Principal Component Analysis

Introduction

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

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

Let $Q = [q(1) \mid \ldots \mid q(n)]$ and $\Lambda$ be a diagonal matrix containing eigenvalues along diagonal.

Then $R = Q \Lambda Q^T$ can be decomposed into eigenvector and eigenvalue decomposition.
First Principal Component

Find 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 \( E(y(i)) = 0 \) and \( E(y(i)^2) = \lambda(i) \).

Note that \( y = Q^T x \) and we can obtain data vector \( x \) from \( y \) by noting that \( x = Q y \). Also have \( E(y(i)y(j)) = 0 \) for \( i \neq j \).

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 \(1/m \mathbf{X}^T \mathbf{X}\) 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.
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 $\alpha/\gamma - X^TXX\alpha = 0$
Let $K = X^TX$ and $\lambda = 1/\gamma$, then we have an eigenvalue/eigenvector problem in dual space, $K\alpha = \lambda\alpha$.

Want to maximize $e^Te = \alpha^T\alpha/\gamma^2 = \lambda_{\text{max}}$ 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 $E(0-w^T (\phi(x) - m_\phi))^2$.
- 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 $\rho(z_x, z_y) = \frac{E(z_x z_y)}{\sqrt{E(z_x z_x) E(z_y z_y)}}$ 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.
Projection of Data

- Principal Component Analysis (PCA) (maximize variance)
- Fisher Discriminant Analysis (FDA) (minimize in class scatter while maximizing distance between means)
- Canonical Correlation Analysis (CCA) (maximize covariance)
- Kernelize PCA, FDA, CCA
- Factor analysis, partial regression analysis
- Projection pursuit
- Independent Component Analysis
Given a signal S1 mixed with additive impairment S2 we can use an adaptive filter (using algorithms such as LMS) to cancel out the effects of the noise.
Modification to Noise Cancelling

Give X1 and X2 with Adaptive Filter 1 having access to S2 and Adaptive Filter 2 having access to S1 we can recover S1 and S2.
Independent Component Analysis

- Let $X = AU$ where $A$ is a square mixing matrix, $U$ is a random $m$ vector, and $X$ is the observed random $m$ vector.
- Can we recover $U$ from $X$ if $A$ and $U$ are unknown?
- Under assumptions that components of $U$ are independent random variables we can recover $U$ from $X$ under certain assumptions. We need to establish optimization criteria to recover $U$ from $X$.
- Closely related to projection pursuit and factor analysis
- Applications: Blind Source Separation, Blind Deconvolution, Feature Extraction
Independent Component Analysis

PCA decorrelates inputs. However in many instances we may want to make outputs independent.

Inputs U assumed independent and user sees X. Goal is to find W so that Y is independent.
Applications of ICA

- Speech Separation: several speech signals are mixed together (cocktail problem)
- Array antenna processing: several narrowband signals mixed together from unknown directions
- Hyperspectral Images: images at multiple wavelengths
- Biomedical information: Brain signals, EEG data, FMRI data
- Financial market data analysis: extract dominant signals
ICA Solution

- $Y = DPU$ where $D$ is a diagonal matrix and $P$ is a permutation matrix.
- Algorithm is unsupervised. What are assumptions where learning is possible? All components of $U$ except possibly one are nongaussian.
- Establish criterion to learn from (use higher order statistics): information based criteria, kurtosis function.
- Kullback–Leibler Divergence:

$$D(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x)} \right) \, dx$$
ICA Information Criterion

- Kullback Leibler Divergence nonnegative
- Mutual Information $I(X;Y) = H(X) - H(X|Y)$ nonnegative
- Set $f$ to joint density of $Y$ and $g$ to products of marginals of $Y$ then

$$D(f,g) = -H(Y) + \sum H(Y_i)$$

which is minimized when components of $Y$ are independent.

- When outputs are independent they will be a permutation and scaled version of $U$. 
ICA Objective Functions

- Kurtosis
- Kullback Leibler Divergence nonnegative
- Mutual Information $I(X:Y) = H(X) - H(X|Y)$ nonnegative. Set $f$ to joint density of $Y$ and $g$ to products of marginals of $Y$ then
  \[ D(f, g) = -H(Y) + \sum H(Y_i) \]
  which is minimized when components of $Y$ are independent.
- Negenentropy
- Contrast functions
ICA Preprocessing

- Signal processing and filtering
- Center data (remove means)
- Decorrelate data (apply PCA). If data is jointly Gaussian cannot do any more
Learning Algorithms

- Can learn weights by approximating divergence cost function established using contrast functions.
- Iterative gradient estimate algorithms can be used.
- Faster convergence can be achieved with fixed point algorithms that approximate Newton’s methods.
- Algorithms have been shown to converge.
ICA Example

- Three signals are linearly mixed

FIGURE 10.13 Waveforms on left-hand side: original source signals. Waveforms on right-hand side: separated source signals.