Principal Component Analysis

Correlation matrix decomposition

Consider a zero mean random vector \( x \in \mathbb{R}^n \) with autocorrelation matrix \( R = E(xx^T) \).

R has eigenvectors \( q(1), \ldots, q(n) \) and associated eigenvalues \( \lambda(1) \geq \ldots \geq \lambda(n) \).

Let \( Q = \begin{bmatrix} q(1) & \ldots & q(n) \end{bmatrix} \) and \( \Lambda \) be a diagonal matrix containing eigenvalues along diagonal.

Then \( R = Q \Lambda Q^T \) can be decomposed into eigenvector and eigenvalue decomposition.
PCA works with data

Given \((x(1), x(2), \ldots, x(m))\) find weight direction that gives most information about data.
Let \( X = [x(1) \mid \ldots \mid x(m)] \) and \( R = (1/m)XX^T \) (sample correlation matrix).

Problem: \( \max w^T R w \) subject to \( \|w\| = 1 \).

Maximum obtained when \( w = q(1) \) as this corresponds to \( w^T R w = \lambda(1) \).

\( q(1) \) is first principal component of \( x \) and also yields direction of maximum variance.

\( y(1) = q(1)^T x \) is projection of \( x \) onto first principal component.
Other Principal Components

ith principal component denoted by \( q(i) \) and projection denoted by \( y(i) = q(i)^T x \) with \( \mathbb{E}(y(i)) = 0 \) and \( \mathbb{E}(y(i)^2) = \lambda(i) \). \( y(i) \)

Note that \( y = Q^T x \) and we can obtain data vector \( x \) from \( y \) by noting that \( x = Qy \).

We can approximate \( x \) by taking first \( m \) principal components (PC) to get \( z: z = q(1)x(1) + \ldots + q(m)x(m) \). Error given by \( e = x-z \). \( e \) is orthogonal to \( q(i) \) when \( 1 \leq i \leq m \).

All PCA give eigenvalue / eigenvector decomposition of \( R \) and is also known as the Discrete Karhunen Loeve Transform
First PC gives more information than second PC.
Learning Principal Components

- Given m inputs \((x(1), x(2), \ldots, x(m))\) how can we find the Principal Components?
- Batch learning: Find sample correlation matrix \(1/m X^TX\) and then find eigenvalue and eigenvector decomposition. Decomposition can be found using SVD methods.
- On-line learning: Oja’s rule learns first PCA. Generalized Hebbian Algorithm, APEX.
Convergence of Oja’s Learning rule

- Stochastic approximation algorithm that converges to with probability 1 under set of assumptions including decreasing step size
- Mean convergence

\[ w(k+1) = w(k) + \mu (x(k) x(k)^T w(k) - w^T(k)x(k)x(k)^T (w(k)w(k))) \]

- Take expectations of both sides and let \( k \) grow large to get that \( Rw = (w^T R w)w \) where \( w = \lim_{k \to \infty} w(k) \)
- Can then show that \( w = q(i) \) where \( q(i) \) is an eigenvector
- Perturb weight and show that \( q(i) \) is first eigenvector
PCA and LS-SVM formulation

Problem: $\max w^TXX^Tw$ subject to $w^Tw=1$.
Reformulated in terms of SV QP methods:

$$\max J(w,e) = \frac{\gamma}{2} e^Te - \frac{1}{2} w^Tw$$
Subject to $e = X^Tw$

Lagrangian becomes

$$\max L(w,e,\alpha) = \frac{\gamma}{2} e^Te - \frac{1}{2} w^Tw - \alpha^T(e - X^Tw)$$

Take derivatives wrt each variable and set to zero to get

that $w = X\alpha$, $\alpha = \gamma e$, and $e - X^Tw = 0$.

Eliminate $e$ and $w$ to get that $\alpha / \gamma - X^TX \alpha = 0$
Dual Space Formulation and Comments

- Let $K = X^T X$ and $\lambda = 1/\gamma$, then we have an eigenvalue/eigenvector problem in dual space, $K \alpha = \lambda \alpha$.
- Want to maximize $e^T e = \alpha^T \alpha / \gamma^2 = \lambda_{\text{max}}$ when eigenvectors are normalized to have magnitude 1.
- Problem can easily be modified if data does not have zero mean and also to include bias term.
- Convergence to first eigenvector of ensemble correlation matrix
- Fisher Discriminant Analysis (FDA) (similar to PCA except FDA has targets to minimize scatter around targets)
Kernel Methods

In many classification and detection problems a linear classifier is not sufficient. However, working in higher dimensions can lead to “curse of dimensionality”.

Solution: Use kernel methods where computations done in dual observation space.

\[ \Phi: X \rightarrow Z \]
Kernel PCA

- Obtain nonlinear features from data
- Can form kernel PCA in primal space (R) or dual space (K).
- Problem closely related to LS SVM
- Must ensure feature data has zero mean
- Applications: Preprocessing data, denoising, compression, image interpretation
KPCA Formulation

- Kernel PCA uses kernels to max \( \mathbb{E}(0-w^T (\phi(x) - m\phi))^2 \).
- Use input data to approximate ensemble average to get the following quantities, \( \Phi(x) = (\phi(x(1), \ldots, \phi(x(m)))^T \), \( R \) is sample covariance matrix, and kernel matrix is
  \[
  K = (\Phi(x) - 1/m 11^T \Phi(x)) (\Phi(x) - 1/m 11^T \Phi(x))^T
  \]
- We can formulate as a QP problem where we
  \[
  \max \frac{1}{2} w^T R w
  \]
  subject to \( w^T w = 1 \) and \( w = (\Phi(x) - 1/m 11^T \Phi(x))^T \alpha \)
- Can solve in primal or dual spaces. In dual space we have another eigenvector /eigenvalue problem \( K \alpha = \lambda \alpha \).
Canonical Correlation Analysis

- Given two pairs of data, can we extract features from different data.
- Problem: Given \( x \in \mathbb{R}^{n_1} \) and \( y \in \mathbb{R}^{n_2} \) are zero mean vectors, find \( z_x = w^T x \) and \( z_y = v^T y \) so that correlation defined by \( \rho(z_x, z_y) = \frac{E(z_x z_y)}{(E(z_x, z_x)E(z_y, z_y))^{1/2}} \) is maximized.
- QP formulation: \( \max J(w, v) = w^T C_{xy} v \) subject to \( w^T C_{xx} w = 1 \) and \( v^T C_{yy} v = 1 \) where \( C_{xx} = E(x x^T) \), \( C_{yy} = E(y y^T) \), and \( C_{xy} = E(x y^T) \).
- CCA can be formulated using dual space and kernel CCA can also be constructed.