A Unified Framework for Probabilistic Component Analysis

Stefanos Zafeiriou
Imperial College London, UK
Mihalis A. Nicolaou
Imperial College London, UK
Maja Pantic
Imperial College London, UK

Abstract

In this paper, we present a unifying framework which reduces the construction of probabilistic component analysis techniques to a mere selection of the latent neighbourhood via the prior, thus providing an elegant and principled framework for creating novel component analysis models. Under our framework, we unify many very popular and well-studied component analysis algorithms, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA). We firstly show that the projection directions produced by all the aforementioned methods are also produced by the Maximum Likelihood (ML) solution of a single joint probability density function (PDF) just by choosing an appropriate prior over the latent space in our framework. Subsequently, we propose novel Expectation Maximization (EM) algorithms utilising the proposed joint PDF. Theoretical analysis and experiments show the usefulness of the proposed framework.

1. Introduction

Unification frameworks in machine learning do not only provide valuable material towards the deeper understanding of various methodologies, but also form a flexible basis upon which further extensions can be easily built. One of the first attempts to unify methodologies was made in (Roweis & Ghahramani, 1999). In this seminal work, models such as Factor analysis (FA), Principal Component Analysis (PCA), mixtures of Gaussian clusters (MGC), vector quantization (VQ), Linear Dynamic Systems (LDS), Hidden Markov Models (HMM) and Independent Component Analysis (ICA) were unified as variations of unsupervised learning under a single basic generative model. A more recent work (Takeda et al., 2012), unifies algorithms such as support vector machine (SVM), minmax probability machine (MPM) and Fischer discriminant analysis (FDA) under a robust classification (uncertainty minimization) scheme.

In this paper we propose a framework which unifies several well-studied component analysis algorithms. In particular, we firstly formulate the joint (complete-data) probability density function (PDF) of a set of observations and latent variables. Subsequently, we show that the Maximum Likelihood (ML) solution of this joint PDF can produce the projection directions of PCA, LDA, LPP and SFA, by changing only the joint prior distribution of the latent variable, which in fact models the latent dependencies. For example, when using a fully connected Markov Random Field (MRF) for the latent prior distribution, we derive the PCA method. When choosing the product of a fully connected MRF and an MRF connected only to within-class data, we derive LDA. LPP is derived by choosing a locally connected MRF, while finally, SFA is produced when the joint prior is a linear Markov-chain. Afterwards, based on the aforementioned PDF we propose Expectation Maximization (EM) algorithms for learning the parameters of the model. Finally, with a set of both synthetic and real data, we demonstrate the usefulness and advantages of this family of probabilistic component analysis methods.
2. Prior Art and Novelties

An important contribution of our paper lies in the proposed unification of probabilistic component techniques. To the best of our knowledge, this is the first framework that reduces the construction of probabilistic component analysis models to the design of a proper prior over the latent space (where essentially the choice is reduced to selecting the latent neighbourhood).

In this section, we review the state-of-the-art in probabilistic alternatives of PCA, LDA, LPP and SFA. While doing so, we highlight the other novelties and advantages that our proposed framework entails wrt. each alternative formulation. Throughout this paper we consider, without any loss of generality, a zero mean set of $F$-dimensional observations of size $T$, \{$x_1, \ldots, x_T$\}, represented by a matrix $X = [x_1, \ldots, x_T]$. All of these methods discover an $N$-dimensional latent space $Y = [y_1, \ldots, y_T]$ which preserves certain properties of $X$.

2.1. Probabilistic PCA

The deterministic model of PCA finds a set of projection bases $W$, with the latent space $Y$ being the projection of the training set $X$ (i.e., $Y = W^TX$). The optimization problem is as follows

$$W_o = \arg \max_W \text{tr}[W^T SW], \text{ s.t. } W^T W = I$$

where $S = \frac{1}{T} \sum_{i=1}^{T} x_i x_i^T$ is the total scatter matrix and $I$ the identity matrix. The optimal $N$ projection basis $W_o$ are recovered (the $N$ eigenvectors of $S$ that correspond to the $N$ largest eigenvalues).

Probabilistic PCA (PPCA) approaches were independently proposed in (Roweis, 1998) and (Tipping & Bishop, 1999). In (Tipping & Bishop, 1999) a probabilistic generative model was adopted as:

$$x_i = Wy_i + \varepsilon_i, \ y_i \sim N(0, I), \ \varepsilon_i \sim N(0, \sigma^2 I)$$

where $W$ is the matrix that relates the latent variable $y_i$, with the observed samples $x_i$ and $\varepsilon_i$ is the noise which is assumed to be an isotropic Gaussian model. The motivation is that, when $N < F$, the latent variables will offer a more parsimonious explanation of the dependencies between the observations. The ML and EM solutions for parameter and moments $E[y_i]$ and $E[y,y^T]$ can be found in (Bishop, 2006; Tipping & Bishop, 1999). Several variations have been proposed since, e.g. by incorporating sparseness and nonnegative constraints (Sigg & Buhmann, 2008) or utilising joint generative/regression frameworks (Yu et al., 2006).

2.2. Probabilistic LDA

Let us now further assume that our data $X$ is further separated into $K$ disjoint classes $C_1, \ldots, C_K$ having $T_i$ samples and $T = \sum_{i=1}^{K} |C_i|$. The Fisher’s Linear Discriminant Analysis (FLDA) finds a set of projection bases $W$ s.t. (Yan et al., 2007)

$$W_o = \arg \min_W \text{tr}[W^T S_w W], \text{ s.t. } W^T S W = I$$

where $S_w = \sum_{i=1}^{K} N(x_i - \mu_{C_i})(x_i - \mu_{C_i})^T$ and $\mu_{C_i}$ the mean of class $i$. The idea is to find a latent space $Y = W^TX$ so that the within-class variance is minimized in a whitened space. The solution is given by the eigenvectors of $S_w$ that correspond to the $N-K$ smallest eigenvectors of the whitened data (i.e. by removing the variance after applying PCA). \(^1\)

Several probabilistic latent variable models which exploit class information have been recently proposed (c.f., (Prince & Elder, 2007; Zhang & Yeung, 2009; Ioffe, 2006)). In (Prince & Elder, 2007; Zhang & Yeung, 2009) another two related attempts were made to formulate a PLDA. Considering $x_i$ to be the $i$-th sample of the $c$-th class, the generative model of (Prince & Elder, 2007) can be described as:

$$x_i = Fh_c + Gw_{ic} + \epsilon_{ic}, \ h_c, w_{ic} \sim N(0, I), \ \epsilon_{ic} \sim N(0, \Sigma)$$

where $h_c$ represents the class-specific weights and $w_{ic}$ the weights of each individual sample, with $G$ and $F$ denoting the corresponding loadings. Regarding (Zhang & Yeung, 2009), the probabilistic model is as follows:

$$x_i = F_c h_c + \epsilon_{ic}, \ h_c, F_c \sim N(0, I), \ \epsilon_{ic} \sim N(0, \Sigma)$$

We note that the two models become equivalent when choosing a common $F$ (Eq. 5) for all classes while also disregarding the matrix $G$. In this case, the ML solution is given by obtaining the largest eigenvectors of $S_w$. Hence, the solution is vastly different than the one obtained by deterministic LDA (which keeps the smallest ones, Eq. 3). When learning a different $F_c$ per class, the model of (Zhang & Yeung, 2009) reduces to applying PPCA per class.

To the best of our knowledge the only probabilistic model where the ML solution is closely related to that of deterministic LDA is (Ioffe, 2006). The probabilistic model is defined as follows: $x \in C_i$, $x \sim N(y, \Phi_w)$, $y \sim N(m, \Phi_b)$, $V^T \Phi_b V = \Psi$ and $V^T \Phi_w V = I$.

\(^1\)We adopt this formulation of LDA instead of the equivalent of maximizing the trace of the between-class scatter matrix (Belhumeur et al., 1997), since this facilitates our following discussion on Prob. LDA alternatives.
Now let us consider the case that the columns of \( y \) appropriate prior over the latent space \( Y \) is for LPPs. In the following (Sec. 3, 4), we show how by assuming again a linear mapping \( Y \) representation \( y \) for these values, defined as the variance of the first feature analysis (SFA) is given if \( w \) is a vector of ones). The objective function with the chosen weights \( w_{ij} \) results in a heavy penalty if the neighboring points \( x_i \) and \( x_j \) are mapped far apart. Therefore, its minimization ensures that if \( x_i \) and \( x_j \) are near, then the projected features \( y_i = W^T x_i \) and \( y_j = W^T x_j \) are near, as well. To the best of our knowledge no probabilistic models exist for LPPs. In the following (Sec. 3, 4), we show how a probabilistic version of LPPs arises by choosing an appropriate prior over the latent space \( y_i \).

2.3. Locality Preserving Projections

Locality Preserving Projections (LPP) is the linear alternative of Laplacian Eigenmaps (Niyogi, 2004). The aim is to obtain a set of projection bases \( W \) and a latent space \( Y = W^T X \) which preserves the neighborhoods of the original samples. First, let us define a set of weights that represent locality. Common choices for the weights are the heat kernel \( w_{ij} = e^{-\frac{|x_i - x_j|^2}{\sigma^2}} \) or a set of constant weights \( (w_{ij} = 1 \text{ if the } i-th \text{ and the } j-th \text{ vectors are adjacent and } w_{ij} = 0 \text{ otherwise}) \). LPP finds a set of projection basis matrix \( W \) by solving the following problem:

\[
W_o = \arg\min_W \sum_{i,j=1}^{T} \sum_{n=1}^{N} w_{ij} |w_{n}^T x_i - w_{n}^T x_j|^2
\]

\[
\text{s.t. } W^T X D X^T W = I
\]

where \( L = D - W \) and \( D = \text{diag}(W1) \) (where \( \text{diag}(a) \) is the diagonal matrix having as main diagonal vector \( a \) and \( 1 \) is a vector of ones). The objective function with the chosen weights \( w_{ij} \) results in a heavy penalty if the neighboring points \( x_i \) and \( x_j \) are mapped far apart. Therefore, its minimization ensures that if \( x_i \) and \( x_j \) are near, then the projected features \( y_i = W^T x_i \) and \( y_j = W^T x_j \) are near, as well. To the best of our knowledge no probabilistic models exist for LPPs. In the following (Sec. 3, 4), we show how a probabilistic version of LPPs arises by choosing an appropriate prior over the latent space \( y_i \).

2.4. Probabilistic Slow Feature Analysis

Now let us consider the case that the columns of \( x_i \) are samples of a time series of length \( T \). The aim of slow feature analysis (SFA) is to find a sequential observation vectors \( X = [x_1 \ldots x_T] \), to find an output signal representation \( Y = [y_1 \ldots y_T] \) for which the features change slowest over time (Wiskott & Sejnowski, 2002). By assuming again a linear mapping \( Y = W^T X \) for the output representation, SFA minimizes the slowness for these values, defined as the variance of the first derivative of \( Y \). Formally, \( W \) of SFA is computed as

\[
W_o = \arg\min_W \text{tr} \left[ W^T \dot{X} X W \right], \text{s.t. } W^T S W = I
\]

where \( \dot{X} \) is the first derivative matrix (usually computed as the first order difference i.e., \( \dot{x}_j = x_j - x_{j-1} \)). An ML solution of the SFA was recently proposed in (Turner & Sahani, 2007). The idea was to incorporate a Gaussian linear dynamical system prior over the latent space \( Y \). The proposed generative model is

\[
\begin{align*}
P(x_t | W, y_t, \sigma_x) & \sim N(W^{-1} y_t, \sigma_x^2 I) \\
P(y_t | y_{t-1}, \lambda_{1:N}, \sigma_{1:N}) & = \prod_{n=1}^{N} P(y_{n,t} | y_{n,t-1}, \lambda_n, \sigma_{n}^2) \\
P(y_{n,t} | y_{n,t-1}, \lambda_n, \sigma_{n}^2) & \sim N(\lambda_n y_{n,t-1}, \sigma_{n}^2) \\
P(y_{n,1} | \sigma_{n,1}^2) & \sim N(0, \sigma_{n,1}^2).
\end{align*}
\]

As we will show, SFA is indeed a special case of our general model.

Summarizing, in the following sections we formulate a unified, probabilistic framework which (a) incorporates PCA as a special case, (b) produces a probabilistic LDA which (1) does not make assumptions regarding the number of samples per class (as in (Ioffe, 2006)) and (2) has an ML solution for the loading matrix \( W \) with similar direction to the deterministic LDA (Eq. 3), (c) provides the first, to the best of our knowledge, probabilistic model that explains LPPs, (d) naturally incorporates the recently proposed ML framework of SFA (Turner & Sahani, 2007) as a special case while solving an EM optimisation algorithm for SFA, and (e) provides variance estimates for each dimension along with the observation variance. This also differentiates our model from existing probabilistic component analysis techniques (PPCA, PLDA, HLDA) by providing more robust estimates.

3. A Maximum Likelihood Framework for Unified Component Analysis

In this section we will present the general framework for PCA and show how each of the set of projections from PCA, LDA, LPP and SFA can be derived in a ML framework. Let us consider the generative model \( x_i = W^{-1} y_i + \epsilon_i \), \( \epsilon_i \sim N(0, \sigma_x^2 I) \). We will prove that by choosing one of the priors defined below as a latent prior and subsequently taking the ML solution wrt. parameters, we end up to the aforementioned family of probabilistic component models. The priors, parametrised by \( \beta = \{\sigma_{1:N}, \lambda_{1:N}\} \), are:

- An MRF with full connectivity - each latent node \( y_i \) is connected to all other latent nodes \( y_{j}, j \neq i \).

\[
P(Y | \beta) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \sum_{n=1}^{N} \frac{1}{\sigma_{n}} \sum_{i,j=1}^{T} \lambda_{n} (y_{n,i} - \lambda_{n} y_{n,j})^2 \right\}
\]

\[
= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \left( \text{tr} \left[ \Lambda^{(1)} Y Y^T \right] + \text{tr} \left[ \Lambda^{(2)} Y M Y^T \right] \right) \right\}
\]

where \( M = -\frac{1}{2} I N^T \), \( \Lambda^{(1)} = \left[ \delta_{mn} \frac{\lambda^2}{\sigma_{n}} + 1 \right], \Lambda^{(2)} = \left[ \frac{1}{2} \right] \).
A product of two MRFs. In the first, each latent node $y_i$ is connected only to other latent nodes in the same class ($y_j, j \in \mathcal{C}_i$). In the second, each latent node ($y_i$) is connected to all other latent nodes ($y_j, j \neq i$).

- A product of two MRFs. In the first, each latent node $y_i$ is connected to all other latent nodes that belong in $y_i$’s neighborhood i.e. $y_j, j \in \mathcal{N}_i$. In the second, we only have individual potentials per node.

$$P(Y|\beta) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right] \right\}$$

$$= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \left( \text{tr} \left[ \Lambda^{(1)} Y Y^T \right] + \text{tr} \left[ \Lambda^{(2)} Y Y^T \right] \right) \right\}$$

where $\Lambda^{(1)} = \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right]$ and $\Lambda^{(2)} = \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right]$.

- A product of two MRFs. In the first, each latent node $y_i$ is connected to all other latent nodes that belong in $y_i$’s neighborhood i.e. $y_j, j \in \mathcal{N}_i$. In the second, we only have individual potentials per node.

$$P(Y|\beta) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right] \right\}$$

$$= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \left( \text{tr} \left[ \Lambda^{(1)} Y Y^T \right] + \text{tr} \left[ \Lambda^{(2)} Y Y^T \right] \right) \right\}$$

where $\Lambda^{(1)} = \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right]$ and $\Lambda^{(2)} = \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right]$.

A linear dynamical system prior over the latent space.

$$P(Y|\beta) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} \sum_{n=1}^{N} \sum_{t=1}^{T} \lambda_x \mathbb{I}_{\mathcal{N}_i} \left[ \frac{1}{\sigma^2_x} (y_{n,i} - y_{n,j})^2 \right] \right\}$$

$$= \frac{1}{Z} \exp \left\{ -\frac{1}{2} \left( \text{tr} \left[ \Lambda^{(1)} Y Y^T \right] + \text{tr} \left[ \Lambda^{(2)} Y Y^T \right] \right) \right\}$$

where $K = PP^T$ and $P$ is a $T \times (T - 1)$ matrix with elements $p_{ii} = 1$ and $p_{(i+1,i)} = -1$ (the rest are zero). The approximation holds when $T \to \infty$. Again, $\Lambda^{(1)}$ and $\Lambda^{(2)}$ are defined as above.

In all cases the partition function $Z$ is defined as

$$Z = \int P(Y) dY.$$
By completing the integrals and taking the logs, we obtain the conditional log-likelihood:

\[
L(\Psi) = \log P(X|\theta) = -\log Z + T \log |W| - \frac{1}{2} \text{tr} \left[ A^{(1)} W X B^{(1)} X^T W^T + A^{(2)} W X B^{(2)} X^T W^T \right]
\]

where \( \log Z \) is a constant term independent of \( W \). By maximising for \( W (\frac{\partial L}{\partial W} = 0) \) we obtain

\[
TW^{-T} - \left( A^{(1)} W X B^{(1)} X^T + A^{(2)} W X B^{(2)} X^T \right) = 0
\]

\[
I = A^{(1)} W X B^{(1)} X^T + A^{(2)} W X B^{(2)} X^T W^T.
\]

It is easy to prove that since \( A^{(1)}, A^{(2)} \) are diagonal matrices, the \( W \) which satisfies (20) is the one that simultaneously diagonalises \( X M_1 X^T \) and \( X M_2 X^T \). By substituting the \( B \) matrices as defined for Eq. 14 for each prior, we now consider all cases separately:

- **PCA**: In case of \( P(Y) \) of Eq. 10 then Eq. 20 is reformulated as \( W^T XX^T W = [A^{(1)}]^{-1} \) hence \( W \) is given by (up to a scale ambiguity) the eigenvectors of total scatter matrix \( S \).

- **LDA**: In case of \( P(Y) \) of Eq. 11 then Eq. 20 is reformulated as \( A^{(1)} W^T X M X^T W + A^{(2)} W^T X X^T W = I \). \( W \) is given by the directions that simultaneously diagonalise \( S \).

- **LPP**: In case of \( P(Y) \) of Eq. 12 then Eq. 20 is reformulated as \( A^{(1)} W^T X L X^T W + A^{(2)} X X^T W = I \). \( W \) is given by the directions that simultaneously diagonalise \( X L X^T \) and \( X X^T \).

- **SFA**: In case of \( P(Y) \) of Eq. 13 then Eq. 20 is reformulated as \( A^{(1)} W^T X K X^T W + A^{(2)} X X^T W = I \). \( W \) is given by the directions that simultaneously diagonalise \( X K X^T \) and \( X X^T \).

The direction of \( W \) does not depend of \( \sigma_n^2 \) and \( \lambda_n \), which can be estimated by optimizing Eq. 19 with regards to these parameters. In this work we will provide update rules for \( \sigma_n \) and \( \lambda_n \) using an EM framework. As we can see, the ML loading \( W \) does not depend on the exact setting of \( \lambda_n \), so long as they are all different. If \( 0 < \lambda_n < 1, \forall n \), then larger values of \( \lambda_n \) correspond to more expressive (in case of PCA), more discriminant (in LDA), more local (in LPP) and slower latents (in case of SFA). This corresponds directly to the ordering of the solutions from PCA, LDA, LPP and SFA. To recover exact equivalence to LDA, LPP, SFA another limit is required that corrects the scales. There are several choices, but a natural one is to let \( \sigma_n^2 = 1 - \lambda_n^2 \). This choice in case of LDA and SFA fixes the prior covariance of the latent variables to be one (\( W^T XX = I \)) and it forces \( W^T XD WX = I \) in case of LPP. This choice of \( \sigma_n \) has been also discussed in (Turner & Sahani, 2007) for slow feature analysis.

4. A Unified Expectation Maximization for Component Analysis

In the following we propose a unified EM framework for component analysis. This framework can treat all priors with undirected links (such as Eq. 10, Eq. 11 and Eq. 12). The EM of the prior in Eq. 13 is different since it contains only directed links with no loops, and thus can be solved (without any approximations) similarly to the EM of a linear dynamical system (Bishop, 2006).

In order to perform EM with an MRF prior we adopt the simple and elegant mean field approximation theory (Qian & Titterington, 1991; Celeux et al., 2003; Zhang, 1992). Let us consider the priors we defined in Sec. 3. Without loss of generality, we now assume the model

\[
x_i = W y_i + \epsilon_i, \epsilon_i \sim N(0, \sigma_x).
\]

For clarity, the set of parameters associated with the prior (i.e. energy function) are denoted as \( \beta = \{\sigma_1:N, \lambda_1:N\} \), the parameters related to the observation model \( \theta = \{W, \sigma_x\} \) while the total set of parameters are denoted as \( \Psi = \{\theta, \beta\} \).

In agreement with (Celeux et al., 2003), we replace the marginal distribution \( P(Y|\beta) \) by the mean-field

\[
P(Y|\beta) \approx \prod_{i=1}^{T} P(y_i|m_i^{(R)}, \beta).
\]

Since different models have different connectivity, the mean-field influence on each latent point \( y_i \) now depends on this specific connectivity through \( m_i^{(R)} \), which is a function of \( E[y_j] \). After calculating the normalising integral for the priors Eq. 10-12 and given the mean-field, we obtain

\[
P(y_i|m_i^{(R)}, \beta) = \begin{cases} N(y_i|m_i^{(PCA)}, \Sigma^{(PCA)}) & \text{for } (10) \\ N(y_i|m_i^{(LDA)}, \Sigma^{(LDA)}) & \text{for } (11) \\ N(y_i|m_i^{(LPP)}, \Sigma^{(LPP)}) & \text{for } (12) \end{cases}
\]

where \( R = \{PCA, LDA, LPP\} \). The means \( m_i^{(R)} \) are
where $\hat{E}^{(\theta)}$ is the likelihood is separated for estimating the maximisation step. By following the approximations in (Celeux et al., 2003), the complete-data likelihood is approximated as:

$$P(Y, X|\Psi) \approx \prod_{i=1}^{T} P(x_i|y_i, \theta) P(y_i|m_i^{(R)}, \beta).$$

(29)

The likelihood is separated for estimating $\theta^{(R)} = \{W^{(R)}, \sigma_z^{(R)}\}$ and $\beta = \{\sigma_1^{(R)}, \lambda_1^{(R)}\}$ as follows:

$$\theta^{(R)} = \arg \max \left\{ \sum_{i=1}^{T} \int_{y_i} P(y_i|x_i, m_i^{(R)}, \Psi) \log P(x_i|y_i, \theta) dy_i \right\}.$$

(30)

$$\beta^{(R)} = \arg \max \left\{ \sum_{i=1}^{T} \int_{y_i} P(y_i|x_i, m_i^{(R)}, \Psi) \log P(y_i|m_i^{(R)}, \beta) dy_i \right\}.$$

(31)

Subsequently, we maximise the log-likelihoods with the parameters, recovering the update equations as detailed in the supplementary material. For $\beta$, by maximising Eq. 30, we obtain:

$$W^{(R)} = \left( \sum_{i=1}^{T} x_i E[R|y_i] \right) \left( \sum_{i=1}^{T} E[R|y_i, y_i^T] \right)^{-1}$$

(32)

$$\sigma_{x}^{(R),2} = \frac{1}{T} \sum_{i=1}^{T} \{ ||x_i||^2 - 2 E[R|y_i]x_i (W^{(R)})^T x_i \}$$

(33)

Similarly, by maximising Eq. 31 for $\beta$, we obtain:

$$\sigma_{n}^{(R),2} = \frac{\zeta^{(R)}}{T} \sum_{i=1}^{T} \{ ||y_i||^2 - 2 E[R][y_i] y_i^T (W^{(R)})^T x_i \}$$

(34)

where

$$\zeta^{(R)} = \begin{cases} 1 & \text{for } R = \text{PCA} \\ \lambda_n + (1 - \lambda_n)^2 & \text{for } R = \text{LDA}, R = \text{LPP} \end{cases}$$

(35)

For $\lambda_n$ we choose the updates as described in Sec. 3.

### 4.1. Comparison to other variants of PPCA

It is clear that for PCA, the updates for $\theta = \{W, \sigma_z^2\}$ as well as the distribution of the latent variable $y_i$ are the same with previous probabilistic approaches (Roweis, 1998; Tipping & Bishop, 1999). The only variation is the mean of $y_i$, which in our case is shifted by the mean field $(\hat{\Sigma}^{(PCA)})^{-1} m_i^{(PCA)}$. In addition, our method models per-dimension variance ($\sigma_n$).

### 4.2. Complexity

The EM algorithm for our models is an iterative procedure for recovering the latent space which preserves the characteristics enforced by the selected latent neighbourhood. Our analysis is similar to PPCA (Roweis, 1998; Tipping & Bishop, 1999). For $N \ll T, F$ the complexity at each iteration is bounded by $O(TNF)$, unlike deterministic models which is $O(T^3)$. This is due to the covariance appearing only in trace operations, and is of high value for our proposed EM based models, since e.g., for LPP there is only the deterministic equivalent, with $O(T^3)$ complexity.

### 5. Experiments

As a proof of concept, in this section, we demonstrate the application of our proposed probabilistic compo-
A Unified Framework for Probabilistic Component Analysis

Figure 1. Synthetic experiments with deterministic LLE, LDA and PCA compared to our proposed probabilistic methods. For the deterministic models, the projections are shown in the 2nd column. For our probabilistic equivalents, we show the $E[y]$ (3rd column) along with the projections (4th column). A neighbourhood of 12 was used in the case of LLE.

6. Conclusions

In this paper we introduced a novel, unifying probabilistic component analysis framework, which reduces the construction of probabilistic component analysis models to essentially selecting the proper latent neighbourhood via the design of the latent connectivity. Our framework can thus be used to introduce novel probabilistic component analysis techniques by formulating new latent priors as products of MRFs. In this work, we have shown specific priors which when used, generate probabilistic models corresponding to PCA, LPP, LDA and SFA. By means of theoretical analysis and experiments, we have demonstrated various advantages that our proposed methods pose against existing probabilistic & deterministic techniques, while to the best of our knowledge, we introduce the first probabilistic equivalent to LPP and discuss the first EM model for SFA.

References

Belhumeur, P.N., Hespanha, J.P., and Kriegman, D.J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

Bishop, Christopher M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

Celeux, G., Forbes, F., and Peyrard, N. Em procedures using mean field-like approximations for markov model-based image segmentation. *Pattern recognition*, 36(1):131–144, 2003.

He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H.J. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

Ioffe, Sergey. Probabilistic Linear Discriminant Analysis. In *Computer Vision ECCV 2006*, pp. 531–542, 2006.

Niyogi, X. Locality preserving projections. In *Advances in neural information processing systems 16: proceedings of the 2003 conference*, volume 16, pp. 153. The MIT Press, 2004.

Prince, S. J. D. and Elder, J. H. Probabilistic linear discriminant analysis for inferences about identity. In *ICCV* 2007.

Qian, W. and Titterington, DM. Estimation of parameters in hidden markov models. *Phil. Trans. of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):407–428, 1991.
Roweis, S. Em algorithms for pca and spca. *Advances in neural information processing systems*, pp. 626–632, 1998.

Roweis, Sam and Ghahramani, Zoubin. A unifying review of linear gaussian models. *Neural Comput.*, 11(2):305–345, February 1999. ISSN 0899-7667.

Sigg, C.D. and Buhmann, J.M. Expectation-maximization for sparse and non-negative pca. In *Proceedings of the 25th international conference on Machine learning*, pp. 960–967. ACM, 2008.

Takeda, Akiko, Mitsugi, Hiroyuki, and Kanamori, Taka-fumi. A unified robust classification model. *CoRR*, abs/1206.4599, 2012.

Tipping, Michael E. and Bishop, Chris M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.

Turner, R. and Sahani, M. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

Wiskott, L. and Sejnowski, T.J. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

Yan, S. et al. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

Yu, Shipeng, Yu, Kai, Tresp, Volker, Kriegel, Hans-Peter, and Wu, Mingrui. Supervised probabilistic principal component analysis. In *KDD*, pp. 464–473, 2006.

Zhang, J. The mean field theory in em procedures for markov random fields. *IEEE Transactions on Signal Processing*, 40(10):2570–2583, 1992.

Zhang, Yu and Yeung, Dit-Yan. Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In *Proc. of the ECML-PKDD ’09: Part II*, pp. 602–616, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04173-0.