COS 424: Interacting with Data

Lecturer: Dave Blei
Scribe: CJ Bell and Ana Pop

Lecture #22
April 24, 2008

1 Principal Component Analysis (PCA)

PCA is one method used to reduce the number of features used to represent data. The benefits of this *dimensionality reduction* include providing a simpler representation of the data, reduction in memory, and faster classification. We accomplish by projecting data from a higher dimension to a lower dimensional manifold such that the error incurred by reconstructing the data in the higher dimension is minimized.

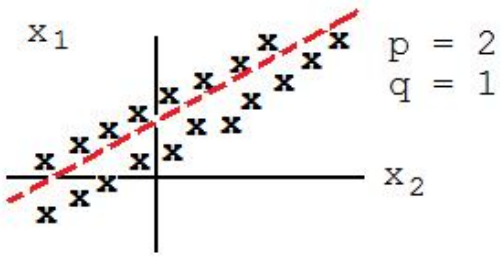

Figure 1: A plot of x's in 2D ($\mathbb{R}^p$) space and an example 1D ($\mathbb{R}^q$) space (dashed line) to which the data can be projected.

An example of this is given by Figure 1, where 2D data can be projected to the 1D space represented by the dashed line with reasonably small error. In general, we want to map $x \in \mathbb{R}^p$ to $\tilde{x} \in \mathbb{R}^q$ where $q < p$.

1.1 Idea Behind PCA

- Draw some lower dimensional space. In Figure 1, this is the dashed line.
- Represent each data point by its projection along the line.

In Figure 1, the free parameter is the slope. We draw the line to minimize the distances to the points. Note that in regression, the distance to the line is vertical, not perpendicular, as shown by Figure 2.

1.2 PCA Interpretation

PCA can be interpreted in three different ways.

- Maximize the variance of projection along each component.
- Minimize the reconstruction error (ie. the squared distance between the original data and its "estimate").
- Some MLE of a parameter in a probabilistic model.

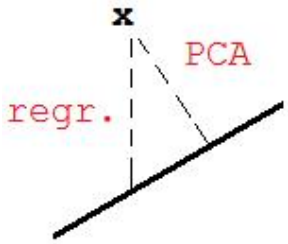

Figure 2: Projecting x to $\mathbb{R}^1$. The vertical line is the regression mapping and the perpendicular line is the PCA projection.

1.3 PCA Details

Given data points $x_1, x_2, ..., x_n \in \mathbb{R}^p$.

We define the reconstruction of data in $\mathbb{R}^q$ to $\mathbb{R}^p$ as

$$f(\lambda) = \mu + v_q \lambda \tag{1}$$

In this rank q model, the mean is $\mu \in \mathbb{R}^p$ and v_q is a $p \times q$ matrix with q orthogonal unit vectors. Finally, $\lambda \in \mathbb{R}^q$ is the low-dimensional data points we are projecting.

Creating a good low-dimensional representation of the data requires that we carefully choose μ, v_q, and λ. One way we can do this is by minimizing the reconstruction error given by

$$\min_{\mu, \lambda_{1...N}, v_q} \sum_{n=1}^{N} ||x_n - \mu - v_q \lambda_n|| \tag{2}$$

In Equation 2, μ is the intercept of the lower space in the higher space. Next, $\lambda_{1...N}$ is the $\mathbb{R}^q$ coordinate of x, or where x lies on the line in Figure 1. We define the $\mathbb{R}^p$ plane using v_q and μ. Last, the quantity inside the sum is the distance between the original data and the low-dimensional representation reconstruction in the original space (the L_2 distance between the original data and the projection).

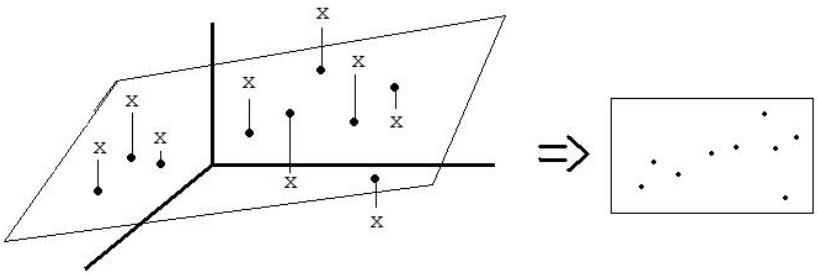

Figure 3: Projecting $\mathbb{R}^3$ data to $\mathbb{R}^2$

We next present an example where the number three is recognized from handwritten text, as shown in Figure 4. Each image is a datapoint in $\mathbb{R}^{256}$ where a pixel is a dimension that varies between white and black. When reducing to two dimensions, the principal components are λ_1 and λ_2. We can reconstruct a $\mathbb{R}^{256}$ datapoint from a $\mathbb{R}^2$ point using

$$\hat{f}(\lambda) = \text{  } + \lambda_1 \cdot \text{  } + \lambda_2 \cdot \text{  } \tag{3}$$

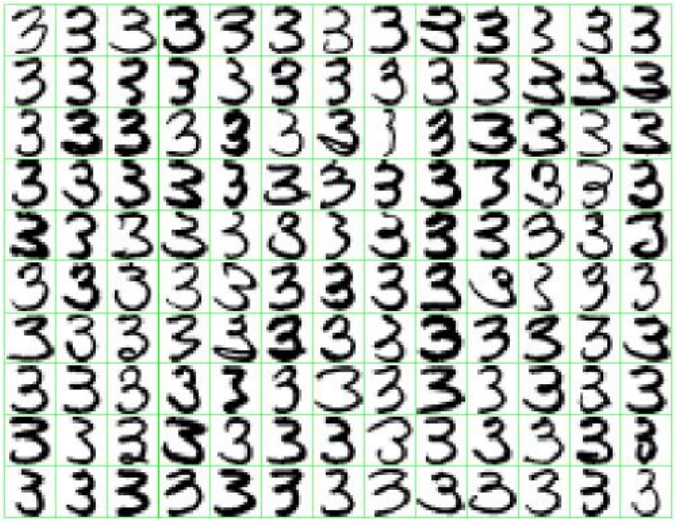

Figure 4: 130 samples of handwritten threes in a variety of writing styles.

Instead of minimizing the reconstruction error, however, we maximize the variance with the objective function

$$\min_{v_q} \sum_{n=1}^{N} ||x_n - v_q v_q^T x_n||^2 \tag{4}$$

From Equation 2, fitting a PCA (Equation 4) is the same as minimizing the reconstruction error. The optimal intercept is the sample mean $\mu^* = \overline{x}$. Without loss of generality, assume $\mu^* = 0$ and $x = x - \mu^*$. The projection v_q on x_n is $\lambda_n = v_q^T x_n$. Now we find the principle components v_q. These are the places where to put the data to reconstruct with minimum error. We get the solution to v_q using singular value decomposition (SVD).

1.4 SVD

Consider

$$X = UDV^T \tag{5}$$

where

- X is an $N \times p$ matrix.
- U is an $N \times p$ orthogonal matrix and the columns of U are linearly independent.
- D is a positive $p \times p$ diagonal matrix with $d_{11} \geq d_{22} \geq ... \geq d_{pp}$.
- V^T is a $p \times p$ orthogonal matrix.

We represent each data point as linear combinations.

$$\begin{aligned} x_1 &= u_{11} d_1 \bar{v_1} + u_{12} d_2 \bar{v_2} + ... + u_{1p} d_p \bar{v_p} \\ x_2 &= u_{21} d_2 \bar{v_1} + u_{22} d_2 \bar{v_2} + ... + u_{1p} d_p \bar{v_p} \\ &... \end{aligned}$$

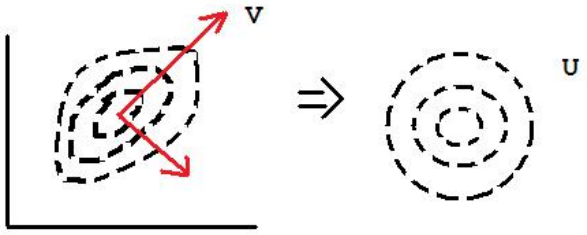

Figure 5:

We can embed x into an orthogonal space via rotation. D scales, V rotates, and U is a perfect circle.

PCA cuts off SVD at q dimensions. In Figure 6, U is a low dimensional representation. Examples 3 and 1.3 use $q = 2$ and $N = 130$. D reflects the variance so we cut off dimensions with low variance (remember $d_{11} \leq d_{22}$...). Lastly, V are the principle components.

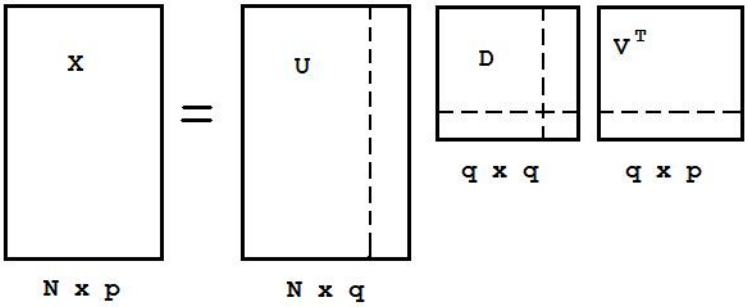

Figure 6:

2 Factor Analysis

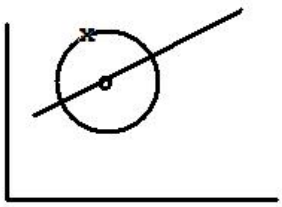

Figure 7: The hidden variable is the point on the hyperplane (line). The observed value is x, which is dependant on the hidden variable.

Factor analysis is another dimension-reduction technique. The low-dimension representation of higher-dimensional space is a hyperplane drawn through the high dimensional space. For each datapoint, we select a point on the hyperplane and choose data from the Gaussian around that point. These chosen points are observable whereas the point on the hyperplane is latent.

2.1 Multivariate Gaussian

This is a Gaussian for p-vectors characterized by

- mean μ, a p-vector
- covariance matrix $\sum$, a $p \times p$ positive-definite, and symmetric

$$\sigma_{ij} = E[x_i x_j] - E[x_i]E[x_j] \tag{6}$$

Some observations:

- A data point is $x :< x_1 ... x_p >$ vector which is also a random variable.
- If x_i, x_j are independent, $\sigma_{ij} = 0$
- σ_{ij} is the covariance between components i and j.
- $\sigma_{ii} = E[x_i^2] - E[x_i]^2 = var(x_i)$

The density function is over vectors of length p.

$$p(x|\mu_i\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma^{1/2}|} exp\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \tag{7}$$

Note that $|\Sigma| = det(\Sigma)$ and that $(x - \mu)$ is a p-vector.

We now define contours of constant probability density as $f(x) = \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$. These are points where the multivariate Gaussian is the same. They are points on an ellipse.

2.2 MLE

The optimal sample mean, $\hat{\mu}$, is a p-vector and $\hat{\Sigma}$ is how often two components are large together or small together for positive covariances.

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{8}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})(x_n - \hat{\mu})^T \tag{9}$$

2.3 Factor Analysis

The parameters are Λ, a q dimensional subspace in p space and a $q \times q$ matrix, and Ψ, a diagonal and positive $p \times p$ matrix.

For each data point,

$Z_n \sim N_q(\vec{0}, I)$ means it has mean of 0 and each component is an independent Gaussian.

$x_n \sim N_p(\Lambda z, \Psi)$ means it has mean of Λz and diagonal covariance matrix Ψ.

In PCA, $x = z_1\overline{\lambda_1} + z_2\overline{\lambda_2} + ... + z_q\overline{\lambda_q}$

In FA, $x \tilde{} N(z_1\overline{\lambda_1} + ... + z_q\overline{\lambda_q}), \Psi$

Fit FA with EM.