Initializing Probabilistic Linear Discriminant Analysis

Stylianos Moschoglou
Department of Computing
Imperial College London
London, UK
s.moschoglou@imperial.ac.uk

Mihalis Nicolaou
Department of Computing
Goldsmiths, University of London
London, UK
m.nicolaou@gold.ac.uk

Yannis Panagakis
Department of Computing
Middlesex University London
London, UK
y.panagakis@mdx.ac.uk

Stefanos Zafeiriou
Department of Computing
Imperial College London
London, UK
s.zafeiriou@imperial.ac.uk

***Abstract*—Component Analysis (CA) consists of a set of statistical techniques that decompose data to appropriate latent components that are relevant to the task-at-hand (e.g., clustering, segmentation, classification, alignment). During the past few years, an explosion of research in probabilistic CA has been witnessed, with the introduction of several novel methods (e.g., Probabilistic Principal Component Analysis, Probabilistic Linear Discriminant Analysis (PLDA), Probabilistic Canonical Correlation Analysis). PLDA constitutes one of the most widely used supervised CA techniques which is utilized in order to extract suitable, distinct subspaces by exploiting the knowledge of data annotated in terms of different labels. Nevertheless, an inherent difficulty in PLDA variants is the proper initialization of the parameters in order to avoid ending up in poor local maxima. In this light, we propose a novel method to initialize the parameters in PLDA in a consistent and robust way. The performance of the algorithm is demonstrated via a set of experiments on the modified XM2VTS database, which is provided by the authors of the original PLDA model.**

I. Introduction

Component Analysis (CA) methods are typically utilized as dimensionality reduction or feature extraction techniques in areas such as machine learning, computer vision and signal processing. Some of the very first and most known CA methods are Principal Component Analysis (PCA) [10], Linear Discriminant Analysis (LDA) [19], [24] and Canonical Correlation Analysis (CCA) [8]. One common attribute among the aforementioned techniques is that they are deterministic. Probabilistic interpretations of CA methods, such as Probabilistic PCA [14], [18], [21], Probabilistic LDA (PLDA) [16], [22], [25], [23], [9], [17] and Probabilistic CCA [11], [2], [15], were introduced in the literature over the past two decades. Probabilistic CA methods possess several advantages over their deterministic counterparts. In particular, they may be utilized as general density models [21], extended to mixture models [20] and Bayesian methodologies [12], used to model variance and handle missing data [1].

PLDA is a supervised probabilistic technique that takes into account the label (e.g., identity of a person depicted in an image) with which each datum is annotated. In particular, in PLDA the following assumption holds: Each datum is generated by two distinct, latent subspaces. The first subspace models the class in which each datum belongs to and the second renders the uniqueness for the specific datum. In face identification tasks, for instance, the first subspace would model the different identities that may exist in the training set while the second one would render the uniqueness of an image of a particular subject.

Nevertheless, initialization of the parameters of PLDA are accomplished utilizing heuristic techniques. To overcome this, we propose a novel algorithm to initialize PLDA. Our method eliminates the chance of ending up in poor local maxima and hence initializes the parameters in a consistent and robust way. We demonstrate the performance of our method by a set of experiments on the modifed XM2VTS database, provided by the authors of the original PLDA model [17], [13].

II. Probabilistic Linear Discriminant Analysis

In this section, we briefly cover the PLDA model introduced in [17], [13]. As stated above, in PLDA data are derived by two distinct, latent subspaces. One that models the class in which the sample belongs to and one that renders the uniqueness for that sample. In more detail, for a training set that consists of a total of I classes with each class i having a total of J images, the j-th sample of the i-th class may be formulated as

$$\mathbf{x}_{i,j} = \boldsymbol{\mu} + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j} + \boldsymbol{\epsilon}_{i,j}, \tag{1}$$

where $\boldsymbol{\mu}$ denotes the global mean of the training set, $\mathbf{F}$ defines the subspace that renders the identities of the subjects included in the training set, with $\mathbf{h}_i$ being the latent variable rendering the identity for the particular image. Moreover, $\mathbf{G}$ is the subspace that captures the specific conditions under which the images were taken (e.g., lighting variations) and $\mathbf{w}_{i,j}$ is the latent variable that renders the specific conditions for the particular image. Finally, $\boldsymbol{\epsilon}_{i,j}$ models a residual noise term which is Gaussian with diagonal covariance $\boldsymbol{\Sigma}$. Assuming that the data is centered (i.e., they are zero-mean), the model in (1) can be formulated as

$$P\left(\mathbf{x}_{i,j}|\mathbf{h}_i, \mathbf{w}_{i,j}, \boldsymbol{\theta}\right) = \mathcal{N}_{\mathbf{x}}\left(\mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j}, \boldsymbol{\Sigma}\right), \tag{2}$$

$$P\left(\mathbf{h}_i\right) = \mathcal{N}_{\mathbf{h}}\left(\mathbf{0}, \mathbf{I}\right), \tag{3}$$

$$P\left(\mathbf{w}_{i,j}\right) = \mathcal{N}_{\mathbf{w}}\left(\mathbf{0}, \mathbf{I}\right), \tag{4}$$

where the set of parameters $\boldsymbol{\theta} = \{\mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$ is optimized during training process via the EM algorithm [5].

ISBN 978-0-9928626-7-1 © EURASIP 2017

III. Initializing PLDA

As aforementioned, $\mathbf{F}$ and $\mathbf{G}$ are the parameters in PLDA which model the two distinct subspaces. Parameter $\mathbf{F}$ is initialized utilizing PCA [17], [13]. Moreover, parameter $\mathbf{G}$ is initialized using a random number generator. This approach has several shortcomings.

- Firstly, PCA is an unsupervised component analysis method. Therefore, when applied, the labels with which the data are annotated are not taken into account. Nevertheless, subspace $\mathbf{F}$ models the different classes in which the samples belong to. By applying PCA to initialize subspace $\mathbf{F}$ essential information is lost, which, had been taken into consideration, would have improved the final inference results.
- Secondly, PCA may be utilized for only a limited number of dimensions (components). The maximum number of principal components that may be used for a matrix $\mathbf{A}_{M\times N}$ is $\min\{M, N\}$. In PLDA [17], [13], the maximum number of components that may be used for $\mathbf{F}$ is equal to the number of identities that exist in the training set. In general, there should not be any restriction on how many dimensions to choose for the subspaces.
- Thirdly, subspace $\mathbf{G}$ is randomly initialized [17], [13]. Therefore, when the same experiment is repeated several times with the same set of parameters, the final result undulates around a mean value and hence some variance is introduced.

In order to tackle the disadvantages discussed above, we conceived and developed a novel algorithm which is utilized for the initialization of the parameters of PLDA. In more detail, our method is a deterministic variant of PLDA and carries the following assumptions. Each datum is generated by three distinct parts:

- a part that renders the identity of the person depicted in an image,
- a part that renders the specific conditions under which each image was taken (e.g., lighting variations) and thus the one that models the uniqueness for each image,
- a part that represents some residual noise.

This can be formulated as

$$\mathbf{x}_{i,j} = \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{i,j} + \mathbf{r}_{i,j}, \tag{5}$$

where $\mathbf{x}_{i,j}$ is the centered (i.e., zero-mean) j-th datum corresponding to the i-th identity, $\mathbf{F}$ is the subspace that renders the different identities that exist in the training set, $\mathbf{h}_i$ is the vector that defines the identity for the particular datum, $\mathbf{G}$ is the subspace that defines the uniqueness for each datum, $\mathbf{w}_{i,j}$ is the vector that renders the uniqueness for the particular datum and $\mathbf{r}_{i,j}$ denotes some residual noise.

A. Problem formulation

Without any loss of generality, suppose that exist a total of J images per identity and a total of I identities. By stacking all images in a column-wise manner, the model in (5) can then be reformulated as

$$\mathbf{X} = \mathbf{F}\begin{bmatrix}\mathbf{h}_1\mathbf{1}^T & \dots & \mathbf{h}_I\mathbf{1}^T\end{bmatrix} + \mathbf{G}\mathbf{W} + \mathbf{R}, \tag{6}$$

where

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_{1,1} & \dots & \mathbf{x}_{1,J} & \dots & \mathbf{x}_{I,1} & \dots & \mathbf{x}_{I,J}\end{bmatrix}, \tag{7}$$

$$\mathbf{1}^T = \underbrace{\begin{bmatrix}1 & \dots & 1\end{bmatrix}}_{J \text{ times}}, \tag{8}$$

$$\mathbf{W} = \begin{bmatrix}\mathbf{w}_{1,1} & \dots & \mathbf{w}_{1,J} & \dots & \mathbf{w}_{I,1} & \dots & \mathbf{w}_{I,J}\end{bmatrix}, \tag{9}$$

$$\mathbf{R} = \begin{bmatrix}\mathbf{r}_{1,1} & \dots & \mathbf{r}_{1,J} & \dots & \mathbf{r}_{I,1} & \dots & \mathbf{r}_{I,J}\end{bmatrix}. \tag{10}$$

In order to render subspaces $\mathbf{F}$ and $\mathbf{G}$ as informative as possible, we should minimize error $\mathbf{R}$. To achieve that, we need to solve

$$\begin{aligned}\min_{\mathbf{R}} \frac{1}{2}\|\mathbf{R}\|_F^2 \\ \text{subject to} \quad \mathbf{X} = \mathbf{F}\begin{bmatrix}\mathbf{h}_1\mathbf{1}^T & \dots & \mathbf{h}_I\mathbf{1}^T\end{bmatrix} + \mathbf{G}\mathbf{W} + \mathbf{R}, \\ \mathbf{F}^T\mathbf{F} = \mathbf{I}, \\ \mathbf{G}^T\mathbf{G} = \mathbf{I}, \\ \mathbf{F}^T\mathbf{G} = \mathbf{0},\end{aligned} \tag{11}$$

where $\|X\|_F \doteq \sqrt{\mathrm{tr}\left(\mathbf{X}^T\mathbf{X}\right)}$ is the Frobenius norm and $\mathrm{tr}\left(\cdot\right)$ denotes the trace of a square matrix. We introduce the orthonormality constraints in the minimization function in order to ensure that the subspaces are not correlated and thus the solution is identifiable. To solve (11), we utilize the Alternating Direction Method of Multipliers (ADMM) algorithm [6], [4], where an augmented Lagrangian is minimized. Let us denote $\mathbf{H} \doteq \begin{bmatrix}\mathbf{h}_1\mathbf{1}^T & \dots & \mathbf{h}_I\mathbf{1}^T\end{bmatrix}$. The augmented Lagrangian can then be written as follows

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}\right) = \min_{\boldsymbol{\theta}} \Bigg\{ \frac{1}{2}\|\mathbf{R}\|_F^2 + \mathrm{tr}\left[\boldsymbol{\Lambda}^T\left(\mathbf{X} - \mathbf{F}\mathbf{H} - \mathbf{G}\mathbf{W} - \mathbf{R}\right)\right] \\ + \frac{\mu}{2}\|\mathbf{X} - \mathbf{F}\mathbf{H} - \mathbf{G}\mathbf{W} - \mathbf{R}\|_F^2 \Bigg\} \\ \text{subject to} \quad \mathbf{F}^T\mathbf{F} = \mathbf{I}, \mathbf{G}^T\mathbf{G} = \mathbf{I}, \mathbf{F}^T\mathbf{G} = \mathbf{0},\end{aligned} \tag{12}$$

where $\boldsymbol{\theta} = \{\mathbf{R}, \mathbf{F}, \mathbf{G}, \mathbf{W}, \mathbf{H}, \boldsymbol{\Lambda}, \mu\}$. By employing an alternating optimization scheme, the iteration of the ADMM is the following.

Update the primal variables:

For obtaining subspace $\mathbf{F}$, we need to solve

$$\begin{aligned}\mathbf{F}_{t+1} = \underset{\mathbf{F}_t}{\mathrm{argmin}} \left\| \mathbf{X} - \mathbf{R}_t + \frac{\boldsymbol{\Lambda}_t}{\mu_t} - \mathbf{F}_t\mathbf{H}_t - \mathbf{G}_t\mathbf{W}_t \right\|_F^2, \\ \text{subject to} \quad \mathbf{F}_t^T\mathbf{F}_t = \mathbf{I}, \mathbf{F}_t^T\mathbf{G}_t = \mathbf{0}.\end{aligned} \tag{13}$$

In order to solve (13), we rely on the $\mathcal{Q}$ operator and Lemma introduced next. The rank-r Singular Value Decomposition (SVD) is defined for any matrix $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{B}\boldsymbol{\Sigma}\mathbf{W}^T$. Moreover, based on the SVD of $\mathbf{Y}$, the Procrustes operator is defined as $\mathcal{Q}\left[\mathbf{Y}\right] \doteq \mathbf{B}\mathbf{W}^T$.

ISBN 978-0-9928626-7-1 © EURASIP 2017

Lemma 1: The constraint minimization problem

$$\begin{aligned}\mathbf{\Omega}^{*} &= \underset{\mathbf{\Omega}}{\operatorname{argmin}} \left\|\mathbf{N}-\mathbf{\Omega}\mathbf{M}-\mathbf{Q}\mathbf{S}\right\|_{F}^{2}, \\ &\text{subject to} \quad \mathbf{\Omega}^{T}\mathbf{\Omega}=\mathbf{I}, \mathbf{Q}^{T}\mathbf{\Omega}=\mathbf{0},\end{aligned} \tag{14}$$

has a closed form solution [3] of the form $\mathbf{\Omega} = \mathcal{Q}\left[\left(\mathbf{I}-\mathbf{Q}\mathbf{Q}^{T}\right)\mathbf{N}\mathbf{M}^{T}\right]$. Therefore, the solution for (13) is given by

$$\mathbf{F}_{t+1} = \mathcal{Q}\left[\left(\mathbf{I}-\mathbf{G}_{t}\mathbf{G}_{t}^{T}\right)\left(\mathbf{X}-\mathbf{R}_{t}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}\right)\mathbf{H}_{t}^{T}\right]. \tag{15}$$

For deriving subspace $\mathbf{G}$, we need to solve

$$\begin{aligned}\mathbf{G}_{t+1} &= \underset{\mathbf{G}_{t}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{R}_{t}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}-\mathbf{G}_{t}\mathbf{W}_{t}-\mathbf{F}_{t+1}\mathbf{H}_{t}\right\|_{F}^{2}, \\ &\text{subject to} \quad \mathbf{G}_{t}^{T}\mathbf{G}_{t}=\mathbf{I}, \mathbf{G}_{t}^{T}\mathbf{F}_{t+1}=\mathbf{I}.\end{aligned} \tag{16}$$

The solution is given by utilizing the operator $\mathcal{Q}$ and Lemma 1.

$$\mathbf{G}_{t+1} = \mathcal{Q}\left[\left(\mathbf{I}-\mathbf{F}_{t+1}\mathbf{F}_{t+1}^{T}\right)\left(\mathbf{X}-\mathbf{R}_{t}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}\right)\mathbf{W}_{t}^{T}\right]. \tag{17}$$

For obtaining error $\mathbf{R}$, we have

$$\mathbf{R}_{t+1} = \underset{\mathbf{R}_{t}}{\operatorname{argmin}} \mathcal{L}\left(\boldsymbol{\theta}\right) \Rightarrow$$

$$\mathbf{R}_{t+1} = \frac{\mu_{t}}{1+\mu_{t}}\left(\mathbf{X}-\mathbf{F}_{t+1}\mathbf{H}_{t}-\mathbf{G}_{t+1}\mathbf{W}_{t}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}\right). \tag{18}$$

For $\mathbf{W}$, the solution is given by solving

$$\mathbf{W}_{t+1} = \underset{\mathbf{W}}{\operatorname{argmin}} \left\|\mathbf{X}-\mathbf{F}_{t+1}\mathbf{H}_{t}-\mathbf{R}_{t+1}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}-\mathbf{G}_{t}\mathbf{W}_{t}\right\|_{F}^{2}, \tag{19}$$

which admits a closed form solution

$$\mathbf{W}_{t+1} = \mathbf{G}_{t+1}^{T}\left(\mathbf{X}-\mathbf{F}_{t+1}\mathbf{H}_{t}-\mathbf{R}_{t+1}+\frac{\mathbf{\Lambda}_{t}}{\mu_{t}}\right). \tag{20}$$

For $\mathbf{H} \doteq \begin{bmatrix}\mathbf{h}_{1}\mathbf{1}^{T} & \dots & \mathbf{h}_{I}\mathbf{1}^{T}\end{bmatrix}$, we need to solve for each individual $\mathbf{h}_{i}, \forall i \in \{1,\dots,I\}$. The solution for each $\mathbf{h}_{i}$ is given by minimizing

$$\begin{aligned}\mathbf{h}_{i,t+1} = \underset{\mathbf{h}_{i,t}}{\operatorname{argmin}} \Bigg\|&\mathbf{X}_{i}-\mathbf{G}_{t+1}\mathbf{W}_{i,t+1}-\mathbf{R}_{i,t+1}+\frac{\mathbf{\Lambda}_{i,t}}{\mu_{t}} \\ &-\mathbf{F}_{t+1}\mathbf{h}_{i,t}\mathbf{1}^{T}\Bigg\|_{F}^{2},\end{aligned} \tag{21}$$

where the subscript i denotes that we consider only the J columns corresponding each time to the i-th subject. The solution for (21) is

$$\mathbf{h}_{i,t+1} = \frac{1}{J}\mathbf{F}_{t+1}^{T}\left(\mathbf{X}_{i}-\mathbf{G}_{t+1}\mathbf{W}_{i,t+1}-\mathbf{R}_{i,t+1}+\frac{\mathbf{\Lambda}_{i,t}}{\mu_{t}}\right)\mathbf{1}. \tag{22}$$

Update the Lagrange multiplier:

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_{t}+\mu_{t}\left(\mathbf{X}-\mathbf{F}_{t+1}\mathbf{H}_{t+1}-\mathbf{G}_{t+1}\mathbf{W}_{t+1}-\mathbf{R}_{t+1}\right). \tag{23}$$

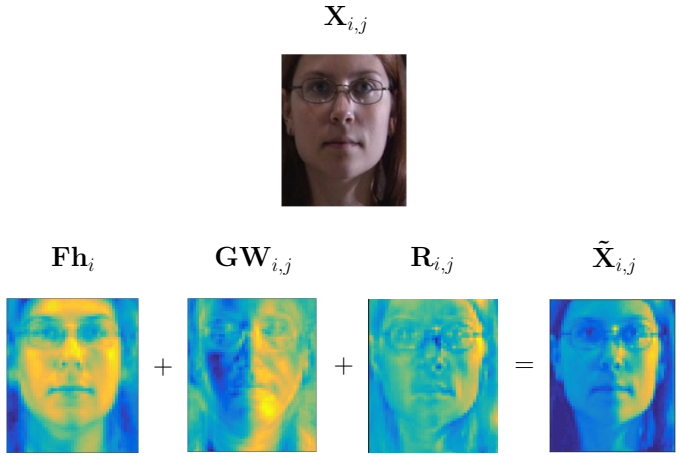

Fig. 1. Visualizing how the different components after the training procedure of ADMM reconstruct the original image ($\mathbf{X}_{i,j}$). Term $\mathbf{F}\mathbf{h}_{i}$ refers to the component that produces the specific identity, *regardless* of the different conditions under which the image was taken (e.g., lighting variations, etc.), term $\mathbf{G}\mathbf{W}_{i,j}$ refers to the component that renders the uniqueness for the specific image (i.e., captures the illumination setting, etc.), term $\mathbf{R}_{i,j}$ refers to the residual noise. Term $\tilde{\mathbf{X}}_{i,j}$ is the reconstructed image. The original image is taken from Multi-PIE [7].

The ADMM algorithm for solving (11) is outlined in Algorithm 1. In Fig. 1, a visualization of the solution derived by the ADMM solver (Algorithm 1) is provided. As shown, the obtained subspaces clearly model identity and illumination.

IV. Experiments

Having described PLDA along with a novel algorithm to initialize the parameters in PLDA, in this section we demonstrate the superiority of our method against the original PLDA model via a set of experiments on the modified XM2VTS database, provided by the authors of the original PLDA model [17], [13].

Modified XM2VTS database

Face identification: For this set of experiments, we utilize the modfied XM2VTS dataset provided by the authors of the original PLDA model [17], [13]. In more detail, we employ our algorithm to initialize the parameters in the PLDA model (dubbed Init. + PLDA) and compare the results against the original PLDA model. For Init. + PLDA, we apply Algorithm 1 on the training set to initialize subspaces $\mathbf{F}$ and $\mathbf{G}$. As mentioned in Section III, when the same experiment with the same set of parameters is repeated several times, variance is introduced in the inference results in PLDA. Therefore, for every distinct set of parameters we exectued 10 trials. Average identification rates per setting of parameters for Init. + PLDA compared to PLDA are presented in Figure 2. Average identification rates along with corresponding standard deviations per distinct set of parameters for PLDA and Init. + PLDA are shown in Tables I and II, respectively. Regarding the dimensionalities of the subspaces, for PLDA we used the same number of dimensions for subspaces $\mathbf{F}$ and $\mathbf{G}$, as stated in [13], due to the fact that the authors report that the maximum

ISBN 978-0-9928626-7-1 © EURASIP 2017

Algorithm 1: ADMM solver for (11)

Input: $\mathbf{X}, d_F, d_G$, where $\mathbf{X}$ is the set of the training data, d_F is the dimensionality of the $\mathbf{F}$ subspace and d_G is the dimensionality of the $\mathbf{G}$ subspace.

Initializations: $\mu_0 = \frac{1.25}{\|\mathbf{X}\|}, \{\mathbf{\Lambda}_0, \mathbf{W}_0, \mathbf{R}_0, \mathbf{H}_0\} = \mathbf{0}, \rho = 1.25, \mu_{\max} = 10^7, \{\mathbf{F}_0, \mathbf{G}_0\} =$ random initializations so that both $\mathbf{F}_0$ and $\mathbf{G}_0$ are orthogonal and $\mathbf{F}_0^T\mathbf{G}_0 = \mathbf{0}$.

Output: $\mathbf{F}_T, \mathbf{G}_T, \mathbf{W}_T, \mathbf{R}_T, \mathbf{H}_T$.

while *not converged or no termination criterion is met* **do**

  Update R:

$$\mathbf{R}_{t+1} = \frac{\mu_t}{1+\mu_t}\left(\mathbf{X} - \mathbf{F}_t\left[\mathbf{h}_{1,t}\mathbf{1}^T \quad \dots \quad \mathbf{h}_{I,t}\mathbf{1}^T\right] - \mathbf{G}_t\mathbf{W}_t + \frac{\mathbf{\Lambda}_t}{\mu_t}\right)$$

  Update F:

$$\mathbf{F}_{t+1} = \mathcal{Q}\left[\left(\mathbf{I} - \mathbf{G}_t\mathbf{G}_t^T\right)\left(\mathbf{X} - \mathbf{R}_{t+1} + \frac{\mathbf{\Lambda}_t}{\mu_t}\right)\left[\mathbf{h}_{1,t}\mathbf{1}^T \quad \dots \quad \mathbf{h}_{I,t}\mathbf{1}^T\right]^T\right]$$

  Update G:

$$\mathbf{G}_{t+1} = \mathcal{Q}\left[\left(\mathbf{I} - \mathbf{F}_{t+1}\mathbf{F}_{t+1}^T\right)\left(\mathbf{X} - \mathbf{R}_{t+1} + \frac{\mathbf{\Lambda}_t}{\mu_t}\right)\mathbf{W}_t^T\right]$$

  Update W:

$$\mathbf{W}_{t+1} = \mathbf{G}_{t+1}^T\left(\mathbf{X} - \mathbf{F}_{t+1}\left[\mathbf{h}_{1,t}\mathbf{1}^T \quad \dots \quad \mathbf{h}_{I,t}\mathbf{1}^T\right] - \mathbf{R}_{t+1} + \frac{\mathbf{\Lambda}_t}{\mu_t}\right)$$

  for $i = 1 : I$ **do**

    Update $\mathbf{h}_i$:

$$\mathbf{h}_{i,t+1} = \frac{1}{J}\mathbf{F}_{t+1}^T\left(\mathbf{X}_i - \mathbf{G}_{t+1}\mathbf{W}_{i,t+1} - \mathbf{R}_{i,t+1} + \frac{\mathbf{\Lambda}_{i,t}}{\mu_t}\right)\mathbf{1}$$

  end

  Update $\mathbf{\Lambda}$:

$$\mathbf{\Lambda}_{t+1} = \mathbf{\Lambda}_t + \mu_t\left(\mathbf{X} - \mathbf{F}_{t+1}\left[\mathbf{h}_{1,t+1}\mathbf{1}^T \quad \dots \quad \mathbf{h}_{I,t+1}\mathbf{1}^T\right] - \mathbf{G}_{t+1}\mathbf{W}_{t+1} - \mathbf{R}_{t+1}\right)$$

  Update μ:

$$\mu_{t+1} = \min\left(\rho\mu_t, \mu_{\max}\right)$$

end

TABLE I
PLDA AVERAGE IDENTIFICATION RATE ± STANDARD DEVIATION OVER 10 TRIALS PER DISTINCT SET OF PARAMETERS d_F, d_G, WHICH DENOTE THE DIMENSIONALITIES OF SUBSPACES $\mathbf{F}$ AND $\mathbf{G}$, RESPECTIVELY.

d_F	d_G	PLDA
16	16	0.69 ± 0.03
32	32	0.81 ± 0.02
48	48	0.81 ± 0.02
64	64	0.83 ± 0.03
80	80	0.86 ± 0.02
96	96	0.86 ± 0.02
112	112	0.85 ± 0.01
128	128	0.87 ± 0.02
144	144	0.87 ± 0.01
160	160	0.87 ± 0.01

TABLE II
INIT. + PLDA AVERAGE IDENTIFICATION RATE PER DISTINCT SET OF PARAMETERS d_F, d_G, WHICH DENOTE THE DIMENSIONALITIES OF SUBSPACES $\mathbf{F}$ AND $\mathbf{G}$, RESPECTIVELY. SINCE BOTH $\mathbf{F}$ AND $\mathbf{G}$ ARE FIXED AT THE INITIALIZATION FOR EVERY SET OF PARAMETERS, THE FINAL IDENTIFICATION RATE WILL BE FIXED AS WELL, THUS NO STANDARD DEVIATION WILL BE INTRODUCED.

d_F	d_G	Init. + PLDA
16	16	0.73
32	32	0.84
48	32	0.86
64	52	0.88
80	44	0.91
96	48	0.90
112	96	0.90
128	72	0.90
144	56	0.90
160	160	0.90

identification rates are attained when subspaces $\mathbf{F}$ and $\mathbf{G}$ have the same dimensionality. For Init. + PLDA we conducted experiments for varying dimensionalities for subspace $\mathbf{G}$ and concluded that the condition reported in [13] (i.e., maximum results are attained when both of the subspaces have the same dimensionality) is no longer valid. In Table II we provide the dimensionalities of subspace $\mathbf{G}$ for which the maximum identification rate is attained for various dimensionalities of subspace $\mathbf{F}$.

V. CONCLUSION

In this paper we introduced an algorithm to initialize the parameters of PLDA in a robust and consistent way. We demonstrate the superiority of our method against the original PLDA model via a series of experiments on the modified XM2VTS database, which is provided by the authors of the original PLDA model.

ISBN 978-0-9928626-7-1

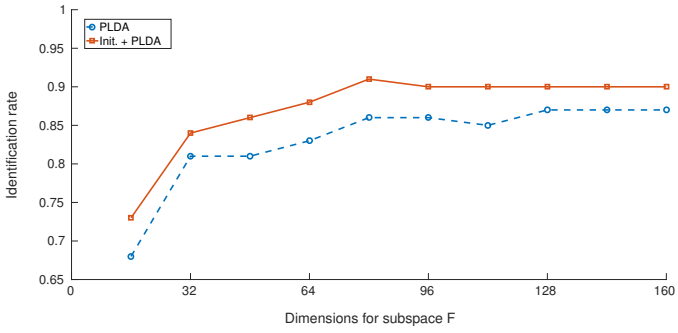

Fig. 2. Average identification rates of PLDA (dashed line) compared to the ones of Init. + PLDA (straight line) over a wide range of dimensions for subspace $\mathbf{F}$. When subspaces $\mathbf{F}$ and $\mathbf{G}$ are initialized utilizing Algorithm 1, final inference results are improved compared to the original PLDA model.

ACKNOWLEDGMENTS

Stylianos Moschoglou is funded by an EPSRC DTA studentship from Imperial College London. The work of Stefanos Zafeiriou was partially funded by the EPSRC Project EP/N007743/1 (FACER2VM).

REFERENCES

[1] C. Archambeau, N. Delannay, and M. Verleysen. Mixtures of robust probabilistic principal component analyzers. *Neurocomputing*, 71(7):1274–1282, 2008.

[2] F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[3] C. Bao, J.-F. Cai, and H. Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3384–3391, 2013.

[4] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[6] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers and Mathematics with Applications*, 2(1):17–40, 1976.

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[8] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[9] S. Ioffe. Probabilistic linear discriminant analysis. In *Proceedings of the European Conference in Computer Vision (ECCV)*, pages 531–542. Springer, 2006.

[10] I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2002.

[11] A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *The Journal of Machine Learning Research*, 14(1):965–1003, 2013.

[12] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

[13] P. Li, Y. Fu, U. Mohammed, J. H. Elder, and S. J. Prince. Probabilistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(1):144–157, 2012.

[14] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):696–710, 1997.

[15] M. A. Nicolaou, V. Pavlovic, and M. Pantic. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(7):1299–1311, 2014.

[16] M. A. Nicolaou, S. Zafeiriou, and M. Pantic. A unified framework for probabilistic component analysis. In *Machine Learning and Knowledge Discovery in Databases*, pages 469–484. Springer, 2014.

[17] S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[18] S. Roweis. Em algorithms for pca and spca. *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

[19] D. L. Swets and J. J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, (8):831–836, 1996.

[20] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[21] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[22] M. E. Wibowo, D. Tjondronegoro, L. Zhang, and I. Himawan. Heteroscedastic probabilistic linear discriminant analysis for manifold learning in video-based face recognition. In *Proceedings of the IEEE Conference Applications of Computer Vision (Workshop)*, pages 46–52. IEEE, 2013.

[23] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 464–473. ACM, 2006.

[24] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki. Regularized kernel discriminant analysis with a robust kernel for face recognition and verification. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):526–534, 2012.

[25] Y. Zhang and D.-Y. Yeung. Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In *Machine Learning and Knowledge Discovery in Databases*, pages 602–616. Springer, 2009.