JOURNAL OF CHEMOMETRICS
J. Chemometrics 2002; **16**: 576–595
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.759

Bayesian principal component analysis

Mohamed N. Nounou[1], Bhavik R. Bakshi[1]*, Prem K. Goel[2] and Xiaotong Shen[2]

[1]Department of Chemical Engineering, The Ohio State University, Columbus, OH 43210, USA
[2]Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Received 2 February 2001; Accepted 12 July 2002

Principal component analysis (PCA) is a dimensionality reduction modeling technique that transforms a set of process variables by rotating their axes of representation. Maximum likelihood PCA (MLPCA) is an extension that accounts for different noise contributions in each variable. Neither PCA nor any of its extensions utilizes external information about the model or data, such as the range or distribution of the underlying measurements. Such prior information can be extracted from measured data and can be used to greatly enhance the model accuracy. This paper develops a Bayesian PCA (BPCA) modeling algorithm that improves the accuracy of estimating the parameters and measurements by incorporating prior knowledge about the data and model. The proposed approach integrates modeling and feature extraction by simultaneously solving parameter estimation and data reconciliation optimization problems. Methods for estimating the prior parameters from available data are discussed. Furthermore, BPCA reduces to PCA or MLPCA when a uniform prior is used. Several examples illustrate the benefits of BPCA versus existing methods even when the measurements violate the assumptions about their distribution. Copyright © 2002 John Wiley & Sons, Ltd.

KEYWORDS: Bayesian analysis; principal component analysis; filtering; latent variables

1. INTRODUCTION

Advances in computing and sensor technology allow the collection and storage of large amounts of measurements from many chemical processes and chemometric tasks. These measured data are a rich source of information, which when used effectively can greatly enhance the performance of these processes. The information embedded in data can be efficiently extracted by constructing accurate models that describe, summarize and predict the process behavior. Principal component analysis (PCA) is a popular modeling technique used to extract information from process data by relating the process variables. PCA has been found useful in many applications, such as process monitoring [1,2], data filtering [3], compression and regression. It transforms the process variables by rotating their axes of representation to capture the variation of the original variables in a lower-dimensional space. The new axes of rotation are represented by the projection directions or principal component loadings. This transformation can equivalently be obtained by minimizing the sum of square errors in all estimated variables. This equally weighted combination of variables means that PCA does not account for different noise contributions in different variables. Maximum likelihood PCA (MLPCA) was developed as a remedy to this drawback. MLPCA accounts for varying noise contributions by minimizing the sum of square errors of all variables normalized by their error covariance matrix. An iterative approach to solving for the MLPCA model was recently developed [4].

In practice, more information about the noise-free data or the PCA model is often available. Such information includes the range of variation and mean value of the principal component loadings and scores. Exploiting this information can enhance the accuracy of the estimated data and model. Unfortunately, neither PCA nor MLPCA accommodates such information, since both techniques consider the projection directions and principal components as fixed quantities to be estimated from the measured data. External information can be incorporated into the PCA modeling problem through a prior density function within a Bayesian framework, in which all quantities, measured and unmeasured, are considered random, having a probability density function that describes their behavior. In a Bayesian setting, the information brought by the data (quantified by the likelihood function) is combined with any external information (quantified by the prior) in a density function called the posterior. A sample is chosen from the posterior as the Bayesian estimate of the PCA model. Therefore PCA model estimation based on this combined knowledge is likely to be more accurate than modeling without the prior knowledge, unless the prior knowledge is totally inaccurate. Bayesian estimation also satisfies the likelihood principle, which states

*Correspondence to: B. R. Bakshi, Department of Chemical Engineering, The Ohio State University, Columbus, OH 43210, USA.
E-mail: bakshi.2@osu.edu
Contract/grant sponsor: National Science Foundation; Contract/grant number: CTS 9733627.

that all information brought by the data about the quantities of interest is included in the likelihood function. Thus, when the likelihood density used in a Bayesian approach is defined as in the MLPCA method, the Bayesian approach can account for noise in all variables and in varying contributions. These attractive advantages of Bayesian estimation [5,6] motivate our work.

Most efforts toward developing Bayesian dimensionality reduction models have been made by econometricians, with emphasis on factor analysis (FA). FA models are very common in the social sciences. They seek to explain the correlation among the original variables in terms of the extracted factors, such that the residual errors are uncorrelated. Contrary to PCA, which provides orthogonal principal components, the factors estimated in FA are not necessarily orthogonal. In addition, the factors computed under different assumptions about the model dimension can be totally different. A maximum likelihood solution to the FA modeling problem is presented by Seber [7]. An early formulation of Bayesian FA [8,9] uses a uniform prior for a subset of the FA model parameters and a zero-mean Gaussian prior for the remaining set. A Bayesian FA technique that avoids non-positive estimates of the data covariance matrix has also been developed [10]. Subsequently, a Bayesian approach that uses a normal prior for the model parameters, an inverted Wishart distribution for the noise covariance matrix and a vague constant prior for the factors has been presented [11]. The authors could obtain analytical large-sample estimates for the factor scores, the factor loading matrix and the noise covariance matrix. The robustness of this Bayesian FA model was later studied [12]. None of the above Bayesian FA formulations incorporates any information about the data or the transformed variables, since they all assume a uniform prior for the factors. Consequently, they do not improve the accuracy of the estimated data.

Dimensionality reduction techniques that improve the estimation of the underlying noise-free data not just by reducing the data dimension through a model that relates the variables, but also by filtering noise within each variable, have also been developed. Examples of such techniques include exponentially weighted moving PCA (EWMPCA) [13] and multiscale PCA (MSPCA) [14]. EWMPCA combines the advantages of PCA with those of EWMA filters to improve data filtering. The EWMPCA model is estimated by recursively forecasting the data using an exponentially weighted filter and updating the PCA model using the new measurements. MSPCA, on the other hand, combines the advantages of multiscale data filtering using wavelets with those of PCA filtering. In MSPCA the data are represented at multiple scales using wavelets, and a PCA model is constructed at each scale. Then the small wavelet coefficients are eliminated and the remaining coefficients are reconstructed back to the time domain. Finally a PCA model is derived using the reconstructed data. MSPCA simultaneously extracts the relationship across variables and across measurements. The advantages of MSPCA models are illustrated through applications to process monitoring. These approaches provide improved PCA models, but they neither account for varying noise contributions in different variables nor allow incorporation of external knowledge about the model.

In this paper a Bayesian principal component analysis (BPCA) modeling technique is developed to improve upon the accuracy of the estimated PCA model and measurements by incorporating external knowledge about these quantities through a prior density function. The approach integrates modeling and feature extraction in a statistically rigorous manner by simultaneously solving parameter estimation and data rectification problems. The BPCA approach is shown to be more general than PCA and MLPCA and to reduce to these methods when a uniform prior is used.

The rest of this paper is organized as follows. Section 2 introduces PCA and MLPCA. A brief description of Bayesian estimation is presented in Section 3. In Section 4 a general formulation of BPCA is presented and a BPCA algorithm is derived under some simplifying assumptions. This is followed by details about methods for estimating the prior and the number of retained principal components. Finally, in Section 5 the advantages of BPCA over existing methods are shown through illustrative examples.

2. PCA AND MLPCA

2.1. Principal component analysis

PCA represents a matrix of process variables as the product of two matrices, one containing the transformed variables (scores) and the other containing the new axes of rotation (loadings or projection directions). Given an $n \times r$ matrix of measured process variables, $\mathbf{X} = \widetilde{\mathbf{X}} + \boldsymbol{\varepsilon}_{\mathbf{x}}$, where $\widetilde{\mathbf{X}}$ is the matrix of underlying noise-free data, $\boldsymbol{\varepsilon}_{\mathbf{x}}$ is the additive noise matrix, r is the number of variables and n is the number of observations, PCA decomposes the matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\alpha}^{\mathrm{T}} \tag{1}$$

where $\mathbf{Z}$ is an $n \times r$ matrix of the principal components or principal component scores and α is an orthogonal $r \times r$ matrix of the loadings or projection directions. This transformation diagonalizes the data covariance matrix as

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \boldsymbol{\alpha}\mathbf{D}\boldsymbol{\alpha}^{\mathrm{T}} \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of the data covariance matrix. Substituting Equation (1) in Equation (2) gives

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = (\mathbf{Z}\boldsymbol{\alpha}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{Z}\boldsymbol{\alpha}^{\mathrm{T}}) = \boldsymbol{\alpha}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\alpha^{\mathrm{T}} = \boldsymbol{\alpha}\mathbf{D}\boldsymbol{\alpha}^{\mathrm{T}} \tag{3}$$

which indicates that the principal components are uncorrelated variables with variances equal to the eigenvalues of the data covariance matrix.

The PCA estimation problem for determining the first component can be formulated as the following optimization problem:

$$\{\hat{\boldsymbol{\alpha}}_1, \hat{\mathbf{Z}}_1\}_{\mathrm{PCA}} = \underset{\hat{\boldsymbol{\alpha}}_1, \hat{\mathbf{Z}}_1}{\operatorname{argmax}} \{\operatorname{var}(\mathbf{X}\hat{\boldsymbol{\alpha}}_1)\} \tag{4a}$$

$$\text{s.t. } \hat{\mathbf{Z}}_1 = \mathbf{X}\hat{\boldsymbol{\alpha}}_1 \text{ and } \hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_1 = 1 \tag{4b}$$

Other components may be found based on the residual error. The loadings maximize the variations captured by the principal components. The solution to this optimization

problem is found to be the singular value decomposition of the matrix $\mathbf{X}$, i.e.

$$\mathbf{X} = \mathbf{U}\mathbf{D}^{1/2}\boldsymbol{\alpha}^{\mathrm{T}} \tag{5}$$

where $\mathbf{U}$ is a unitary matrix containing the left eigenvectors, α is a unitary matrix containing the right eigenvectors and $\mathbf{Z} = \mathbf{U}\mathbf{D}^{1/2}$. The dimensionality of the data matrix can be reduced by retaining p principal components ($p < r$) with the largest eigenvalues that capture most of the variations in the data, assuming that the remaining principal components capture the contaminating noise.

The PCA estimation problem shown in Equation (4) can be equivalently formulated as the following optimization problem, in which the sum of estimation errors from all variables is minimized [15]:

$$\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{z}}_i\}_{\mathrm{PCA}} = \underset{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{z}}_i}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \tag{6a}$$

$$\text{s.t. } \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}}\hat{\mathbf{z}}_i \text{ and } \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\hat{\boldsymbol{\alpha}} = \mathbf{I} \tag{6b}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, which are $r \times 1$ vectors, are the ith measured and estimated observations respectively and the quantity $\hat{\mathbf{z}}_i$ is a $p \times 1$ vector of the estimated principal component at the observation $\mathbf{x}_i$. For notational purposes, these vectors are the transposed rows of the matrices $\mathbf{X}$ and $\mathbf{Z}$ respectively. The use of an identity-normalizing matrix in Equation (6a) shows that PCA implicitly assumes equal noise contribution in all variables. This assumption may not hold for many measured process data owing to the use of different sensors with different calibrations. In such cases the noise variation across the variables is interpreted as variation in the noise-free data, resulting in poor PCA models. One way to account for varying noise contributions in different variables is by using maximum likelihood principal component analysis (MLPCA) [4].

2.2. Maximum likelihood PCA

MLPCA estimates the model that maximizes the likelihood of estimating the true principal components and projection directions given the measured variables, or equivalently maximizing the probability density function of the measurements given the noise-free principal components, projection directions and the true rank of the data matrix, $\widetilde{p}$, as

$$\begin{aligned}\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}\}_{\mathrm{MLPCA}} &= \underset{\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}}{\operatorname{argmax}}\ L(\widetilde{\boldsymbol{\alpha}}, \widetilde{\mathbf{Z}}, \widetilde{p}; \mathbf{X}) \\ &= \underset{\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}}{\operatorname{argmax}}\ P(\mathbf{X}|\widetilde{\boldsymbol{\alpha}}, \widetilde{\mathbf{Z}}, \widetilde{p})\end{aligned} \tag{7}$$

subject to the constraint given in Equation (6b). If the distribution of the contaminating noise is assumed to be zero-mean Gaussian, i.e. $\boldsymbol{\varepsilon}_{\mathbf{x}} \sim N(\mathbf{0}, \mathbf{Q}_{\varepsilon_{\mathbf{x}}})$, maximizing this likelihood function is equivalent to minimizing the sum of square errors normalized by the noise covariance matrix. Since the noise-free model and data are not available, the minimization is performed with respect to the estimated data and thus the MLPCA solution is obtained by solving the following optimization problem:

$$\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}\}_{\mathrm{MLPCA}} = \underset{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\varepsilon_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \tag{8}$$

where $\mathbf{Q}_{\varepsilon_{\mathbf{x}}}$ is the noise covariance matrix, which is assumed to be known, and subject to the constraints given in Equation (6b). This minimization problem requires an iterative procedure to solve for the MLPCA model. One such algorithm [4] alternates between minimizing the objective function in the row and column spaces of the data matrix. In Equation (8) the noise distribution is assumed to be fixed, which means that all noise observations are assumed to have the same mean and covariance matrices. A more general MLPCA approach that accounts for correlated noise observations with possibly different variances has also been developed [4].

Alternatively, the MLPCA model can also be obtained by solving two simultaneous optimization problems: one solves for the principal component loadings or projection directions (a parameter estimation problem) and the other solves for the principal component scores (a data reconciliation problem) as

$$\begin{aligned}\{\hat{\boldsymbol{\alpha}}\}_{\mathrm{MLPCA}} &= \underset{\hat{\boldsymbol{\alpha}}}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\varepsilon_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \\ \text{s.t. } \{\hat{\mathbf{z}}_i\}_{\mathrm{MLPCA}} &= \underset{\hat{\mathbf{z}}_i}{\operatorname{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\varepsilon_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i)\end{aligned} \tag{9}$$

subject to the constraints given in Equation (6b). The data reconciliation problem (the inner minimization problem) has been studied extensively [4,16] and has the following closed form solution, as shown in Appendix I:

$$\{\hat{\mathbf{z}}_i\}_{\mathrm{MLPCA}} = (\hat{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{Q}_{\varepsilon_{\mathbf{x}}}^{-1} \hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\alpha}}^{\mathrm{T}} \mathbf{Q}_{\varepsilon_{\mathbf{x}}}^{-1} \mathbf{x}_i \tag{10}$$

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, having a joint probability density function that describes their behavior [17,18]. This is a different perspective from that adopted by most 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. This assumption of Bayesian methods permits incorporation of external prior knowledge about the quantities of interest into the estimation problem. To estimate the quantity $\widetilde{\theta}$ 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, $P(\widetilde{\theta}|y)$, which is called the posterior. The *posterior* is a density function that describes the behavior of the quantity $\widetilde{\theta}$ *after* observing the measurements. Using Bayes rule, the posterior can be written as

$$P(\widetilde{\theta}|y) = \frac{P(y|\widetilde{\theta})P(\widetilde{\theta})}{P(y)} \tag{11}$$

The first term in the numerator of Equation (11) 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 the 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. Thus the posterior density can be written as

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

or

$$\text{posterior} \propto \text{likelihood} \times \text{prior} \tag{12}$$

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 $\tilde{\theta}$. Contrary 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 of Bayesian estimation can be outlined as follows [18].

1. Set up a full probability model (a joint probability density function) of all observable and unobservable quantities. This is possible based on the assumption that all variables are random.
2. Calculate the conditional density of the variables to be estimated given the observed data (posterior).
3. Evaluate the implication of the posterior and check the accuracy of the estimated quantities.

The second step is a mathematical one which involves computing the posterior density function. When the likelihood and 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) [19]. The third step is more judgemental since it requires a decision about the sample to be selected from 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 $L(\tilde{\theta};\hat{\theta})$ corresponds to a utility function that decides which sample from the posterior is to be selected as the Bayesian estimate. Here $\hat{\theta}$ and $\tilde{\theta}$ denote the Bayesian estimate and true value of the quantity θ respectively. Many loss functions have been suggested, such as quadratic and zero–one loss functions [20]. 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.

$$L(\hat{\theta};\tilde{\theta}) = \begin{cases} 0 & \text{when } \{\hat{\theta}\}_{\text{Bayesian}} = \tilde{\theta} \\ 1 & \text{otherwise} \end{cases} \tag{13}$$

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

$$\{\hat{\theta}\}_{\text{MAP}} = \underset{\tilde{\theta}}{\text{argmax}}\ P(y|\tilde{\theta})P(\tilde{\theta}) \tag{14}$$

The BPCA algorithm developed in this paper uses the zero–one loss function. One advantage of using this loss function is that it reduces the Bayesian PCA modeling to a minimization problem, which permits comparison between BPCA and other existing methods. Furthermore, a zero–one loss function is often more computationally efficient, as the Bayesian estimate of the data has a closed form solution.

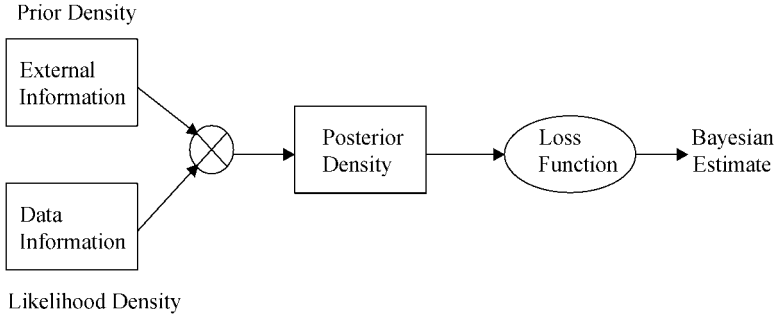

Figure 1. Schematic diagram of main steps in Bayesian estimation.

4. BAYESIAN PRINCIPAL COMPONENT ANALYSIS (BPCA)

4.1. General formulation

Defining the PCA model from a data matrix requires estimating the projection directions, principal components and true model rank (or number of retained principal components). Therefore, within a Bayesian framework, the posterior should be defined as the conditional density of these quantities given the measured data. This can be written using Bayes' rule as

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\mathbf{X}) = \frac{P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}, \boldsymbol{\alpha}, \widetilde{p})}{P(\mathbf{X})} \tag{15}$$

The first term in the numerator is the likelihood function, which is the conditional density of the measured variables given the noise-free PCA model and data, while the second term is the prior. The unnormalized posterior can be written as

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\mathbf{X}) \propto P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}) \tag{16}$$

4.1.1. *The prior density function*

The prior is the joint density of the noise-free principal components, projection directions and true PCA model rank and is a very complicated function. However, the density function of the principal components and projection directions depends on the model rank. Thus the prior can be written as

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}) = P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}|\widetilde{p})P(\widetilde{p}) \tag{17}$$

Note that $P(\widetilde{p})$ is a discrete density function which can be defined as

$$P(\widetilde{p} = j) = k_j \quad \text{such that} \quad \sum_{j=1}^{r} k_j = \mathbf{1} \tag{18}$$

Furthermore, the joint density function of the principal components and projection directions can be expressed using the multiplication rule of probabilities as

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}|\widetilde{p}) = P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\boldsymbol{\alpha}}|\widetilde{p}) \tag{19}$$

Thus the unnormalized posterior can be written as

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\mathbf{X}) \propto P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\boldsymbol{\alpha}}|\widetilde{p})P(\widetilde{p}) \tag{20}$$

4.2. Simplifying assumptions

Computing the posterior density shown in Equation (20) requires defining the prior and likelihood densities, which depend on the nature of the noise-free data and the contaminating noise. Therefore assumptions about the data need to be made in order to define the structures of these densities. In this subsection the assumptions and their implications are described.

4.2.1. *Known true model rank*

Most applications of PCA and MLPCA determine the model rank before developing the model. The BPCA method also assumes that the model rank $\widetilde{p}$ is known. As shown in Section 5, the impact of this assumption is less severe for BPCA than for PCA or MLPCA. Under this assumption the rank portion of the prior density becomes

$$P(\widetilde{p}) = \mathbf{1} \tag{21}$$

reducing the prior to

$$P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\boldsymbol{\alpha}}|\widetilde{p})P(\widetilde{p}) = P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})P(\widetilde{\boldsymbol{\alpha}}) \tag{22}$$

and simplifying the posterior to

$$P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})P(\widetilde{\boldsymbol{\alpha}}) \tag{23}$$

In practice, however, the true rank of the PCA model is unknown and needs to be estimated. A technique for estimating the model rank is presented in Section 4.5.

4.2.2. *Loss function*

In this work a zero–one loss function of the form

$$L(\hat{\mathbf{Z}}, \hat{\boldsymbol{\alpha}}; \widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}) = \begin{cases} 0 & \text{when } \{\hat{\mathbf{Z}}, \hat{\boldsymbol{\alpha}}\}_{\text{Bayesian}} = \widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}} \\ 1 & \text{otherwise} \end{cases} \tag{24}$$

is used. Consequently, the BPCA solution can be obtained by solving the following optimization problem:

$$\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}\}_{\text{Bayesian}} = \underset{\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}}{\text{argmax}}\ P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})P(\widetilde{\boldsymbol{\alpha}}) \tag{25}$$

Such a formulation results in a closed form solution for the estimated data, which is computationally very efficient, and allows direct comparison with existing methods such as PCA and MLPCA.

4.2.3. *The likelihood density function*

The structure of the likelihood function depends on the nature of the noise. If the measured process variables are assumed to be contaminated with zero-mean additive Gaussian noise, i.e. $\mathbf{X} = \widetilde{\mathbf{X}} + \boldsymbol{\varepsilon}_{\mathbf{x}}$, where $\boldsymbol{\varepsilon}_{\mathbf{x}} \sim N(\mathbf{0}, \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}})$, then the likelihood function will also be normal, with the following moments:

$$E[\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}] = E[\widetilde{\mathbf{X}} + \boldsymbol{\varepsilon}_{\mathbf{x}}] = \widetilde{\mathbf{X}} \tag{26}$$

$$\text{cov}[\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}] = E[(\mathbf{X} - \widetilde{\mathbf{X}})^{\mathrm{T}}(\mathbf{X} - \widetilde{\mathbf{X}})] = E[(\boldsymbol{\varepsilon}_{\mathbf{x}})^{\mathrm{T}}(\boldsymbol{\varepsilon}_{\mathbf{x}})] = \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}} \tag{27}$$

These moments are assumed to be known. Therefore

$$P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}) \sim N(\widetilde{\mathbf{X}}, \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}) \tag{28}$$

Note that this is the same density function as used in MLPCA.

4.2.4. *Multivariate Gaussian noise-free data*

The structure of the densities $P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})$ and $P(\widetilde{\boldsymbol{\alpha}})$ depends on the nature of the noise-free variables. In general, the density $P(\widetilde{\boldsymbol{\alpha}})$ is a complicated function, and most attempts made toward deriving its structure have relied on the assumption that the underlying noise-free data follow a multivariate normal distribution. Even under this normality assumption and for distinct eigenvalues, only asymptotic results have been obtained [21]. In this work we will also assume that the noise-free data follow a Gaussian distribution. As the illustrative example in this subsection indicates, this assumption seems to be reasonable, as the distributions of

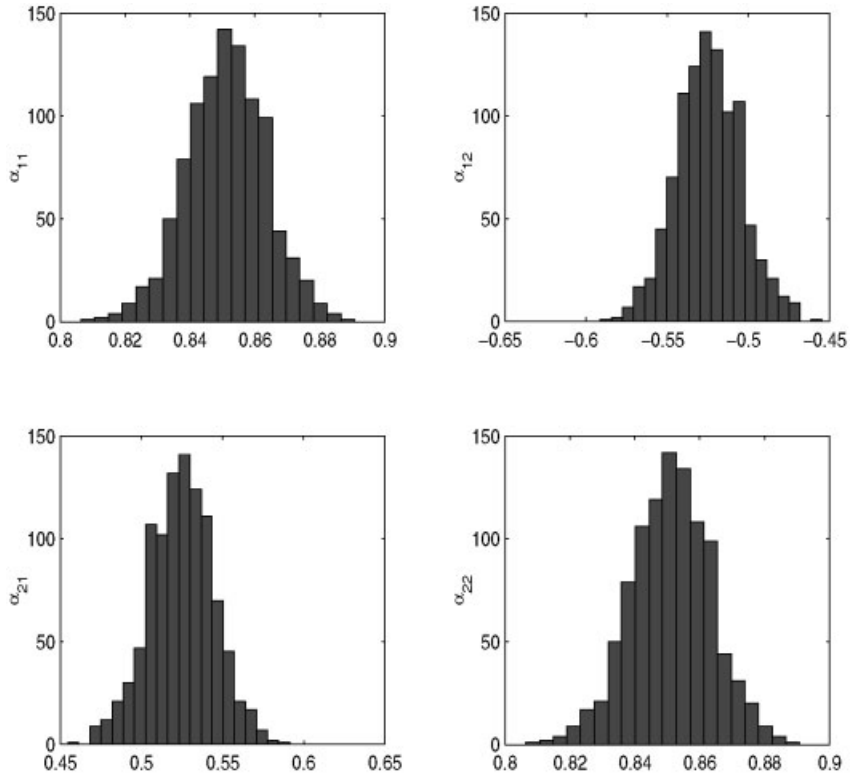

Figure 2. Histograms of elements of projection directions for Gaussian data. The Gaussian distribution confirms Girshick's results [21].

many types of data which do not follow a Gaussian distribution can still be reasonably approximated by a Gaussian density. Therefore each noise-free observation in the data matrix is assumed to be a sample from a multivariate normal distribution, i.e.

$$\tilde{\mathbf{x}}_i = [\tilde{x}_{i1} \quad \dots \quad \tilde{x}_{ir}]^{\mathrm{T}} \sim MVN(\boldsymbol{\mu}_{\tilde{\mathbf{x}}}, \mathbf{Q}_{\tilde{\mathbf{x}}}), \quad i = 1, \dots, n \tag{29}$$

It has been shown [21] that under this normality assumption, and if the eigenvalues of the covariance matrix of the noise-free data are distinct, the eigenvalues and eigenvectors of the sample covariance matrix are asymptotically multivariate normal and the eigenvalues are independent of the eigenvectors. The following asymptotic moments of each projection direction $\tilde{\boldsymbol{\alpha}}_j$ have also been presented:

$$E[\tilde{\boldsymbol{\alpha}}_j] = \boldsymbol{\alpha}_j + O(n^{-1}) \tag{30}$$

$$\mathrm{cov}[\tilde{\boldsymbol{\alpha}}_j] = \frac{1}{n} \sum_{j \neq k} \frac{\lambda_j \lambda_k}{(\lambda_j - \lambda_k)^2} \boldsymbol{\alpha}_j \boldsymbol{\alpha}_k^{\mathrm{T}} + O(n^{-2}) \tag{31}$$

where the λs and αs are the eigenvalues and eigenvectors respectively of the matrix

$$E[\tilde{\mathbf{X}}^{\mathrm{T}} \tilde{\mathbf{X}}] = \mathbf{Q}_{\tilde{\mathbf{x}}} + \boldsymbol{\mu}_{\tilde{\mathbf{x}}} \boldsymbol{\mu}_{\tilde{\mathbf{x}}}^{\mathrm{T}} \tag{32}$$

4.2.5. *Illustrative example*

To illustrate Girshick's results [21], consider the projection direction matrix α of a data matrix $\tilde{\mathbf{X}}$ having two variables and 1000 observations, in which each observation is a sample from the Gaussian distribution $N(\boldsymbol{\mu}_{\tilde{\mathbf{x}}}, \mathbf{Q}_{\tilde{\mathbf{x}}})$, where $\boldsymbol{\mu}_{\tilde{\mathbf{x}}} = [2 \quad 1]^{\mathrm{T}}$ and $\mathbf{Q}_{\tilde{\mathbf{x}}} = \mathrm{diag}(1 \quad 2)$. In this example the α matrix is of size 2×2 and can be written as

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \tag{33}$$

To investigate the distribution of the projection directions, a Monte Carlo simulation is performed with 1000 realizations. In each realization a matrix $\tilde{\mathbf{X}}$ containing 1000 samples drawn from the normal distribution described above is generated and the projection directions are computed. Then histograms for all elements α_{ij} of the matrix α are produced, as shown in Figure 2. This figure illustrates that the distribution of α_{ij} is close to normal. The accuracy of Girshick's estimator of the means and variances of the elements of the projection direction matrix is shown in Table I, which compares the means and variances of the elements α_{ij} obtained by simulation and by Girshick's theorem. Since

Table I. Comparison of means and variances of elements of projection direction matrix obtained using a Monte Carlo simulation and Girshick's theorem

	Mean		Variance ($\times 10^{-4}$)	
Element	Girshick	Monte Carlo	Girshick	Monte Carlo
α_{11}	0.8507	0.8514	1.52	1.52
α_{21}	0.5257	0.5240	3.98	4.01
α_{12}	−0.5257	−0.5240	3.98	4.01
α_{22}	0.8507	0.8514	1.52	1.52

Girshick's results are asymptotic, the distribution of the projection directions tends toward normal as the number of observations increases.

Thus, from Girshick's results, it follows that if we define the vector $\widetilde{\mathbf{a}} \equiv \begin{bmatrix} \widetilde{\boldsymbol{\alpha}}_1^{\mathrm{T}} & \widetilde{\boldsymbol{\alpha}}_2^{\mathrm{T}} & \dots & \widetilde{\boldsymbol{\alpha}}_p^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ of size $rp \times 1$, where p is the number of retained projection directions, then the vector $\widetilde{\mathbf{a}}$ will asymptotically follow a multivariate normal distribution, i.e. $\widetilde{\mathbf{a}} \sim MVN(\boldsymbol{\mu}_{\widetilde{\mathbf{a}}}, \mathbf{Q}_{\widetilde{\mathbf{a}}})$. The density $P(\widetilde{\mathbf{a}})$ is degenerate, since some elements in the vector $\widetilde{\mathbf{a}}$ are dependent on others owing to the orthogonality constraint imposed on the projection direction matrix $\widetilde{\boldsymbol{\alpha}}$.

For the density $P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})$, on the other hand, since the noise-free principal components, process variables and projection directions are linearly related as $\widetilde{\mathbf{Z}} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\alpha}}$, and since $\widetilde{\mathbf{X}}$ follows a multivariate normal distribution, then the density of the noise-free principal components given the projection directions is also multivariate normal, with the following moments:

$$E[\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}] = E[\widetilde{\mathbf{X}}]\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\mu}_{\widetilde{\mathbf{x}}}\widetilde{\boldsymbol{\alpha}} \tag{34}$$

$$\mathrm{Cov}[\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}] = \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}} E[(\widetilde{\mathbf{X}} - \boldsymbol{\mu}_{\widetilde{\mathbf{x}}})^{\mathrm{T}}(\widetilde{\mathbf{X}} - \boldsymbol{\mu}_{\widetilde{\mathbf{x}}})]\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\widetilde{\mathbf{x}}}\widetilde{\boldsymbol{\alpha}} \tag{35}$$

Therefore

$$P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}) \sim MVN(\boldsymbol{\mu}_{\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}}, \mathbf{Q}_{\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}}) = MVN(\boldsymbol{\mu}_{\widetilde{\mathbf{x}}}\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\widetilde{\mathbf{x}}}\widetilde{\boldsymbol{\alpha}}) \tag{36}$$

4.3. The BPCA algorithm

The MAP solution of the BPCA problem can be obtained by solving Equation (25), which is equivalent to solving the following simultaneous parameter estimation and data reconciliation problems similar to those solved in MLPCA:

$$\begin{aligned} \{\hat{\boldsymbol{\alpha}}\}_{\mathrm{MAP}} &= \underset{\widetilde{\boldsymbol{\alpha}}}{\mathrm{argmax}}\ P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{a}}) \\ \text{s.t.} \{\widetilde{\mathbf{Z}}\}_{\mathrm{MAP}} &= \underset{\widetilde{\mathbf{Z}}}{\mathrm{argmax}}\ P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}) \\ \widetilde{\mathbf{X}} &= \widetilde{\boldsymbol{\alpha}}\widetilde{\mathbf{Z}} \text{ and } \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}} = \mathbf{I} \end{aligned} \tag{37}$$

Based on the simplifying assumptions made in Section 4.2, all densities in the posterior are defined as multivariate normal. Thus the MAP solution can be equivalently obtained by solving the following simultaneous minimization problems for the projection directions and the reconciled data:

$$\begin{aligned} &\hat{\boldsymbol{\alpha}}_{\mathrm{MAP}} = \\ &\underset{\hat{\boldsymbol{\alpha}}}{\mathrm{argmin}} \left\{ \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{a}} - \boldsymbol{\mu}_{\hat{\mathbf{a}}})^{\mathrm{T}} \mathbf{Q}_{\hat{\mathbf{a}}}^{-1} (\hat{\mathbf{a}} - \boldsymbol{\mu}_{\hat{\mathbf{a}}}) \right\} \\ &\text{s.t.} \quad \{\hat{\mathbf{z}}_i\}_{\mathrm{MAP}} = \underset{\hat{\mathbf{z}}_i}{\mathrm{argmin}} \left\{ \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \right. \\ &\left. (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}})^{\mathrm{T}} \mathbf{Q}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}}) \right\} \widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}} = \mathbf{I} \text{ and } \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}}\hat{\mathbf{z}}_i \end{aligned} \tag{38}$$

The data reconciliation problem has the following closed form solution, as shown in Appendix II:

$$\{\hat{\mathbf{z}}_i\}_{\mathrm{MAP}} = (\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}^{-1}\hat{\boldsymbol{\alpha}} + \mathbf{Q}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}}^{-1})^{-1}(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}^{-1}\mathbf{x}_i + \mathbf{Q}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}}^{-1}\boldsymbol{\mu}_{\hat{\mathbf{z}}|\hat{\boldsymbol{\alpha}}}) \tag{39}$$

This BPCA algorithm reduces to MLPCA if the prior terms are set to zero, i.e. when the prior is uniform. Owing to the degeneracy of the distribution of $\widetilde{\mathbf{a}}$, the covariance matrix $\mathbf{Q}_{\widetilde{\mathbf{a}}}$ is singular. One way to approximate its inverse is by neglecting the off-diagonal elements, which represent the cross relationship between the elements of the projection directions. This assumption is not bad, since the relationship is already captured by the orthogonality constraint $\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}} = \mathbf{I}$ imposed on the minimization problem. Alternatively, the generalized inverse in the reduced space may be used. The number of independent elements in the projection direction matrix can be calculated as follows:

$$\begin{aligned} &\text{number of independent elements} = r \times p \\ &\quad - \text{number of normality constraints} \\ &\quad - \text{number of orthogonality constraints} \\ &= r \times p - p + \binom{p}{2} = r \times p - \frac{p(p+1)}{2} \end{aligned} \tag{40}$$

However, the problem with this alternative is that the effective rank of the matrix $\mathbf{Q}_{\widetilde{\mathbf{a}}}$ is usually much less than the number of independent elements computed using Equation (40) owing to the non-linearity of the orthonormality constraints $\widetilde{\boldsymbol{\alpha}}^{\mathrm{T}}\widetilde{\boldsymbol{\alpha}} = \mathbf{I}$. Thus reducing the dimension of the matrix $\mathbf{Q}_{\widetilde{\mathbf{a}}}$ to the number of its independent elements does not guarantee its inversion.

4.4. Estimating the prior density

In the BPCA algorithm described in Section 4.3, the structures of the densities $P(\widetilde{\mathbf{a}})$ and $P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}})$ were assumed to be multivariate normal and the parameters $\boldsymbol{\mu}_{\widetilde{\mathbf{x}}}$, $\mathbf{Q}_{\widetilde{\mathbf{x}}}$, $\boldsymbol{\mu}_{\widetilde{\mathbf{a}}}$ and $\mathbf{Q}_{\widetilde{\mathbf{a}}}$ (which are called the prior hyperparameters) were also assumed to be known. In other words, the entire prior density was assumed to be defined *a priori*. Such a fully predefined prior density is commonly used in Bayesian analysis. In practice, however, parts of or the entire prior distribution might be unspecified, for which the observed data are usually used in their estimation. Such an approach is called empirical Bayesian (EB) analysis [18,22].

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, then the data are used to estimate its hyperparameters. In the non-parametric approach, on the other hand, the entire prior distribution is estimated from the data, which is usually a much more challenging and computationally more demanding task than the parametric approach [23]. For BPCA the parametric approach will be used, since under the simplifying assumption described earlier the structures of all parts of the prior distribution are known, and owing to the computational burden expected in the non-parametric approach. Empirical estimation of the prior from a parametric point of view simply corresponds to estimating its hyperparameters. Denoting the set of hyperparameters, $\{\boldsymbol{\mu}_{\widetilde{\mathbf{x}}}, \mathbf{Q}_{\widetilde{\mathbf{x}}}, \boldsymbol{\mu}_{\widetilde{\mathbf{a}}}, \mathbf{Q}_{\widetilde{\mathbf{x}}}\}$, by η, the posterior for this EBPCA problem becomes

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\mathbf{X}, \boldsymbol{\eta}) = \frac{P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\boldsymbol{\eta})}{P(\mathbf{X})} \tag{41}$$

Now the prior is dependent on the set of hyperparameters, η,

which is unknown. When these hyperparameters are actually known, η drops from the preceding expression, as there is no need to express conditioning on a constant, and Equation (41) reduces to the posterior density shown in Equation (15). The basic idea here is to estimate the set of hyperparameters. η, from the data using maximum likelihood estimation and then use the empirically estimated prior to solve for the BPCA model. Therefore the EBPCA problem is solved in three steps.

1. Solve for the MLPCA model using the available data.
2. Use the MLPCA solution to estimate the set of hyperparameters, $\hat{\boldsymbol{\eta}}$, as follows.
   (a) Set $\hat{\boldsymbol{\mu}}_{\tilde{\mathbf{a}}} = \{\hat{\mathbf{a}}\}_{\text{MLPCA}}$.
   (b) Solve for $\hat{\mathbf{Q}}_{\tilde{\mathbf{a}}}$ using Equation (31).
   (c) Estimate $\hat{\boldsymbol{\mu}}_{\tilde{\mathbf{x}}}$ as $E[\{\hat{\mathbf{X}}\}_{\text{MLPCA}}]$.
   (d) Estimate $\hat{\mathbf{Q}}_{\tilde{\mathbf{x}}}$ as $\text{cov}[\{\hat{\mathbf{X}}\}_{\text{MLPCA}}]$.

   Now the prior is defined in terms of the set $\hat{\boldsymbol{\eta}}$ as $P(\widetilde{\mathbf{Z}},\ \widetilde{\boldsymbol{\alpha}},\ \widetilde{p}|\hat{\boldsymbol{\eta}})$.
3. Solve the BPCA problem using the following posterior:

$$P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\mathbf{X}, \hat{\boldsymbol{\eta}}) = \frac{P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})P(\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p}|\hat{\boldsymbol{\eta}})}{P(\mathbf{X})} \tag{42}$$

This process of empirically estimating the prior can be repeated several times by using the Bayesian estimate of the PCA model to recalculate the prior, which is then used in the next Bayesian estimate. Such an iterative approach may improve the BPCA estimate. However, the solution may diverge for too many iterations. The examples in this paper estimate the prior from the MLPCA solution without iterations.

4.5. Estimating the PCA model rank

One of the challenges in applying PCA to practical problems is determining the number of retained principal components or the PCA model rank. This is a model selection problem, for which many techniques have been developed. Some of these approaches are heuristic and rely on the relative magnitude of the eigenvalues to estimate the number of retained principal components [24,25]. Other approaches rely on cross-validation [26,27] or on modifications of the likelihood function [25]. As shown through illustrative examples in Section 5, the likelihood function increases by retaining more principal components. Consequently, maximizing the likelihood to infer the model dimension always yields the largest model possible. Therefore some techniques [25] seek to modify the likelihood function by penalizing high-dimensional models. However, the accuracy of these techniques depends on the penalty used and the nature of the problem.

An intuitive but incorrect approach for estimating the model rank is selecting the BPCA model that maximizes the posterior. For numerical purposes the BPCA problem may be expressed in terms of the posterior natural logarithm as

$$\begin{aligned}\{\hat{\mathbf{Z}}, \hat{\boldsymbol{\alpha}}, \hat{p}\}_{\text{Bayesian}} = \underset{\tilde{\mathbf{Z}},\tilde{\alpha},\tilde{p}}{\text{argmax}}\ \{&\ln(P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{p})) \\ &+ \ln(P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}, \widetilde{p})) + \ln(P(\widetilde{\boldsymbol{\alpha}}|\widetilde{p})) + \ln(P(\widetilde{p}))\}\end{aligned} \tag{43}$$

Since $P(\widetilde{p} = j) = k_j$ (see Equation (18)), then the posterior natural logarithm at a particular model rank j, denoted by LogPost_j, can be written as

$$\begin{aligned}\text{LogPost}_j = &\ln(P(\mathbf{X}|\widetilde{\mathbf{Z}}, \widetilde{\alpha}, \widetilde{p} = j)) \\ &+ \ln(P(\widetilde{\mathbf{Z}}|\widetilde{\boldsymbol{\alpha}}, \widetilde{p} = j)) + \ln(P(\widetilde{\boldsymbol{\alpha}}|\widetilde{p} = j)) + \ln(k_j)\end{aligned} \tag{44}$$

Then the MAP estimate of the BPCA model rank can be determined by selecting the BPCA model that maximizes the natural logarithm of the posterior function evaluated at all ranks, i.e.

$$\begin{aligned}\{\hat{p}\}_{\text{Bayesian}} &= \underset{\tilde{p}}{\text{argmax}}\ \{\text{LogPost}_1, \text{LogPost}_2, \ldots, \text{LogPost}_r\}\end{aligned} \tag{45}$$

where LogPost_j is given in Equation (44). Any external information about the model rank can be incorporated through the last term of Equation (44), which becomes a constant when no prior preference is given to any specific model.

When no preference is given to any particular rank, and using empirical priors, maximizing the posterior is shown through a simulated example in the next section to work only at moderate noise contents; that is, at very low and very high signal-to-noise ratios the MAP estimator of the number of retained principal components is shown to be ineffective without incorporating external information about the true model rank. The reason behind this poor performance of the MAP estimator of the model rank is that it is meaningless to compare values of the posterior density at different model dimensions, as they quantify totally different models. More details about this MAP estimator of the model rank are presented later through a simulated example.

In this work a hypothesis testing approach is used to approximate the dimensions of the MLPCA model [4]. This approach is based on the fact that the sum of square approximation errors obtained in MLPCA, i.e.

$$S = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{x}}}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \tag{46}$$

should follow a χ^2 distribution with the appropriate degree of freedom, $(r-p)(n-p)$, when the true model rank is used. Therefore, if Pr is the probability of realizing a value of S below the observed one using MLPCA, then for a confidence of $(1-\alpha)\times 100\%$ a Pr value higher than $1-0.5\alpha$ would reject the null hypothesis that the model is correct. It has been illustrated that when the correct model dimension is reached, a significant drop in the value of Pr is observed [4]. Note that this approach is only valid when the noise covariance matrix is known, which is assumed in this paper.

This hypothesis testing approach can be summarized as follows.

1. For each possible model dimension j, solve for the MLPCA model and compute the corresponding value S_j.
2. For each value S_j, compute the probability Pr_j of realizing a lower value than S_j using a χ^2 distribution with the appropriate degree of freedom, $(n-j)(r-j)$.
3. Select the smallest model dimension at which the value Pr_j drops below the confidence limit, which for a $(1-\alpha)\times 100\%$ confidence equals $1-0.5\alpha$.

5. ILLUSTRATIVE EXAMPLES

A variety of examples are presented in this section to illustrate and compare the performance of the Bayesian PCA technique with that of PCA and MLPCA. The accuracy of estimated data is determined by computing the mean square errors between the estimated data and the noise-free data for the various techniques. This is possible since the noise-free data are known in the synthetic examples. The accuracy of the estimated loadings or projection directions, on the other hand, can be determined by computing the mean square errors of the estimated regression parameters of the last $r-p$ variables on the first p variables. For example, a data matrix with three variables and a rank of 2 can be written in terms of the two independent variables as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{47}$$

where the regression parameters a_1 and a_2 relate the last variable to the first two variables.

The regression parameters relating the last $r-p$ variables to the first p variables can be computed using the estimated projection direction matrix as follows:

$$\begin{bmatrix} \mathbf{I}_p \\ \hat{\mathbf{a}} \end{bmatrix} = \hat{\boldsymbol{\alpha}} \underbrace{(\hat{\boldsymbol{\alpha}}_{\mathrm{U}})^{-1}}_{p \times p}, \quad \text{where } \hat{\boldsymbol{\alpha}} = \underbrace{\begin{bmatrix} \hat{\boldsymbol{\alpha}}_{\mathrm{U}} \\ \hat{\boldsymbol{\alpha}}_{\mathrm{L}} \end{bmatrix}}_{r \times p} \tag{48}$$

Inverting the upper part of the projection direction matrix may not always be possible. In such cases the generalized inverse may be used. Another criterion for comparing the model accuracy is by computing the angular deviation between each noise-free projection direction and the subspace spanned by the estimated projection directions. This metric can be computed as [4]

$$\gamma_j = \cos^{-1}\left(\frac{\tilde{\boldsymbol{\alpha}}_j^{\mathrm{T}} \hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^{\mathrm{T}} \tilde{\boldsymbol{\alpha}}_j}{\|\tilde{\boldsymbol{\alpha}}_j^{\mathrm{T}}\| \cdot \|\hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}^{\mathrm{T}} \tilde{\boldsymbol{\alpha}}_j\|} \right) \tag{49}$$

5.1. Example 1. Stationary Gaussian data contaminated by white noise

The data matrix considered in this example consists of three variables and 50 observations. The first two noise-free variables are independent and are drawn from the Gaussian distributions

$$\tilde{x}_1 \sim N(3,1) \quad \text{and} \quad \tilde{x}_2 \sim N(1,4) \tag{50}$$

where the variances of the two variables are 1 and 4 respectively and the third variable is a linear combination of the first two, i.e.

$$\tilde{x}_3 = a_1 \tilde{x}_1 + a_2 \tilde{x}_2, \quad \text{where} \quad a_1 = a_2 = 1 \tag{51}$$

Therefore the rank of the noise-free data matrix is 2, which is assumed to be known. The noise-free data are then contaminated with additive zero-mean white noise with the covariance matrix

$$\mathbf{Q}_{\varepsilon_{\tilde{\mathbf{x}}}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix} \tag{52}$$

which is also assumed to be known. The signal-to-noise ratio is unity for all variables.

The performance of BPCA is studied and compared with that of PCA and MLPCA using different priors. Case I uses a perfect prior, i.e.

$$\mathbf{P}(\tilde{\mathbf{Z}}|\tilde{\boldsymbol{\alpha}}) \sim MVN(\boldsymbol{\mu}_{\tilde{\mathbf{x}}}\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\tilde{\mathbf{x}}}\tilde{\boldsymbol{\alpha}}) \tag{53}$$

where

$$\boldsymbol{\mu}_{\tilde{\mathbf{x}}} = [3 \quad 1 \quad 4]^{\mathrm{T}} \text{ and } \mathbf{Q}_{\tilde{\mathbf{x}}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 4 \\ 1 & 4 & 5 \end{bmatrix} \tag{54}$$

and the prior of the eigenvectors is computed using Equations (30) and (31), assuming that the number of observations is 500. This case represents the best case scenario for the performance of BPCA. Case II determines the prior empirically from 500 external noisy observations available from historical data. Case III represents empirical BPCA (EBPCA), since the prior is computed empirically from the same 50 noisy observations used in modeling. This case represents the worst case scenario for using BPCA, since no external information about the noise-free PCA models or data is used.

The performance of various techniques is compared via a Monte Carlo simulation of 100 realizations. The results in Table II show that BPCA outperforms PCA and MLPCA. The results of Case I show that with perfect prior knowledge highly accurate results can be obtained. Although such a perfect prior is usually not available in practice, the results of Case I indicate the extent of possible improvement by BPCA. The results of Case II show that significantly better performance may be obtained by utilizing the information in historical data. Finally, the results of EBPCA in Case III show that even with no external information an empirically estimated prior can still provide an improvement over PCA and MLPCA. This case does not show any improvement

Table II. PCA modeling and rectification of stationary Gaussian noise-free data (Example 1): Case I, perfect prior; Case II, estimated using 500 external observations; Case III, empirical prior

	PCA	MLPCA	BPCA (Case I)	BPCA (Case II)	EBPCA (Case III)
Prior	Uniform	Uniform	Perfect	From historical data	From data being modeled
MSE(x_1)	1.546	0.902	0.475	0.514	0.537
MSE(x_2)	3.520	2.615	1.498	1.623	1.715
MSE(x_3)	3.087	2.816	1.694	1.741	1.948
$\gamma_1 \pm \sigma$	2.9 ± 2.0	2.2 ± 1.6	0.018 ± 0.012	0.29 ± 0.19	2.2 ± 1.6
$\gamma_2 \pm \sigma$	16.9 ± 11.0	9.0 ± 8.1	0.037 ± 0.027	4.50 ± 0.06	8.9 ± 8.1
MSE(a_1)	0.182	0.041	1.7×10^{-5}	3.1×10^{-4}	0.041
MSE(a_2)	0.192	0.145	3.0×10^{-5}	2.5×10^{-3}	0.143

over MLPCA in the estimated projection directions. This lack of improvement in the parameters is analogous to that of James–Stein (JS) estimators [28]. James and Stein have shown that shrinkage methods can result in lower risk (mean square error) than maximum likelihood methods for models with rank *greater than* 2. This property also applies to the proposed BPCA approach, since JS estimators are shown to be similar to empirical Bayesian estimators [29]. This property indicates that EBPCA should yield better results than MLPCA for the parameters for models with dimensions higher than two, as shown next.

5.1.1. *Improvement in EBPCA model parameters*

To examine the level of improvement in model parameters obtained by EBPCA, the effect of the model rank on the accuracy of EBPCA models is studied. To illustrate this effect, nine data sets, each with a different rank, having 10 variables and 50 observations, are used to derive EBPCA models. The rank of the data sets ranges from 1 to 9. The noise-free data of these data sets are generated as follows. Each of the first p noise-free variables in the pth data set, which is of rank p, is generated from the Gaussian distribution

$$\widetilde{\mathbf{X}}_i \sim N(0, Q_i), \quad \text{where} \quad Q_i \sim U(1,2), \quad i = 1, ..., p \qquad (55)$$

Next the last $m - p$ variables are generated by multiplying the first p variables by a $p \times (r - p)$ matrix, each of whose entries is drawn from the uniform distribution $U(0.5,1)$. Then the data are contaminated with noise such that the signal-to-noise ratio of all variables is 3. A Monte Carlo simulation of 100 realizations is performed for this analysis, and the results are schematically illustrated in Figure 3, which shows that the performance of EBPCA improves at higher model ranks. The percentage improvement shown in Figure 3 is computed as follows:

$$\%\ \text{improvement} = \frac{\text{MSE}_{\text{MLPCA}} - \text{MSE}_{\text{BPCA}}}{\text{MSE}_{\text{MLPCA}}} \times 100 \qquad (56)$$

Even when there is little improvement in the model parameter estimates at low ranks, EBPCA still provides better accuracy in estimating the noise-free data. This is an important advantage of EBPCA, since in many applications, such as data rectification and process monitoring, good estimation of the underlying noise-free data is essential.

Furthermore, the extent of improvement achieved by EBPCA in estimating the model parameters is larger for large MLPCA parameter errors. This is illustrated in Figure 4, which plots the EBPCA parameter MSE versus the MLPCA parameter MSE. The diagonal line represents equal MLPCA and EBPCA errors. Since most points in Figure 4 are below the diagonal, it indicates that the improvement in parameter estimation by EBPCA is greater when MLPCA does not do very well. Figure 4 also shows that most of the parameter errors lie below the equal error line, indicating that on average EBPCA results in a smaller parameter MSE than MLPCA.

5.1.2. *Estimating the PCA model rank*

The results reported in Table II were obtained with a known model rank. The performance of empirical methods of MAP and hypothesis for estimating the model rank is compared in Figure 5. This plot represents a Monte Carlo simulation of 100 realizations for different signal-to-noise ratios. It shows that the hypothesis testing approach, even though not perfect, is much more consistent than the MAP technique at various noise contents and that the MAP technique works only within a small range of signal-to-noise ratios, as discussed in Section 4.5. The percentage accuracy reported

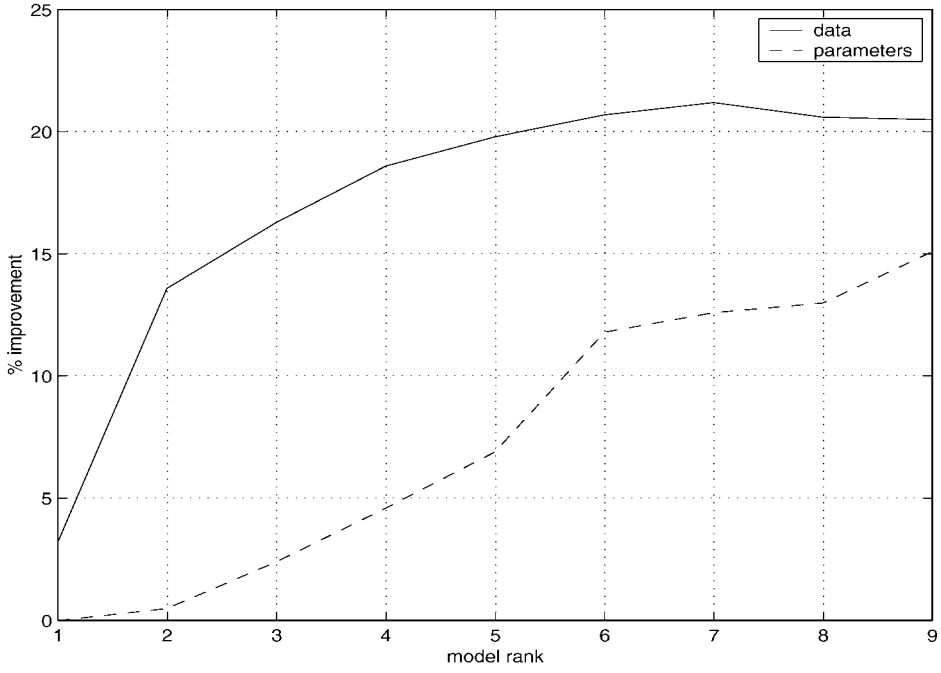

Figure 3. Percentage improvement achieved by EBPCA over MLPCA versus model rank. The improvement in parameter estimates is significant only for ranks greater than 2.

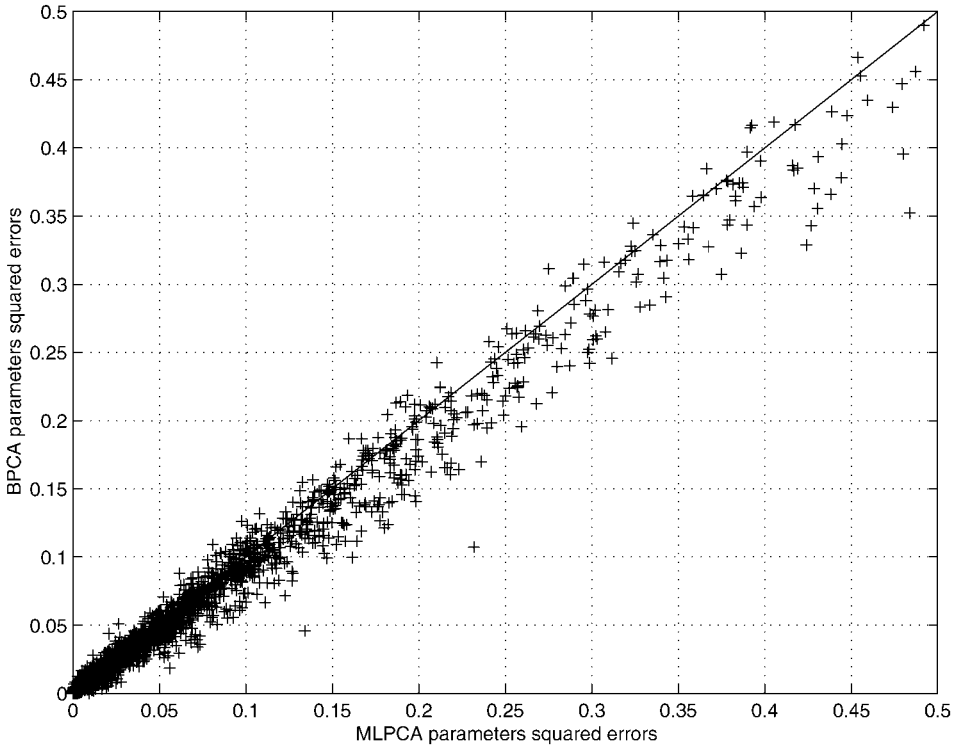

Figure 4. Comparison of parameter square errors obtained by EBPCA and MLPCA. The diagonal line represents equal MLPCA and EBPCA errors. Points below the diagonal indicate better performance of BPCA.

in Figure 5 is computed as follows:

$$\% \text{ accuracy} = \frac{\text{number of realizations the model rank is estimated correctly}}{\text{total number of realizations}} \times 100 \quad (57)$$

This poor performance of the MAP estimator of the model rank can be understood by comparing the relative magnitudes of the likelihood and prior terms of Equation (44). When more principal components are retained, the likelihood term increases and the empirical prior terms decrease, as shown in Figures 6(c) and 6(d). The likelihood increases

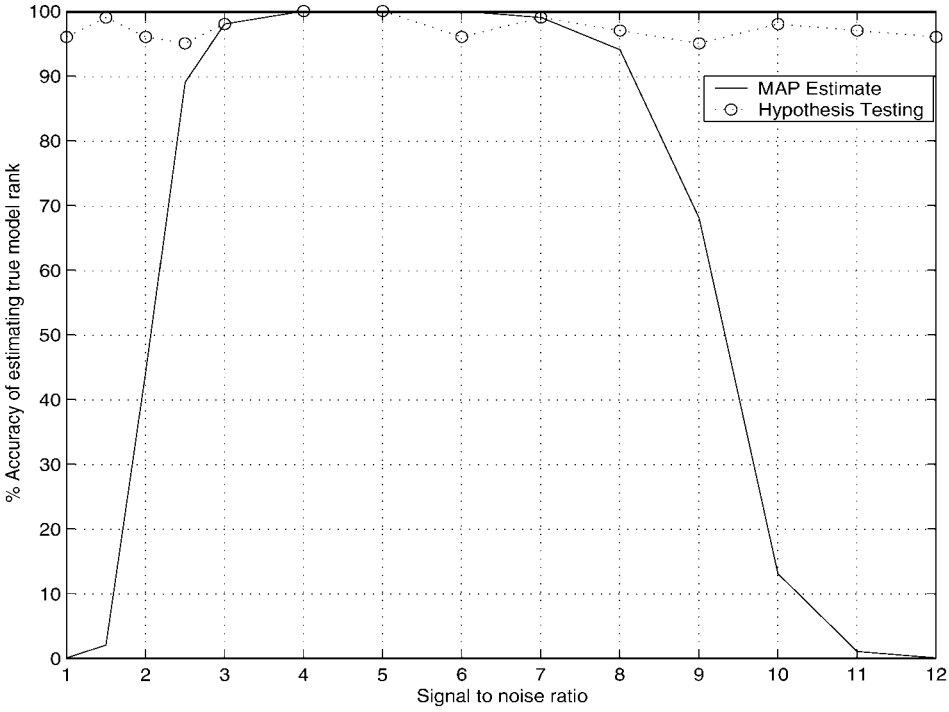

Figure 5. Percentage accuracy in estimating model rank for Gaussian data in Example 1.

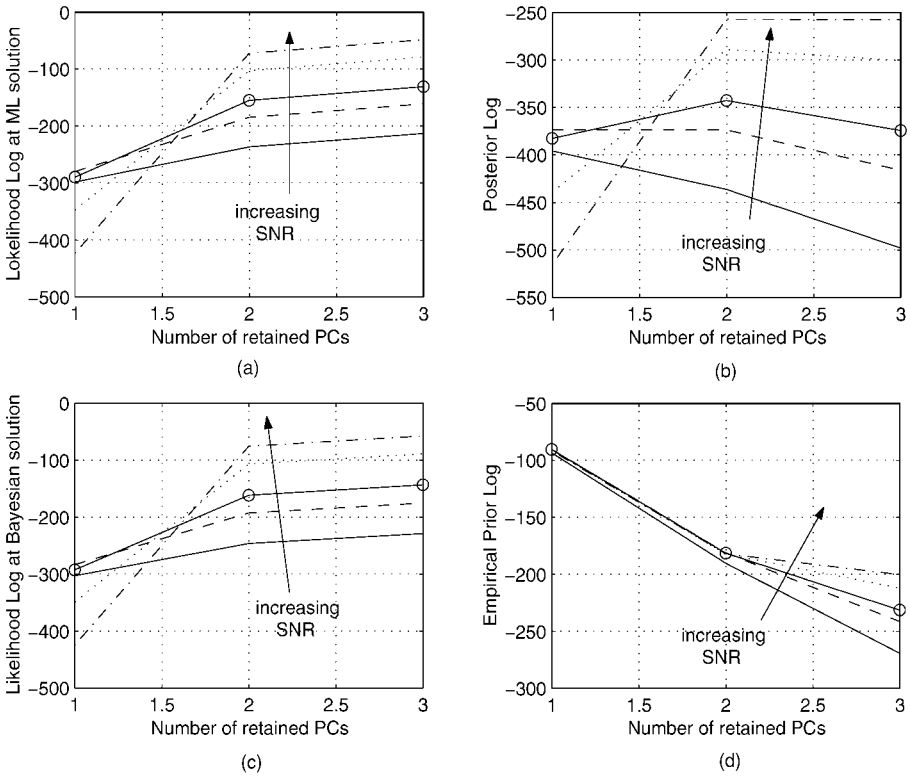

Figure 6. Performance of MAP method for estimating model rank for Example 1: (a) logarithm of likelihood function at MLPCA solution; (b) logarithm of posterior density; (c) logarithm of likelihood function at EBPCA solution; (d) logarithm of prior density. Each plot is versus different numbers of retained principal components and at different signal-to-noise ratios (SNR values are 1, 3, 6, 9 and 12).

because the likelihood function is an exponential function of the negative data mean square error, which decreases as more principal components are retained. As the mean square error decreases, its likelihood function increases.

On the other hand, the prior, which is an exponential function of the negative parameters and prior data mean square errors, decreases as these quantities increase at higher model dimensions. When the data have a moderate noise content (a signal-to-noise ratio in the range of 3–8 for this example), the posterior logarithm will have a maximum at the correct rank. At high signal-to-noise ratios, however, the likelihood term increases faster than the prior terms, resulting in an increasing posterior function that cannot be used for inference about the model rank. On the other hand, at low signal-to-noise ratios the prior terms dominate the posterior, which becomes a decreasing function that also cannot be used in this regard. This behavior of the posterior is illustrated in Figure 6(b).

The effectiveness of the hypothesis testing approach is demonstrated in Figure 7, which shows the sorted probabilities $\Pr_j$ for each retained principal component. Figure 7 shows that for most realizations the probabilities for the first principal component are above the 95% confidence line, while those corresponding to the second component are below the line. This means that in most cases the procedure is capable of identifying the correct model rank, which is 2 in this example, despite the small signal-to-noise ratio.

5.2. Example 2. Uniform data from a reactor operating at steady state

This example illustrates the performance of BPCA for data violating the normality assumption made in deriving the BPCA algorithm. The noise-free variables represent the stream flow rates for the reactor shown in Figure 8. A steady state material balance results in the following model [30]:

$$\begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{58}$$

The data matrix, which consists of five variables and 50 observations, is generated as follows. The first two noise-free variables F_1 and F_2 follow the uniform distributions $U(15,40)$ and $U(1,5)$ respectively, and the remaining variables are computed to satisfy the steady state model shown in Equation (58). Thus the actual rank of the data matrix is 2. Then all variables are contaminated with zero—mean Gaussian noise with the covariance matrix $\mathbf{Q}_{\varepsilon_x} = \text{diag}$ (1 9 16 16 1).

The results of a Monte Carlo simulation of 100 realizations with known model rank are summarized in Table III. These results illustrate the advantage of EBPCA over existing methods in estimating the underlying noise-free data, but no improvement in the model parameters over MLPCA, which

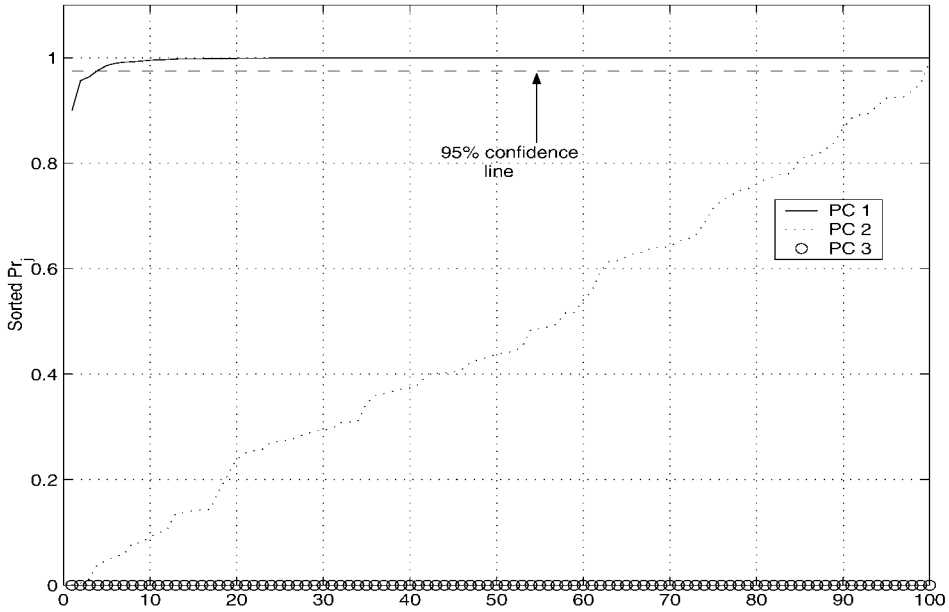

Figure 7. Sorted probabilities Pr_j for 100 realizations of hypothesis testing for Example 1. The *x*-axis is the index of sorted probabilities.

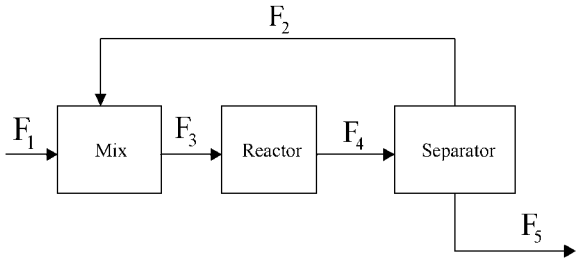

Figure 8. Flow sheet for Example 2.

is consistent with the results of the previous example for a model rank of less than 3. This example demonstrates that BPCA can outperform PCA and MLPCA even when the underlying assumptions of Gaussian distributions are violated.

Another important advantage of BPCA is its robustness to the number of retained principal components. This property is illustrated in Figure 9, which compares the mean square errors of the estimated variables for different techniques and different numbers of retained principal components. These plots show that EBPCA results in much smaller data mean square errors than PCA and MLPCA, especially when the model rank is overestimated. These plots also show that keeping only one principal component results in the least mean square errors for all techniques, indicating that the best model rank might be 1, even though the true mathematical rank of the noise-free data is 2. This is due to the large noise content on one of the independent variables, F_2, which makes it effectively like noise in the data.

Table III. PCA modeling and rectification of steady state reactor data (Example 2)

	PCA	MLPCA	EBPCA
MSE(F_1)	2.254	0.591	0.519
MSE(F_2)	3.333	5.032	2.465
MSE(F_3)	11.176	6.096	3.161
MSE(F_4)	9.904	5.381	2.916
MSE(F_5)	2.259	0.600	0.524
MSE(**F**)	5.786	3.540	1.917
$\gamma_1 \pm \sigma$	0.51 ± 0.25	0.56 ± 0.29	0.56 ± 0.29
$\gamma_2 \pm \sigma$	68.8 ± 16.3	33.5 ± 19.2	33.5 ± 19.2

5.2.1. *Estimating the model rank*

The results of hypothesis testing to estimate the actual model rank are shown in Figure 10 as the sorted probabilities Pr_j of 100 realizations for different principal components. Figure 10 shows that probability for the first principal component is noticeably smaller than unity for most realizations, indicating that the estimated model rank is 1, which agrees with the earlier observation that retaining one principal component results in the least mean square error.

5.3. Example 3. Dynamic non-stationary data

The objective of this example is to show the performance of BPCA for data violating most of the assumptions made in its derivation. The noise-free data are generated using the

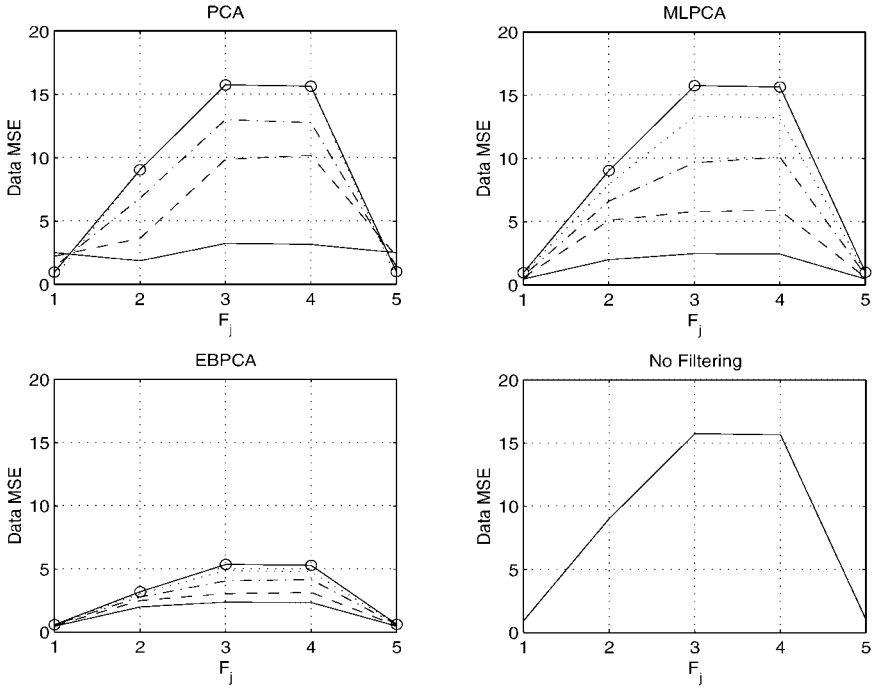

Figure 9. Data mean square errors versus flow rates (F_j) for different numbers of principal components for Example 2. The true model rank is 2. The narrower range of variation for EBPCA indicates greater robustness to errors in estimating the model rank. Legend: full line, one PC; broken line, two PCs; chain line, three PCs; dotted line, four PCs; circles, five PCs.

following dynamic model:

$$\tilde{y}(k) = 0.8\tilde{y}(k-1) + \tilde{u}(k)$$

$$\text{where} \quad \tilde{u}(k) \sim \begin{cases} N(0,2), & 1 \leqslant k \leqslant 15 \\ N(5,2), & 16 < k \end{cases} \tag{59}$$

Then the variables $\tilde{y}$ and $\tilde{u}$ are contaminated with zero-mean Gaussian noise with variances 2 and 4 respectively. To account for the dynamics in the data, the matrix $\tilde{\mathbf{X}}$, which contains 64 observations, is constructed as

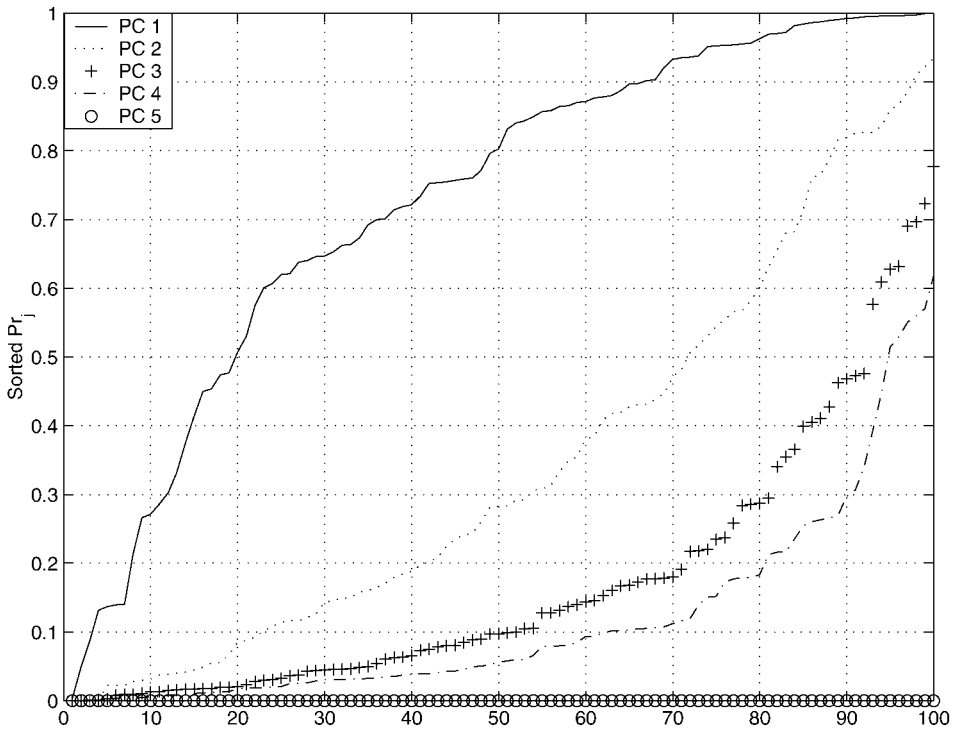

Figure 10. Sorted probabilities Pr_j for hypothesis testing for Example 2. The *x*-axis is the index of sorted probabilities.

follows:

$$\mathbf{X} = [\,\mathbf{Y}(k-1) \quad \mathbf{U}(k) \quad \mathbf{Y}(k)\,] = \begin{bmatrix} y(1) & u(2) & y(2) \\ . & . & . \\ y(k-1) & u(k) & y(k) \\ . & . & . \\ y(64) & u(65) & y(65) \end{bmatrix} \tag{60}$$

Thus the true rank of the noise-free data is 2 to satisfy Equation (59), and the corresponding noise covariance matrix is

$$\mathbf{Q}_{\varepsilon_x} = \text{diag}\,(4 \quad 2 \quad 4) \tag{61}$$

which is assumed to be known. Since the input $\widetilde{u}$ contains a step change, the measurements are far from Gaussian. The model dynamics also results in autocorrelated measurements.

The results of a Monte Carlo simulation of 100 realizations, summarized in Table IV, show a clear advantage of EBPCA over both PCA and MLPCA. These results are obtained under the assumption that the true model rank of 2 is known. As illustrated in Section 5.1, if a more accurate prior is used or historical data are available, BPCA can perform even better.

The results of hypothesis testing to estimate the model rank are shown in Figure 11 as a plot of the sorted probabilities Pr_j. These results shows that for a confidence of about 95% the hypothesis testing approach has successfully estimated the true rank in more than 95% of all realizations.

Table IV. Rectification of non-stationary dynamic data using dynamic PCA (Example 3)

MSE	PCA	MLPCA	EBPCA
$\mathbf{Y}(k-1)$	3.26	2.77	2.46
$\mathbf{U}(k)$	1.49	1.59	1.07
$\mathbf{Y}(k)$	2.70	2.18	2.04
$\mathbf{X}$	2.48	2.18	1.86
$\gamma_1 \pm \sigma$	0.32 ± 0.20	0.33 ± 0.21	0.33 ± 0.21
$\gamma_2 \pm \sigma$	12.8 ± 7.0	6.0 ± 4.5	6.0 ± 4.5

5.4. Example 4. PCA filtering of temperature data from a distillation column

In this example, temperature measurements from a distillation column are used to illustrate the performance of EBPCA in estimating the underlying noise-free data. The noise-free data consist of six variables and 50 observations representing temperature measurements from six different trays in a 30-tray distillation column used to separate methanol and ethanol from propanol and *n*-butanol. The feed stream enters the distillation column at the 15th tray and is equimolar in the four components. The data used in this example are simulated under a temperature-controlled operation of the distillation column [31]. The data are then contaminated with zero-mean Gaussian noise with the covariance matrix $\mathbf{Q}_{\varepsilon_x}$ = diag (0.05 0.1 0.05 0.1 0.05 0.1). A Monte Carlo simulation is performed assuming that the actual model rank is 3, and the results are summarized in Table V. Again, EBPCA has a smaller data MSE than existing techniques.

This example also confirms the robustness of EBPCA to errors in the number of retained principal components. A plot of the mean square errors versus T_j for different numbers of components is shown in Figure 12. The smaller variation of the plots for EBPCA for different numbers of

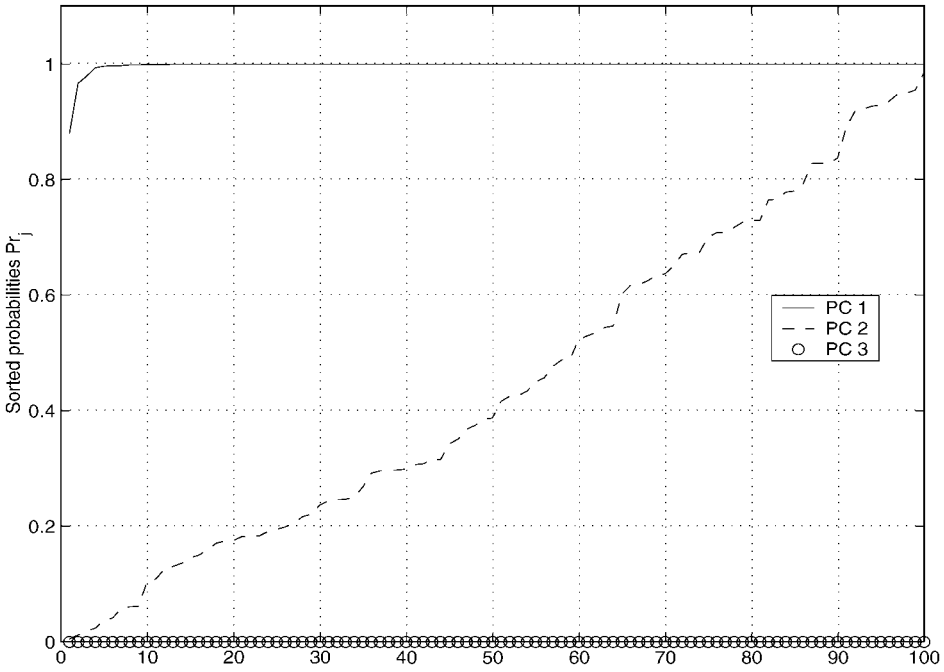

Figure 11. Sorted probabilities Pr_j for hypothesis testing for Example 3. The *x*-axis is the index of sorted probabilities.

Table V. Data filtering of temperature data from a distillation column (Example 4)

	PCA	MLPCA	EBPCA
MSE(T_1)	0.034	0.035	0.021
MSE(T_2)	0.064	0.034	0.020
MSE(T_3)	0.043	0.044	0.031
MSE(T_4)	0.079	0.072	0.059
MSE(T_5)	0.030	0.032	0.015
MSE(T_6)	0.084	0.078	0.058
MSE(**T**)	0.056	0.049	0.034
$\gamma_1 \pm \sigma$	0.010 ± 0.006	0.010 ± 0.006	0.010 ± 0.006
$\gamma_2 \pm \sigma$	4.2 ± 1.6	4.5 ± 2.0	4.5 ± 2.0
$\gamma_3 \pm \sigma$	52.3 ± 17.5	38.1 ± 21.3	38.1 ± 21.3

selected components indicates that EBPCA is more robust to errors in estimating the model rank. This is due to the fact that when the model dimension is overestimated the data mean square errors for the different variables increase, until they become the noise variance when all the principal components are retained. In EBPCA, however, the data mean square errors are much smaller than the noise variance even when a full-rank model is used. Estimating the model rank by hypothesis testing yields a rank of 3 for most realizations, as portrayed in Figure 13.

5.5. Example 5. PCA filtering of UV absorption data

In this example, industrial UV absorption data are used to illustrate the performance of EBPCA in estimating the underlying noise-free data. The data consist of 35 observations and four variables representing the absorption of four solutions of 1-fluoro-3-nitrobenzene and dimethyl phthalate at 35 wavelengths in the range of 215–385 nm [32]. These data, plotted in Figure 14, show that there are two peaks corresponding to the two compounds and that the data distribution is far from normal.

The data are then contaminated with zero-mean Gaussian noise such that the signal-to-noise ratio in all variables is 2. A Monte Carlo simulation of 100 realizations is performed assuming that the actual model rank is 1. The results in Table VI show that EBPCA does better even when the assumptions made in the derivation of the BPCA algorithm are violated and even when only 35 observations are available. The mean square errors obtained using different numbers of components, plotted in Figure 15, again show that EBPCA is more robust to errors in estimating the model rank. The model rank estimated by hypothesis testing is found to be 1, as illustrated in Figure 16. For a larger signal-to-noise ratio, EBPCA can still benefit from the use of prior knowledge.

This example is repeated *without* adding extra noise, but assuming that the real data are already noisy with a signal-to-noise ratio of 2. The purpose of this repetition is to visually compare the performance of the different methods, although the underlying data are not known. The performances of the different methods are illustrated in Figure 17, which shows that PCA, MLPCA and EBPCA are comparable in this case, which makes sense, since the data are noise-free and all techniques should perform similarly. The results for other signal-to-noise ratios are similar.

6. CONCLUSIONS

This paper presents a Bayesian approach to the popular technique of principal component analysis. Unlike previous related research, the approach developed in this paper uses prior knowledge about the parameters *and* measurements and integrates Bayesian parameter estimation with Bayesian reconciliation problems while retaining the orthogonality features of PCA. Consequently, BPCA can improve the

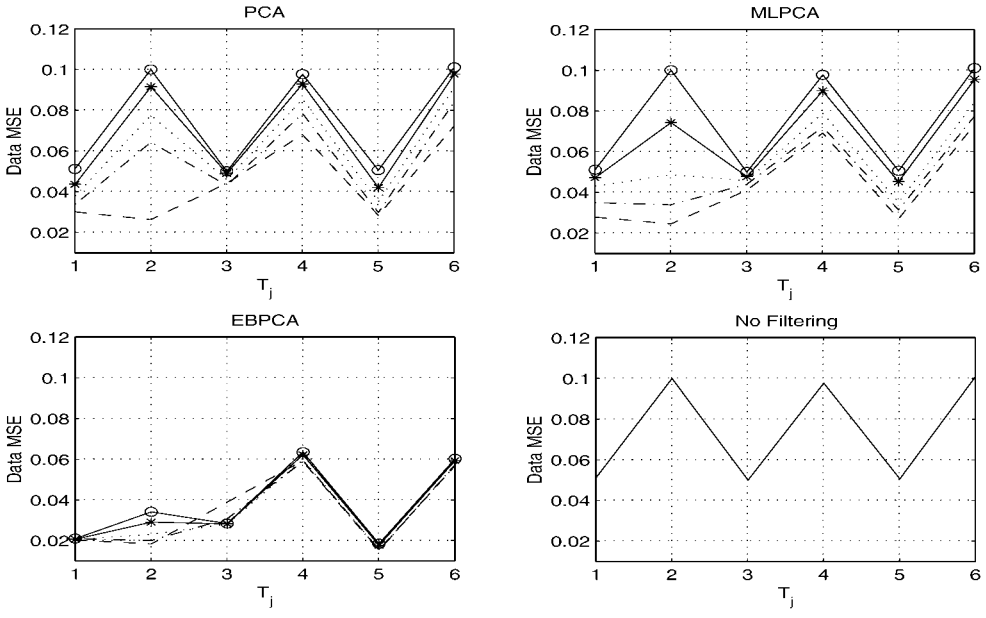

Figure 12. Data mean square errors versus temperatures (T_j) obtained by various techniques for different numbers of retained principal components for Example 4. Legend: broken line, two PCs; chain line, three PCs; dotted line, four PCs; asterisks, five PCs; circles, six PCs.

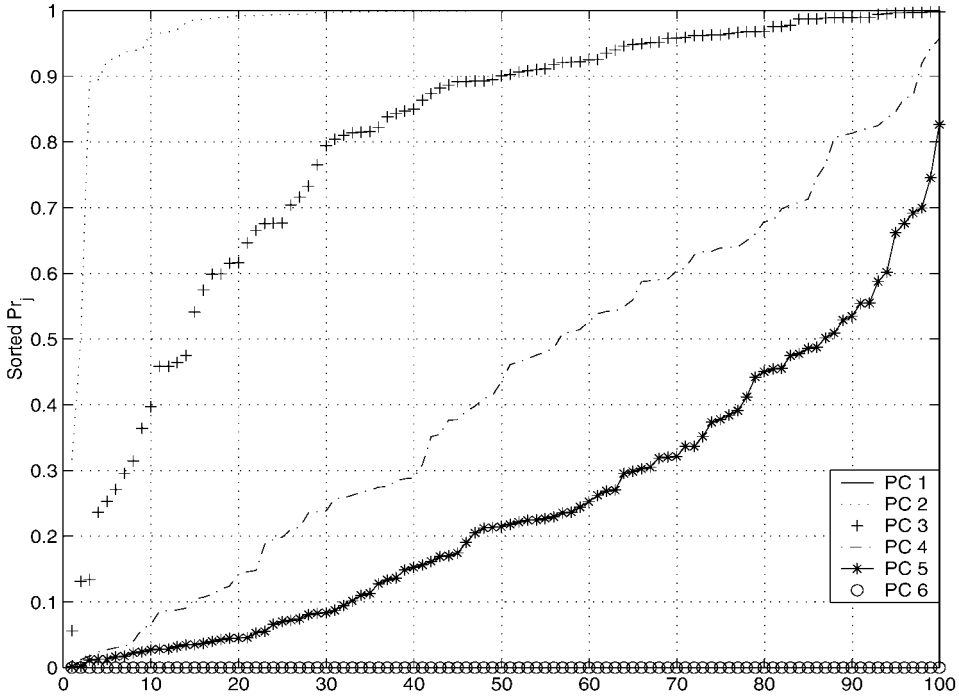

Figure 13. Sorted probabilities Pr_j from hypothesis testing for Example 4. The *x*-axis is the index of sorted probabilities.

accuracy of both the estimated parameters and measurements. The formulation of the BPCA approach is shown to be more general than existing methods and to reduce to these techniques under special conditions. For example, a uniform prior converts BPCA to MLPCA [4]. In addition, if the noise covariance matrix is assumed to be a multiple of identity, BPCA reduces to PCA.

The BPCA algorithm is derived based on the assumptions that the model rank is known or can be estimated by other methods and that the noise and underlying measurements are Gaussian. The last assumption permits the use of Gaussian priors for the loadings and scores, and the development of a computationally efficient algorithm. Since the performance of any Bayesian approach depends on the quality of the prior, techniques are developed for estimating the prior parameters from the available measurements. The

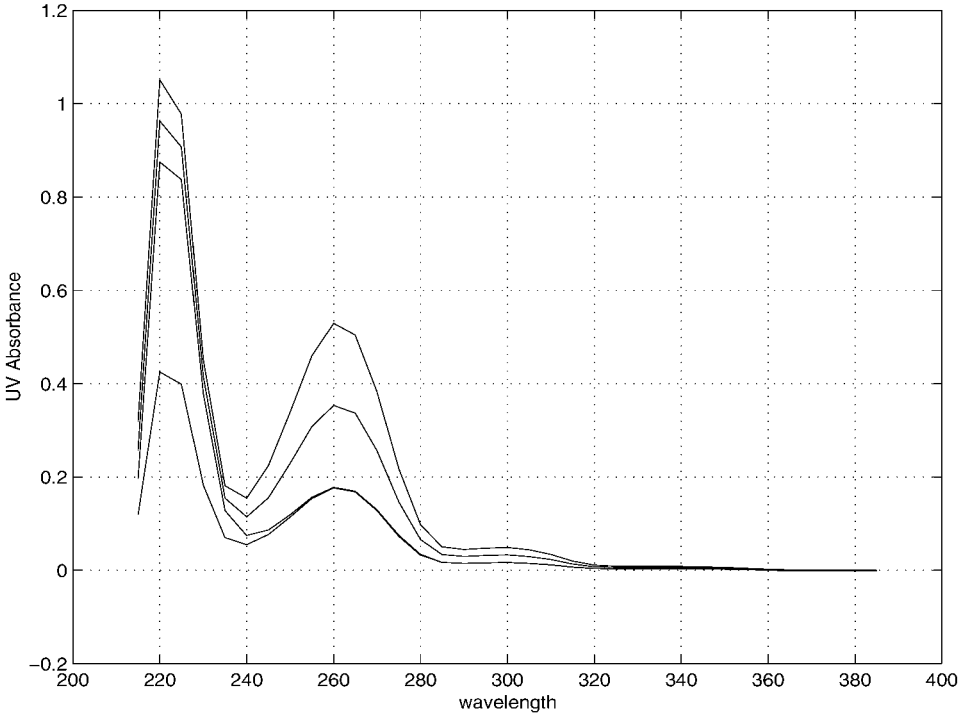

Figure 14. UV absorption data used in Example 5 for four solutions versus wavelength.

Table VI. PCA filtering of UV absorption data (Example 5)

	PCA	MLPCA	EBPCA
MSE($\hat{x}_1$)	0.0018	0.0016	0.0014
MSE($\hat{x}_2$)	0.0089	0.0076	0.0073
MSE($\hat{x}_3$)	0.0151	0.0120	0.0103
MSE($\hat{x}_4$)	0.0094	0.0076	0.0068
MSE($\hat{x}$)	0.0088	0.0072	0.0065
$\gamma \pm \sigma$	5.2 ± 2.2	4.8 ± 2.2	4.8 ± 2.2

resulting empirical BPCA (EBPCA) approach can utilize historical data or only the data for which the model is being developed. Several illustrative examples demonstrate the superior performance of BPCA over PCA and MLPCA even when the underlying assumptions of Gaussian distributions are violated. Furthermore, BPCA is also shown to be more robust to errors in estimating the model rank.

The proposed BPCA algorithm is expected to be useful in any PCA or MLPCA problem that permits estimation of a

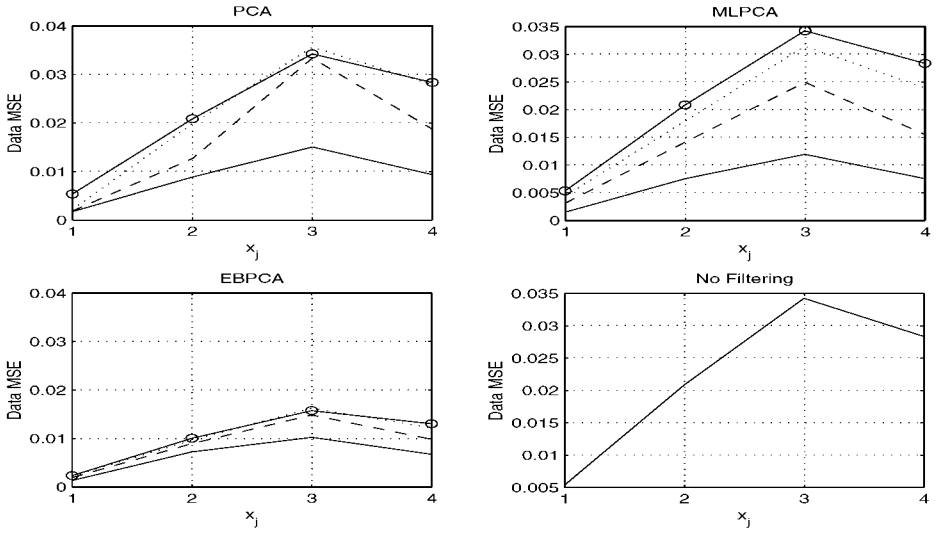

Figure 15. Data mean square errors versus variables (x_j) for various techniques at different numbers of retained principal components for Example 5. Legend: full line, one PC; broken line, two PCs; chain line, three PCs; circles, four PCs.

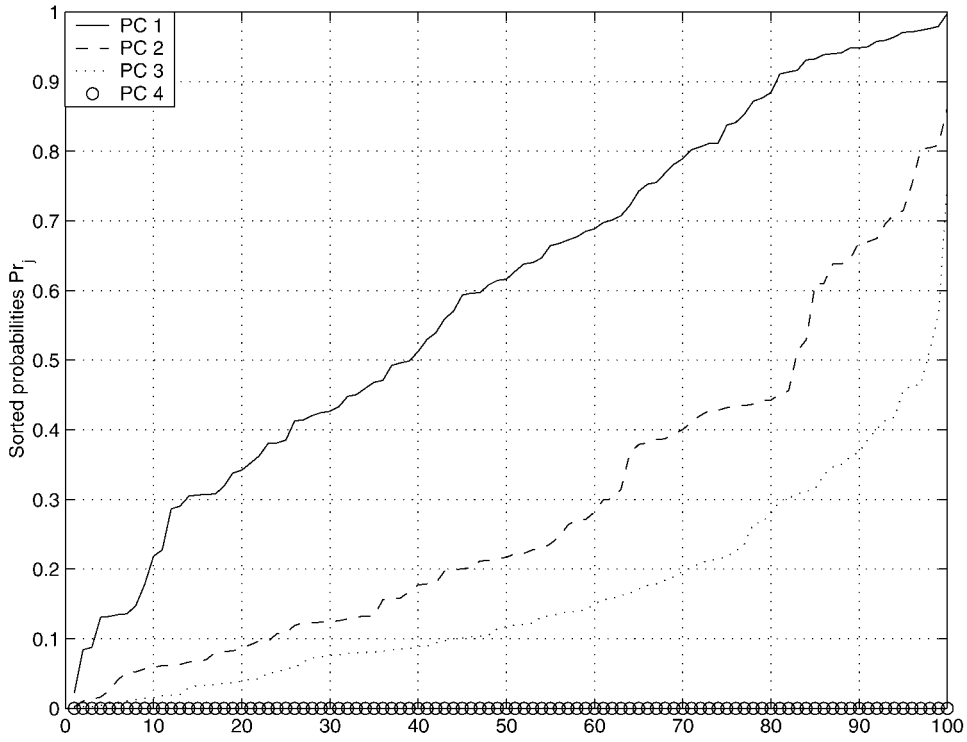

Figure 16. Sorted probabilities Pr_j from hypothesis testing for Example 5. The *x*-axis is the index of sorted probabilities.

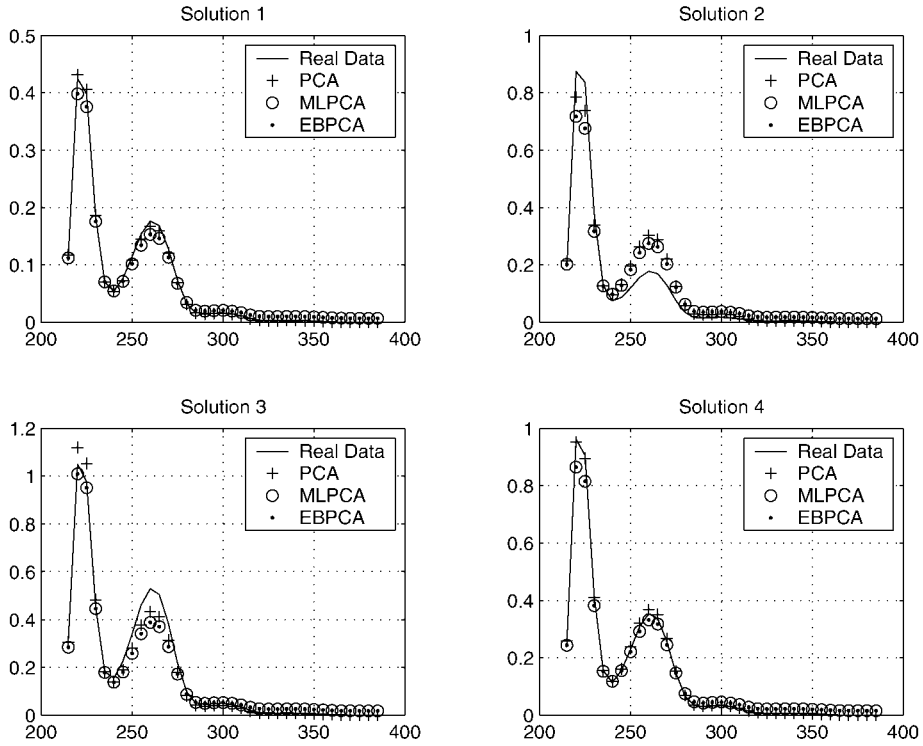

Figure 17. Comparison of PCA, MLPCA and EBPCA using real UV absorption data for Example 5.

reasonably accurate prior. It can also provide the foundation for Bayesian latent variable regression (BLVR) methods, resulting in Bayesian analogues of existing regression methods. Like BPCA, these Bayesian regression methods are expected to perform better than their non-Bayesian counterparts. Indeed, such a Bayesian linear regression approach has been developed recently [33]. Recent work also shows that the challenge of estimating an accurate prior distribution may be addressed by combining wavelets with Bayesian analysis [34] or by Monte Carlo methods [35]. These and other research advances along with increasing computational ability are expected to increase the popularity of Bayesian methods for a variety of statistical and chemometric tasks [36].

Acknowledgements

The authors thank the National Science Foundation (CAREER award CTS 9733627) for financial support, Dr Manabu Kano for the distillation data and Dr C. H. Lochmuller for the UV absorption data.

APPENDIX I. DERIVATION OF THE MLPCA DATA RECTIFICATION SOLUTION

The MLPCA data reconciliation problem can be formulated as follows:

$$\{\hat{\mathbf{z}}_i\}_{\text{MLPCA}} = \underset{\hat{\mathbf{z}}_i}{\text{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i)$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i \tag{62}$$

Solution

Define the Lagrange function as

$$L = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) \tag{63}$$

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

$$\frac{\partial L}{\partial \hat{\mathbf{x}}_i} = -2\mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + {}^{\text{T}} = 0 \tag{64}$$

$$\frac{\partial L}{\partial \hat{\mathbf{z}}_i} = -\hat{\boldsymbol{\alpha}} = 0 \tag{65}$$

$$\frac{\partial L}{\partial} = \hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i = 0 \tag{66}$$

Substituting Equation (64) in Equation (65), get

$$\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = 0 \tag{67}$$

Substituting Equation (66) in Equation (67), get

$$\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) = 0 \tag{68}$$

Rearranging Equation (68), get the MLPCA solution

$$\{\hat{\mathbf{z}}_i\}_{\text{MLPCA}} = (\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} \hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} \mathbf{x}_i \tag{69}$$

APPENDIX II. DERIVATION OF THE BPCA DATA RECTIFICATION SOLUTION

The BPCA data reconciliation problem can be formulated as

follows:

$$\{\hat{\mathbf{z}}_i\}_{\text{MAP}} = \underset{\hat{\mathbf{z}}_i}{\text{argmin}} \left\{ (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}})^{\text{T}} \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) \right\}$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i \tag{70}$$

Solution

Define the Lagrange function as

$$L = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}})^{\text{T}} \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) + (\hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) \tag{71}$$

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

$$\frac{\partial L}{\partial \hat{\mathbf{x}}_i} = -2\mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + {}^{\text{T}} = 0 \tag{72}$$

$$\frac{\partial L}{\partial \hat{\mathbf{z}}_i} = 2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - \hat{\boldsymbol{\alpha}}^{\text{TT}} = 0 \tag{73}$$

$$\frac{\partial L}{\partial} = \hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i = 0 \tag{74}$$

Substituting Equation (72) in Equation (73), get

$$2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - 2\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = 0 \tag{75}$$

Substituting Equation (74) in Equation (75), get

$$2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - 2\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) = 0 \tag{76}$$

Rearranging Equation (76), get the MAP solution

$$\{\hat{\mathbf{z}}_i\}_{\text{MAP}} = (\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} \hat{\boldsymbol{\alpha}} + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1})^{-1} (\hat{\boldsymbol{\alpha}}^{\text{T}} \mathbf{Q}_{\varepsilon_x}^{-1} \mathbf{x}_i + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) \tag{77}$$

REFERENCES

1. Kresta JV, MacGregor JF and Marlin TE. Multivariate Statistical Process Monitoring of Process Operating Performance. *Can. J. Chem. Eng.* 1991; **69**: 35–47.
2. Wise BM, Ricker NL, Veltkamp DF and Kowalski BR. A Theoretical Basis for the Use of Principal Component Models for Monitoring Multivariate Processes. *Process Control Qual.* 1990; **1**: 41.
3. Kramer MA and Mah RSH. Model Based Monitoring. In *Proceedings of the International Conference on Foundations of Computer Aided Process Operations*, Rippin D, Hale J, Davis J (eds). CACHE: Austin, TX 1994; 45–68.
4. Wentzell PD, Andrews D, Hamilton DC, Faber K and Kowalski BR. Maximum Likelihood Principal Component Analysis. *J. Chemometrics* 1997; **11**: 339–366.
5. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer: New York, 1985.
6. West M and Harrison J. *Forecasting and Dynamic Models*. Springer: New York, 1997.
7. Seber GAE. *Multivariate Observations*. Wiley: New York, 1984.
8. Press SJ. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd edn). Krieger: Malabar, Florida, 1982.
9. Press SJ. *Applied Multivariate Analysis*. Holt, Rinehart and Winston: New York, 1972.
10. Martin JK and McDonald RP. Bayesian Estimation in Unrestricted Factor Analysis: A Treatment for Heywood Cases. *Psychometrika* 1975; **40**: 505–517.
11. Press SJ and Shigemasu K. In *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*. 1989; 271–278.
12. Lee SE and Press SJ. Robustness of Bayesian Factor Analysis Estimates. *Commun. Statist. – Theory Methods* 1998; **27**: 1871–1893.
13. Wold S. Exponentially Weighted Moving Principal Components Analysis and Projections to Latent Structures. *Chemometrics Intell. Lab. Syst.* 1994; **23**: 149–161.
14. Bakshi BR. Multiscale PCA with Application to Multivariate Statistical Process Monitoring. *AIChE J.* 1998; **44**: 1596–1610.
15. Hastie TJ and Stuetzle W. Principal Curves. *J. Am. Statist. Assoc.* 1989; **84**: 505–516.
16. Mah RSH. *Chemical Process Structures and Information Flows*. Butterworths: Boston, MA, 1990.
17. Kadane JB. Prime Time for Bayes. *Controlled Clin. Trials* 1995; **16**: 313–318.
18. Gelman A, Carlin JB, Stern HS and Rubin D. *Bayesian Data Analysis*. Chapman and Hall: London, 1995.
19. Gilks WR, Richardson S, Spiegelhalter D (eds). *Practical Markov Chain Monte Carlo*. Chapman and Hall: New York, 1996.
20. Robert CP. *The Bayesian Choice: a Decision Theoretic Motivation*. Springer: New York, 1994.
21. Girshick MA. On the Sampling Theory of Roots of Determinantal Equations. *Ann. Math. Statist.* 1939; **10**: 203–224.
22. Maritz JS. *Empirical Bayes Methods*. Methuen: London, 1970.
23. Carlin BB and Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall: London, 1996.
24. Ku W, Storer RH and Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics Intell. Lab. Syst.* 1995; **30**: 179–196.
25. Basilevsky A. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley: New York, 1994.
26. Wold S. Cross-Validatory Estimation of the Number of Components in actor and Principal Component Models. *Technometrics* 1978; **20**: 397–405.
27. Eastment HT and Krzanowski WJ. Cross-Validatory Choice of the Number of Components from Principal Component Analysis. *Technometrics* 1982; **24**: 73–77.
28. James W and Stein C. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematics and Statistics*, vol. **1**. University of California Press: Berkeley, CA, 1961; 361–379.
29. Gruber MH. *Improving Efficiency by Shrinkage: the James–Stein and Ridge Regression Estimators*. Marcel Dekker: New York, 1998.
30. Johnston L and Kramer M. Maximum Likelihood Data Rectification: Steady State Systems. *AIChE J.* 1995; **41**: 2415–2426.
31. Kano M, Miyazaki K, Hasebe S and Hashimoto I. Inferential Control System of Distillation Compositions using Dynamic Partial Least Squares Regression. *J. Process Control* 2000; **10**: 157–166.
32. Luchmuller CH and Reese CE. Introduction to Factor Analysis. *Crit. Rev. Anal. Chem.* 1998; **28**: 21–49.
33. Nounou MN, Bakshi BR, Goel PK and Shen X. Process Modeling by Bayesian Latent Variable Regression. *AIChE J.* 2002; **48**: 1775–1793.
34. Nounou MN, Bakshi BR, Goel PK and Shen X. Multiscale Bayesian Data Rectification of Linear Steady-State and Dynamic Systems with Steady-State Models. *Ind. Eng. Chem. Res.* 2001; **40**: 261–274.
35. Chen W-S, Bakshi BR, Goel PK and Ungarala S. Bayesian Rectification of Unconstrained Nonlinear Dynamic Systems by Sequential Monte Carlo Sampling. Tech. Rep., Ohio State University, Columbus, OH, 2002.
36. Malakoff D. Bayes offers a 'new' way to make sense of numbers. *Science* 1999; **286**: 1460–1464.