# A Unified Framework for Probabilistic Component Analysis

Mihalis A. Nicolaou[1], Stefanos Zafeiriou[1], and Maja Pantic[1,2]

[1] Department of Computing, Imperial College London, UK
[2] EEMCS, University of Twente, Netherlands (NL)
{mihalis,s.zafeiriou,m.pantic}@imperial.ac.uk

**Abstract.** We present a unifying framework which reduces the construction of probabilistic component analysis techniques to a mere selection of the latent neighbourhood, thus providing an elegant and principled framework for creating novel component analysis models as well as constructing probabilistic equivalents of deterministic component analysis methods. 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), some of which have no probabilistic equivalents in literature thus far. We firstly define the Markov Random Fields (MRFs) which encapsulate the latent connectivity of the aforementioned component analysis techniques; subsequently, we show that the projection directions produced by all PCA, LDA, LPP and SFA are also produced by the Maximum Likelihood (ML) solution of a single joint probability density function, composed by selecting one of the defined MRF priors while utilising a simple observation model. Furthermore, we propose novel Expectation Maximization (EM) algorithms, exploiting the proposed joint PDF, while we generalize the proposed methodologies to arbitrary connectivities via parametrizable MRF products. Theoretical analysis and experiments on both simulated and real world data show the usefulness of the proposed framework, by deriving methods which well outperform state-of-the-art equivalents.

**Keywords:** Unifying Framework, Probabilistic Methods, Component Analysis, Dimensionality Reduction, Random Fields.

## 1 Introduction

Unification frameworks in machine learning provide valuable material towards the deeper understanding of various methodologies, while also they form a flexible basis upon which further extensions can be easily built. One of the first attempts to unify methodologies was made in [17]. In this seminal work, models such as Factor Analysis (FA), Principal Component Analysis (PCA), mixtures of Gaussian clusters, Linear Dynamic Systems, Hidden Markov Models and Independent Component Analysis were unified as variations of unsupervised learning under a single basic generative model.

T. Calders et al. (Eds.): ECML PKDD 2014, Part II, LNCS 8725, pp. 469–484, 2014.

Deterministic Component Analysis (CA) unification frameworks proposed in previous works, such as [1], [10], [4], [23] and [21], provide significant insights on how CA methods such as Principal Component Analysis, Linear Discriminant Analysis, Laplacian Eigenmaps and others can be jointly formulated as, e.g., least squares problems under mild conditions or general trace optimisation problems. Nevertheless, while several probabilistic equivalents of, e.g. PCA have been formulated (c.f., [22] [16]), to this date no unification framework has been proposed for *probabilistic* component analysis. Motivated by the latter, in this paper we propose the *first* unified framework for probabilistic component analysis. Based on Markov Random Fields (MRFs), our framework unifies *all* component analysis techniques whose corresponding deterministic problem is solved as a trace optimisation problem without domain constraints for the parameters, such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locality Preserving Projections (LPP) and Slow Feature Analysis (SFA). Our framework provides further insight on component analysis methods from a probabilistic perspective. This entails providing probabilistic explanations for the data at hand with explicit variance modelling, as well as reduced complexity compared to the deterministic equivalents. These features are especially useful in case of methods for which no probabilistic equivalent exists in literature so far, such as LPP. Furthermore, under our framework one can generate *novel* component analysis techniques by merely combining products of MRFs with arbitrary connectivity.

The rest of this paper is organised as follows. We initially introduce previous work on CA, highlighting the properties of the proposed framework (Sec. 2). Subsequently, we formulate the joint complete-data Probability Density Function (PDF) of observations and latent variables. We show that the Maximum Likelihood (ML) solution of this joint PDF is co-directional to the solutions obtained via deterministic PCA, LDA, LPP and SFA, by changing only the prior latent distribution (Sec. 3), which, as we show, models the latent dependencies and thus determines the resulting CA technique. E.g, when using a fully connected MRF, we obtain PCA. 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. Based on the aforementioned PDF we subsequently propose Expectation Maximization (EM) algorithms (Sec. 4). Finally in Sec. 5, utilising 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, giving rise to the first framework that reduces the construction of probabilistic component analysis models to the design of a appropriate prior, thus defining only the latent neighbourhood. Nevertheless, other novelties arise in methods generated via our framework. In this section, we review the state-of-the-art in deterministic and probabilistic PCA, LDA, LPP and SFA.

While doing so, we highlight novelties and advantages that our proposed framework entails wrt. each alternative formulation. Throughout this paper we consider, a zero mean set of $F$-dimensional observations of length $T$, represented by the matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$. All CA methods discover an $N$-dimensional latent space $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ which preserves certain properties of $\mathbf{X}$.

### 2.1 Principal Component Analysis (PCA)

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

$$\mathbf{W}_o = \arg\max_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{S}\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{1}$$

where $\mathbf{S} = \frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i\mathbf{x}_i^T$ is the total scatter matrix and $\mathbf{I}$ the identity matrix. The optimal $N$ projection basis $\mathbf{W}_o$ are recovered (the $N$ eigenvectors of $\mathbf{S}$ that correspond to the $N$ largest eigenvalues). Probabilistic PCA (PPCA) approaches were independently proposed in [16] and [22]. In [22] a probabilistic generative model was adopted as:

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \boldsymbol{\epsilon}_i,\ \mathbf{y}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{F\times N}$ is the matrix that relates the latent variable $\mathbf{y}_i$ with the observed samples $\mathbf{x}_i$ and $\boldsymbol{\epsilon}_i$ 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 arising in observations.

### 2.2 Linear Discriminant Analysis (LDA)

Let us now further assume that our data $\mathbf{X}$ is further separated into $K$ disjoint classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$ with $T = \sum_{c=1}^{K}|\mathcal{C}_c|$. The Fisher's Linear Discriminant Analysis (LDA) finds a set of projection bases $\mathbf{W}$ s.t. [26]

$$\mathbf{W}_o = \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{S}_w\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{S}\mathbf{W} = \mathbf{I} \tag{3}$$

where $\mathbf{S}_w = \sum_{c=1}^{K}\sum_{\mathbf{x}_i\in\mathcal{C}_c}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{C}_i})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{C}_i})^T$ and $\boldsymbol{\mu}_{\mathcal{C}_i}$ the mean of class $i$. The aim is to find the latent space $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ such that the within-class variance is minimized in a whitened space. The solution is given by the eigenvectors of $\mathbf{S}_w$ corresponding to the $N - K$ smallest eigenvalues of the whitened data. [1]

Several probabilistic latent variable models which exploit class information have been recently proposed (c.f., [14,29,8]). In [14,29] another two related attempts were made to formulate a PLDA. Considering $\mathbf{x}_i$ to be the $i$-th sample of the $c$-th class, the generative model of [14] can be described as:

$$\mathbf{x}_i = \mathbf{F}\mathbf{h}_c + \mathbf{G}\mathbf{w}_{ic} + \boldsymbol{\epsilon}_{ic},\ \mathbf{h}_c, \mathbf{w}_{ic} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_{ic} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4}$$

[1] We adopt this formulation of LDA instead of the equivalent of maximizing the trace of the between-class scatter matrix [2], since this facilitates our following discussion on Probabilistic LDA alternatives.

where $\mathbf{h}_c$ represents the class-specific weights and $\mathbf{w}_{ic}$ the weights of each individual sample, with $\mathbf{G}$ and $\mathbf{F}$ denoting the corresponding loadings. Regarding [29], the probabilistic model is as follows:

$$\mathbf{x}_i = \mathbf{F}_c\mathbf{h}_c + \boldsymbol{\epsilon}_{ic},\ \mathbf{h}_c, \mathbf{F}_{ic} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),\ \boldsymbol{\epsilon}_{ic} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{5}$$

We note that the two models become equivalent when choosing a common $\mathbf{F}$ (Eq. 5) for all classes while also disregarding the matrix $\mathbf{G}$. In this case, the ML solution is given by obtaining the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w$. Hence, the solution is vastly different than the one obtained by deterministic LDA (which keeps the smallest ones, Eq. 3), resembling more to the solution of problems which retain the maximum variance. In fact, when learning a different $\mathbf{F}_c$ per class, the model of [29] 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 [8]. The probabilistic model is defined as follows: $\mathbf{x} \in \mathcal{C}_i$, $\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\mathbf{y}, \boldsymbol{\Phi}_w)$, $\mathbf{y} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Phi}_b)$, $\mathbf{V}^T\boldsymbol{\Phi}_b\mathbf{V} = \boldsymbol{\Psi}$ and $\mathbf{V}^T\boldsymbol{\Phi}_w\mathbf{V} = \mathbf{I}$, $\mathbf{A} = \mathbf{V}^{-T}$, $\boldsymbol{\Phi}_w = \mathbf{A}\mathbf{A}^T$ $\boldsymbol{\Phi} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A}^T$, where the observations are generated as:

$$\mathbf{x}_i = \mathbf{A}\mathbf{u},\ \mathbf{u} \sim \mathcal{N}(\mathbf{V}, \mathbf{I}),\ \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}). \tag{6}$$

The drawback of [8] is the requirement for all classes to contain the same number of samples. As we show, we overcome this limitation in our formulation.

### 2.3 Locality Preserving Projections (LPP)

Locality Preserving Projections (LPP) is the linear alternative to Laplacian Eigenmaps [13]. The aim is to obtain a set of projections $\mathbf{W}$ and a latent space $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ which preserves the locality of the original samples. First, let us define a set of weights that represent locality. Common choices for the weights are the heat kernel $u_{ij} = e^{-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\gamma}}$ or a set of constant weights ($u_{ij} = 1$ if the $i$-th and the $j$-th vectors are adjacent and $u_{ij} = 0$ otherwise, while $u_{ij} = u_{ji}$). LPP finds a set of projection basis matrix $\mathbf{W}$ by solving the following problem:

$$\begin{aligned}\mathbf{W}_o &= \arg\min_{\mathbf{W}} \textstyle\sum_{i,j=1}^{T}\sum_{n=1}^{N} u_{ij}||\mathbf{w}_n^T\mathbf{x}_i - \mathbf{w}_n^T\mathbf{x}_j||^2 \\ &= \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}\right] \text{ s.t. } \mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{W} = \mathbf{I}\end{aligned} \tag{7}$$

where $\mathbf{U} = [u_{ij}]$, $\mathbf{L} = \mathbf{D} - \mathbf{U}$ and $\mathbf{D} = \operatorname{diag}(\mathbf{U}\mathbf{1})$ (where $\operatorname{diag}(\mathbf{a})$ is the diagonal matrix having as main diagonal vector $\mathbf{a}$ and $\mathbf{1}$ is a vector of ones). The objective function with the chosen weights $w_{ij}$ results in a heavy penalty if the neighbouring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Therefore, its minimization ensures that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are near, then the projected features $\mathbf{y}_i = \mathbf{W}^T\mathbf{x}_i$ and $\mathbf{y}_j = \mathbf{W}^T\mathbf{x}_i$ 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 $\mathbf{y}_i$.

### 2.4 Slow Feature Analysis

Now let us consider the case that the columns of $\mathbf{x}_i$ are samples of a time series of length $T$. The aim of Slow Feature Analysis (SFA) is, given $T$ sequential observation vectors $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_T]$, to find an output signal representation $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_T]$ for which the features change slowest over time [25], [9]. By assuming again a linear mapping $\mathbf{Y} = \mathbf{W}^T\mathbf{X}$ for the output representation, SFA minimizes the *slowness* for these values, defined as the variance of the first derivative of $\mathbf{Y}$. Formally, $\mathbf{W}$ of SFA is computed as

$$\mathbf{W}_o = \arg\min_{\mathbf{W}} \operatorname{tr}\left[\mathbf{W}^T\dot{\mathbf{X}}\dot{\mathbf{X}}\mathbf{W}\right], \text{ s.t. } \mathbf{W}^T\mathbf{S}\mathbf{W} = \mathbf{I}, \tag{8}$$

where $\dot{\mathbf{X}}$ is the first derivative matrix (usually computed as the first order difference i.e., $\dot{\mathbf{x}}_j = \mathbf{x}_j - \mathbf{x}_{j-1}$). An ML solution of SFA was recently proposed in [24], by incorporating a Gaussian linear dynamic system prior over the latent space $\mathbf{Y}$. The proposed generative model is

$$\begin{aligned} P(\mathbf{x}_t|\mathbf{W}, \mathbf{y}_t, \sigma_x) &= \mathcal{N}(\mathbf{W}^{-1}\mathbf{y}_t, \sigma_x^2\mathbf{I}) \\ P(\mathbf{y}_t|\mathbf{y}_{t-1}, \lambda_{1:N}, \sigma_{1:N}) &= \textstyle\prod_{n=1}^{N} P(y_{n,t}|y_{n,t-1}, \lambda_n, \sigma_n^2) \end{aligned} \tag{9}$$

with $P(y_{n,t}|y_{n,t-1}, \lambda_n, \sigma_n^2) = \mathcal{N}(\lambda_n y_{n,t-1}, \sigma_n^2)$ and $P(y_{n,1}|\sigma_{n,1}^2) = \mathcal{N}\left(0, \sigma_{n,1}^2\right)$. 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 for component analysis which: (1) incorporates PCA as a special case, (2) produces a Probabilistic LDA which (i) has an ML solution for the loading matrix $\mathbf{W}$ with similar direction to the deterministic LDA (Eq. 3) and (ii) does not make assumptions regarding the number of samples per class (as in [8]), (3) provides the first, to the best of our knowledge, probabilistic model that explains LPP, (4) naturally incorporates SFA as a special case, (5) provides variance estimates not only for observations but also per latent dimension (differentiating our approach from existing probabilistic CA (e.g., PPCA, PLDA), and (6) provides a straightforward framework for producing novel component analysis techniques.

## 3 A Unified ML Framework for Component Analysis

In this section, we will present the proposed Maximum Likelihood (ML) framework for probabilistic component analysis and show how PCA, LDA, LPP and SFA can be generated within this framework, also proving equivalence with known deterministic models. Firstly, to ease computation, we assume the generative model for the $i$-th observation, $\mathbf{x}_i$, is defined as

$$\mathbf{x}_i = \mathbf{W}^{-1}\mathbf{y}_i + \boldsymbol{\epsilon}_i, \;\; \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_x^2\mathbf{I}). \tag{10}$$

In order to fully define the likelihood we need to define a prior distribution on the latent variables $\mathbf{y}$. We will prove that by choosing one of the priors defined below and subsequently taking the ML solution wrt. parameters, we end up

generating the aforementioned family of probabilistic component models. The priors, parametrised by $\beta = \{\sigma_{1:N}, \lambda_{1:N}\}$, are as follows (see also Fig. 1).

• An MRF with full connectivity - each latent node $\mathbf{y}_i$ is connected to all other latent nodes $\mathbf{y}_j, j \neq i$.

$$\begin{aligned} P(\mathbf{Y}|\beta) &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{1}{T-1}\sum_{j=1,j\neq i}^{T}\tfrac{1}{\sigma_n^2}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \\ &\approx \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{1}{T}\sum_{j=1}^{T}\tfrac{1}{\sigma_n^2}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \\ &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{M}\mathbf{Y}^T\right]\right)\right\}, \end{aligned} \tag{11}$$

where $\mathbf{M} \triangleq -\frac{1}{T}\mathbf{1}\mathbf{1}^T$, $\mathbf{\Lambda}^{(1)} \triangleq \left[\delta_{mn}\frac{\lambda_n^2+1}{\sigma_n^2}\right], \mathbf{\Lambda}^{(2)} \triangleq \left[\delta_{mn}\frac{\lambda_n}{\sigma_n^2}\right]$.

• A product of two MRFs. In the first, each latent node $\mathbf{y}_i$ is connected only to other latent nodes in the same class ($\mathbf{y}_j, j \in \tilde{\mathcal{C}}_i$). In the second, each latent node ($\mathbf{y}_i$) is connected to all other latent nodes ($\mathbf{y}_j, j \neq i$).

$$\begin{aligned} P(\mathbf{Y}|\beta) &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{1}{|\tilde{\mathcal{C}}_i|}\sum_{j\in\tilde{\mathcal{C}}_i}\tfrac{\lambda_n}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right\} \\ &\quad \exp\left\{-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{1}{T-1}\sum_{j=1}^{T}\tfrac{(1-\lambda_n)^2}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right\} \\ &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{M}_c\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{M}_t\mathbf{Y}^T\right]\right)\right\}, \end{aligned} \tag{12}$$

where $\mathbf{M}_c \triangleq \mathbf{I} - \mathrm{diag}[\mathbf{C}_1, \ldots, \mathbf{C}_C]$, $\mathbf{C}_c \triangleq \frac{1}{|\mathcal{C}_c|}\mathbf{1}_c\mathbf{1}_c^T$, $\mathbf{M}_t \triangleq \mathbf{I} + \mathbf{M}$, $\mathbf{\Lambda}^{(1)} \triangleq \left[\delta_{mn}(\frac{\lambda_n}{\sigma_n^2})\right]$, $\mathbf{\Lambda}^{(2)} \triangleq \left[\delta_{mn}\frac{(1-\lambda_n)^2}{\sigma_n^2}\right]$, while $\tilde{\mathcal{C}}_i = \{j : \exists\, \mathcal{C}_l \text{ s.t. } \{\mathbf{x}_j, \mathbf{x}_i\} \in \mathcal{C}_l, i \neq j\}$.

• A product of two MRFs. In the first, each latent node $\mathbf{y}_i$ is connected to all other latent nodes that belong in $\mathbf{y}_i$'s neighbourhood (symmetrically defined as $\mathcal{N}_i^s = \mathcal{N}_j^s = \{i \in \mathcal{N}_j \cup j \in \mathcal{N}_i\}$). In the second, we only have individual potentials per node.

$$\begin{aligned} P(\mathbf{Y}|\beta) &= \tfrac{1}{Z}\exp\left(-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{1}{|\mathcal{N}_i^s|}\sum_{j\in\mathcal{N}_i^s}\tfrac{\lambda_n}{\sigma_n^2}(y_{n,i}-y_{n,j})^2\right) \\ &\quad \exp\left(-\tfrac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{T}\tfrac{(1-\lambda_n)^2}{\sigma_n^2}y_{n,i}^2\right) \\ &= \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\tilde{\mathbf{L}}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\tilde{\mathbf{D}}\mathbf{Y}^T\right]\right)\right\} \end{aligned} \tag{13}$$

where $\tilde{\mathbf{L}} = \mathbf{D}^{-1}\mathbf{L}$ and $\tilde{\mathbf{D}} = \mathbf{I}$ ($\mathbf{L}$ and $\mathbf{D}$ are defined in Sec. 2.3 referring to LPPs). $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are defined as above.

• A linear dynamical system prior over the latent space.

$$\begin{aligned} P(\mathbf{Y}|\beta) &= \tfrac{1}{Z}\exp\left\{-\sum_{n=1}^{N}\left(\tfrac{1}{2\sigma_{n,1}^2}y_{n,1}^2+\tfrac{1}{2\sigma_n^2}\sum_{t=2}^{T}[y_{n,t}-\lambda_n y_{n,t-1}]^2\right)\right\} \\ &\approx \tfrac{1}{Z}\exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{K_1}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{Y}^T\right]\right)\right\} \end{aligned} \tag{14}$$

where $\mathbf{K}_1 = \mathbf{P}_1\mathbf{P}_1^T$ and $\mathbf{P}_1$ 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, $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$ are defined as above.

In all cases the partition function $Z$ is defined as $Z = \int P(\mathbf{Y})d\mathbf{Y}$. The motivation behind choosing the above latent priors was given by the influential

analysis made in [7] where the connection between (deterministic) LPP, PCA and LDA was explored. A further piece of the puzzle was added by the recent work [24] where the linear dynamical system prior (Eq. 14) was used in order to provide a derivation of SFA in a ML framework. By formulating the appropriate priors for these models we unify these subspace methods in a single probabilistic framework of a linear generative model along with a prior of the form

$$P(\mathbf{Y}) \propto \exp\left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(2)}\mathbf{Y}^T\right]\right)\right\}. \tag{15}$$

The differentiation amongst these models lies in the neighbourhood over which the potentials are defined. In fact, the varying neighbouring system is translated into the matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ in the functional form of the potentials, essentially encapsulating the latent covariance connectivity. E.g., for Eq. 11, $\mathbf{B}^{(1)} = \mathbf{I}$ and $\mathbf{B}^{(2)} = \mathbf{M}$, for Eq. 12, $\mathbf{B}^{(1)} = \mathbf{M}_c$ and $\mathbf{B}^{(2)} = \mathbf{M}_t$, for Eq. 13, $\mathbf{B}^{(1)} = \tilde{\mathbf{L}}$ and $\mathbf{B}^{(2)} = \tilde{\mathbf{D}}$ and finally for Eq. 14, $\mathbf{B}^{(1)} = \mathbf{K}$ and $\mathbf{B}^{(2)} = \mathbf{I}$. In the following we will show that ML estimation using these potentials is equivalent to the deterministic formulations of PCA, LDA and LPP. SFA is a special case for which it was already shown in [24] that a potential of the form of Eq. 14 with an ML framework produces a projection with the same direction as Eq. 8.

Adopting the linear generative model in Eq. 10, the corresponding conditional data (observation) probability is a Gaussian,

$$P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma_x^2) = \mathcal{N}(\mathbf{W}^{-1}\mathbf{y}_t, \sigma_x^2). \tag{16}$$

Having chosen a prior of the form described in Eq. 15 (e.g., as defined in Eq. 11,12,13,14) we can now derive the likelihood of our model as follows:

$$P(\mathbf{X}|\Psi) = \int \prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2) P(\mathbf{Y}|\sigma_{1:N}^2, \lambda_{1:N}) d\mathbf{Y}, \tag{17}$$

where the model parameters are defined as $\Psi = \{\sigma_x^2, \mathbf{W}, \sigma_{1:N}^2, \lambda_{1:N}\}$. In the following we will show that by substituting the above priors in Eq. 17 and maximising the likelihood we obtain loadings $\mathbf{W}$ which are co-directional (up to a scale ambiguity) to deterministic PCA, LDA and LPPs and SFA. Firstly, by substituting the general prior (Eq. 15) in the likelihood (Eq. 17), we obtain

$$\begin{aligned} P(\mathbf{X}|\Psi) = \int \textstyle\prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{y}_t, \mathbf{W}, \sigma^2) \tfrac{1}{Z} \exp \\ \left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}\mathbf{Y}^T\right]\right)\right\} d\mathbf{Y}. \end{aligned} \tag{18}$$

In order to obtain a zero-variance limit ML solution, we map $\sigma_x \to 0$

$$\begin{aligned} P(\mathbf{X}|\Psi) = \int \textstyle\prod_{t=1}^{T} \delta(\mathbf{x}_t - \mathbf{W}^{-1}\mathbf{y}_t) \tfrac{1}{Z} \exp \\ \left\{-\tfrac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{Y}\mathbf{B}^{(1)}\mathbf{Y}^T\right] + \mathrm{tr}\left[\mathbf{\Lambda}^{(2)}\mathbf{Y}\mathbf{B}^{(2)}Y^T\right]\right)\right\} d\mathbf{Y} \end{aligned} \tag{19}$$

By completing the integrals and taking the log, we obtain the conditional log-likelihood:

$$\begin{aligned} L(\Psi) = \log P(\mathbf{X}|\theta) = -\log Z + T \log|\mathbf{W}| - \tfrac{1}{2} \\ \mathrm{tr}\left[\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\mathbf{W}^T\right] \end{aligned} \tag{20}$$

where $Z$ is a constant term independent of $\mathbf{W}$. By maximising for $\mathbf{W}$ we obtain

$$\begin{aligned} T\mathbf{W}^{-T} - \left(\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\right) = \mathbf{0}, \\ \mathbf{I} = \mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T\mathbf{W}^T. \end{aligned} \tag{21}$$

It is easy to prove that since $\mathbf{\Lambda}^{(1)}, \mathbf{\Lambda}^{(2)}$ are diagonal matrices, the $\mathbf{W}$ which satisfies Eq. 21 simultaneously diagonalises (up to a scale ambiguity) $\mathbf{X}\mathbf{B}^{(1)}\mathbf{X}^T$ and $\mathbf{X}\mathbf{B}^{(2)}\mathbf{X}^T$. By substituting the $\mathbf{B}$ matrices as defined above in Eq. 21, we now consider all cases separately. For PCA, by utilising Eq. 11, Eq. 21 is reformulated as $\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \left[\mathbf{\Lambda}^{(1)}\right]^{-1}$ hence $\mathbf{W}$ is given by the eigenvectors of the total scatter matrix $\mathbf{S}$. For LDA (Eq. 12), Eq. 21 is reformulated as $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$. Thus, $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{S}$ and $\mathbf{S}_w$. For LPP (Eq. 13 ), Eq. 21 yields $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{D}^T\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$, therefore $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{X}\tilde{\mathbf{L}}\mathbf{X}^T$ and $\mathbf{X}\tilde{\mathbf{D}}\mathbf{X}^T$. Finally, for SFA, by utilising Eq. 14, Eq. 21 becomes $\mathbf{\Lambda}^{(1)}\mathbf{W}\mathbf{X}\mathbf{K}\mathbf{X}^T\mathbf{W}^T + \mathbf{\Lambda}^{(2)}\mathbf{W}\mathbf{X}\mathbf{X}^T\mathbf{W}^T = \mathbf{I}$, and $\mathbf{W}$ is given by the directions that simultaneously diagonalise $\mathbf{X}\mathbf{K}\mathbf{X}^T$ and $\mathbf{X}\mathbf{X}^T$.

The above shows that the ML solution following our framework is equivalent to the deterministic models of PCA, LDA, LPP and SFA. The direction of $\mathbf{W}$ does not depend of $\sigma_n^2$ and $\lambda_n$, which can be estimated by optimizing Eq. 20 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 observe, the ML loading $\mathbf{W}$ 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 ($\mathbf{W}^T\mathbf{X}\mathbf{X}\mathbf{W} = \mathbf{I}$) and it forces $\mathbf{W}^T\mathbf{X}\mathbf{D}\mathbf{X}\mathbf{W} = \mathbf{I}$ in case of LPP. This choice of $\sigma_n$ has been also discussed in [24] for SFA. We note that in case of PCA, we should set $\sigma_n$ to be analogous to the corresponding eigenvalue of the covariance matrix, since otherwise the method will result to a *minor* component analysis.

## 4 A Unified EM Framework 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. 11, Eq. 12 and Eq. 13). The EM of the prior in Eq. 14 contains only directed links with no loops, and thus can be solved (without any approximations) similarly to the EM of a linear dynamical system [3]. If we treat the SFA links as undirected, we end up with an autoregressive component analysis (see Section 4.1).

In order to perform EM with an MRF prior we adopt the simple and elegant mean field approximation theory [15,5,27], which essentially allows computationally favourable factorizations within an EM framework. Let us consider a generalisation of the priors we defined in Sec. 3 to $\mathcal{M}$ MRFs:

$$P(\mathbf{Y}|\beta) = \prod_{\mu\in\mathcal{M}} \frac{1}{Z^\mu} \exp\{Q^\mu\} \tag{22}$$

$$Q^\mu = -\sum_{n=1}^{N} \frac{f_\mu(\lambda_n)}{2\sigma_n^2} \frac{1}{c} \sum_{i\in\omega_i} \frac{1}{c_j^\mu} \sum_{j\in\omega_j^\mu} (y_{n,i} - \phi_\mu(\lambda_n) y_{n,j})^2$$

where $c$ and $c_j$ are normalisation constants, while $f_\mu$ and $\phi_\mu$ are functions of $\lambda_n$. Without loss of generality and for clarity of notation, we assume that $c = 1$, $c_j^\mu = |\omega_j^\mu|$ and $\omega_i^\mu = [1, \ldots, T]$. Furthermore, we now assume the linear model

$$\mathbf{x}_i = \mathbf{W}\mathbf{y}_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_x^2). \tag{23}$$

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 $\theta = \{\mathbf{W}, \sigma_x\}$, while the total parameter set is denoted as $\Psi = \{\theta, \beta\}$. In agreement with [5], we replace the marginal distribution $P(\mathbf{Y}|\beta)$ by the mean-field

$$P(\mathbf{Y}|\beta) \approx \prod_{i=1}^{T} P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}). \tag{24}$$

Since different CA models have different latent connectivities (and thus different MRF configurations), the mean-field influence on each latent point $\mathbf{y}_i$ now depends on the model-specific connectivity via $\mathbf{m}_i^{\mathcal{M}}$, a function of $\mathbb{E}[\mathbf{y}_j]$. After calculating the normalising integral for the priors Eq. 11-13 and given the mean-field, it can be easily shown that Eq. 22 follows a Gaussian distribution,

$$P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta) = \mathcal{N}(\mathbf{m}_i^{\mathcal{M}}, \mathbf{\Sigma}^{\mathcal{M}}), \tag{25}$$

$$\mathbf{m}_i^{\mathcal{M}} = \sum_{\mu\in\mathcal{M}} \left( \frac{f_\mu(\lambda_n)\phi_\mu(\lambda_n)}{F^M(\lambda_n)} \boldsymbol{\mu}_{\omega_j^\mu} \right) = \sum_{\mu\in\mathcal{M}} \mathbf{\Lambda}^\mu \boldsymbol{\mu}_{\omega_j^\mu}, \tag{26}$$

$$\mathbf{\Sigma}^{\mathcal{M}} = \left[ \delta_{mn} \frac{\sigma_n^2}{F^M(\lambda_n)} \right] \tag{27}$$

with $\boldsymbol{\mu}_{\omega_j^\mu} = \frac{1}{|\omega_j^\mu|} \sum_{j\in\omega_j^\mu} \mathbb{E}[\mathbf{y}_{n,j}]$ and $F^M(\lambda_n) = \sum_{\mu\in\mathcal{M}} f_\mu(\lambda_n)$.

Therefore, by simply replacing the parametrisation of the priors we defined in Eq. 11 (PCA), 12 (LDA) and 13 (LPP) (see also Tab. 1) for the mean and

**Table 1.** MRF configuration for PCA, LDA and LPP

| $\mathcal{M} = \{\alpha, \beta\}$ | $F^{\mathcal{M}} = \sum_\mu f_\mu$ | $f_a$ | $\phi_\alpha$ | $\omega_j^\alpha$ | $f_\beta$ | $\phi_\beta$ | $\omega_j^\beta$ |
|---|---|---|---|---|---|---|---|
| PCA (11) | 1 | 1 | $\lambda_n$ | $\{1 \ldots T\} \setminus \{i\}$ | | | |
| LDA (12) | $\lambda_n + (1-\lambda_n)^2$ | $\lambda_n$ | 1 | $\tilde{\mathcal{C}}_i$ | $(1-\lambda_n)^2$ | 1 | $\{1 \ldots T\} \setminus \{i\}$ |
| LPP (13) | $\lambda_n + (1-\lambda_n)^2$ | $\lambda_n$ | 1 | $\mathcal{N}_i^s$ | $(1-\lambda_n)^2$ | 0 | $\{1\}$ |

variance (Eq. 26 and Eq. 27), we obtain the distribution for each CA method we propose. The means $\mathbf{m}_i^{\mathcal{M}}$ for PCA, LDA and LPP are obtained as

$$\mathbf{m}_i^{(\text{PCA})} = \boldsymbol{\Lambda}\boldsymbol{\mu}_{-i}, \mathbf{m}_i^{(\text{LDA})} = \boldsymbol{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{-i} + \boldsymbol{\Lambda}^{(\beta)}\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i}, \mathbf{m}_i^{(\text{LPP})} = \boldsymbol{\Lambda}^{(\alpha)}\boldsymbol{\mu}_{\mathcal{N}_i^s} \tag{28}$$

and the variances $\boldsymbol{\Sigma}^{\mathcal{M}}$ as

$$\boldsymbol{\Sigma}^{(\text{PCA})} = \left[\delta_{mn}\sigma_n^2\right], \boldsymbol{\Sigma}^{(\text{LDA})} = \boldsymbol{\Sigma}^{(\text{LPP})} = \left[\delta_{mn}\left(\frac{\sigma_n^2}{\lambda_n+(1-\lambda_n)^2}\right)\right] \tag{29}$$

where $\boldsymbol{\mu}_{-i} = \frac{1}{T-1}\sum_{j\neq i}^{T}\mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ is the mean, $\boldsymbol{\mu}_{\tilde{\mathcal{C}}_i} = \frac{1}{|\tilde{\mathcal{C}}_i|}\sum_{j\in\tilde{\mathcal{C}}_i}\mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ the class mean, and $\boldsymbol{\mu}_{\mathcal{N}_i^s} = \frac{1}{|\mathcal{N}_i^s|}\sum_{j\in\mathcal{N}_i^s}^{T}\mathbb{E}^{\mathcal{M}}[\mathbf{y}_j]$ the neighbourhood mean. Furthermore, $\boldsymbol{\Lambda} = [\delta_{mn}\lambda_n]$, $\boldsymbol{\Lambda}^{(\alpha)} = \left[\delta_{mn}\left(\frac{\lambda_n}{\lambda_n+(1-\lambda_n)^2}\right)\right]$ and $\boldsymbol{\Lambda}^{(\beta)} = \left[\delta_{mn}\left(\frac{(1-\lambda_n)^2}{\lambda_n+(1-\lambda_n)^2}\right)\right]$.

In order to complete the expectation step, we infer the first order moments of the latent posterior, defined as

$$P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) = \frac{P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}})P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}})}{\int_{\mathbf{y}_i}P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}})P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}})d\mathbf{y}_i}. \tag{30}$$

Since the posterior is a product of Gaussians[2] , we have

$$P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}}) = \mathcal{N}(\mathbf{y}_i|(\mathbf{W}^T\mathbf{x}_i + \boldsymbol{\Sigma}^{\mathcal{M}^{-1}}\mathbf{m}_i^{\mathcal{M}})\mathbf{A}, \sigma_x^{\mathcal{M}^2}\mathbf{A}) \tag{31}$$

with $\mathbf{A} = (\mathbf{W}^T\mathbf{W} + (\hat{\boldsymbol{\Sigma}}^{\mathcal{M}})^{-1})^{-1}$ and $\hat{\boldsymbol{\Sigma}}^{\mathcal{M}} = \left[\delta_{mn}(\Sigma_{mn}^{\mathcal{M}}/\sigma_x^{\mathcal{M}^2})\right]$. Therefore $\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]$ is equal to the mean, and $\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i\mathbf{y}_i^T] = \sigma_x^{\mathcal{M}^2}\mathbf{A} + \mathbb{E}[\mathbf{y}_i]\mathbb{E}[\mathbf{y}_i]^T$.

Having recovered the first order moments, we move on to the maximisation step. In order to maximize the marginal log-likelihood, $\log P(\mathbf{X}|\Psi^{\mathcal{M}})$, we adopt the usual EM bound [17], $\int_{\mathbf{Y}}P(\mathbf{Y}|\mathbf{X}, \Psi^{\mathcal{M}})\log P(\mathbf{X}, \mathbf{Y})d\mathbf{Y}$. By adopting the approximation proposed in [5], the complete-data likelihood is factorised as

$$P(\mathbf{Y}, \mathbf{X}|\Psi^{\mathcal{M}}) \approx \prod_{i=1}^{T}P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}})P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}}). \tag{32}$$

and therefore, the maximisation term (EM bound) becomes

$$\sum_{i=1}^{T}\int_{\mathbf{y}_i}P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}})\log P(\mathbf{x}_i, \mathbf{y}_i|\Psi^{\mathcal{M}})d\mathbf{y}_i. \tag{33}$$

As can be seen the likelihood can be separated due to the logarithm for estimating $\theta^{\mathcal{M}} = \{\mathbf{W}^{\mathcal{M}}, \sigma_x^{\mathcal{M}}\}$ and $\beta = \{\sigma_{1:N}^{\mathcal{M}}, \lambda_{1:N}^{\mathcal{M}}\}$ as follows:

$$\theta^{\mathcal{M}} = \arg\max\left\{\sum_{i=1}^{T}\int_{\mathbf{y}_i}P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}})\log P(\mathbf{x}_i|\mathbf{y}_i, \theta^{\mathcal{M}})d\mathbf{y}_i\right\}. \tag{34}$$

$$\beta^{\mathcal{M}} = \arg\max\left\{\sum_{i=1}^{T}\int_{\mathbf{y}_i}P(\mathbf{y}_i|\mathbf{x}_i, \mathbf{m}_i^{\mathcal{M}}, \Psi^{\mathcal{M}})\log P(\mathbf{y}_i|\mathbf{m}_i^{\mathcal{M}}, \beta^{\mathcal{M}})d\mathbf{y}_i\right\}. \tag{35}$$

[2] The result can be easily obtained by completing the square for $\mathbf{y}_i$.

Subsequently, we maximise the log-likelihoods wrt. the parameters, recovering the update equations. For $\theta$, by maximising Eq. 34, we obtain

$$\mathbf{W}^{\mathcal{M}} = \left(\sum_{i=1}^{T} \mathbf{x}_i \mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]^T\right) \left(\sum_{i=1}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_i \mathbf{y}_i^T]\right)^{-1} \tag{36}$$

$$\begin{aligned} \sigma_x^{\mathcal{M}^2} = \tfrac{1}{FT} \textstyle\sum_{i=1}^{T} \{||\mathbf{x}_i||^2 - 2\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i]^T (\mathbf{W}^{\mathcal{M}})^T \mathbf{x}_i \\ + \mathrm{Tr}[\mathbb{E}^{\mathcal{M}}[\mathbf{y}_i \mathbf{y}_i^T](\mathbf{W}^{\mathcal{M}})^T \mathbf{W}^{\mathcal{M}}]\}. \end{aligned} \tag{37}$$

Similarly, by maximising Eq. 35 for $\beta$, we obtain:

$$\sigma_n^{\mathcal{M}^2} = \frac{F^{\mathcal{M}}(\lambda_n)}{T} \sum_{i=1}^{T} (\mathbb{E}^{\mathcal{M}}[y_{n,i}^2] - 2\mathbb{E}^{\mathcal{M}}[y_{n,i}] m_{n,i}^{\mathcal{M}} + m_{n,i}^{\mathcal{M}^2}) \tag{38}$$

where, as defined in Eq. 27, for PCA $F^{\mathcal{M}}(\lambda_n) = 1$, and for LDA and LPP $F^{\mathcal{M}}(\lambda_n) = \lambda_n + (1 - \lambda_n)^2$. For $\lambda_n$ we choose the updates as described in Sec. 3. In what follows, we discuss some further points wrt. the proposed EM framework.

### 4.1 Further Discussion

**Comparison to other Probabilistic Variants of PCA.** It is clear that regarding the proposed EM-PCA, the updates for $\theta = \{\mathbf{W}, \sigma_x^2\}$ as well as the distribution of the latent variable $\mathbf{y}_i$ are the same with previously proposed probabilistic approaches [16],[22]. The only variation is the mean of $\mathbf{y}_i$, which in our case is shifted by the mean field, $\hat{\mathbf{\Sigma}}^{(\mathrm{PCA})^{-1}} \mathbf{m}_i^{(\mathrm{PCA})}$, while in addition, our method models per-dimension variance ($\sigma_n$). Note that in order to fully identify with the PPCA proposed in [22], we can set $\lambda_n = 0$ and $\sigma_n = 1$.

**EM for SFA.** The SFA prior in Eq. 14 allows for two interpretations of the SFA graphical model: both as an undirected MRF and a directed Dynamic Bayesian Network (DBN). Based on the undirected MRF interpretation, SFA trivially fits into the EM framework described in this section, leading to an auto-regressive SFA model [19], able to learn bi-directional latent dependencies. When considering the SFA prior as a directed Markov chain, one can resort to exact inference techniques applied on DBNs. In fact, the EM for SFA can be straightforwardly reduced to solving a standard Linear Dynamical System (Chap. 13 [3]), while also enforcing diagonal transition matrices and setting $\sigma_n^2 = 1 - \lambda_n^2$.

**Complexity.** The proposed EM algorithm iteratively recovers the latent space preserving the characteristics enforced by the selected latent neighbourhood. Similarly to PCCA [16,22], for $N \ll T, F$ the complexity of each iteration is bounded by $O(TNF)$, unlike deterministic models ($\mathcal{O}(T^3)$). This is due to the covariance appearing only in trace operations, and is of high value for our proposed models, especially in case where no other probabilistic equivalent exists.

**Probabilistic LDA Classification.** We can exploit the probabilistic nature of the proposed EM-LDA in order to probabilistically infer the most likely class

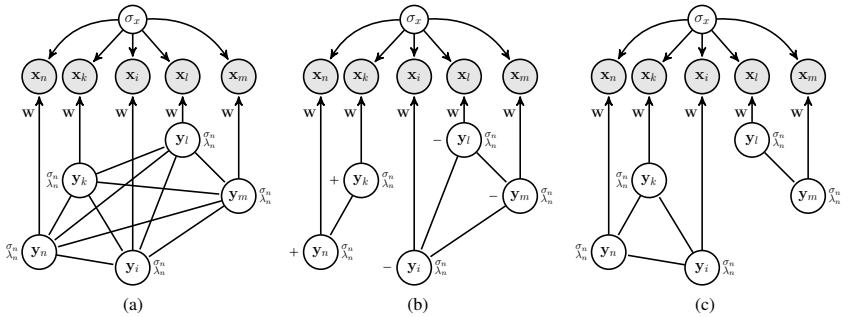


**Fig. 1.** MRF connectivities used for PCA, LDA and LPP under our unifying framework, with shaded nodes representing observations. (a) Fully connected MRF (PCA), (b) within-class connected MRF (LDA), and (c) a locally connected MRF (LPP).

assignment for unseen data. Instead of using the inferred projection, we can essentially utilise the log-likelihood of the model. In more detail, we can estimate the marginal log-likelihood for each test point $\mathbf{x}^*$ being assigned to each class $c$:

$$\arg_c \max \left\{ \log P(\mathbf{x}^* | \mathbf{m}^{\mathcal{M}_c}, \Psi^{\mathcal{M}}) \right\} \tag{39}$$

where by adopting the usual EM bound (as shown in Eq. 33) this boils down to

$$\arg_c \max \int_{\mathbf{y}_i^*} P(\mathbf{y}_i^* | \mathbf{x}_i^*, \mathbf{m}^{\mathcal{M}_c}, \Psi^{\mathcal{M}}) \log P(\mathbf{x}_i^*, \mathbf{y}_i^* | \Psi^{\mathcal{M}}) d\mathbf{y}_i^* \tag{40}$$

where $P(\mathbf{y}_i^* | \mathbf{x}_i^*, \mathbf{m}^{\mathcal{M}}, \Psi^{\mathcal{M}})$ is estimated as in Eq. 31, by utilising the inferred model parameters ($\Psi^{\mathcal{M}}$) along with the class model. Note that since the posterior mean given $\mathbf{x}_i$ depends on all *other* observations excluding $i$ (Eq. 28), we only need to store the class mean estimated as a weighted average of all training data and all training data in class $c$, as

$$\mathbf{m}^{\mathcal{M}_c} = \mathbf{\Lambda}^{(\alpha)} \frac{1}{T} \sum_{j=1}^{T} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j] + \mathbf{\Lambda}^{(\beta)} \frac{1}{|\mathcal{C}_c|} \sum_{j \in \mathcal{C}_c} \mathbb{E}^{\mathcal{M}}[\mathbf{y}_j] \tag{41}$$

This is in contrast to traditional methods where all the (projected) training data have to be kept. Furthermore, during evaluation, we only need to estimate the likelihood of each test datum's assignment to each class ($\mathcal{O}(|C|)$), rather than compare each test datum to the entire training set ($\mathcal{O}(T)$).

## 5 Experiments

As proof of concept, we provide experiments both on synthetic and real-world data. We aim to (i) experimentally validate the equivalence of the proposed probabilistic models to other models belonging in the same class, and (ii) experimentally evaluate the performance of our models against others in the same class.

**Synthetic Data.** We demonstrate the application of our proposed probabilistic CA techniques on a set of synthetic data (see Fig. 2), generated utilising the Dimensionality Reduction Toolbox. In more detail, we compare the corresponding deterministic formulations of PCA, LDA and LLE to the proposed probabilistic models. The aim is mainly to qualitatively illustrate the equivalence of the proposed methods (by observing that the probabilistic projections match the deterministic equivalents). Furthermore, the variance modelling per latent dimension in our EM-LDA is clear in $\mathbb{E}[\mathbf{y}]$ of LDA (Fig. 2, Col. 3). This will prove beneficial prediction-wise, as we show in the following section.

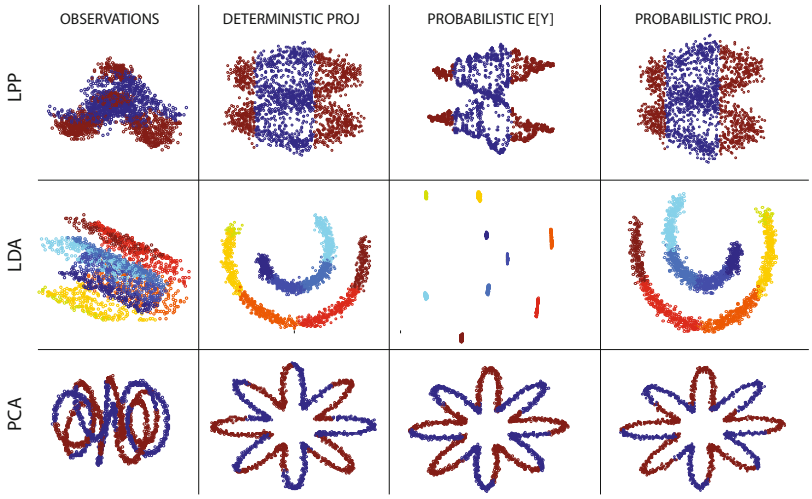


**Fig. 2.** Synthetic experiments on deterministic LLE, LDA and PCA (2nd col.) compared to the proposed probabilistic methods ($\mathbb{E}[y]$ in 3rd col., projections in 4th col.)

**Real Data: Face Recognition via EM-LDA.** One of the most common applications of LDA is face recognition. Therefore, we utilise various databases in order to verify the performance of our proposed EM-LDA. In more detail, we utilise the popular Extended Yale B database [6], as well as the PIE [20] and AR databases [12]. The experiments span a wide range of variability, such as various facial expressions, illumination changes, as well as pose changes. In more detail from the CMU PIE database [20] we used a total of 170 images near frontal images for each subject. For training, we randomly selected a subset consisting of 5 images per subject, while for testing the remaining images were used. For the extended Yale B database [6], we utilised a subset of 64 near frontal images per subject, where a random selection of 5 images per subject was used for training, while the rest of the images where used for testing. Regarding AR [12], we focus

on facial expressions. We firstly randomly select 100 subjects. Subsequently, use the images which portray varying facial expressions from session 1, while using the corresponding images from session 2 for testing. In related experiments, we compared our EM-LDA against deterministic LDA, the Fukunaga-Koontz variant (FK-LDA) [28] and PLDA [14] (which has been shown to outperform other probabilistic methods such as [8] in [11]) under the presence of Gaussian noise. We used the gradients of each image pixel as features, since as we experimentally verified, this improved the results for all compared methods. The errors of each compared method for each database, accompanied by increasing Gaussian noise in the input, is shown in Fig. 3. Although PLDA offers a substantial improvement wrt. deterministic LDA and performs better than FK-LDA, it is clear that the proposed EM-LDA outperforms other compared LDA variants. This can be attributed to the explicit variance modelling (both for observations and per dimension) in our models, which appears to enable more robust classification.

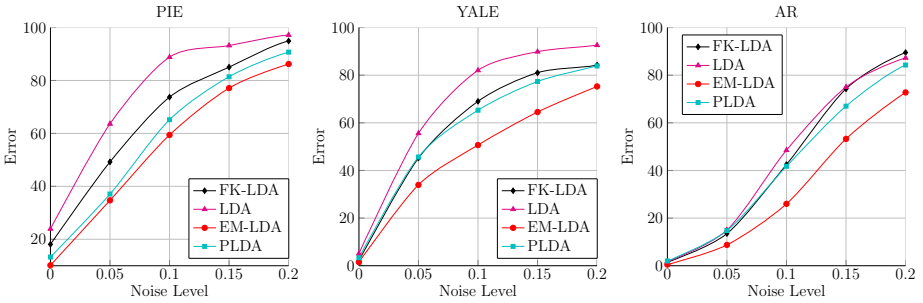


**Fig. 3.** Recognition error on PIE, YALE and AR under increasing Gaussian noise, comparing LDA, FK-LDA [28] the proposed EM-LDA and PLDA [14]

**Real Data: Face Visualisation via EM-LPP.** One of the typical applications of Neighbour Embedding methods is the visualisation of, usually high-dimensional, data at hand. In particular, LPPs have often been used in visualising faces, providing an intuitive understanding of the variance and structural properties of the data [16], [7]. In order to evaluate the proposed EM-LPP, which is to the best of our knowledge the first probabilistic equivalent to LPP [13], we experiment on the Frey Faces database [18], which contains 1965 images, captured as sequential frames of a video sequence. We apply a similar experiment to [7]. We firstly perturbed the images with random Gaussian noise, while subsequently we apply EM-LPP and LPP. The resulting space is illustrated in Fig. 4. It is clear that the deterministic LPP was unable to cope with the added Gaussian noise, failing to capture a meaningful data clustering. Note that the proposed EM-LPP was able to well capture the structure of the input data, modelling both pose and expression within the inferred latent space.

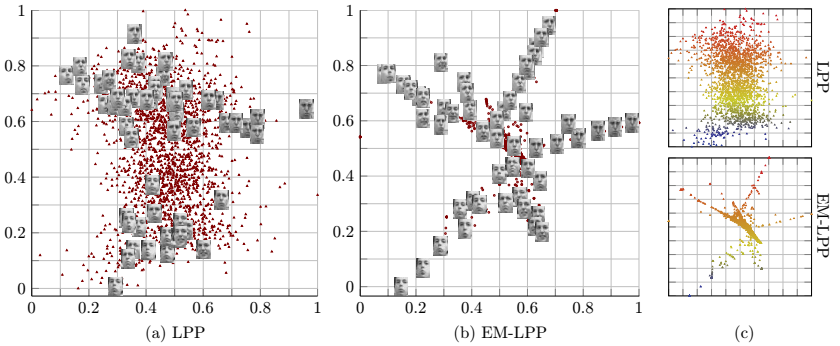


**Fig. 4.** Latent projections obtained by applying the proposed EM-LPP and LPP [13] to the Frey Faces database, with each image perturbed with random Gaussian noise

## 6 Conclusions

In this paper we introduced a novel, unifying probabilistic component analysis framework, reducing the construction of probabilistic component analysis models to 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. We have shown specific priors which when used, generate probabilistic models corresponding to PCA, LPP, LDA and SFA, and by doing so, we introduced the first, favourable complexity-wise, probabilistic equivalent to LPP. Finally, by means of theoretical analysis and experiments, we have demonstrated various advantages that our proposed methods pose against existing probabilistic and deterministic techniques.

**Acknowledgements.** This work has been funded by the European Union's 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG) and the EPSRC project EP/J017787/1 (4DFAB).

## References

1. Akisato, K., Masashi, S., Hitoshi, S., Hirokazu, K.: Designing various multivariate analysis at will via generalized pairwise expression. JIP 6(1), 136–145 (2013)
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE TPAMI 19(7), 711–720 (1997)
3. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus (2006)
4. Borga, M., Landelius, T., Knutsson, H.: A unified approach to PCA, PLS, MLR and CCA (1997)
5. Celeux, G., Forbes, F., Peyrard, N.: EM procedures using mean field-like approximations for Markov model-based image segmentation. Pattern Recogn. 36(1), 131–144 (2003)

6. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE TPAMI 23(6), 643–660 (2001)
7. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianfaces. IEEE TPAMI 27(3), 328–340 (2005)
8. Ioffe, S.: Probabilistic Linear Discriminant Analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 531–542. Springer, Heidelberg (2006)
9. Klampfl, S., Maass, W.: Replacing supervised classification learning by slow feature analysis in spiking neural networks. In: Advances in NIPS, pp. 988–996 (2009)
10. Kokiopoulou, E., Chen, J., Saad, Y.: Trace optimization and eigenproblems in dimension reduction methods. Numer. Linear Algebra Appl. 18(3), 565–602 (2011)
11. Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J.: Probabilistic models for inference about identity. IEEE TPAMI 34(1), 144–157 (2012)
12. Martinez, A.M.: The AR face database. CVC Technical Report 24 (1998)
13. Niyogi, X.: Locality preserving projections. In: NIPS 2003, vol. 16, p. 153 (2004)
14. Prince, S.J.D., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: ICCV (2007)
15. Qian, W., Titterington, D.: 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)
16. Roweis, S.: EM algorithms for PCA and SPCA. In: NIPS 1998, pp. 626–632 (1998)
17. Roweis, S., Ghahramani, Z.: A unifying review of linear gaussian models. Neural Comput. 11(2), 305–345 (1999)
18. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
19. Rue, H., Held, L.: Gaussian Markov random fields: Theory and applications. CRC Press (2004)
20. Sim, T., Baker, S., Bsat, M.: The CMU Pose, Illumination, and Expression Database. In: Proc. of the IEEE FG 2002 (2002)
21. Sun, L., Ji, S., Ye, J.: A least squares formulation for a class of generalized eigenvalue problems in machine learning. In: ICML 2009, pp. 977–984. ACM (2009)
22. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B 61, 611–622 (1999)
23. De la Torre, F.: A least-squares framework for component analysis. IEEE TPAMI 34(6), 1041–1055 (2012)
24. Turner, R., Sahani, M.: A maximum-likelihood interpretation for slow feature analysis. Neural Computation 19(4), 1022–1038 (2007)
25. Wiskott, L., Sejnowski, T.: Slow feature analysis: Unsupervised learning of invariances. Neural Computation 14(4), 715–770 (2002)
26. Yan, S., et al.: Graph embedding and extensions: A general framework for dimensionality reduction. IEEE TPAMI 29(1), 40–51 (2007)
27. Zhang, J.: The mean field theory in EM procedures for Markov random fields. IEEE Transactions on Signal Processing 40(10), 2570–2583 (1992)
28. Zhang, S., Sim, T.: Discriminant subspace analysis: A fukunaga-koontz approach. IEEE TPAMI 29(10), 1732–1745 (2007)
29. Zhang, Y., Yeung, D.-Y.: Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part II. LNCS, vol. 5782, pp. 602–616. Springer, Heidelberg (2009)