Principal Component Analysis & Independent Component Analysis

Overview

Principal Component Analysis
• Purpose
• Derivation
• Examples

Independent Component Analysis
• Generative Model
• Purpose
• Restrictions
• Computation
• Example

Literature
Notation & Basics

\( \mathbf{u} \) \hspace{1cm} \text{Vector}

\( \mathbf{U} \) \hspace{1cm} \text{Matrix}

\( \mathbf{u}^T \mathbf{v}, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^N \) \hspace{1cm} \text{Dot product written as matrix product}

\( \mathbf{u}a \) \hspace{1cm} \text{Product of a row vector with scalar as matrix product, and not } a\mathbf{u}

\( \mathbf{u}^2 = \|\mathbf{u}\|^2 = \mathbf{u}^T \mathbf{u} \) \hspace{1cm} \text{squared norm}

Rules for matrix multiplication:

\( \mathbf{U} \mathbf{V} \mathbf{W} = (\mathbf{U} \mathbf{V}) \mathbf{W} = \mathbf{U}(\mathbf{V} \mathbf{W}) \)

\( (\mathbf{U} + \mathbf{V}) \mathbf{W} = \mathbf{U} \mathbf{W} + \mathbf{V} \mathbf{W} \)

\( (\mathbf{U} \mathbf{V})^T = \mathbf{V}^T \mathbf{U}^T \)
Principal Component Analysis (PCA)

**Purpose**

For a set of samples of a random vector $\mathbf{x} \in \mathbb{R}^N$, discover or reduce the dimensionality and identify meaningful variables.

$$ y = \mathbf{Ux}, \quad \mathbf{U} : p \times N, \quad p < N $$
Principal Component Analysis (PCA)

PCA by Variance Maximization

Find the vector $\mathbf{u}_1$, such that the variance of the data along this direction is maximized:

$$\sigma^2_{\mathbf{u}_1} = E\left\{ (\mathbf{u}_1^T \mathbf{x})^2 \right\}, \ E\{\mathbf{x}\} = 0, \|\mathbf{u}_1\| = 1$$

$$\sigma^2_{\mathbf{u}_1} = E\left\{ (\mathbf{u}_1^T \mathbf{x})(\mathbf{x}^T \mathbf{u}_1) \right\} = \mathbf{u}_1^T E\{\mathbf{x}\mathbf{x}^T\} \mathbf{u}_1,$$

$$\sigma^2_{\mathbf{u}_1} = \mathbf{u}_1^T \mathbf{C} \mathbf{u}_1, \quad \mathbf{C} = E\{\mathbf{x}\mathbf{x}^T\}$$

The solution is the eigenvector $\mathbf{e}_1$ of $\mathbf{C}$ with the largest eigenvalue $\lambda_1$.

$$\mathbf{C}\mathbf{e}_1 = \mathbf{e}_1 \lambda_1, \quad \lambda_1 = \mathbf{e}_1^T \mathbf{C} \mathbf{e}_1 \iff \sigma^2_{\mathbf{u}_1} = \lambda_1$$
Principal Component Analysis (PCA)

PCA by Variance Maximization

For a given $p < N$, find $p$ orthonormal basis vectors $\mathbf{u}_i$ such that the variance of the data along these vectors is maximally large, under the constraint of decorrelation:

$$E\left\{ (\mathbf{u}_i^T \mathbf{x})(\mathbf{u}_n^T \mathbf{x}) \right\} = 0, \quad \mathbf{u}_i^T \mathbf{u}_n = 0 \quad n \neq p$$

The solution are the eigenvectors of $\mathbf{C}$ ordered according to decreasing eigenvalues $\lambda$:

$$\mathbf{u}_1 = \mathbf{e}_1, \mathbf{u}_2 = \mathbf{e}_2, \ldots, \mathbf{u}_p = \mathbf{e}_p, \lambda_1 > \lambda_2 \ldots > \lambda_p$$

Proof of decorrelation for eigenvectors:

$$E\left\{ (\mathbf{e}_i^T \mathbf{x})(\mathbf{e}_n^T \mathbf{x}) \right\} = \mathbf{e}_i^T E\left\{ \mathbf{x}\mathbf{x}^T \right\} \mathbf{e}_n = \mathbf{e}_i^T \mathbf{C} \mathbf{e}_n = \mathbf{e}_i^T \mathbf{e}_n \quad \lambda_n = 0 \quad \text{orthogonal}$$
Principal Component Analysis (PCA)

PCA by Mean Square Error Compression

For a given $p < N$, find $p$ orthonormal basis vectors such that the $mse$ between $\mathbf{x}$ and its projection $\hat{\mathbf{x}}$ into the subspace spanned by the $p$ orthonormal basis vectors is minimum:

$$mse = E\left\{ \| \mathbf{x} - \hat{\mathbf{x}} \|^2 \right\}, \quad \hat{\mathbf{x}}_i = \sum_{k=1}^{p} u_k \left( x_k^T u_k \right), \quad u_k^T u_m = \delta_{k,m}$$
Principal Component Analysis (PCA)

\[
mse = E \left\{ \| x - \hat{x} \|^2 \right\} = E \left\{ \left\| x - \sum_{k=1}^{p} u_k \left( x^T u_k \right) \right\|^2 \right\}
\]

\[
= E \left\{ \| x \|^2 \right\} - 2E \left\{ \sum_{k=1}^{p} x^T u_k (x^T u_k) \right\} + E \left\{ \sum_{n=1}^{p} (x^T u_n) u_n^T \sum_{k=1}^{p} u_k (x^T u_k) \right\}
\]

\[
= E \left\{ \| x \|^2 \right\} - 2E \left\{ \sum_{k=1}^{p} (x^T u_k)^2 \right\} + E \left\{ \sum_{k=1}^{p} (x^T u_k)^2 \right\}
\]

\[
= E \left\{ \| x \|^2 \right\} - E \left\{ \sum_{k=1}^{p} (x^T u_k)^2 \right\}
\]

\[
= \text{trace} (C) - \sum_{k=1}^{p} u_k^T C u_k, \quad C = E \left\{ xx^T \right\}
\]
Principal Component Analysis (PCA)

\[
mse = \text{trace}(C) - \sum_{k=1}^{P} u_k^T Cu_k, \quad C = E\{xx^T\}
\]

Solution to minimizing \(mse\) is any (orthonormal) basis of the subspace spanned by the \(p\) first eigenvectors \(e_{1}, ..., e_{p}\) of \(C\).

\[
 mse = \text{trace}(C) - \sum_{k=1}^{P} \lambda_k = \sum_{k=p+1}^{N} \lambda_k
\]

The \(mse\) is the sum of the eigenvalues corresponding to the discarded eigenvectors \(e_{p+1}, ..., e_{N}\) of \(C\):

\[
 mse = \sum_{k=p+1}^{N} \lambda_k
\]
Principal Component Analysis (PCA)

How to determine the number of principal components $p$?
Linear signal model with unknown number $p < N$ of signals:

$$x = As + n, \quad A = \begin{pmatrix} a_{1,1} & a_{1,p} \\ \vdots & \vdots \\ a_{N,1} & a_{N,p} \end{pmatrix} \quad N \times p$$

Signal $s_i$ have 0 mean and are uncorrelated, $n$ is white noise:

$$E\{ss^T\} = I, \quad E\{nn^T\} = \sigma_n^2 I$$

$$C = E\{xx^T\} = E\{As(As)^T\} + E\{nn^T\} + E\{Asn^T\}$$

$$= AA^T + \sigma_n^2 I$$

$$d_1 > d_2 > \ldots > d_p > d_{p+1} = d_{p+2} = \ldots = d_N = \sigma_n^2$$

→ cut off when eigenvalues become constants
Principal Component Analysis (PCA)

Computing the PCA

Given a set of samples \( \{x_1, \ldots, x_M\} \) of a random vector \( x \) calculate mean and covariance.

\[
\tilde{\mu} = \frac{1}{M} \sum_{i=1}^{M} x_i, \quad x \rightarrow x - \tilde{\mu}
\]

\[
\tilde{C} = \frac{1}{M} \sum_{i=1}^{M} (x_i - \tilde{\mu})(x_i - \tilde{\mu})^T
\]

Compute eigenvectors of \( \tilde{C} \) e.g. with QR algorithm
Principal Component Analysis (PCA)

Computing the PCA

If the number of samples $M$ is smaller than the dimensionality $N$ of $x$:

$$
B = \begin{pmatrix}
  x_{1,1} & \cdots & x_{1,M} \\
  \vdots & \ddots & \vdots \\
  x_{N,1} & \cdots & x_{N,M}
\end{pmatrix}, \tilde{C} = BB^T, \ B : N \times M, \ B^T : M \times N
$$

$$
BB^T e = e\lambda
$$

$$
B^T Be' = e'\lambda'
$$

$$
BB^T (Be') = (Be')\lambda' \quad \rightarrow \text{Reducing complexity from $O(N^2)$ to $O(M^2)$}
$$

$e = Be'$, $\lambda' = \lambda$
Principal Component Analysis (PCA)

Examples

Eigenfaces for face recognition (Turk&Pentland):

Training:
- Calculate the eigenspace for all faces in the training database
- Project each face into the eigenspace → feature reduction

Classification:
- Project new face into eigenspace
- Nearest neighbor in the eigenspace
Principal Component Analysis (PCA)

Examples cont.

Feature reduction/extraction

Original

Reconstruction with 20 PC

http://www.nist.gov/
Independent Component Analysis (ICA)

Generative model

Noise free, linear signal model:

\[ x = As = \sum_{i=1}^{N} a_i s_i \]

\[ x = (x_1, \ldots, x_N)^T \quad \text{Observed variables} \]

\[ s = (s_1, \ldots, s_N)^T \quad \text{latent signals, independent components} \]

\[ A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots \cdots \vdots \\ a_{N,1} & \cdots & a_{N,N} \end{pmatrix} \quad \text{unknown mixing matrix, } a_i = (a_{1,i}, \ldots, a_{N,i})^T \]
Independent Component Analysis (ICA)

**Task**
For the linear, noise free signal model, compute $A$ and $s$ given the measurements $x$.

Blind source separation: separate the three original signals $s_1, s_2,$ and $s_3$ from their mixtures $x_1, x_2,$ and $x_3$.

Figure by MIT OCW.
Independent Component Analysis (ICA)

Restrictions

1.) Statistical independence
The signals $s_i$ must be statistically independent:

$$p(s_1, s_2, ..., s_N) = p_1(s_1)p_2(s_2)...p_N(s_N)$$

Independent variables satisfy:

$$E\{g_1(s_1)g_2(s_2)...g_N(s_N)\} = E\{g_1(s_1)\} E\{g_2(s_2)\}...E\{g_N(s_N)\}$$

for any $g_i(s) \in L^1$

$$E\{g_1(s_1)g_2(s_2)\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(s_1)g_2(s_2)p(s_1, s_2)ds_1ds_2$$

$$= \int_{-\infty}^{\infty} g_1(s_1)p(s_1)ds_1 \int_{-\infty}^{\infty} g_2(s_2)p(s_2)ds_2 = E\{g_1(s_1)\} E\{g_2(s_2)\}$$
**Independent Component Analysis (ICA)**

**Restrictions**

Statistical independence cont.

\[
E \left\{ g_1(s_1)g_2(s_2)\ldots g_N(s_N) \right\} = E \left\{ g_1(s_1) \right\} E \left\{ g_2(s_2) \right\} \ldots E \left\{ g_N(s_N) \right\}
\]

Independence includes uncorrelatedness:

\[
E \left\{ (s_i - \mu_i)(s_n - \mu_n) \right\} = E \left\{ (s_i - \mu_i) \right\} E \left\{ (s_n - \mu_n) \right\} = 0
\]
2.) Nongaussian components
The components $s_i$ must have a nongaussian distribution otherwise there is no unique solution.

Example:

given $A$ and two gaussian signals:

$$ p(s_1, s_2) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(- \frac{s_1^2}{2\sigma_1^2} - \frac{s_2^2}{2\sigma_2^2}\right) $$

generate new signals

$$ s' = \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{pmatrix} s \Rightarrow p(s'_1, s'_2) = \frac{1}{2\pi} \exp\left(- \frac{s'_1^2}{2}\right) \exp\left(- \frac{s'_2^2}{2}\right) $$

Scaling matrix $S$
Independent Component Analysis (ICA)

Restrictions

Nongaussian components cont.

under rotation the components remain independent:

\[ s'' = \begin{pmatrix} \cos \varphi & \sin \varphi \\ -\sin \varphi & \cos \varphi \end{pmatrix} s', \quad p(s_1'', s_2'') = \frac{1}{2\pi} \exp\left(-\frac{s_1''^2}{2}\right) \exp\left(-\frac{s_2''^2}{2}\right) \]

combine whitening and rotation \( B = RS \):

\[ x = As = AB^{-1}s'' \]

\( AB^{-1} \) is also a solution to the ICA problem.
Independent Component Analysis (ICA)

Restrictions

3.) Mixing matrix must be invertible
The number of independent components is equal to
the number of observed variables.
Which means that there are no redundant mixtures.

In case mixing matrix is not invertible apply PCA
on measurements first to remove redundancy.
Independent Component Analysis (ICA)

Ambiguities

1.) Scale

\[ x = \sum_{i=1}^{N} a_i s_i = \sum_{i=1}^{N} \left( \frac{1}{\alpha_i} a_i \right) (\alpha_i s_i) \]

Reduce ambiguity by enforcing \( E\{s_i^2\} = 1 \)

2.) Order

We cannot determine an order of the independant components
Independent Component Analysis (ICA)

Computing ICA

a) Minimizing mutual information:
\[ \hat{s} = \hat{A}^{-1}x \]

Mutual information: 
\[ I(\hat{s}) = \sum_{i=1}^{N} H(\hat{s}_i) - H(\hat{s}) \]

\( H \) is the differential entropy:
\[ H(\hat{s}) = -\int p(\hat{s}) \log_2 (p(\hat{s})) \, d\hat{s} \]

\( I \) is always nonnegative and 0 only if the \( \hat{s}_i \) are independent.

Iteratively modify \( \hat{A}^{-1} \) such that \( I(\hat{s}) \) is minimized.
b) Maximizing Nongaussianity

\[ s = A^{-1}x \]

introduce \( y \) and \( b: y = b^T x = b^T As = q^T s \)

From central limit theorem:

\( y = q^T s \) is more gaussian than any of the \( s_i \) and becomes least gaussian if \( y = s_i \).

Iteratively modify \( b^T \) such that the "gaussianity" of \( y \) is minimized. When a local minimum is reached, \( b^T \) is a row vector of \( A^{-1} \).
PCA Applied to Faces

Each pixel is a feature, each face image a point in the feature space. Dimension of feature vector is given by the size of the image.

$\mathbf{x}_{i}$ are the eigenvectors which can be represented as pixel images in the original coordinate system $x_1 \ldots x_N$. 

$\mathbf{x}_1, \ldots, \mathbf{x}_M$
ICA Applied to Faces

Now each image corresponds to a particular observed variable measured over time ($M$ samples). $N$ is the number of images.

$$ x = As = \sum_{i=1}^{N} a_i s_i $$

$$ x = (x_1, \ldots, x_N)^T \quad \text{Observed variables} $$

$$ s = (s_1, \ldots, s_N)^T \quad \text{latent signals, independent components} $$
PCA and ICA for Faces

Features for face recognition

Image removed due to copyright considerations. See Figure 1 in: Baek, Kyungim et. al. "PCA vs. ICA: A comparison on the FERET data set." International Conference of Computer Vision, Pattern Recognition, and Image Processing, in conjunction with the 6th JCIS. Durham, NC, March 8-14 2002, June 2001.