Bayesian Principal Component Analysis

Mohamed N. Nounou, Bhavik R. Bakshi*

Department of Chemical Engineering

Prem K. Goel and Xiaotong Shen

Department of Statistics

The Ohio State University

Columbus, OH 43210, USA

Abstract

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 its extensions utilize 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.

KEY WORDS: Bayesian analysis; principal component analysis; filtering; latent variables.

* Correspondence should be addressed to Bhavik R. Bakshi.

Fax: 1-614-292-3769, Email: bakshi.2@osu.edu

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 its 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 dimension 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 solve 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 that all information brought by the data about the quantities of interest are 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 towards 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 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] They could obtain analytical large sample estimates for the factor scores, 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 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 the 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 application 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 reduces to these methods when a uniform prior is used.

The rest of this paper is organized as follows. The next section, introduces PCA and MLPCA. A brief description of Bayesian estimation is presented next. Then, 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 number of principal components. Finally, 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 a $n \times r$ matrix of measured process variables, $\mathbf{X} = \tilde{\mathbf{X}} + \boldsymbol{\varepsilon}_X$, where

$\tilde{\mathbf{X}}$ 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 Z is a $n \times r$ matrix of the principal components or the principal component scores, and $\boldsymbol{\alpha}$ 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) into Equation (2) gives

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

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,

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

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

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.,

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

where, $\mathbf{U}$ is a unitary matrix containing the left eigenvectors, $\boldsymbol{\alpha}$ is a unitary matrix containing the right eigenvectors, and $\mathbf{Z} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$. 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]

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

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

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$, which are $(r \times 1)$ vectors, are the i-th 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 purpose, 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 due 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)

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 "$\tilde{p}$", as

$$\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}\}_{MLPCA} = \underset{\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}}{\arg\max}\, L(\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{p}}; \mathbf{X}) = \underset{\hat{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}}{\arg\max}\, P(\mathbf{X} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}, \tilde{p}) \tag{7}$$

subject to the constraint given in Equation (6b). If the distribution of the contaminating noise is assumed to be zero-mean Gaussian, i.e., $\boldsymbol{\varepsilon}_X \sim N(\mathbf{0}, \mathbf{Q}_{\varepsilon_X})$, 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,

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

where $\mathbf{Q}_{\varepsilon_X}$ 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 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

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

$$\text{s.t. } \{\hat{\mathbf{z}}_i\}_{\text{MLPCA}} = \underset{\hat{\mathbf{z}}_i}{\text{argmin}} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \tag{9}$$

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 from solution as shown in Appendix I,

$$\{\hat{\mathbf{z}}_i\}_{\text{MLPCA}} = \left(\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_x}^{-1} \hat{\boldsymbol{\alpha}}\right)^{-1} \hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_x}^{-1} \mathbf{x}_i . \tag{10}$$

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 $\tilde{\theta}$, from a set of measurements of the quantity, y, Bayesian estimation starts by defining the conditional density of the variable to be estimated given the measurements, $P(\tilde{\theta} \mid y)$, which is called the posterior. The *posterior* is a density

function that describes the behavior of the quantity, $\tilde{\theta}$, *after* observing the measurements. Using Bayes rule, the posterior can be written as follows.

$$P(\tilde{\theta} \mid \mathrm{y}) = \frac{P(\mathrm{y} \mid \tilde{\theta}) P(\tilde{\theta})}{P(\mathrm{y})}. \tag{11}$$

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,

$$P(\tilde{\theta} \mid \mathrm{y}) \propto P(\mathrm{y} \mid \tilde{\theta}) P(\tilde{\theta})$$

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 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]:

i. 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.
ii. Calculate the conditional density of the variables to be estimated given the observed data (posterior).

iii. Evaluate the implication of the posterior and check the accuracy of the estimated quantities.

The second step is a mathematical one, which involves computing the posterior density function. When the likelihood and the prior densities are mathematically simple, such computation can be done analytically. However, for more complicated problems, it is usually done empirically by some sampling algorithm, such as Markov Chain Monte Carlo (MCMC).[19] The third step is more judgmental, 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 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,

$$\{\hat{\theta}\}_{\text{MAP}} = \arg\max_{\tilde{\theta}} P(\mathrm{y} \mid \tilde{\theta}) P(\tilde{\theta}). \tag{14}$$

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 into 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.

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

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \mathbf{X}\right) = \frac{P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right)}{P\left(\mathbf{X}\right)}. \tag{15}$$

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

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \mathbf{X}\right) \propto P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right). \tag{16}$$

The Prior Density Function

The prior is the joint density of the noise-free, principal components, projection directions, and rank of true PCA model, 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

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) = P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}\right) P(\tilde{\mathrm{p}}). \tag{17}$$

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

$$P(\tilde{\mathrm{p}} = \mathrm{j}) = \mathrm{k}_{\mathrm{j}}, \quad \text{such that, } \sum_{\mathrm{j}=1}^{\mathrm{r}} \mathrm{k}_{\mathrm{j}} = 1. \tag{18}$$

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

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}\right) = P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}\right). \tag{19}$$

Thus, the unnormalized posterior can be written as,

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \mathbf{X}\right) \propto P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}\right) P(\tilde{\mathrm{p}}). \tag{20}$$

4.2 Simplifying Assumptions

Computing the posterior density shown in equation (20) requires defining the prior and the 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 section, the assumptions and their implications are described.

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, $\tilde{\mathrm{p}}$, is known. As shown in Section 5, the impact of this assumption is less severe for BPCA than PCA or MLPCA. Under this assumption, the rank portion of the prior density becomes

$$P(\tilde{\mathrm{p}}) = 1, \tag{21}$$

reducing the prior to

$$P(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}) P(\tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}) P(\tilde{\mathrm{p}}) = P(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}) P(\tilde{\boldsymbol{\alpha}}) \tag{22}$$

and simplifying the posterior to

$$P(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}) P(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}) P(\tilde{\boldsymbol{\alpha}}). \tag{23}$$

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.

Loss function

In this work, a zero-one loss function of the form,

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

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

$$\{\hat{\boldsymbol{\alpha}}, \hat{\mathbf{Z}}\}_{\text{Bayesian}} = \arg\max_{\tilde{\boldsymbol{\alpha}}, \tilde{\mathbf{Z}}} P(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}) P(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}) P(\tilde{\boldsymbol{\alpha}}), \tag{25}$$

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.

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., $\mathbf{X} = \tilde{\tilde{\mathbf{X}}} + \boldsymbol{\varepsilon}_{\mathbf{X}}$, where, $\boldsymbol{\varepsilon}_{\mathbf{X}} \sim N\left(0, \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{X}}}\right)$, then the likelihood function will also be normal with the following moments,

$$E\left[\mathbf{X} \mid \tilde{\tilde{\mathbf{Z}}}, \tilde{\tilde{\boldsymbol{\alpha}}}, \tilde{\tilde{\mathrm{p}}}\right] = E\left[\tilde{\tilde{\mathbf{X}}} + \boldsymbol{\varepsilon}_{\mathbf{X}}\right] = \tilde{\tilde{\mathbf{X}}} \tag{26}$$

and, $Cov\left[\mathbf{X} \mid \tilde{\tilde{\mathbf{Z}}}, \tilde{\tilde{\boldsymbol{\alpha}}}, \tilde{\tilde{\mathrm{p}}}\right] = E\left[\left(\mathbf{X} - \tilde{\tilde{\mathbf{X}}}\right)^{\mathrm{T}}\left(\mathbf{X} - \tilde{\tilde{\mathbf{X}}}\right)\right] = E\left[\left(\boldsymbol{\varepsilon}_{\mathbf{X}}\right)^{\mathrm{T}}\left(\boldsymbol{\varepsilon}_{\mathbf{X}}\right)\right] = \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{X}}}$, (27)

These moments are assumed to be known. Therefore,

$$\mathrm{P}\left(\mathbf{X} \mid \tilde{\tilde{\mathbf{Z}}}, \tilde{\tilde{\boldsymbol{\alpha}}}, \tilde{\tilde{\mathrm{p}}}\right) \sim N\left(\tilde{\tilde{\mathbf{X}}}, \mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathbf{X}}}\right). \tag{28}$$

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

Multivariate Gaussian noise-free data

The structure of the densities, $P\left(\tilde{\tilde{\mathbf{Z}}} \mid \tilde{\tilde{\boldsymbol{\alpha}}}\right)$ and $P\left(\tilde{\tilde{\boldsymbol{\alpha}}}\right)$, depend on the nature of the noise-free variables. In general, the density, $P\left(\tilde{\tilde{\boldsymbol{\alpha}}}\right)$, 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 examples in this section indicate, this assumption seems to be reasonable as the distributions of many types of data, which do not follow Gaussain distributions, can still be reasonably approximated by Gaussian density. Therefore, each noise-free observation in the data matrix is assumed to be a sample from a multivariate normal distribution, i.e.,

$$\tilde{\tilde{\mathbf{x}}}_{\mathrm{i}} = \left[\tilde{\tilde{\mathrm{x}}}_{\mathrm{i1}} \quad . \quad . \quad \tilde{\tilde{\mathrm{x}}}_{\mathrm{ir}}\right]^{\mathrm{T}} \sim MVN\left(\boldsymbol{\mu}_{\tilde{\tilde{\mathbf{X}}}}, \mathbf{Q}_{\tilde{\tilde{\mathbf{X}}}}\right), \text{ and } \mathrm{i} = 1,..., \mathrm{n}. \tag{29}$$

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 the eigenvectors of the sample covariance matrix are asymptotically multivariate normal, and that the eigenvalues are independent of the eigenvectors. The following asymptotic moments of each projection direction, $\tilde{\tilde{\boldsymbol{\alpha}}}_{\mathrm{j}}$, have also been presented

$$E\left[\tilde{\tilde{\boldsymbol{\alpha}}}_{\mathrm{j}}\right] = \boldsymbol{\alpha}_{\mathrm{j}} + O\left(\mathrm{n}^{-1}\right) \tag{30}$$

and, $Cov\left[\tilde{\boldsymbol{\alpha}}_j\right] = \frac{1}{n}\sum_{j\neq k}\frac{\lambda_j\lambda_k}{\left(\lambda_j-\lambda_k\right)^2}\boldsymbol{\alpha}_j\boldsymbol{\alpha}_k^T + O\left(n^{-2}\right)$ (31)

where the $\boldsymbol{\lambda}$'s and $\boldsymbol{\alpha}$'s are the eigenvalues and the eigenvectors of the matrix

$$E\left[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right] = \mathbf{Q}_{\tilde{\mathbf{X}}} + \boldsymbol{\mu}_{\tilde{\mathbf{X}}}\boldsymbol{\mu}_{\tilde{\mathbf{X}}}^T. \quad (32)$$

Illustrative Example

To illustrate Girshick's results, consider the projection directions matrix, $\boldsymbol{\alpha}$, of a data matrix $\tilde{\mathbf{X}}$ having two variables and 1000 observations, in which each observation is a sample from the following Gaussian distribution, $N\left(\boldsymbol{\mu}_{\tilde{\mathbf{X}}}, \mathbf{Q}_{\tilde{\mathbf{X}}}\right)$, where $\boldsymbol{\mu}_{\tilde{\mathbf{X}}} = \begin{bmatrix}2 & 1\end{bmatrix}^T$, and $\mathbf{Q}_{\tilde{\mathbf{X}}} = diag\begin{pmatrix}1 & 2\end{pmatrix}$. In this example, the $\boldsymbol{\alpha}$ matrix is of size (2×2), which can be written as,

$$\boldsymbol{\alpha} = \begin{bmatrix}\alpha_{11} & \alpha_{12}\\ \alpha_{21} & \alpha_{22}\end{bmatrix}. \quad (33)$$

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 of the matrix $\boldsymbol{\alpha}$ (α_{ij}) 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 mean and variances of the elements of the projection directions matrix is tabulated in Table 1, which compares the means and variances of the elements (α_{ij}), obtained by simulation and by Girshick's theorem. Since Girshick's results are asymptotic, the distribution of the projection directions tends towards normal as the number of observations increases.

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

For the density, $P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}\right)$, on the other hand, since the noise-free principal components, process variables, and projection directions are linearly related as $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\alpha}}$, and since $\tilde{\mathbf{X}}$ 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,

$$E\left[\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}\right] = E\left[\tilde{\mathbf{X}}\right]\tilde{\boldsymbol{\alpha}} = \boldsymbol{\mu}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}} \tag{34}$$

$$\text{and,} \quad Cov\left[\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}\right] = \tilde{\boldsymbol{\alpha}}^{\mathrm{T}} E\left[\left(\tilde{\mathbf{X}} - \boldsymbol{\mu}_{\tilde{\mathbf{X}}}\right)^{\mathrm{T}}\left(\tilde{\mathbf{X}} - \boldsymbol{\mu}_{\tilde{\mathbf{X}}}\right)\right]\tilde{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}}. \tag{35}$$

Therefore,

$$\mathrm{P}\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}\right) \sim MVN\left(\boldsymbol{\mu}_{\tilde{\mathbf{Z}}|\tilde{\boldsymbol{\alpha}}}, \mathbf{Q}_{\tilde{\mathbf{Z}}|\tilde{\boldsymbol{\alpha}}}\right) = MVN\left(\boldsymbol{\mu}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}}\right). \tag{36}$$

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,

$$\begin{aligned}
&\left\{\hat{\boldsymbol{\alpha}}\right\}_{\mathrm{MAP}} = \arg\max_{\tilde{\boldsymbol{\alpha}}} P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{a}}\right) \\
&\text{s.t.} \quad \left\{\tilde{\mathbf{Z}}\right\}_{\mathrm{MAP}} = \arg\max_{\tilde{\mathbf{Z}}} P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}\right) \\
&\qquad \tilde{\mathbf{X}} = \tilde{\boldsymbol{\alpha}}\tilde{\mathbf{Z}} \text{ and } \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} = \mathbf{I}.
\end{aligned} \tag{37}$$

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 as follows,

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

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

$$\left\{\hat{\mathbf{z}}_{\mathrm{i}}\right\}_{\mathrm{MAP}} = \left(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathrm{X}}}^{-1}\hat{\boldsymbol{\alpha}} + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1}\right)^{-1}\left(\hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\boldsymbol{\varepsilon}_{\mathrm{X}}}^{-1}\mathbf{x}_{\mathrm{i}} + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1}\boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}\right). \tag{39}$$

This BPCA algorithm reduces to MLPCA if the prior terms are set to zero, that is, when the prior is uniform. Due to the degeneracy of the distribution of $\tilde{\mathbf{a}}$, the covariance matrix, $\mathbf{Q}_{\tilde{\mathbf{a}}}$, is singular. One way to approximate it's 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, $\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} = \mathbf{I}$, 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,

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

However, the problem with this alternative is that the effective rank of the matrix, $\mathbf{Q}_{\tilde{\mathbf{a}}}$, is usually much less than the number of independent elements computed using equation (40) due to the nonlinearity of the orthonomality constraints, $\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} = \mathbf{I}$. Thus, reducing the dimension of the matrix, $\mathbf{Q}_{\tilde{\mathbf{a}}}$, 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, $P(\tilde{\mathbf{a}})$ and $P(\tilde{\mathbf{Z}}|\tilde{\boldsymbol{\alpha}})$, were assumed to be multivariate normal, and the parameters $\boldsymbol{\mu}_{\tilde{\mathbf{x}}}$, $\mathbf{Q}_{\tilde{\mathbf{x}}}$, $\boldsymbol{\mu}_{\tilde{\mathbf{a}}}$, and $\mathbf{Q}_{\tilde{\mathbf{a}}}$ (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 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, and 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 a more demanding task than the

parametric approach.[22] 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 due 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, $\{\boldsymbol{\mu}_{\tilde{\mathbf{x}}}$, $\mathbf{Q}_{\tilde{\mathbf{x}}}$, $\boldsymbol{\mu}_{\tilde{\mathbf{a}}}$, and $\mathbf{Q}_{\tilde{\mathbf{a}}}\}$, by $\boldsymbol{\eta}$, the posterior for this EBPCA problem becomes,

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \mathbf{X}, \boldsymbol{\eta}\right) = \frac{P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \boldsymbol{\eta}\right)}{P(\mathbf{X})}. \tag{41}$$

Now, the prior is dependent on the set of hyperparameters, $\boldsymbol{\eta}$, which is unknown. When these hyperparameters are actually known, $\boldsymbol{\eta}$ 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 the hyperparameters, $\boldsymbol{\eta}$, 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:

I. Solve for the MLPCA model using the available data.

II. Use the MLPCA solution to estimate the set of hyperparameters, $\hat{\boldsymbol{\eta}}$, as follows:

1. Set $\hat{\boldsymbol{\mu}}_{\tilde{\mathbf{a}}} = \{\hat{\mathbf{a}}\}_{\mathrm{MLPCA}}$,
2. Solve for $\hat{\mathbf{Q}}_{\tilde{\mathbf{a}}}$ using equation (31),
3. Estimate $\hat{\boldsymbol{\mu}}_{\tilde{\mathbf{x}}}$ as $E\left[\{\hat{\mathbf{X}}\}_{\mathrm{MLPCA}}\right]$,
4. Estimate $\hat{\mathbf{Q}}_{\tilde{\mathbf{x}}}$ as $Cov\left[\{\hat{\mathbf{X}}\}_{\mathrm{MLPCA}}\right]$.

Now, the prior is defined in terms of the set, $\hat{\boldsymbol{\eta}}$, as $P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \hat{\boldsymbol{\eta}}\right)$.

III. Solve the BPCA problem using the following posterior,

$$P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \mathbf{X}, \hat{\boldsymbol{\eta}}\right) = \frac{P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right) P\left(\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} \mid \hat{\boldsymbol{\eta}}\right)}{P(\mathbf{X})}. \tag{42}$$

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,

$$\left\{\hat{\mathbf{Z}}, \hat{\boldsymbol{\alpha}}, \hat{\mathrm{p}}\right\}_{\text{Bayesian}} = \underset{\tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}}{\arg\max} \left\{ \ln\left(P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right)\right) + \ln\left(P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}}\right)\right) + \ln\left(P\left(\tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}}\right)\right) + \ln\left(P\left(\tilde{\mathrm{p}}\right)\right) \right\} \tag{43}$$

Since $P\left(\tilde{\mathrm{p}} = \mathrm{j}\right) = \mathrm{k}_{\mathrm{j}}$ (see Equation (18)), then the posterior natural logarithm at a particular model rank, j, denoted by $LogPost_{\mathrm{j}}$, can be written as

$$LogPost_{\mathrm{j}} = \ln\left(P\left(\mathbf{X} \mid \tilde{\mathbf{Z}}, \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} = \mathrm{j}\right)\right) + \ln\left(P\left(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}, \tilde{\mathrm{p}} = \mathrm{j}\right)\right) + \ln\left(P\left(\tilde{\boldsymbol{\alpha}} \mid \tilde{\mathrm{p}} = \mathrm{j}\right)\right) + \ln\left(\mathrm{k}_{\mathrm{j}}\right). \tag{44}$$

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.,

$$\left\{\hat{\mathrm{p}}\right\}_{\text{Bayesian}} = \arg\max_{\tilde{\mathrm{p}}} \left\{ LogPost_1, LogPost_2, ..., LogPost_{\mathrm{r}} \right\} \tag{45}$$

where, $LogPost_{\mathrm{j}}$, 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 MLPCA model.[4] This approach is based on the fact that the sum of square approximation errors obtained in MLPCA

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

should follow a chi-square 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 of 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:

I. For each possible model dimensions (j), solve for the MLPCA model and compute the corresponding value S_j.

II. For each value S_j, compute the probability (Pr_j) of realizing a lower value than S_j using a chi-square distribution with the appropriate degree of freedom, $(n - j)(r - j)$.

III. 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 rank of two can be written in terms of the two independent variables as follows,

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

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,

$$\begin{bmatrix} \mathbf{I}_p \\ \hat{\mathbf{a}} \end{bmatrix} = \hat{\boldsymbol{\alpha}} \underbrace{\left(\hat{\boldsymbol{\alpha}}_U\right)^{-1}}_{p \times p}, \quad \text{where, } \hat{\boldsymbol{\alpha}} = \underbrace{\begin{bmatrix} \hat{\boldsymbol{\alpha}}_U \\ \hat{\boldsymbol{\alpha}}_L \end{bmatrix}}_{r \times p}. \tag{48}$$

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 follows,[4]

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

5.1 Stationary Gaussian data contaminated by white noise

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

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

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

$$\mathbf{Q}_{\varepsilon_X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \tag{52}$$

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, that is,

$$\mathrm{P}(\tilde{\mathbf{Z}} \mid \tilde{\boldsymbol{\alpha}}) \sim MVN\left(\boldsymbol{\mu}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{Q}_{\tilde{\mathbf{X}}}\tilde{\boldsymbol{\alpha}}\right) \tag{53}$$

where,

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

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 2 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 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 two*. 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 should, as shown next.

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 ten variables and fifty observations, are used to derive EBPCA models. The rank of the data sets ranges from one to nine. The noise free data of these data sets are generated as follows: each of the first p noise-free variables in the p^{th} data set, which is of rank p, is generated from the following Gaussian distribution,

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

Then, the last $(m - p)$ variables are generated by multiplying the first p variables by a $p \times (r - p)$ matrix each of its 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 show that the performance of EBPCA improves at higher model ranks. The percent improvement shown in Figure 3 is computed as follows,

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

Even when there is a 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 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 in average EBPCA results in a smaller parameter MSE than MLPCA.

Estimating the PCA model rank

The results reported in Table 2 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 percent accuracy reported in Figure 5 is computed as follows,

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

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 6c,d. The likelihood increases because the likelihood function is an exponential function of the negative data mean squared 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 squared 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 can not 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 can not be used in this regard. This behavior of the posterior is illustrated in Figure 6b.

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 components are above the 95% confidence line and 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 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]

$$\begin{bmatrix} 0 & 0 & 1 & -1 & 0 \\ 1 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{F}_1 \\ \mathrm{F}_2 \\ \mathrm{F}_3 \\ \mathrm{F}_4 \\ \mathrm{F}_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{58}$$

The data matrix, which consists of 5 variables and 50 observations, is generated as follows. The first two noise-free variables, $\mathbf{F}_1$ and $\mathbf{F}_2$, 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 following covariance matrix, $\mathbf{Q}_{\varepsilon_x} = diag(1 \quad 9 \quad 16 \quad 16 \quad 1)$.

The results of a Monte Carlo simulation of 100 realizations with known model rank are summarized in Table 3. 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 is consistent with the results of the previous example for model rank of less than

three. 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 is 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 one, even though the true mathematical rank of the noise-free data is two. This is due to the large noise content on one of the independent variable, F_2, which makes it effectively like noise in the data.

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 estimated model rank is one, which agrees with the earlier observation that retaining one principal component results in the least mean squared error.

5.3 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 following dynamic model,

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

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

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, $\tilde{X}$, which contain 64 observations, is constructed as follows,

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

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

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

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

The results of a Monte Carlo simulation of 100 realizations summarized in Table 4 show a clear advantage of EBPCA over both, PCA and MLPCA. These results are obtained under the assumption that the true model rank of two 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.

5.4 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 6 variables and 50 observations representing temperature measurements from 6 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 of 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 following covariance matrix, $\mathbf{Q}_{\varepsilon_X} = diag(0.05 \quad 0.1 \quad 0.05 \quad 0.1 \quad 0.05 \quad 0.1)$. A Monte Carlo simulation is performed assuming that the actual model rank is 3, and the results are summarized in Table 5. 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 number of components is shown in Figure 12. The smaller variation of the plots for EBPCA for different number of selected components indicates that EBPCA is more robust to errors in estimating the model rank. This is due to the fact when the model dimension in overestimated, the data mean squared 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 squared 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 three for most realizations, as portrayed in Figure 13.

5.5 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 4 variables representing the absorption of 4 solutions of 1-fluoro-3-nitrobenzene and dimethy phthalate at 35 wavelengths in the range of 215-385.[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 one. The results in Table 6 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 number 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 one 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 is already noisy with 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 in 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 accuracy of both, the estimated parameters and measurements. The formulation of the BPCA approach is shown to be more general than existing methods and reduces 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 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 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 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

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 maximum likelihood PCA data reconciliation problem can be formulated as follows:

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

$$\text{s.t.} \quad \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i . \tag{A1.1}$$

Solution:

Define the Lagrange function as,

$$L = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \boldsymbol{\lambda}(\hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) \tag{A1.2}$$

Taking the partial derivatives of L with respect to $\hat{\mathbf{x}}_i$, $\hat{\mathbf{z}}_i$, and $\boldsymbol{\lambda}$, and setting them to zeros,

$$\frac{\partial L}{\partial \hat{\mathbf{x}}_i} = -2\mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \boldsymbol{\lambda}^T = 0 \tag{A1.3}$$

$$\frac{\partial L}{\partial \hat{\mathbf{z}}_i} = -\boldsymbol{\lambda} \hat{\boldsymbol{\alpha}} = 0 \tag{A1.4}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i = 0 . \tag{A1.5}$$

Substituting equation A1.3 in A1.4, get

$$\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = 0 . \tag{A1.6}$$

Substituting equation A1.5 in A1.6, get

$$\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_x}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) = 0 \tag{A1.7}$$

Rearranging equation A1.7, get the MLPCA solution

$$\{\hat{\mathbf{z}}_i\}_{\text{MLPCA}} = \left(\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} \hat{\boldsymbol{\alpha}}\right)^{-1} \hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} \mathbf{x}_i \tag{A1.8}$$

Appendix II

Derivation of the BPCA Data Rectification Solution

The BPCA data reconciliation problem can be formulated as follows:

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

$$\text{s.t. } \hat{\mathbf{x}}_i = \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i . \tag{A2.1}$$

Solution:

Define the Lagrange function as,

$$L = (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}})^T \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) + \boldsymbol{\lambda}(\hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) \tag{A2.2}$$

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

$$\frac{\partial L}{\partial \hat{\mathbf{x}}_i} = -2\mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) + \boldsymbol{\lambda}^T = 0 \tag{A2.3}$$

$$\frac{\partial L}{\partial \hat{\mathbf{z}}_i} = 2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - \hat{\boldsymbol{\alpha}}^T \boldsymbol{\lambda}^T = 0 \tag{A2.4}$$

$$\frac{\partial L}{\partial \boldsymbol{\lambda}} = \hat{\mathbf{x}}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i = 0 . \tag{A2.5}$$

Substituting equation A2.3 in A2.4, get

$$2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - 2\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) = 0 . \tag{A2.6}$$

Substituting equation A2.5 in A2.6, get

$$2\mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} (\hat{\mathbf{z}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}) - 2\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\alpha}} \hat{\mathbf{z}}_i) = 0 \tag{A2.7}$$

Rearranging A.7, get the MAP solution

$$\{\hat{\mathbf{z}}_i\}_{\text{MAP}} = \left(\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} \hat{\boldsymbol{\alpha}} + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1}\right)^{-1} \left(\hat{\boldsymbol{\alpha}}^T \mathbf{Q}_{\varepsilon_X}^{-1} \mathbf{x}_i + \mathbf{Q}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}^{-1} \boldsymbol{\mu}_{\tilde{\mathbf{z}}|\tilde{\boldsymbol{\alpha}}}\right). \tag{A2.8}$$

References

1. J. V. Kresta, J. F. MacGregor, and T. E. Marlin, *Can. J. Chem. Eng.*, 69, 35-47 (1991).
2. B. M. Wise, N. L. Ricker, D. F. Veltkamp, and B. R. Kowalski, *Proc. Cont. Qual.*, 1, 41 (1990).
3. M. A. Kramer and R. S. H. Mah, *Proc. Int. Conf. On Foundations of Computer Aided Process Operations*, D. Rippin, J. Hale, J. Davis, eds. CACHE (1994).
4. P. D. Wentzell, D. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, *J. of Chemometrics*, 11, 339-366 (1997).
5. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlar, New York (1985).
6. M. West, and J. Harrison, *Forecasting and Dynamic Models*", Springer, New York (1997).
7. G. A. E. Seber, *Multivariate Observations*, Wiley, New York (1984).
8. Press, S. James, *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, second edition, Robert E. Krieger Publishing Company, Florida (1982).
9. S. J. Press, *Applied Multivariate Analysis*, New York: Holt, Rinehart and Winston, Inc. (1972).
10. J. K. Martin and R. P. McDonald, *Psychometrika*, 40, 4, 505-517 (1975).
11. S. J. Press and K. Shigemasu, *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, 271-278 (1989).
12. S. E. Lee and S. J. Press, *Commun. Stat.-Theory Meth.*, 27, 8, 1871-1893 (1998).
13. S. Wold, *Chemometrics and Intelligent Laboratory Systems*, 23, 149-161 (1994).
14. B. R. Bakshi, *AIChE Journal*, 44, 7, 1596-1610 (1998).
15. T. J. Hastie and W. Stuetzle, *J. of American Statistical Association*, 84, 406, 505-516 (1989).
16. R. S. H. Mah, *Chemical Process Structures and Information Flows*, Butterworths, Boston (1990).
17. J. B. Kadane, *Controlled Clinical Trials*, 16, 313-318 (1995).
18. A. Gelman, J. B. Carlin, H. S. Stern, and D. Rubin, *Bayesian Data Analysis*, Chapman and Hall, London (1995).
19. W. R. Gilks, S. Richardson, and D. Spiegelhalter eds., *Practical Markov Chain Monte Carlo*, Chapman And Hall, New York (1996).

20. C. P. Robert, *The Bayesian Choice: A Decision Theoretic Motivation*, Springer-Verlag, New York (1994).
21. M. A. Girshick, *Ann. Math. Stat.*, 10, 203-224 (1939).
22. J.S. Maritz, *Empirical Bayes Methods*, Methuen & CO., London (1970).
23. B. B. Carlin and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, First edition, Monographs on Statistics and Applied Probability 69, Chapman & Hall (1996).
24. W. Ku, R.H. Storer, and C. Georgakis, *Chemometrics and Intellegent Laboratory Systems*, 30, 179-196 (1995).
25. A. Basilevsky, *Statistical Factor Analysis and Related Methods: Theory and Applications*, Wiley Series in Probability and Mathematical Statistics, New York (1994).
26. S. Wold, *Technometrics*, 20, 4, 397-405 (1978).
27. H. T. Eastment and W. J. Krzanowski, *Technometrics*, 24,1, 73-77 (1982).
28. W. James and C. Stein, *Proceedings of the Fourth Berkeley Symposium on Mathematics and Statistics*, Berkeley: University of California Press 1, 361-379 (1961).
29. M. H. Gruber, *Improving efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*, Marcel Dekker, New York (1998).
30. L. Johnston and M. Kramer, *AIChE Journal*, 41, 11 (1995).
31. M. Kano, K. Miyazaki, S. Hasebe, and I. Hashimoto, *J. Process Control*, 10, 157-166 (2000).
32. C. H. Luchmuller and C. E. Reese, *Critical Reviews In Analytical Chemistry*, 28,1, 21-49 (1998).
33. M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen, *AIChE Journal*, accepted (2002)
34. M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen, *Industrial and Engineering Chemistry Research*, 40, 1, 261 –274 (2001)
35. W.-S. Chen, B. R. Bakshi, P. K. Goel, and S. Ungarala, *Technical Report*, Ohio State University (2002).
36. D. Malakoff, *Science*, 286, 1460 (1999).

Table 1. Comparison of the mean and variances of the elements of the projection directions matrix obtained using a Monte Carlo simulation and Girshick's theorem.

Element	Mean		Variance $\times 10^{-4}$	
	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

Table 2. PCA modeling and rectification of stationary Gaussian noise-free data (Example 5.1), 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

Table 3. PCA modeling and rectification of steady state reactor data (Example 5.2).

	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

Table 4. Rectification of non-stationary dynamic data using dynamic PCA (Example 5.3).

MSE	PCA	MLPCA	EBPCA
$Y(k-1)$	3.26	2.77	2.46
$U(k)$	1.49	1.59	1.07
$Y(k)$	2.70	2.18	2.04
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

Table 5. Data filtering of temperature data from a distillation column (Example 5.4).

	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

Table 6. PCA filtering of UV absorption data (Example 5.5).

	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

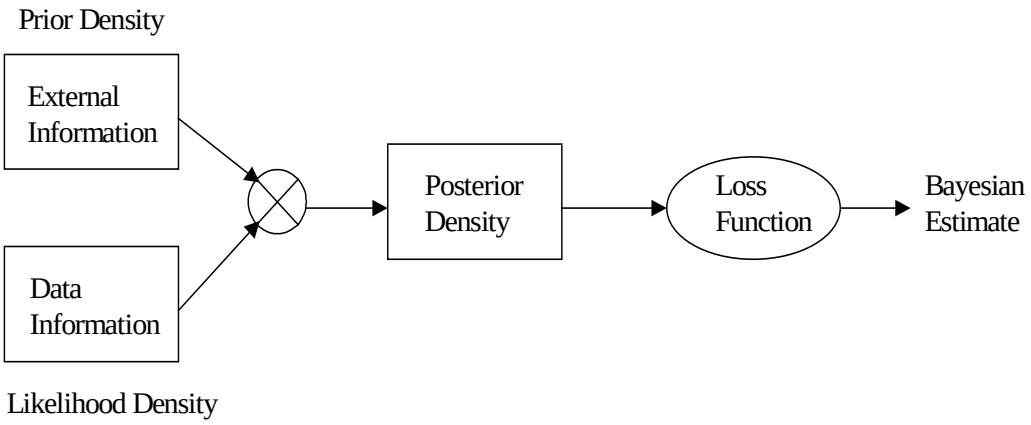

Figure 1. A schematic diagram of the main steps in Bayesian estimation.

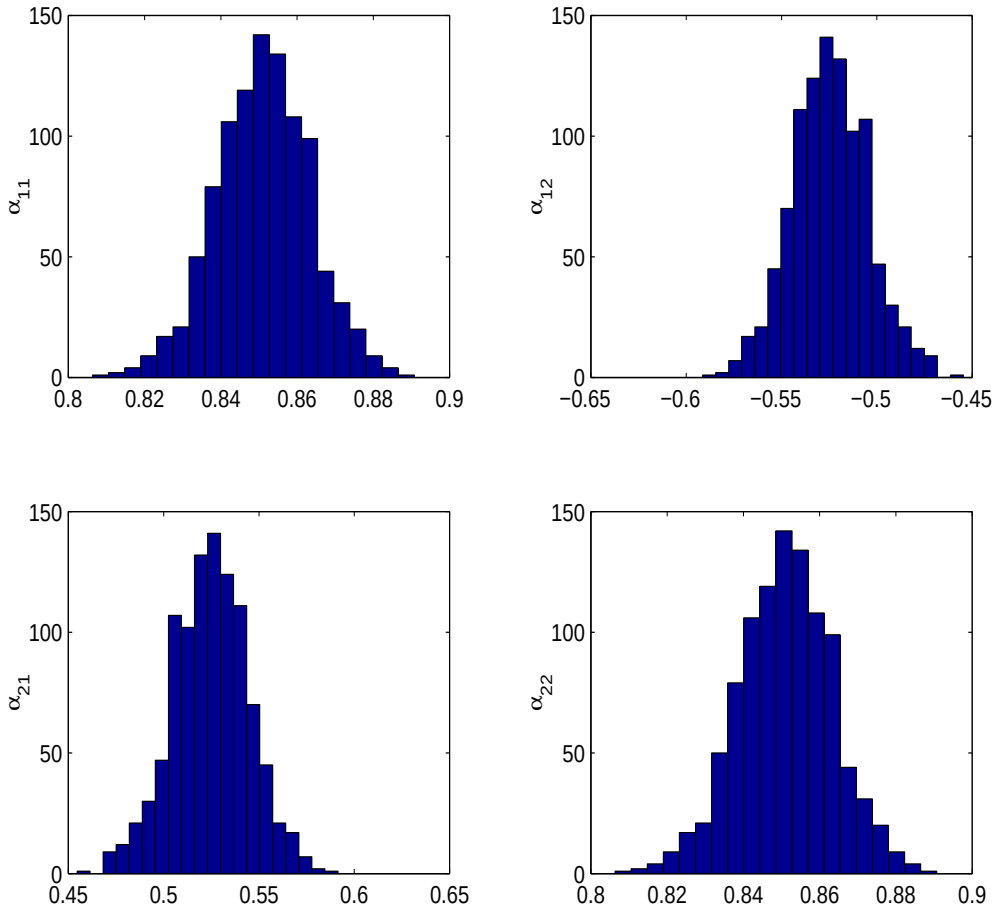

Figure 2. Histograms of the elements of the projection directions for Gaussian data. Gaussian distribution confirms Girschick's results.[21]

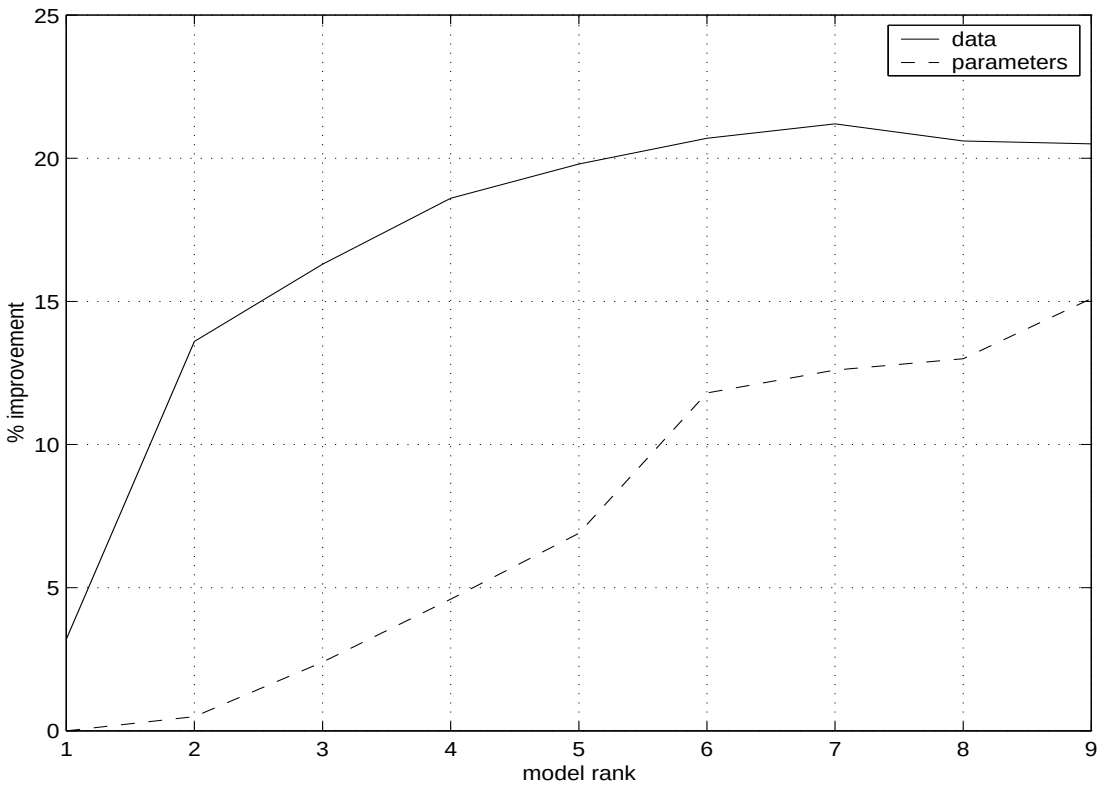

Figure 3. Percent improvement achieved by EBPCA over MLPCA versus model rank. Improvement in parameter estimates is significant only for rank greater than two.

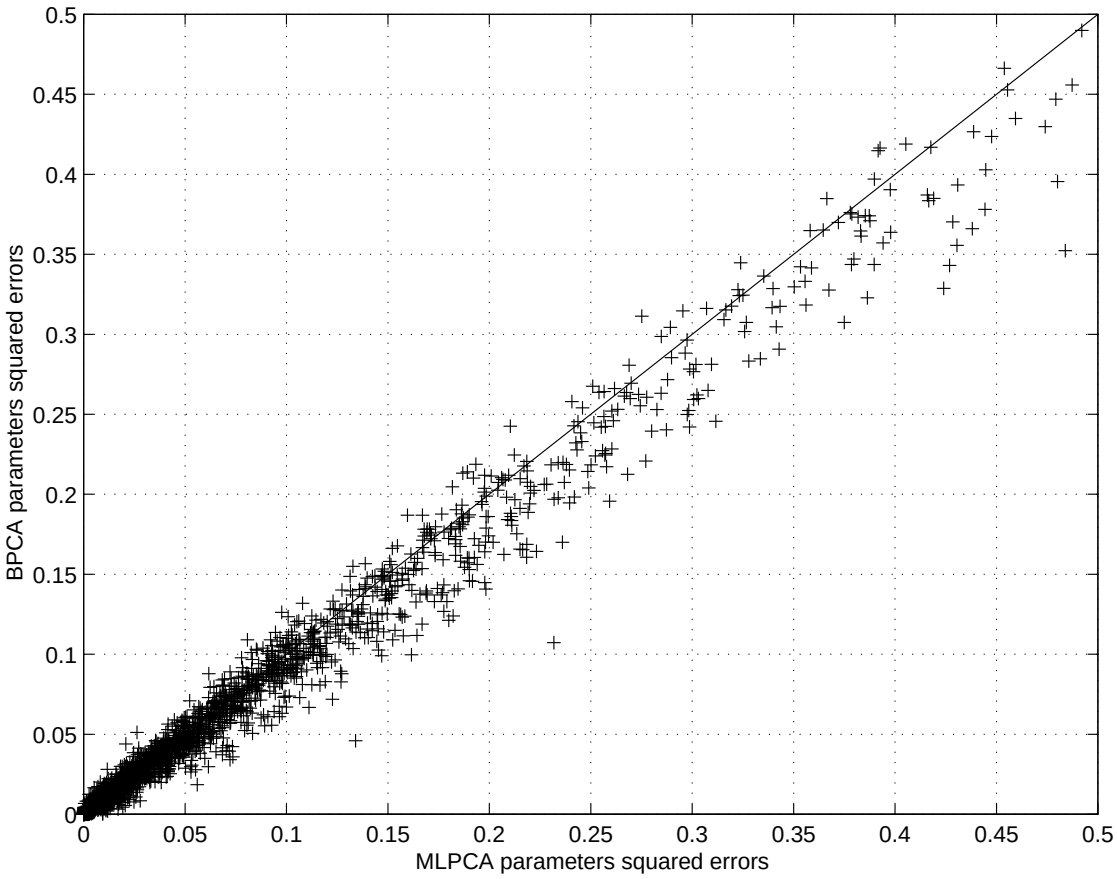

Figure 4. Comparison of the parameter squared errors obtained by EBPCA and MLPCA. Diagonal line represents equal MLPCA and EBPCA errors. Points below diagonal indicate better performance of BPCA.

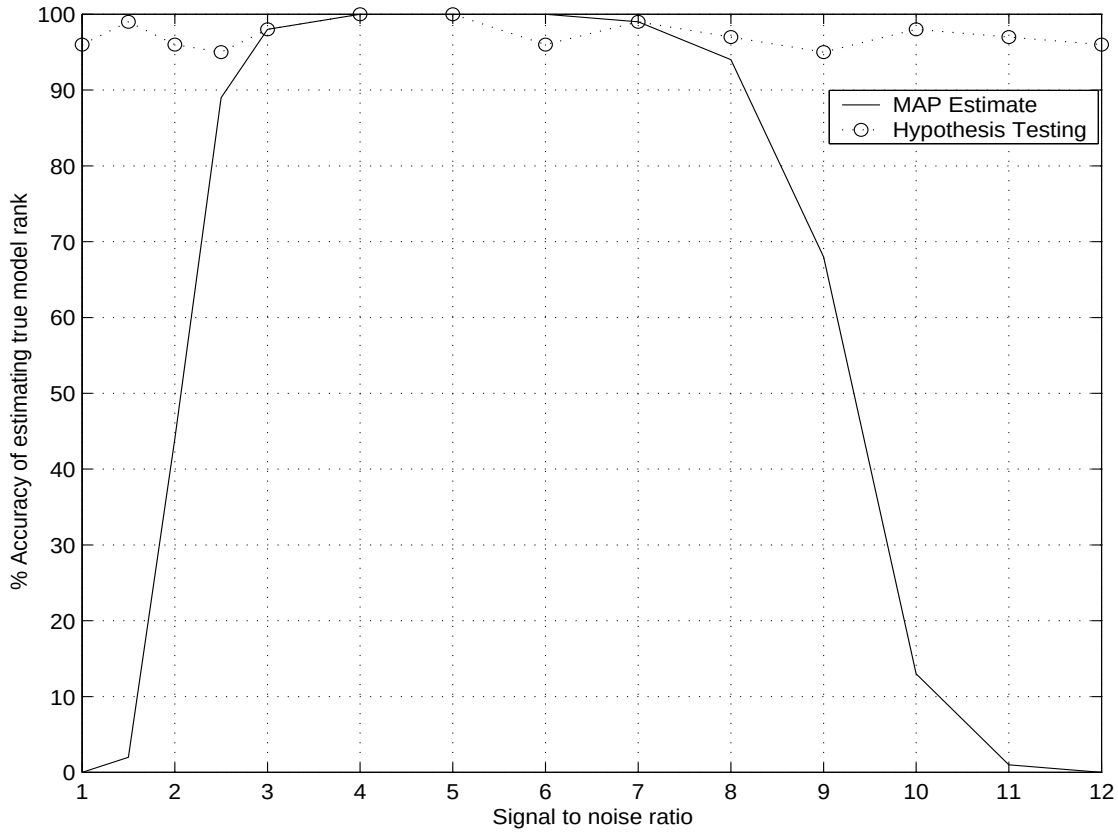

Figure 5. Percent accuracy in estimating the model rank for the Gaussian data in Example 5.1.

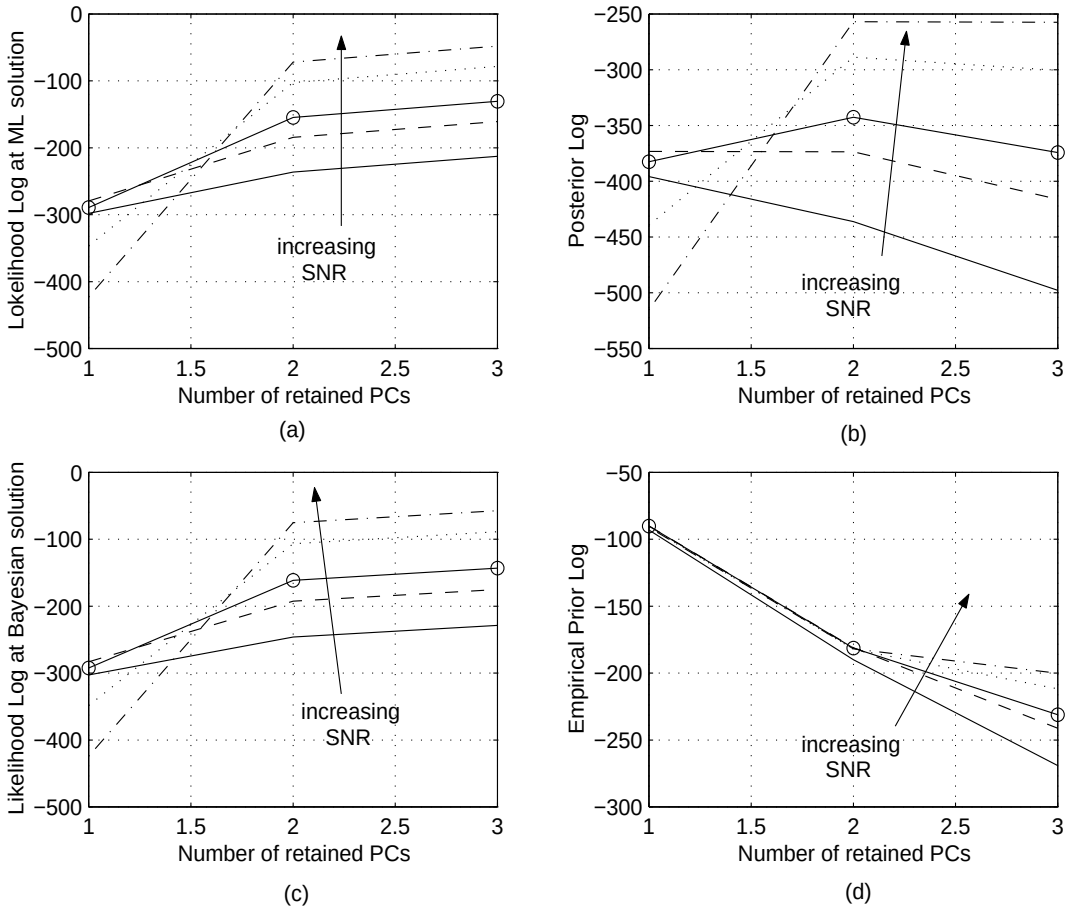

Figure 6. Performance of MAP method for estimating model rank for Example 5.1. (a) Logarithm of the likelihood function at the MLPCA solution, (b) Logarithm of the posterior density, (c) Logarithm of the likelihood function at the EBPCA solution, and (d) Logarithm of the prior density. Each plot versus different numbers of retained principal components and at different signal-to-noise ratios (SNR values are 1,3,6,9,12).

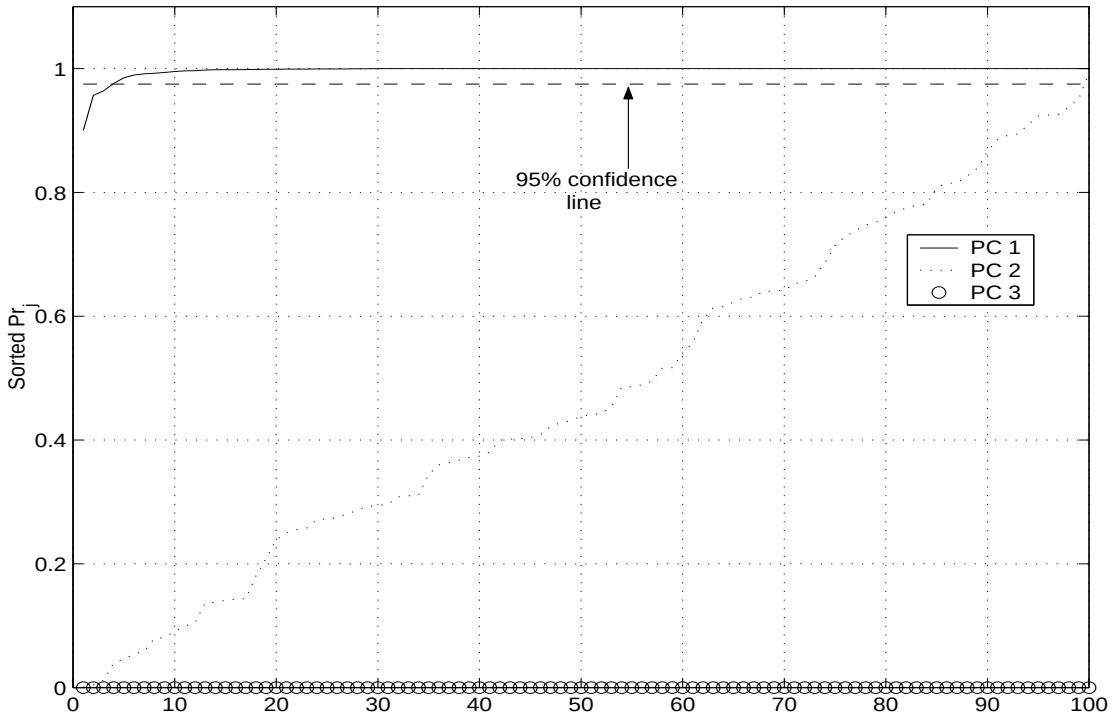

Figure 7. The sorted probabilities, Pr_j, for 100 realizations of hypothesis testing for Example 5.1. The x-axis is the index of the sorted probabilities.

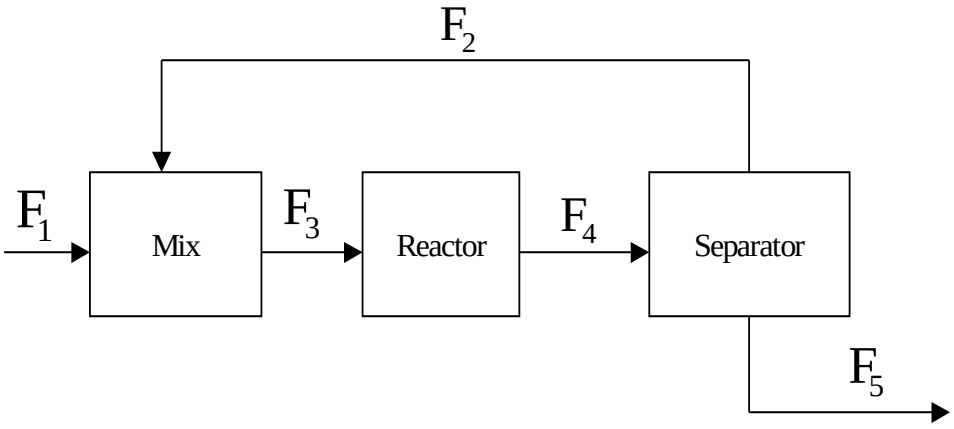

Figure 8. Flowsheet for Example 5.2.

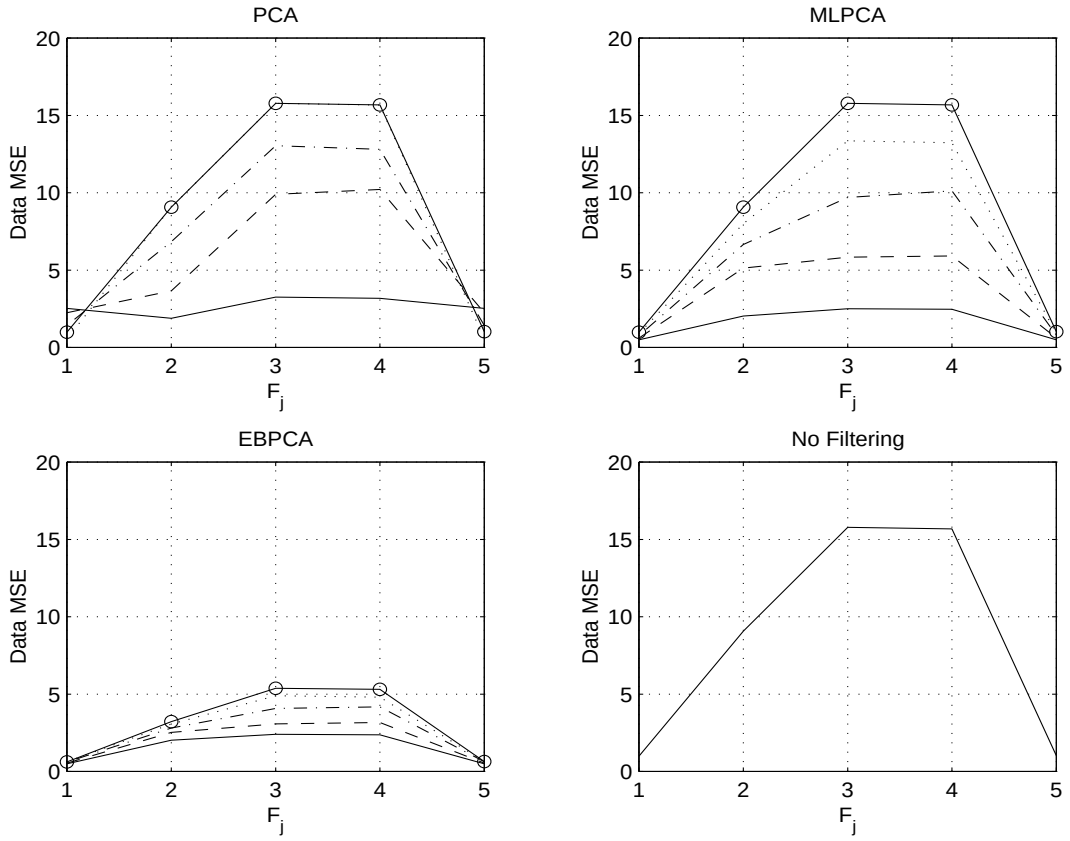

Figure 9. Data MSE versus flowrates (F_j) for different number of principal components for Example 5.2. True model rank is two. Narrower range of variation for EBPCA indicates greater robustness to errors in estimating the model rank. Legend - solid line: 1 PC; dashed line: 2 PC's; dash-dot line: 3 PC's; dot line: 4 PC's; solid-circle line: 5 PC's.

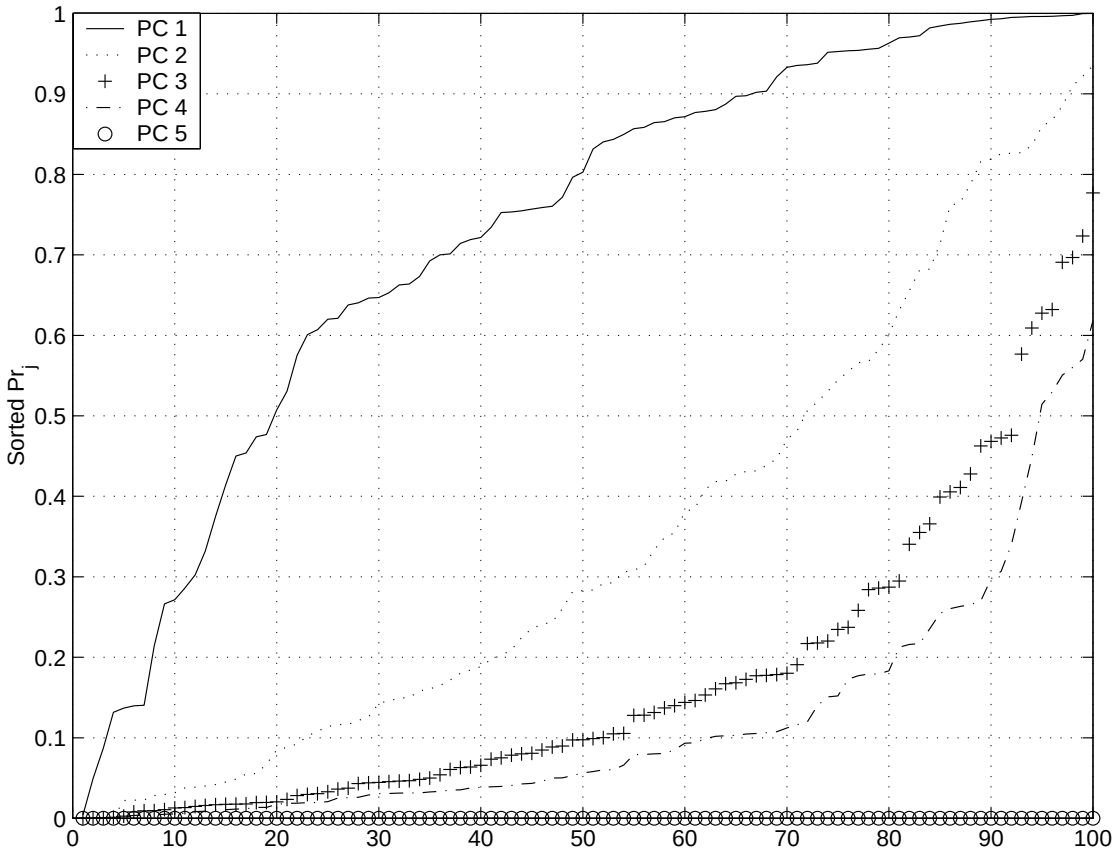

Figure 10. The sorted probabilities, Pr_j, for hypothesis testing for Example 5.2. The x-axis is the index of the sorted probabilities.

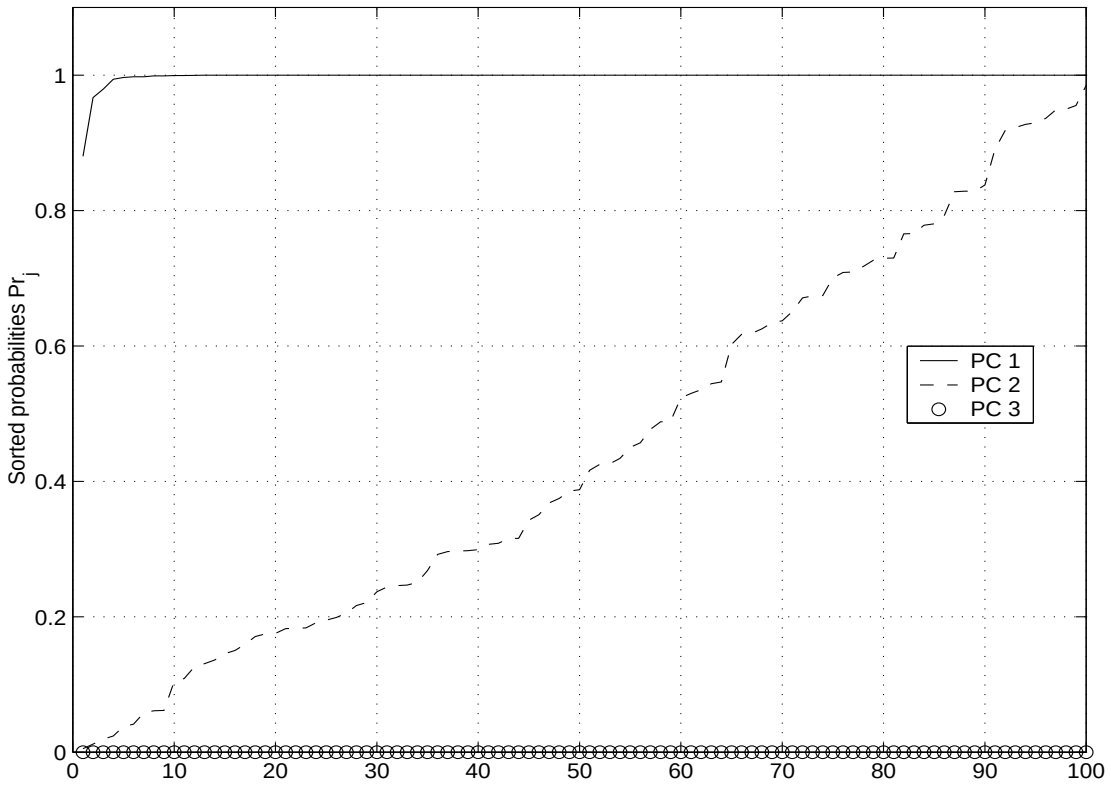

Figure 11. The sorted probabilities, Pr_j, for hypothesis testing for Example 5.3. The x-axis is the index of the sorted probabilities.

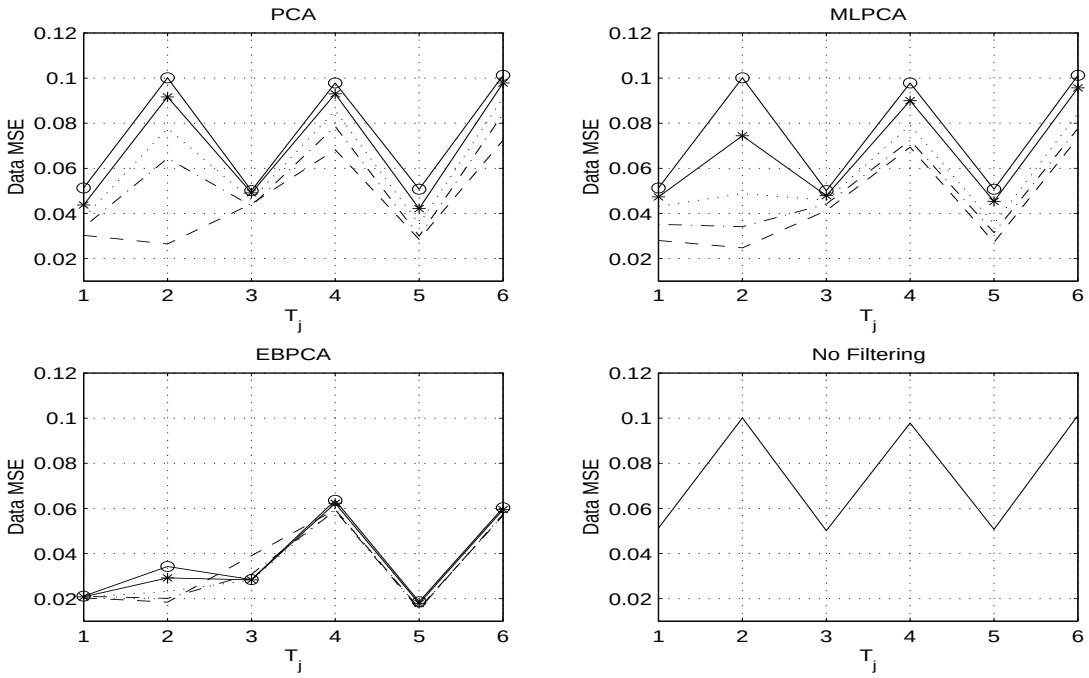

Figure 12. Data mean square errors versus temperatures (T_j) obtained by the various techniques for different numbers of retained principal components for Example 5.4. Legend - dashed line: 2 PC; dash-dot line: 3 PC's; dot-line: 4 PC's; stars: 5 PC's; circles: 6 PC's.

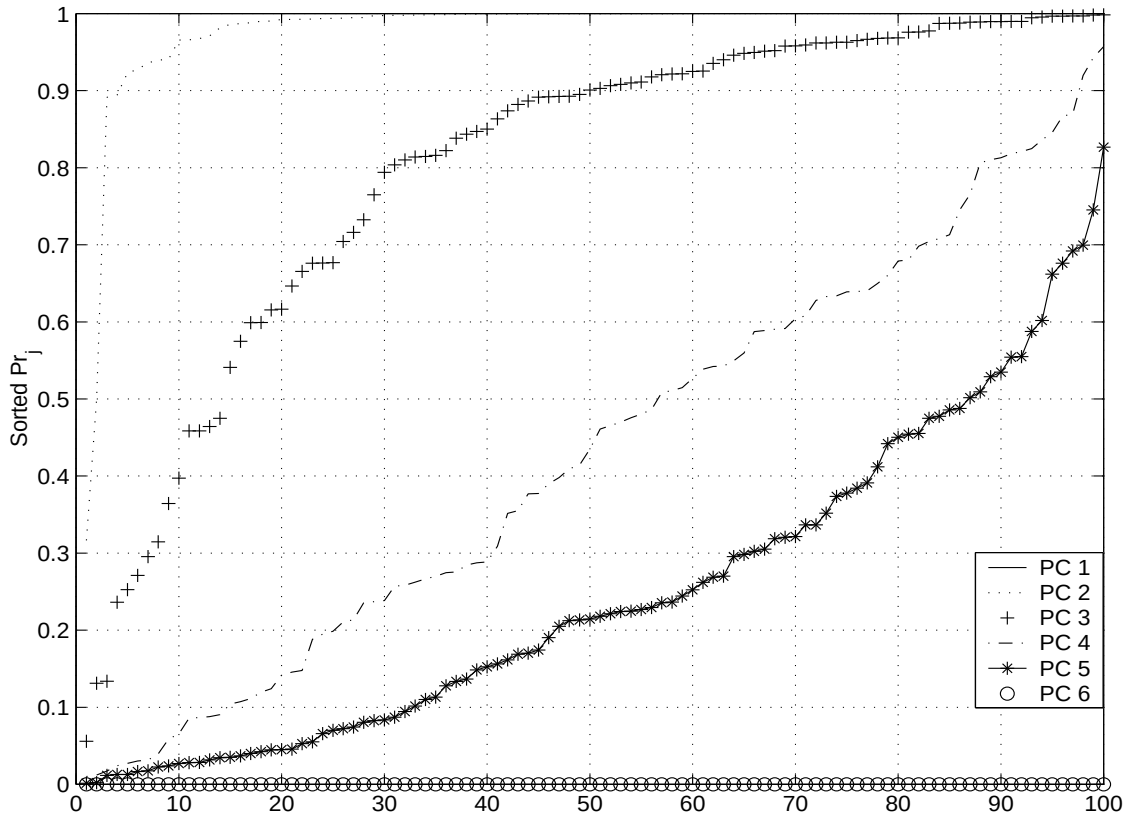

Figure 13. The sorted probabilities, Pr_j, from hypothesis testing for Example 5.4. The x-axis is the index of the sorted probabilities.

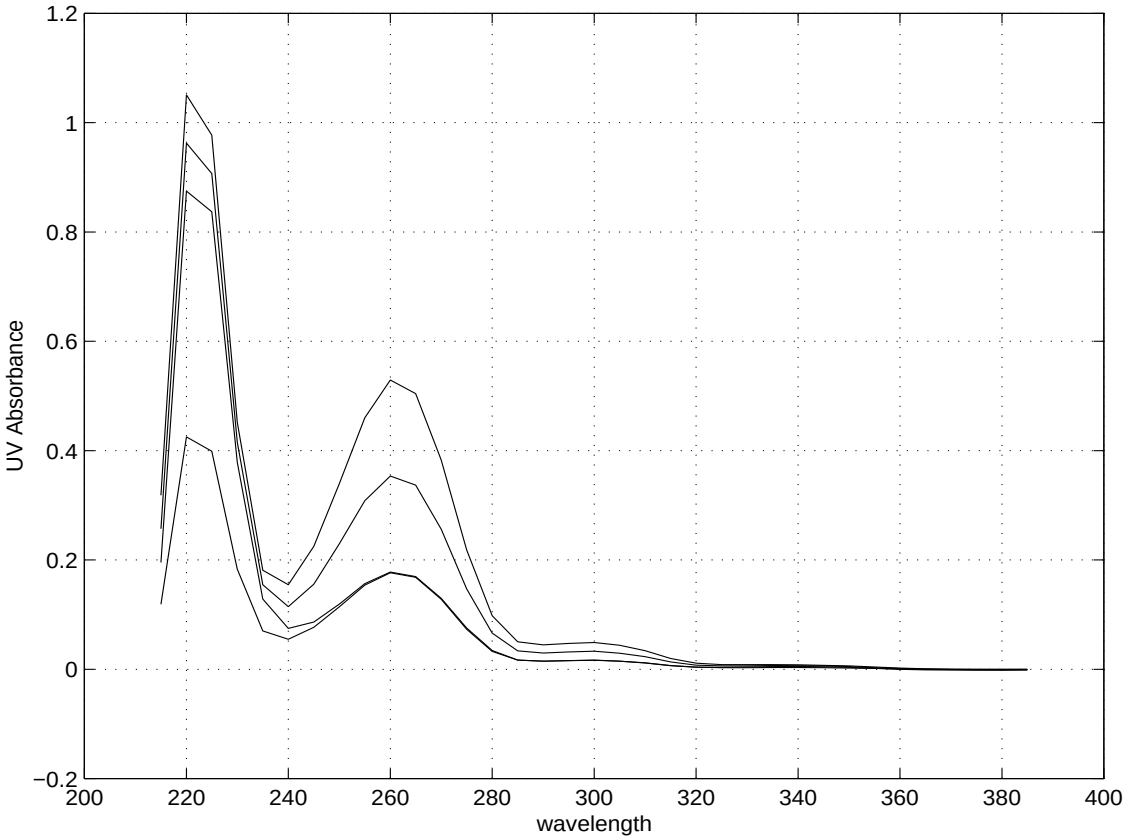

Figure 14. UV absorption data used in Example 5.5 for the four solutions vs. wavelength.

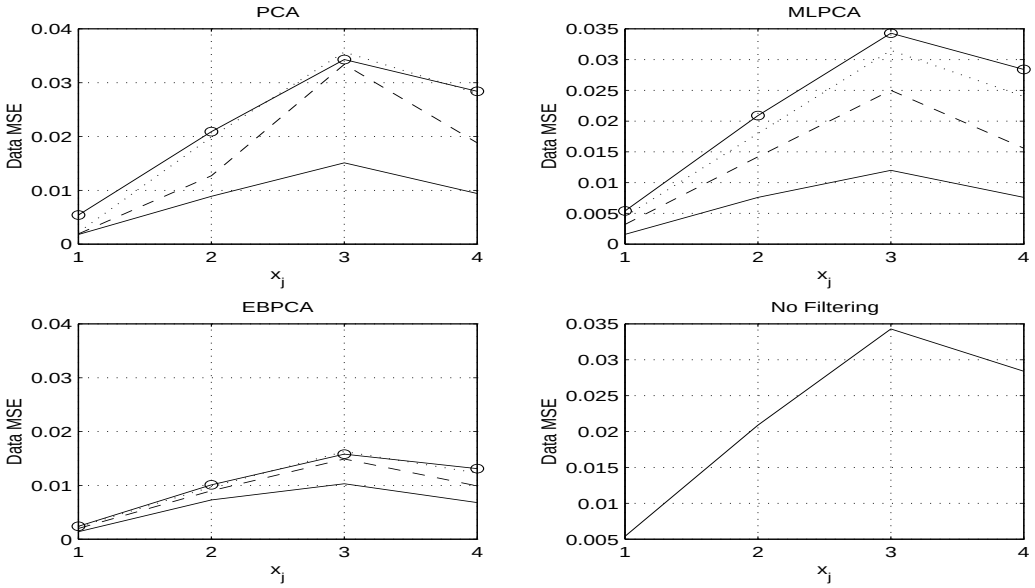

Figure 15. Data mean square errors versus variables (x_j) for various techniques at different numbers of retained principal components for Example 5.5. Legend - solid line: 1 PC; dashed line: 2 PC's; dash-dot line: 3 PC's; solid-circle line: 4 PC's.

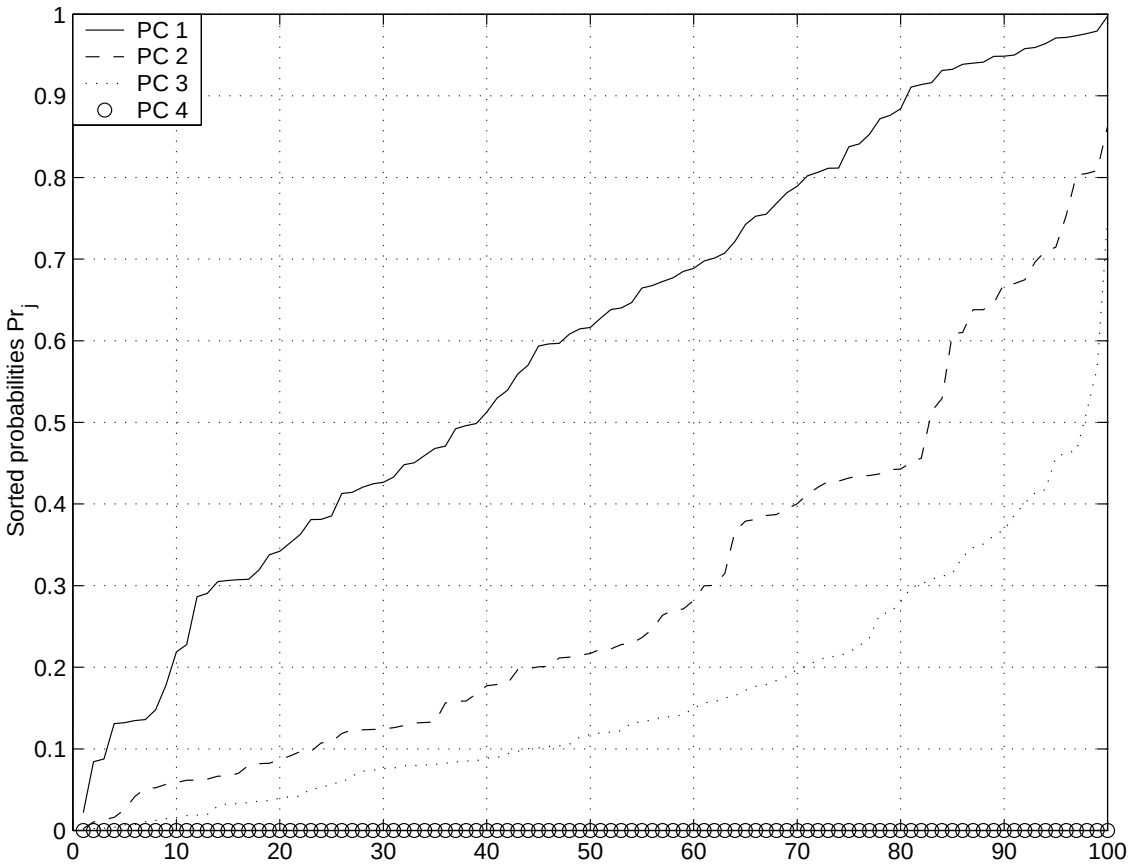

Figure 16. The sorted probabilities, Pr_j, from hypothesis testing for Example 5.5. The x-axis is the index of the sorted probabilities.

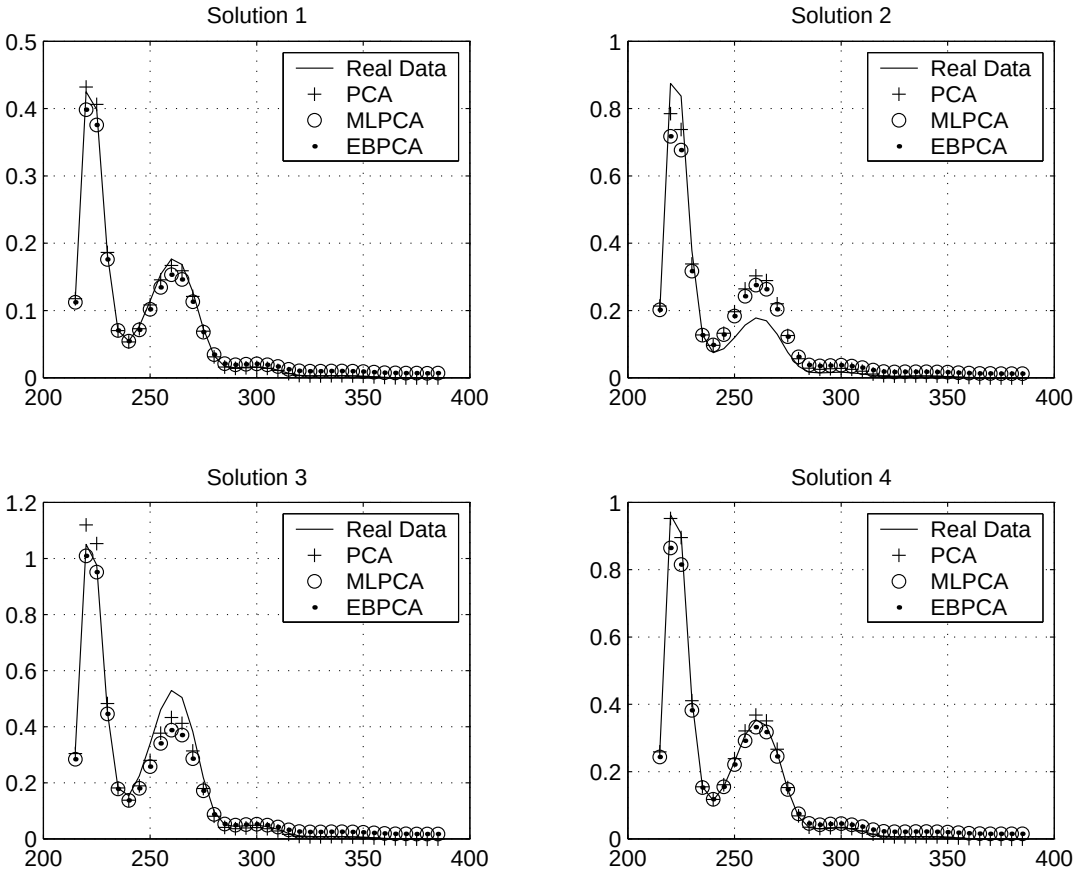

Figure 17. Comparison of PCA, MLPCA, and EBPCA using real UV absorption data for Example 5.5.