Dimension Reduction in LDA

Stephen Vardeman
Analytics Iowa LLC
ISU Statistics and IMSE

"Sphering"

Returning specifically to LDA, let

$$\bar{\boldsymbol{\mu}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_k$$

and note that one is free to replace $\mathbf{x}$ and all K means $\boldsymbol{\mu}_k$ with respectively

$$\mathbf{x}^* = \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \bar{\boldsymbol{\mu}}) \quad \text{and} \quad \boldsymbol{\mu}_k^* = \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})$$

This produces

$$\ln \left(\frac{P[y = k|\mathbf{x}^*]}{P[y = l|\mathbf{x}^*]} \right) = \ln \left(\frac{\pi_k}{\pi_l} \right) - \frac{1}{2} \left\| \mathbf{x}^* - \boldsymbol{\mu}_k^* \right\|^2 + \frac{1}{2} \left\| \mathbf{x}^* - \boldsymbol{\mu}_l^* \right\|^2$$

1st sphered form of LDA

In sphered form, the theoretically optimal (LDA) classifier/decision rule can be described as

$$f(\mathbf{x}) = \arg\max_{k} \left[\ln(\pi_k) - \frac{1}{2} \left\| \mathbf{x}^* - \boldsymbol{\mu}_k^* \right\|^2 \right]$$

That is, in terms of $\mathbf{x}^*$, optimal decisions are based on ordinary Euclidian distances to the transformed means $\boldsymbol{\mu}_k^*$. Further, this form can often be made even simpler.

2^{nd} sphered form of LDA

The μ_k^* typically span a subspace of $\Re^p$ of dimension $\min(p, K-1)$. For

$$\underset{p \times K}{\mathbf{M}} = (\mu_1^*, \mu_2^*, \ldots, \mu_K^*)$$

let $\mathbf{P_M}$ be the $p \times p$ matrix of projection onto $C(\mathbf{M})$ (the column space of $\mathbf{M}$ in $\Re^p$). Since $(\mathbf{P_M x}^* - \mu_k^*) \in C(\mathbf{M})$ and $(\mathbf{I} - \mathbf{P_M})\mathbf{x}^* \in C(\mathbf{M})^\perp$

$$\begin{aligned}\|\mathbf{x}^* - \mu_k^*\|^2 &= \|(\mathbf{P_M x}^* - \mu_k^*) + (\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2 \\ &= \|\mathbf{P_M x}^* - \mu_k^*\|^2 + \|(\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2\end{aligned}$$

Further, since $\|(\mathbf{I} - \mathbf{P_M})\mathbf{x}^*\|^2$ doesn't depend upon k, an optimal classifier can be described as

$$f(\mathbf{x}) = \arg\max_k \left[\ln(\pi_k) - \frac{1}{2}\|\mathbf{P_M x}^* - \mu_k^*\|^2\right]$$

in terms of the projection of $\mathbf{x}^*$ onto $C(\mathbf{M})$ and its distances to the μ_k^*.

Eigen analysis for covariance matrix of sphered means

$\frac{1}{K}\mathbf{MM}'$ is the sample covariance matrix of the $\boldsymbol{\mu}_k^*$ (typically of rank $\min(p, K-1)$) and has eigen decomposition

$$\frac{1}{K}\mathbf{MM}' = \mathbf{VDV}'$$

for $\mathbf{D} = \mathbf{diag}\,(d_1, d_2, \ldots, d_p)$ where $d_1 \geq d_2 \geq \cdots \geq d_p$ are the eigenvalues, and the columns of $\mathbf{V}$ are orthonormal eigenvectors corresponding in order to the eigenvalues. These $\mathbf{v}_k$ with $d_k > 0$ specify linear combinations of the coordinates of the $\boldsymbol{\mu}_l^*$, $\langle \mathbf{v}_k, \boldsymbol{\mu}_l^* \rangle$, with the largest sample variances subject to the constraints that $\|\mathbf{v}\| = 1$ and $\langle \mathbf{v}_l, \mathbf{v}_k \rangle = 0$ for all $l < k$. (These $\mathbf{v}_k$ are $\perp$ vectors in successive directions of most important unaccounted-for spread of the $\boldsymbol{\mu}_k^*$.) This suggests the possibility of "reduced rank" LDA.

Dimension reduction in LDA

That is, for $l \leq \text{rank}(\mathbf{MM}')$ define

$$\mathbf{V}_l = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_l)$$

let

$$\mathbf{P}_l = \mathbf{V}_l \mathbf{V}_l'$$

be the matrix projecting onto $C(\mathbf{V}_l)$ in $\Re^p$. A possible "reduced rank" approximation to the theoretically optimal LDA classification rule is

$$f_l(\mathbf{x}) = \arg\max_k \left[\ln(\pi_k) - \frac{1}{2} \left\| \mathbf{P}_l \mathbf{x}^* - \mathbf{P}_l \boldsymbol{\mu}_k^* \right\|^2 \right]$$

and l becomes a complexity parameter that one might optimize to regularize the method.

Graphical representation of reduced rank LDA

Note also that for $\mathbf{w} \in \Re^p$

$$\mathbf{P}_l \mathbf{w} = \sum_{k=1}^{l} \langle \mathbf{v}_k, \mathbf{w} \rangle \mathbf{v}_k$$

For purposes of graphical representation of what is going on in these computations, one might replace the p coordinates of $\mathbf{x}$ and the means $\boldsymbol{\mu}_k$ with the l (so-called "canonical") coordinates of

$$\left(\langle \mathbf{v}_1, \mathbf{x}^* \rangle , \langle \mathbf{v}_2, \mathbf{x}^* \rangle , \ldots , \langle \mathbf{v}_l, \mathbf{x}^* \rangle \right)' \tag{1}$$

and of the

$$\left(\langle \mathbf{v}_1, \boldsymbol{\mu}_k^* \rangle , \langle \mathbf{v}_2, \boldsymbol{\mu}_k^* \rangle , \ldots , \langle \mathbf{v}_l, \boldsymbol{\mu}_k^* \rangle \right)' \tag{2}$$

It seems to be essentially ordered pairs of these coordinates that are plotted in HTF in their Figures 4.8 and 4.11.

Arbitrary signs

Regarding this graphical method, we need to point out that since any eigenvector $\mathbf{v}_k$ could be replaced by $-\mathbf{v}_k$ without any fundamental effect in the above development, the vector (1) and all of the vectors (2) could be altered by multiplication of any particular set of coordinates by -1. (Whether a particular algorithm for finding eigenvectors produces $\mathbf{v}_k$ or $-\mathbf{v}_k$ is not fundamental, and there seems to be no standard convention in this regard.) It appears that the pictures in HTF might have been made using the R function `lda` and its choice of signs for eigenvectors.

Basis functions/transforms

The form $\mathbf{x}'\beta$ is (of course and by design) linear in the coordinates of $\mathbf{x}$. An obvious natural generalization of this discussion is to consider discriminants that are linear in some (non-linear) functions of the coordinates of $\mathbf{x}$. This is simply choosing some M basis functions/transforms/features $h_m(\mathbf{x})$ and replacing the p coordinates of $\mathbf{x}$ with the M coordinates of $(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_M(\mathbf{x}))$ in the development of LDA.

Of course, upon choosing basis functions that are all coordinates, squares of coordinates, and products of coordinates of $\mathbf{x}$, one produces *linear* (in the basis functions) discriminants that are general *quadratic* functions of $\mathbf{x}$. The possibilities opened here are myriad and (as always) "the devil is in the details."