Note that the PCs decompose the total variance in the data in the following way :

$$\sum_{i=1}^{d} Var(\mathbf{u_i}^T X) = \mathbf{u_i}^T S \mathbf{u_i} = \lambda_i$$

$$= \sum_{i=1}^{d} (\lambda_i)$$

$$= \mathbf{Tr}(S)$$

$$= Var(X)$$

$Var(\boldsymbol{u_i}^\top \boldsymbol{u_i}) = \boldsymbol{u_i}^\top S \boldsymbol{u_i} = \lambda_i$ where λ_i is an eigenvalue of the sample covariance matrix S and $\boldsymbol{u_i}$ is its corresponding eigenvector.

$Var(\boldsymbol{u_1}^T X)$ is maximized if $\boldsymbol{u_1}$ is the eigenvector of S with the corresponding maximum eigenvalue.

Each successive PC can be generated in the above manner by taking the eigenvectors of S that correspond to the eigenvalues:

$\lambda_1 \geq ... \geq \lambda_d$

such that

$Var(\boldsymbol{u_1}^T X) \geq ... \geq Var(\boldsymbol{u_d}^T X)$

Algorithm 1

Recover basis (PCs): Calculate $XX^\top = \sum_{i=1}^{n} x_i x_i^\top$ and let $U =$ eigenvectors of $XX^\top$ corresponding to the top p eigenvalues.

Encode training data: $Y = U^\top X$ where Y is a $p \times n$ matrix of encodings of the original data.

Reconstruct training data: $\hat{X} = UY = UU^\top X$.

Encode test example: $y = U^\top x$ where y is a p-dimensional encoding of x.

Reconstruct test example: $\hat{x} = Uy = UU^\top x$.

Table: *Direct PCA Algorithm*

- A unique solution can be obtained by finding the singular value decomposition of X
- For each rank p, U consists of the first p columns of U.

$$X = U\Sigma V^T$$

- The columns of U in the SVD contain the eigenvectors of XX^T

PCA vs. Fisher's Linear Discriminant Analysis

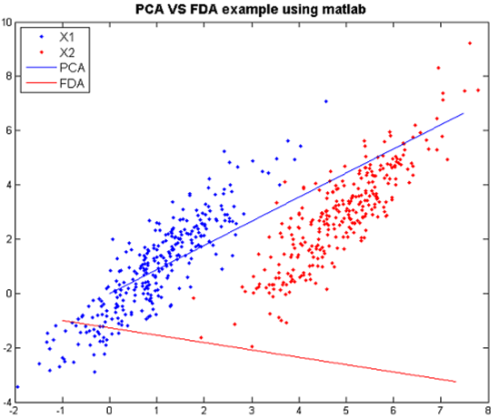

Figure: Projection of data from two classes onto various directions.

Fisher's Linear Discriminant Analysis (Two class problem)

- Assume we have only two classes.
- The idea behind Fisher's Linear Discriminant Analysis is to reduce the dimensionality of the data to one dimension. That is, to take d-dimensional $\mathbf{x} \in \mathbb{R}^d$ and map it to one dimension by finding $\mathbf{w}^T\mathbf{x}$:

$$z = \mathbf{w}^T\mathbf{x} = \begin{bmatrix} w_1 & \cdots & w_d \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \sum_{i=1}^{d} w_i x_i$$

- The one-dimensional z is then used for classification.

Fisher's Linear Discriminant Analysis

Goal: To find a direction such that projected data $\mathbf{w}^T\mathbf{x}$ are well separated.

Consider the two-class problem:

$$\mu_0 = \frac{1}{n_0}\sum_{i:y_i=0} x_i \qquad \mu_1 = \frac{1}{n_1}\sum_{i:y_i=1} x_i$$

We want to:

1. Maximize the distance between projected class means.
2. Minimize the within class variance.

Fisher's Linear Discriminant Analysis

The distance between projected class means is:

$$
\begin{aligned}
(\mathbf{w}^T\mu_0 - \mathbf{w}^T\mu_1)^2 &= (\mathbf{w}^T\mu_0 - \mathbf{w}^T\mu_1)^T(\mathbf{w}^T\mu_0 - \mathbf{w}^T\mu_1) \\
&= (\mu_0 - \mu_1)^T\mathbf{w}\mathbf{w}^T(\mu_0 - \mu_1) \\
&= \mathbf{w}^T(\mu_0 - \mu_1)(\mu_0 - \mu_1)^T\mathbf{w} \\
&= \mathbf{w}^T S_B \mathbf{w}
\end{aligned}
$$

where S_B is the between-class variance (known).

Fisher's Linear Discriminant Analysis

Minimizing the within-class variance is equivalent to minimizing the sum of all individual within-class variances. Thus the within class variance is:

$$\begin{aligned} \mathbf{w}^T\Sigma_0\mathbf{w} + \mathbf{w}^T\Sigma_1\mathbf{w} &= \mathbf{w}^T(\Sigma_0 + \Sigma_1)\mathbf{w} \\ &= \mathbf{w}^T S_W \mathbf{w} \end{aligned}$$

where S_W is the within-class covariance (known).

Fisher's Linear Discriminant Analysis

To maximize the distance between projected class means and minimize the within-class variance, we can maximize the ratio:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

This is equivalent to finding:

$$\max_{\mathbf{w}} \mathbf{w}^T S_B \mathbf{w}$$

subject to the constraint:

$$\mathbf{w}^T S_W \mathbf{w} = 1$$

Fisher's Linear Discriminant Analysis

To turn this constraint optimization problem into a non-constranst optimization problem, we apply Lagrange multipliers:

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T S_B \mathbf{w} - \lambda(\mathbf{w}^T S_W \mathbf{w} - 1)$$

Differentiating with respect to $\mathbf{w}$ we get:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} = 2S_B\mathbf{w} - \lambda 2 S_W \mathbf{w} &= 0 \\ S_B \mathbf{w} &= \lambda S_W \mathbf{w} \end{aligned}$$

Fisher's Linear Discriminant Analysis

This is a generalized eigenvector problem that is equivalent to (if S_W is not singular):

$$S_W^{-1} S_B \mathbf{w} = \lambda \mathbf{w}$$

where λ and $\mathbf{w}$ are the eigenvalues and eigenvectors of $S_W^{-1} S_B$ respectively.

$\mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue of $S_W^{-1} S_B$.