Multi-Attribute Probabilistic Linear Discriminant Analysis for 3D Facial Shapes*

Stylianos Moschoglou[1], Stylianos Ploumpis[1], Mihalis A. Nicolaou[2], and Stefanos Zafeiriou[1]

[1] Imperial College London, London, UK
{s.moschoglou, s.ploumpis, s.zafeiriou}@imperial.ac.uk
[2] Computation-based Science and Technology Research Centre, The Cyprus Institute
m.nicolaou@cyi.ac.cy

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). During the past 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). A particular subset of CA methods such as PLDA, inspired by the classical Linear Discriminant Analysis, incorporate the knowledge of data labeled in terms of an attribute in order to extract a suitable discriminative subspace. Nevertheless, while many modern datasets incorporate labels with regards to multiple attributes (e.g., age, ethnicity, weight), existing CA methods can exploit at most a single attribute (i.e., one set of labels) per model. That is, in case multiple attributes are available, one needs to train a separate model per attribute, in effect not exploiting knowledge of other attributes for the task-at-hand. In this light, we propose the first, to the best of our knowledge, Multi-Attribute Probabilistic LDA (MAPLDA), that is able to jointly handle data annotated with multiple attributes. We demonstrate the performance of the proposed method on the analysis of 3D facial shapes, a task with increasing value due to the rising popularity of consumer-grade 3D sensors, on problems such as ethnicity, age, and weight identification, as well as 3D facial shape generation.

Keywords: Multi-Attribute · PLDA · Component Analysis · 3D shapes.

1 Introduction

Component Analysis (CA) techniques such as Principal Component Analysis (PCA) [10], Linear Discriminant Analysis (LDA) [23] and Canonical Correlation Analysis (CCA) [8] are among the most popular methods for feature extraction

* Supported by an EPSRC DTA studentship from Imperial College London, EPSRC Project EP/N007743/1 (FACER2VM) and a Google Faculty Award.

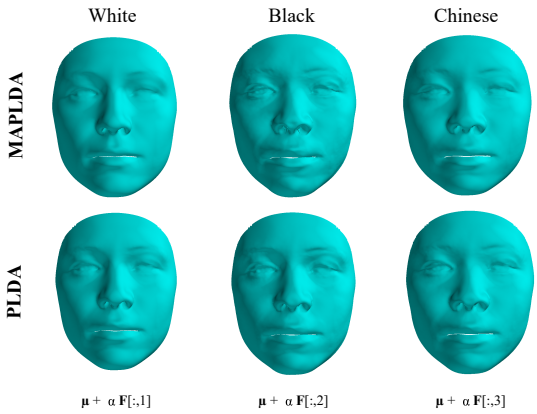

Fig. 1. Visualization of recovered components by MAPLDA as compared to PLDA, highlighting the improvement induced by explicitly accounting for multiple attributes. We denote with $\boldsymbol{\mu}$ the global mean, and with $\mathbf{F}$ the learned subspace of the *ethnicity* attribute where $\alpha \geq 1$ is used to accentuate the component visualization. MAPLDA is trained by jointly taking into account the *ethnicity* and *age-group* attributes. As can be clearly seen, this leads to a more accurate representation of the *ethnicity attribute* in MAPLDA, which is more prominent for the *Black* class.

and dimensionality reduction, typically utilized in a wide range of applications in computer vision and machine learning. While CA methods such as PCA have been introduced in the literature more than a century ago, it was only during the last two decades that *probabilistic* interpretations of CA techniques have been introduced in the literature, with examples of such efforts including Probabilistic PCA (PPCA) [18, 22, 25], Probabilistic LDA (PLDA) [19, 28, 30, 29, 9, 20] and Probabilistic CCA (PCCA) [12, 3]. The rise in popularity of probabilistic CA methods can be attributed to several appealing properties, such as explicit variance modeling and inherent handling of missing data [2]. Furthermore, probabilistic CA models may be easily extended to mixture models [24] and Bayesian methodologies [13], while they can also be utilized as general density models [25].

While many CA methods such as PCA and CCA are typically considered to be unsupervised, methods such as LDA assume knowledge of labeled data in order to derive a discriminative subspace based on attribute values (labels), that can subsequently be utilized for predictive analysis e.g., classification of unlabeled data. Probabilistic LDA (PLDA) [20, 14] constitutes one of the first attempts towards formulating a probabilistic generative CA model that incorporates information regarding data labels (e.g., the identity of a person in an image). In more detail, each datum is generated by two distinct subspaces: a subspace that incorporates information among instances belonging to the same class, and a subspace that models information that is unique to each datum. Put simply in the context of face recognition, all images of a specific subject share

the same identity, while each image may carry its own particular variations (e.g., in terms of illumination, pose and so on).

Nevertheless, a feature of PLDA and other probabilistic LDA variants that can be disadvantageous is the *single-attribute* assumption. In other words, PLDA is limited to the knowledge of one attribute, effectively disregarding knowledge of any other attributes available for the data-at-hand that may prove beneficial for a given task. For example, it is reasonable to assume that knowledge of attributes such as *pose*, *expression* and *age* may be deemed beneficial in terms of determining the identity of a person in a facial image. By incorporating knowledge of multiple attributes, we would expect a generative model to better explain the observation variance, by decomposing the observation space into multiple components conditioned on the attributes at-hand. Fig. 1 illustrates the more accurate representations we can obtain in this way.

In the past, PLDA was successfully applied to tasks such as face recognition and speaker verification [20, 11]. The advent of Deep Convolutional Neural Networks (DCNNs) provided models that overperformed linear CA techniques with respect to feature extraction in computer vision applications that involve intensity images and video, mainly due to the complex variations introduced by the texture and the geometric transformations. Nevertheless, linear CA techniques remain prominent and powerful techniques for tasks related to the analysis of 3D shapes, especially in case that dense correspondences have been established among them. Recently, very powerful frameworks have been proposed for establishing dense correspondences in large scale databases of 3D faces [16, 6], 3D bodies [15] and 3D hands [21].

Given that several modern databases of 3D shapes are annotated in terms of multiple attributes, and further motivated by the aforementioned shortcomings of single-attribute methods, in this paper we propose a Multi-Attribute generative probabilistic variant of LDA, dubbed Multi-Attribute Probabilistic LDA (MAPLDA). The proposed MAPLDA is able to *jointly* model the influence of multiple attributes on observed data, thus effectively decomposing the observation space into a set of subspaces depending on multiple attribute instantiations. As shown via a set of experiments on age, ethnicity and age group identification, the joint multi-attribute modeling embedded in MAPLDA appears highly beneficial, outperforming other single-attribute approaches in an elegant probabilistic framework. In what follows, we briefly summarize the contributions of our paper.

- We present MAPLDA, the first, to the best of our knowledge, probabilistic variant of LDA that is *inherently* able to *jointly* model multiple attributes.
- We provide a probabilistic formulation and optimization procedure for training, as well as a flexible framework for performing inference on *any* subset of the multiple attributes available during training.
- We demonstrate the advantages of joint-attribute modelling by a set of experiments on the MeIn3D dataset [6], in terms of ethnicity, age and weight group identification, as well as facial shape generation.

The rest of the paper is organized as follows. In Section 2, we briefly introduce PLDA, a generative counterpart to LDA. MAPLDA is introduced in Section 3, along with details on optimization and inference. Finally, experimental evaluation is detailed in Section 4.

2 Probabilistic Linear Discriminant Analysis

In this section, we briefly review the PLDA model introduced in [20, 14]. As aforementioned, PLDA carries the assumption that data are generated by two different subspaces: one that depends on the class and one that depends on the sample. That is, assuming we have a total of I classes, with each class i containing a total of J samples, then the j-th datum of the i-th class is defined 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 capturing the identity of every subject, with $\mathbf{h}_i$ being the latent identity variable representing the position in the particular subspace. Furthermore, $\mathbf{G}$ defines the subspace modeling variations among data, with $\mathbf{w}_{i,j}$ being the associated latent variable. Finally, $\boldsymbol{\epsilon}_{i,j}$ is a residual noise term which is Gaussian with diagonal covariance $\boldsymbol{\Sigma}$. Assuming zero-mean observations, the model in (1) can be described 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 via EM [7]. In the training process, EM is applied and the optimal set of parameters, $\boldsymbol{\theta} = \{\mathbf{F}, \mathbf{G}, \boldsymbol{\Sigma}\}$, is recovered.

3 Multi-Attribute PLDA (MAPLDA)

Let us consider a generalization of the single-attribute setting, as described in Section 2. In particular, let us assume that the data at-hand is labeled in terms of a total of N attributes, where each attribute may take K_i discrete instantiations (labels/classes), that is $a_i \in \{1, \cdots, K_i\}$[3]. We further assume that a set of J data available during training for any distinct combination of attribute instantiations. The generative model for MAPLDA corresponding to the j-th observation (datum) can then be described as:

$$\mathbf{x}_{a_{1:N},j} = \boldsymbol{\mu} + \sum_{i=1}^{N} \mathbf{F}_i \mathbf{h}_{i,a_i} + \mathbf{G}\mathbf{w}_{a_{1:N},j} + \boldsymbol{\epsilon}_{a_{1:N},j} \tag{5}$$

[3] For brevity of notation, we denote $a_1, \ldots, a_N$ as $a_{1:N}$.

where $\boldsymbol{\mu}$ denotes the training set global mean, $\mathbf{F}_1, \ldots, \mathbf{F}_N$ are loadings that define the subspace bases for each particular attribute (e.g., $\mathbf{F}_1$ may be the basis for the attribute age-group, $\mathbf{F}_2$ the basis for the attribute ethnicity, etc.) and $\mathbf{h}_{1,a_1}, \ldots, \mathbf{h}_{N,a_N}$ are selectors that define the position in each subspace, respectively (e.g., selector $\mathbf{h}_{1,a_1}$ will render the distinct age-group instantiation with which each datum is annotated). Furthermore, matrix $\mathbf{G}$ defines a basis for the subspace that models the variations among the data and $\mathbf{w}_{a_{1:N},j}$ defines the position in that subspace for the j-th datum. Finally, random noise is captured through the term $\boldsymbol{\epsilon}_{a_1:N,j}$ which is specific for each datum and is set as a Gaussian with diagonal covariance $\boldsymbol{\Sigma}$. Note that from here on, to avoid cluttering the notation we omit dependence on attribute instantiations (unless specified otherwise), that is we denote $\mathbf{x}_{a_{1:N},j}$ as $\mathbf{x}_j$, $\mathbf{w}_{a_{1:N},j}$ as $\mathbf{w}_j$ and $\boldsymbol{\epsilon}_{a_{1:N},j}$ as $\boldsymbol{\epsilon}_j$. Moreover, by assuming zero-mean observations, the model in (5) can be written more clearly as:

$$\mathbf{x}_j = \sum_{i=1}^{N} \mathbf{F}_i \mathbf{h}_{i,a_i} + \mathbf{G}\mathbf{w}_j + \boldsymbol{\epsilon}_j \tag{6}$$

while the prior probabilities of (6) can be written as:

$$P\left(\mathbf{h}_{i,a_i}\right) = \mathcal{N}_{\mathbf{h}}\left(\mathbf{0}, \mathbf{I}\right), \quad \forall i \in \{1, \ldots, N\} \tag{7}$$

$$P\left(\mathbf{w}_j\right) = \mathcal{N}_{\mathbf{w}}\left(\mathbf{0}, \mathbf{I}\right) \tag{8}$$

and the posterior as:

$$P\left(\mathbf{x}_j | \mathbf{h}_{1,a_1}, \ldots, \mathbf{h}_{N,a_N}, \mathbf{w}_j, \boldsymbol{\theta}\right) = \mathcal{N}_{\mathbf{x}}\left(\sum_{i=1}^{N} \mathbf{F}_i \mathbf{h}_{i,a_i} + \mathbf{G}\mathbf{w}_j, \boldsymbol{\Sigma}\right) \tag{9}$$

where $\boldsymbol{\theta} = \{\mathbf{F}_1, \ldots, \mathbf{F}_N, \mathbf{G}, \boldsymbol{\Sigma}\}$ is the set of parameters. Having defined our model, in the next subsections we detail both the training and inference procedures of MAPLDA in the presence of multiple attributes. For further clarification, we note that the graphical model of MAPLDA is illustrated in Fig. 2.

3.1 Training with Multiple Attributes

In this section, we detail the estimation of both the latent variables and parameters involved in MAPLDA. We assume that we are interested in making predictions regarding a subset of available attributes. While any subset can be chosen, for purposes of clarity and without loss of generality, we assume this set consists of the first $N-1$ attributes. That is, when given a test datum we can assign any of the $N-1$ attributes to classes $a_i, i \in \{1, \ldots, K_i\}$, while exploiting the knowledge of the remaining attributes (e.g., by marginalization during inference). Furthermore, without loss of generality, assume that there is a total of M data for each distinct combination of the $N-1$ attributes instantiations. We denote $\mathbf{F} \doteq \left[\mathbf{F}_1\ \mathbf{F}_2 \ldots \mathbf{F}_{N-1}\right]$, and $\mathbf{h} \doteq \left[\mathbf{h}_{1,a_1}^T\ \mathbf{h}_{2,a_2}^T \ldots \mathbf{h}_{N-1,a_{N-1}}^T\right]^T$ the block matrices consisting of loadings and variables for the first N-1 attributes, and

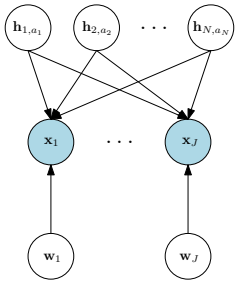

Fig. 2. Graphical model for J observed data of the training set (i.e., $\mathbf{x}_1, \ldots, \mathbf{x}_J$) for a distinct combination of attribute instantiations. The positions of the data in the subspaces $\mathbf{F}_1, \ldots, \mathbf{F}_N$ are given by the latent variables $\mathbf{h}_{1,a_1}, \ldots, \mathbf{h}_{N,a_N}$, respectively, while the position in subspace $\mathbf{G}$ is given by the latent variables $\mathbf{w}_1, \ldots, \mathbf{w}_J$, respectively.

$\hat{\mathbf{h}}_N \doteq \left[\mathbf{h}_{N,1}^T \, \mathbf{h}_{N,2}^T \ldots \mathbf{h}_{N,K_N}^T\right]^T$ the latent variable block matrix for all attribute values of the N-th attribute. Following a block matrix formulation, we group the M data samples as follows,

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_M \end{bmatrix} = \begin{bmatrix} \mathbf{F} & e_{1,a_N} \otimes \mathbf{F}_N & \mathbf{G} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{F} & e_{2,a_N} \otimes \mathbf{F}_N & \mathbf{0} & \mathbf{G} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \mathbf{F} & e_{M,a_N} \otimes \mathbf{F}_N & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \hat{\mathbf{h}}_N \\ \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix} \tag{10}$$

where $\otimes$ denotes the Kronecker product, and $e_{i,a_N} \in \mathbb{R}^{1 \times K_N}$ is a one-hot embedding of the value of attribute a_N for datum $\mathbf{x}_i$ (recall that $a_N \in \{1, \ldots, K_N\}$). For example, assume that for $\mathbf{x}_1$, $a_N = K_N$. Then, $e_{1,N} = [0, \ldots, 0, 1] \in \mathbb{R}^{1 \times K_N}$ and $e_{1,N} \otimes \mathbf{F}_N = [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{F}_N]$. Furthermore, (10) can be written compactly as:

$$\mathbf{x}' = \mathbf{A}\mathbf{y} + \epsilon' \tag{11}$$

where the prior and conditional probabilities of (11) can now be written as:

$$P\left(\mathbf{x}'|\mathbf{y}, \boldsymbol{\theta}\right) = \mathcal{N}_{\mathbf{x}'}\left(\mathbf{A}\mathbf{y}, \boldsymbol{\Sigma}'\right) \tag{12}$$

$$P\left(\mathbf{y}\right) = \mathcal{N}_{\mathbf{y}}\left(\mathbf{0}, \mathbf{I}\right) \tag{13}$$

where:

$$\boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma} \end{bmatrix} \tag{14}$$

Following EM and given an instantiation of the model parameters $\boldsymbol{\theta} = \{\mathbf{F}_1, \ldots, \mathbf{F}_N, \mathbf{G}, \boldsymbol{\Sigma}\}$, we need to estimate the sufficient statistics, that is the

first and second moments of the posterior latent distribution $P(\mathbf{y}|\mathbf{x}', \boldsymbol{\theta})$. Since both (12) and (13) refer to Gaussian distributions, it can easily be shown [5] that the posterior also follows a Gaussian distribution:

$$P(\mathbf{y}|\mathbf{x}', \boldsymbol{\theta}) = \mathcal{N}_{\mathbf{y}}\left(\hat{\mathbf{A}}\mathbf{A}^T \boldsymbol{\Sigma}'^{-1}\mathbf{x}', \hat{\mathbf{A}}\right) \tag{15}$$

where $\hat{\mathbf{A}} \doteq \left(\mathbf{A}^T \boldsymbol{\Sigma}'^{-1}\mathbf{A} + \mathbf{I}\right)^{-1}$, and thus:

$$\mathbb{E}[\mathbf{y}] = \hat{\mathbf{A}}\mathbf{A}^T \boldsymbol{\Sigma}'^{-1}\mathbf{x}' \tag{16}$$

$$\mathbb{E}\left[\mathbf{y}\mathbf{y}^T\right] = \hat{\mathbf{A}} + \mathbb{E}[\mathbf{y}]\,\mathbb{E}[\mathbf{y}]^T \tag{17}$$

Having derived the sufficient statistics of MAPLDA, we carry on to the maximization step. In order to recover the parameter updates, we take the partial derivatives of the conditional (on the posterior) expectation of the complete-data log likelihood of MAPLDA with regards to parameters $\boldsymbol{\theta} = \{\mathbf{F}_1, \ldots, \mathbf{F}_N, \mathbf{G}, \boldsymbol{\Sigma}\}$. In order to do so, we firstly rewrite (6) as follows:

$$\mathbf{x}_j = \begin{bmatrix}\mathbf{F}_1 \ldots \mathbf{F}_N\ \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{h}_{1,a_1} \\ \vdots \\ \mathbf{h}_{N,a_N} \\ \mathbf{w}_j\end{bmatrix} + \boldsymbol{\epsilon}_j \tag{18}$$

where (18) can be compactly written as:

$$\mathbf{x}_j = \mathbf{B}\mathbf{z}_j + \boldsymbol{\epsilon}_j. \tag{19}$$

By adopting the aforementioned grouping, our set of parameters is now denoted as $\boldsymbol{\theta} = \{\mathbf{B}, \boldsymbol{\Sigma}\}$, and the complete-data log likelihood conditioned on the posterior is formulated as:

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}\right) = \sum_{\mathbf{Z}} P\left(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}\right) \ln\left[P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})\right] \tag{20}$$

where the joint can be decomposed as:

$$P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{a_1=1}^{K_1} \cdots \prod_{a_N=1}^{K_N} \prod_{j=1}^{J} P\left(\mathbf{x}_{a_{1:N},j}|\mathbf{z}_{a_{1:N},j}\right) P\left(\mathbf{z}_{a_{1:N},j}\right) \tag{21}$$

It can be easily shown [5] that the updates are as follows:

$$\mathbf{B} = \left(\sum_{a_1=1}^{K_1} \cdots \sum_{a_N=1}^{K_N} \sum_{j=1}^{J} \mathbf{x}_{a_{1:N},j} \mathbb{E}\left[\mathbf{z}_{a_{1:N}}\right]^T\right) \left(\sum_{a_1=1}^{K_1} \cdots \sum_{a_N=1}^{K_N} \mathbb{E}\left[\mathbf{z}_{a_{1:N}} \mathbf{z}_{a_{1:N}}^T\right]\right)^{-1} \tag{22}$$

$$\boldsymbol{\Sigma} = \frac{1}{\mathcal{K}J}\mathrm{Diag}\left(\mathbf{S}_t - \mathbf{B} \sum_{a_1=1}^{K_1} \cdots \sum_{a_N=1}^{K_N} \sum_{j=1}^{J} \mathbb{E}\left[\mathbf{z}_{a_{1:N}}\right] \mathbf{x}_{a_{1:N},j}^T\right), \tag{23}$$

with $\mathbf{S}_t = \sum_{a_1=1}^{K_1} \cdots \sum_{a_N=1}^{K_N} \sum_{j=1}^{J} \mathbf{x}_{a_{1:N},j} \mathbf{x}_{a_{1:N},j}^T$ being the total covariance matrix and $\mathcal{K} = \prod_{i=1}^{N} K_i$.

3.2 Inference

Having completed the training process and derived the optimal MAPLDA parameters, we can proceed with inferences on unseen data on the first $N-1$ attributes. That is, given a datum (probe) from a test set, we aim to classify the datum into the appropriate classes for each of the corresponding $N-1$ attributes.

Since we do not have any prior knowledge of the conditions under which the data that belong to the test set may have been captured, it is very likely that the data may be perturbed by noise. Therefore, in order to determine the appropriate class, we compare the probe ($\mathbf{x}_p$) with a number of different data from a gallery in order to find the most likely match, in a similar manner to [20]. In essence, this boils down to maximum likelihood estimation under M (i.e., the total number of data in the gallery) different models. That is, for every model $m, m \in \{1, \ldots, M\}$, we calculate the log likelihood that the datum $\mathbf{x}_k$ in the gallery matches with the probe $\mathbf{x}_p$ and finally, we keep the pair that gives the largest log likelihood. This process falls under the so-called closed-set identification task, where a probe dataum has to be matched with a gallery datum. The algorithm can be extended to cover other scenarios such as verification or open-set identification.

Without loss of generality, let us assume a gallery with M data, all of which are labeled with different instantiations per attribute. Our aim is to find the pair that produces the maximum likelihood between the probe datum and one of the M gallery data. More formally, this corresponds to:

$$M_v \equiv \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \{\ln P(\mathcal{M}_m|\mathbf{X})\} \tag{24}$$

where $\mathbf{X} \doteq \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_M^T, \mathbf{x}_p^T\right]^T$. The optimal set of instantiations is described by the model M_v. If we consider a uniform prior for the selection of each model (i.e., $P(\mathcal{M}_m)$ is a constant for all $m \in \{1, \ldots, M\}$), then the actual log likelihood in (24) can calculated using Bayes' theorem as follows:

$$P(\mathcal{M}_m|\,\mathbf{X}) = \frac{P(\mathbf{X}|\mathcal{M}_m)\, P(\mathcal{M}_m)}{\sum_{m=1}^{M} P(\mathbf{X}|\mathcal{M}_m)\, P(\mathcal{M}_m)} \tag{25}$$

where the denominator is simply a normalizing constant, ensuring the probabilities sum to 1. Therefore, inference boils down to calculating:

$$\ln P(\mathbf{X}|\mathcal{M}_m) = \sum_{q=1, q\neq m}^{M} \ln P(\mathbf{x}_q) + \ln P(\mathbf{x}_p, \mathbf{x}_m) \tag{26}$$

where for each model m, the probe is paired with the m-th datum in the gallery and an individual marginal is added for the rest of the gallery data.

As aforementioned, and without loss of generality, we assume that inference is conducted for the first $N-1$ attributes. In order to perform inference without disregarding knowledge of attributes not required for inference, the sensible approach is to marginalize out the remaining N-th attribute. Then, following the process described above, we recover the optimal instantiations of attributes explained by model $\mathcal{M}_v$, utilizing (24), (25) and (26). The joint probabilities in (26) are Gaussians, and therefore, they can be estimated as:

$$P\left(\mathbf{x}_q\right) \sim \mathcal{N}_{\mathbf{x}_q}\left(\mathbf{0}, \mathbf{F}\mathbf{F}^T + \mathbf{F}_N\mathbf{F}_N^T + \mathbf{G}\mathbf{G}^T + \boldsymbol{\Sigma}\right) \tag{27}$$

where $\mathbf{F} \doteq \left[\mathbf{F}_1\, \mathbf{F}_2 \ldots \mathbf{F}_{N-1}\right]$. By assigning $\mathbf{x}' \doteq \left[\mathbf{x}_p^T, \mathbf{x}_m^T\right]^T$ and using the "completing-the-square" method, the marginals can be estimated as:

$$P\left(\mathbf{x}'\right) = \mathcal{N}_{\mathbf{x}'}\left(\mathbf{0}, \mathbf{A}\mathbf{A}^T + \boldsymbol{\Sigma}'\right) \tag{28}$$

where:

$$\mathbf{A} = \begin{bmatrix} \mathbf{F} & \mathbf{G} & \mathbf{0} \\ \mathbf{F} & \mathbf{0} & \mathbf{G} \end{bmatrix}, \quad \boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma} + \mathbf{F}_N\mathbf{F}_N^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} + \mathbf{F}_N\mathbf{F}_N^T \end{bmatrix} \tag{29}$$

A graphical representation for this case can be found in Fig. 3.

Regarding the special case where inference about *only one* attribute is required, the marginals have the same form as in (27). The joint distribution, given that the attribute of interest is denoted as $i \in \{1, \ldots, N\}$, follows the form:

$$P\left(\mathbf{x}'\right) \sim \mathcal{N}_{\mathbf{x}'}\left(\mathbf{0}, \mathbf{A}\mathbf{A}^T + \boldsymbol{\Sigma}'\right) \tag{30}$$

where in this case:

$$\mathbf{A} = \begin{bmatrix} \mathbf{F}_i & \mathbf{G} & \mathbf{0} \\ \mathbf{F}_i & \mathbf{0} & \mathbf{G} \end{bmatrix}, \quad \boldsymbol{\Sigma}' = \begin{bmatrix} \boldsymbol{\Sigma} + \sum\limits_{i=1, i\neq n}^{N} \mathbf{F}_i\mathbf{F}_i^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} + \sum\limits_{i=1, i\neq n}^{N} \mathbf{F}_i\mathbf{F}_i^T \end{bmatrix} \tag{31}$$

We finally note that MAPLDA is a generalization of PLDA; in the degenerate case where only one attribute is available during training, MAPLDA reduces to PLDA.

3.3 3D Facial Shape Generation

We can exploit the generative property of MAPLDA, alongside the multi-attribute aspect of the model, to generate data with respect to different combinations of attribute values. Data generation can be accomplished as follows:

- Firstly, without loss of generality, we train a MAPLDA model with regards to two attributes we are interested in (e.g., attributes ethnicity and age, weight and age, etc.). After the training process, we recover the optimal $\mathbf{F}_1$, $\mathbf{F}_2$, $\mathbf{G}$ subspaces and noise diagonal covariance $\boldsymbol{\Sigma}$.

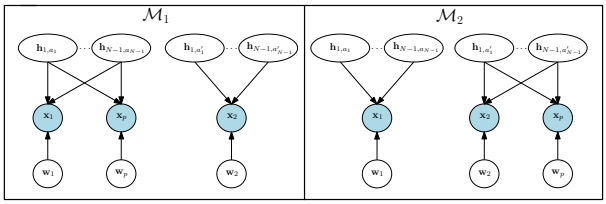

Fig. 3. Inference for *some* attributes (in this case, the first $N-1$ attributes). For this particular case, only two data exist in the gallery, so the probe datum $\mathbf{x}_p$ can be matched with either datum $\mathbf{x}_1$ or datum $\mathbf{x}_2$. In case it does match with datum $\mathbf{x}_1$, then it is assigned labels $\{a_1, \ldots, a_{N-1}\}$ (model $\mathcal{M}_1$). Otherwise, it receives labels $\{a'_1, \ldots, a'_{N-1}\}$ (model $\mathcal{M}_2$).

- Secondly, we pick the distinct instantiations of attributes we are interested in generating (e.g., *Chinese* ethnic group and *18-24* age group) and stack row-wise all the training data pertaining to these instantiations, creating a new vector $\mathbf{x}'$.
- Thirdly, if $\mathbf{h}_{i,a_i}$ and $\mathbf{h}_{j,a_j}$ are the selectors corresponding to the particular attributes, we stack them row-wise, i.e., $\mathbf{h}^T \doteq \begin{bmatrix}\mathbf{h}_{i,a_i} & \mathbf{h}_{j,a_j}\end{bmatrix}$, and calculate the posterior $\mathbb{E}\left[P\left(\mathbf{h}|\mathbf{x}'\right)\right]$ as

$$\mathbb{E}\left[P\left(\mathbf{h}|\mathbf{x}'\right)\right] = \mathbf{C}\mathbf{A}^T\mathbf{D}^{-1}\mathbf{x}', \tag{32}$$

  where $\mathbf{A} = \begin{bmatrix}\mathbf{F}_1 & \mathbf{F}_2\end{bmatrix}$, $\mathbf{C} = \left(\mathbf{I} + \mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}\right)^{-1}$ and $\mathbf{D} = \left(\boldsymbol{\Sigma}' + \mathbf{G}'\mathbf{G}'^T\right)^{-1}$, where $\boldsymbol{\Sigma}'$ is defined as in (14), and $\boldsymbol{G}'$ is a block-diagonal matrix with copies of $\boldsymbol{G}$ on the diagonal.
- Finally, for selector $\mathbf{w}$, we choose a random vector from the multivariate normal distribution and the generated datum will be rendered as

$$\mathbf{x}_g = \mathbf{A}\mathbb{E}\left[P\left(\mathbf{h}|\mathbf{x}'\right)\right] + \mathbf{G}\mathbf{w}. \tag{33}$$

  Examples of generated shapes are provided in the next section.

4 Experiments

Having described the training and inference procedure for MAPLDA, in this section we demonstrate the effectiveness of MAPLDA against PLDA [20], DS-LDA [27], Ioffe's PLDA variant [9], the Bayesian approach [17], LDA [4] and PCA [26], by performing several experiments on facial shapes from MeIn3D dataset [6]. In these experiments we only take into account the 3D shape of the human face *without* any texture information.

MeIn3d Dataset

MeIn3D dataset [6] consists of $10,000$ raw facial scans that describe a large variation of the population. More specifically, MeIn3D dataset [6] consists of data

Table 1. *Ethnicity* identification. Average identification rates ± standard deviations per method. MAPLDA outperforms all of the compared methods.

Method	Mean	Std
MAPLDA	**0.990**	0.051
PLDA	0.927	0.084
DS-LDA	0.919	0.073
PLDA (Ioffe)	0.917	0.089
Bayesian	0.911	0.077
LDA	0.878	0.079
PCA	0.634	0.083

annotated with multiple attributes (i.e., ethnicity, age, weight), thus it is highly appropriate for evaluating MAPLDA. Before performing any type of training or inference the scans are consistently re-parametrized into a form where the number of vertices, the triangulation and the anatomical meaning of each vertex are made consistent across all meshes. In this way all the training and the test meshes are brought into dense correspondence. In order to achieve this task we employ an optimal step non-rigid ICP algorithm [1]. We utilize the full spectrum of $10,000$ meshes where each mesh is labelled for a specific identity, age and ethnicity. The training and the inference is performed directly on the vectorized re-parametrized mesh of the form $\mathbb{R}^{3*N\times 1}$, where N is the distinct number of vertices.

4.1 Ethnicity Identification

In this experiment we identify the *ethnicity* attribute for a given 3D shape based on its shape features regardless of the *age-group* attribute (i.e., by marginalizing out the attribute *age-group*). We split the *ethnicity* attribute into three groups consisting of *White*, *Black* and *Asian* ethnic groups. We used 85% of the MeIn3D data for training and the rest for testing. Moreover, for each experiment, we used three random test data, with each test datum belonging in a different ethnic group. For the gallery we use the same set of distinct ethnic groups used in test samples from three random identities. We execute a total of 100 random experiments (i.e., we repeat the aforementioned process 100 times for randomly chosen test data and galleries in every experiment). Average identification rates along with the corresponding standard deviations per setting are shown in Table 1. Confusion matrices for MAPLDA and PLDA are provided in Table 2. As can be seen, MAPLDA outperforms all of the compared methods, thus demonstrating the advantages of joint attribute modeling.

4.2 Age-group Identification

In this experiment we identify the *age-group* for a given datum regardless of the *ethnicity* attribute (i.e., by marginalizing out the *ethnicity* attribute). We split the *age-group* attribute into four groups consisting of under 18 years old

Table 2. Confusion matrices of MAPLDA and PLDA for the *ethnicity* identification experiment. By incorporating the knowledge of the *age-group* attribute in the training phase, MAPLDA is able to better discriminate between the different ethnicities. In particular, MAPLDA classifies correctly all of the *Black* people in contrast with PLDA.

Actual	Predicted			Acc
	White	Black	Chinese	
White	**0.99**	0.00	0.01	0.99
Black	0.00	**1.00**	0.00	1.00
Chinese	0.02	0.00	**0.98**	0.98

(a) MAPLDA

Actual	Predicted			Acc
	White	Black	Chinese	
White	**0.97**	0.01	0.02	0.97
Black	0.04	**0.89**	0.07	0.89
Chinese	0.05	0.02	**0.93**	0.93

(b) PLDA [20]

Table 3. *Age-group* identification. Average identification rates ± standard deviations per method. MAPLDA outperforms all of the compared methods.

Method	Mean	Std
MAPLDA	**0.695**	0.063
PLDA [20]	0.540	0.079
PLDA (Ioffe) [9]	0.534	0.068
DS-LDA [27]	0.531	0.059
Bayesian [17]	0.529	0.071
LDA [4]	0.464	0.065
PCA [26]	0.327	0.074

(*<18*), *18-24*, *24-31* and *31-60* years old groups. We used 85% of the MeIn3D data for training and the rest for testing. Moreover, for each experiment we used four different random test data, with each test datum belonging in a different age group. For the gallery we use the same set of distinct age groups used in the test data from four random identities. We execute 100 random experiments per setting (i.e., we repeat the aforementioned process 100 times for randomly chosen probes and galleries in every experiment). Average identification rates along with the corresponding standard deviations per setting are shown in Table 3. Confusion matrices for MAPLDA and PLDA are provided in Table 4. The identification rates are considerably lower compared to the *ethnicity* experiment and that demonstrates that the task of inferring the age of a certain face by the shape of it is a challenging one. Nevertheless, our proposed framework exhibits performance that outperforms all of the compared methods by a large margin.

4.3 Weight-group Identification

In this experiment we identify the *weight-group* attribute for a given datum regardless of *age-group* attribute (i.e., by marginalizing out the attribute *age-group*). We split the weight attribute into five groups consisting of *30-45* kg, *45-55* kg, *55-62* kg, *62-70* kg and *70-80* kg groups. We used 85% of the MeIn3D data for training and the rest for testing. Similarly to our previous experiments,

Table 4. Confusion matrices of MAPLDA and PLDA for the *age-group* identification experiment. By incorporating the knowledge of the *ethnicity* attribute in the training phase, MAPLDA is able to better discriminate between the different age-groups.

Actual	Predicted				Acc
	< 18	18-24	24-31	31-60	
< 18	**0.77**	0.18	0.05	0	0.77
18-24	0.14	**0.62**	0.23	0.01	0.62
24-31	0.02	0.20	**0.66**	0.12	0.66
31-60	0	0.06	0.19	**0.75**	0.75

(a) MAPLDA

Actual	Predicted				Acc
	< 18	18-24	24-31	31-60	
< 18	**0.59**	0.27	0.13	0.01	0.59
18-24	0.17	**0.48**	0.31	0.04	0.48
24-31	0.02	0.24	**0.52**	0.22	0.52
31-60	0.02	0.13	0.28	**0.57**	0.57

(b) PLDA [20]

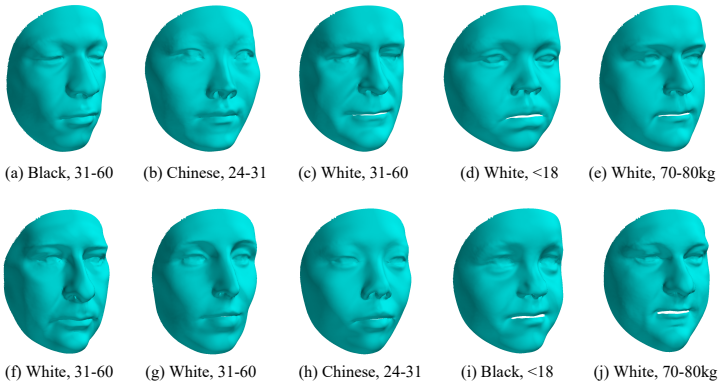

Fig. 4. 3D facial shapes generated via MAPLDA for different attribute combinations. Figures (a-d) and (f-i) visualize different instantiations of attributes *ethnicity* and *age group*, while Figures (e,j) of attributes *ethnicity* and *weight group*.

we use five different random test data, with each test datum belonging in a different weight group. For the gallery we use the same set of distinct weight groups used in the test samples from five random identities. We execute 100 random experiments per setting (i.e., we repeat the aforementioned process 100 times for randomly chosen test data and galleries in every experiment). Average identification rates along with the corresponding standard deviations per setting are shown in Table 5. Confusion matrices for MAPLDA and PLDA are provided in Table 6. Weight identification is considered to be the most challenging experiment of all three, as predicting the correct *weight group* solely from 3D facial shapes without considering the scaling factor is a very difficult problem. Nevertheless, as it can be seen in Table 5, the top performance is given by MAPLDA which is 51.6%, outperforming the other methods by a large margin.

Table 5. *Weight-group* identification. Average identification rates ± standard deviations per method. MAPLDA outperforms all of the compared methods.

Method	Mean	Std
MAPLDA	**0.516**	0.051
PLDA [20]	0.380	0.084
PLDA (Ioffe) [9]	0.373	0.049
DS-LDA [27]	0.368	0.054
Bayesian [17]	0.364	0.071
LDA [4]	0.346	0.059
PCA [26]	0.197	0.062

Table 6. Confusion matrices of MAPLDA and PLDA for the *weight-group* identification experiment. By incorporating the knowledge of the *age-group* attribute in the training phase, MAPLDA is able to better discriminate between the different weight-groups.

Actual	Predicted					Acc
	30-45	45-55	55-62	62-70	70-80	
30-45	**0.55**	0.26	0.14	0.04	0.01	0.55
45-55	0.23	**0.58**	0.11	0.05	0.03	0.58
55-62	0.09	0.15	**0.46**	0.23	0.07	0.46
62-70	0.02	0.10	0.19	**0.53**	0.16	0.53
70-80	0.02	0.08	0.17	0.24	**0.49**	0.49

(a) MAPLDA

Actual	Predicted					Acc
	30-45	45-55	55-62	62-70	70-80	
30-45	**0.41**	0.31	0.19	0.06	0.03	0.41
45-55	0.26	**0.44**	0.20	0.07	0.03	0.44
55-62	0.10	0.22	**0.32**	0.28	0.08	0.32
62-70	0.04	0.12	0.25	**0.38**	0.21	0.38
70-80	0.06	0.11	0.18	0.30	**0.35**	0.35

(b) PLDA [20]

4.4 Generating data

As thoroughly described in Section 3.3, the novel, multi-attribute nature of MAPLDA can be exploited to generate data with regards to a particular combination of attributes. By utilizing MeIn3D [6] dataset, we can train a multi-attribute model with regards to e.g., the *ethnicity* and *age-group* attributes and thus generate bespoke shapes that belong in a specific combination of attribute instantiations (e.g., ethnic group *Asian* and age group *24-31*). In Fig. 4, we visualize some examples of generated shapes belonging to a distinct combination of attributes such as *ethnicity* and *age-group* and *ethnicity* and *weight-group*.

5 Conclusions

In this paper, we introduced Multi-Attribute PLDA (MAPLDA), a novel component analysis method that is able to *jointly* model observations enriched with labels in terms of multiple attributes. We provide a probabilistic formulation and optimization procedure for training, as well as a flexible and efficient framework for inference on any subset of the attributes available during training. Evaluation is performed via several experiments on 3D facial shapes, namely ethnicity, age, and weight identification as well as 3D face generation under arbitrary instantiations of attributes. Results show that MAPLDA outperforms all compared methods, deeming the advantages of *joint* attribute modelling apparent.

References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007)
2. Archambeau, C., Delannay, N., Verleysen, M.: Mixtures of robust probabilistic principal component analyzers. Neurocomputing **71**(7), 1274–1282 (2008)
3. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis (2005)
4. Belhumeur, P.N., Hespanha, J.P., 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)
5. Bishop, C.M.: Pattern recognition & machine learning. Machine Learning (2006)
6. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5543–5552 (2016)
7. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological) pp. 1–38 (1977)
8. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16**(12), 2639–2664 (2004)
9. Ioffe, S.: Probabilistic linear discriminant analysis. In: Proceedings of the European Conference on Computer Vision, pp. 531–542. Springer (2006)
10. Jolliffe, I.: Principal component analysis. Wiley Online Library (2002)
11. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **16**(5), 980–988 (2008)
12. Klami, A., Virtanen, S., Kaski, S.: Bayesian canonical correlation analysis. The Journal of Machine Learning Research **14**(1), 965–1003 (2013)
13. Lawrence, N.: Probabilistic non-linear principal component analysis with gaussian process latent variable models. The Journal of Machine Learning Research **6**, 1783–1816 (2005)
14. Li, P., Fu, Y., Mohammed, U., Elder, J.H., Prince, S.J.: Probabilistic models for inference about identity. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 144–157 (2012)
15. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 248 (2015)
16. Lüthi, M., Gerig, T., Jud, C., Vetter, T.: Gaussian process morphable models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
17. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian face recognition. Pattern Recognition **33**(11), 1771–1782 (2000)
18. Moghaddam, B., Pentland, A.: Probabilistic visual learning for object representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(7), 696–710 (1997)
19. Nicolaou, M.A., Zafeiriou, S., Pantic, M.: A unified framework for probabilistic component analysis. In: Machine Learning and Knowledge Discovery in Databases, pp. 469–484. Springer (2014)

20. Prince, S.J., Elder, J.H.: Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1–8. IEEE (2007)
21. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) **36**(6), 245 (2017)
22. Roweis, S.: Em algorithms for pca and spca. Advances in Neural Information Processing Systems pp. 626–632 (1998)
23. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence (8), 831–836 (1996)
24. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. Neural Computation **11**(2), 443–482 (1999)
25. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622 (1999)
26. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–591. IEEE (1991)
27. Wang, X., Tang, X.: Dual-space linear discriminant analysis for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. II–II. IEEE (2004)
28. Wibowo, M.E., Tjondronegoro, D., Zhang, L., Himawan, I.: Heteroscedastic probabilistic linear discriminant analysis for manifold learning in video-based face recognition. In: IEEE Workshop on Applications of Computer Vision (WACV). pp. 46–52. IEEE (2013)
29. Yu, S., Yu, K., Tresp, V., Kriegel, H.P., Wu, M.: Supervised probabilistic principal component analysis. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining. pp. 464–473. ACM (2006)
30. Zhang, Y., Yeung, D.Y.: Heteroscedastic probabilistic linear discriminant analysis with semi-supervised extension. In: Machine Learning and Knowledge Discovery in Databases, pp. 602–616. Springer (2009)