A Probabilistic Interpretation of Canonical Correlation Analysis

Francis R. Bach
Computer Science Division
University of California
Berkeley, CA 94114, USA
fbach@cs.berkeley.edu

Michael I. Jordan
Computer Science Division
and Department of Statistics
University of California
Berkeley, CA 94114, USA
jordan@cs.berkeley.edu

April 21, 2005

Technical Report 688
Department of Statistics
University of California, Berkeley

Abstract

We give a probabilistic interpretation of canonical correlation (CCA) analysis as a latent variable model for two Gaussian random vectors. Our interpretation is similar to the probabilistic interpretation of principal component analysis (Tipping and Bishop, 1999, Roweis, 1998). In addition, we cast Fisher linear discriminant analysis (LDA) within the CCA framework.

1 Introduction

Data analysis tools such as principal component analysis (PCA), linear discriminant analysis (LDA) and canonical correlation analysis (CCA) are widely used for purposes such as dimensionality reduction or visualization (Hotelling, 1936, Anderson, 1984, Hastie et al., 2001). In this paper, we provide a probabilistic interpretation of CCA and LDA. Such a probabilistic interpretation deepens the understanding of CCA and LDA as model-based methods, enables the use of local CCA models as components of a larger probabilistic model, and suggests generalizations to members of the exponential family other than the Gaussian distribution.

In Section 2, we review the probabilistic interpretation of PCA, while in Section 3, we present the probabilistic interpretation of CCA and LDA, with proofs presented in Section 4. In Section 5, we provide a CCA-based probabilistic interpretation of LDA.

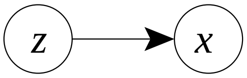

Figure 1: Graphical model for factor analysis.

2 Review: probabilistic interpretation of PCA

Tipping and Bishop (1999) have shown that PCA can be seen as the maximum likelihood solution of a factor analysis model with isotropic covariance matrix. More precisely, let $x = (x^1, \dots, x^n)$ denote n i.i.d. observations of an m-dimensional random vector, where $x^j = (x^j_1, \dots, x^j_m)^\top$ denotes the jth observation. The sample mean $\tilde{\mu}$ and sample covariance matrix $\widetilde{\Sigma}$ are defined as:

$$\tilde{\mu} = \frac{1}{n}\sum_{j=1}^{n} x^j \qquad \text{and} \qquad \widetilde{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}(x^j - \tilde{\mu})(x^j - \tilde{\mu})^\top.$$

PCA is concerned with finding a linear transformation $A \in \mathbb{R}^{d\times m}$ that renders the data uncorrelated with unit marginal variances. The linear transformation is equal to $A = R\Lambda_d^{-1/2}U_d$, where the d column vectors in the $m \times d$ matrix U_d are the d principal eigenvectors of $\widetilde{\Sigma}$, corresponding to eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_d$ and Λ_d is a $d \times d$ diagonal matrix with diagonal $\lambda_1, \dots, \lambda_d$. The matrix R is an arbitrary $d \times d$ orthogonal matrix.

Tipping and Bishop (1999) proved the following theorem:

Theorem 1 *The maximum likelihood estimates of the parameters W, μ and σ^2 of the following model (see the graphical model in Figure 1):*

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d) \\ x|z &\sim \mathcal{N}(Wz + \mu, \sigma^2 I_m), \quad \sigma > 0, \quad W \in \mathbb{R}^{md} \end{aligned}$$

are

$$\hat{\mu} = \tilde{\mu}, \quad \widehat{W} = U_d(\Lambda_d - \hat{\sigma}^2 I)^{1/2}R, \quad \hat{\sigma}^2 = \frac{1}{m-d}\sum_{i=d+1}^{m}\lambda_i, \tag{1}$$

where the d column vectors in the $m\times d$ matrix U_d are the d principal eigenvectors of $\widetilde{\Sigma}$, corresponding to eigenvalues $\lambda_1, \dots, \lambda_d$ in the $d\times d$ diagonal matrix Λ_d. R is an arbitrary $d\times d$ orthogonal matrix.

The posterior expectation of z given x is an affine function of x; with ML estimates for the parameters, we have:

$$E(z|x) = R^\top(\Lambda_d - \hat{\sigma}^2 I)^{1/2}\Lambda_d^{-1}U_d^\top(x - \hat{\mu}).$$

This yields the same linear subspace as PCA, and the projections are the same (up to rotation) if the discarded eigenvalues are zero.

3 Probabilistic interpretation of CCA

3.1 Definition of CCA

Given a random vector x, (PCA) is concerned with finding a linear transformation such that the components of the transformed vector are uncorrelated. Thus PCA diagonalizes the covariance matrix of x. Similarly, given two random vectors, x_1 and x_2, of dimension m_1 and m_2, *canonical correlation analysis (CCA)* is concerned with finding a pair of linear transformations such that one component within each set of transformed variables is correlated with a single component in the other set. The correlation matrix between x_1 and x_2 is reduced to a block diagonal matrix with blocks of size two, where each block is of the form $\begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$, potentially padded with the identity matrix if $m_1 \neq m_2$. The nonnegative numbers ρ_i, at most $p = \min\{m_1, m_2\}$ of which are nonzero, are called the *canonical correlations* and are usually ordered from largest to smallest.

We let $\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & \widetilde{\Sigma}_{22} \end{pmatrix}$ denote the $m \times m$ (where $m = m_1 + m_2$) sample covariance matrix obtained from data x_1^j, x_2^j, $j = 1, \ldots, n$. It is defined by blocks according to the partition $m = m_1 + m_2$. The canonical pairs of directions (u_{1i}, u_{2i}), $i = 1, \ldots, m$, are equal to $(u_{1i}, u_{2i}) = ((\widetilde{\Sigma}_{11})^{-1/2} v_{1i}, (\widetilde{\Sigma}_{22})^{-1/2} v_{2i})$, where (v_{1i}, v_{2i}) is a pair of left and right singular vectors of $(\widetilde{\Sigma}_{11})^{-1/2} \widetilde{\Sigma}_{12} (\widetilde{\Sigma}_{22})^{-1/2}$, with singular value equal to the canonical correlation ρ_i for $i = 1 \ldots, p$, and equal to 0, for $i > p$. In this note, we assume that the $p = \min\{m_1, m_2\}$ canonical correlations are distinct and nonzero, so that the normalized singular vectors are unique, up to a sign change [1]. We also assume that the sample covariance matrix $\widetilde{\Sigma}$ is invertible. If $U_1 = (u_{11}, \ldots, u_{1m})$ and $U_2 = (u_{21}, \ldots, u_{2m})$, then we have $U_1^\top \widetilde{\Sigma}_{11} U_1 = I_m$, $U_2^\top \widetilde{\Sigma}_{22} U_2 = I_m$, $U_2^\top \widetilde{\Sigma}_{21} U_1 = P$, where P is a $m_2 \times m_1$ diagonal matrix with canonical correlations on the diagonal.[2]

Note that due to the equivalence of the singular value decomposition of a rectangular matrix M and the eigen-decomposition of the matrix $\begin{pmatrix} 0 & M \\ M^\top & 0 \end{pmatrix}$ the CCA direction can also be obtained by solving the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \widetilde{\Sigma}_{12} \\ \widetilde{\Sigma}_{21} & 0 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \rho \begin{pmatrix} \widetilde{\Sigma}_{11} & 0 \\ 0 & \widetilde{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}. \tag{2}$$

In the next section, we show that the canonical directions emerge from fitting a certain latent variable model.

3.2 Latent variable interpretation

In this section, we consider model depicted in Figure 2 and show that maximum likelihood estimation based on this model leads to the canonical correlation directions.

Theorem 2 *The maximum likelihood estimates of the parameters W_1, W_2, Ψ_1, Ψ_2, μ_1 and μ_2 for the model defined in Figure 2 and by*

$$\begin{aligned} z &\sim \mathcal{N}(0, I_d), \quad \min\{m_1, m_2\} \geqslant d \geqslant 1 \\ x_1 | z &\sim \mathcal{N}(W_1 z + \mu_1, \Psi_1), \quad W_1 \in \mathbb{R}^{m_1 \times d}, \Psi_1 \succcurlyeq 0 \\ x_2 | z &\sim \mathcal{N}(W_2 z + \mu_2, \Psi_2), \quad W_2 \in \mathbb{R}^{m_2 \times d}, \Psi_2 \succcurlyeq 0 \end{aligned}$$

[1] The assumption of distinctness of the singular values is made to simplify the notation and the proof in Section 4, since with distinct singular values, canonical directions are uniquely defined. An extension of the result to the case of coalescing eigenvalues is straightforward.

[2] A rectangular diagonal matrix is defined as having nonzero elements only for equal row and column indices.

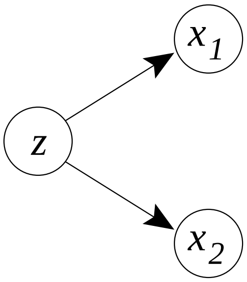

Figure 2: Graphical model for canonical correlation analysis.

are

$$\begin{aligned}
\widehat{W}_1 &= \widetilde{\Sigma}_{11} U_{1d} M_1 \\
\widehat{W}_2 &= \widetilde{\Sigma}_{22} U_{2d} M_2 \\
\hat{\Psi}_1 &= \widetilde{\Sigma}_{11} - \widehat{W}_1 \widehat{W}_1^\top \\
\hat{\Psi}_2 &= \widetilde{\Sigma}_{22} - \widehat{W}_2 \widehat{W}_2^\top \\
\hat{\mu}_1 &= \tilde{\mu}_1 \\
\hat{\mu}_2 &= \tilde{\mu}_2,
\end{aligned}$$

where $M_1, M_2 \in \mathbb{R}^{d \times d}$ are arbitrary matrices such that $M_1 M_2^\top = P_d$ and the spectral norms of M_1 and M_2 are smaller than one, where the ith columns of U_{1d} and U_{2d} are the first d canonical directions, and where P_d is the diagonal matrix of the first d canonical correlations.

The posterior expectations and variances of z given x_1 and x_2 are:

$$\begin{aligned}
E(z|x_1) &= M_1^\top U_{1d}^\top (x_1 - \hat{\mu}_1) \\
E(z|x_2) &= M_2^\top U_{2d}^\top (x_2 - \hat{\mu}_2) \\
\mathrm{var}(z|x_1) &= I - M_1 M_1^\top \\
\mathrm{var}(z|x_2) &= I - M_2 M_2^\top \\
E(z|x_1, x_2) &= \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^\top \begin{pmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{pmatrix} \begin{pmatrix} U_{1d}^\top (x_1 - \hat{\mu}_1) \\ U_{2d}^\top (x_2 - \hat{\mu}_2) \end{pmatrix} \\
\mathrm{var}(z|x_1, x_2) &= I - \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}^\top \begin{pmatrix} (I - P_d^2)^{-1} & (I - P_d^2)^{-1} P_d \\ (I - P_d^2)^{-1} P_d & (I - P_d^2)^{-1} \end{pmatrix} \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}.
\end{aligned}$$

Whatever M_1 and M_2 are, the projections $E(z|x_1)$ and $E(z|x_2)$ lie the d-dimensional subspaces of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$ that are identical to those obtained from CCA.

Note that among all solutions, the solutions that minimize $-\log|\Psi| = -\log|\Psi_1| - \log|\Psi_2|$ (i.e., that maximize the conditional entropy of x given z), are such that $M_1 = M_2 = M$ is a square root of S, i.e., $M = S^{1/2} R$, where R is a rotation matrix, and the solutions are then:

$$\begin{aligned}
\widehat{W}_1 &= \widetilde{\Sigma}_{11} U_{1d} M R \\
\widehat{W}_2 &= \widetilde{\Sigma}_{22} U_{2d} M R \\
\widehat{\Psi}_1 &= (\widetilde{\Sigma}_{11})^{1/2} (I - P_d) (\widetilde{\Sigma}_{11})^{1/2} \\
\widehat{\Psi}_2 &= (\widetilde{\Sigma}_{22})^{1/2} (I - P_d) (\widetilde{\Sigma}_{22})^{1/2}
\end{aligned}$$

4 Proof of Theorem 2

The proof follows along the lines of the proof of Tipping and Bishop (1999): we first show that stationary points of the likelihood are combinations of canonical directions and then that the canonical correlations must be largest for the maximum likelihood estimates. The marginal mean and covariance matrix of $x = (x_1, x_2)$ under our model are $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} W_1 W_1^\top + \Psi_1 & W_1 W_2^\top \\ W_2 W_1^\top & W_2 W_2^\top + \Psi_2 \end{pmatrix}$.

The negative log likelihood of the data is equal to (with $|A|$ denoting the determinant of a square matrix A):

$$\begin{aligned}
\ell_1 &= \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{1}{2}\sum_{j=1}^{n} \operatorname{tr}\Sigma^{-1}(x_j-\mu)(x_j-\mu)^\top \\
&= \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{1}{2}\sum_{j=1}^{n} \left(\operatorname{tr}\Sigma^{-1}x_j x_j^\top - 2x_j^\top\Sigma^{-1}\mu\right) + \frac{n}{2}\mu^\top\Sigma^{-1}\mu \\
&= \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{n}{2}\operatorname{tr}\Sigma^{-1}(\widetilde{\Sigma} - \tilde{\mu}\tilde{\mu}^\top) - n\tilde{\mu}^\top\Sigma^{-1}\mu + \frac{n}{2}\mu^\top\Sigma^{-1}\mu \\
&= \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{n}{2}\operatorname{tr}\Sigma^{-1}\widetilde{\Sigma} + \frac{n}{2}(\tilde{\mu}-\mu)^\top\Sigma^{-1}(\tilde{\mu}-\mu).
\end{aligned}$$

We first maximize with respect to μ, which yields a maximum at $\mu = \tilde{\mu}$ (the sample mean). Plugging back into the log likelihood, we obtain the following profile negative log likelihood:

$$\ell_2 = \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{n}{2}\operatorname{tr}\Sigma^{-1}\widetilde{\Sigma},$$

with $\Sigma = WW^\top + \Psi$, if we define $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ and $\Psi = \begin{pmatrix} \Psi_1 & 0 \\ 0 & \Psi_2 \end{pmatrix}$.

The value of the negative log likelihood is infinite if the covariance model $\Sigma = WW^\top + \Psi$ is non-invertible, thus we can restrict ourselves to $\Sigma \succ 0$. Stationary points of the likelihood are defined by the following equations (obtained by computing derivatives):

$$(\Sigma^{-1} - \Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1})W = 0 \tag{3}$$

$$(\Sigma^{-1} - \Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1})_{11} = 0 \tag{4}$$

$$(\Sigma^{-1} - \Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1})_{22} = 0, \tag{5}$$

where Eq. (3) is obtained by differentiating with respect to $W = (W_1, W_2)$, Eq. (4) is obtained by differentiating with respect to Ψ_1 and where Eq. (5) is obtained by differentiating with respect to Ψ_2. A_{11} denotes the upper left block (of size $m_1 \times m_1$) of A, and similarly for A_{22}.

We now assume that we have a stationary point (W, Ψ_1, Ψ_2) of the log likelihood. We have the following lemmas:

Lemma 1 $\widetilde{\Sigma} \succcurlyeq WW^\top$.

Proof Since Σ is invertible, Eq. (3) implies

$$W = \widetilde{\Sigma}(WW^\top + \Psi)^{-1}, \tag{6}$$

which implies by right-multiplication by $W^\top\Psi^{-1/2}$:

$$\widetilde{W}\widetilde{W}^\top = \Phi^{-1/2}\widetilde{\Sigma}\Phi^{-1/2}(\widetilde{W}\widetilde{W}^\top + I)^{-1}\widetilde{W}\widetilde{W}^\top,$$

where $\widetilde{W} = \Psi^{-1/2}W$. If we let $ASA^\top$ denote the eigenvalue decomposition of $\widetilde{W}\widetilde{W}^\top$, with $A^\top A = I$ and S diagonal, then this implies:

$$\widetilde{\Sigma} = \Phi^{1/2}A(S+I)A^\top\Phi^{1/2} \succcurlyeq \Phi^{1/2}ASA^\top\Phi^{1/2} = WW^\top.$$

■

Lemma 2 $\widetilde{\Sigma}\Sigma^{-1} = (\widetilde{\Sigma} - WW^\top)\Psi^{-1}$.

Proof By the matrix inversion lemma, we have

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1}W(I + W^\top\Psi^{-1}W)^{-1}W^\top\Psi^{-1}$$

and

$$\Sigma^{-1}W = \Psi^{-1}W(I + W^\top\Psi^{-1}W)^{-1}.$$

Thus,

$$\begin{aligned}
\widetilde{\Sigma}\Sigma^{-1} &= \widetilde{\Sigma}\Psi^{-1} - \widetilde{\Sigma}\Psi^{-1}W(I + W^\top\Psi^{-1}W)W^\top\Psi^{-1} \\
&= \widetilde{\Sigma}\Psi^{-1} - \widetilde{\Sigma}\Sigma^{-1}W^\top\Psi^{-1} \\
&= \widetilde{\Sigma}\Psi^{-1} - WW^\top\Psi^{-1},
\end{aligned}$$

by Eq. (6). ■

Lemma 3 $\Psi(\Sigma^{-1} - \Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1})\Psi = \Psi - (\widetilde{\Sigma} - WW^\top)$.

Proof We have:

$$\begin{aligned}
\Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1} &= \Sigma^{-1}(\widetilde{\Sigma} - WW^\top)\Psi^{-1} \qquad \text{(by Lemma 2)} \\
&= \left[(\widetilde{\Sigma} - WW^\top)\Psi^{-1}\right]^\top \Psi^{-1} - \Sigma^{-1}WW^\top\Psi^{-1} \qquad \text{(by Lemma 2)} \\
&= \Psi^{-1}\widetilde{\Sigma}\Psi^{-1} - \Psi^{-1}WW^\top\Psi^{-1} - \Sigma^{-1}(WW^\top + \Psi - \Psi)\Psi^{-1} \\
&= \Psi^{-1}\widetilde{\Sigma}\Psi^{-1} - \Psi^{-1}WW^\top\Psi^{-1} - \Psi^{-1} + \Sigma^{-1},
\end{aligned}$$

which implies the lemma. ■

Lemma 4 *$\Psi_1 = \widetilde{\Sigma}_{11} - W_1W_1^\top \succcurlyeq 0$, and $\Psi_2 = \widetilde{\Sigma}_2 - W_2W_2^\top \succcurlyeq 0$.*

Proof Taking the first diagonal block of the previous lemma, we obtain:

$$\Psi_1(\Sigma^{-1} - \Sigma^{-1}\widetilde{\Sigma}\Sigma^{-1})_{11}\Psi_1 = \Psi_1 - (\widetilde{\Sigma}_{11} - W_1W_1^\top).$$

By Eq. (4), we get $\Psi_1 - (\widetilde{\Sigma}_{11} - W_1W_1^\top)$, i.e., $\Psi_1 - \widetilde{\Sigma}_{11} - W_1W_1^\top$. The matrix is positive semidefinite by Lemma 1. The proof is the same for Ψ_2. ■

Lemma 5

$$W = (\widetilde{\Sigma} - WW^\top)\begin{pmatrix} (\widetilde{\Sigma}_{11} - W_1W_1^\top)^{-1} & 0 \\ 0 & (\widetilde{\Sigma}_{22} - W_2W_2^\top)^{-1} \end{pmatrix} W.$$

Proof Simply plug the expressions for Ψ_1 and Ψ_2 in Lemma 4 into Lemma 2. ■

Lemma 6 *Let $\Sigma_{11}^{-1/2}W_1 = A_1S_1B_1^\top$ and $\Sigma_{22}^{-1/2}W_2 = A_2S_2B_2^\top$ be the singular value decompositions of $\Sigma_{11}^{-1/2}W_1$ and $\Sigma_{22}^{-1/2}W_2$, where A_1 and A_2 are $m_2 \times d$ matrices with orthonormal columns, where B_1 and B_2 are $d \times d$ orthogonal matrices, and where S_1 and S_2 are $d \times d$ diagonal matrices. Let $\widetilde{C}_{12} = (\widetilde{\Sigma}_{11})^{-1/2}\widetilde{\Sigma}_{12}(\widetilde{\Sigma}_{11})^{-1/2}$. We then have:*

$$\begin{aligned}
\widetilde{C}_{12}A_2 &= A_1S_1B_1^\top B_2S_2 && (7)\\
\widetilde{C}_{12}^\top A_1 &= A_2S_2B_2^\top B_1S_1 && (8)\\
\widetilde{C}_{12}^\top\widetilde{C}_{12}A_2 &= A_2S_2B_2^\top B_1S_1^2B_1^\top B_2S_2 && (9)
\end{aligned}$$

Proof Lemma 5 can be rewritten using the singular value decomposition as follows:

$$\begin{aligned}
\begin{pmatrix} A_1S_1B_1^\top \\ A_2S_2B_2^\top \end{pmatrix} &= \begin{pmatrix} I - A_1S_1^2A_1^\top & \widetilde{C}_{12} - A_1S_1B_1^\top B_2S_2A_2^\top \\ \widetilde{C}_{21} - A_2S_2B_2^\top B_1S_1A_1^\top & I - A_2S_2^2A_2^\top \end{pmatrix}\\
&\qquad \times \begin{pmatrix} I - A_1S_1^2A_1^\top & 0 \\ 0 & I - A_2S_2^2A_2^\top \end{pmatrix}^{-1} \begin{pmatrix} A_1S_1B_1^\top \\ A_2S_2B_2^\top \end{pmatrix}.
\end{aligned}$$

Considering the first block we obtain:

$$A_1S_1B_1^\top = A_1S_1B_1^\top + (I - A_1S_1^2A_1^\top)^{-1}(\widetilde{C}_{12} - A_1S_1B_1^\top B_2S_2A_2^\top)(I - A_2S_2^2A_2^\top)^{-1},$$

which implies Eq. (7). Eq. (8) can be obtained using the second block, while Eq. (9) is implied by Eq. (7) and Eq. (8). ■

We can now rewrite the likelihood ℓ_2 for a stationary point:

$$\ell_3 = \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log\begin{vmatrix} \widetilde{\Sigma}_{11} & W_1W_2^\top \\ W_2W_1^\top & \widetilde{\Sigma}_{22} \end{vmatrix} + \frac{n}{2}\operatorname{tr}\begin{pmatrix} \widetilde{\Sigma}_{11} & W_1W_2^\top \\ W_2W_1^\top & \widetilde{\Sigma}_{22} \end{pmatrix}^{-1}\widetilde{\Sigma}$$

Using the Schur complement lemma for determinants, we obtain:

$$\begin{aligned}
\ell_3 &= \frac{(m_1+m_2)n}{2}\log 2\pi + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}|\\
&\quad + \frac{n}{2}\log|I - (\widetilde{\Sigma}_{22})^{-1/2}W_2W_1^\top(\widetilde{\Sigma}_{11})^{-1}W_1W_2^\top(\widetilde{\Sigma}_{11})^{-1/2}| + \frac{n}{2}(m_1+m_2)\\
&= \frac{(m_1+m_2)n}{2}\log 2\pi e + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}| + \frac{n}{2}\log|I - A_2S_2B_2^\top B_1S_1^2B_1^\top B_2S_2A_2^\top|\\
&= \frac{(m_1+m_2)n}{2}\log 2\pi e + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}|\\
&\quad + \frac{n}{2}\log|I - A_2A_2^\top + A_2(I - S_2B_2^\top B_1S_1^2B_1^\top B_2S_2)A_2^\top|\\
&= \frac{(m_1+m_2)n}{2}\log 2\pi e + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}| + \frac{n}{2}\log|I - S_2B_2^\top B_1S_1^2B_1^\top B_2S_2|\\
&= \frac{(m_1+m_2)n}{2}\log 2\pi e + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}| + \frac{n}{2}\log|I - A_2^\top\widetilde{C}_{12}^\top\widetilde{C}_{12}A_2|
\end{aligned}$$

The term $\log|I - A_2^\top \widetilde{C}_{12}^\top \widetilde{C}_{12} A_2|$, for A_2 an $m_2 \times d$ matrix with orthonormal columns, is lower bounded by $\sum_{i=1}^d \log(1-\rho_i^2)$ where $\{\rho_i\}_{i=1}^d$ are the d largest canonical correlations, and there is equality if and only if $A_2 = V_{2d}R_2$, where R_2 is an arbitrary $d \times d$ orthogonal matrix and the columns of V_{2d} are the d largest singular vector of $\widetilde{C}_{12}$.

The minimum of the likelihood is thus $\frac{(m_1+m_2)n}{2} \log 2\pi e + \frac{n}{2}\log|\widetilde{\Sigma}_{11}| + \frac{n}{2}\log|\widetilde{\Sigma}_{22}| + \frac{n}{2}\sum_i \log(1-\rho_i^2)$.

The singular value decomposition of $(\widetilde{\Sigma}_{11})^{-1/2}\widetilde{\Sigma}_{12}(\widetilde{\Sigma}_{22})^{-1/2}$ is $(\widetilde{\Sigma}_{11})^{-1/2}\widetilde{\Sigma}_{12}(\widetilde{\Sigma}_{22})^{-1/2} = V_1 P V_2^\top$, where V_1 is $m_1 \times m_1$ orthonormal, where P is an $m_1 \times m_2$ diagonal matrix (recall our earlier definition of a diagonal rectangular matrix), and where V_2 is an $m_2 \times m_2$ orthogonal matrix. We let V_{1d} and V_{2d} denote the first d singular vectors and let P_d denote the diagonal matrix of the largest d canonical correlations.

The minimum is attained at all points such that $A_1 = V_{1d}R_1$, $A_2 = V_{2d}R_2$, where R_1 and R_2 are $d \times d$ orthogonal matrices. Plugging into the stationary equations Eq. (7) and Eq. (8), we get:

$$\begin{aligned} V_1 P V_2^\top V_{2d} R_2 &= V_{1d} R_1 S_1 B_1^\top B_2 S_2 \\ V_2 P^\top V_1^\top V_{1d} R_1 &= V_{2d} R_2 S_2 B_2^\top B_1 S_1. \end{aligned}$$

Left-multiplying the first equation by $V_{1d}^\top$ and the second equation by $V_{2d}^\top$, and right-multiplying by $R_2^\top$ and $R_1^\top$, we obtain

$$\begin{aligned} P_d &= R_1 S_1 B_1^\top B_2 S_2 R_2^\top \\ P_d^\top &= R_2 S_2 B_2^\top B_1 S_1 R_1^\top. \end{aligned}$$

Maximum likelihood solutions are therefore of the form

$$\begin{aligned} \hat{W}_1 &= (\widetilde{\Sigma}_{11})^{1/2} V_{1d} M_1 = \widetilde{\Sigma}_{11} U_{1d} M_1 \\ \hat{W}_2 &= (\widetilde{\Sigma}_{22})^{1/2} V_{2d} M_2 = \widetilde{\Sigma}_{22} U_{2d} M_2, \end{aligned}$$

where $M_1 \in \mathbb{R}^{d\times d}$ and $M_2 \in \mathbb{R}^{d\times d}$ are such that $M_1 M_2^\top = P_d$ and the spectral norms of M_1 and M_2 are smaller than one (so that $\Psi_1 \succcurlyeq 0$ and $\Psi_2 \succcurlyeq 0$). The matrices U_{1d} and U_{2d} are composed of the first canonical directions.

4.1 EM algorithm

The EM algorithm provides a general framework for fitting the parameters of latent variable models (Dempster et al., 1977). In particular, our latent variable formulation of CCA readily yields the following EM update equations:

$$\begin{aligned} W_{t+1} &= \widetilde{\Sigma}\Psi_t^{-1} W_t M_t (M_t + M_t W_t^\top \Psi_t^{-1} \widetilde{\Sigma} \Psi_t^{-1} W_t M_t)^{-1} \\ \Psi_{t+1} &= \begin{pmatrix} (\widetilde{\Sigma} - \widetilde{\Sigma}\Psi_t^{-1} W_t M_t W_{t+1}^\top)_{11} & 0 \\ 0 & (\widetilde{\Sigma} - \widetilde{\Sigma}\Psi_t^{-1} W_t M_t W_{t+1}^\top)_{22} \end{pmatrix}, \end{aligned}$$

where $M_t = (I + W_t^\top \Psi_t^{-1} W_t)^{-1}$.

The EM algorithm always converges to a solution of the form described above, where the specific solution that is found among solutions of this form depends on the initialization.

4.2 CCA, maximum likelihood and low-rank approximation

Comparing the theorem for PCA of Tipping and Bishop (1999) with the analogous result for CCA that we have provided in Theorem 2, we see that for PCA the maximum likelihood covariance matrix

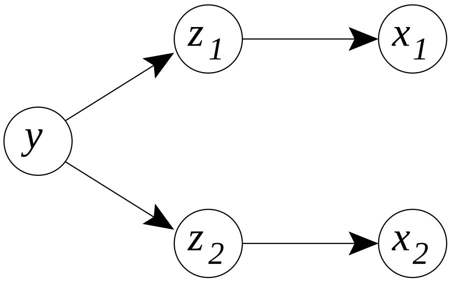

Figure 3: Alternative graphical model for canonical correlation analysis.

has the form “rank d + constant × identity” and is obtained from the principal component directions, whereas for CCA the maximum likelihood joint covariance matrix is such that the cross-covariance matrix has rank d and is obtained from the canonical correlation directions. The following theorem makes this precise:

Theorem 3 *If the sample joint covariance matrix has full rank and the canonical correlations are distinct, then the maximum likelihood covariance matrix with rank d cross-covariance is*

$$\left(\begin{array}{cc} \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{11} U_{1d} P_d U_{2d}^\top \widetilde{\Sigma}_{22} \\ \widetilde{\Sigma}_{22} U_{2d} P_d U_{1d}^\top \widetilde{\Sigma}_{11} & \widetilde{\Sigma}_{22} \end{array} \right).$$

Proof The theorem follows from the fact that any covariance matrix of rank d can be written in the form $\left(\begin{array}{cc} W_1 W_1^\top + \Psi_1 & W_1 W_2^\top \\ W_2 W_1^\top & W_2 W_2^\top + \Psi_2 \end{array} \right)$.

■

4.3 Alternative model

In this section, we consider an alternative model that also leads to the canonical directions. The graphical model representation is shown in Figure 3. We have the following theorem:

Theorem 4 *The maximum likelihood estimates of the parameters* $W_1, W_2, \Psi_1, \Psi_2, \mu_1, \mu_2$, Φ_1 *and* Φ_2 *for the model defined by (see Figure 3):*

$$\begin{array}{rcl} y & \sim & \mathcal{N}(0, I_d), \quad \min\{m_1, m_2\} \geqslant d \geqslant 1 \\ z_1|y & \sim & \mathcal{N}(y, \Phi_1), \quad \Phi_1 \succcurlyeq 0 \\ z_2|y & \sim & \mathcal{N}(y, \Phi_2), \quad \Phi_2 \succcurlyeq 0 \\ x_1|z_1 & \sim & \mathcal{N}(W_1 z_1 + \mu_1, \Psi_1), \quad W_1 \in \mathbb{R}^{m_1 \times d}, \Psi_1 \succcurlyeq 0 \\ x_2|z_2 & \sim & \mathcal{N}(W_2 z_2 + \mu_2, \Psi_2), \quad W_2 \in \mathbb{R}^{m_2 \times d}, \Psi_2 \succcurlyeq 0 \end{array}$$

are such that

$$\begin{aligned}
\hat{W}_1 &= \widetilde{\Sigma}_{11} U_{1d} M_1 \\
\hat{W}_2 &= \widetilde{\Sigma}_{22} U_{2d} M_2 \\
\hat{\Psi}_1 + \hat{W}_1 \hat{\Phi}_1 \hat{W}_1^\top &= \widetilde{\Sigma}_{11} - \hat{W}_1 \hat{W}_1^\top \\
\hat{\Psi}_2 + \hat{W}_2 \hat{\Phi}_1 \hat{W}_2^\top &= \widetilde{\Sigma}_{22} - \hat{W}_2 \hat{W}_2^\top \\
\hat{\mu}_1 &= \tilde{\mu}_1 \\
\hat{\mu}_2 &= \tilde{\mu}_2,
\end{aligned}$$

where $M_1 \in \mathbb{R}^{d \times d}$ and $M_2 \in \mathbb{R}^{d \times d}$ are such that $M_1 M_2^\top = P_d$ and the spectral norms of M_1 and M_2 are smaller than one, where the ith columns of U_{1d} and U_{2d} are the first d canonical directions, and where P_d is the diagonal matrix of the first d canonical correlations.

Proof The joint covariance of the data under model is:

$$\begin{pmatrix} W_1 W_1^\top + \Psi_1 + W_1 \Phi_1 W_1^\top & W_1 W_2^\top \\ W_2 W_1^\top & W_2 W_2^\top + \Psi_2 + W_2 \Phi_2 W_2^\top \end{pmatrix},$$

and the result follows from the previous theorem.

■

The maximum likelihood estimates such that the conditional variance Ψ_1 and Ψ_2 are minimal are of the form:

$$\begin{aligned}
\hat{W}_1 &= \widetilde{\Sigma}_{11} U_{1d} M_1 \\
\hat{W}_2 &= \widetilde{\Sigma}_{22} U_{2d} M_2 \\
\hat{\Psi}_1 &= \widetilde{\Sigma}_{11} - \widetilde{\Sigma}_{11}^{1/2} V_{1d} V_{1d}^\top \widetilde{\Sigma}_{11}^{1/2} = \widetilde{\Sigma}_{11} - U_{1d} U_{1d}^\top \\
\hat{\Psi}_2 &= \widetilde{\Sigma}_{22} - \widetilde{\Sigma}_{22}^{1/2} V_{2d} V_{2d}^\top \widetilde{\Sigma}_{22}^{1/2} = \widetilde{\Sigma}_{22} - U_{2d} U_{2d}^\top \\
\hat{\Phi}_1 &= M_1^{-1} M_1^{-\top} - I \\
\hat{\Phi}_2 &= M_2^{-1} M_2^{-\top} - I \\
\hat{\mu}_1 &= \tilde{\mu}_1 \\
\hat{\mu}_2 &= \tilde{\mu}_2.
\end{aligned}$$

For those estimates, we have the following posterior expectations:

$$\begin{aligned}
E(z_1|x_1) &= M_1^{-1} U_{1d}^\top (x_1 - \hat{\mu}_1) \\
E(z_2|x_2) &= M_2^{-1} U_{2d}^\top (x_2 - \hat{\mu}_2).
\end{aligned}$$

5 LDA as CCA

Given random vectors $y^1, \ldots, y^j \in \mathbb{R}^p$ and labels $t^1, \ldots, t^j \in \{1, \ldots, k\}$, Fisher linear discriminant analysis is a dimensionality reduction technique that works as follows: partition the data into the k classes, compute the sample means $m_i \in \mathbb{R}^p$ and sample covariance matrices $S_i \in \mathbb{R}^{p \times p}$ of the data in each class, $i = 1, \ldots, k$, define $n_i = \#\{j, t^j = i\}$ as the number of data points belonging to class i, let $\pi_i = n_i/n$, $i = 1, \ldots, k$ and let $M = (m_1, \ldots, m_k) \in \mathbb{R}^{p \times k}$. Define the "within-class covariance matrix" $S_W = \sum_i \pi_i S_i$, and the "between-class covariance matrix" $S_B = M(\mathrm{diag}(\pi) - \pi\pi^\top)M^\top$. Finally, solve the generalized eigenvalue problem $S_B\alpha = \lambda S_W \alpha$.

We now define two random vectors as follows: $x_1 = y \in \mathbb{R}^p$, and $x_2 \in \mathbb{R}^k$, defined as $(x_2)_i = 1$ if and only if $t = i$. A short calculation shows that the joint sample covariance matrix of x_1, x_2 is equal to

$$\widetilde{\Sigma} = \left(\begin{array}{cc} S_W + S_B & M(\mathrm{diag}(\pi) - \pi\pi^\top) \\ (\mathrm{diag}(\pi) - \pi\pi^\top)M^\top & \mathrm{diag}(\pi) - \pi\pi^\top \end{array} \right)$$

and thus CCA aims to find singular vectors of $\widetilde{C}_{12} = \widetilde{\Sigma}_{11}^{-1/2}\widetilde{\Sigma}_{12}\widetilde{\Sigma}_{11}^{-1/2}$, which is equivalent to finding eigenvectors of

$$\begin{aligned} \widetilde{C}_{12}\widetilde{C}_{12}^\top &= \widetilde{\Sigma}_{11}^{-1/2}\widetilde{\Sigma}_{12}\widetilde{\Sigma}_{11}^{-1}\widetilde{\Sigma}_{12}\widetilde{\Sigma}_{11}^{-1/2} \\ &= (S_W + S_B)^{-1/2} M(\mathrm{diag}(\pi) - \pi\pi^\top)M^\top (S_W + S_B)^{-1/2} \\ &= (S_W + S_B)^{-1/2} S_B (S_W + S_B)^{-1/2}, \end{aligned}$$

which is itself equivalent to solving the generalized eigenvalue problem $S_B\beta = \lambda(S_W + S_B)\beta$, i.e., $S_B\beta = \frac{\lambda}{1-\lambda} S_W \beta$. This shows that LDA for (y, t) is equivalent to CCA for (x_1, x_2).

References

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley & Sons, 1984.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:185–197, 1977.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, 2001.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

S. Roweis. EM algorithms for PCA and SPCA. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998.

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