Proceedings of the American Control Conference
Arlington, VA June 25-27, 2001

Improving Principal Component Analysis using Bayesian Estimation

Mohamed N. Nounou, Bhavik R. Bakshi
Department of Chemical Engineering
Prem K. Goel, Xiaotong Shen
Department of Statistics
The Ohio State University, Columbus, OH. 43210

Abstract

Bayesian estimation is used in this paper to derive a new PCA modeling algorithm that improves the estimation accuracy by incorporating prior knowledge about the data and model. It is shown that the algorithm is more general than existing methods, PCA and MLPCA, and reduces to these techniques when a uniform prior is used. It is also shown that when no external information is available, an empirically estimated prior from the available data can still provide improved accuracy over non-Bayesian methods.

1 Introduction

Principal Component Analysis (PCA) is a popular modeling technique used to summarize a set of process variables. PCA has been found useful in many applications, such as process monitoring (Kresta et. al., 1991; Wise, et. al., 1990) and data rectification (Kramer and Mah, 1994). PCA transforms the process variables by rotating their axes of representation to capture the variations in the data in a smaller number of transformed variables. A drawback of PCA is that it assumes equal error contribution in all variables. This drawback is accounted for in Maximum Likelihood PCA (MLPCA), (Basilevsky, 1994).

In practice, however, more information about the noise-free data and/or the PCA model is often available. Exploiting this information can greatly enhance the accuracy of the estimated model and data. Unfortunately, neither PCA nor MLPCA can accommodate such information. External information can be incorporated into the PCA modeling problem using a prior density within a Bayesian framework, in which all quantities, measured and unmeasured are considered random. Bayesian estimation combines the data information represented by the likelihood function and any external information represented by the prior density to improve the accuracy of the estimated PCA model.

In this paper, a Bayesian Principal Component Analysis (BPCA) modeling technique is developed to improve the performances of PCA and MLPCA when prior information about the noise-free data and/or PCA models is available. The BPCA approach is shown to be more general than PCA and MLPCA, and reduces to these methods when a uniform prior density is used.

The rest of this paper is organized as follows. A brief description of PCA and MLPCA, and an introduction to Bayesian estimation are presented in Section 2. Then, the Bayesian PCA algorithm is derived in Section 3, followed by a methodology for estimating the prior empirically. Then, in Section 4, the advantages of BPCA over existing methods are shown through illustrative steady state and dynamic examples. Finally, the paper in concluded with few remarks in Section 5.

2 Background

2.1 Principal Component Analysis (PCA)

Given a $n \times r$ matrix of measured process variables, $X = \tilde{X} + \varepsilon_X$, where $\tilde{X}$ is the noise-free data matrix, ε_X is the additive noise, r is

the number of variables, and n is the number of observations, PCA decomposes the matrix X as

$$X = Z\alpha^T \quad (1)$$

where Z is a $n \times r$ matrix of the principal components, and α is an orthonormal $r \times r$ matrix of the projection directions or rotated axes. The PCA estimation problem can be formulated as follows,

$$\{\hat{\alpha}, \hat{Z}\}_{PCA} = \underset{\hat{\alpha},\hat{z}}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \hat{x}_i)^T (x_i - \hat{x}_i) \quad (2a)$$

$$\text{s.t.} \quad \hat{x}_i = \hat{\alpha}\hat{z}_i, \text{ and } \hat{\alpha}^T\hat{\alpha} = I \quad (2b)$$

where x_i is the i^{th} measurement vector (of size $r \times 1$) and $\hat{x}_i$ is its estimate. This formulation of PCA shows that it assumes equal noise content in all variables. The solution to this optimization problem is the well known singular value decomposition of the matrix, X.

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 variation in the data, assuming that the remaining principal components capture the contaminating noise.

2.2 Maximum Likelihood PCA (MLPCA)

MLPCA extends the performance of PCA by accounting for different noise contributions in different variables. It estimates the model that maximizes the likelihood of estimating the true principal components and projection directions given the measured variables, which is the probability density function of the measurements given the true principal components, projection directions, and the true rank of the process variables "p", as

$$\{\hat{\alpha}, \hat{Z}\}_{MLPCA} = \underset{\tilde{\alpha},\tilde{Z}}{\arg\max}\, L(\tilde{\alpha}, \tilde{Z}, \tilde{p}; X) = \underset{\tilde{\alpha},\tilde{Z}}{\arg\max}\, P(X \mid \tilde{\alpha}, \tilde{Z}, \tilde{p}) \quad (3)$$

subject to the constraint given in (2b). If the distribution of the errors is assumed to be Gaussian, maximizing the likelihood is equivalent to minimizing the sum of square errors normalized by the noise covariance matrix, and thus the MLPCA solution can be obtained as follows,

$$\{\hat{\alpha}, \hat{Z}\}_{MLPCA} = \underset{\hat{\alpha},\hat{Z}}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) \quad (4)$$

where Q_{ε_x} is the noise covariance matrix, which is assumed to be known. This optimization problem requires an iterative procedure to be solved, which can be performed by solving two simultaneous optimization problems: one solves for the projection directions (parameter estimation), and the other solves for the principal components (data reconciliation) as follows

$$\{\hat{\alpha}\}_{MLPCA} = \underset{\hat{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i)$$

$$\text{s.t.} \quad \{\hat{z}_i\}_{MLPCA} = \underset{\hat{z}_i}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i - \hat{x}_i)^T Q_{\varepsilon_x}^{-1} (x_i - \hat{x}_i) \quad (5)$$

subject to the constraints shown in equation (2b). The data reconciliation problem has the following closed from solution (see Appendix I),

$$\{\hat{z}_i\}_{MLPCA} = (\hat{\alpha}^T Q_{\varepsilon_x}^{-1} \hat{\alpha})^{-1} \hat{\alpha}^T Q_{\varepsilon_x}^{-1} x_i. \quad (6)$$

Another MLPCA algorithm has also been developed by Wentzell et. al. (1997), which uses iterative least squares and singular value decomposition.

2.3 Bayesian Estimation

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 (Gelman et. al., 1995). Let's assume that it is desired to estimate the quantity $\tilde{\theta}$, form 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

$$P(\tilde{\theta} \mid y) \propto P(y \mid \tilde{\theta}) P(\tilde{\theta}). \quad (7)$$

The first term in equation (7) is the *likelihood* function, which contains the information brought by the observations, y, about the quantity, $\tilde{\theta}$. The second term, on the other hand, is the "*prior*", which quantifies our knowledge and belief about $\tilde{\theta}$ *before* observing the measurements, and through which external knowledge about the quantity $\tilde{\theta}$ can be incorporated into the estimation problem. Thus, the posterior density combines the data information and any external information to improve the estimation accuracy. A good description of Bayesian estimation is presented by Gelman et. al. (1995).

3 Bayesian PCA (BPCA)

3.1 General Formulation

In PCA, it is desired to estimate the noise-free projection directions, principal components, and true model rank. Therefore, the posterior is defined as the conditional density of these quantities given the measured variables, which can be written using Bayes rule as,

$$P(\tilde{Z},\tilde{\alpha},\tilde{p}\,|\,X)=\frac{P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z},\tilde{\alpha},\tilde{p})}{P(X)}. \quad (8)$$

The first term in the numerator is the likelihood function, which is maximized in MLPCA, and the second term is the prior density, which quantifies our prior knowledge about the true PCA model. The denominator, on the other hand, is the density of the measurements, which can be considered constant after collecting the data, and therefore,

$$P(\tilde{Z},\tilde{\alpha},\tilde{p}\,|\,X)\propto P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z},\tilde{\alpha},\tilde{p}). \quad (9)$$

The Prior Density Function

The prior, $P(\tilde{Z},\tilde{\alpha},\tilde{p})$, is a complicated joint density function. However, since the principal components and projection directions depend on the model rank, the prior can be written as

$$P(\tilde{Z},\tilde{\alpha},\tilde{p})=P(\tilde{Z},\tilde{\alpha}\,|\,\tilde{p})P(\tilde{p}) \quad (10)$$

which, using the multiplication rule of probabilities, can be expressed as,

$$P(\tilde{Z},\tilde{\alpha},\tilde{p})=P(\tilde{Z}\,|\,\tilde{\alpha},\tilde{p})P(\tilde{\alpha}\,|\,\tilde{p})P(\tilde{p}). \quad (11)$$

3.2 Simplifying Assumptions

Computing the posterior density requires computing the prior and the likelihood densities, which depend on the nature of the noise-free data and the contaminating noise, respectively. Therefore, assumptions need to be made to define the structure of these densities. The assumptions and their implications are described below:

I. Known True Rank

Knowing the model rank implies that $P(\tilde{p})=1$, which reduces the posterior to

$$P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z}|\tilde{\alpha})P(\tilde{\alpha}). \quad (12)$$

In practice, however, the model rank is usually not known. A method for estimating the model rank is described in (Nounou et. al., 2001).

II. Zero-One Loss Function

A 0-1 loss function is defined as,

$$L(\hat{Z},\hat{\alpha};\tilde{Z},\tilde{\alpha})=\begin{cases}0 & \text{when}\ \{\hat{Z},\hat{\alpha}\}_{\text{Bayesian}}=\tilde{Z},\tilde{\alpha}\\ 1 & \text{otherwise}\end{cases}. \quad (13)$$

This type of a loss function selects the posterior mode as the Bayesian estimate of the PCA model. Thus, the BPCA solution can be obtained by solving the following optimization problem

$$\{\hat{\alpha},\hat{Z}\}_{\text{Bayesian}}=\underset{\tilde{\alpha},\tilde{Z}}{\text{argmax}}P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z}|\tilde{\alpha})P(\tilde{\alpha}) \quad (14)$$

which is referred to as the maximum a posteriori (MAP) estimate.

III. Gaussian Likelihood Density

If the measured process variables are assumed to be contaminated with additive zero mean Gaussian noise, i.e., $\varepsilon_X \sim N(0,Q_{\varepsilon_X})$, then the likelihood becomes the following normal density function,

$$P(X|\tilde{Z},\tilde{\alpha},\tilde{p})\sim N(\tilde{X},Q_{\varepsilon_X}). \quad (15)$$

IV. Multivariate Gaussian Noise-Free Data

Girshick (1939) has shown that the eigenvectors of multivariate Gaussian data are independent and asymptotically follow a multivariate normal distribution, with the following moments,

$$E[\tilde{\alpha}_j] = \alpha_j + O(n^{-1}), \text{ and}$$

$$Cov[\tilde{\alpha}_j] = \frac{1}{n}\sum_{j\neq k}\frac{\lambda_j\lambda_k}{(\lambda_j-\lambda_k)^2}\alpha_j\alpha_k^T + O(n^{-2}). \quad (16)$$

Thus, if we define the vector, $\tilde{A} \equiv [\tilde{\alpha}_1^T \ \tilde{\alpha}_2^T \ . \ \tilde{\alpha}_p^T]^T$, (which is of size $(rp\times 1)$, where p is the number of retained projection directions), then $\tilde{A}$ will also follow a multivariate normal distribution, i.e., $\tilde{A} \sim MVN(\mu_{\tilde{A}}, Q_{\tilde{A}})$. Furthermore, since the data and the principal components are linearly related, then

$$\tilde{Z}|\tilde{\alpha} \sim MVN(\mu_{\tilde{Z}|\tilde{\alpha}}, Q_{\tilde{Z}|\tilde{\alpha}}) = MVN(\mu_{\tilde{X}}\tilde{\alpha}, \tilde{\alpha}^T Q_{\tilde{X}}\tilde{\alpha}). \quad (17)$$

3.3 The BPCA (MAP) Algorithm

The MAP solution of the BPCA problem can be obtained by solving the following simultaneous parameter estimation and data reconciliation problems similar to those solved in MLPCA,

$$\{\hat{\alpha}\}_{MAP} = \arg\max_{\tilde{\alpha}} P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{A})$$

s.t.

$$\{\tilde{Z}\}_{MAP} = \arg\max_{\tilde{Z}} P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z}|\tilde{\alpha})$$

and, $\tilde{X} = \tilde{\alpha}\tilde{Z}$, $\tilde{\alpha}^T\tilde{\alpha} = I$. (18)

Based on the simplifying assumptions made earlier, all densities in the posterior are now defined as multivariate normal, and thus the MAP solution can be obtained by solving the following simultaneous minimization problems for the parameters and the reconciled data as follows,

$$\{\hat{\alpha}\}_{BPCA} = \underset{\hat{\alpha}}{\text{argmin}}\left\{\sum_{i=1}^{n}(x_i-\hat{x}_i)^T Q_{\varepsilon_X}^{-1}(x_i-\hat{x}_i) + (\hat{A}-\mu_{\tilde{A}})^T Q_{\tilde{A}}^{-1}(\hat{A}-\mu_{\tilde{A}})\right\}$$

s.t.

$$\{\hat{z}_i\}_{BPCA} = \underset{\hat{z}_i}{\text{argmin}}\left\{\sum_{i=1}^{n}(z_i-\hat{z}_i)^T Q_{\varepsilon_Z}^{-1}(z_i-\hat{z}_i) + (\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}})^T Q_{\tilde{z}|\tilde{\alpha}}^{-1}(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}})\right\} \quad (19)$$

subject to the constraints shown in equation (2b). The data reconciliation problem has the following closed form solution (see Appendix II),

$$\{\hat{z}_i\}_{MAP} = (\hat{\alpha}^T Q_{\varepsilon_X}^{-1}\hat{\alpha} + Q_{\tilde{z}|\tilde{\alpha}}^{-1})^{-1}(\hat{\alpha}^T Q_{\varepsilon_X}^{-1} x_i + Q_{\tilde{z}|\tilde{\alpha}}^{-1}\mu_{\tilde{z}|\tilde{\alpha}}). \quad (20)$$

3.4 Empirical Prior Estimation

The preceding BPCA formulation assumes the availability of the prior density. Unfortunately, sometimes the data are the only available source of information. In such cases, the prior is estimated empirically from the data itself. In BPCA, since a structural prior is available, we only need to estimate its hyperparameters, $\{\mu_{\tilde{X}}, Q_{\tilde{X}}, \mu_{\tilde{A}}, \text{and } Q_{\tilde{A}}\}$, which we will denote by η. Thus, the posterior becomes

$$P(\tilde{Z},\tilde{\alpha},\tilde{p}|X,\eta) \propto P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z},\tilde{\alpha},\tilde{p}|\eta). \quad (21)$$

Note that now the prior is dependent on the set of hyperparameters, η. The basic idea here is to estimate η from the data using maximum likelihood estimation, and then use the empirically estimated prior to solve for the BPCA model using the following posterior,

$$P(\tilde{Z},\tilde{\alpha},\tilde{p}|X,\hat{\eta}) \propto P(X|\tilde{Z},\tilde{\alpha},\tilde{p})P(\tilde{Z},\tilde{\alpha},\tilde{p}|\hat{\eta}). \quad (22)$$

The BPCA model obtained using an empirically estimated prior is referred to as empirical BPCA (EBPCA).

4. Illustrative Examples

Example 1: *Gaussian Steady State Data*

The data matrix used in this example consists of 3 variables and fifty observations, and the noise-free data are generated as follows:

$$\tilde{x}_1 \sim N(3,1),\ \tilde{x}_2 \sim N(1,4),$$
$$\text{and, } \tilde{x}_3 = a_1\tilde{x}_1 + a_2\tilde{x}_2,$$
$$\text{where } a_1 = a_2 = 1. \quad (23)$$

Therefore, the model tank of the noise-free data is 2, which is assumed o be known. Then, the data are contaminated with zero mean Gaussian noise with the following covariance matrix, $Q_{\varepsilon_X} = diag(1 \quad 4 \quad 5)$.

To illustrate the performance of BPCA, three cases are considered. In the first case (I), the perfect prior is used, which is known in this synthetic example. Note that this case represents the best case scenario for the performance of BPCA. In the second case (II), the prior is

computed using 500 external data points generated the same way as described above. And finally, in the third case (III), the prior is calculated empirically as described in Section 3.4. The results of a Monte Carlo simulation using 100 realizations are shown in Table 1, which compares the mean square errors of the estimated data, the model parameters (a_1 and a_2), and the angular deviations, (γ_1 and γ_2) between the noise-free projection directions and the estimated spaces. The angular deviation of the j^{th} projection direction is defined as follows:

$$\gamma_j = \cos^{-1}\left(\frac{\tilde{\alpha}_j^T \hat{\alpha}\hat{\alpha}^T \tilde{\alpha}_j}{\left\|\tilde{\alpha}_j^T\right\|\left\|\hat{\alpha}\hat{\alpha}^T \tilde{\alpha}_j\right\|}\right). \qquad (24)$$

Table 1 shows that incorporating external information can greatly improve the estimated BPCA model (as in cases I and II), and that even when no external information is used, an empirically estimated prior can still help, especially in estimating the noise-free data (as in case III).

Table 1. PCA modeling of Gaussian data

MSE	PCA	ML-PCA	BPCA (I)	BPCA (II)	EBPCA (III)
x_1	1.55	0.90	0.48	0.51	0.54
x_2	3.52	2.62	1.50	1.62	1.72
x_3	3.09	2.82	1.69	1.74	1.95
γ_1	2.9	2.2	0.018	0.29	2.2
γ_2	16.9	9.0	0.037	4.50	8.9
a_1	0.18	0.04	1.7e-5	3.1e-4	0.04
a_2	0.19	0.15	3.0e-5	2.5e-3	0.14

***Example 2:** Dynamic Data*

In this example, EBPCA is applied to dynamic data, and its performance is compared to those of PCA and MLPCA. The noise-free data are generated using the following model.

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

$$\text{where, } \tilde{u}(k) \sim \begin{cases} N(0,2) & 1 \le k \le 15 \\ N(5,2) & 16 < k \end{cases}. \qquad (25)$$

Then, the input and output variables are contaminated with zero-mean Gaussian noise with variances 2 and 4, respectively. The data matrix is organized as follows:

$$X = [Y(k-1) \quad U(k) \quad Y(k)], \qquad (26)$$

and thus, the corresponding noise covariance matrix is $Q_{\varepsilon_X} = diag(4 \quad 2 \quad 4)$. The results of a Monte Carlo simulation, using 100 realizations assuming a model of rank 2, are summarized in Table 2. This table shows that, similar to example 1, EBPCA outperforms existing methods in estimating the underlying noise-free data as well as the PCA model. More details about these examples are presented in (Nounou, 2000).

Table 2. PCA modeling of dynamic data

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
γ_1	0.32	0.33	0.33
γ_2	12.8	6.0	6.0

5. Conclusions

This paper presented a Bayesian PCA approach that allows incorporating external knowledge about the noise-free projection directions and data to improve its accuracy. The BPCA approach is shown to be more general than existing methods and reduces to these techniques when no prior information is incorporated. When no external information is available, it is shown that the approach can still do better than the existing methods by empirically estimating a prior using the available data. This approach has been extended to Bayesian latent variable regression and shows similar benefits (Nounou, 2000).

References

Basilevsky, A., "Statistical Factor Analysis and Related Methods: Theory and Applications", Wiley Series in Probability and Mathematical Statistics, New York, 1994.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D., "Bayesian Data Analysis", Chapman and Hall, London, 1995.

Girshick, M.A., "On the Sampling Theory of Roots of Determinantal Equations", Ann. Math. Stat., 10, 203-224, 1939.

Kramer, M.A. and R.S.H. Mah, "Model-Based Monitoring", Proc. Int. Conf. On Foundations of Computer Aided Process Operations, D. Rippin, J. Hale, J. Davis, eds. CACHE, 1994.

Kresta, J.V., J.F. MacGregor, and T.E. Marlin, "Multivariate Statistical Process Operation Performance", Can. J. Chem. Eng., 69, 35-47, 1991.

Nounou, M. N., Bakshi, B. R., Goel, P. K., and Shen, X., "Bayesian Principal Component Analysis", (submitted to J. Chemometrics, 2001).

Nounou, M. N., "Multiscale Bayesian Linear Modeling and Applications", Ph.D. Thesis, Department of Chemical Engineering, The Ohio State University, 2000.

Wentzell P.D., Andrews, D., Hamilton, D. C., Faber, K., and Kowalski, B.R., "Maximum Likelihood Principal Component Analysis", J. of Chemometrics, 11, 339-366, 1997.

Wise, B.M., N.L. Ricker, D.F. Veltkamp, and B.R. Kowalski, "A Theoretical Basis for the Use of Principal Component Models for Monitoring Multivariate Processes", Proc. Cont. Qual., 1, 41, 1990.

Appendices

Appendix I: *derivation of the MLPCA data rectification solution*

Define the Lagrange function as,

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

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

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

$$\frac{\partial L}{\partial \hat{z}_i}=-\lambda\hat{\alpha}=0 \tag{A1.3}$$

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

Substituting equation A1.2 in A1.3, get

$$\hat{\alpha}^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\left(\mathbf{x}_i-\hat{\mathbf{x}}_i\right)=0. \tag{A1.5}$$

Substituting equation A1.4 in A1.5, get

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

Rearranging equation A1.6, get the MLPCA solution

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

Appendix II: *derivation of the BPCA data rectification solution*

Define the Lagrange function as,

$$L=\left(\mathbf{x}_i-\hat{\mathbf{x}}_i\right)^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\left(\mathbf{x}_i-\hat{\mathbf{x}}_i\right)+\left(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}}\right)^{\mathrm{T}}\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}\left(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}}\right)+\lambda\left(\hat{\mathbf{x}}_i-\hat{\alpha}\hat{z}_i\right). \tag{A2.1}$$

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

$$\frac{\partial L}{\partial \hat{\mathbf{x}}_i}=-2\mathbf{Q}_{\varepsilon_x}^{-1}\left(\mathbf{x}_i-\hat{\mathbf{x}}_i\right)+\lambda^{\mathrm{T}}=0 \tag{A2.2}$$

$$\frac{\partial L}{\partial \hat{z}_i}=2\mathbf{Q}_{\tilde{z}}^{-1}\left(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}}\right)-\hat{\alpha}^{\mathrm{T}}\lambda^{\mathrm{T}}=0 \tag{A2.3}$$

$$\frac{\partial L}{\partial \lambda}=\hat{\mathbf{x}}_i-\hat{\alpha}\hat{z}_i=0. \tag{A2.4}$$

Substituting equation A2.2 in A2.3, get

$$2\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}\left(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}}\right)-2\hat{\alpha}^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\left(\mathbf{x}_i-\hat{\mathbf{x}}_i\right)=0. \tag{A2.5}$$

Substituting equation A2.4 in A2.5, get

$$2\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}\left(\hat{z}_i-\mu_{\tilde{z}|\tilde{\alpha}}\right)-2\hat{\alpha}^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\left(\mathbf{x}_i-\hat{\alpha}\hat{z}_i\right)=0 \tag{A2.6}$$

Rearranging A2.6, get the MAP solution

$$\left\{\hat{z}_i\right\}_{\mathrm{MAP}}=\left(\hat{\alpha}^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\hat{\alpha}+\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}\right)^{-1}\left(\hat{\alpha}^{\mathrm{T}}\mathbf{Q}_{\varepsilon_x}^{-1}\mathbf{x}_i+\mathbf{Q}_{\tilde{z}|\tilde{\alpha}}^{-1}\mu_{\tilde{z}|\tilde{\alpha}}\right). \tag{A2.7}$$