Theoretical Connection between Locally Linear Embedding, Factor Analysis, and Probabilistic PCA

Benyamin Ghojogh[†,*], Ali Ghodsi[‡], Fakhri Karray[†], Mark Crowley[†]
[†]Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada
[‡]Department of Statistics and Actuarial Science, University of Waterloo, ON, Canada

Abstract

Locally Linear Embedding (LLE) is a nonlinear spectral dimensionality reduction and manifold learning method. It has two main steps which are linear reconstruction and linear embedding of points in the input space and embedding space, respectively. In this work, we look at the linear reconstruction step from a stochastic perspective where it is assumed that every data point is conditioned on its linear reconstruction weights as latent factors. The stochastic linear reconstruction of LLE is solved using expectation maximization. We show that there is a theoretical connection between three fundamental dimensionality reduction methods, i.e., LLE, factor analysis, and probabilistic Principal Component Analysis (PCA). The stochastic linear reconstruction of LLE is formulated similar to the factor analysis and probabilistic PCA. It is also explained why factor analysis and probabilistic PCA are linear and LLE is a nonlinear method. This work combines and makes a bridge between two broad approaches of dimensionality reduction, i.e., the spectral and probabilistic algorithms.

Keywords: locally linear embedding, factor analysis, probabilistic principal component analysis, manifold learning, dimensionality reduction

1. Introduction

Dimensionality reduction and manifold learning methods are widely useful for feature extraction, manifold unfolding, and data visualization. The dimensionality reduction methods can be divided into three main categories, i.e., spectral methods, probabilistic methods, and neural network-based methods. An example for spectral methods is Locally Linear Embedding (LLE) [1–3]. Examples for probabilistic methods are factor analysis [4, 5] and probabilistic Principal Component Analysis (PCA) [6, 7]. An example for neural network-based methods is variational autoencoder [8] which formulates variational inference [9–12] in an autoencoder framework. In this short paper, we show theoretical connections between the fundamental methods of LLE, factor analysis, and probabilistic PCA. Hence, we connect the spectral and probabilistic approaches of dimensionality reduction. We also explain why factor analysis and probabilistic PCA are linear and LLE is a nonlinear method. In Section 2, we review the required background. Section 3 models the linear reconstruction of LLE stochastically. Finally, section 4 concludes the paper and discusses the connections of factor analysis, probabilistic PCA, and LLE algorithms.

2. Technical Background and Preliminaries

2.1. Marginal Multivariate Gaussian Distribution

Consider two random variables $\boldsymbol{x}_1 \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{d_2}$ and let $\boldsymbol{x}_3 := [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top]^\top \in \mathbb{R}^{d_1+d_2}$. Assume that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are jointly multivariate Gaussian, i.e., $\boldsymbol{x}_3 \sim \mathcal{N}(\boldsymbol{x}_3; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$. The

*Corresponding author: bghojogh@uwaterloo.ca

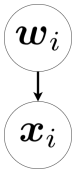

Figure 1. The probabilistic graphical model for stochastic linear reconstruction in LLE.

mean and covariance can be decomposed as:

$$\boldsymbol{\mu}_3 = [\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top]^\top \in \mathbb{R}^{d_1+d_2}, \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \in \mathbb{R}^{(d_1+d_2)\times(d_1+d_2)}, \tag{2.1}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^{d_1}$, $\boldsymbol{\mu}_2 \in \mathbb{R}^{d_2}$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{d_1 \times d_2}$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{d_2 \times d_2}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{d_1 \times d_2}$, and $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^\top$.

It can be shown that the marginal distributions for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are Gaussian distributions where $\mathbb{E}[\boldsymbol{x}_1] = \boldsymbol{\mu}_1$ and $\mathbb{E}[\boldsymbol{x}_2] = \boldsymbol{\mu}_2$. The covariance matrix of the joint distribution can be simplified as [12]:

$$\boldsymbol{\Sigma}_3 = \mathbb{E}\left[\begin{bmatrix} (\boldsymbol{x}_1 - \boldsymbol{\mu}_1)(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^\top, (\boldsymbol{x}_1 - \boldsymbol{\mu}_1)(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^\top \\ (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^\top, (\boldsymbol{x}_2 - \boldsymbol{\mu}_2)(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)^\top \end{bmatrix} \right], \tag{2.2}$$

where $\mathbb{E}[.]$ is the expectation operator. According to the definition of the multivariate Gaussian distribution, the conditional distribution is also a Gaussian distribution, i.e., $\boldsymbol{x}_2|\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{x}_2; \boldsymbol{\mu}_{x_2|x_1}, \boldsymbol{\Sigma}_{x_2|x_1})$ where [12]:

$$\mathbb{R}^{d_2} \ni \boldsymbol{\mu}_{x_2|x_1} := \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1), \tag{2.3}$$

$$\mathbb{R}^{d_2 \times d_2} \ni \boldsymbol{\Sigma}_{x_2|x_1} := \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}, \tag{2.4}$$

and likewise we have for $\boldsymbol{x}_1|\boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\mu}_{x_1|x_2}, \boldsymbol{\Sigma}_{x_1|x_2})$. Also, note that the probability density function of d-dimensional Gaussian distribution is:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\Big(-\frac{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}{2}\Big), \tag{2.5}$$

where $|.|$ denotes the determinant of matrix.

2.2. **Variational Inference**

Assume data $\boldsymbol{x}_i$ is conditioned on a latent factor $\boldsymbol{w}_i$, as shown in Fig. 1. Let the parameters of model be denoted by $\boldsymbol{\theta}$. In variational inference, the Evidence Lower Bound (ELBO) is a lower bound on the log likelihood of data and is defined as minus Kullback-Leibler (KL) divergence between a distribution $q(.)$ on $\boldsymbol{w}_i$ and the joint distribution of $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, i.e., $\mathcal{L}(q, \boldsymbol{\theta}) := -\mathrm{KL}\big(q(\boldsymbol{w}_i) \,\|\, \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{w}_i \,|\, \boldsymbol{\theta})\big)$ [8, 9, 12]. Maximizing this lower bound results in maximization of likelihood of data [10, 11]. Variational inference uses EM for Maximum Likelihood Estimation (MLE) [12]:

$$q^{(t)}(\boldsymbol{w}_i) \leftarrow \mathbb{P}(\boldsymbol{w}_i \,|\, \boldsymbol{x}_i, \boldsymbol{\theta}^{(t-1)}), \tag{2.6}$$

$$\boldsymbol{\theta}^{(t)} \leftarrow \arg\max_{\boldsymbol{\theta}} \; \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}\big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{w}_i \,|\, \boldsymbol{\theta})\big], \tag{2.7}$$

where $\mathbb{E}[.]$ denotes expectation. Here, we use the EM approach of variational inference for stochastic linear reconstruction in LLE.

2.3. **Factor Analysis**

Factor analysis [4, 5] assumes that every data point $\boldsymbol{x}_i$ is generated from a latent factor $\boldsymbol{w}_i$. Its probabilistic graphical model is similar to Fig. 1 but with a small difference [13].

It assumes $\boldsymbol{x}_i$ is obtained by linear projection of $\boldsymbol{w}_i$ onto the space of data by projection matrix $\boldsymbol{\Lambda}$, then applying some linear translation, and finally adding a Gaussian noise $\boldsymbol{\epsilon}$ with covariance matrix $\boldsymbol{\Psi}$. This addition of noise is the main difference from the model depicted in Fig. 1. If $\boldsymbol{\mu}$ denotes the mean of data, factor analysis considers:

$$\boldsymbol{x}_i := \boldsymbol{\Lambda}\boldsymbol{w}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{2.8}$$

$$\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{w}_i, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\Psi}) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\Lambda}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}), \tag{2.9}$$

where $\mathbb{P}(\boldsymbol{w}_i) = \mathcal{N}(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{I})$ and $\mathbb{P}(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{\Psi})$. Factor analysis uses EM algorithm for finding optimum $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ (see [12] for details of EM in factor analysis). We will show that the stochastic linear reconstruction in LLE, for modeling the relation between a data point and its latent reconstruction weights, is similar to Eq. (2.8).

2.4. **Probabilistic PCA**

Probabilistic PCA [6, 7] is a special case of factor analysis where the variance of noise is equal in all dimensions of data space with covariance between dimensions, i.e.:

$$\boldsymbol{\Psi} = \sigma^2 \boldsymbol{I}. \tag{2.10}$$

In other words, probabilistic PCA considers an isotropic noise model. Similar to factor analysis, it can be solved iteratively using EM [6]. However, one can also find a closed-form solution to its EM approach [7]. Hence, by restricting the noise covariance to be isotropic, its solution becomes simpler and closed-form. See [12] for details of derivations and solutions for probabilistic PCA.

3. **Stochastic Modeling of Linear Reconstruction in LLE**

3.1. **Notations and Joint and Conditional Distributions**

LLE [1, 2] has two main steps which are linear reconstruction and linear embedding [3]. As Fig. 1 depicts, the linear reconstruction step of LLE can be seen stochastically where every point $\boldsymbol{x}_i$ is conditioned on and generated by its reconstruction weights $\boldsymbol{w}_i$ as a latent factor. Therefore, $\boldsymbol{x}_i$ can be written as a stochastic function of $\boldsymbol{w}_i$ where we assume $\boldsymbol{w}_i$ has a multivariate Gaussian distribution:

$$\boldsymbol{x}_i = \boldsymbol{X}_i \boldsymbol{w}_i + \boldsymbol{\mu}, \tag{3.1}$$

$$\mathbb{P}(\boldsymbol{w}_i) = \mathcal{N}(\boldsymbol{w}_i; \boldsymbol{0}, \boldsymbol{\Omega}_i) \implies \mathbb{E}[\boldsymbol{w}_i] = \boldsymbol{0},\ \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] = \boldsymbol{\Omega}_i, \tag{3.2}$$

where $\boldsymbol{\Omega}_i \in \mathbb{R}^{k \times k}$ is covariance of $\boldsymbol{w}_i$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean of data because $(1/n) \sum_{i=1}^n \boldsymbol{x}_i = \mathbb{E}[\boldsymbol{x}_i] = \boldsymbol{X}_i \mathbb{E}[\boldsymbol{w}_i] + \boldsymbol{\mu} \overset{(3.2)}{=} \boldsymbol{0} + \boldsymbol{\mu} = \boldsymbol{\mu}$. According to Eq. (2.2), for the joint distribution of $[\boldsymbol{x}_i^\top, \boldsymbol{w}_i^\top]^\top \in \mathbb{R}^{d+k}$, we have:

$$\boldsymbol{\Sigma}_{11} = \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top] \overset{(3.1)}{=} \mathbb{E}[(\boldsymbol{X}_i \boldsymbol{w}_i)(\boldsymbol{X}_i \boldsymbol{w}_i)^\top] = \boldsymbol{X}_i \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] \boldsymbol{X}_i^\top \overset{(3.2)}{=} \boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top,$$

$$\boldsymbol{\Sigma}_{12} = \mathbb{E}[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{w}_i - \boldsymbol{0})^\top] \overset{(3.1)}{=} \boldsymbol{X}_i \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] \overset{(3.2)}{=} \boldsymbol{X}_i \boldsymbol{\Omega}_i,$$

$$\boldsymbol{\Sigma}_{22} = \mathbb{E}[(\boldsymbol{w}_i - \boldsymbol{0})(\boldsymbol{w}_i - \boldsymbol{0})^\top] = \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] \overset{(3.2)}{=} \boldsymbol{\Omega}_i.$$

Hence:

$$\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{w}_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{w}_i \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top & \boldsymbol{X}_i \boldsymbol{\Omega}_i \\ \boldsymbol{\Omega}_i^\top \boldsymbol{X}_i^\top & \boldsymbol{\Omega}_i \end{bmatrix} \right). \tag{3.3}$$

We have:

$$\mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{w}_i, \boldsymbol{\Omega}_i) \overset{(a)}{=} \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{X}_i \boldsymbol{w}_i + \boldsymbol{\mu}, \boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top), \tag{3.4}$$

where (a) is because of Eqs. (3.1) and (3.3). We can use EM for MLE in stochastic linear reconstruction of LLE. In the following, the steps of EM are explained.

3.2. **E-Step in Expectation Maximization**

As we will see later in the M-step of EM, we will have two expectation terms which need to be computed in the E-step. These expectations, which are over the $q(\boldsymbol{w}_i) := \mathbb{P}(\boldsymbol{w}_i \,|\, \boldsymbol{x}_i)$ distribution, are $\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}[\boldsymbol{w}_i] \in \mathbb{R}^k$ and $\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] \in \mathbb{R}^{k \times k}$ where t denotes the iteration index in EM iterations. According to Eqs. (2.3), (2.4), and (3.3), we have:

$$\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}[\boldsymbol{w}_i] = \boldsymbol{\mu}_{w|x} = \boldsymbol{\Omega}_i^\top \boldsymbol{X}_i^\top (\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)^\dagger (\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{3.5}$$

$$\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] = \boldsymbol{\Sigma}_{w|x} = \boldsymbol{\Omega}_i - \boldsymbol{\Omega}_i^\top \boldsymbol{X}_i^\top (\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)^\dagger \boldsymbol{X}_i \boldsymbol{\Omega}_i, \tag{3.6}$$

where † denotes either inverse or pseudo-inverse of matrix.

3.3. **M-Step in Expectation Maximization**

In M-step of EM, we maximize the joint likelihood of data and weights over all n data points where the optimization variable is the covariance of prior distribution of weights:

$$\begin{aligned}
&\max_{\{\boldsymbol{\Omega}_i\}_{i=1}^n} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\log \mathbb{P}(\boldsymbol{x}_i, \boldsymbol{w}_i \,|\, \boldsymbol{\Omega}_i)\big] \\
&\overset{(a)}{=} \max_{\{\boldsymbol{\Omega}_i\}_{i=1}^n} \sum_{i=1}^n \Big(\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\log \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{w}_i, \boldsymbol{\Omega}_i)\big] + \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\log \mathbb{P}(\boldsymbol{w}_i)\big]\Big) \\
&\overset{(b)}{=} \max_{\{\boldsymbol{\Omega}_i\}_{i=1}^n} \sum_{i=1}^n \Big(\mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\log \mathcal{N}(\boldsymbol{X}_i \boldsymbol{w}_i + \boldsymbol{\mu}, \boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)\big] + \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\log \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_i))\big]\Big) \\
&\overset{(2.5)}{=} \max_{\{\boldsymbol{\Omega}_i\}_{i=1}^n} \Big(\underbrace{-\frac{d\,n}{2} \log(2\pi)}_{\text{constant}} - \frac{n}{2} \log |\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top| \\
&- \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[(\boldsymbol{x}_i - \boldsymbol{X}_i \boldsymbol{w}_i - \boldsymbol{\mu})^\top (\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)^{-1} (\boldsymbol{x}_i - \boldsymbol{X}_i \boldsymbol{w}_i - \boldsymbol{\mu})\big] \\
&- \underbrace{\frac{k\,n}{2} \log(2\pi)}_{\text{constant}} - \frac{n}{2} \log |\boldsymbol{\Omega}_i| - \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\boldsymbol{w}_i^\top \boldsymbol{\Omega}_i^{-1} \boldsymbol{w}_i\big]\Big) \\
&\overset{(c)}{=} \max_{\{\boldsymbol{\Omega}_i\}_{i=1}^n} \Big(- \frac{n}{2} \log |\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top| - \frac{n}{2} \mathbf{tr}\big((\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)^{-1} \boldsymbol{S}_1\big) - \frac{n}{2} \log |\boldsymbol{\Omega}_i| - \frac{n}{2} \mathbf{tr}\big(\boldsymbol{\Omega}_i^{-1} \boldsymbol{S}_2\big)\Big),
\end{aligned} \tag{3.7}$$

where (a) is because of the chain rule $\mathbb{P}(\boldsymbol{x}_i, \boldsymbol{w}_i \,|\, \boldsymbol{\Omega}_i) = \mathbb{P}(\boldsymbol{x}_i \,|\, \boldsymbol{w}_i, \boldsymbol{\Omega}_i)\, \mathbb{P}(\boldsymbol{w}_i)$, and (b) is because of Eqs. (3.2) and (3.4), and (c) is because we define the scatters $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ as:

$$\begin{aligned}
&\mathbb{R}^{d \times d} \ni \boldsymbol{S}_1 := \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[(\boldsymbol{x}_i - \boldsymbol{X}_i \boldsymbol{w}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{X}_i \boldsymbol{w}_i - \boldsymbol{\mu})^\top\big] \\
&= \frac{1}{n} \sum_{i=1}^n \Big((\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top - 2 \boldsymbol{X}_i \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\boldsymbol{w}_i\big] (\boldsymbol{x}_i - \boldsymbol{\mu})^\top + \boldsymbol{X}_i \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)} \big[\boldsymbol{w}_i \boldsymbol{w}_i^\top\big] \boldsymbol{X}_i^\top \Big),
\end{aligned} \tag{3.8}$$

$$\mathbb{R}^{k\times k} \ni \boldsymbol{S}_2 := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\sim q^{(t)}(\boldsymbol{w}_i)}\big[\boldsymbol{w}_i \boldsymbol{w}_i^\top\big], \tag{3.9}$$

where the expectation terms are found by Eqs. (3.5) and (3.6). The gradient of the joint likelihood is:

$$\begin{aligned}\mathbb{R}^{k\times k} \ni \frac{\partial \text{ Eq. (3.7)}}{\partial \boldsymbol{\Omega}_i^{-1}} =& \frac{n}{2}\Big[\mathbf{vec}_{k\times k}^{-1}\big[\boldsymbol{T}_i\, \mathbf{vec}_{d^2\times 1}(\boldsymbol{X}_i \boldsymbol{\Omega}_i \boldsymbol{X}_i^\top)\big] \\ & - \mathbf{vec}_{k\times k}^{-1}\big[\boldsymbol{T}_i\, \mathbf{vec}_{d^2\times 1}(\boldsymbol{S}_1)\big] + \boldsymbol{\Omega}_i - \boldsymbol{S}_2\Big],\end{aligned} \tag{3.10}$$

where we use the Magnus-Neudecker convention in which matrices are vectorized, $\mathbf{vec}(.)$ vectorizes the matrix, $\mathbf{vec}_{k\times k}^{-1}(.)$ is de-vectorization to $k\times k$ matrix, $\otimes$ denotes the Kronecker product, and $\mathbb{R}^{k^2\times d^2} \ni \boldsymbol{T}_i := \boldsymbol{X}_i^\top \otimes \boldsymbol{X}_i^\top$. In the M-step, one can update the variables $\{\boldsymbol{\Omega}_i\}_{i=1}^n$ using gradient descent with the gradient in Eq. (3.10). However, we can relax the covariance matrix and simplify the algorithm.

3.4. **Relaxation of Covariance**

Inspired by relaxation of factor analysis for probabilistic PCA (see Eq. (2.10)), we can relax the covariance matrix to be spherical, i.e., diagonal and the variance of weights to be equal in all k dimensions:

$$\boldsymbol{\Omega}_i = \sigma_i \boldsymbol{I} \in \mathbb{R}^{k\times k}. \tag{3.11}$$

Substituting this covariance into Eq. (3.7) and noticing the properties of determinant and trace gives:

$$\begin{aligned}&\max_{\{\sigma_i\}_{i=1}^n} \Big(-\frac{n}{2}\log(\sigma_i^d |\boldsymbol{X}_i\boldsymbol{X}_i^\top|) - \frac{n}{2}\sigma_i^{-1}\mathbf{tr}\big((\boldsymbol{X}_i\boldsymbol{X}_i^\top)^{-1}\boldsymbol{S}_1\big) - \frac{n}{2}\log\sigma_i^k - \frac{n}{2}\sigma_i^{-1}\mathbf{tr}\big(\boldsymbol{S}_2\big)\Big) \\ &\overset{(a)}{=} \max_{\{\sigma_i\}_{i=1}^n} \Big(-\frac{n}{2}(d+k)\log(\sigma_i) - \frac{n}{2}\Big[\mathbf{tr}\big((\boldsymbol{X}_i\boldsymbol{X}_i^\top)^{-1}\boldsymbol{S}_1\big) - \mathbf{tr}\big(\boldsymbol{S}_2\big)\Big]\sigma_i^{-1}\Big),\end{aligned} \tag{3.12}$$

where (a) is because $\log(\sigma_i^d|\boldsymbol{X}_i\boldsymbol{X}_i^\top|) = d\log(\sigma_i) + \log(|\boldsymbol{X}_i\boldsymbol{X}_i^\top|)$ whose second term is a constant w.r.t. σ_i. Setting the gradient of the joint likelihood to zero gives:

$$\begin{aligned}&\mathbb{R} \ni \frac{\partial \text{ Eq. (3.12)}}{\partial \sigma_i^{-1}} = \frac{n}{2}\Big[(d+k)\sigma_i - \Big(\mathbf{tr}\big((\boldsymbol{X}_i\boldsymbol{X}_i^\top)^{-1}\boldsymbol{S}_1\big) + \mathbf{tr}(\boldsymbol{S}_2)\Big)\Big] \overset{\text{set}}{=} 0 \\ &\implies \sigma_i = (d+k)^{-1}\Big(\mathbf{tr}\big((\boldsymbol{X}_i\boldsymbol{X}_i^\top)^{\dagger}\boldsymbol{S}_1\big) + \mathbf{tr}(\boldsymbol{S}_2)\Big),\end{aligned} \tag{3.13}$$

where † denotes either inverse or pseudo-inverse of matrix. As we also have in probabilistic PCA, the relaxation of covariance matrix results in the closed-form solution of Eq. (3.13). Without this relaxation, the solution of M-step in EM algorithm of LLE is solved iteratively by the gradient in Eq. (3.10).

The EM algorithm for stochastic linear reconstruction in LLE is summarized in Algorithm 1. We can sample $\{\boldsymbol{w}_i\}_{i=1}^n$ with the following prior and conditional distributions:

$$\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{w}_i; \boldsymbol{0}, \sigma_i \boldsymbol{I}), \tag{3.14}$$

$$\boldsymbol{w}_i \,|\, \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{w}_i; \boldsymbol{\mu}_{w|x}, \boldsymbol{\Sigma}_{w|x}), \tag{3.15}$$

where $\boldsymbol{\mu}_{w|x}$ and $\boldsymbol{\Sigma}_{w|x}$ are defined in Eqs. (3.5) and (3.6), respectively.

4. **Conclusion and Discussion on the Connection of Methods**

Comparing Eqs. (2.8) and (3.1) shows that data point $\boldsymbol{x}_i$ is conditioned on some latent variable $\boldsymbol{w}_i$, in all methods of factor analysis [4, 5], probabilistic PCA [6, 7], and LLE [1, 2]. In these methods, the latent variable $\boldsymbol{w}_i$ is related to data $\boldsymbol{x}_i$ using a transformation

Input: $\{\boldsymbol{x}_i\}_{i=1}^n$, kNN graph or $\{\boldsymbol{X}_i\}_{i=1}^n$
Initialize: $\{\boldsymbol{\Omega}_i\}_{i=1}^n = \boldsymbol{I}$
while *not converged* **do**
  // E-step:
  for *every data point* i *from* 1 *to* n **do**
    Calculate expectations by Eqs. (3.5) and (3.6)
  // Sampling:
  Sample weights $\{\boldsymbol{w}_i\}_{i=1}^n$ using Eq. (3.15)
  // M-step:
  Calculate $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ using Eqs. (3.8) and (3.9)
  for *every data point* i *from* 1 *to* n **do**
    Calculate σ_i by Eq. (3.13)
    Calculate $\boldsymbol{\Omega}_i$ using Eq. (3.11)
Return weights $\{\boldsymbol{w}_i\}_{i=1}^n$

Algorithm 1: Stochastic Linear Reconstruction in LLE with EM

matrix. In factor analysis and probabilistic PCA (see Eq. (2.8)), this transformation matrix is a global matrix $\boldsymbol{\Lambda}$ so it is data-independent in the sense that it is the same matrix for all data points. However, the transformation matrix of LLE is $\boldsymbol{X}_i$ (see Eq. (3.1)) which is local and data-dependent in the sense that it is different for every data point. This explains why factor analysis and probabilistic PCA are linear methods and LLE is a nonlinear algorithm. Moreover, in this framework, if the covariance matrix is relaxed to be spherical (see Eqs. (2.10) and (3.11)), the solution becomes closed-form. This relaxation happens in probabilistic PCA and can also be used in LLE to have a closed-form solution (see Eq. (3.13)). Finally, the introduced relation of the three algorithms opens the door for more investigation in relation of spectral and probabilistic approaches of machine learning.

References

[1] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding". In: *Science* 290.5500 (2000), pp. 2323–2326.

[2] L. K. Saul and S. T. Roweis. "Think globally, fit locally: unsupervised learning of low dimensional manifolds". In: *Journal of machine learning research* 4.Jun (2003), pp. 119–155.

[3] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. "Locally Linear Embedding and its Variants: Tutorial and Survey". In: *arXiv preprint arXiv:2011.10925* (2020).

[4] B. Fruchter. *Introduction to factor analysis*. Van Nostrand, 1954.

[5] D. Child. *The essentials of factor analysis*. Cassell Educational, 1990.

[6] S. Roweis. "EM algorithms for PCA and SPCA". In: *Advances in neural information processing systems* 10 (1997), pp. 626–632.

[7] M. E. Tipping and C. M. Bishop. "Probabilistic principal component analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 611–622.

[8] D. P. Kingma and M. Welling. "Auto-encoding variational Bayes". In: *International Conference on Learning Representations*. 2014.

[9] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[10] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[11] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.

[12] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. "Factor Analysis, Probabilistic Principal Component Analysis, Variational Inference, and Variational Autoencoder: Tutorial and Survey". In: *arXiv preprint arXiv:2101.00734* (2021).

[13] Z. Ghahramani and G. E. Hinton. *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto, 1996.