Variational Bayes Based Approach to Robust Subspace Learning

Takayuki Okatani and Koichiro Deguchi
Tohoku University
6-6-01 Aramaki Aza Aoba, Aoba-ku Sendai, Japan
okatani@fractal.is.tohoku.ac.jp

Abstract

This paper presents a new algorithm for the problem of robust subspace learning (RSL), i.e., the estimation of linear subspace parameters from a set of data points in the presence of outliers (and missing data). The algorithm is derived on the basis of the variational Bayes (VB) method, which is a Bayesian generalization of the EM algorithm. For the purpose of the derivation of the algorithm as well as the comparison with existing algorithms, we present two formulations of the EM algorithm for RSL. One yields a variant of the IRLS algorithm, which is the standard algorithm for RSL. The other is an extension of Roweis's formulation of an EM algorithm for PCA, which yields a robust version of the alternated least squares (ALS) algorithm. This ALS-based algorithm can only deal with a certain type of outliers (termed vector-wise outliers). The VB method is used to resolve this limitation, which results in the proposed algorithm. Experimental results using synthetic data show that the proposed algorithm outperforms the IRLS algorithm in terms of the convergence property and the computational time.

1. Introduction

The estimation of linear subspace parameters from a set of data points in the presence of outliers is a fundamental problem in computer vision and many other fields. Its applications include the structure from motion (SFM) problem under the assumption of affine cameras, and the image based rendering (IBR) using images of a scene taken under different illumination conditions. In this paper, this problem is termed *robust subspace learning* (RSL) on the basis of [4]. A formal statement of the non-robust version of the problem is as follows. Assume that an observation y_{ij} ($i = 1, \ldots, m$ and $j = 1, \ldots, n$) follows a model given by

$$y_{ij} = \mathbf{u}_i^\top \mathbf{v}_j + \mu_j + \varepsilon_{ij}, \tag{1}$$

where $\mathbf{u}_i$ and $\mathbf{v}_j$ are r-vectors and ε_{ij} is observation noise. We estimate $\mathbf{u}_i$, $\mathbf{v}_j$, and μ_j. Defining $\mathbf{Y}$ to be an $m \times n$ matrix $\{y_{ij}\}$, $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]^\top$, $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]^\top$, and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$, Eq.(1) can be rewritten as follows:

$$\mathbf{Y} = \mathbf{U}\mathbf{V}^\top + \mathbf{1}_m \boldsymbol{\mu}^\top + \mathbf{E}, \tag{2}$$

where $\mathbf{1}_m$ is an m-vector $[1, \ldots, 1]^\top$, and $\mathbf{E}$ is an $m \times n$ matrix $\{\varepsilon_{ij}\}$ of noise. Then, the problem is to estimate $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\mu}$, given the data $\mathbf{Y}$.

Assuming that ε_{ij} is an iid Gaussian noise with a zero mean, a maximum likelihood estimation of the parameters is performed as follows:

$$\|\mathbf{Y} - \mathbf{U}\mathbf{V}^\top - \mathbf{1}_m \boldsymbol{\mu}^\top\|_F^2 \to \min, \tag{3}$$

where $\|\cdot\|_F^2$ represents the Frobenius norm. This minimization is reduced to a basic eigenvalue problem of a matrix, and its computation is easy.

The problem becomes difficult (1) when some of the data are missing and increases in difficulty (2) when there are outliers in the data. The missing components are dealt with as follows. We define h_{ij} such that $h_{ij} = 1$ if y_{ij} exists and $h_{ij} = 0$ if it is missing. Instead of (3), we then consider a minimization of the sum over only the non-missing components:

$$\|\mathbf{H} \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^\top - \mathbf{1}_m \boldsymbol{\mu}^\top)\|_F^2 \to \min. \tag{4}$$

where $\mathbf{H}$ is an $m \times n$ matrix $\{h_{ij}\}$ and $\odot$ represents the component-wise product. Unlike (3), the minimization is essentially nonlinear and a direct nonlinear computation is necessary. There are several algorithms for this nonlinear computation, such as ALS and the Wiberg algorithm [10]; also see [7, 3].

In this paper, we consider the case in which there exist outliers in the data in addition to the missing components; we refer to the corresponding problem as RSL. In this problem, some of the non-missing components are inliers (datum following Eq.(1)), whereas some are outliers (datum not following Eq.(1)); however which of these are inliers and which are outliers is not known in advance.

A standard numerical algorithm for the problem is the iterative reweighted least squares (IRLS) algorithm [4]. In this algorithm, the problem is converted to a weighted least squares problem, in which a weight is defined for each datum. This weight represents the probability of each datum being an inlier (or outlier). The weights are first determined using the current estimates of the subspace parameters, and then the linear subspace parameters are updated by solving the weighted linear least squares problem. This two steps are performed alternately until convergence is achieved. Recently, Torre et al. provided a rigorous interpretation of this algorithm within the framework of the M-estimators [4].

1-4244-1180-7/07/$25.00 ©2007 IEEE

In this paper, we present a new algorithm for RSL. It is derived on the basis of the *variational Bayes* (VB) method [1]. The VB method can be regarded as a Bayesian generalization of the expectation maximization (EM) algorithms, and it is used especially for problems in which the expectations are mathematically intractable (and thus, no feasible EM algorithm is available).

In Section 2, we show two different formulations to derive the EM algorithms for RSL. In the first formulation, an EM algorithm that coincides with the traditional IRLS algorithm is derived. The second formulation is an extension of Roweis's formulation for deriving an EM algorithm for principal component analysis (PCA) [8]. In this formulation, a feasible EM algorithm is derived only in the case in which the outliers emerge as entire row vectors of $\mathbf{Y}$ (we will term these vector-wise outliers). In Section 3, by employing the VB method, we show that it is possible to derive a promising algorithm on the basis of the second formulation that can deal with the general case in which each component of $\mathbf{Y}$ can be an outlier (we term these component-wise outliers). We observe through experiments that this algorithm provides a better performance than the existing algorithms in terms of the convergence property as well as the computational time. In Section 4, we report this observation along with experimental results.

2. EM Algorithms for Robust Subspace Learning

2.1. The IRLS algorithm

First, we present the standard formulation of an EM algorithm for RSL. The resulting algorithm is similar to the IRLS algorithm, as shown below. Although Torre et al. used the M-estimators to provide a rigorous interpretation of the IRLS algorithm [4], we would like to argue that the EM framework enables a more rigorous discussion and yields a more flexible algorithm, although the details are omitted here due to the lack of space.

Let z_{ij} represent an indicator variable for the inliers; $z_{ij} = 1$ indicates that y_{ij} is an inlier, whereas $z_{ij} = 0$ indicates y_{ij} is an outlier. The variable z_{ij} is a hidden variable that is not observed. The joint density of y_{ij} and z_{ij} can be formally represented as

$$p(y_{ij}, z_{ij}) = \Big\{p(y_{ij}|z_{ij}=1)\,p(z_{ij}=1)\Big\}^{z_{ij}} \times \Big\{p(y_{ij}|z_{ij}=0)\,p(z_{ij}=0)\Big\}^{1-z_{ij}}. \quad (5)$$

We denote the proportion of the mixture of inliers (i.e. how frequently inliers will appear in the data) by $\alpha \equiv p(z_{ij}=1)$. Then, the proportion of the mixture of outliers is expressed as $1-\alpha = p(z_{ij}=0)$. α is treated as an unknown parameter and is estimated along with the other parameters. Assuming a normal density $N(0,\sigma^2)$ for the noise that contaminate the inliers, the density of an inlier is denoted by $p(y_{ij}\,|\,z_{ij}=1) = N(\mathbf{u}_i^\top\mathbf{v}_j+\mu_j,\sigma^2)$. In the case of that for the outliers, we assume a uniform density $p(y_{ij}\,|\,z_{ij}=0)=\gamma$, where γ is a fixed constant that is specified beforehand. We denote all the parameters by $\mathbf{\Theta} = \{\mathbf{U},\mathbf{V},\boldsymbol{\mu},\alpha,\sigma^2\}$.

The goal is to estimate $\mathbf{\Theta}$ from the given data $\mathbf{Y}$. A standard solution is to maximize the marginal likelihood, which is obtained by the marginalization of the above-mentioned joint density with respect to the hidden variable z_{ij}. However, in this case, thecalculation of the marginal likelihood is intractable, and an EM algorithm is used. In the EM algorithm, the following conditional expectation of the log-likelihood of the joint density is maximized:

$$Q(\mathbf{\Theta}';\mathbf{\Theta}) \equiv E\left[\log p(\mathbf{y},\mathbf{z};\mathbf{\Theta}')\,|\,\mathbf{y};\mathbf{\Theta}\right]. \quad (6)$$

The maximization is followed by updating $\mathbf{\Theta}' \to \mathbf{\Theta}$, and these two steps are repeated alternately until convergence is achieved.

When the joint density (5) is substituted into Q, the conditional expectation $E[\cdot]$ affects only z_{ij}. We denote $w_{ij} \equiv E[z_{ij}\,|\,\mathbf{y},\mathbf{\Theta}]$. (We set $w_{ij}=0$ for (i,j) of $h_{ij}=0$.) Note that w_{ij} can take a continuous value from 0 to 1. In the E-step, merely the evaluation of w_{ij} is performed. In the subsequent M-step, Q is maximized w.r.t. $\mathbf{\Theta}$. The values of $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\mu}$ are determined such that the following function is maximized:

$$\phi(\mathbf{U},\mathbf{V},\boldsymbol{\mu}) = \sum_{i,j} w_{ij}(y_{ij} - \mathbf{u}_i^\top\mathbf{v}_j - \mu_j)^2. \quad (7)$$

In the case of the remaining parameters α and σ^2, explicit solutions are available. Thus, the resulting algorithm shown in Algorithm 1 is obtained. This algorithm includes a nonlinear minimization of Eq.(7), which can be performed by several algorithms, such as the weighted ALS (W-ALS), the weighted Wiberg (W-Wiberg), as well as general Newton-based methods. Note that Algorithm 1 is almost equivalent to that derived on the basis of the M-estimators [4], except for a few (minor) differences. We will use Algorithm 1 in Section 4 for performance comparisons.

Algorithm 1 EM algorithm for RSL (EM-IRLS)

1: Initialize $\mathbf{U}$, $\mathbf{V}$, $\boldsymbol{\mu}$, α, and σ^2.
2: For $i=1,\ldots,m$ and $j=1,\ldots,n$, set $w_{ij}=0$ if $h_{ij}=0$; otherwise update w_{ij} as

$$w_{ij} = \frac{\alpha(2\pi\sigma^2)^{-\frac{1}{2}}\exp\{-e_{ij}^2/(2\sigma^2)\}}{\alpha(2\pi\sigma^2)^{-\frac{1}{2}}\exp\{-e_{ij}^2/(2\sigma^2)\}+(1-\alpha)\gamma} \quad (8)$$

where $e_{ij} \equiv y_{ij} - \mathbf{u}_i^\top\mathbf{v}_j - \mu_j$.
3: Minimize Eq.(7) with respect to $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\mu}$. Update α and σ^2 by using the minimization solution for $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\mu}$ as

$$\alpha = \frac{\sum_{i,j} w_{ij}}{\sum_{i,j} h_{ij}}, \quad \sigma^2 = \frac{\sum_{i,j} w_{ij}(y_{ij} - \mathbf{u}_i^\top\mathbf{v}_j - \mu_j)^2}{\sum_{i,j} w_{ij}}. \quad (9)$$

4: Go to 2 until convergence is achieved.

2.2. Factor-analysis-based EM algorithm for RSL

In [8], Roweis derives an EM algorithm for PCA by transforming an EM algorithm tailored for the factor anal-

ysis (FA). (We do not distinguish PCA from the problem considered here.) Here, we present a robust version of the algorithm. (It is not shown in [8].) It should be noted that the resulting algorithm has a limitation: it can deal with only a special type of outliers, such that the entire row vectors of $\mathbf{Y}$ can be either inliers or outliers. We use the term *vector-wise* outliers to represent these outliers. The algorithm is still important, since a close examination of the algorithm leads to the derivation of the new algorithm, which is shown in Section 3.

The factor analysis assumes a data model in which a datum $\mathbf{y}_i$ (the row vector of $\mathbf{Y}$) is generated as follows:

$$\mathbf{y}_i = \mathbf{V}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \tag{10}$$

where $\mathbf{u}_i$ is assumed to follow $N(0, \sigma_u^2\mathbf{I})$, and $\boldsymbol{\varepsilon}_i$ is the noise following a normal density with a zero mean and a variance $\boldsymbol{\Psi}$. As compared to PCA, one difference is that $\mathbf{u}_i$ is not considered to be a parameter but a probabilistic variable whose density is specified. (We omit the mean vector $\boldsymbol{\mu}$ in this model for the sake of simplicity.) The joint density $p(\mathbf{y}, \mathbf{u})$ is given by $p(\mathbf{y}, \mathbf{u}) = p(\mathbf{y}\,|\,\mathbf{u})p(\mathbf{u})$, where $p(\mathbf{y}\,|\,\mathbf{u}; \mathbf{V}) = N(\mathbf{V}\mathbf{u}, \boldsymbol{\Psi})$ and $p(\mathbf{u}) = N(\mathbf{0}, \sigma_u^2\mathbf{I})$. By considering $\mathbf{u}_i$ to be a hidden variable, an EM algorithm can be derived as shown in Algorithm 2 [5].

Algorithm 2 EM algorithm for the factor analysis

1: [E-step] Compute $E[\mathbf{u}|\mathbf{y}_i] = \mathbf{B}\mathbf{y}_i$ and $E[\mathbf{u}\mathbf{u}^\top|\mathbf{y}_i] = \mathbf{I} - \mathbf{B}\mathbf{V} + \mathbf{B}\mathbf{y}_i\mathbf{y}_i^\top\mathbf{B}^\top$ for each $\mathbf{y}_i$ $(i = 1, \ldots, m)$, where $\mathbf{B} = \mathbf{V}^\top(\boldsymbol{\Psi} + \mathbf{V}\mathbf{V}^\top)^{-1}$.
2: [M-step] Compute $\mathbf{V}'$ and $\boldsymbol{\Psi}'$ as

$$\mathbf{V}' = \left(\sum_{i=1}^{m} \mathbf{y}_i E[\mathbf{u}|\mathbf{y}_i]^\top\right)\left(\sum_{l=1}^{m} E[\mathbf{u}\mathbf{u}^\top|\mathbf{y}_l]\right)^{-1}, \tag{11}$$

$$\boldsymbol{\Psi}' = \frac{1}{n}\mathrm{diag}\left(\sum_{i=1}^{m} \mathbf{y}_i\mathbf{y}_i - \mathbf{V}'E[\mathbf{u}|\mathbf{y}_i]\mathbf{y}_i^\top\right), \tag{12}$$

where $\mathrm{diag}(\cdot)$ is an operator that forces every off-diagonal element to be 0. Then, update $\mathbf{V} = \mathbf{V}'$ and $\boldsymbol{\Psi} = \boldsymbol{\Psi}'$.

Now, we choose $\boldsymbol{\Psi} = \epsilon\mathbf{I}$ and take a limit as $\epsilon \to 0$. Since $\mathbf{B} = \lim_{\epsilon\to 0} \mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top + \epsilon\mathbf{I})^{-1} = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top$, the conditional expectations are given as $E[\mathbf{u}|\mathbf{y}_i] = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{y}_i$ and $E[\mathbf{u}\mathbf{u}^\top|\mathbf{y}_i] = E[\mathbf{u}|\mathbf{y}_i]E[\mathbf{u}|\mathbf{y}_i]^\top$. In the limit, Algorithm 2 is converted into a simple algorithm that performs subspace learning (or PCA), shown in Algorithm 3. It should be noted that it coincides with the alternated least squares (ALS) algorithm.

Algorithm 3 EM algorithm for PCA

1: [E-step] Update $\mathbf{U}$ by $\mathbf{U} = \mathbf{Y}\mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}$.
2: [M-step] Update $\mathbf{V}$ by $\mathbf{V} = \mathbf{Y}^\top\mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}$.

In [9, 6], the same formulation is applied to the mixture models in which each datum is assumed to be generated from one of the several different subspaces. (The formulation is applied to the multi-body SFM problem [6].) By replacing (one of) the multiple subspace models with a model for the outliers, an algorithm for RSL can be derived. Specifically, by defining an indicator variable z_i such that $z_i = 1$ indicates that $\mathbf{y}_i$ is an inlier and $z_i = 0$ indicates that $\mathbf{y}_i$ is an outlier, the joint density can be written as

$$p(\mathbf{y}_i, \mathbf{u}_i, z_i) = \{p(\mathbf{y}_i\,|\,\mathbf{u}_i, z_i = 1)p(\mathbf{u}_i)p(z_i = 1)\}^{z_i} \times \{p(\mathbf{y}_i\,|\,z_i = 0)p(z_i = 0)\}^{1-z_i}. \tag{13}$$

As mentioned earlier, we define $\alpha \equiv p(z_i = 1)$ $(1 - \alpha = p(z_i = 0))$ and assume a uniform density for the outliers: $\gamma \equiv p(\mathbf{y}_i\,|\,z_i = 0)$. Then, by considering z_i to be a hidden variable, we can derive an EM algorithm, shown in Algorithm 4.

Algorithm 4 EM algorithm for RSL with data including only the vector-wise outliers

1: Initialize $\mathbf{V}$, α, and σ^2.
2: Compute $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_m]^\top$ from

$$\mathbf{u}_i = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}\mathbf{y}_i, \tag{14}$$

and compute w_i $(i = 1, \ldots, m)$ as

$$w_i = \frac{\alpha(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\mathbf{e}_i^2/(2\sigma^2))}{\alpha(2\pi\sigma^2)^{-\frac{n}{2}}\exp(-\mathbf{e}_i^2/(2\sigma^2)) + (1-\alpha)\gamma}, \tag{15}$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{V}\mathbf{u}_i = (\mathbf{I} - \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top)\mathbf{y}_i$.
3: Compute $\mathbf{V}$ from

$$\mathbf{V} = \mathbf{Y}^\top\mathbf{W}\mathbf{U}(\mathbf{U}^\top\mathbf{W}\mathbf{U})^{-1}, \tag{16}$$

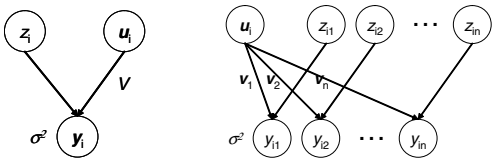
where $\mathbf{W} = \mathrm{diag}[w_1, \ldots, w_m]$, and then α and σ^2 from

$$\alpha = \frac{\sum_i w_i}{\sum_i 1} \quad \text{and} \quad \sigma^2 = \frac{\sum_i w_i\{(\mathbf{I} - \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top)\mathbf{y}_i\}^2}{n\sum_i w_i}. \tag{17}$$

It should be noted that there is an important difference between the algorithmic structures of Algorithms 1 and 4 other than the fact that the latter can deal with only the vector-wise outliers. This difference is that the former requires an external nonlinear minimization subroutine, whereas the latter does not and can stand by itself. Except for the apparent fact that the computational cost per iteration is significantly smaller for the latter than for the former, we cannot assert that this is an advantage; however, this is a remarkable property. In the next section, we will consider extending the above formulation to the general type of outliers, namely, thecomponent-wise outliers (i.e., each component of $\mathbf{Y}$ can be an outlier).

3. Variational Bayes based approach to RSL

As mentioned above, Algorithm 4 can deal with only the vector-wise outliers. Is it possible to derive an EM algo-



(a) (b)

Figure 1. Generative models for (a) the data with vector-wise outliers and (b) the data with component-wise outliers.

rithm that can deal with the component-wise outliers, based on the same formulation. The answer to this question is negative; a feasible EM algorithm cannot be obtained. In order to deal with the component-wise outliers, we need to introduce a component-wise indicator z_{ij} of the inlier/outlier for y_{ij}, as was done in the derivation of Algorithm 1. Then, the joint density is given as

$$p(\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i) = \prod_j \left\{ p(y_{ij} \,|\, \mathbf{u}_i, z_{ij} = 1) p(\mathbf{u}_i) p(z_{ij} = 1) \right\}^{z_{ij}} \times \left\{ p(y_{ij} \,|\, z_{ij} = 0) p(z_{ij} = 0) \right\}^{1 - z_{ij}}. \quad (18)$$

Then, we need to evaluate the corresponding conditional expectation Q, as was done in the derivation of Algorithm 4. However, it is mathematically intractable due to the following reason. Figure 1 shows the generative models for (a) the data with vector-wise outliers and for (b) the data with component-wise outliers. In case (a) (Algorithm 4), the datum $\mathbf{y}_i$ depends on a single indicator, whereas in case (b), $\mathbf{y}_i$ (and also y_{ij}) depends on multiple indicators $[z_{i1}, \ldots, z_{in}]$. Therefore, in order to calculate Q, a sum over all the combinations of assigning 0 and 1 to $[z_{i1}, \ldots, z_{in}]$ needs to be evaluated. The number of terms in the sum will be of the order of 2^n.

The variational Bayes (VB) method, which can be regarded as a generalization of the EM algorithms, has been developed to resolve such a difficulty, i.e., the intractability of Q. Because of the constraint of space, we will explain only the spirit of the VB algorithm below.

We begin with a marginal density, in which the hidden variables $\mathbf{Z}$ and the parameters $\boldsymbol{\Theta}$ are treated as variables and are marginalized:

$$\mathcal{L}(\mathbf{Y}) = \log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\Theta}) d\mathbf{Z} d\boldsymbol{\Theta}, \quad (19)$$

where $\boldsymbol{\Theta} = \{\mathbf{U}, \mathbf{V}, \alpha, \sigma^2\}$. The following holds for an arbitrary density $q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})$:

$$\mathcal{L}(\mathbf{Y}) = \mathcal{F}[q] + \mathcal{D}[q, p], \quad (20)$$

where

$$\mathcal{F}[q] \equiv \left\langle \log \frac{p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\Theta})}{q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})} \right\rangle_{\mathbf{Z}, \boldsymbol{\Theta}} \quad \text{and} \quad (21a)$$

$$\mathcal{D}[q, p] \equiv \left\langle \log \frac{q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})}{p(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})} \right\rangle_{\mathbf{Z}, \boldsymbol{\Theta}}; \quad (21b)$$

here, $\langle \cdot \rangle_{\mathbf{Z}, \boldsymbol{\Theta}}$ represents an expectation w.r.t. $q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})$. It should be noted that $\mathcal{D}[q, p]$ indicates the KL divergence between q and p. Therefore, since $\mathcal{L}(\mathbf{Y})$ is constant, the value of q that maximizes $\mathcal{F}[q]$ gives an optimal approximation to the true posterior p, in the sense that the KL information is minimized. Such an optimal q is searched for by using the variational method, which is the main idea of the VB method.

Usually, we need some additional assumption on q to make the computation of the optimal q tractable. It is often assumed that q is factorizable w.r.t. each variable, that is,

$$q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y}) = q(\mathbf{Z} \,|\, \mathbf{Y}) q(\boldsymbol{\Theta} \,|\, \mathbf{Y}). \quad (22)$$

Since the true posterior is not necessarily factorizable in this manner, the best density q calculated under this assumption is only an approximation of the true posterior p. In other words, we intend to find the best approximation q to p within the class of the functions factorizable as shown above.

When assuming that q is factorizable, the variational method derives the following solution. For the posterior $q(\mathbf{Z} \,|\, \mathbf{Y})$ of the hidden variable, we solve $\delta\mathcal{F} / \delta q(\mathbf{Z} \,|\, \mathbf{Y}) = 0$, and for the posterior $q(\boldsymbol{\Theta} \,|\, \mathbf{Y})$ of the parameters, we solve $\delta\mathcal{F} / \delta q(\boldsymbol{\Theta} \,|\, \mathbf{Y}) = 0$ [1], from which we obtain

$$q(\mathbf{Z} \,|\, \mathbf{Y}) = C_{\mathbf{Z}} \exp\langle \log p(\mathbf{Y}, \mathbf{Z} \,|\, \boldsymbol{\Theta}) \rangle_{\boldsymbol{\Theta}} \quad \text{and} \quad (23a)$$

$$q(\boldsymbol{\Theta} \,|\, \mathbf{Y}) = C_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}) \exp\langle \log p(\mathbf{Y}, \mathbf{Z} \,|\, \boldsymbol{\Theta}) \rangle_{\mathbf{Z}}, \quad (23b)$$

where $C_{\mathbf{Z}}$ and $C_{\boldsymbol{\Theta}}$ are the normalization constants and $p(\boldsymbol{\Theta})$ is a prior of $\boldsymbol{\Theta}$, which is assumed to be of a uniform density in our subsequent derivation.

The first equation (23a) calculates a posterior of $\mathbf{Z}$ given that of $\boldsymbol{\Theta}$, while the second equation (23b) calculates a posterior of $\boldsymbol{\Theta}$ given that of $\mathbf{Z}$. Thus, these two equations depend on each other. In order to resolve this mutual dependency, the VB method alternately performs the two calculations: first, $q(\mathbf{Z} \,|\, \mathbf{Y})$ is calculated using the latest estimate $q(\boldsymbol{\Theta} \,|\, \mathbf{Y})$ and then, $q(\boldsymbol{\Theta} \,|\, \mathbf{Y})$ is calculated using $q(\mathbf{Z} \,|\, \mathbf{Y})$. The two steps are sometimes referred to as VB-E and VB-M steps. The iteration of these two steps until convergence constitutes an algorithm for estimating the posteriors. It should be noted that the algorithm does not estimate the parameters $\boldsymbol{\Theta}$ themselves; it estimates their posterior densities. Fortunately, as long as the elementary densities such as the exponential family are assumed for the data models, the resulting posteriors q will, in most cases, be elementary densities, too. Thus, the estimation of the posteriors is reduced to the estimation of their sufficient statistics. The resulting algorithm iteratively updates these sufficient statistics of the posteriors.

By applying the above-meneioned explanation to our RSL problem, we obtain Algorithm 5; the detailed derivation is shown in Appendix A. The notation used here is the same as that used in Section 2.1. For a version of the model with a mean $y_{ij} \leftrightarrow \mathbf{u}_i^\top \mathbf{v}_j + \mu_j$, an algorithm can similarly be derived. However, it is omitted here due to the lack of space.

It should be noted that as is the case with Algorithm 4, Algorithm 5 consists only of linear computations. Its struc-

ture is similar to that of the ALS algorithm. Therefore, unlike Algorithm 1, Algorithm 5 does not require any external minimization subroutine. We can thus conclude that the computational cost per iteration in this algorithm will be significantly lower than that in the IRLS algorithm.

However, it is known [2, 7, 4] that the ALS algorithm sometimes suffers from a slow convergence: the ALS algorithm often requires a large number of iterations to converge. Therefore, the total computational cost of the ALS algorithm can be (much) larger than that of the Newton-based algorithms (Levenberg-Marquardt and Wiberg). Hence, we intend to examine whether Algorithm 5 exhibits the same characterstic. According to our experiments, the answer to this question is fortunately negative. In fact, the new algorithm— Algorithm 5— shows a rather faster convergence than the existing methods. This will be demonstrated in the next section.

Algorithm 5 VB-based algorithm for RSL

1: For $i = 1, \ldots, m$ and $j = 1, \ldots, n$, initialize $\mathbf{u}_i$, $\mathbf{v}_j$, α (e.g., 0.5), and σ^2. Set $\mathbf{\Psi}_i = \mathbf{u}_i\mathbf{u}_i^\top$ and $\mathbf{\Phi}_j = \mathbf{v}_j\mathbf{v}_j^\top$.

2: [VB-E] For $i = 1, \ldots, m$ and $j = 1, \ldots, n$, set $w_{ij} = 0$ if $h_{ij} = 0$, and if $h_{ij} = 1$, set

$$w_{ij} = \frac{\alpha(2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-e_{ij}^2/(2\sigma^2)\}}{\alpha(2\pi\sigma^2)^{-\frac{1}{2}} \exp\{-e_{ij}^2/(2\sigma^2)\} + (1-\alpha)\gamma} \tag{24}$$

where

$$e_{ij}^2 = y_{ij}^2 - 2y_{ij}\mathbf{u}_i^\top\mathbf{v}_j + \mathrm{tr}(\mathbf{\Psi}_i\mathbf{\Phi}_j). \tag{25}$$

3: [VB-M] Update α and σ^2 by

$$\alpha = \frac{\sum_{i,j} w_{ij} + 1}{\sum_{i,j} h_{ij} + 2} \quad \text{and} \quad \sigma^2 = \frac{\sum_{i,j} w_{ij} e_{ij}^2}{\sum_{i,j} w_{ij}}. \tag{26}$$

Next, update $\mathbf{u}_i$ and then $\mathbf{\Psi}_i$ for $i = 1, \ldots, m$ by

$$\mathbf{u}_i = \left\{\sum_j w_{ij}\mathbf{\Phi}_j\right\}^{-1} \left\{\sum_j w_{ij}y_{ij}\mathbf{v}_j\right\} \quad \text{and} \tag{27a}$$

$$\mathbf{\Psi}_i = \sigma^2 \left\{\sum_j w_{ij}\mathbf{\Phi}_j\right\}^{-1} + \mathbf{u}_i\mathbf{u}_i^\top. \tag{27b}$$

Note that $\mathbf{u}_i$ on the rhs of (27b) is the latest $\mathbf{u}_i$ computed from (27a). Similarly, update $\mathbf{v}_j$ and then $\mathbf{\Phi}_j$ for $j = 1, \ldots, n$ by

$$\mathbf{v}_j = \left\{\sum_i w_{ij}\mathbf{\Psi}_i\right\}^{-1} \left\{\sum_i w_{ij}y_{ij}\mathbf{u}_i\right\} \quad \text{and} \tag{28a}$$

$$\mathbf{\Phi}_j = \sigma^2 \left\{\sum_i w_{ij}\mathbf{\Psi}_i\right\}^{-1} + \mathbf{v}_j\mathbf{v}_j^\top. \tag{28b}$$

4. Experimental results

We choose 30×20 as the dimension of $\mathbf{Y}$, i.e., $m = 30$ and $n = 20$, and $r = 3$ for the dimension of the linear subspace (the number of columns of $\mathbf{U}$ and $\mathbf{V}$). Then, the data are randomly generated as follows. Random values are generated according to a normal density $N(0, 1)$, and they are assigned to the components of $\mathbf{U}$ and $\mathbf{V}$. Then, we compute $\mathbf{Y}_0 \equiv \mathbf{U}\mathbf{V}^\top$ and generate $\mathbf{Y}$ by adding the noise obeying $N(0, \sigma^2)$ to each component of $\mathbf{Y}_0$. Subsequently, missing components are randomly chosen. Let R_{miss} be the supposed proportion of the missing components in the size mn of $\mathbf{Y}$. We randomly choose $mn \times R_{\mathrm{miss}}$ components out of $\mathbf{Y}$ and consider them to be the missing components. The outliers are then chosen. Let R_{out} be the supposed proportion of the outliers in the number of non-missing components. We randomly choose $mn(1 - R_{\mathrm{miss}})R_{\mathrm{out}}$ components from the non-missing components of $\mathbf{Y}$. Their values are overwritten by a uniform random value chosen from the range $[-5.0 : 5.0]$. This range indicates that $\gamma = 0.1$, and it is used in all the algorithms.

Since the missing components and the outliers are randomly chosen, their distribution in $\mathbf{Y}$ can be nonuniform such that the corresponding problem could be indeterminate (i.e., an infinite number of solutions). Consequently, it becomes very difficult to correctly assess the performance of the algorithms. To avoid this, we check the following two conditions: (1) whether the number of inliers in each row and column of $\mathbf{Y}$ is greater or equal to $2r$, and (2) whether the Hessian $\partial^2 J/\partial\mathbf{x}^2$ of the function $J(\mathbf{x}) = J(\mathbf{U}, \mathbf{V}) \equiv \|\mathbf{H}' \odot (\mathbf{Y} - \mathbf{U}\mathbf{V}^\top)\|_F^2$ that is to be minimized is numerically non-singular, where $\mathbf{H}'$ is an indicator matrix of the missing components as well as the outliers. If either of the two conditions is not met, we regenerate the data from the beginning. Condition (1) is for eliminating nearly ill-conditioned data. Condition (2) is a necessary condition, and it is reduced to whether the rank of $\mathbf{Q}_F\mathbf{G}$ is equal to $r(n - r)$, where the notations are borrowed from [7]. (The details are omitted here.)

We applied the following three algorithms to the data: Algorithm 5 (VB); Algorithm 1, in which the minimization is performed by the weighted ALS algorithm (EM-ALS); and the weighted Wiberg algorithm (EM-Wiberg). Each of the algorithms is run for 100 trials. The initial values of $\mathbf{U}$ and $\mathbf{V}$ are chosen randomly, the initial value of the outlier proportion α is set to 0.5, and that of the noise variance σ^2 is set to 100 (in general, a large value is preferred, since otherwise, every component will be identified as an outlier in the first iteration). For the sake of fair comparison, a common set of data and initial values are fed to each of the algorithms.

We first examine the number of iterations required for convergence. Figures 2, 3, and 4 show the iterations for VB, EM-ALS, and EM-Wiberg algorithms. For the EM-ALS and EM-Wiberg algorithms, the iteration counts of the subroutines (ALS and Wiberg, respectively) are shown. In the figures, the minimum, median, and maximum of the number of iterations are shown; here the proportion of the missing components and that of the outliers vary. From these results, it is evident that VB requires the least iteration

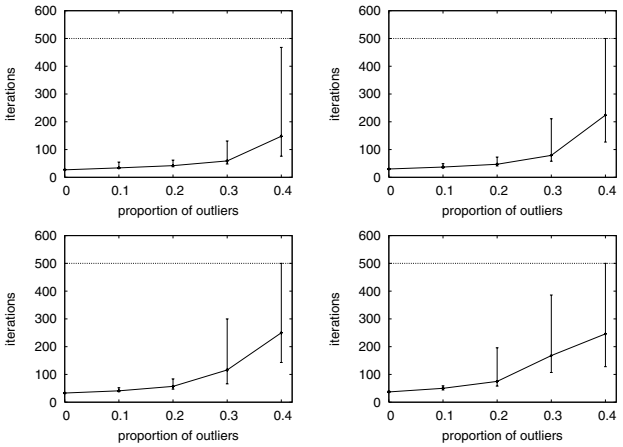


Figure 2. The number of iterations for VB. The upper limit is 500 iterations. The bars represents the minimum, median, and maximum of 100 trials. From the upper left to the lower right, the proportion of the missing components varies as 0%, 10%, 20%, and 30%, respectively.

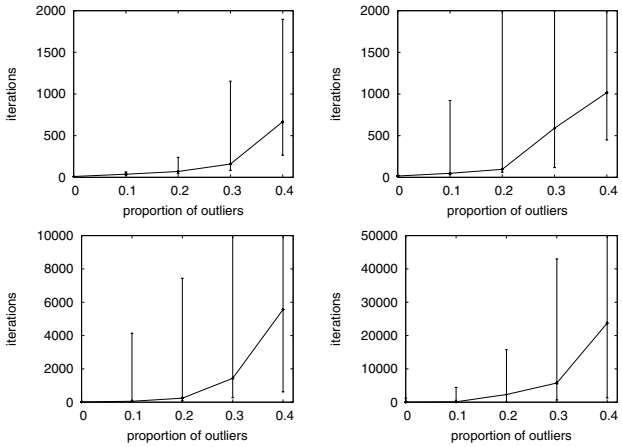


Figure 3. The number of iterations for EM-ALS. The upper limits of the main loop (EM) and the sub-loop (ALS) are 200 and 300 iterations, respectively.

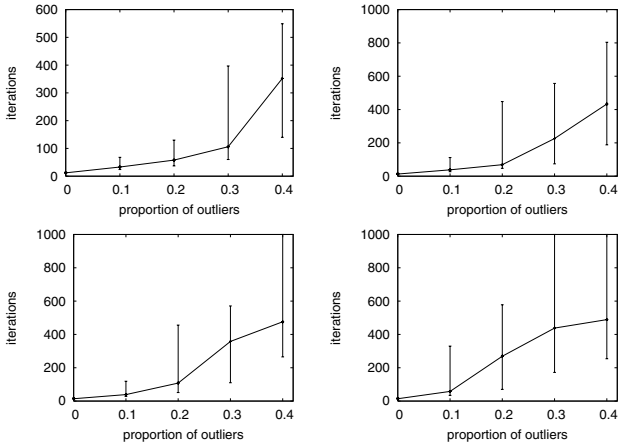


Figure 4. The number of iterations for EM-Wiberg. The upper limits of the main loop (EM) and the sub-loop (ALS) are set to 200 and 300 iterations, respectively.

counts among all the cases. Although EM-Wiberg exhibits only slightly larger iteration counts than VB, the computational cost of a single iteration is significantly larger for EM-Wiberg than for VB. At every iteration, EM-Wiberg solves a linear problem whose matrix size is $p \times nr$, where p is the number of non-missing components, whereas VB solves a problem whose matrix size is only $r \times r$ (equivalent to the ALS algorithm). Thus, when comparing the computational time, VB is drastically faster than EM-Wiberg as a result.

Next, we examine how well the algorithms converge to the correct solution (when starting from random initial values). Figure 5 shows the proportion of successful trials out of 100 trials. A successful trial is defined as one for which the proportion of misidentified inliers (i.e., outliers that are wrongly identified as inliers) to the total number of outliers is less than 5%. (Thus, this is a fairly tight criterion.) From the plots, it can be seen that the three algorithms share the same tendency, namely, their performance deteriorates for difficult sets of data. However, the extent of the deterioration is different. While the results for EM-ALS and EM-Wiberg are almost identical, VB shows better results. This is clearly confirmed in the fourth plot (lower right); for 20% missing components, the number of successful trials is nearly double, and for 30% missing components, it is quadruple or more.

5. Summary

As shown previously, there are two formulations for deriving an EM algorithm for RSL. The difference between the two is whether or not the subspace parameters are treated as probabilistic variables. If not, an EM algorithm that coincides with the standard IRLS algorithm is derived. Otherwise, an EM algorithm that can only deal with the vector-wise outliers is derived. We show that by applying the variational Bayes (VB) method, the component-wise outliers can be dealt with in a similar formulation, in which all the parameters are treated as probabilistic variables, as in usual Bayesian formulation. The resulting algorithm differs from any of the existing algorithms. We have found through experiments that this new algorithm shows a better performance in terms of the convergence property and the computational time, as compared to the existing algorithms. The structure of the algorithm might contribute to these good qualities, although the exact mechanisms are unclear. We would like to examine these features in the future.

A. Derivation of the variational-Bayes-based algorithm

By choosing the last two parameters from $\boldsymbol{\Theta} = [\mathbf{U}, \mathbf{V}, \alpha, \sigma^2]$, we sometimes write $\boldsymbol{\theta} = [\alpha, \sigma^2]$ below. When $\mathbf{U}$, $\mathbf{V}$, and $\boldsymbol{\theta}$ are specified, each component (i, j) of $\mathbf{Y}$ and $\mathbf{Z}$ is independent of the others. Therefore, the conditional

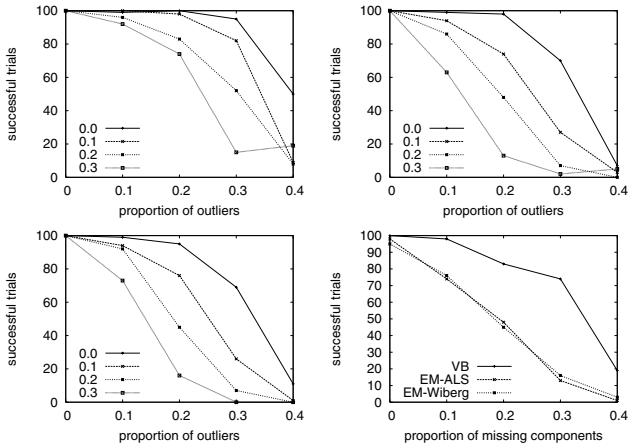


Figure 5. The plot of the number of successful trials out of 100. See the text for details. Upper left: VB. Upper right: EM-ALS. Lower left: EM-Wiberg. The four lines in each plot correspond to different proportions of missing components. On the lower right is a plot of the successful trials vs. the proportion of the missing elements for the case of 20% outliers.

density is written as

$$p(\mathbf{Y}, \mathbf{Z} \,|\, \mathbf{U}, \mathbf{V}, \boldsymbol{\theta}) = \prod_{ij} p(y_{ij}, z_{ij} \,|\, \mathbf{u}_i, \mathbf{v}_j, \boldsymbol{\theta}). \tag{29}$$

The density of (y_{ij}, z_{ij}) is given by Eq.(5). We denote $\alpha = p(z_{ij} = 1 \,|\, \boldsymbol{\theta})$; then, we obtain $p(z_{ij} = 0 \,|\, \boldsymbol{\theta}) = 1 - \alpha$. The conditional densities of an inlier y_{ij} and an outlier y_{ij} are given by

$$p(y_{ij} \,|\, z_{ij} = 1, \mathbf{u}_i, \mathbf{v}_j, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2\right\}, \tag{30a}$$

$$p(y_{ij} \,|\, z_{ij} = 0, \boldsymbol{\theta}) = \gamma. \tag{30b}$$

The substitution of these into Eq.(29) yields

$$p(\mathbf{Y}, \mathbf{Z} \,|\, \boldsymbol{\Theta}) = \prod_{ij} \left[\frac{\alpha}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2\right\}\right]^{z_{ij}} \times \{(1-\alpha)\gamma\}^{1-z_{ij}}. \tag{31}$$

As described in Secion 3, we introduce the variational posterior $q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})$ that is intended to approximate the true posterior $p(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y})$. We consider a class of separable q, i.e.,

$$q(\mathbf{Z}, \boldsymbol{\Theta} \,|\, \mathbf{Y}) = q(\mathbf{Z} \,|\, \mathbf{Y}) q(\boldsymbol{\Theta} \,|\, \mathbf{Y}) \tag{32}$$

We further assume that $q(\boldsymbol{\Theta} \,|\, \mathbf{Y})$ is also separable w.r.t. each parameter: $q(\boldsymbol{\Theta} \,|\, \mathbf{Y}) = q(\mathbf{U}) q(\mathbf{V}) q(\alpha) q(\sigma^2)$. Then, we intend to find an optimal q within this class that is closest to the true posterior p. An optimal q is obtained by solving the variational equation, which is given by Eqs.(23) [1]. The posterior $q(\mathbf{Z} \,|\, \mathbf{Y})$ of the hidden variable is given by

$$q(\mathbf{Z} \,|\, \mathbf{Y}) \propto \exp\langle \log p(\mathbf{Y}, \mathbf{Z} \,|\, \boldsymbol{\Theta}) \rangle_{\boldsymbol{\Theta}}. \tag{33}$$

The logarithm is expanded as

$$\log p(\mathbf{Y}, \mathbf{Z} \,|\, \boldsymbol{\Theta}) = \sum_{i,j} \left[z_{ij} \left\{ \log \frac{\alpha}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(y_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right\} + (1 - z_{ij}) \log\{(1-\alpha)\gamma\} \right], \tag{34}$$

Therefore, the posterior is represented as

$$q(\mathbf{Z} \,|\, \mathbf{Y}) \propto \prod_{i,j} \exp\{a_{ij} z_{ij} + b_{ij}(1 - z_{ij})\}, \tag{35}$$

where

$$a_{ij} = -\frac{1}{2}\log 2\pi + \langle \log \alpha \rangle - \frac{1}{2}\left\langle \log \sigma^2 \right\rangle - \frac{1}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \left\langle (y_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \right\rangle$$

and

$$b_{ij} = \langle \log(1 - \alpha) \rangle \langle \log \gamma \rangle,$$

where $\langle \cdot \rangle$ represents $\langle \cdot \rangle_{\boldsymbol{\Theta}}$. All the terms in $\langle \cdot \rangle$ will be evaluated below by using the parameter posterior $q(\boldsymbol{\Theta} \,|\, \mathbf{Y})$. Similarly, when evaluating the parameter posterior, the expectation $\langle z_{ij} \rangle$ is required. It is calculated from the identity $q(z_{ij} = 1) + q(z_{ij} = 0) = 1$ as

$$\langle z_{ij} \rangle = 1 \cdot q(z_{ij} = 1) = \frac{\exp(a_{ij})}{\exp(a_{ij}) + \exp(b_{ij})} \tag{36}$$

An optimal solution for the parameter posterior $q(\boldsymbol{\Theta} \,|\, \mathbf{Y}) = q(\mathbf{U}) q(\mathbf{V}) q(\sigma^2) q(\alpha)$ is given by Eq.(23b). Owing to the separability, we may consider each parameter in turn. As for $\mathbf{U}$, its posterior is given by

$$q(\mathbf{U} \,|\, \mathbf{Y}) \propto \exp\langle \log p(\mathbf{Y}, \mathbf{Z} \,|\, \mathbf{U}, \mathbf{V}, \boldsymbol{\theta}) \rangle_{\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}} = \prod_{ij} \exp\left\{ -\frac{1}{2}\left\langle \frac{1}{\sigma^2} \right\rangle \langle z_{ij} \rangle (y_{ij}^2 - 2 y_{ij} \mathbf{u}_i^\top \langle \mathbf{v}_j \rangle + \mathbf{u}_i^\top \langle \mathbf{v}_j \mathbf{v}_j^\top \rangle \mathbf{u}_i \right\}.$$

Note that the expectation $\langle \cdot \rangle_{\mathbf{Z}, \mathbf{V}, \boldsymbol{\theta}}$ is w.r.t. not just $q(\mathbf{Z})$ but $q(matV)$ and $q(\boldsymbol{\theta})$. (Every $\langle \cdot \rangle$ above represents an expectation with the same significance.) Thus, $\mathbf{U}$ ($\mathbf{u}_i$) follows a normal density, whose sufficient statistics, $\langle \mathbf{u}_i \rangle$ and $\langle \mathbf{u}_i \mathbf{u}_i^\top \rangle$, are computed as in Eqs.(27a) and (27b). ($\langle \mathbf{u}_i \rangle$ and $\langle \mathbf{u}_i \mathbf{u}_i^\top \rangle$ are expressed as $\mathbf{u}_i$ and $\boldsymbol{\Phi}_i$, respectively.) By a similar procedure, we find that $\mathbf{V}$ follows a normal density and its sufficient statistics are computed as in Eqs.(28a) and (28b). ($\langle \mathbf{v}_j \rangle$ and $\langle \mathbf{v}_j \mathbf{v}_j^\top \rangle$ are expressed as $\mathbf{v}_j$ and $\boldsymbol{\Psi}_j$, respectively.)

Here, we do not treat σ as a variable, but rather its square σ^2. Similarly, its posterior is given by $q(\sigma^2) \propto \exp\langle \log p(\mathbf{Y}, \mathbf{Z} \,|\, \mathbf{U}, \mathbf{V}, \boldsymbol{\theta}) \rangle_{\mathbf{Z}, \mathbf{U}, \mathbf{V}, \alpha}$. For simplicity, we define $n_s \equiv \sum_{i,j} \langle z_{ij} \rangle$ and $r_s \equiv \sum_{i,j} \langle z_{ij} (y_{ij} - \mathbf{u}_i^\top \mathbf{v}_j)^2 \rangle$. Then, the posterior is given by

$$q(\sigma^2) = C_\sigma (\sigma^2)^{-\frac{1}{2} n_s} \exp\left(-\frac{1}{2\sigma^2} r_s\right), \tag{37}$$

where C_σ is a normalization factor. From $\int_0^\infty q(\sigma^2)d\sigma^2 = 1$, C_σ is given by

$$C_\sigma = \frac{1}{(\frac{2}{r_s})^{\frac{1}{2}n_s-1}\Gamma\left(\frac{1}{2}n_s-1\right)}, \tag{38}$$

where $\Gamma(x)$ is the gamma function ($\Gamma(x) \equiv \int_0^\infty t^{x-1}\exp(-t)dt$). Thus, we have

$$q(\sigma^2) = \frac{(\sigma^2)^{-\frac{1}{2}n_s}\exp\left(-\frac{1}{2\sigma^2}r_s\right)}{\left(\frac{2}{r_s}\right)^{\frac{n_s}{2}-1}\Gamma\left(\frac{n_s}{2}-1\right)} \tag{39}$$

Using this, the expectations $\langle 1/\sigma^2\rangle$ and $\langle \log\sigma^2\rangle$ that appeared on a_{ij} can be evaluated as follows:

$$\left\langle \frac{1}{\sigma^2}\right\rangle = \frac{2\Gamma\left(\frac{n_s}{2}\right)}{r_s\Gamma\left(\frac{n_s}{2}-1\right)} = \frac{n_s-2}{r_s}, \tag{40a}$$

$$\langle \log\sigma^2\rangle = \log\left(\frac{r_s}{2}\right) - \psi\left(\frac{n_s}{2}-1\right) \approx \log\left(\frac{r_s}{n_s-2}\right), \tag{40b}$$

where $\psi(\cdot)$ is the digamma function defined as $\psi(x) \equiv \frac{d}{dx}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. The approximation $\psi(x) \approx \log(x)$ holds when x is large.

By a similar procedure, it is shown that the last parameter α follows a Dirichlet distribution:

$$q(\alpha) \propto \alpha^{n_s}(1-\alpha)^{mn-n_s}. \tag{41}$$

Using the formulae for the Dirichlet distribution ($E[\log\alpha] = \psi(r_1) - \psi(r_1+r_2)$ for $p(\alpha) \propto (\alpha)^{r_1-1}(1-\alpha)^{r_2-1}$), we have

$$\langle \log\alpha\rangle = \psi(n_s+1) - \psi(mn+2) \approx \log\frac{n_s+1}{mn+2}, \tag{42a}$$

$$\begin{aligned}\langle \log(1-\alpha)\rangle &= \psi(mn-n_s+1) - \psi(mn+2)\\ &\approx \log\frac{mn-n_s+1}{mn+2},\end{aligned} \tag{42b}$$

where $\approx$ holds when mn is sufficiently large.

In the VB methodology, the evaluation of the hidden variable posterior $q(\mathbf{Z})$ and that of the parameter posterior $q(\mathbf{\Theta})$ are performed in an alternate manner. The computations for each of the evaluation steps are given by the above-mentioned derivation. Although they are somewhat complicated in their original form, by assuming that n_s and mn are large and by then employing the above-mentioned approximations for σ^2 and α, they can be reduced to a compact form as follows. We first denote $w_{ij} \equiv \langle z_{ij}\rangle$ and

$$\begin{aligned}e_{ij}^2 &\equiv \left\langle (y_{ij} - \mathbf{u}_i^\top\mathbf{v}_j)^2\right\rangle\\ &= y_{ij}^2 - 2y_{ij}(\langle\mathbf{u}_i^\top\rangle\langle\mathbf{v}_j\rangle) + \mathrm{tr}(\langle\mathbf{u}_i\mathbf{u}_i^\top\rangle\langle\mathbf{v}_j\mathbf{v}_j^\top\rangle).\end{aligned}$$

When the approximation is valid, the expectations for σ^2 in Eqs.(40) yield an identical equation. By redefining $1/\sigma^2 \equiv \langle 1/\sigma^2\rangle$, we may write

$$\sigma^2 = \frac{\sum_{i,j} w_{ij}e_{i,j}^2}{\sum_{i,j} w_{ij}}.$$

Similarly, Eqs.(42) are identical equations; we may write

$$\alpha = \frac{\sum_{i,j} w_{ij} + 1}{\sum_{i,j} 1 + 2}.$$

Using these results, we can also rewrite a_{ij} and b_{ij} in Eq.(36) as

$$a_{ij} = \alpha\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{e_{ij}^2}{2\sigma^2}\right\} \quad \text{and} \tag{43a}$$

$$b_{ij} = (1-\alpha)\gamma. \tag{43b}$$

Further, by rewriting $\mathbf{u}_i = \langle\mathbf{u}_i\rangle$, $\mathbf{\Psi}_i = \langle\mathbf{u}_i\mathbf{u}_i^\top\rangle$, $\mathbf{v}_j = \langle\mathbf{v}_j\rangle$, and $\mathbf{\Phi}_j = \langle\mathbf{v}_j\mathbf{v}_j^\top\rangle$, we arrive at Algorithm 5.

References

[1] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999. 2, 4, 7

[2] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2005. 5

[3] P. Chen and D. Suter. Recovering the missing components in a large noisy low-rank matrix: application to SFM. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(8):1051–1063, 2004. 1

[4] F. de la Torre and M. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1):117–142, 2003. 1, 2, 5

[5] Z. Ghahramani and G. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, Dept. Of Computer Science, University of Tronto, 1997. 3

[6] A. Gruber and Y. Weiss. Multibody factorization with uncertainty and missing data using the EM algorithm. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2004. 3

[7] T. Okatani and K. Deguchi. On the Wiberg algorithm for matrix factorization in the presence of missing components. *International Journal of Computer Vision*, page (online version), 2006. 1, 5

[8] S. Roweis. EM algorithms for PCA and SPCA. In *NIPS*, 1997. 2, 3

[9] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analysers. *Neural computation*, 11(2):443–482, 1999. 3

[10] T. Wiberg. Computation of principal components when data are missing. In *Proceedings Symposium of Comp. Stat.*, pages 229–326, 1976. 1