norm of the residual output until the difference between two successive model dimensions becomes very small, and using cross validation. Cross validation has been commonly used for many reduced rank models such as, PCR, PLS, and PCA models (Stone and Brooks, 1990; Wold, 1978; Eastment and Krzanowski, 1982). It is a powerful technique that selects the number of principal components that minimizes the output prediction mean squared error for unseen data. It starts by splitting the data in two sets: training and testing, uses the training set to develop the model, and then uses the testing set to test the model and decide its optimum dimension. The number of principal components that

minimizes the testing mean squared error or that at which the testing error stops decreasing is usually selected as the optimum model dimension. The examples in Section 5 use cross validation to estimate the model dimension.

5. Illustrative Examples

This section illustrates the performance of BLVR, and compares it to that of existing methods. The performance of BLVR-I and BLVR-II is evaluated for the following four cases.

- *Case (a)* uses a perfect prior. This case represents the best case scenario for the Bayesian algorithms.
- *Case (b)* estimates the prior from 500 external noisy observations, which are assumed to be available as historic data, using the technique described in Section 4.5.
- *Case (c)* estimates the prior empirically from the data available for modeling.
- *Case (d)* uses a uniform prior, resulting in the maximum likelihood estimate.

Case (a) represents the best case for BLVR since it assumes a perfect prior. This case is not practical, but provides a benchmark for more practical methods. Cases (b), (c) and (d) are practically relevant. Case (d) represents the worst case scenario for BLVR since no prior knowledge is used. Case (c) is usually better than Case (d) since it uses the available measured data for obtaining the prior.

Example 1: Steady State Gaussian Data

In this example, the advantage of Bayesian modeling is illustrated for a simple steady state 3-input, 1-output model. The noise-free input variables are generated as follows:

$$\begin{aligned}\tilde{x}_1 &\sim N(3,2), \quad \tilde{x}_2 \sim N(1,4), \\ \tilde{x}_3 &= \tilde{x}_1 + \tilde{x}_2,\end{aligned}\tag{59}$$

and the noise-free output is assumed to be: $\tilde{y} = 0.4\tilde{x}_1 + 0.4\tilde{x}_2 + 0.4\tilde{x}_3$. The true model rank in this example is 2, which is assumed to be known. This model can also be written in terms of the independent variables, \tilde{x}_1 and \tilde{x}_2 , as:

$$\tilde{y} = a_1\tilde{x}_1 + a_2\tilde{x}_2, \text{ where } a_1 = a_2 = 0.8.\tag{60}$$

The inputs and outputs, which consist of 64 observations, are contaminated with zero-mean Gaussian noise with covariance matrices,

$$\mathbf{Q}_{\varepsilon_x} = \begin{bmatrix} 2/3 & 0 & 0 \\ 0 & 4/3 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \text{and} \quad \mathbf{Q}_{\varepsilon_y} = 1.28, \quad (61)$$

which are assumed to be known. These covariance matrices result in a signal-to-noise ratio of 3 in all variables.

To illustrate the performance of the Bayesian algorithms, a Monte Carlo simulation of 100 realizations is performed assuming a model dimension of 2. The results of this simulation are summarized in Table 2, which lists the output training and testing errors, the input training and testing errors, as well as the model parameters, a_1 and a_2 mean square errors all with respect to their noise-free values.

These results show the benefits of the Bayesian algorithms in prediction as well as model parameter estimation. For example, in Case (a), where perfect knowledge about the data and model is incorporated in the modeling algorithms, very high accuracy models are obtained using both Bayesian algorithms. Even though this perfect knowledge about the true model is usually not available in practice, the results reported in Case (a) indicate the extent of possible improvement by Bayesian algorithms. The results of Case (b) show that BLVR does not need a perfect prior to achieve significant improvement over existing methods. In this case, only historical noisy data from the process are used to obtain an accuracy comparable to Case (a). Such data are readily available for many processes.

The results of Case (c) show that even when no external information about the true model is available, empirically estimated priors from the data being modeled can still provide improved prediction over existing methods. Finally, as Case (d) shows, when a uniform prior is used, the Bayesian algorithms provide the maximum likelihood solution. This solution is closer to existing methods, but with some important differences. For example, even though BLVR-II is similar to PLS, it outperforms PLS since unlike PLS, BLVR-II accounts for the different noise contents in each variable. For BLVR-I (d), on the other hand, the estimated model parameters and output training errors are even better than those obtained by BLVR-II (d), but the testing error is a little worse. In fact BLVR-I (d) performs a little worse than OLS, and only slightly better than PCR and PLS. The reason behind this performance is that BLVR-I (d), as in EIV, does not focus on just good output prediction when estimating the model parameters since it minimizes the combined input and output errors. Thus, in BLVR-I as well as EIV, the output is

considered as important as any other input. Therefore, the estimated parameters by BLVR-I (d) may not necessarily provide good output prediction for unseen data. However, BLVR-I has an advantage over EIV since it can handle collinearity, which is not accounted for by EIV.

Therefore, BLVR-I (d) with a uniform prior can be thought of as a reduced-rank version of the EIV method. This example also shows that the training (data reconciliation) mean squared errors obtained by BLVR-I are the least among all other techniques. This is not surprising since BLVR-I simultaneously minimizes the data reconciliation errors when estimating the model.

Both BLVR algorithms do not exhibit much improvement in the accuracy of the model parameters between Case (c), which is Bayesian, and Case (d), which is maximum likelihood. This lack of improvement may be explained by the observation that for models with ranks *higher than two*, James-Stein (JS) estimators have been shown to provide parameter estimates with lower risk (mean square error) than those obtained by the maximum likelihood solution. For simple problems such as those encountered in linear modeling, JS estimators are shown to be similar to empirical Bayesian estimators (Gruber, 1998). This discussion implies that the Bayesian parameter estimates should be more accurate than maximum likelihood parameter estimates for models of rank greater than two. This is confirmed by Nounou et al. (2001) for Bayesian Principal Component Analysis (BPCA).

The results reported in Table 2 also show that the OLS output testing error is even less than that of PCR and PLS. However, the relative performance of the various techniques is known to be a function of the ratio (n/p) , where n is the number of observations and p is the number of variables (Wold, 1982). Therefore, for a fair look at all techniques, Example 1 is repeated several times using different (n/p) ratios by changing the number of data points. The results, shown in Figure 2, indicate that both Bayesian algorithms are consistent over the entire range of (n/p) values. Figure 2a shows that when PLS and PCR outperform OLS at low (n/p) values, BLVR-I (d) using a uniform prior is better than OLS. When OLS outperforms PCR and PLS at high (n/p) values, however, BLVR-I (d) is comparable to PCR. However, when an empirical prior is used, BLVR-I (c) outperforms all other techniques at all (n/p) values. Figure 2b shows that both, BLVR-II (c) and (d) are consistently better than all other techniques at all (n/p) values.

The results listed in Table 2 were obtained under the assumption that the actual model rank of 2 is known. To estimate the optimum model dimension for both algorithms, cross validation is used, and the results are shown in Figure 3. These results show that the testing error either has a minimum or flattens after retaining two principal components in both algorithms, confirming that the model dimension is 2.

Example 2: Dynamic FIR Model with Non-Gaussian Measurements

This example compares the performances of the proposed methods with existing methods for the common process task of developing a finite impulse response model. To represent a more realistic situation, the underlying variables in this example are nonstationary due to a changing mean. The model used for simulating the data is of the following form,

$$y(t) = b_1 u_1(t-1) + b_2 u_2(t-2) + b_3 u_2(t-1) \quad (62)$$

where, u_1 and u_2 represent inputs, and y represents the output, with,

$$b_1 = 0.5, \quad b_2 = 0.3, \quad \text{and} \quad b_3 = 0.4. \quad (63)$$

The noise-free data are generated as follows,

$$\tilde{u}_1 = \begin{cases} N(0,0.5) & 1 \leq t \leq 20 \\ N(3,0.5) & 21 \leq t \leq 40, \text{ and} \\ N(0,0.5) & 41 \leq t \leq 64 \end{cases}, \quad \tilde{u}_2 = \begin{cases} N(0,0.5) & 1 \leq t \leq 15, 46 \leq t \leq 64 \\ N(5,0.5) & 16 \leq t \leq 30, 31 \leq t \leq 45 \end{cases}. \quad (64)$$

The two inputs and output are contaminated with zero-mean Gaussian noise with variances 1, 2 and 1, respectively. Thus, the input matrix, which is of size (64×3) , is constructed to account for the dynamic relationship as,

$$x(t) = [u_1(t-1) \quad u_1(t-2) \quad u_2(t-1)] \quad (65)$$

In this example, three variations of BLVR-I and II are considered. In Case (b), the prior is estimated from 500 external data points. In Case (c), the prior is estimated empirically from the data itself, and in Case (d), a uniform prior is used. The results of Monte Carlo simulation with 100 realizations is performed, and the estimated model parameters and input and output testing errors are compared for the various techniques, assuming that the model rank is 2. These results, summarized in Table 3, show that incorporating knowledge into the modeling problem by BLVR has a significant advantage. For example, in Cases (b) and (c), both BLVR algorithms provide better output prediction and parameter estimation than conventional methods. However, in Case (d), the output testing error obtained by BLVR-I is smaller than that of EIV, but larger than that

obtained by other predictive models, such as OLS, PLS, and PCR. As discussed in Section 4, this performance should not be surprising since BLVR-I considers the output to be as important as any other input when estimating the model. For maximum likelihood estimation, it is advantageous to only consider the output when estimating the model parameters. Such an approach is adopted in BLVR-II, which does better than BLVR-I and similar to existing methods. Finally, the cross validation errors confirm that the optimum model dimension is indeed 2, as shown Figure 4. These errors are determined according to the approach discussed in Section 4.6.

Example 3: Inferential Modeling of Distillation Column Compositions

This example develops an inferential model to estimate the composition in a distillation column based on temperature measurements. The data are simulated by Kano et al. (2000) using a detailed SPEEDUP[®] model for a 30-tray distillation column. The feed consists of equimolar quantities of methanol, ethanol, 1-propanol, and n-butanol, and is introduced on the 15th tray. The total flow rate is 128 kmol/h. The control system set points are mole fractions of propanol and ethanol at the top and bottom of 0.0010. Data for inferential modeling are collected by varying the component flow rates in the feed stream as PRBS signals, with the total feed flow changing stepwise by $\pm 10\%$ every 2 hours. The entire simulation is run for 20 hours. Additional details and analysis of the process are provided by Kano et. al. (2000).

The objective of this distillation process is to maintain high purity separation of the light and heavy components. Since on-line composition measurement is usually very expensive, it is common to develop inferential models that estimate the product composition in the distillate and bottom streams. In this example, an inferential model is constructed to estimate the composition of ethanol in the distillate stream from temperature measurements at different trays. The product compositions are estimated from temperature measurements at the fourth, ninth, twenty second, and twenty seventh trays. This corresponds to Case B3 of Kano et al. (2000). Thus, the input data matrix has the following structure,

$$T = [T_4 \quad T_9 \quad T_{22} \quad T_{27}]. \quad (66)$$

The simulated data, which consists of 64 observations, are assumed to represent the underlying noise-free behavior of the column. The measured input and output data are contaminated with additive zero-mean Gaussian noise with the following covariance matrix,

$$\mathbf{Q}_{\varepsilon_T} = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}, \text{ and } \mathbf{Q}_{\varepsilon_Y} = 10^{-4}. \quad (67)$$

Since the data used in this example are obtained under temperature control, the distributions of some noise-free input variables are not Gaussian, as shown in Figure 5.

Like the previous examples, Case (b) uses a prior estimated from 100 external or historical data points. In Case (c), the prior is estimated empirically from the data itself, and Case (d) represents the maximum likelihood solution. A Monte Carlo simulation of 100 realizations is performed, assuming that the actual model rank is one, and the results are shown in Table 4. In this example, since the actual model parameters relating the noise-free inputs and outputs are not known, only the output and input testing errors are reported. These results confirm the benefits of BLVR, and show that even when some of the noise-free inputs do not follow a Gaussian distribution, both BLVR algorithms achieve a noticeable improvement over existing methods.

7. Discussion and Conclusions

This paper presents a novel approach for the common but important task of developing a linear model from measured data. The proposed Bayesian Latent Variable Regression (BLVR) method is able to extract a more accurate model than existing methods from the same set of measured data. This advantage of BLVR is due to its use of a Bayesian framework for incorporating prior knowledge about the data and variables to influence the model. Existing methods commonly used for process modeling such as, OLS, PCR, and PLS implicitly assume that such knowledge is not available, while existing Bayesian regression methods are not able to handle collinear variables or errors in all the variables.

For any Bayesian approach, the quality of the prior knowledge is crucial in determining the accuracy of the model. In BLVR, empirical Bayes methods are used to extract prior distributions from historical data or from the data being used for modeling. Assumptions are also made about the prior distributions and noise being Gaussian. These assumptions seem to be reasonable based on the illustrative examples, since BLVR outperforms traditional methods even when the prior distributions are not Gaussian. Furthermore, the proposed approach can easily handle other types of distributions, if necessary. Two variations of BLVR are developed and studied. BLVR-I is a Bayesian and reduced-rank version of the Error-In-Variables (EIV) method, while BLVR-II

is analogous to a Bayesian variation of PLS. Illustrative examples demonstrate the improved accuracy of BLVR modeling versus existing latent variable modeling methods.

A natural question in response to any new method is how it is related to existing methods. The relationship between existing methods and BLVR may be understood by generalizing the BLVR problem formulation presented by Equation Set (47) to,

$$\{\hat{\alpha}\}_{\text{Bayes}} = \underset{\tilde{\alpha}}{\operatorname{argmax}} P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{11}} P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{12}}$$

s.t.

$$\{\hat{\beta}\}_{\text{Bayes}} = \underset{\tilde{\beta}}{\operatorname{argmax}} P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{21}} P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{22}} P(\tilde{\beta} | \tilde{\alpha})^{w_{23}}$$

$$\{\hat{\mathbf{Z}}\}_{\text{Bayes}} = \underset{\tilde{\mathbf{Z}}}{\operatorname{argmax}} P(\mathbf{X} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{31}} P(\mathbf{Y} | \tilde{\mathbf{Z}}, \tilde{\alpha}, \tilde{\beta}, \tilde{\mathbf{r}})^{w_{32}} P(\tilde{\mathbf{Z}} | \tilde{\alpha})^{w_{33}}$$

s.t.

$$\hat{\mathbf{Y}} = \hat{\mathbf{Z}}\hat{\beta}, \quad \hat{\mathbf{X}} = \hat{\mathbf{Z}}\hat{\alpha}^T, \quad \hat{\alpha}^T\hat{\alpha} = \mathbf{I} \quad (68)$$

In Equation Set (68), separate objective functions are written for all the parameters and variables to be estimated. Furthermore, the probabilities are raised to a power, w_{ij} . This “power probability” representation may be specialized to existing methods depending on selected values of w_{ij} . For example, BLVR-I is obtained by setting all weights to unity, whereas BLVR-II is obtained by setting w_{21} and w_{32} to zero, and all other weights equal one. A Bayesian formulation of OLS may be obtained if the weights, w_{12} , w_{22} , and w_{31} , are unity, and the others are zero. If the distributions corresponding to $w_{ij}=1$ are Gaussian, it leads to traditional OLS. Similarly, EIV, TLS, RR, PCR, PLS, and CR may be obtained from Equation Set (68) depending on the values selected for the weights (Nounou, 2000). When $w_{ij} = 1$, the corresponding density remains unchanged, while $w_{ij} = 0$ flattens the density to make it uniformly distributed. The insight provided by such a general Bayesian framework may be useful for improving existing methods or for developing new methods. An intriguing question is whether the weights can be related to each other, and adapted to the nature of the modeling problem. Such a method could specialize to the best existing method, or to new Bayesian methods, just as Continuum Regression can specialize to PCR, PLS, OLS, or methods in between. Further exploration of this general Bayesian framework is in progress.

Many opportunities exist for the development and use of Bayesian methods for solving chemical and process engineering problems. Just as BLVR can improve upon existing linear modeling methods, Bayesian methods for tasks such as, system identification, process monitoring, fault diagnosis, data rectification and nonlinear modeling, are expected to perform better than existing methods. For system identification, subspace identification methods have become quite popular (Verhaegen and Dewilde, 1992). These methods are related to latent variable modeling methods (Shi and MacGregor, 2000), and should be amenable to a Bayesian approach similar to BLVR to obtain more accurate models. Computationally efficient statistical techniques such as, Markov Chain Monte Carlo methods, are also available for relaxing the assumption of Gaussian distributions, and for dealing with nonlinear problems. These methods are being used for recursive and Bayesian methods state and parameter estimation in nonlinear dynamic systems (Chen et al., 2001). The combination of multiscale and Bayesian methods is also attractive for handling non-Gaussian and autocorrelated systems (Bakshi et al., 2001). Many of these topics are the focus of on-going research. It is hoped that this paper will trigger additional research on tapping the benefits of the Bayesian approach for more efficient chemical processing.

Acknowledgements

Financial support from the National Science Foundation (CTS-9733627), and data for the distillation example from Dr. Manabu Kano are gratefully acknowledged.

References

- Albuquerque, J. S., and L. T. Biegler "Data Reconciliation and Gross Error Detection in Dynamic Systems", *AICHE J.*, 42, 2841-2856, (1996).
- Bakshi, B. R., and G. Stephanopoulos "Representation of Process Trends Part III. Multi-Scale Extraction of Trends from Process Data", *Comp. Chem. Eng.*, 18, 4, 267-302, (1994).
- Bakshi, B. R., M. N. Nounou, P. K. Goel, P. K. and X. Shen "Multiscale Bayesian Data Rectification of Linear Steady-State and Dynamic Systems without Accurate Models", *Ind. Eng. Chem. Res.*, 40, 1, 261 -274, (2001).
- Box, G. E. P., G. M. Jenkins and G. C. Reinsel *Time Series Analysis*, Prentice-Hall, Englewood Cliffs, NJ, (1994)
- Chen, W-S., S. Ungarala, B. R. Bakshi and P. K. Goel "Bayesian Rectification of Nonlinear Dynamic Processes by the Weighted Bootstrap", *AICHE Annual Meeting*, Paper 275f, Reno, NV (2001).
- Congdon, P., *Bayesian Statistical Modeling*, John Wiley and Sons, West Sussex, England, (2001).
- De Alba, E., and O. Anguilar "Constrained Forecasts in ARMA Models: A Bayesian Approach", Technical Report, Department of Statistics, Duke University, (1995).

- de Freitas, J. F. G., M. Niranjan, A. H. Gee and A. Doucet “Sequential Monte Carlo Methods to Train Neural Network Models”, *Neural Computation*, 12, 995, (2000).
- Eastment, H. T. and Krazanowski, W. J., “Cross-Validatory Choice of the Number of Components from a Principal Component Analysis”, *Technometrics*, 24, 1, 73-77, (1982).
- Frank, I.E., and J.H. Friedman, “A Statistical View of Some Chemometrics Regression Tools”, *Technometrics*, 35, 2, 109-148, (1993).
- Geladi, P. and B. Kowalski “Partial Least Squares Regression: A Tutorial”, *Analytica Chimica Acta*, 185, 1-17, (1986).
- Gelman, A., Carlin, J. B., Stern, H. S. and D. Rubin, *Bayesian Data Analysis*, Chapman and Hall, London, (1995).
- Gilks, W.R., S. Richardson, and D. Spiegelhalter eds., “*Practical Markov Chain Monte Carlo*”, Chapman And Hall, New York, (1996).
- Golub, G.H., M. Heath, and G. Wahba “Generalized Corss-Validation as a Method for Choosing a Good Ridge Parameter”, *Technometrics*, 21, 215-224, (1979).
- Gruber, M. H., *Improving efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*, Marcel Dekker, New York, (1998).
- Hoerl, A.E. and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, *Technometrics*, 8, 27-51, (1970).
- Johnston, L. P. M., and M. A. Kramer “Estimating state probability distributions from noisy and corrupted data” *AIChE J.*, **44**, 591, (1998).
- Kadane, Joseph B., “Prime Time for Bayes”, *Controlled Clinical Trials*, 16, 313-318, (1995).
- Kano, M., K. Miyazaki, S. Hasebe, and I. Hashimoto, “Inferential Control System of Distillation Compositions using Dynamic Partial Least Squares Regression”, *J. Process Control*, 10, 157-166, (2000).
- Kaspar, M. H. and W. H. Ray “Dynamic PLS Modeling for Process Control”, *Chem. Engg. Sci.*, 48, 3447, (1993).
- Kim, I.W., S. Kang, S. Park and T. Edgar “Robust Data Reconciliation and Gross Error Detection: The Modified MIMT using NLP”, *Computers and Chemical Engineering*, 21, 7, 775-782, (1997).
- Kim, I.-W., M. Liebman, and T. Edgar “Robust Error-in-Variables Estimation Using Nonlinear Programming Techniques.”, *AIChE J.*, 36, 7, (1990).
- Kresta, J. V., J. F. MacGregor and T. E. Marlin “Multivariate Statistical Monitoring of Process Operating Performance”, *Can. J. Chem. Eng.*, 69, 35, (1991).
- Kresta, J. V., T. E. Marlin, and J. F. MacGregor “Development of Inferential Process Models Using PLS”, *Comp. Chem. Eng.*, 18, 7, 597-611, (1994).
- Ku, W. F., R. Storer, R., and C. Georgakis “Disturbance Detection and Isolation by Dynamic Principal Component Analysis”, *Chemometrics Intell. Lab. Syst.*, 30, 179, (1995).
- Lakshminarayanan, S., S. Shah and K. Nandakumar “Modeling and Control of Multivariate Processes: Dynamic PLS Approach”, *AIChE J.*, 43, 2307, (1997).
- Leamer, E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York, (1978).
- Ljung, L., *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, (1999).

- Lorber, A., L.E. Wangen, and B. R. Kowalski, "A Theoretical Foundation for PLS Algorithm", *Journal of Chemom.*, 1, 19-31, (1987).
- MacGregor, J. F., T. Kourti, and J. V. Kresta "Multivariate Identification: A Study of Several Methods", *IFAC Symp. ADCHEM-91*, Toulouse, (1991).
- Malakoff, D., "Bayes Offers a 'New' Way to Make Sense of Numbers", *Science*, 286, 1460-1464, (1999).
- Malthouse, E. C. , A. C. Tamhane, and R. S. H. Mah "Nonlinear Partial Least Squares", *Computers and Chemical Engineering*, 21, 8, 875-890, (1997).
- Maritz, J.S., *Empirical Bayes Methods*, Methuen & CO., London, (1970).
- Massy, W.F., "Principal components Regression in Exploratory Statistical Research", *Journal of the American Statistical Association*, 60, 234-246, (1965).
- McCulloch, R. E. and R. S. Tsay "Bayesian Analysis of Autoregressive Time Series Via the Gibbs Sampler", *Journal of Time Series Analysis*, 15, 235-250, (1994).
- Mejdell, T., and S. Skogestad "Estimation of Distillation Compositions from Multiple Temperature Measurements Using Partial-Least-Squares Regression", *Ind. Eng. Chem. Res.*, 30, 2543-2555, (1991).
- Neal, R. M., *Bayesian Learning for Neural Networks*, Springer-Verlag, New York (1996).
- Negiz, A. and A. Cinar, "Statistical Monitoring of Multivariable Dynamic Process with State-Space Models", *AIChE J.*, 43, 2002, (1997).
- Nomikos, P. and J. F. MacGregor "Monitoring Batch Process Using Multiway Principal Component Analysis", *AIChE J.*, 40, 1361, (1994).
- Nounou, M. N., "Multiscale Bayesian Linear Modeling and Applications", *Ph.D. Dissertation*, The Ohio State University, (2000).
- Nounou, M. N., B. R. Bakshi, P. K. Goel and X. Shen "Bayesian Principal Component Analysis", submitted to *J. Chemometrics*, (2001).
- Pliz, J., "Bayesian Estimation and Experimental Design in Linear Regression Models", *Teubner-Texte zur Mathematik*, (1983).
- Press, S. J., *Bayesian Statistics: Principles, Models, and Applications*, Wiley, New York, (1989).
- Rengaswamy, R., and V. Venkatasubramanian "A Syntactic Pattern-Recognition Approach for Process Monitoring and Fault-Diagnosis", *Eng. Appl. Artif. Intel.*, 8, 1, 35-51, (1995).
- Robert, C.P., *The Bayesian Choice: A Decision Theoretic Motivation*, Springer-Verlag, New York, (1994).
- Shi, R. and J. F. MacGregor, "Modeling of Dynamic Systems Using Latent Variable and Subspace Methods", *J. Chemometrics*, 14, 423-439, (2000).
- Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, (1986).
- Stone, M. and R. J. Brooks, "Continuum Regression: Cross-Validated Sequentially Constructed Prediction embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression", *J. R. Statist. Soc. B*, 52, 2, 237-269, (1990).
- Tamhane, A. C., C. Iordache, and R. S. H. Mah, "A Bayesian Approach to Gross Error Detection in Chemical Process Data. Part I: Model Development", *Chemometrics and Intell. Lab. Sys.*, 4, 131-146, (1988).

- Valko, P. and S. Vajda “An Extended Marquardt-Type Procedure for Fitting Error-in-Variables Models.”, *Computers and Chemical Engineers*, 11, 1, 37-43, (1987).
- Van Huffel, S. and J. Vandewalle “The Total Least Squares Problem: Computational Aspects and Analysis”, *Frontiers in Applied Mathematics*, Philadelphia, PA., (1991).
- Van Huffel, S., *Recent Advances in Total Least Squares Techniques and Error-In-Variables Modeling*, Proceedings of the second Int. Workshop on Total Least Squares and Error-In-Variables Modeling, SIAM, Leuven, Belgium, (1997).
- Verhaegen, M. and P. DeWilde “Subspace Model Identification .1. The Output-Error State-Space Model Identification Class Of Algorithms”, *Int. J. Control*, 56, 1187, (1992).
- West, M. and J. Harrison, *Bayesian and Dynamic Models*, Springer-Verlag, New York, (1989).
- Whiteley, J. R. and J. F. Davis “Knowledge-Based Interpretation of Sensor Patterns”, *Comp. Chem. Eng.*, 16, 4, 329-346, (1992).
- Wise, B. M., and N. L. Ricker “Identification of Finite Impulse Response Models by Principal Components Regression: Frequency Response Properties”, *Proc. Cont. Qual.*, 4, 77-86, (1992).
- Wold, S., “Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models”, *Technometrics*, 20, 4, 397-405, (1978).
- Wold, S., “Nonlinear PLS Modeling II: Spline inner relations. *Chemom. Intell. Lab. Sys.*, 14, 71-84, (1992).
- Wold, S., “Soft Modeling. The basic design and some extensions”, in *Systems under Indirect Observations*, editors, K. Joreskog and H. Wold, Eslevier, Amsterdam (1982).
- Wong, J. C., K. A. McDonald and A. Palazoglu “Classification of Process Trends based on Fuzzified Symbolic Representation and Hidden Markov Models”, *J. Proc. Cont.*, 8, 5-6, 395-408, (1998).
- Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York, (1971).

Appendices

Appendix I: Derivation of the EIV data reconciliation solution

The EIV data reconciliation problem can be formulated as follows,

$$\begin{aligned} \{\hat{x}_i, \hat{y}_i\}_{\text{EIV}} &= \arg \min_{\hat{y}_i, \hat{x}_i} \left\{ (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) \right\} \\ \text{s.t.} \quad \hat{y}_i &= \hat{x}_i^T \hat{\mathbf{b}}. \end{aligned} \quad (\text{A1.1})$$

Solution:

Define the Lagrange function as,

$$L = (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + \lambda (\hat{y}_i - \hat{x}_i^T \hat{\mathbf{b}}). \quad (\text{A1.2})$$

Taking the partial derivatives of L with respect to \hat{x}_i , \hat{y}_i , and λ , and setting them to zeros,

$$\frac{\partial L}{\partial \hat{x}_i} = -2Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) - \hat{\mathbf{b}} \lambda = 0. \quad (\text{A1.3})$$

$$\frac{\partial L}{\partial \hat{y}_i} = -2Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + \lambda = 0. \quad (\text{A1.4})$$

$$\frac{\partial L}{\partial \lambda} = \hat{y}_i - \hat{x}_i^T \hat{\mathbf{b}} = 0. \quad (\text{A1.5})$$

Substituting Equation A1.5 in Equation A1.4, get

$$\lambda = 2Q_{\varepsilon_y}^{-1} (y_i - \hat{x}_i^T \hat{\mathbf{b}}). \quad (\text{A1.6})$$

Substituting Equation A1.6 in Equation A1.3, get

$$-2Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) - 2\hat{\mathbf{b}} Q_{\varepsilon_y}^{-1} (y_i - \hat{x}_i^T \hat{\mathbf{b}}) = 0. \quad (\text{A1.7})$$

Rearranging Equation A1.7, get the EIV data reconciliation solution

$$\{\hat{x}_i\}_{\text{EIV}} = \left[Q_{\varepsilon_x}^{-1} + \hat{\mathbf{b}} Q_{\varepsilon_y}^{-1} \hat{\mathbf{b}}^T \right]^{-1} \left(Q_{\varepsilon_x}^{-1} x_i + \hat{\mathbf{b}} Q_{\varepsilon_y}^{-1} y_i \right) \quad (\text{A1.8})$$

and

$$\{\hat{y}_i\}_{\text{EIV}} = \hat{\mathbf{b}}^T \hat{x}_i. \quad (\text{A1.9})$$

Appendix II: Derivation of the data reconciliation solution for BLVR-I

The data reconciliation problem for BLVR-I can be formulated as follows,

$$\begin{aligned} \{\hat{z}_i, \hat{y}_i\}_{\text{MAP}} &= \arg \min_{\hat{z}_i, \hat{y}_i} \left\{ (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + (\hat{z}_i - \mu_{z_i \tilde{\alpha}})^T Q_{z_i \tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{z_i \tilde{\alpha}}) \right\} \\ \text{s.t.} \quad \hat{y}_i &= \hat{z}_i^T \hat{\beta}, \quad \hat{x}_i = \hat{\alpha} \hat{z}_i \end{aligned} \quad (\text{A2.1})$$

Solution:

Define the Lagrange function as,

$$L = (z_i - \hat{z}_i)^T Q_{\varepsilon_z}^{-1} (z_i - \hat{z}_i) + (y_i - \hat{y}_i)^T Q_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + (\hat{z}_i - \mu_{z_i \tilde{\alpha}})^T Q_{z_i \tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{z_i \tilde{\alpha}}) + \lambda (\hat{y}_i - \hat{z}_i^T \hat{\beta}). \quad (\text{A2.2})$$

Taking the partial derivatives of L with respect to \hat{z}_i , \hat{y}_i , and λ , and setting them to zeros, get

$$\frac{\partial L}{\partial \hat{z}_i} = -2\mathbf{Q}_{\varepsilon_z}^{-1}(z_i - \hat{z}_i) + 2\mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1}(z_i - \hat{z}_i) - \hat{\beta}\lambda = 0. \quad (\text{A2.3})$$

$$\frac{\partial L}{\partial \hat{y}_i} = -2\mathbf{Q}_{\varepsilon_y}^{-1}(y_i - \hat{y}_i) + \lambda = 0. \quad (\text{A2.4})$$

$$\frac{\partial L}{\partial \lambda} = \hat{y}_i - \hat{z}_i^T \hat{\beta} = 0. \quad (\text{A2.5})$$

Substituting Equation A2.5 in A2.4, get

$$\lambda = 2\mathbf{Q}_{\varepsilon_y}^{-1}(y_i - \hat{\beta}^T \hat{z}_i). \quad (\text{A2.6})$$

Substituting Equation A2.6 in A2.3, get

$$-2\mathbf{Q}_{\varepsilon_z}^{-1}(z_i - \hat{z}_i) + 2\mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1}(z_i - \hat{z}_i) - 2\hat{\beta}\mathbf{Q}_{\varepsilon_y}^{-1}(y_i - \hat{\beta}^T \hat{z}_i) = 0. \quad (\text{A2.7})$$

Rearranging Equation A2.7, get the data reconciliation solution

$$\{\hat{z}_i\}_{\text{MAP}} = [\mathbf{Q}_{\varepsilon_z}^{-1} + \mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1} + \hat{\beta}\mathbf{Q}_{\varepsilon_y}^{-1}\hat{\beta}^T]^{-1}(\mathbf{Q}_{\varepsilon_z}^{-1}z_i + \mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1}\mu_{\tilde{z}_i\tilde{\alpha}} + \hat{\beta}\mathbf{Q}_{\varepsilon_y}^{-1}y_i) \quad (\text{A2.8})$$

and

$$\{\hat{y}_i\}_{\text{MAP}} = \hat{\beta}^T \{\hat{z}_i\}_{\text{MAP}}. \quad (\text{A2.9})$$

Appendix III: Derivation of the data reconciliation solution for BLVR-II

The data reconciliation can be formulated as follows,

$$\begin{aligned} \{\hat{z}_i\} = \underset{\hat{z}_i}{\text{argmin}} \left\{ (x_i - \hat{x}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (\hat{z}_i - \mu_{\tilde{z}_i\tilde{\alpha}})^T \mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{\tilde{z}_i\tilde{\alpha}}) \right\} \\ \text{s.t.} \quad \hat{x}_i = \hat{\alpha}\hat{z}_i. \end{aligned} \quad (\text{A3.1})$$

Solution:

Define the Lagrange function as,

$$L = (x_i - \hat{x}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) + (\hat{z}_i - \mu_{\tilde{z}_i\tilde{\alpha}})^T \mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1} (\hat{z}_i - \mu_{\tilde{z}_i\tilde{\alpha}}) + \lambda(\hat{x}_i - \hat{\alpha}\hat{z}_i). \quad (\text{A3.2})$$

Taking the partial derivatives of L with respect to \hat{x}_i , \hat{z}_i , and λ , and setting them to zeros, get

$$\frac{\partial L}{\partial \hat{x}_i} = -2\mathbf{Q}_{\varepsilon_x}^{-1}(x_i - \hat{x}_i) + \lambda^T = 0. \quad (\text{A3.3})$$

$$\frac{\partial L}{\partial \hat{z}_i} = 2\mathbf{Q}_{\tilde{z}_i\tilde{\alpha}}^{-1}(\hat{z}_i - \mu_{\tilde{z}_i\tilde{\alpha}}) - \hat{\alpha}^T \lambda^T = 0. \quad (\text{A3.4})$$

$$\frac{\partial L}{\partial \lambda} = \hat{x}_i - \hat{\alpha}\hat{z}_i = 0. \quad (\text{A3.5})$$

Substituting Equation A3.3 in A3.4, get

$$2\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}(\hat{z}_i - \mu_{\tilde{z}|\tilde{\alpha}}) - 2\hat{\alpha}^T \mathbf{Q}_{\varepsilon_x}^{-1}(x_i - \hat{x}_i) = 0. \quad (\text{A3.6})$$

Substituting Equation A3.5 in A3.6, get

$$2\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}(\hat{z}_i - \mu_{\tilde{z}|\tilde{\alpha}}) - 2\hat{\alpha}^T \mathbf{Q}_{\varepsilon_x}^{-1}(x_i - \hat{\alpha}\hat{z}_i) = 0. \quad (\text{A3.7})$$

Rearranging Equation A3.7, get the data reconciliation solution

$$\{\hat{z}_i\} = (\hat{\alpha}^T \mathbf{Q}_{\varepsilon_x}^{-1} \hat{\alpha} + \mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1})^{-1} (\hat{\alpha}^T \mathbf{Q}_{\varepsilon_x}^{-1} x_i + \mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1} \mu_{\tilde{z}|\tilde{\alpha}}). \quad (\text{A3.8})$$

Appendix IV: Model parameter solution for BLVR-II

The parameter estimation problem for the second Bayesian algorithm can be formulated as follows,

$$\begin{aligned} \{\hat{\beta}\} = \underset{\hat{\beta}}{\operatorname{argmin}} & \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^T \mathbf{Q}_{\varepsilon_y}^{-1} (y_i - \hat{y}_i) + (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}})^T \mathbf{Q}_{\tilde{\beta}|\tilde{\alpha}}^{-1} (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}}) \right\} \\ \text{s.t.} \quad \hat{y}_i &= \hat{z}_i^T \hat{\beta} \end{aligned} \quad (\text{A4.1})$$

Solution:

For single output models, the covariance matrix $\mathbf{Q}_{\varepsilon_y}$ is a scalar, and thus the optimization problem shown in Equation A4.1 can be rewritten as follows,

$$\{\hat{\beta}\} = \underset{\hat{\beta}}{\operatorname{argmin}} \left\{ (\mathbf{Y} - \hat{\mathbf{Z}}\hat{\beta})^T (\mathbf{Y} - \hat{\mathbf{Z}}\hat{\beta}) + (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}})^T \mathbf{Q}_h^{-1} (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}}) \right\} \quad (\text{A4.2})$$

where, $\mathbf{Q}_h = \mathbf{Q}_{\tilde{\beta}|\tilde{\alpha}}^{-1} \mathbf{Q}_{\varepsilon_y}$.

Define the Lagrange function as,

$$L = (\mathbf{Y} - \hat{\mathbf{Z}}\hat{\beta})^T (\mathbf{Y} - \hat{\mathbf{Z}}\hat{\beta}) + (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}})^T \mathbf{Q}_h^{-1} (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}}). \quad (\text{A4.3})$$

Taking the partial derivative of L with respect to $\hat{\beta}$ and setting it to zero, get

$$\frac{\partial L}{\partial \hat{\beta}} = -2\hat{\mathbf{Z}}^T (\mathbf{Y} - \hat{\mathbf{Z}}\hat{\beta}) + 2\mathbf{Q}_h^{-1} (\hat{\beta} - \mu_{\tilde{\beta}|\tilde{\alpha}}) = 0. \quad (\text{A4.4})$$

Rearranging the terms, get

$$(\hat{\mathbf{Z}}^T \mathbf{Y} + \mathbf{Q}_h^{-1} \mu_{\tilde{\beta}|\tilde{\alpha}}) = (\hat{\mathbf{Z}}^T \hat{\mathbf{Z}} + \mathbf{Q}_h^{-1}) \hat{\beta}. \quad (\text{A4.5})$$

Solving for the model parameters, get

$$\hat{\beta} = (\hat{\mathbf{Z}}^T \hat{\mathbf{Z}} + \mathbf{Q}_h^{-1})^{-1} (\hat{\mathbf{Z}}^T \mathbf{Y} + \mathbf{Q}_h^{-1} \mu_{\tilde{\beta}|\tilde{\alpha}}). \quad (\text{A4.6})$$

Table 1. Properties of various modeling methods

	Handle Collinearity	Account for input and output noise	Incorporate knowledge about measured variables	Incorporate knowledge about model parameters
OLS	No	No	No	No
TLS	No	Yes	No	No
EIV	No	Yes	No	No
RR	Yes	No	No	Yes
PCR	Yes	Yes	No	No
PLS	Yes	Yes	No	No
BLVR	Yes	Yes	Yes	Yes

Table 2. Monte Carlo simulation results for steady state example. Case (a): perfect prior; Case (b): prior estimated from 500 external data points; Case (c): empirical prior; Case (d): uniform prior

MSE	Prior	Y (training)	Y (testing)	X (training)	X (testing)	a_1 $\times 10^3$	a_2 $\times 10^3$
OLS	uniform	0.62	0.66	1.32	1.32	3.0	17.2
EIV	uniform	0.90	2.52	0.99	1.32	29	211
PCR	uniform	0.67	0.70	0.92	0.92	3.1	16.1
PLS	uniform	0.64	0.71	1.34	1.18	4.8	19.8
BLVR-I (a)	perfect	0.39	0.55	0.55	0.66	0.3	0.3
BLVR-I (b)	external data	0.41	0.57	0.59	0.68	0.3	1.2
BLVR-I (c)	empirical	0.43	0.60	0.60	0.69	4.4	8.9
BLVR-I (d)	uniform	0.47	0.67	0.71	0.82	4.4	9.0
BLVR-II (a)	perfect	0.52	0.55	0.64	0.66	0.3	0.3
BLVR-II (b)	external data	0.58	0.58	0.69	0.69	1.3	8.9
BLVR-II (c)	empirical	0.59	0.62	0.69	0.70	3.1	16.6
BLVR-II (d)	uniform	0.61	0.63	0.83	0.82	3.1	17.2

Table 3. Monte Carlo simulation results for the dynamic example. Case (b): prior estimated from 500 external data points; Case (c): empirical prior; Case (d): uniform prior.

MSE	Prior	y (testing)	X (testing)	b_1 $\times 10^2$	b_2 $\times 10^3$	b_3 $\times 10^3$
OLS	uniform	0.68	1.34	1.51	9.3	4.9
EIV	uniform	1.20	1.34	2.07	24.1	2.3
PCR	uniform	0.65	1.24	1.83	5.7	4.9
PLS	uniform	0.65	1.24	3.35	1.0	3.2
BLVR-I (b)	external data	0.51	1.01	0.94	8.1	0.3
B LVR I (c)	empirical	0.58	1.04	1.11	9.0	4.0
B LVR I (d)	uniform	0.70	1.23	1.12	9.0	4.0
BLVR-II (b)	external data	0.54	1.01	1.47	4.4	3.4
B LVR II (c)	empirical	0.58	1.04	1.70	5.4	4.8
B LVR II (d)	uniform	0.65	1.23	1.75	5.4	5.0

Table 4. Monte Carlo simulation results of MSE for the distillation column inferential modeling example. Case (b): prior estimated from 100 external data points; Case (c): empirical prior; Case (d): uniform prior.

Method	y (testing) $\times 10^5$	X (testing)
OLS	5.66	0.35
EIV	5.66	0.35
PCR	5.36	0.16
PLS	5.36	0.16
BLVR-I (b)	4.89	0.09
BLVR-I (c)	4.93	0.09
BLVR-I (d)	5.18	0.15
BLVR-II (b)	4.88	0.09
BLVR-II (c)	4.93	0.09
BLVR-II (d)	5.18	0.15

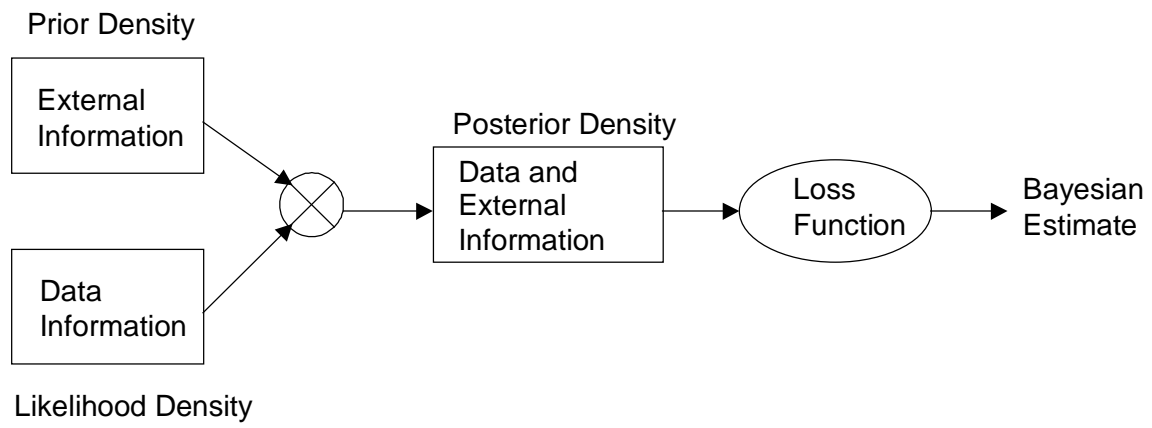


Figure 1. A schematic diagram of the main steps in Bayesian estimation.

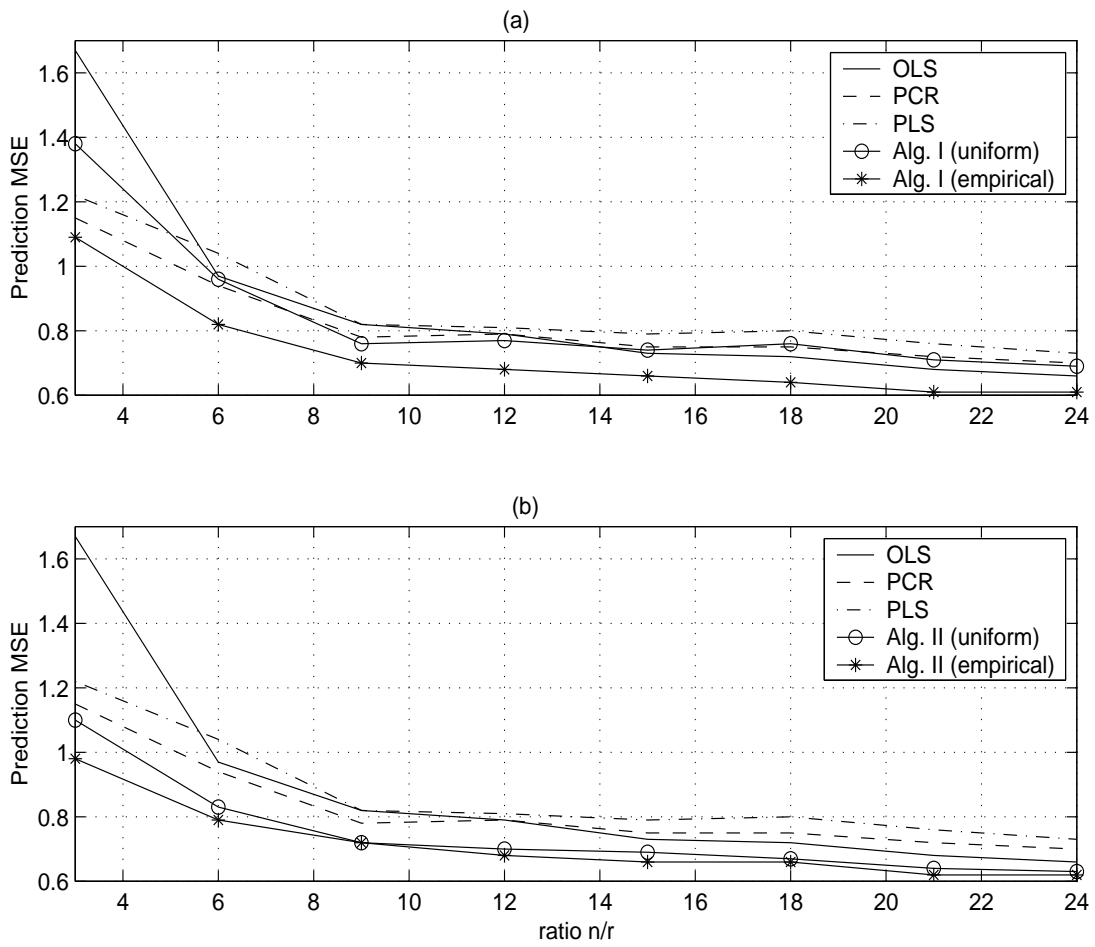


Figure 2. Comparisons of the prediction mean squared error for the various techniques at different number of samples to number of variables (n/p) ratios.

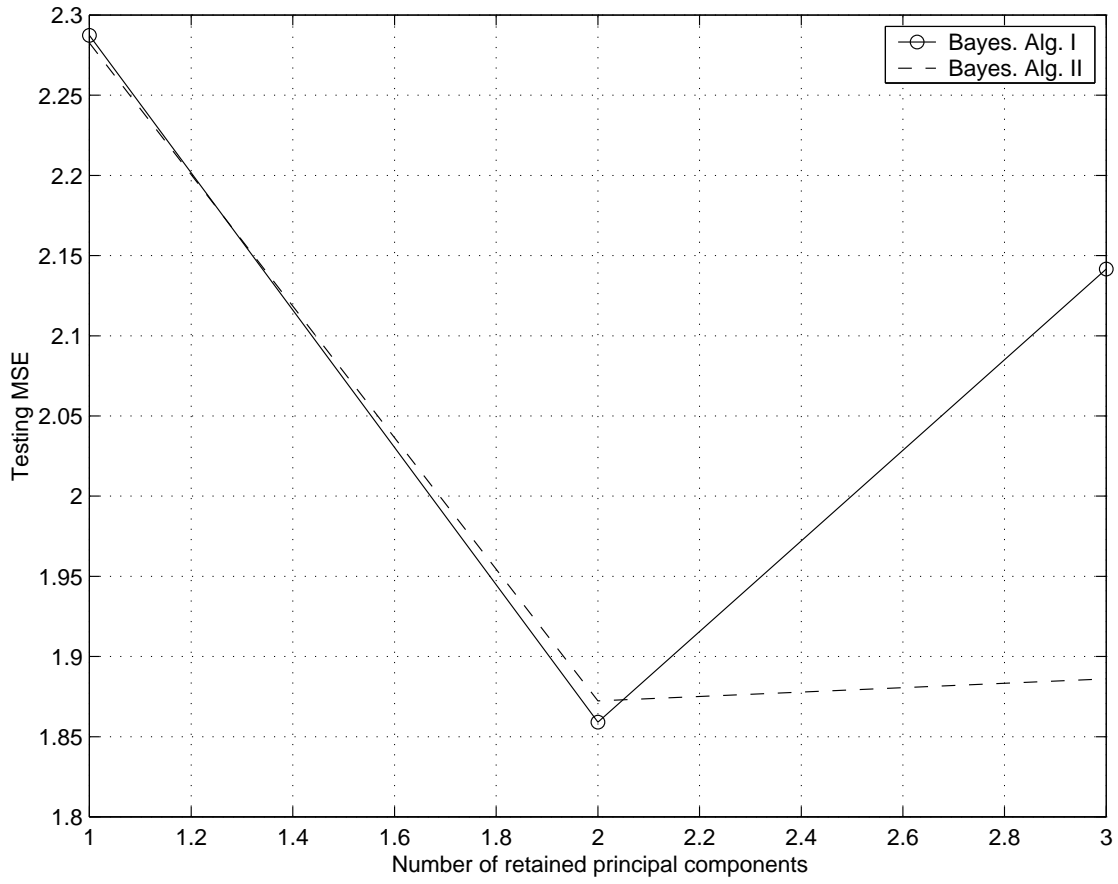


Figure 3. Estimating the model rank using cross validation for the steady state example.

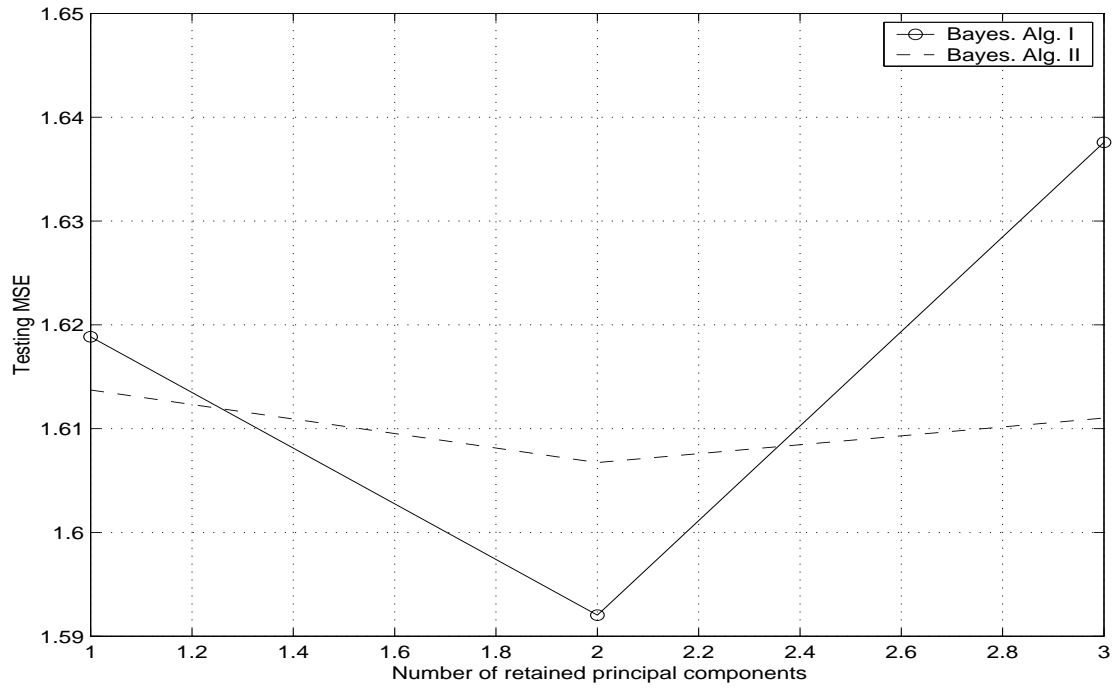


Figure 4. Estimating the model rank using cross validation for the dynamic example.

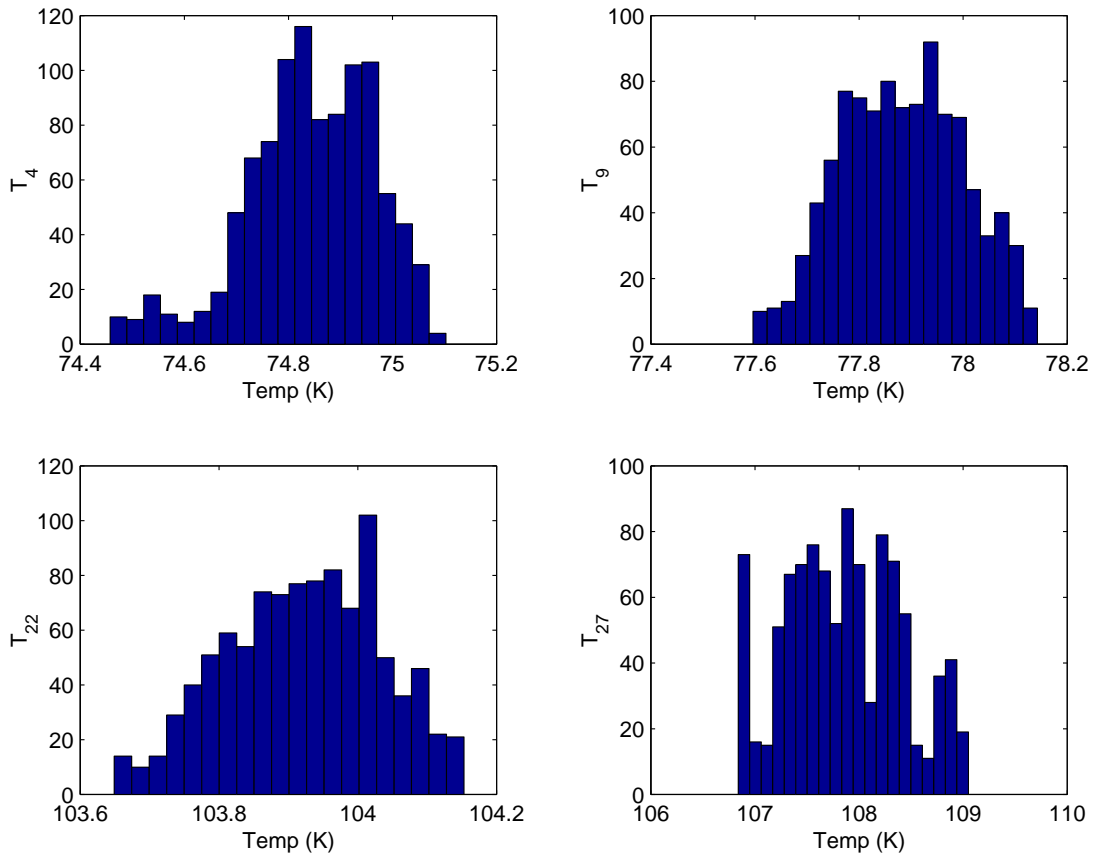


Figure 5. The distributions of the noise-free inputs (temperature data) for the distillation column example.