AIChE

Integrated Estimation of Measurement Error with Empirical Process Modeling—A Hierarchical Bayes Approach

Hongshu Chen and Bhavik R. Bakshi
William G. Lowrie Dept. of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, OH 43210

Prem K. Goel
Dept. of Statistics, The Ohio State University, Columbus, OH 43210

DOI 10.1002/aic.11918
Published online August 24, 2009 in Wiley InterScience (www.interscience.wiley.com).

Advanced empirical process modeling methods such as those used for process monitoring and data reconciliation rely on information about the nature of noise in the measured variables. Because this likelihood information is often unavailable for many practical problems, approaches based on repeated measurements or process constraints have been developed for their estimation. Such approaches are limited by data availability and often lack theoretical rigor. In this article, a novel Bayesian approach is proposed to tackle this problem. Uncertainty about the error variances is incorporated in the Bayesian framework by setting noninformative priors for the noise variances. This general strategy is used to modify the Sampling-based Bayesian Latent Variable Regression (Chen et al., J Chemom.*, 2007) approach, to make it more robust to inaccurate information about the likelihood functions. Different noninformative priors for the noise variables are discussed and unified in this work. The benefits of this new approach are illustrated via several case studies.* © 2009 American Institute of Chemical Engineers *AIChE J,* 55: 2883–2895, 2009

Keywords: Bayesian statistics, latent variable, regression, partial least squares, principal components analysis, Gibbs sampling, noninformative prior

Introduction

Empirical models are essential for several process engineering tasks including process control, abnormal situation management, and data rectification. Such models are also popular in many other disciplines including chemometrics, image processing, bioinformatics, and financial engineering. Linear modeling methods such as ordinary least squares regression (OLS), principal component analysis/regression (PCA/PCR),[1] and partial least squares regression (PLS)[2,3] continue to remain popular for these tasks. These methods usually find the least-squares solution by minimizing the mean squared error between the measured data and estimated values. This approach works best when the measurement noise is independent and identically distributed Gaussian with equal variance. If this assumption is not satisfied, as is often the case, methods such as weighted regression and maximum likelihood principal component analysis/regression (MLPCA/MLPCR),[4] Bayesian principal component analysis (BPCA),[5] and Bayesian latent variable regression (BLVR)[6] may be used. These methods require information about the covariance matrix of the measurement noise. When such information is available, these methods can provide more accurate models. Bayesian approaches can also make use of prior knowledge about the unknown variables. However, information about the measurements errors (likelihood) or

Correspondence concerning this article should be addressed to B. R. Bakshi at bakshi.2@osu.edu

prior is not easily available, which can pose a significant challenge for the application of more advanced modeling methods. Inaccurate information can worsen the model results when compared with the results of simpler methods.

In response to this challenge, many efforts have devised ways of obtaining likelihood information from measured data and other sources. For example, Leger et al.[7] have proposed to estimate the error covariance matrices from repeated measurements and use the estimated covariance matrices in MLPCA. This method is applicable only when enough replicated measurements are available, which may not happen very often due to higher experimentation costs. Research on data reconciliation has also focused on methods for estimating the error covariance for weighting the objective function.[8] The error covariance matrices are usually directly estimated based on the measurements similar to the method used by Leger et al.[7] or indirectly estimated by incorporating additional process information. Almasy et al. and Mah[9] estimated the error covariance matrices from the constraint residuals of process data. Based on their work, Darouach et al.[10] used a maximum likelihood estimator to estimate the error variances by solving a nonlinear optimization problem that couples the estimation problem with the data reconciliation problem. Keller et al.[11] further extended this approach to estimate both the variances and covariances of measurement error. Because these methods are sensitive to outliers, Chen et al.[12] developed an M-estimator to estimate the error covariance matrices. In this approach, the observations are weighted based on their Mahalanobis distances, hence the M-estimator is more robust. Morad et al.[13] developed another robust M-estimator which directly estimates the covariance matrices from the measurements. Maquin et al.[14] applied a direct method to simultaneously estimate the variances of measurement errors and reconcile the data with respect to balance equations. Mirabedini and Hodouin[15] used a state-space model to estimate the variance and covariance of sampling errors in complex dynamic mineral systems. These methods often rely on known process constraints, such as mass balance; however, constraints may not be available or may not even exist for a given process data set. Furthermore, methods like those discussed in this paragraph can be computationally expensive due to the need for mathematical programming algorithms. These hassles limit the applications of most existing error estimation methods.

Most of the previous work focuses on the estimation of the parameters in the likelihood functions. However, getting the estimates of those parameters is not the ultimate goal. Eventually, the estimates are used in the modeling methods and the model quality is what matters. Relying on a single estimate of the error covariance matrix in modeling is inherently vulnerable to the quality of estimation and is not robust. Hence, this two-step strategy is not ideal for solving this problem. From a Bayesian perspective, this challenge can be tackled in a rigorous way without requiring extra measurements, process constraints, or a heavy computational load. Bayesian statistics provides a statistically sound way to combine information from different sources and with uncertainties. When information about the measurement error or likelihood is not available, the noise covariance matrix may be treated as a set of unknown model parameters that are inherently stochastic. The uncertainties of these parameters before modeling can be captured by their prior distributions. When little is known about the prior distribution of a variable, a noninformative prior[16,17] can be assumed, and is a popular approach in Bayesian estimation. For the problem considered in this article, the prior for the noise covariance matrix may be considered to be noninformative. This layer of prior for the noise variance avoids the problem of having to assign a fixed number to the parameter. The prior distributions of model parameters are combined with data measurements by Bayes rule, resulting in a posterior distribution. This posterior distribution contains all the available information about the model parameters. The noise variances and other model parameters can be estimated altogether from the posterior distribution. The uncertainty information of the estimates can also be obtained.

In this article, this strategy is adopted to modify the BLVR approach, such that it can be used for process modeling with inaccurate likelihood information. The original optimization-based BLVR algorithm[6] relies on nonlinear programming for obtaining the maximum a posterior (MAP) estimate, making it computationally expensive and impractical for modeling high dimensional data sets. In addition, optimization-based methods usually only provide point estimates for parameters and require additional effort to obtain uncertainty information about the estimates. In contrast, most of the recent work on the use of Bayesian methods in process engineering rely on sampling-based algorithms.[18–23] The recent surge of interest in Bayesian modeling is in large part due to the adoption of Monte Carlo sampling techniques. To overcome the problems of the optimization-based BLVR, a sampling-based BLVR (BLVR-S)[24] was developed. Instead of solving an optimization problem, it draws samples from the posterior distribution and use samples to estimate the model parameters. The confidence intervals of the estimates can also be easily constructed. BLVR-S implements a Gibbs sampling algorithm to draw samples. The sequential sampling strategy in Gibbs sampling makes it easy to draw samples from the high dimensional posterior distribution. In this article, the BLVR-S is modified to deal with the common situation when likelihood information is not available. An extra step is added in the Gibbs sampler of BLVR-S to draw samples of the noise variances of input and output variables. The introduction of the noninformative priors for the noise variances makes BLVR-S more immune to inaccurate likelihood information. The resulting approach integrates Bayesian modeling with likelihood estimation.

Using noninformative priors to account for uncertainty of information is not novel in Bayesian statistics. However, this work seems to be the first at least among engineering methods that uses the Bayesian approach for estimating the likelihood information in conjunction with latent variable modeling. An important barrier to the use of Bayesian methods is that, in practice, people often do not have all the information needed for applying these methods. In this regard, this extended BLVR-S approach allows the building of Bayesian models with uncertain likelihood information. This makes the proposed approach relevant to many existing Bayesian and maximum likelihood methods including those used for tasks such as data rectification, system identification and chemometric modeling. Furthermore, it can be relatively easily applied in real laboratory and industrial settings. This

article also provides new insight into the practical meaning and differences between methods for choosing a noninformative prior. By addressing such practical challenges, it is hoped that Bayesian methods become more accessible to the engineering community.

The rest of this article is organized as follows. The next section provides background about Bayesian estimation, Gibbs sampling, and BLVR-S. This is followed by a description of details about the algorithm of BLVR-S with noninformative priors for noise variances. Next, several case studies are presented to compare the performance of different modeling methods and to illustrate the characteristics of the proposed method. Finally, the last section gives a summary of this work and discusses possible future research directions.

Background

Bayesian estimation

In Bayesian models, all the model parameters and data are treated as stochastic variables, with a joint distribution. Data $\mathbf{D}$ has a distribution conditional on the model parameters $\boldsymbol{\theta}$, denoted by $P(\mathbf{D} \mid \boldsymbol{\theta})$. Considering this distribution as a function of the parameters given the data, it is called the likelihood function. Maximizing this likelihood function results in the maximum likelihood estimate (MLE) of parameters. In contrast, in Bayesian methods, both the likelihood and prior information about model parameters are used to get estimates by combining them to get the updated posterior distribution of parameters. The Bayes rule is shown as follows,

$$P(\boldsymbol{\theta}|\mathbf{D}) = \frac{P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{D})}, \tag{1}$$

where, $P(\boldsymbol{\theta})$ is the prior distribution, which expresses domain knowledge about the model parameters before the measured data are obtained. The posterior distribution contains all the information available about the problem. The Bayes point estimates from the posterior distribution depend on the choice of loss function. A loss function $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is defined to measure the consequence of a discrepancy between the true parameter $\boldsymbol{\theta}$ and the estimate $\hat{\boldsymbol{\theta}}$. By minimizing the expected loss, the Bayes estimate $\hat{\boldsymbol{\theta}}_L$ can be found,

$$\hat{\boldsymbol{\theta}}_L = \arg\min_{\hat{\boldsymbol{\theta}}} \mathrm{E}(L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})|\mathbf{D}). \tag{2}$$

Bayes estimates for some of the most popular loss functions correspond to some common characteristics of the posterior distribution. For example, for the 0–1 loss function, the Bayes estimate is the posterior mode, the so-called maximum a posteriori (MAP) estimate, for the absolute error loss function, the estimate is the posterior median, and for the squared error loss function, it is the posterior mean. Like other statistical models, a Bayesian model provides a mathematical framework that is not guaranteed to reflect the physical nature of the underlying system. Nonetheless, it offers one way to estimate or forecast variables of the system with measurements and prior knowledge.

Gibbs sampling

Bayesian point estimates or moments may either be obtained through numerical routines, such as nonlinear programming, or by Monte Carlo Sampling.[25,26] For example, given the samples $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n\}$ from $P(\boldsymbol{\theta} \mid \mathrm{D})$, the posterior mean can be approximated as,

$$E(\boldsymbol{\theta}|\mathbf{D}) \approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_j \tag{3}$$

According to the law of large numbers, as n goes to infinity, this approximation converges to the true mean value. It is often much more convenient to use Monte Carlo samples from the posterior distribution to approximate such characteristics, especially when the posterior distribution is high dimensional.

Gibbs sampling is among the most popular for Bayesian computation and is a special type of Markov Chain Monte Carlo (MCMC) approach.[25] A Markov Chain is a sequence of random variables, $\{x_0, x_1, \ldots, x_{k-1}, x_k, \ldots\}$, in which, given the whole past history of the chain, the distribution of the current variable depends only on the immediate past, that is,

$$P(x_i|x_{i-1}, x_{i-2}, ..., x_1, x_0) = P(x_i|x_{i-1}) \tag{4}$$

In MCMC, the random variables in a Markov Chain are drawn sequentially and each one of them follows the same distribution, which is called the stationary distribution of the Markov Chain. Therefore, although those random variables are not independent, they can be treated as independent samples from that stationary distribution and be used to approximate this distribution. Hence, the goal of Bayesian computation is to make the stationary distribution of a Markov Chain be the posterior distribution of interest. Gibbs sampling provides an easy and efficient way to achieve this goal.

For a posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{D})$, elements of the parameter vector $\boldsymbol{\theta}$ are drawn sequentially, from their corresponding full conditional distributions. For example, at the i-th time step of the Markov Chain, a sample of the first element of the m dimensional θ is drawn from its conditional distribution $P(\theta_1 \mid \mathbf{D}, \theta_2^{(i-1)}, \theta_3^{(i-1)},\ldots,\theta_m^{(i-1)})$, where $\theta_j^{(i-1)}$ is the value of the j-th element of θ at the previous time step in the Markov Chain. After the new sample, θ_1 is drawn, a sample of $\boldsymbol{\theta}_2$ can be drawn from its distribution conditional on the most recent samples of other elements of $\boldsymbol{\theta}$, that is, $P(\theta_2 \mid \mathbf{D}, \theta_1^{(i)}, \theta_3^{(i-1)},\ldots,\theta_m^{(i-1)})$. In this manner, all the samples of all elements of $\boldsymbol{\theta}$ can be drawn sequentially at this time step. The detailed algorithm is as follows:

- for $i = 1 : n$
  - for $j = 1 : m$
    - draw $\theta_j^{(i)}$ from $P(\theta_j \mid \mathbf{D}, \theta_1^{(i)}, \theta_2^{(i)},\ldots,\theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)},\ldots, \theta_m^{(i-1)})$
  - end for
- end for

The elements of $\boldsymbol{\theta}$ can also be divided into several blocks, and samples of blocks instead of individual elements, can be drawn sequentially to greatly reduce the number of iterations in this algorithm.

Gibbs sampling is a special case of the more general Metropolis-Hastings sampling[25] algorithm. This algorithm guarantees that the created Markov Chain has a desired stationary distribution. The advantage of Gibbs sampling over other types of Metropolis-Hastings sampling methods[27] is that it does not have a rejection step unlike other methods, making it much more efficient. Also the sequential approach in Gibbs sampling is very effective for dealing with high dimensional distributions which are broken into many lower dimensional distributions. This is similar in principle to the univariate search optimization method.[28]

Because the starting point of the Markov Chain is arbitrary, it usually takes some time steps for the chain to converge to its stationary distribution. Therefore, the first k (such as 1000) samples are regarded as "burn-in" and discarded. Only samples thereafter are used for Monte Carlo approximation. Another practical problem of MCMC arises from the fact that the samples are actually not independent. Although some statisticians[29] argue that this is not a problem, in practice, a thinning step is often considered as a standard way to reduce the correlation among samples. Every r-th (such as 5th or 10th) sample is chosen instead of all the samples after burn-in. This thinning step is applied in the popular MCMC software BUGS.[30,31]

Sampling-based BLVR with known noise variances

In the optimization-based BLVR[6] method, both input $\boldsymbol{X}$ ($n \times m$) and output y ($n \times 1$) are assumed to be contaminated by measurement noise. The noise-free input $\tilde{\boldsymbol{X}}$ and output $\tilde{\boldsymbol{y}}$ are modeled as,

$$\tilde{\boldsymbol{X}} = \boldsymbol{Z}\boldsymbol{\alpha}^T \tag{5}$$

$$\tilde{\boldsymbol{y}} = \boldsymbol{Z}\boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{Z}$ is an $n \times p$ score matrix, $\boldsymbol{\alpha}$ is an $m \times p$ loading matrix, p is the rank of the model, and $\boldsymbol{\beta}$ ($p \times 1$) is the regression parameter vector of output variable on the score vectors. Loading vectors form an orthogonal basis of a lower dimensional space and satisfy the following constraint,

$$\boldsymbol{\alpha}^T\boldsymbol{\alpha} = \mathbf{I}. \tag{7}$$

The regression parameter vector $\boldsymbol{b}$ ($m \times 1$) of output variable on input variables is calculated as,

$$\boldsymbol{b} = \boldsymbol{\alpha}\boldsymbol{\beta}. \tag{8}$$

The measurement noise for each observation is assumed to be independent and identical, following a Gaussian distribution. Also, the measurement noise in input variables is assumed to be independent from the noise in the output variable. Denoting $\boldsymbol{x}_i = \boldsymbol{X}(i,:)^T$, where $\boldsymbol{X}(i,:)$ is the i-th row of $\boldsymbol{X}$, and $y_i = \boldsymbol{y}(i)$, the likelihoods of $\boldsymbol{x}_i$ and y_i are assumed to be,

$$\boldsymbol{x}_i \sim \text{Normal}(\tilde{\boldsymbol{x}}_i, \boldsymbol{R}_x) \tag{9}$$

$$y_i \sim \text{Normal}(\tilde{y}_i, R_y), \tag{10}$$

where $\boldsymbol{R}_x$ ($m \times m$) and R_y (a scalar) denote the covariance matrices of the input and output noise, respectively.

In BLVR, prior information for $\tilde{\boldsymbol{X}}$ and $\boldsymbol{b}$ are combined with the likelihood information to provide estimates. These prior distributions are assumed to be either uniform or Gaussian to permit its formulation as a least squares minimization problem. In case of a uniform prior, the maximum a posteriori (MAP) solution is the same as the maximum likelihood solution. In case of a Gaussian prior, the prior of $\tilde{\boldsymbol{X}}$ is assumed to be

$$\tilde{\boldsymbol{x}}_i \sim \text{Normal}(\boldsymbol{\mu}_x, \boldsymbol{Q}_x), \tag{11}$$

where $\boldsymbol{\mu}_x$ ($m \times 1$) is the prior mean of noise free input variables, and $\boldsymbol{Q}_x$ ($m \times m$) is the prior covariance matrix of noise free input variables. Equations 5 and 7 result in a linear relationship between $\tilde{\boldsymbol{X}}$ and $\boldsymbol{Z}$,

$$\boldsymbol{Z} = \tilde{\boldsymbol{X}}\boldsymbol{\alpha}. \tag{12}$$

Because the latent variables are just a linear transformation of the noise free input variables, the prior for z_i ($z_i = \boldsymbol{Z}(i,:)^T$) is also Gaussian, given by,

$$z_i|\boldsymbol{\alpha} \sim \text{Normal}(\boldsymbol{\alpha}^T\boldsymbol{\mu}_x, \boldsymbol{\alpha}^T\boldsymbol{Q}_x\boldsymbol{\alpha}). \tag{13}$$

The Gaussian prior of $\boldsymbol{b}$ is assumed to be,

$$\boldsymbol{b} \sim \text{Normal}(\boldsymbol{\mu}_b, \boldsymbol{Q}_b), \tag{14}$$

where $\boldsymbol{\mu}_b$ ($m \times 1$) is the prior mean for $\boldsymbol{b}$ and $\boldsymbol{Q}_b$ ($m \times m$) is the prior covariance matrix of $\boldsymbol{b}$. Equations 7 and 8 lead to the following linear relationship between $\boldsymbol{\beta}$ and $\boldsymbol{b}$,

$$\boldsymbol{\beta} = \boldsymbol{\alpha}^T\boldsymbol{b}. \tag{15}$$

Therefore, the prior for $\boldsymbol{\beta}$ is also Gaussian,

$$\boldsymbol{\beta} \sim \text{Normal}(\boldsymbol{\alpha}^T\boldsymbol{\mu}_b, \boldsymbol{\alpha}^T\boldsymbol{Q}_b\boldsymbol{\alpha}). \tag{16}$$

The prior distribution of $\boldsymbol{\alpha}$ is assumed to be uniform. Based on the likelihood functions and prior distributions given above, the posterior distribution is,

$$\begin{aligned} P(\boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) &\propto P(\boldsymbol{X}, \boldsymbol{y}|\boldsymbol{\alpha}, Z, \boldsymbol{\beta})P(\boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\beta}) \\ &\propto P(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{Z})P(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\beta})P(\boldsymbol{Z}|\boldsymbol{\alpha})P(\boldsymbol{\beta}|\boldsymbol{\alpha}). \end{aligned} \tag{17}$$

The right hand side of Eq. 17 is the multiplication of likelihood of data and prior distribution of model parameters, details about the derivation can be found in the original paper of optimization-based BLVR.[6] In that approach, a 0–1 loss function is chosen for obtaining the Bayesian point estimate, which is the posterior mode. This requires solution of the following optimization problem,

$$\{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{\beta}}\} = \arg\max_{\boldsymbol{\alpha},\boldsymbol{Z},\boldsymbol{\beta}} P(\boldsymbol{\alpha}, \boldsymbol{Z}, \boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{y}) \quad \text{s.t. } \boldsymbol{\alpha}^T\boldsymbol{\alpha} = \boldsymbol{I}, \tag{18}$$

which may be solved via a numerical optimization routine, which is often computationally expensive. Furthermore, this approach does not provide error bounds for the estimates. To address these shortcomings, Chen et al.[24] developed a BLVR algorithm based on Monte Carlo sampling instead of

Table 1. Algorithm of Drawing Samples of Z and β by Gibbs Sampling

Target Distribution: $P(\mathbf{Z}, \boldsymbol{\beta}\|\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\alpha}})$
• For $l = 1 : K$ draw $\mathbf{Z}^{(l)}$ from $P(\mathbf{Z}\|\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}^{(l-1)})$ draw $\boldsymbol{\beta}^{(l)}$ from $P(\boldsymbol{\beta}\|\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\alpha}}, \mathbf{Z}^{(l)})$ • End for

optimization. In this BLVR-S approach, the overall objective function is decomposed as follows,

$$\{\hat{\mathbf{Z}}, \hat{\boldsymbol{\beta}}\} = \arg\max_{\mathbf{Z}, \boldsymbol{\beta}} P(\mathbf{Z}, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\alpha}}) \tag{19}$$

$$\text{s.t. } \{\hat{\boldsymbol{\alpha}}\} = \arg\max_{\boldsymbol{\alpha}} P(\boldsymbol{\alpha}|\mathbf{X}, \mathbf{y}, \hat{\mathbf{Z}}, \hat{\boldsymbol{\beta}}) \tag{20}$$

$$\boldsymbol{\alpha}^T \boldsymbol{\alpha} = \mathbf{I}.$$

The orthogonality constraint for $\boldsymbol{\alpha}$ makes it difficult to directly draw samples of the loading vectors. Therefore, to obtain the solution, first, an estimate of $\boldsymbol{\alpha}$ is obtained via PCR, MLPCR, or PLS. With the estimated $\boldsymbol{\alpha}$, the problem in Eq. 19 can be solved by Monte Carlo approximation. Gibbs sampling is chosen to draw samples of $\boldsymbol{\beta}$ and $\mathbf{Z}$ from $P(\mathbf{Z},\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \hat{\boldsymbol{\alpha}})$, those samples are used to estimate the score vectors and regression parameters. The global maximum can be reached by iteration of these two steps. But empirical studies[24] have shown that even without iteration, BLVR-S can provide better models than the optimization-based BLVR and other traditional methods.

To implement the Gibbs sampling algorithm in BLVR-S, full conditional distributions of $\mathbf{Z}$ and $\boldsymbol{\beta}$ have to be derived. Details about those conditional distributions for uniform and Gaussian priors are described in[24] and the resulting BLVR-S algorithm is shown in Table 1.

The Gaussian assumptions for likelihood and prior are not essential for BLVR-S, but they can greatly simplify the algorithm and save computation time. These assumptions can be relaxed by advanced methods, such as Importance Sampling[21] within Gibbs sampling.[32] In the rest of this article, all the BLVR notation and discussion refers to the Gibbs sampling-based BLVR approach, unless stated otherwise.

Method

Setup of noninformative priors

The method developed in this section makes the following simplifying assumptions, which may be relaxed quite easily, if needed. The measurement noise for each variable is assumed to be independent. Hence, the covariance matrices are diagonal and there are only $m + 1$ parameters to be estimated in $\mathbf{R}_x$ and R_y, namely, the diagonal elements of these covariance matrices. This assumption makes it relatively easy and less subjective to define noninformative priors for the unknown noise variances because there may be very few or no user-specified parameters. Noninformative priors are often specified in the form of uniform distributions; however, this is not a requirement. Even if a prior distribution is uniform for one variable, it could be a different type of distribution when the variable is transformed, as illustrated in the following example.

For a Continuous Stirred Tank Reactor (CSTR) of volume (V) 1 m^3, assume the inlet flow rate (F) is stochastic and uniformly distributed between 1 and 1.1 m^3/hr. Then the probability density function (PDF) of F is:

$$P(F) = 10, \quad 1 \le F \le 1.1. \tag{21}$$

The mean of the residence time τ is also stochastic, but since $\tau = \frac{V}{F}$, τ is not a linear function of F, and it is not uniformly distributed. In fact, the PDF of τ can be obtained by applying the following theorem:

Theorem 1. [33]*For a continuous random variable x with a PDF $P_x(x)$, $x \in \chi$, another random variable $y = g(x)$ ($g(\cdot)$ is a monotone function), and $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, then y has a PDF of:*

$$P_y(y) = P_x(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|, \quad y \in \mathcal{Y}. \tag{22}$$

Thus, the PDF of τ (given $V = 1$) is:

$$\begin{aligned} P(\tau) &= 10 \times \left|\frac{\mathrm{d}\frac{1}{\tau}}{\mathrm{d}\tau}\right| \\ &= \frac{10}{\tau^2}, \quad \frac{1}{1.1} \le \tau \le 1. \end{aligned} \tag{23}$$

The distributions of F and τ, shown in Figure 1 indicate that the distribution of τ is far from uniform. Suppose the measured variable is τ instead of F, and if it is observed to be uniformly distributed, then the assumption of F being uniform should be changed accordingly. This example illustrates the importance of choosing an appropriate variable or metric for assigning the uniform prior. Once this is determined, by applying Theorem 1, the prior distribution of other variables or under another metric can be derived.

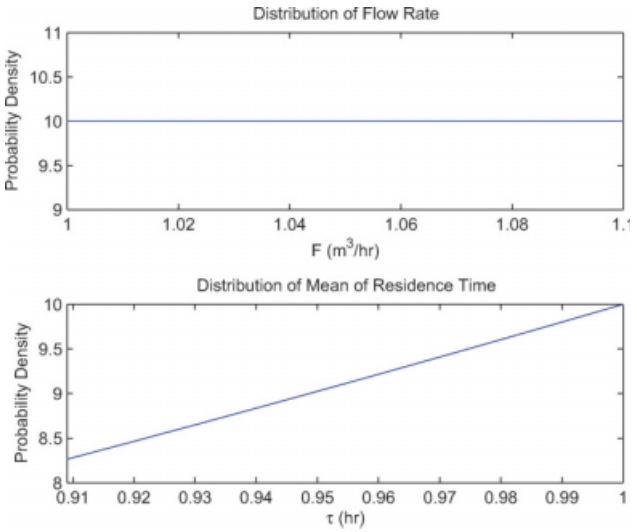

Figure 1. Probability distributions of flow rate and mean of residence time in the CSTR example.

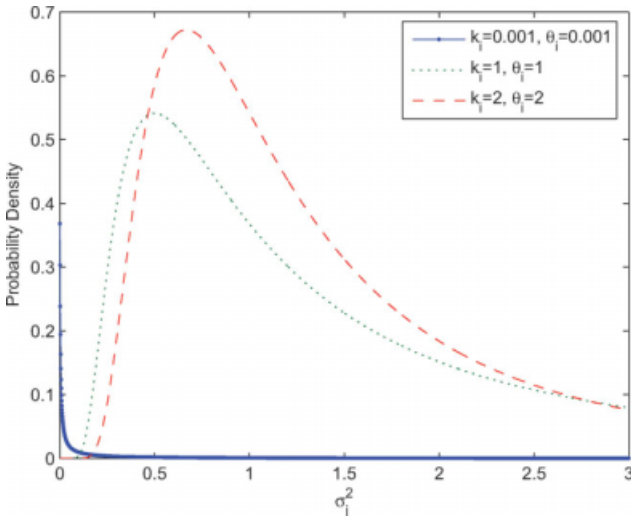

Figure 2. Probability density functions of Inverse-Gamma distributions with different parameters.

A convenient way to come up with the noninformative prior of the noise variances in BLVR is to assume that we have no information about the standard deviations of the noise. Let σ_j denote the standard deviation of noise ε_j of the j-th variable of the data set, that is,

$$\epsilon_j \sim \text{Normal}(0, \sigma_j^2), \quad j = 1, 2, ..., m, m+1, \tag{24}$$

where the first m variables are inputs and the $m+1$-th variable is the output. We may consider these variables to be uniformly distributed. Thus,

$$P(\sigma_j) \propto 1, \quad \sigma_j > 0. \tag{25}$$

By applying the above theorem, the prior distribution of σ_j^2 may be derived as,

$$\begin{aligned} P(\sigma_j^2) &\propto \left|\frac{d\sigma_j}{d\sigma_j^2}\right| \\ &\propto \left|\frac{d\sqrt{\sigma_j^2}}{d\sigma_j^2}\right| \\ &\propto \frac{1}{\sqrt{\sigma_j^2}}. \end{aligned} \tag{26}$$

However, because we can choose a different metric or variable for the uniform distribution, this noninformative prior is not the only one that is acceptable. Another popular choice is to assume that $\log \sigma_j$ is uniformly distributed, that is,

$$P(\log \sigma_j) \propto 1, \quad \sigma_j > 0. \tag{27}$$

Again, by applying Theorem 1, the prior distribution for σ_j^2 is:

$$\begin{aligned} P(\sigma_j^2) &\propto \left|\frac{d\log \sigma_j}{d\sigma_j^2}\right| \\ &\propto \left|\frac{d\frac{\log \sigma_j^2}{2}}{d\sigma_j^2}\right| \\ &\propto \frac{1}{\sigma_j^2}. \end{aligned} \tag{28}$$

In fact, this noninformative prior satisfies Jeffreys' rule and it is called Jeffreys prior.[34] More details are in Appendix.

In addition to the above two noninformative priors, the prior distributions of the variances of Gaussian noise are often assumed to be Inverse-Gamma, which has the PDF of the form:

$$P(\sigma_j^2|k_j, \theta_j) = \frac{\theta_j^{k_j}}{\Gamma(k_j)}(\sigma_j^2)^{-k_j-1}\exp\left(\frac{-\theta_j}{\sigma_j^2}\right), \tag{29}$$

where $\Gamma(\cdot)$ is a Gamma function. Given a Gaussian likelihood function, the Inverse-Gamma prior of σ_j^2 results in an Inverse-Gamma posterior distribution for σ_j^2. The Inverse-Gamma prior of σ_j^2 is called conjugate to Gaussian likelihood functions. Conjugate priors bring convenience to Bayesian computation, hence the Inverse-Gamma prior is a good candidate for the noninformative prior of σ_j^2. To make an Inverse-Gamma distribution resemble a uniform prior, a common choice is to set its parameters k_j and θ_j close to zero. In the popular Bayesian modeling software BUGS,[30] they are both set to be equal to 0.001. Figure 2 shows the PDF for several Inverse-Gamma distributions with different sets of parameters. Clearly, as k_j and θ_j decrease, the Inverse-Gamma distribution tends to become more flat.

In fact, the two types of noninformative priors discussed earlier in this section can be regarded as special cases of the Inverse-Gamma distribution. To obtain the noninformative prior in Eq. 26, $k_j \to -\frac{1}{2}, \theta_j \to 0$, and for the noninformative prior in Eq. 28, $k_j \to 0$, $\theta_j \to 0$. Three noninformative priors for σ_j^2 are summarized in Table 2, and are adopted in the applications of this work. They are also the types of priors considered in the work by Gelman.[17] Because they can be unified under the Inverse-Gamma distribution family, we only need to develop one extended sampling-based BLVR approach with Inverse-Gamma prior distributions for the noise variances, and this method is described next.

Table 2. Summary of Different Types of Noninformative Priors for σ_j^2

Type	Parameters of the Inverse-Gamma Distribution		Uniform (or Approximately Uniform) in the Space of
	k_j	θ_j	
1	0.001	0.001	σ_j^2
2	0	0	$\log \sigma_j$
3	$-\frac{1}{2}$	0	σ_j

BLVR-S with noninformative priors for noise variances

In the BLVR-S approach described in,[24] the noise variances for input and output variables are considered to be deterministic, and they are fixed throughout in the Gibbs sampling steps. By setting noninformative priors for these variances, as done in this article, they are treated as stochastic and are estimated along with the score vectors and regression parameters. Hence, an additional step is needed in the BLVR-S approach to draw samples of the noise variances. In this step, samples of noise variances are drawn from their full conditional distributions. The distributions of $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ remain the same as those described in,[24] except that the noise variance parameters in those distributions need to be updated in every time step.

To sample the noise variances by Gibbs sampling, their full conditional distributions have to be derived. As discussed in the previous subsection, the three types of noninformative priors of noise variance considered in this work can be unified under the Inverse-Gamma family. Because Inverse-Gamma is a conjugate prior for Gaussian likelihood functions, the full conditional distribution of σ_j^2 is also an Inverse-Gamma distribution. Without loss of generality, assume $j < m + 1$, that is, the j-th variable is an input variable. Because the measurement noise are independent, the conditional posterior distribution of σ_j^2 only depends on the likelihood of the j-th input variable, and the conditional distribution of σ_j^2 is:

$$P(\sigma_j^2|\boldsymbol{X}(:,j),\boldsymbol{Z},\boldsymbol{\alpha}) \propto P(\boldsymbol{X}(:,j)|\boldsymbol{Z},\boldsymbol{\alpha},\sigma_j^2)P(\sigma_j^2|k_j,\theta_j)$$

$$\propto \frac{1}{(2\pi)^{n/2}\sigma_j^n}\exp\left(\frac{-\sum_{i=1}^{n}(x_{ij}-\tilde{x}_{ij})^2}{2\sigma_j^2}\right)\frac{\theta_j^{k_j}}{\Gamma(k_j)}(\sigma_j^2)^{-k_j-1}\exp\left(\frac{-\theta_j}{\sigma_j^2}\right)$$

$$\propto (\sigma_j^2)^{-(\frac{n}{2}+k_j)-1}\exp\left(-\frac{\frac{\sum_{i=1}^{n}(x_{ij}-\tilde{x}_{ij})^2}{2}+\theta_j}{\sigma_j^2}\right), \quad (30)$$

where $\boldsymbol{X}(:, j)$ is the j-th column of $\boldsymbol{X}$, x_{ij} is the measurement of the j-th input variable in the i- th observation, $\tilde{x_{ij}}$ is its noise free value, which depends on the loading and score vectors, $\boldsymbol{Z}$ and α, respectively. Comparing Eq. 30 with Eq. 29, it is obvious that the conditional distribution of σ_j^2 indeed is another Inverse-Gamma distribution. Denoting $\sum_{i=1}^{n}(x_{ij}-\tilde{x}_{ij})^2$ as SSE_j, the conditional distribution of σ_j^2 is:

$$\sigma_j^2|\boldsymbol{X}(:,j),\boldsymbol{Z},\alpha \sim \mathrm{InverseGamma}\left(k_j'=\frac{n}{2}+k_j,\theta_j'=\frac{\mathrm{SSE}_j}{2}+\theta_j\right), \quad j=1,2,...,m. \quad (31)$$

Equation 31 provides some insight into the effect of the parameters in the Inverse-Gamma prior. When there are a large number of observations (n is large), $k_j' \approx \frac{n}{2}$ and k_j has little effect on the posterior distribution of σ_j^2. When the magnitude of noise in the j-th variable is large (SSE_j is large), $\theta_j' \approx \frac{\mathrm{SSE}_j}{2}$ and θ_j has little effect on the posterior distribution of σ_j^2. Because the assumption is that we know little about the true noise variances (that is why noninformative priors are needed), it is desirable to make k_j much smaller when compared with n and θ_j much smaller when compared with SSE_j, such that the prior distribution of σ_j^2 has little influence on its posterior distribution, that is, it is indeed "noninformative." This should be considered as a general guideline of determining the hyper-parameters of the noninformative priors. In practice, because n is known, it is easy to find a proper value of k_j. SSE_j may be estimated by approximating the "true" value as the result of simpler approaches such as PCR. Given the estimated SSE_j, a θ_j can be selected such that it is several magnitudes smaller than SSE_j. Similarly, the conditional distribution of σ_{m+1}^2 (or R_y) is,

$$\sigma_{m+1}^2|\boldsymbol{y},\boldsymbol{Z},\boldsymbol{\beta} \sim \mathrm{InverseGamma}\left(k_{m+1}'=\frac{n}{2}+k_{m+1},\theta_{m+1}'=\frac{\mathrm{SSE}_{m+1}}{2}+\theta_{m+1}\right), \quad (32)$$

where $\mathrm{SSE}_{m+1}=\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2$, in which y_i is the measured value of output.

Table 3. BLVR-S Algorithm

- Get $\hat{\boldsymbol{\alpha}}$ by applying PCA, MLPCA, or PLS to $(\boldsymbol{X}, \boldsymbol{y})$
- While not converge
  1. Get $\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{R}}_x$, and $\hat{R}_y$ by Gibbs sampling with K samples in the Markov Chain.
     For $s = 1 : K$
       Draw $\boldsymbol{Z}^{(s)}$ from $P(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{y},\hat{\boldsymbol{\alpha}},\boldsymbol{\beta}^{(s-1)})$
       Draw $\boldsymbol{\beta}^{(s)}$ from $P(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y},\hat{\boldsymbol{\alpha}},\boldsymbol{Z}^{(s)})$
       Draw $\sigma_j^{2(s)}$ from InverseGamma $(\frac{n}{2}+k_j, \frac{\mathrm{SSE}_j^{(s)}}{2}+\theta_j)$, $(j = 1,2,...,m+1)$
     $\boldsymbol{R}_x^{(s)} = \mathrm{diag}[\sigma_1^{2(s)}, \sigma_2^{2(s)},...,\sigma_m^{2(s)}]$, $R_y^{(s)} = \sigma_{m+1}^{2\,(s)}$
     End for
     Estimate $\boldsymbol{Z}$, $\boldsymbol{\beta}$, $\boldsymbol{R}_x$, and R_y based on $\{(\boldsymbol{Z}^{(1)},\boldsymbol{\beta}^{(1)}),(\boldsymbol{Z}^{(2)},\boldsymbol{\beta}^{(2)}),\ldots, (\boldsymbol{Z}^{(K)},\boldsymbol{\beta}^{(K)})\}$
  2. Get $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{y}}$ based on $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{Z}}$, and $\hat{\boldsymbol{\beta}}$
  3. Get $\hat{\boldsymbol{\alpha}}$ by applying PCA, MLPCA, or PLS to $(\hat{\boldsymbol{X}},\hat{\boldsymbol{y}})$
- End while

Table 3 shows the detailed algorithm of BLVR-S with the Inverse-Gamma priors for the noise variances. In the Gibbs sampling steps, the SSE_j depends on the most recent samples of $\boldsymbol{Z}$ and β. The SSE_j at the s-th time step is denoted as $\mathrm{SSE}_j^{(s)}$ and is calculated with $\beta^{(s)}$ and $\boldsymbol{Z}^{(s)}$.

As described at the beginning of this section, three different noninformative priors are considered in this work, the noise variance of each variable can choose one of them as its prior distribution. Hence, the parameter k_j and θ_j in the priors can be different for each σ_j^2. But for simplicity, the same noninformative type of prior is used for all the σ_j^2 here. BLVR-S with the Type 1, Type 2, or Type 3 prior in Table 2 for noise variance is denoted as BLVR-v1, BLVR-v2, and BLVR-v3. In the next section, the performance of those three approaches is compared with other methods via various case studies.

Experiments

Simulated high dimensional data

This example is adapted from a simulated example in Ref. 24. The data set contains 50 observations, 15 input variables

Table 4. Summary of the Different Settings in Each Case of the Simulated High Dimensional Data Example

Case	l	SNR of the Inputs	SNR of the Output
1	2	3	3
2	2	$1 \sim 15$	3
3	3	$1 \sim 15$	3
4	0	3	3
5	0	$1 \sim 15$	3

and 1 output variable, and the true rank of the input matrix is known to be 10. Both input and output variables are contaminated by Gaussian measurement noise, and variances of the noise are determined by the Signal to Noise Ratio (SNR) of variables. The loading vectors for BLVR are obtained from PCR. The historical Gaussian priors for the noise free input variables and the regression parameters of BLVR, BLVR-v1, BLVR-v2, and BLVR-v3 are generated based on the results of applying PCR to a historical data set with 600 observations. The true rank 10 is used for all methods. The performance of PCR, MLPCR, BLVR, BLVR-v1, BLVR-v2, BLVR-v3 are compared. The four variations of BLVR are evaluated for uniform or historical Gaussian priors. With the uniform prior, they are denoted as BLVR(u), BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3. With the Gaussian prior based on historical data, they are denoted as BLVR(h), BLVR(h)-v1, BLVR(h)-v2, and BLVR(h)-v3. Results are based on 50 realizations.

To mimic our ignorance about the measurement error, the initial guess for σ_j^2 is assumed to be:

$$\sigma_j^{2(0)} = \sigma_j^2 \exp(u_j), \tag{33}$$

where $u_j \sim$ Uniform$(-l, l)$ $(l \leq 0)$. The variable l denotes inaccuracy of the initial guess for σ_j^2. The larger the l is, the bigger the possibility of having a guess that is far from the truth. When $l = 0$, the guess is the same as the true noise variance, which means that accurate likelihood information is available. This initial guess of the noise variance is used in MLPCR and is also used as the first sample in the Markov Chain for σ_j^2 in BLVR, BLVR-v1, BLVR-v2, and BLVR-v3.

Five different cases are studied in this example by varying the SNR and l. Parameter settings for these cases are summarized in Table 4.

- Case I. The SNR for all the input and output variables are set to be 3 with $l = 2$.
- Case II. The SNR in input variables varies from 1 to 15, which means that the magnitudes of measurement noise are quite different among the variables. Other conditions remain the same as in Case I.
- Case III. In this case, $l = 3$, which means that there is larger error in the initial guess of the measurement noise variance. Other conditions are the same as Case II.
- Case IV. This case is the same as Case I except that $l = 0$, which means that the true noise variances are used as the initial guesses.
- Case V. All settings are the same as in Case 2, except that $l = 0$.

Table 5 shows the average, over 50 realizations, of the normalized mean squared error (MSE) of the output variable for different methods in each of the five cases. The MSE is calculated by comparing the model output with the true noise-free output, the same method is applied in all examples of this article. The MSE are then normalized by the MSE of PCR in each realization, because PCR is one of the most popular approaches in practice and it is reasonable to use it as a benchmark for comparison.

As shown in Table 5, in Case 1, the performance of MLPCR and BLVR(u) are worse than PCR. This is not surprising because they make use of inaccurate error variances provided by Eq. 33, while PCR assumes same error variances for all variables. In this case, the assumption made by PCR is not that bad because the differences of the actual error variances of different variables are relatively small. Although BLVR(h) also makes use of the inaccurate likelihood information, its performance is slightly better than PCR, due to the utilization of extra information in the historical prior. As for BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3, they have similar performance which are almost the same or slightly worse than PCR but are all better than BLVR(u) without the noninformative prior for noise variances. The estimates, BLVR(h)-v1, BLVR(h)-v2, and BLVR(h)-v3 using the noninformative priors on the noise

Table 5. Simulated High Dimensional Data Example: Average of Output MSE for Different Methods, Normalized by the MSE of PCR in Each Realization; 50 Realizations

MSE	MLPCR	BLVR(u)	BLVR(u)-v1	BLVR(u)-v2	BLVR(u)-v3	BLVR(h)	BLVR(h)-v1	BLVR(h)-v2	BLVR(h)-v3
Case 1: $SNR_x = 3$, $SNR_y = 3$, $l = 2$									
Y(testing)	1.2420	1.1801	1.0575	1.0573	1.0064	0.9059	0.7463	0.7559	0.7370
Y(training)	1.1863	0.9038	0.7924	0.7990	0.7446	0.8096	0.7510	0.7622	0.7333
Case 2: $SNR_x = 1,2,\ldots,15$, $SNR_y = 3$, $l = 2$									
Y(testing)	0.8997	0.9234	0.8846	0.9061	0.8441	0.7303	0.7343	0.7498	0.7158
Y(training)	0.9299	0.9017	0.8513	0.8219	0.7941	0.7663	0.7733	0.7940	0.7469
Case 3: $SNR_x = 1,2,\ldots,15$, $SNR_y = 3$, $l = 3$									
Y(testing)	1.2559	1.2189	0.9494	0.9607	0.8866	0.9274	0.8025	0.8240	0.7843
Y(training)	1.2190	1.1174	0.8040	0.8040	0.7731	0.9735	0.8563	0.8840	0.8201
Case 4: $SNR_x = 3$, $SNR_y = 3$, $l = 0$									
Y(testing)	0.9545	0.9734	1.0696	1.0828	1.0148	0.8141	0.8222	0.8365	0.8118
Y(training)	0.9622	0.7019	0.8393	0.8461	0.7649	0.6852	0.8121	0.8298	0.7899
Case 5: $SNR_x = 1,2,\ldots,15$, $SNR_y = 3$, $l = 0$									
Y(testing)	0.6919	0.7143	0.8832	0.8868	0.8368	0.5194	0.6813	0.6984	0.6684
Y(training)	0.7461	0.6933	0.7926	0.7939	0.7756	0.5426	0.7063	0.7349	0.6875

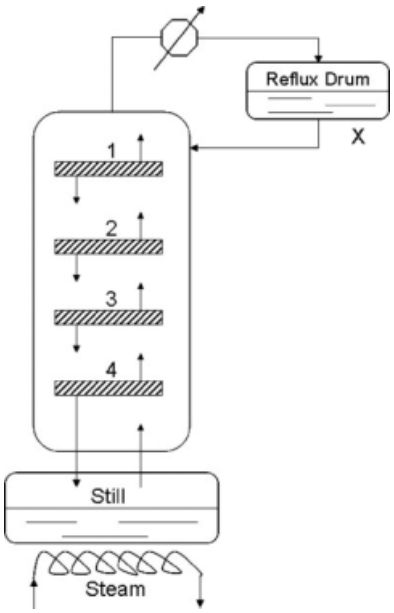

Figure 3. Illustration of the batch distillation process.

variances outperform all other methods. This is as expected because these methods make use of the most information, along with noninformative priors for the noise variances. The three types of noninformative priors have similar effects on the performance of BLVR. This makes sense because the number of observations and the measurement noise are relatively large comparing to the hyper-parameters in the all the noninformative priors, and as discussed in the previous section, the choice of the hyper-parameters in the Inverse-Gamma prior distributions does not have much effect on the performance of BLVR.

A similar trend is observed for Case 2, except that both MLPCR and BLVR(u) now outperform PCR. This can be explained by the fact that in this case, the noise variances of the input variables vary a lot more than in Case 1 due to SNR_x being between 1 and 15. Thus, the PCR assumption of equal noise variance in each variable is violated, whereas the MLPCR assumption of unequal variance is satisfied. Even though MLPCR and BLVR(u) use inaccurate noise variances, they still do a better job than PCR. Better performance of the three versions of hierarchical BLVR is also observed in this case.

As for Case 3, the trend is similar to that in Case 1 and PCR again outperforms MLPCR and BLVR(u). In this case, the variation among the noise variances of different variables is the same as in Case 2. However, the quality of likelihood information is worse as indicated by the larger l. Hence, PCR yields a better model than MLPCR or BLVR(u) with inaccurate likelihood information. The inferior likelihood also worsens the performance of various BLVR versions, but they perform better than conventional BLVR without noise estimation.

In Case 4, because accurate noise variances are available, BLVR(h) has the best performance, but the performance of BLVR(h)-v1, BLVR(h)-v2, and BLVR(h)-v3 are not far behind. It is worth noting that the latter three approaches are not aware of the fact that accurate noise information is available, they only use the accurate noise variances as the starting point in the Markov chain and allow their values to change in the Gibbs sampling process. The fact that they still achieve comparable performance as BLVR(h) (which uses the accurate noise variances throughout Gibbs sampling) demonstrates the ability of the proposed approach to estimate noise information. The performance of MLPCR, BLVR(u), BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3 are also similar.

In Case 5, when accurate noise variances are also available, BLVR(h) again has the best performance. When compared with the results in Case 4, there is a larger difference in performance when comparing BLVR(u) with BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3, or comparing BLVR(h) with BLVR(h)-v1, BLVR(h)-v2, and BLVR(h)-v3. This is expected because the variation among the noise variances is much larger than in Case 4 and having accurate noise variances information about the noise variances throughout Gibbs sampling can greatly improve the model quality.

Inferential modeling of a batch distillation process

This problem originated from a 4-stage batch distillation column described in Ref. 35. Binous[36] coded a program to simulate the separation of the mixture of methanol and water by this distillation column, assuming that the reflux ratio is 10, the pressure is 1 bar, and the initial molar fraction of methanol is 80%. Constant molar holdup is assumed at each stage to be 20 mol, and in the still is 100 mol. The vapor flow rate is 10 mol/min. Figure 3 illustrates this batch distillation process. A simulation using Binous' program is run to obtain the temperature of each stage and the still, and the methanol molar fraction of the distillate in the reflux drum.

The temperature and molar fraction are sampled with a time step of 1 min from 1 to 200 min. Change of temperature with time in the distillation process is shown in Figure 4. As the concentration of methanol in the still decreases, the temperatures of the stages also decrease and become homogeneous while the temperature of the still slightly

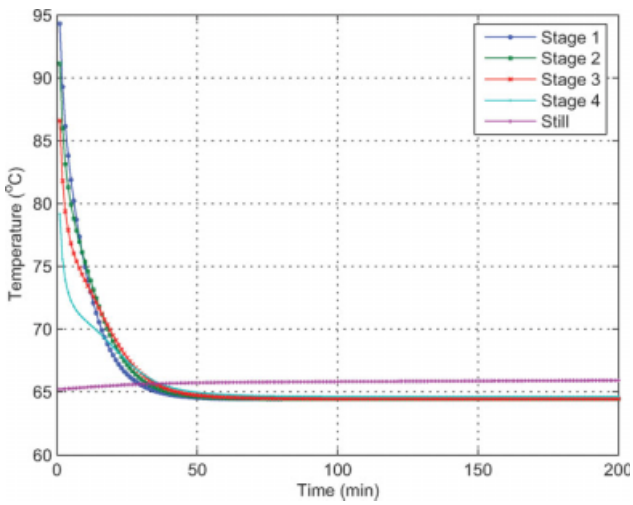

Figure 4. Change of temperature with time in the batch distillation process.

Table 6. Batch Distillation Example: Average of Output MSE for Different Methods of the Batch Distillation Problem, Normalized by the MSE of PCR in Each Realization; Rank 2, 50 Realizations

MSE	PLS	MLPCR	BLVR(u)	BLVR(u)-v1	BLVR(u)-v2	BLVR(u)-v3
			Case 1: $l = 3$			
Y(testing)	0.9960	0.9811	0.9596	0.5251	0.5226	0.5272
Y(training)	0.9275	1.0486	0.6118	0.3578	0.3699	0.3466
			Case 2: $l = 0$			
Y(testing)	0.9960	0.5306	0.5281	0.5273	0.5236	0.5303
Y(training)	0.9256	0.4602	0.2267	0.3376	0.3477	0.3252

increases. Independent Gaussian noise are then added to the true values of temperature and molar fraction, those contaminated numbers are considered as measurements of the system. The variances of the measurement noise for the temperature of the four stages and the still are 1^2, 0.8^2, 0.6^2, 0.4^2, and 0.2^2, respectively; the variance of the measurement noise for the molar fraction is 0.03^2.

Measurements at the 200 time steps are randomized and divided into training and testing sets, each containing 100 observations. There are five input variables (temperature at four stages and the still) and one output variable (methanol molar fraction of the distillate). Although this is a dynamic system, the dynamics are not considered in this modeling problem because the available measurements of the current temperature at the stages and the still are strongly related to the current molar fraction of the distillate. PCR, PLS, MLPCR, BLVR(u), BLVR(u)-v1, BLVR(u)-v2, BLVR(u)-v3 are applied to model this process. Rank 2 is used for all methods. In the first case, the initial guess of the variances of measurement noise is obtained in the same way as in previous examples, where $l = 3$; in the second case, the true variances of measurement noise are used, that is, $l = 0$. Table 6 shows the average normalized output MSE of different methods over 50 realizations in these two cases. The MSE is calculated by comparing the model output with the true noise-free output, then normalized by the MSE of PCR.

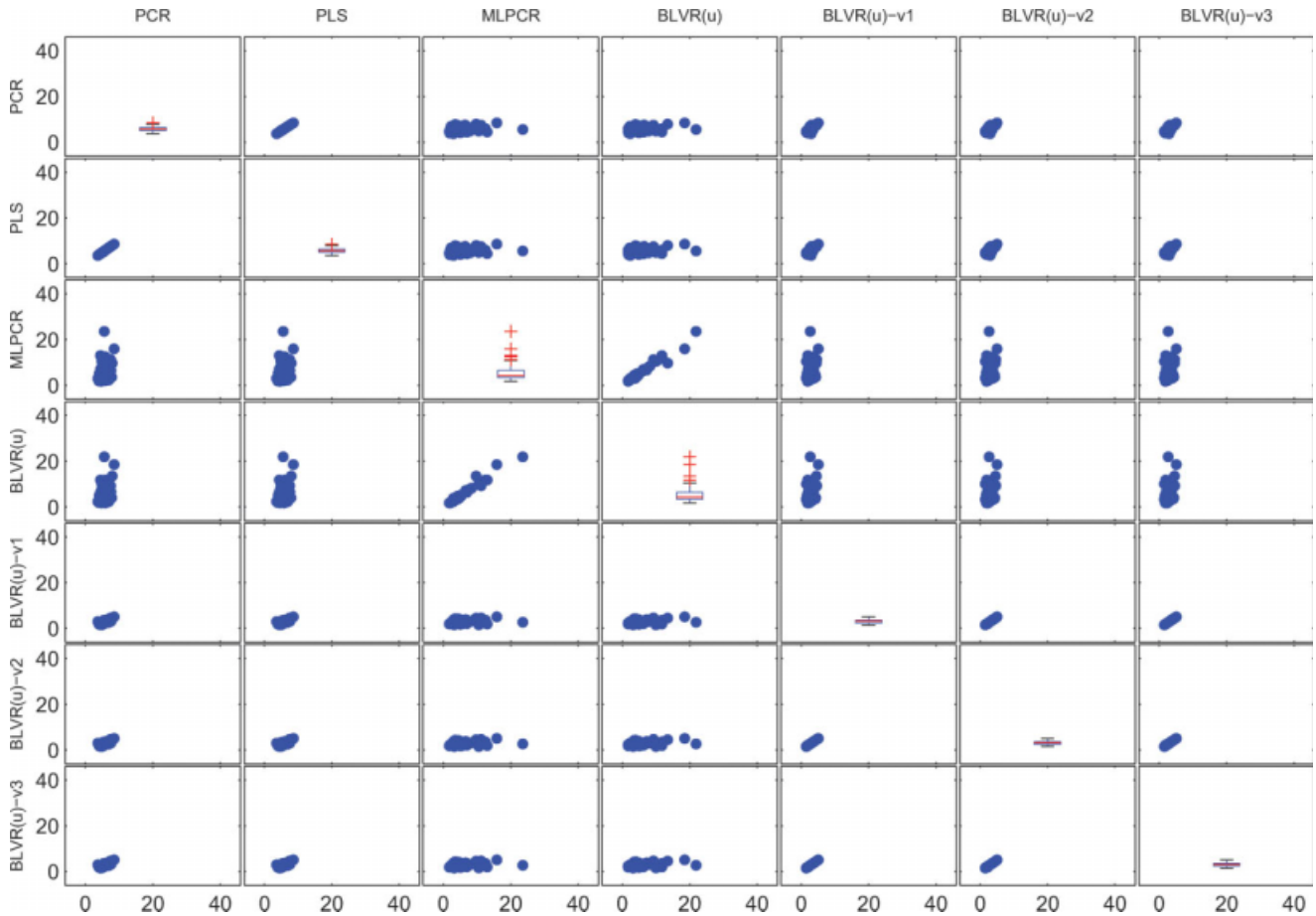

Figure 5. Case 1 of the batch distillation example: Matrix plot of testing MSE of output, normalized by the variance of noise in the output; $l = 3$, 50 realizations.

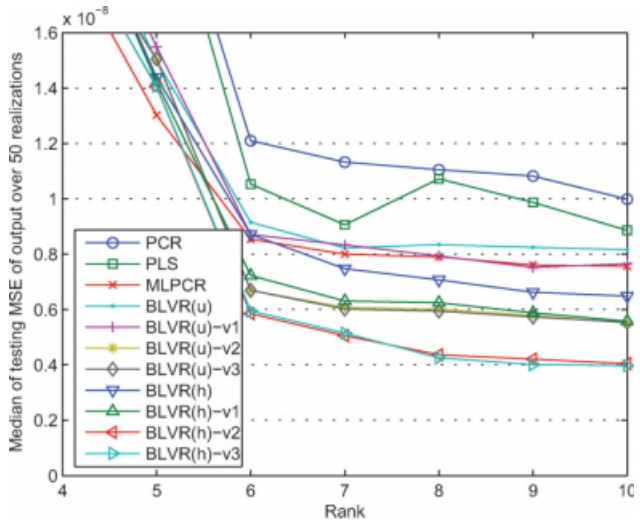

Figure 6. Case 1 of the continuous distillation column example: Median of testing MSE of output over 50 realizations, $l = 3$.

When the noise variance is not known very well, as indicated by $l = 3$, BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3 have significantly higher accuracy for the output as compared to other methods. When the error is known accurately ($l = 0$), MLPCR, BLVR(u), BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3 have similar performance. Figure 5 shows the matrix plot of the testing MSE of the output in the 50 realizations for the first case, normalized by the variance of measurement noise. The diagonal of Figure 5 presents the box plots of the normalized MSE for each method. The paired plots of BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3 confirms that the three type of noninformative priors have similar performance. These plots are almost straight lines on the diagonal, which means that in each realization, they have similar MSE. With no surprise, BLVR(u) and MLPCR also have similar performance and most points also line on the diagonal in their paired plots. It is clear that in the first case there are many outliers for MLPCR and BLVR(u) over 50 realizations, and the box plots also show that they have relatively large variance of MSE because they are vulnerable to bad likelihood information; while the variances of MSE for BLVR(u)-v1, BLVR(u)-v2, and BLVR(u)-v3 are much smaller.

Inferential modeling of a continuous distillation process

This inferential modeling problem is derived from a problem that has been thoroughly studied by Skogestad et al.[4,37–39] There are 40 theoretical stages and a total condenser in this distillation column. It is used to separate a binary mixture with a constant relative volatility of 1.5. The difference in the boiling points of the two components is 13.5°C. The feed flow rate is 1 kmol/min. In this study, the feed composition of the light component is set to be stochastic, normally distributed with mean of 0.5 and variance of 0.01. This composition could also be set to have other types of features such as the deterministic features in the previous example and Figure 4, and still get similar improvement. A nonlinear dynamic simulation of 500 min is run with the Matlab and Simulink files provided by Skogestad[39] to obtain the composition of the mixture at each stage and the condenser with a time step of 1 min. The temperature at each stage is calculated by a linear approximation[39] with the composition data. Those temperatures are further contaminated by Gaussian measurement noise with mean zero and different variances. The variance of the measurement noise of the first and last 10 stages is 9×10^{-6}, the variance of the measurement noise of the stages $16 \sim 25$ is 1×10^{-4}, and the variance of the measurement noise of the remaining 10 stages is 2.5×10^{-5}. The composition of the distillate is also contaminated by Gaussian noise with mean of zero and variance of 4×10^{-8}. Like the previous example, the dynamics of this system are ignored in the modeling process. The 500 observations are randomized and divided into training and testing sets with 250 observations in each. Two cases have been considered in this simulation study. In the first case, the initial guess of the variance of measurement noise is obtained in the same way as in the previous examples, where $l = 3$; whereas in the second case, the true variances of measurement noise are used, that is, $l = 0$. Another simulation of 500 min is run to get a historical data set, which is used to generate the historical prior distributions by PCR. Then we apply all the methods to estimate the composition of the distillate with the temperature measurements at the 40 stages.

Figures 6 and 7 show the median training and testing MSE of the output over 50 realizations of the above methods with different ranks in the two cases. The MSE is calculated by comparing the model output with the true noise-free output. According to Figure 6, in the first case, BLVR(h)-v2 and BLVR(h)-v3 have the best performance while apparently BLVR(h)-v1 performs worse. Also BLVR(u)-v2 and BLVR(h)-v3 perform better than BLVR(u)-v1. This is a little different from the trend observed in previous case studies, where the differences in performance among different

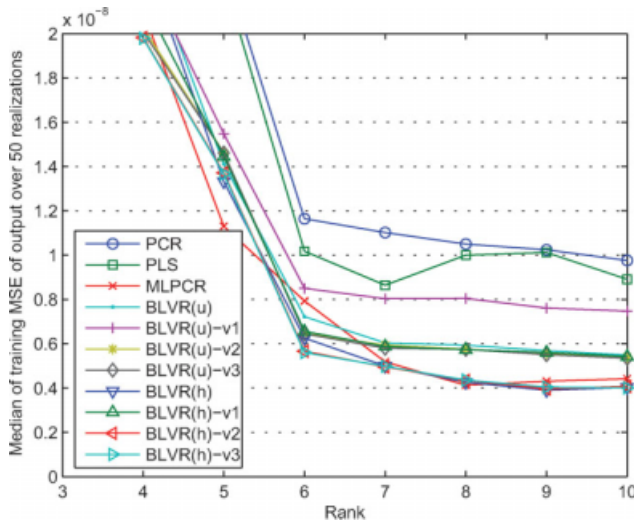

Figure 7. Case 2 of the continuous distillation column example: Median of testing MSE of output over 50 realizations, $l = 0$.

noninformative priors are less significant and visible. This difference can be explained by recalling the effect of θ_j on the posterior distribution of σ_j^2. As discussed in the third section, θ_j has little effect only when SSE_j is much larger than θ_j. Although in this example, the noise variances of most variables are much smaller than 0.001, which is the value of θ_j in the Type 1 prior. Hence, the Type 1 prior has significant effect on the posterior distribution. Therefore, it is more biased than the Type 2 and Type 3 prior in this problem, which leads to worse performance. Hence, when the measurement noise are very small, Type 2 and Type 3 priors should be preferred. Figure 7 show that BLVR(h), BLVR(h)-v2, and BLVR(h)-v3 perform equally well and slightly better than MLPCR in Case 2. It also shows that Type 1 prior performs worse than Type 2 and Type 3 priors.

Conclusions

This article proposes a general Bayesian approach to address the challenge posed by a lack of accurate information about measurement errors or likelihood in process modeling. This approach sets up a Bayesian model by treating the noise variance as a stochastic variable with corresponding uncertainty information. It is applied to developing an extended BLVR-S method with noninformative prior for the variances in the likelihood functions. Case studies show that the proposed approach can effectively handle the problem of inaccurate information about measurement noise, and that by setting noninformative priors for the noise variances, the model quality is, in general, better than what may be obtained by fixing the noise variances in BLVR-S. This is especially true when the available information regarding the noise variance may be far from the underlying truth. Although PCR or PLS seem attractive in that situation, they are ill-suited for utilizing prior information about the parameters and variables being estimated. Also the variances of the measurement noise can be estimated from BLVR-S with a noninformative prior approach. This is an additional benefit that PCR or PLS cannot readily provide.

As shown in this article, a prior distribution can be setup for the parameters in the likelihood functions to reflect the uncertainty about the likelihood information. A noninformative prior can be used when there is little knowledge about the measurement noise. Three different types of noninformative priors suggested in this paper are unified under the Inverse-Gamma distribution family. Because the hyper-parameters of the Inverse-Gamma prior distributions in practice can be different from the priors discussed in this article, the choice of priors in this framework is not limited to those three priors. Although different priors usually have similar performance, in some cases, as illustrated in this work, the performance of BLVR-S is sensitive to the choice of the prior hyper-parameters. Thus it is important to choose the prior hyper-parameters according to the number of observations in the data set and the magnitude of measurement noise. The proposed approach is general and should be useful for improving the performance of other process modeling methods that require likelihood information that is not readily available. This work should also convey the many benefits of Bayesian modeling, thus encouraging their use.

Acknowledgements

Financial support from the US National Science Foundation (CTS-0321911) and the distillation simulation programs provided by Dr. Binous and Dr. Skogestad are gratefully acknowledged.

Literature Cited

1. Jackson JE. *A User's Guide to Principal Components*. New York: Wiley, 1991.
2. Lindberg W, Persson JA, Wold S. Partial least-squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Anal Chem.* 1983;55:643–648.
3. Mejdell T, Skogestad S. Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression. *Ind Eng Chem Res.* 1991;30:2543–2555.
4. Wentzell PD, Andrews DT, Hamilton DC, Faber K, Kowalski BR. Maximum likelihood principal component analysis. *J Chemom.* 1997;11:339–366.
5. Nounou MN, Bakshi BR, Goel PK, Shen X. Bayesian principal component analysis. *J Chemom.* 2002;16:576–595.
6. Nounou MN, Bakshi BR, Goel PK, Shen X. Process modeling by Bayesian latent variable regression. *AIChE J*. 2002;48:1775–1793.
7. Leger MN, Vega-Montoto L, Wentzell PD. Methods for systematic investigation of measurement error covariance matrices. *Chemom Intell Lab Syst*. 2005;77:181–205.
8. Crowe CM. Data reconciliation—progress and challenges. *J Process Control*. 1996;6:89–98.
9. Almasy GA, Mah RSH. Estimation of measurement error variances from process data. *Ind Eng Chem Res.* 1984;23:779–784.
10. Darouach M, Ragot R, Zasadzinski M, Krzakala G. Maximum likelihood estimator of measurement error variances in data reconciliation IFAC. *ALPAC Symp*. 1992;2:135–139.
11. Keller JY, Zasadzinski M, Darouach M. Analytical estimator of measurement error variances in data reconciliation. *Comput Chem Eng*. 1992;16:185–188.
12. Chen J, Bandoni A, Romagnoli JA. Robust estimation of measurement error variance/covariance from process sampling data. *Comput Chem Eng*. 1996;21:593–600.
13. Morad K, Svrcek WY, McKay I. A robust direct approach for calculating measurement error covariance matrix. *Comput Chem Eng*. 1999;23:889–897.
14. Maquin D, Narasimhan S, Ragot J. Data validation with unknown variance matrix. *Comput Chem Eng*. 1999;23:S609–S612.
15. Mirabedini A, Hodouin D. Calculation of variance and covariance of sampling errors in complex mineral processing systems, using state-space dynamic models. *Int J Miner Process*. 1998;55:1–20.
16. Toussaint UV, Dose V. Bayesian inference in surface physics. *Appl Phys A*. 2006;82:403–413.
17. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal*. 2006;1:515–533.
18. Coleman MC, Block DE. Bayesian parameter estimation with informative priors for nonlinear systems. *AIChE J*. 2006;52:651–667.
19. Park ES, Oh MS, Guttorp P. Multivariate receptor models and model uncertainty. *Chemom Intell Lab Syst*. 2002;60:49–67.
20. Lin HI, Berzins DW, Myers L, George WT, Abdelghani A, Watanabe KH. A Bayesian approach to parameter estimation for a crayfish (Procambarus spp) bioaccumulation model. *Environ Toxicol Chem*. 2004;23:2259–2266.
21. Chen WS, Bakshi BR, Goel PK, Ungarala S. Bayesian estimation of unconstrained nonlinear dynamic systems via sequential Monte Carlo sampling. *Ind Eng Chem Res*. 2004;43:4012–4025.
22. Jitjareonchai JJ, Reilly PM, Duever TA, Chambers DB. Parameter estimation in the Error-in-Variables models using the Gibbs sampler. *Can J Chem Eng*. 2006;84:125–138.
23. Ungarala S, Chen ZZ, Li KY. Bayesian state estimation of nonlinear systems using approximate aggregate Markov chains. *Ind Eng Chem Res*. 2006;45:4208–4221.
24. Chen H, Bakshi BR, Goel PK. Bayesian latent variable regression via Gibbs sampling: methodology and practical aspects. *J Chemom.* 2007;21:578–591.
25. Gamerman D. *Markov Chain Monte Carlo*. New York: Chapman & Hall, 1997.

26. Chen H. Tutorial on Monte Carlo Sampling Technical Report. Department of Chemcal & Biomolecular Engineering, The Ohio State University, Columbus, OH, 2005.
27. Tierney L. Markov chain for exploring posterior distributions. *Ann Stat*. 1994;22:1701–1762.
28. Edgar TF, Himmelblau DM, Lasdon LS. *Optimization of Chemical Processes*, 2nd ed. New York: McGraw-Hill, 2001.
29. MacEachern SN, Berliner LM. Subsampling the Gibbs sampler. *Am Statistician*. 1994;48:188–190.
30. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR, Lunn D. BUGS: Bayesian inference using Gibbs sampling. Available at: http://www.mrc-bsu.cam.ac.uk/bugs/1994, 2003.
31. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput*. 2000;10:325–337.
32. Chen H. Sampling-Based Bayesian Latent Variable Regression Methods with Applications in Process Engineering, PhD Thesis, The Ohio State University, 2007.
33. Casella G, Berger RL. *Statistical Inference*, 2nd ed., North Scituate, MA: Duxbury Press, 2001.
34. Jeffreys H. *Theory of Probability*. New York: Oxford University Press, 1961.
35. Ingham J, Dunn IJ, Heinzle E, Prenosil JE. *Chemical Engineering Dynamics*, 2nd ed. New York: Wiley-VCH, 2000.
36. Binous H. Separation of a water-methanol mixture using a four stage batch distillation column. Available at: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=10169&objectType=file 2006.
37. Skogestad S, Morari M. Understanding the dynamic behavior of distillation columns. *Ind Eng Chem Res*. 1988;27:1848–1862.
38. Skogestad S, Postlethwaite I. *Multivariable Feedback Control*. New York: Wiley, 1996.
39. Skogestad S. MATLAB distillation column model ("Column A"). Available at: http://www.nt.ntnu.no/users/skoge/book/matlab_m/cola/cola.html 1997.
40. Schervish MJ. *Theory of Statistics*. New York: Springer, 1995.

Appendix

Proof of the Type 2 Prior is the Jeffreys Prior for Noise Variance in BLVR

Jeffreys proposed an noninformative prior which is proportional to the square root of the Fisher information,[40] that is,

$$P(\theta) \propto \sqrt{I(\theta|\mathbf{D})}, \tag{34}$$

where $P(\theta)$ is a prior distribution of a model parameter θ, $I(\theta \mid \mathbf{D})$ is the Fisher information. This is called Jeffreys prior.[34] The Fisher information is calculated as:

$$I(\theta|\mathbf{D}) = \mathbf{E}\left\{ \left[\frac{d \log P(\mathbf{D}|\theta)}{d\theta}\right]^2 |\theta \right\} \tag{35}$$

Because Fisher information is based only on the data likelihood, hence no extra information is used in Jeffreys prior. In addition, it is invariant under reparameterization of parameters.

Without loss of generality, let $j \in 1,2,\ldots,m$, the Fisher information of σ_j^2 is:

$$\begin{aligned} I(\sigma_j^2|\boldsymbol{X}(:,j)) &= \mathbf{E}\left[\frac{(\sum_{i=1}^{n}(x_{ij}-\tilde{x}_{ij})^2)^2}{4(\sigma_j^2)^4}\right] \\ &= \frac{1}{4(\sigma_j^2)^4}\mathbf{E}\left[\left(\sum_{i=1}^{n}(x_{ij}-\tilde{x}_{ij})^2\right)^2\right] \\ &= \frac{3n^2(\sigma_j^2)^2}{4(\sigma_j^2)^4} \\ &= \frac{3n^2}{4(\sigma_j^2)^2}. \end{aligned} \tag{36}$$

Thus, the Jeffreys prior is,

$$\begin{aligned} P(\sigma_j^2) &\propto \sqrt{\frac{3n^2}{4(\sigma_j^2)^2}} \\ &\propto \frac{1}{\sigma_j^2}, \end{aligned} \tag{37}$$

This is the same as in Eq. 28. Therefore, the Type 2 prior for σ_j^2 in Table 2 is the Jeffreys prior.

Manuscript received June 4, 2008, and revision received Mar. 18, 2009.