MATH 567: Mathematical Techniques in Data Science
Principal component analysis

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

April 3, 2017

- High-dimensional data often has a low-rank structure.
- Most of the "action" may occur in a subspace of $\mathbb{R}^p$.
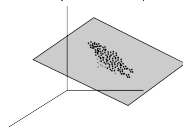
Problem: How can we discover low dimensional structures in data?

- Principal components analysis: construct projections of the data that capture most of the *variability* in the data.
- Provides a low-rank approximation to the data.
- Can lead to a significant dimensionality reduction.

Principal component analysis (PCA)

- Let $X \in \mathbb{R}^{n \times p}$ with rows $x_1, \ldots, x_n \in \mathbb{R}^p$. We think of X as n observations of a random vector $(X_1, \ldots, X_p) \in \mathbb{R}^p$.
- Suppose each column has mean 0, i.e., $\sum_{i=1}^n x_i = \mathbf{0}_{1 \times p}$.
- We want to find a linear combination $w_1 X_1 + \cdots + w_p X_p$ with maximum variance. (Intuition: we look for a direction in $\mathbb{R}^p$ where the data varies the most.)

We solve:
$$w = \operatorname*{argmax}_{\|w\|_2=1} \sum_{i=1}^n (x_i^T w)^2.$$

(Note: $\sum_{i=1}^n (x_i^T w)^2$ is proportional to the sample variance of the data since we assume each column of X has mean 0.)

Equivalently, we solve:
$$w = \operatorname*{argmax}_{\|w\|_2=1} (Xw)^T (Xw) = \operatorname*{argmax}_{\|w\|_2=1} w^T X^T X w$$

Claim: w is an eigenvector associated to the largest eigenvalue of $X^T X$.

Proof of claim: Rayleigh quotients

Let $A \in \mathbb{R}^{p \times p}$ be a symmetric (or Hermitian) matrix. The *Rayleigh quotient* is defined by
$$R(A, x) = \frac{x^T A x}{x^T x} = \frac{\langle Ax, x \rangle}{\langle x, x \rangle}, \qquad (x \in \mathbb{R}^p, x \neq \mathbf{0}_{p \times 1}).$$

Observations:

1. If $Ax = \lambda x$ with $\|x\|_2 = 1$, then $R(A, x) = \lambda$. Thus,
$$\sup_{x \neq 0} R(A, x) \geq \lambda_{\max}(A).$$

2. Let $\{\lambda_1, \ldots, \lambda_p\}$ denote the eigenvalues of A, and let $\{v_1, \ldots, v_p\} \subset \mathbb{R}^p$ be an orthonormal basis of eigenvectors of A. If $x = \sum_{i=1}^p \theta_i v_i$, then $R(A, x) = \frac{\sum_{i=1}^p \lambda_i \theta_i^2}{\sum_{i=1}^p \theta_i^2}$. It follows that
$$\sup_{x \neq 0} R(A, x) \leq \lambda_{\max}(A).$$

Thus, $\sup_{x \neq 0} R(A, x) = \sup_{\|x\|_2=1} x^T A x = \lambda_{\max}(A)$.

Previous argument shows that

$$w^{(1)} = \operatorname*{argmax}_{\|w\|_2=1} \sum_{i=1}^{n} (x_i^T w)^2 = \operatorname*{argmax}_{\|w\|_2=1} w^T X^T X w$$

is an eigenvector associated to the largest eigenvalue of $X^T X$.

First principal component:
- The linear combination $\sum_{i=1}^{p} w_i^{(1)} X_i$ is the *first principal component* of $(X_1, \ldots, X_p)$.
- Alternatively, we say that $X w^{(1)}$ is the first (sample) principal component of X.
- It is the linear combination of the columns of X having the "most variance".

Second principal component: We look for a new linear combination of the X_i's that
1. Is orthogonal to the first principal component, and
2. Maximizes the variance.

In other words:

$$w^{(2)} := \operatorname*{argmax}_{\substack{\|w\|_2=1 \\ w \perp w^{(1)}}} \sum_{i=1}^{n} (x_i^T w)^2 = \operatorname*{argmax}_{\substack{\|w\|_2=1 \\ w \perp w^{(1)}}} w^T X^T X w.$$

- Using a similar argument as before with Rayleigh quotients, we conclude that $w^{(2)}$ is an eigenvector associated to the second largest eigenvalue of $X^T X$.
- Similarly, given $w^{(1)}, \ldots, w^{(k)}$, we define

$$w^{(k+1)} := \operatorname*{argmax}_{\substack{\|w\|_2=1 \\ w \perp w^{(1)}, w^{(2)}, \ldots, w^{(k)}}} \sum_{i=1}^{n} (x_i^T w)^2 = \operatorname*{argmax}_{\substack{\|w\|_2=1 \\ w \perp w^{(1)}, w^{(2)}, \ldots, w^{(k)}}} w^T X^T X w.$$

As before, the vector $w^{(k+1)}$ is an eigenvector associated to the $(k+1)$-th largest eigenvalue of $X^T X$.

In summary, suppose

$$X^T X = U \Lambda U^T$$

where $U \in \mathbb{R}^{p \times p}$ is an orthogonal matrix and $\Lambda \in \mathbb{R}^{p \times p}$ is diagonal. (Eigendecomposition of $X^T X$.)
- Recall that the columns of U are the eigenvectors of $X^T X$ and the diagonal of Λ contains the eigenvalues of $X^T X$ (i.e., the (square of the) singular values of X).
- Then the *principal components* of X are the columns of XU.
- Write $U = (u_1, \ldots, u_p)$. Then the variance of the i-th principal component is

$$(X u_i)^T (X u_i) = u_i^T X^T X u_i = (U^T X^T X U)_{ii} = \Lambda_{ii}.$$

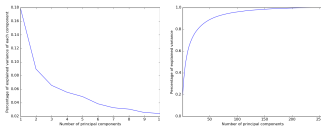**Conclusion:** The variance of the i-th principal component is the i-th eigenvalue of $X^T X$.
- We say that the first k PCs *explain* $(\sum_{i=1}^{k} \Lambda_{ii})/(\sum_{i=1}^{p} \Lambda_{ii}) \times 100$ percent of the variance.

Recall the zip dataset:
1. 9298 images of digits $0 - 9$.
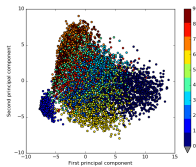2. Each image is in black/white with $16 \times 16 = 256$ pixels.

We use PCA to project the data onto a 2-dim subspace of $\mathbb{R}^{256}$.

Projecting the data on the first two principal components:

- Note: $\approx 27\%$ variance explained by the first two PCAs.
- $\approx 90\%$ variance explained by first 55 components.

- PCAs can be directly used in a regression context.
- Principal component regression: $y \in \mathbb{R}^{n \times 1}$, $X \in \mathbb{R}^{n \times p}$.
  1. Center y and each column of X (i.e., subtract mean from the columns)
  2. Compute the eigen-decomposition of $X^T X$:
  $$X^T X = U \Lambda U^T.$$
  3. Compute $k \geq 1$ principal components:
  $$W_k := (Xu_1, \ldots, Xu_k) = X U_k,$$
  where $U = (u_1, \ldots, u_p)$, and $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{p \times k}$.
  4. Regress y on the principal components:
  $$\hat{\gamma}_k := (W_k^T W_k)^{-1} W_k^T y.$$
  5. The PCR estimator is:
  $$\hat{\beta}_k := U_k \hat{\gamma}_k, \qquad \hat{y}^{(k)} := X \hat{\beta}_k = X U_k \hat{\gamma}_k.$$

Note: k is a parameter that needs to be chosen (using CV or another method). Typically, one picks k to be significantly smaller than p.

- PCA looks for subspaces with the most variance.
- Can also optimize other criteria.

Projection pursuit (PP):
  1. Set up a projection "index" to judge the merit of a particular one or two-dimensional projection of a given set of multivariate data.
  2. Use an optimization algorithm to find the global and local extrema of that projection index over all 1/2-dimensional projections of the data.

Example: (Izenman, 2013) The absolute value of kurtosis, $|\kappa_4(Y)|$, of the one-dimensional projection $Y = w^T X$ has been widely used as a measure of non-Gaussianity of Y.

- Recall: The marginals of the multivariate Gaussian distribution are Gaussian.
- Can maximize/minimize the kurtosis to find subspaces where data looks Gaussian/non-Gaussian (e.g. to detect outliers).