A Scalable Formulation of Probabilistic Linear Discriminant Analysis: Applied to Face Recognition

Laurent El Shafey, Chris McCool, Roy Wallace, and Sébastien Marcel

APPENDIX A
MATHEMATICAL DERIVATIONS

The goal of the following section is to provide more detailed proofs of the formulae given in the article for both training and computing the likelihood.

The following proofs make use of a formulation of the inverse of a block matrix that uses the Schur complement. The corresponding identity can be found in [1] (Equations 1.11 and 1.10),

$$\begin{bmatrix} \boldsymbol{L} & \boldsymbol{M} \\ \boldsymbol{N} & \boldsymbol{O} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{R}, & -\boldsymbol{R}\boldsymbol{M}\boldsymbol{O}^{-1} \\ -\boldsymbol{O}^{-1}\boldsymbol{N}\boldsymbol{R}, & \boldsymbol{O}^{-1} + \boldsymbol{O}^{-1}\boldsymbol{N}\boldsymbol{R}\boldsymbol{M}\boldsymbol{O}^{-1} \end{bmatrix}, \tag{51}$$

where we have substituted $\boldsymbol{R} = \left(\boldsymbol{L} - \boldsymbol{M}\boldsymbol{O}^{-1}\boldsymbol{N}\right)^{-1}$.

Another related expression is the Woodbury matrix identity (Equation $C.7$ of [2]), which states that,

$$\begin{aligned}(\boldsymbol{L} + \boldsymbol{M}\boldsymbol{O}\boldsymbol{N})^{-1} &= \\ \boldsymbol{L}^{-1} - \boldsymbol{L}^{-1}&\boldsymbol{M}\left(\boldsymbol{O}^{-1} + \boldsymbol{N}\boldsymbol{L}^{-1}\boldsymbol{M}\right)^{-1}\boldsymbol{N}\boldsymbol{L}^{-1}.\end{aligned} \tag{52}$$

A. Scalable training

The bottleneck of the training procedure is the expectation step (E-Step) of the Expectation-Maximization algorithm. This E-Step requires the computation of the first and second order moments of the latent variables.

1) Estimating the first order moment of the Latent Variables: The most computationally expensive part when estimating the latent variables is the inversion of the matrix $\mathring{\tilde{\boldsymbol{\mathcal{P}}}}$ (Equation (27)). This matrix is block diagonal, the two blocks being $\boldsymbol{\mathcal{P}}_0$ (Equation (28)) and (a repetition of) $\boldsymbol{\mathcal{P}}_1$ (Equation (29)),

$$\mathring{\tilde{\boldsymbol{\mathcal{P}}}} = \begin{bmatrix} \boldsymbol{\mathcal{P}}_0 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\mathcal{P}}_1 & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\mathcal{P}}_1 \end{bmatrix}. \tag{53}$$

The inverse of $\boldsymbol{\mathcal{P}}_1$ is equal to the matrix $\boldsymbol{\mathcal{G}}$, defined by (30). This matrix is of constant size ($D_G \times D_G$), irrespective of the number of training samples for the class. In addition, the inversion of $\boldsymbol{\mathcal{P}}_0$ can be further optimised using the block matrix inversion identity introduced at the beginning of this section, leading to

$$\boldsymbol{\mathcal{P}}_0^{-1} = \begin{bmatrix} \boldsymbol{\mathcal{F}}_{J_i} & \sqrt{J_i}\boldsymbol{\mathcal{H}}^T \\ \sqrt{J_i}\boldsymbol{\mathcal{H}} & \left(\boldsymbol{I}_{D_G} - J_i\boldsymbol{\mathcal{H}}\boldsymbol{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G}\right)\boldsymbol{\mathcal{G}} \end{bmatrix}, \tag{54}$$

where $\boldsymbol{\mathcal{F}}_{J_i}$ is defined by (33) and $\boldsymbol{\mathcal{H}}$ by (37).

Then, the computation of $\mathring{\tilde{\boldsymbol{\mathcal{P}}}}^{-1}\mathring{\tilde{\boldsymbol{A}}}^T\tilde{\boldsymbol{\Sigma}}^{-1}$ gives a block diagonal matrix, the first block being

$$\begin{bmatrix} \sqrt{J_i}\boldsymbol{\mathcal{F}}_{J_i}\boldsymbol{F}^T\boldsymbol{\mathcal{S}} \\ \boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{I}_{D_x} - J_i\boldsymbol{F}\boldsymbol{\mathcal{F}}_{J_i}\boldsymbol{F}^T\boldsymbol{\mathcal{S}}\right) \end{bmatrix},$$

and the other ones being equal to $\boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}$.

As explained in section III.B.a of the article, $\boldsymbol{h}_i$ corresponds to the upper sub-vector of $\mathring{\tilde{\boldsymbol{y}}}_i$ and is not affected by the change of variable, as depicted in (21). Therefore, the first order moment of $\boldsymbol{h}_i$ is directly obtained by multiplying the first block-rows of the matrix $\mathring{\tilde{\boldsymbol{\mathcal{P}}}}^{-1}\mathring{\tilde{\boldsymbol{A}}}^T\tilde{\boldsymbol{\Sigma}}^{-1}$ with $\mathring{\tilde{\boldsymbol{x}}}_i$, which gives (31).

Considering only the $\mathring{\tilde{\boldsymbol{w}}}_i$ (lower) sub-vector of $\mathring{\tilde{\boldsymbol{y}}}_i$, the corresponding (lower) part $\mathring{\tilde{\boldsymbol{\mathcal{B}}}}$ of the matrix $\mathring{\tilde{\boldsymbol{\mathcal{P}}}}^{-1}\mathring{\tilde{\boldsymbol{A}}}^T\tilde{\boldsymbol{\Sigma}}^{-1}$ can be decomposed into a sum of two matrices, the first one being sparse with a single non-zero block (upper left) equal to $\boldsymbol{\mathcal{B}}_0 = -J_i\boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F}\boldsymbol{\mathcal{F}}_{J_i}\boldsymbol{F}^T\boldsymbol{\mathcal{S}}$, and the second one being diagonal by blocks with identical blocks $\boldsymbol{\mathcal{B}}_1 = \boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}$,

$$\mathring{\tilde{\boldsymbol{\mathcal{B}}}} = \begin{bmatrix} \boldsymbol{\mathcal{B}}_0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mathcal{B}}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{\mathcal{B}}_1 \end{bmatrix}. \tag{55}$$

Furthermore, the first order moment of the variables $\tilde{\boldsymbol{w}}_i$ is given by

$$\begin{aligned}E\left[\tilde{\boldsymbol{w}}_i|\tilde{\boldsymbol{x}}_i, \boldsymbol{\Theta}\right] &= \left(\tilde{\boldsymbol{U}}^T \otimes \boldsymbol{I}_{D_G}\right)\begin{bmatrix} \boldsymbol{\mathcal{B}}_0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\mathring{\tilde{\boldsymbol{x}}}_i \\ &+ \left(\tilde{\boldsymbol{U}}^T \otimes \boldsymbol{I}_{D_G}\right)\begin{bmatrix} \boldsymbol{\mathcal{B}}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{\mathcal{B}}_1 \end{bmatrix}\left(\tilde{\boldsymbol{U}} \otimes \boldsymbol{I}_{D_x}\right)\tilde{\boldsymbol{x}}_i.\end{aligned} \tag{56}$$

The previous decomposition greatly simplifies the computation, and leads to the following expression for each $\boldsymbol{w}_{i,j}$,

$$\begin{aligned}E\left[\boldsymbol{w}_{i,j}|\tilde{\boldsymbol{x}}_i, \boldsymbol{\Theta}\right] &= \boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}_{i,j} \\ &- \boldsymbol{\mathcal{G}}\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{F}\boldsymbol{\mathcal{F}}_{J_i}\boldsymbol{F}^T\boldsymbol{\mathcal{S}}\sum_j \bar{\boldsymbol{x}}_{i,j}\end{aligned} \tag{57}$$

L. El Shafey is with Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne, Switzerland e-mail: laurent.el-shafey@idiap.ch

C. McCool, R. Wallace and S. Marcel are with Idiap Research Institute, Martigny, Switzerland e-mail: {christopher.mccool,roy.wallace,sebastien.marcel}@idiap.ch

The research leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7) under grant agreements 238803 (BBfor2) and 257289 (TABULA RASA).