Pattern
Recognition
Winter 93
W. Stuetzle

Connections between Canonical Correlation Analysis, Linear Discriminant Analysis, and Optimal Scaling

As ususal, let X be the $n \times p$ matrix of predictor variables, and let Y be the $n \times K$ matrix of dummy response variables. (In principle, $K-1$ dummy variables would be enough, but having K of them will be convenient below). We assume that the predictors are centered ($\bar{\mathbf{x}} = 0$). Let W and B be the within-class and between-class ssp matrices:

$$\begin{aligned} W &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T \\ B &= \sum_{k=1}^{K} n_k \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^T, \end{aligned}$$

where $\mathbf{x}_{ki}, i = 1, \ldots, n_k$ are the class k predictor vectors. Note that $\text{rank}(B) \leq K-1$ because the K class mean vectors are linearly dependent: $\sum_{k=1}^{K} n_k \bar{\mathbf{x}}_k = \mathbf{0}$.

Linear Discriminant Analysis (LDA)

The first *discriminant direction* $\mathbf{a}_1$ is the unit vector that maximizes the ratio of between class sum of squares to within class sum of squares of the projected data:

$$\mathbf{a}_1 = \text{argmax}_{\mathbf{a}} \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T W \mathbf{a}}$$

The i-th discriminant direction maximizes the ratio subject to being orthogonal to all previous $i-1$ directions. In general, there will be $\min(p, K-1)$ discriminant directions.

Canonical Correlation Analysis (CCA)

The first *canonical correlation* ρ_1 is the maximum correlation between a vector in X-space (the space spanned by the columns $X_1, \ldots X_p$ of X) and a vector in Y-space:

$$\rho_1 = \max_{\mathbf{a},\theta} \mathbf{cor}\,(X\mathbf{a}, Y\theta)$$

The vectors $\mathbf{a}_1$ and θ_1 maximizing the correlation are only determined up to a multiplicative constant. We will refer to the corresponding unit vectors as the first *canonical direction* and the first *canonical scores*. The i-th canonical correlation is the maximum correlation between a vector in X-space uncorrelated with $X\mathbf{a}_1, \ldots, X\mathbf{a}_{i-1}$, and a vector in Y-space uncorrelated with $Y\theta_1, \ldots, Y\theta_{i-1}$. The corresponding unit vectors $\mathbf{a}_i$ and θ_i are called the i-th canonical direction and the i-th canonical scores. In general, there are $\min(K, p)$ canonical correlations. At least one of those will vanish, however, because the Y-space contains the constant vector $\mathbf{1}$, to which the columns of X are orthogonal.

Optimal Scaling

Optimal scaling is frequently used by social scientists to extend extend statistical procedures like principal component analysis and regression to categorical data. We discuss it here for the classification problem, i.e. regression with categorical response. The goal is to find *scores* $\theta_1, \ldots, \theta_K$ for the categories $1, \ldots, K$ such that the transformed response vector $Y\theta$ can be predicted as well as possible by a linear model:

$$\hat{\theta} = \text{argmin}_\theta \min_{\mathbf{a}} \|Y\theta - X\mathbf{a}\|^2$$

To prevent the degenerate solution $\theta = \mathbf{0}$, we have to impose a constraint, like $\|Y\theta\|^2 = 1$. Optimal scaling is just another way of looking at the canonical correlation problem, that makes it look less symmetric and more like a regression problem.

Equivalence of CCA and LDA

To show the equivalence between CCA and LDA, we need the following

Lemma: Let $\mathbf{x}$ be a centered response vector, and let Y be a design matrix, which does not have to be centered, but whose column space contains the constant vector $\mathbf{1}$. Then a coefficient vector θ maximizing $\mathbf{cor}\,(\mathbf{x}, Y\theta)$ can be found by minimizing $\|\mathbf{x} - Y\theta\|$.

Proof: The squared correlation between $\mathbf{x}$ and $Y\theta$ is

$$\mathbf{cor}^{\,2}(\mathbf{x}, Y\theta) = \frac{\langle \mathbf{x}, Y\theta \rangle^2}{\|\mathbf{x}\|^2 \ \mathbf{var}\,(Y\theta)}$$

The centering terms disappear because $\mathbf{x}$ is already centered. Note that the optimal $Y\theta$ is only determined up to a constant vector, because

$$\mathbf{cor}^{\,2}(\mathbf{x}, Y\theta) = \mathbf{cor}^{\,2}(\mathbf{x}, Y\theta + c\mathbf{1}).$$

We thus need to maximize only over the subspace S of θ's with $\mathbf{mean}\,(Y\theta) = 0$, which implies that $\mathbf{var}\,(Y\theta) = \|Y\theta\|^2$:

$$\max_{\theta} \mathbf{cor}^{\,2}(\mathbf{x}, Y\theta) = \max_{\theta \in S} \frac{\langle \mathbf{x}, Y\theta \rangle^2}{\|x\|^2 \|Y\theta\|^2}$$

It is easy to see that the unconstrained optimum

$$\max_{\theta} \frac{\langle \mathbf{x}, Y\theta \rangle^2}{\|x\|^2 \|Y\theta\|^2}$$

can be found by minimizing $\|\mathbf{x} - Y\theta\|$. Moreover,

$$\hat{\theta} = \mathrm{argmin}_{\theta} \|\mathbf{x} - Y\theta\|$$

also satisfies the constraint, because $0 = \bar{\mathbf{x}} = \mathbf{mean}\,(Y\hat{\theta})$. (In the last step we use the fact that the column space of Y contains the constant vector). Therefore, $\hat{\theta}$ is also a solution to the constrained problem.

Here now is the main result.

Prop: The first discriminant direction is the same as the first canonical direction.

Proof: The maximum squared correlation between $X\mathbf{a}$ and $Y\theta$ is

$$\rho_1^2 = \max_{\mathbf{a},\theta} \frac{\langle X\mathbf{a}, Y\theta\rangle}{\|X\mathbf{a}\|^2 \, \mathbf{var}\,(Y\theta)}$$

As shown in Lemma 1, an optimal θ for given $\mathbf{a}$ can be found by linear regression of $X\mathbf{a}$ on on the Y-space:

$$\text{argmax}_\theta \mathbf{cor}^{\,2}(X\mathbf{a}, Y\theta) = (Y^TY)^{-1}Y^TX\mathbf{a}$$

Substituting into the correlation gives

$$\begin{aligned} \rho_1^2 &= \max_{\mathbf{a}} \frac{\langle X\mathbf{a}, Y(Y^TY)^{-1}Y^TX\mathbf{a}\rangle^2}{\|X\mathbf{a}\|^2 \, \|Y(Y^TY)^{-1}Y^TX\mathbf{a}\|^2} \\ &= \frac{\mathbf{a}^TX^TY(Y^TY)^{-1}Y^TX\mathbf{a}}{\|X\mathbf{a}\|^2} \end{aligned}$$

The key point is that the matrix $X^TY(Y^TY)^{-1}Y^TX$ in the numerator actually is the between class ssp matrix. Notice that Y^TX is a $K \times p$ matrix whose k-th row is $n_k\bar{\mathbf{x}}_k$, $(Y^TY)^{-1} = \text{diag}(1/n_1, \ldots, 1/n_K)$, and thus $X^TY(Y^TY)^{-1}Y^TX$ is indeed the properly weighted ssp matrix of the class means.

As a consequence,

$$\begin{aligned} \rho_1^2 &= \max_{\mathbf{a}} \frac{\mathbf{a}^TB\mathbf{a}}{\mathbf{a}^TX^TX\mathbf{a}} \\ &= \max_{\mathbf{a}} \frac{\mathbf{a}^TB\mathbf{a}}{\mathbf{a}^T(B+W)\mathbf{a}} \\ &= \max_{\mathbf{a}} (1 + \frac{\mathbf{a}^TW\mathbf{a}}{\mathbf{a}^TB\mathbf{a}})^{-1} \end{aligned}$$

This shows that the vector $\mathbf{a}$ giving the maximum correlation also maximizes the ratio of between class sum of squares to within class sum of squares of the projected data.