PCA, SVD, and LDA

Shanshan Ding

Spring 2015

Dimension Reduction

We generally do not want to feed a large number of features directly into a machine learning algorithm because:

- They are expensive to store.
- They slow down computations.
- Large samples are required to avoid overfitting.
- In algorithms like k-nearest neighbors, distances in high dimensions are distorted.

Principal component analysis is one method of reducing the number of dimensions in the raw data.

Intuition: Change of Basis

To capture as much information from data as possible in a low number of dimensions, we find a basis of **principal components**. Each principal component is the vector along which variance is maximized, conditioning on it being orthogonal to all preceding principal components.

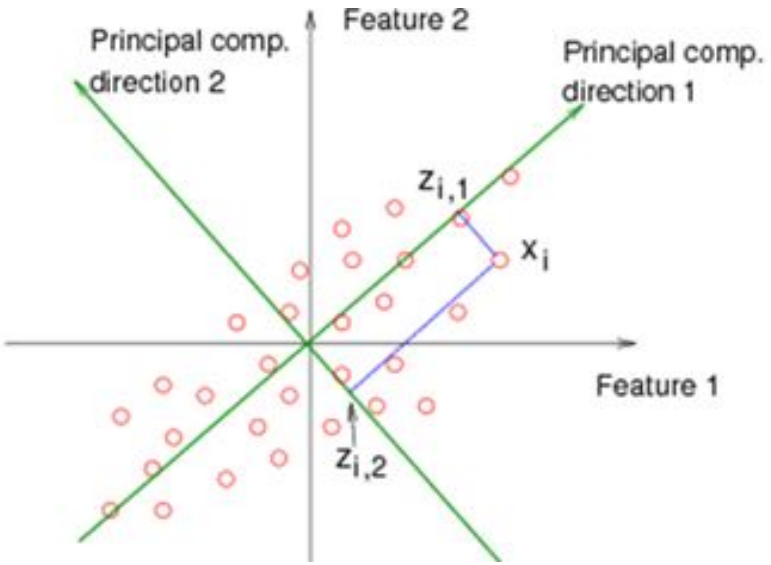


Figure: Courtesy of https://onlinecourses.science.psu.edu/stat857

Variance of Projection

Let $\phi(x_i)$ be the feature vector of x_i, and suppose that data has been centered and normalized in the feature space. The variance of the projection of $\phi(x)$ onto a unit vector w is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|P_{\mathrm{w}}(\phi(x_i))\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{\phi(x_i)\cdot \mathrm{w}}{\mathrm{w}\cdot\mathrm{w}}\mathrm{w}\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathrm{w}^T\phi(x_i)\phi(x_i)^T\mathrm{w}) \\ &= \mathrm{w}^T\hat{\mathbb{E}}(\phi(x)\phi(x)^T)\mathrm{w} = \mathrm{w}^T\mathrm{C}\mathrm{w},\end{aligned}$$

where $\mathrm{C} = \frac{1}{n}\mathrm{X}^T\mathrm{X}$ is the empirical covariance matrix of the dataset X. As C is positive semi-definite, it has a full orthonormal set of eigenvectors with eigenvalues that are all real and positive.

First Principal Component

The direction along which variance is maximized is the solution to

$$\max_{\mathrm{w}} \mathrm{w}^T \mathrm{C}\mathrm{w}, \text{ subject to } \|\mathrm{w}\| = 1,$$

or equivalently,

$$\max_{\mathrm{w}} \frac{\mathrm{w}^T \mathrm{C}\mathrm{w}}{\mathrm{w}^T \mathrm{w}}.$$

Solving with Lagrange multipliers gives that

$$\mathrm{C}\mathrm{w} = \lambda \mathrm{w},$$

so that w is an eigenvector of C. Since

$$\mathrm{w}^T \mathrm{C}\mathrm{w} = \mathrm{w}^T \lambda \mathrm{w} = \lambda,$$

we see that the first principal component w_1 is in fact the eigenvector corresponding to the largest eigenvalue of C.

Further Principal Components

To find the k-th principal component, consider the deflated matrix

$$\mathrm{C}_k = \mathrm{C} - \sum_{i=1}^{k-1} \mathrm{C}\mathrm{w}_i\mathrm{w}_i^T.$$

As $\mathrm{C}_k\mathrm{w}_j = 0$ for $j < k$, deflation reduces the first $k-1$ eigenvalues of C to 0. The eigenvector corresponding to the largest remaining eigenvalue is then the k-th principal component of X.

Let W be the matrix whose columns are the w_j. Since $\phi(x_i) \cdot \mathrm{w}_j$ gives the signed magnitude of $P_{\mathrm{w}_j}\phi(x_i)$, the score matrix

$$\mathrm{M} = \mathrm{X}\mathrm{W}$$

gives the coordinates of the data matrix w.r.t the w_j.

Singular Value Decomposition

PCA is often performed via **singular value decomposition**, because forming $\mathrm{X}^T\mathrm{X}$ would not be required:

Theorem

For any matrix $\mathrm{X} \in \mathbb{R}^{n \times d}$, there exist orthogonal matrices $\mathrm{U} \in \mathbb{R}^{n \times n}$, $\mathrm{W} \in \mathbb{R}^{d \times d}$ and (rectangular) diagonal matrix $\Sigma \in \mathbb{R}^{n \times d}$ with non-negative entries such that

$$\mathrm{X} = \mathrm{U}\Sigma\mathrm{W}^T.$$

Since $\mathrm{X}^T\mathrm{X} = \mathrm{W}\Sigma^T\Sigma\mathrm{W}^T$,

- The diagonal entries σ_j of Σ, known as the singular values of X, are square roots of the eigenvalues of $\mathrm{X}^T\mathrm{X}$. By convention, these are listed in descending order.
- The columns of W are the eigenvectors of $\mathrm{X}^T\mathrm{X}$.

PCA in terms of SVD

The score matrix M can be expressed as

$$\mathrm{M} = \mathrm{XW} = \mathrm{U}\Sigma\mathrm{W}^T\mathrm{W} = \mathrm{U}\Sigma.$$

To consider only the first k principal components, we compute

$$\mathrm{M}_k = \mathrm{XW}_k = \mathrm{U}\Sigma_k = \mathrm{U}_k\Sigma_k,$$

where the subscript k denotes the matrix formed by zeroing out all columns after the k-th. The *proportion of variance* explained by the first k principal components is $\sum_{j=1}^{k} \sigma_j^2 / \sum_{j=1}^{d} \sigma_j^2$.

It can be shown that $\mathrm{U}\Sigma_k\mathrm{W}^T$ is the closest rank-k approximation to X in the Frobenius norm, so that M_k is the k-column matrix with smallest reconstruction error $\|\mathrm{MW}^T - \mathrm{M}_k(\mathrm{W}_k)^T\|_F$.

SVD and Linear Regression

Recall the objective function of linear regression

$$(\mathrm{y} - \mathrm{X}\beta)^T(\mathrm{y} - \mathrm{X}\beta),$$

which, assuming that $\mathrm{X}^T\mathrm{X}$ is invertible, has solution

$$\hat{\beta}^{\mathrm{ls}} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y}.$$

Substituting $\mathrm{U}\Sigma\mathrm{W}^T$ for X,

$$\begin{aligned}\hat{\mathrm{y}}^{\mathrm{ls}} = \mathrm{X}\hat{\beta}^{\mathrm{ls}} &= \mathrm{U}\Sigma\mathrm{W}^T(\mathrm{W}\Sigma^T\Sigma\mathrm{W}^T)^{-1}\mathrm{W}\Sigma^T\mathrm{U}^T\mathrm{y} \\ &= \mathrm{U}\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T\mathrm{U}^T\mathrm{y} = \sum_{j=1}^{d}\mathrm{u}_j\mathrm{u}_j^T\mathrm{y}.\end{aligned}$$

Hence $\hat{\mathrm{y}}^{\mathrm{ls}}$ is a linear combination of projections of y onto the u_j, which themselves can be interpreted as normalized projections of X onto the j-th principal component of the data.

SVD and L^2 Regularization

Now recall the objective function of ridge regression

$$(\mathrm{y} - \mathrm{X}\beta)^T(\mathrm{y} - \mathrm{X}\beta) + \lambda \mathrm{X}^T\mathrm{X},$$

which has solution

$$\hat{\beta}^{\text{ridge}} = (\mathrm{X}^T\mathrm{X} + \lambda \mathrm{I})^{-1}\mathrm{X}^T\mathrm{y}.$$

Since

$$\begin{aligned}\hat{\mathrm{y}}^{\text{ridge}} = \mathrm{X}\hat{\beta}^{\text{ridge}} &= \mathrm{U}\Sigma\mathrm{W}^T(\mathrm{W}\Sigma^T\Sigma\mathrm{W}^T + \lambda \mathrm{I})^{-1}\mathrm{W}\Sigma^T\mathrm{U}^T \\ &= \mathrm{U}\Sigma(\Sigma^T\Sigma + \lambda \mathrm{I})^{-1}\Sigma^T\mathrm{U}^T\mathrm{y} = \sum_{j=1}^{d} \mathrm{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda}\mathrm{u}_j^T\mathrm{y},\end{aligned}$$

more shrinkage is applied along principal components with less variance. Finally, let H_λ be the hat matrix of the regression. As

$$\mathrm{df}(\lambda) := \mathrm{tr}(\mathrm{H}_\lambda) = \sum_{j=1}^{d} \frac{\sigma_j^2}{\sigma_j^2 + \lambda},$$

regularization reduces the *effective* number of parameters.

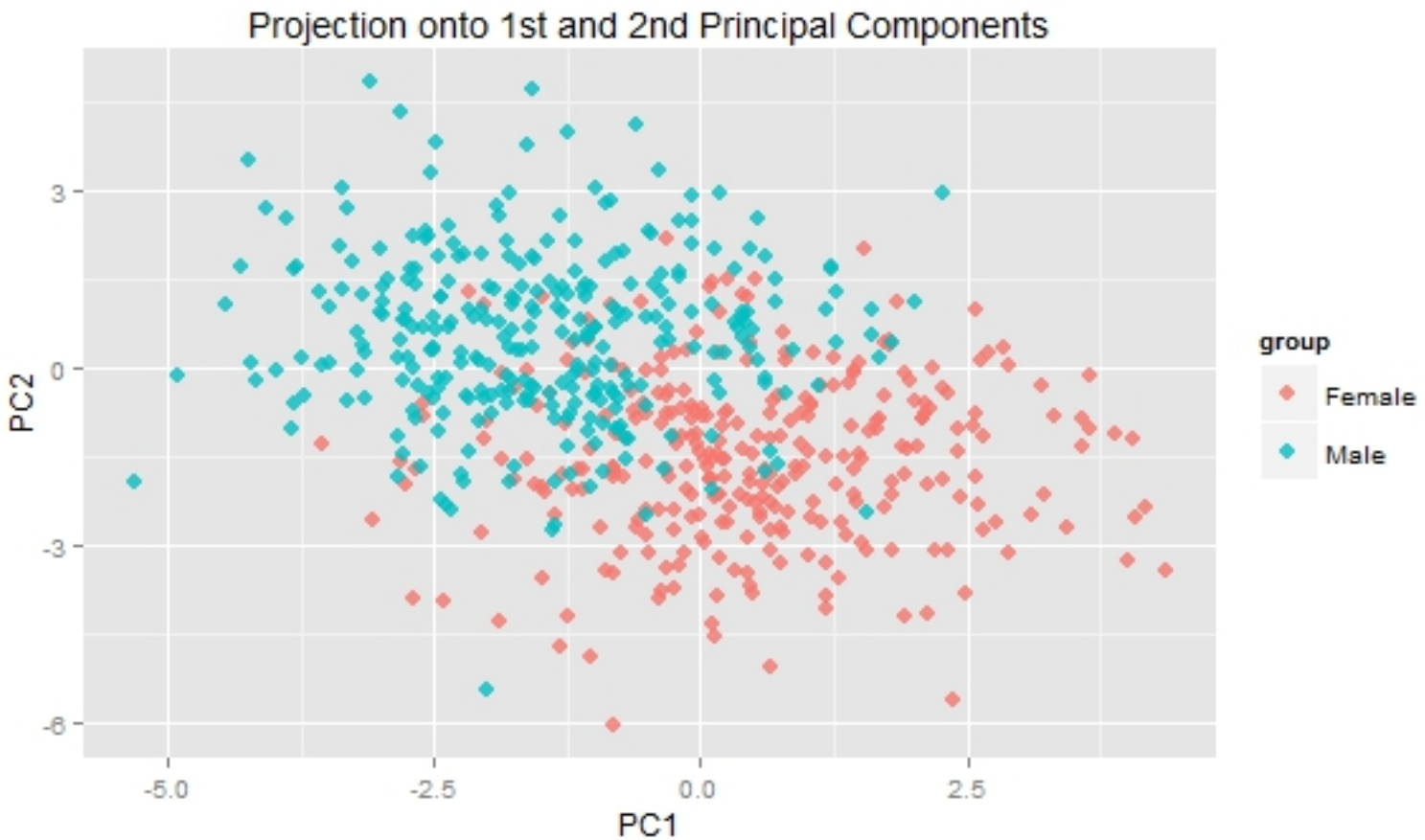
Example: User Study

Task: Highlight how different demographic groups (e.g. male vs. female) use an app differently.

One approach:

1. Perform PCA (via SVD) on the data.
2. Plot the data along the first two or three principal components, i.e. plot the columns of $\mathrm{U}_k\Sigma_k$ for $k = 2$ or 3.
3. Interpret results according to which features are weighted most heavily in these components.

User Study: Visualization



User Study: Interpreting Results

Suppose that the raw features were

(profile length, # pictures, # friends, frequency of log-ins, average session length, # messages initiated, # messages responded),

with

$$w_1 = (0.71, -0.42, -0.34, -0.17, -0.16, -0.02, 0.38) \text{and}$$

$$w_2 = (-0.01, 0.13, 0.42, -0.84, 0.17, -0.11, 0.23).$$

Then "profile length" and "frequency of log-ins" are resp. the features weighted most heavily in w_1 and w_2. Thus the previous plot suggests that

- Female users tend to have longer profiles.
- Female users tend to log in more frequently (since the weight of the feature in w_2 is negative, being more negative along w_2 is correlated with being larger in the feature).

Supervised Dimension Reduction

Performing PCA on user-level data is

- a good first pass at exploring differences between user groups as SVD only needs to be done once for multiple sets of comparisons, but
- principal components highlight features that generate maximal variance among all users, not necessarily between the groups we are interested in comparing.

When the goal of dimension reduction is maximal separation between labeled classes, we seek projections that maximize between-class variance (as normalized by within-class variance).

Fisher's Linear Discriminant Analysis

Recall the PCA optimization problem

$$\max_{\mathrm{w}} \frac{\mathrm{w}^T \mathrm{C} \mathrm{w}}{\mathrm{w}^T \mathrm{w}}, \text{ where } \mathrm{C} = \frac{1}{n} \mathrm{X}^T \mathrm{X}.$$

Suppose now we have classes $\mathcal{C}_1, \ldots, \mathcal{C}_N$, where $\mathcal{C}_i$ has n_i data points with centroid μ_i in the feature space, and let μ be the center of all data points. The **Fisher-LDA** optimization problem is

$$\max_{\mathrm{w}} \frac{\mathrm{w}^T \mathrm{S}_b \mathrm{w}}{\mathrm{w}^T \mathrm{S}_w \mathrm{w}},$$

where S_b and S_w are between- and within-class *scatter matrices*

$$\mathrm{S}_b = \sum_{i=1}^{N} n_i (\mu_i - \mu)(\mu_i - \mu)^T \text{ and } \mathrm{S}_w = \sum_{i=1}^{N} \sum_{\mathrm{x} \in \mathcal{C}_i} (\mathrm{x} - \mu_i)(\mathrm{x} - \mu_i)^T.$$

Its solutions can be shown to be the eigenvectors of $\mathrm{S}_w^{-1} \mathrm{S}_b$, in descending order of the corresponding eigenvalues.