MMAT 5320 Computational Mathematics - Part 2 Applications

Andrew Lam

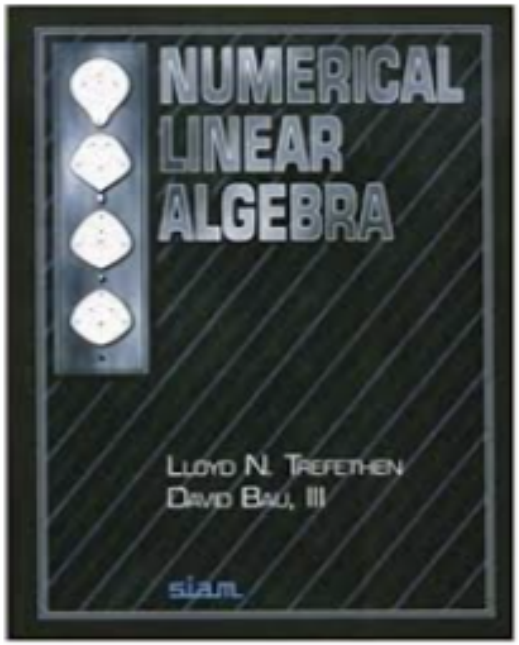
Topics

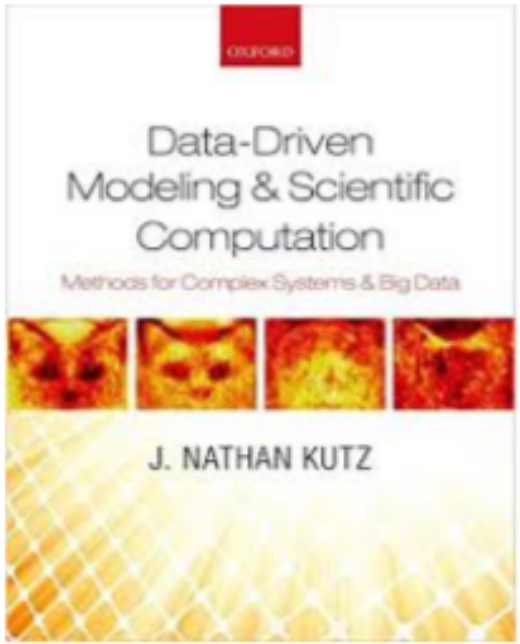
Principal component analysis (PCA)

Independent component analysis (ICA)

Reference books

- Numerical Linear Algebra by Trefethen and Bau (1997)
- Data-Driven Modeling & Scientific Computation by Kutz (2013)





Topics

Topics: Part 1 – Numerical Linear Algebra by Trefethen and Bau (TB)

- Review of Linear algebra
- Singular value decomposition (SVD)
- QR factorization (Gram–Schmidt/Householder)
- Least squares problem
- Eigenvalue problems
- Eigenvalue algorithms

Topics: Part 2 – Data-Driven Modeling & Scientific Computation by Kutz (K)

- Principal component analysis (PCA)
- Independent component analysis (ICA)
- Compress sensing
- Time frequency analysis
- Image denoising and processing
- Data assimilation

§7 - Principal component analysis (PCA)

What is PCA?

Definition from Wikipedia: "PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components."

Unpacking the definition:

- Prerequisite: a collection of observation/data.
- Employ an orthogonal transformation to convert the (correlated) data into a new set of data that is not correlated with each other.
- Statistical procedure - measurement of effectiveness/error involves quantities from statistics.

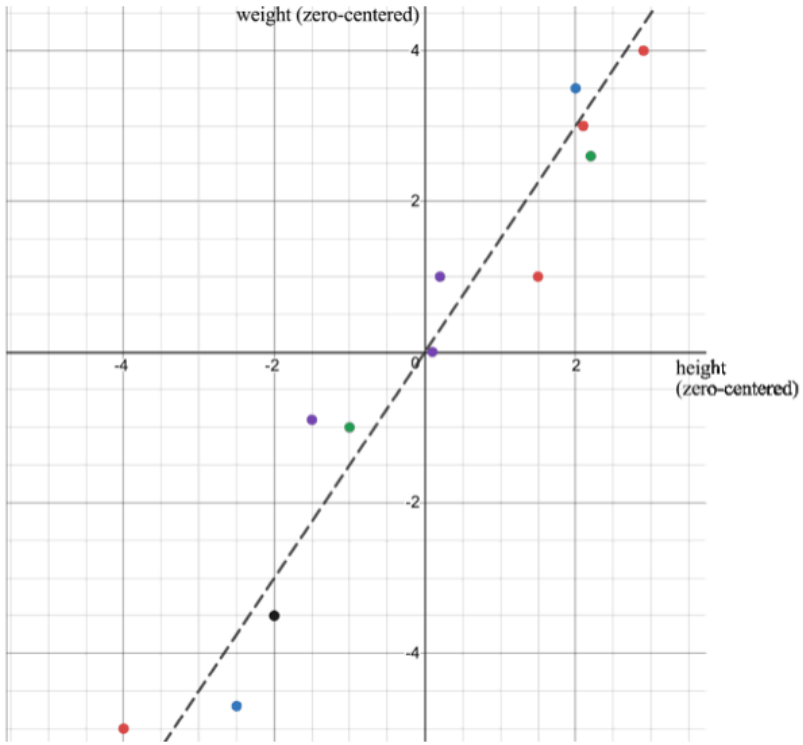
PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it.

PCA motivating example

Consider a sample of heights and weights of 12 people, where we summarised the adjusted mean-zero data in the matrix A:

$$\begin{pmatrix} 2.9 & -1.5 & 0.1 & -1.0 & 2.1 & -4.0 & -2.0 & 2.2 & 0.2 & 2.0 & 1.5 & -2.5 \\ 4.0 & -0.9 & 0.0 & -1.0 & 3.0 & -5.0 & -3.5 & 2.6 & 1.0 & 3.5 & 1.0 & -4.7 \end{pmatrix}$$

The top row is the adjusted height and the bottom row is the adjusted weight.



A plot shows a positive correlation between height and weight. How do we quantify this?

PCA motivating example II

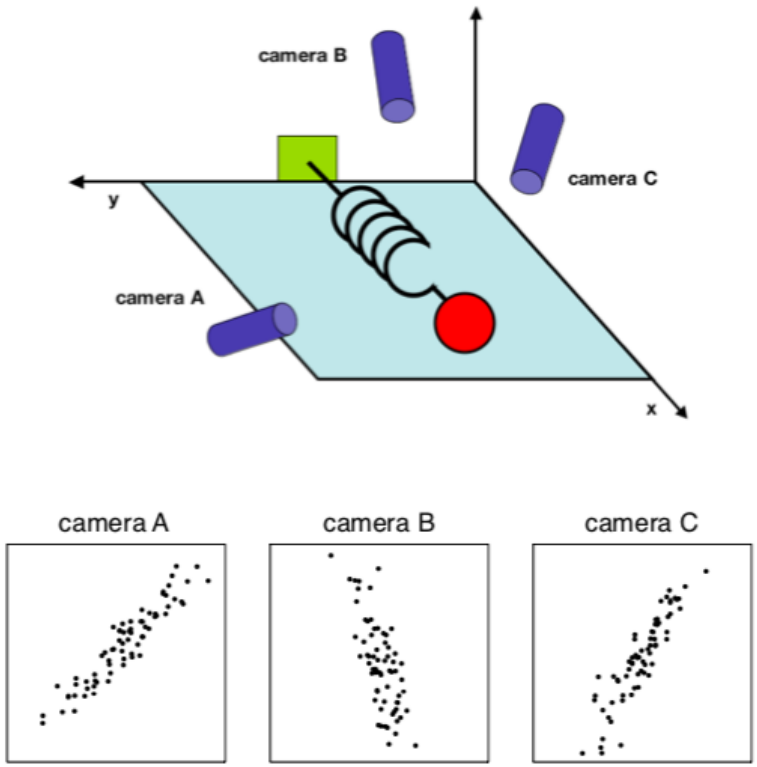


Figure: Taken from J. Shlens. A tutorial on Principal Component Analysis.

Three cameras recording the position of a ball attached to an oscillating spring moving only in the x-axis. Can we reveal this hidden structure from the data obtained by the three cameras?

More general setting

Consider

- m features (e.g. height, weight, number of siblings, etc.), and
- n samples (e.g. number of individuals in the survey).

We collect these in a data matrix $A \in \mathbb{R}^{m \times n}$, $A = (x_{ij})_{1 \le i \le m, 1 \le j \le n}$ where

- row i containing the data for the ith feature,
- column j represents the jth sample of data.

The notation $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{mj})^\top$ denotes the jth column of A.

Some statistical definitions:

1. The feature sample mean vector $\bar{\boldsymbol{x}} \in \mathbb{R}^m$ is the vector whose jth entry is the average value of the n samples of feature j:

$$\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m)^\top, \quad \bar{x}_j = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \in \mathbb{R} \text{ for } j = 1, \ldots, m.$$

2. The sample variance covariance matrix $C = (c_{pq}) \in \mathbb{R}^{m \times m}$ is defined as

$$c_{pq} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{pj} - \bar{x}_p)(x_{qj} - \bar{x}_q).$$

If the data has been preprocessed to have mean zero, i.e., $\bar{x}_i = 0$, we say A has been adjusted, and consequently

$$C = \frac{AA^\top}{n-1}.$$

Properties of the sample covariance matrix

For an adjusted data matrix $A \in \mathbb{R}^{m \times n}$, the sample covariance matrix is

$$C = \frac{AA^\top}{n-1}.$$

Properties:

- $C \in \mathbb{R}^{m \times m}$ is symmetric and positive semi-definite.
- The diagonal entries C_{ii} $(i = 1, \ldots, m)$ represents the sample variance of the ith random variable:

$$\sigma_i^2 = C_{ii} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2 = \frac{1}{n-1} \sum_{j=1}^{n} x_{ij}^2.$$

- The off-diagonal entries C_{ik} represent the sample covariance between the ith and kth random variables.
- If C_{ik} is positive, then we say the ith and kth random variables are positively correlated; if C_{ik} is negative, then they are negatively correlated. If $C_{ik} = 0$, then they are not correlated (hence independent).

Related concept: The correlation matrix $R = (r_{pq}) \in \mathbb{R}^{m \times m}$ is obtained from normalising:

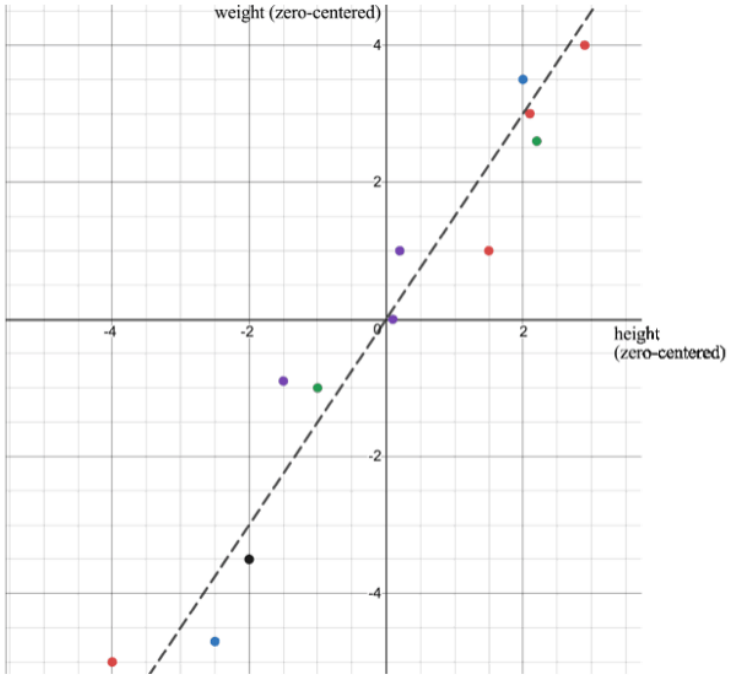
$$r_{pq} = \frac{c_{pq}}{\sigma_p \sigma_q} \in [-1, 1].$$

Back to PCA motivating example I

Back to the example with height and weight data, computing the sample covariance matrix gives

$$C = \frac{AA^\top}{n-1} = \frac{1}{11}\begin{pmatrix} 53.46 & 73.42 \\ 73.42 & 107.16 \end{pmatrix}.$$

$C_{12} = C_{21} > 0$ implies height and weight are positively correlated. This is evident from a line of best fit in the following plot of the data.



But what does the line of best fit represent? The equation of the line is $a(\text{height}) + b(\text{weight}) = 0$ for some $a, b \in \mathbb{R}$. This gives us a new variable!

From correlated to uncorrelated

Remember the definition from wikipedia: "PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components."

- Observations summarised in matrix $A \in \mathbb{R}^{m \times n}$ (with $m = 2$ features "height" and "weight") and $n = 12$ samples.
- Possibly correlated variables/features: from covariance matrix

$$C = \frac{AA^\top}{n-1} = \frac{1}{11}\begin{pmatrix} 53.46 & 73.42 \\ 73.42 & 107.16 \end{pmatrix},$$

  height and weight are correlated.

It remains to find variables/features (call them X, Y) that are uncorrelated from "height" and "weight". In particular, the covariance matrix associated to (X, Y), denoted by $\widehat{C}$, should look like

$$\widehat{C} = \begin{pmatrix} \widehat{c}_{11} & 0 \\ 0 & \widehat{c}_{22} \end{pmatrix}.$$

This means we should seek a transformation $C \mapsto \widehat{C}$.

Main idea of PCA

Since C is symmetric and positive semi-definite, it is diagonalisable with real and nonnegative eigenvalues. Then, C admits an eigenvalue decomposition $C = Q\widehat{C}Q^{-1}$ with orthogonal matrix Q and the eigenvalues listed on the diagonal of $\widehat{C}$.

Therefore the new uncorrelated variables (X, Y) should correspond to the eigenvectors of C. Namely, if

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix},$$

then the new variables are

$$X = q_{11}(\text{height}) + q_{21}(\text{weight}), \quad Y = q_{12}(\text{height}) + q_{22}(\text{weight}),$$

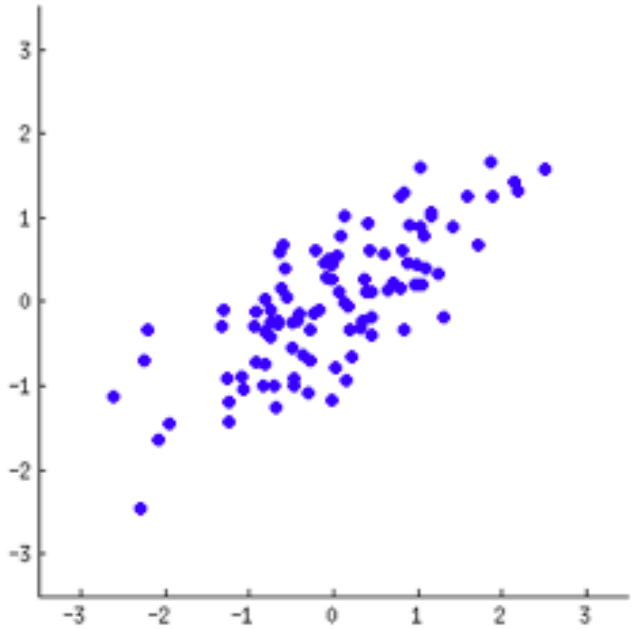
which don't have any real physical meaning, since they are linear combinations of the initial variables "height" and "weight".

We call these new uncorrelated variables the principal components (p.c.). Note that the total number of p.c. equals the dimension of the covariance matrix C!

The idea of PCA is to rank the p.c. by how much of the data is captured along each p.c.. The first p.c. captures the maximum possible information of the data, the second p.c. would capture the maximum remaining information, and so on...

Maximum possible information[1]

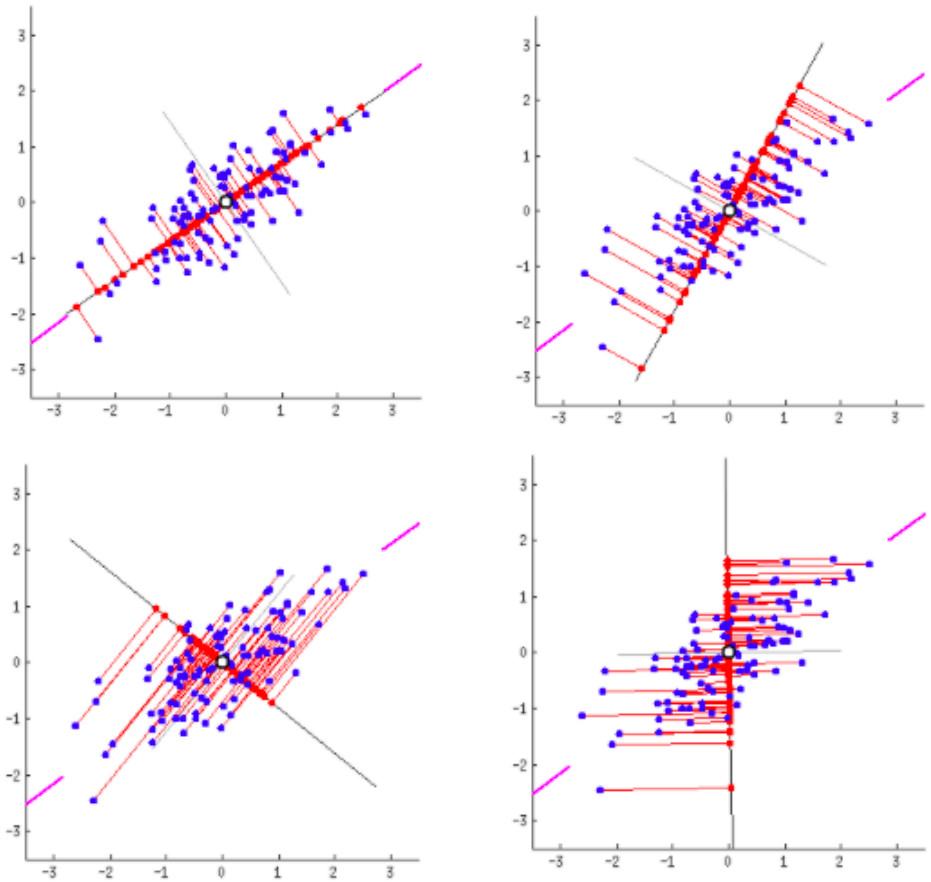
What is a good notion of "maximum possible information"? Let consider the following data plot



[1] Figures in this and the next slides are taken from https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Maximum possible information II

We can draw many lines through the data points:

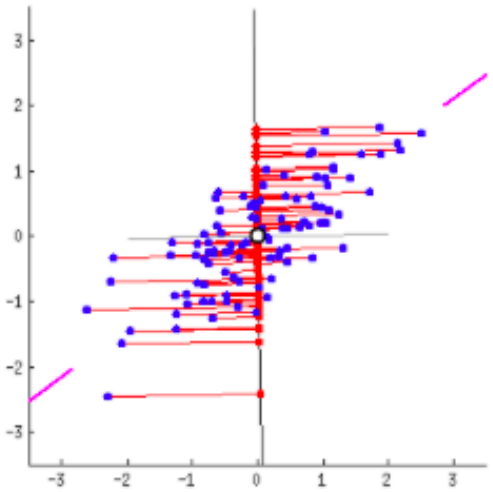


The red dots are projections of the data (blue dots) onto the line. The "spread" of the red dots on the line captures the variance, and the error between the red dot and its corresponding blue dot is measured by the length of the connecting red line.

Choice of maximum possible information

So, by "maximum possible information of the data" we can choose to mean:

- total reconstruction error E_R, given by the average squared length of the red lines, is minimised; or
- the variance Var, measured as the average squared distance from the origin to each red dot, is maximised.



It turns out they are equivalent! Heuristic explanation:

- The angle between the black line and red line is always 90 degrees.
- By Pythagoras theorem, the sum Var $+ E_R$ is the average squared distance from the origin to each blue dot, which is fixed!

Hence, maximising variance is the same as minimising reconstruction error. A more rigorous proof on the next slide.

Equivalence of PCA objectives I

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be an adjusted collection of n samples, i.e., $\sum_{i=1}^n \boldsymbol{x}_i = 0$. Let $\boldsymbol{v}$ be a unit vector and L be the line passing through the origin in direction $\boldsymbol{v}$.

What is the projection of $\boldsymbol{x}_i$ onto the line L? Let $\boldsymbol{p}_i \in L$ be the projection of $\boldsymbol{x}_i$. Then:

- $\boldsymbol{p}_i - \boldsymbol{x}_i$ is orthogonal to the line L, and so $(\boldsymbol{p}_i - \boldsymbol{x}_i) \cdot \boldsymbol{v} = 0$.
- $\boldsymbol{p}_i \in L$ implies $\boldsymbol{p}_i = c\boldsymbol{v}$ for some constant $c \in \mathbb{R}$.
- $\boldsymbol{v}$ is a unit vector and so $\boldsymbol{p}_i \cdot \boldsymbol{v} = c\boldsymbol{v} \cdot \boldsymbol{v} = c$, i.e., $\boldsymbol{p}_i = (\boldsymbol{p}_i \cdot \boldsymbol{v})\boldsymbol{v}$.
- Hence, $\boldsymbol{p}_i = (\boldsymbol{x}_i \cdot \boldsymbol{v})\boldsymbol{v}$.

Note that the projected points $\{\boldsymbol{p}_i\}_{i=1}^n$ are also centered around the origin:

$$\mu := \sum_{i=1}^n \boldsymbol{p}_i = \left(\left[\sum_{i=1}^n \boldsymbol{x}_i\right] \cdot \boldsymbol{v}\right)\boldsymbol{v} = 0.$$

The variance/spread of the projection of $\{\boldsymbol{x}_i\}_{i=1}^n$ on L is measured by

$$\mathrm{Var}(\boldsymbol{v}) = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{p}_i - \boldsymbol{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^n \boldsymbol{p}_i^2 = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i \cdot \boldsymbol{v})^2.$$

Let us rewrite this in terms of the covariance matrix.

Equivalence of PCA objectives II

If A is the data matrix, whose ith column is $\boldsymbol{x}_i$. Then, $(\boldsymbol{x}_i \cdot \boldsymbol{v})^2 = \boldsymbol{v} \cdot (\boldsymbol{x}_i \boldsymbol{x}_i^\top)\boldsymbol{v}$, and

$$\text{Var}(\boldsymbol{v}) = \boldsymbol{v} \cdot \left(\frac{1}{n-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{v} = \boldsymbol{v} \cdot \frac{1}{n-1} AA^\top \boldsymbol{v} = \boldsymbol{v} \cdot C\boldsymbol{v},$$

where C is the covariance matrix.

Next, the error between the data $\boldsymbol{x}_i$ and its projection $\boldsymbol{p}_i$ is

$$\|\boldsymbol{x}_i - \boldsymbol{p}_i\| = \|\boldsymbol{x}_i - (\boldsymbol{x}_i \cdot \boldsymbol{v})\boldsymbol{v}\|.$$

But recall from slide Decomposition of a vector II

$$(\boldsymbol{x}_i \cdot \boldsymbol{v})\boldsymbol{v} = (\boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{x}_i \quad \Longrightarrow \quad \boldsymbol{x}_i - (\boldsymbol{x}_i \cdot \boldsymbol{v})\boldsymbol{v} = (I - \boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{x}_i.$$

The reconstruction error can be expressed as

$$E_R(\boldsymbol{v}) = \sum_{i=1}^{n} \|\boldsymbol{x}_i - (\boldsymbol{x}_i \cdot \boldsymbol{v})\boldsymbol{v}\|_2^2 = \|(I - \boldsymbol{v}\boldsymbol{v}^\top)A\|_F^2,$$

with the Frobenius norm $\|\cdot\|_F$.

Equivalence of PCA objectives III

Theorem: Let $A \in \mathbb{R}^{m \times n}$ be a data matrix with zero row sum, and let $C = \frac{1}{n-1} AA^\top$ be its covariance matrix. Let $\boldsymbol{v}$ be a unit vector. Then,

$$\min_{\boldsymbol{v}} E_R(\boldsymbol{v}) \quad \Leftrightarrow \quad \max_{\boldsymbol{v}} \text{Var}(\boldsymbol{v}).$$

Proof [2]: (1) Let $\text{tr}(B) = \sum_{i=1}^{k} B_{ii}$ denote the trace of the square matrix $B \in \mathbb{R}^{k \times k}$. Then, the Frobenius norm $\|\cdot\|_F$ has the alternate characterisation: $\|A\|_F^2 = \text{tr}(A^\top A)$ for $A \in \mathbb{R}^{m \times n}$.

(2) From the reconstruction error, we see (using $(\boldsymbol{v}\boldsymbol{v}^\top)$ is symmetric)

$$\begin{aligned}
E_R(\boldsymbol{v}) &= \|(I - \boldsymbol{v}\boldsymbol{v}^\top)A\|_F^2 = \text{tr}\Big((A - (\boldsymbol{v}\boldsymbol{v}^\top)A)^\top (A - (\boldsymbol{v}\boldsymbol{v}^\top)A)\Big) \\
&= \text{tr}\Big(A^\top A\Big) - 2\text{tr}\Big(A^\top (\boldsymbol{v}\boldsymbol{v}^\top)A\Big) + \text{tr}\Big(A^\top (\boldsymbol{v}\boldsymbol{v}^\top)(\boldsymbol{v}\boldsymbol{v}^\top)A\Big) \\
&= \text{tr}\Big(A^\top A\Big) - \text{tr}\Big(A^\top (\boldsymbol{v}\boldsymbol{v}^\top)A\Big) \quad \text{since } (\boldsymbol{v}\boldsymbol{v}^\top)(\boldsymbol{v}\boldsymbol{v}^\top) = \boldsymbol{v}\boldsymbol{v}^\top \\
&= \text{tr}\Big(A^\top A\Big) - \text{tr}\Big(\boldsymbol{v}^\top AA^\top \boldsymbol{v}\Big) \\
&= \text{tr}\Big(A^\top A\Big) - (n-1)\boldsymbol{v}^\top C\boldsymbol{v} = \text{tr}\Big(A^\top A\Big) - (n-1)\text{Var}(\boldsymbol{v}).
\end{aligned}$$

The first term $\text{tr}\Big(A^\top A\Big)$ is independent of $\boldsymbol{v}$, and so $\min_{\boldsymbol{v}} E_R(\boldsymbol{v})$ is equivalent to $\max_{\boldsymbol{v}} \text{Var}(\boldsymbol{v})$. □

[2] https://stats.stackexchange.com/questions/32174/pca-objective-function-what-is-the-connection-between-maximizing-variance-and-m

Variance maximisation

Given a covariance matrix C, finding the direction of maximal variance (i.e., the first principal component) corresponds to

$$\max_{\boldsymbol{v}} \mathrm{Var}(\boldsymbol{v}) = \boldsymbol{v} \cdot C\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{v}\|_2 = 1.$$

We use the Lagrange multiplier method, and introduce the Lagrangian

$$L(\boldsymbol{v}, \mu) = \boldsymbol{v} \cdot C\boldsymbol{v} - \mu(\boldsymbol{v}^\top \boldsymbol{v} - 1),$$

for $\mu \in \mathbb{R}$ (known as the Lagrange multiplier for the constraint $\|v\|_2 = 1$). Computing the partial derivatives shows

$$\frac{\partial L}{\partial \boldsymbol{v}} = 2(C\boldsymbol{v} - \mu\boldsymbol{v}), \quad \frac{\partial L}{\partial \mu} = \boldsymbol{v}^\top \boldsymbol{v} - 1 = \|\boldsymbol{v}\|_2^2 - 1.$$

Hence, at a critical point $(\boldsymbol{v}^*, \mu^*)$ of L we get

$$C\boldsymbol{v}^* = \mu^* \boldsymbol{v}^*, \quad \|\boldsymbol{v}^*\|_2^2 = 1,$$

i.e., $\boldsymbol{v}^*$ is a unit eigenvector of C with corresponding eigenvalue μ^*. Substituting $(\boldsymbol{v}^*, \mu^*)$ back into the Lagrangian gives

$$L(\boldsymbol{v}^*, \mu^*) = \boldsymbol{v}^* \cdot C\boldsymbol{v}^* = \mathrm{Var}(\boldsymbol{v}^*) = \boldsymbol{v}^* \cdot \mu^* \boldsymbol{v}^* = \mu^* \|\boldsymbol{v}^*\|_2^2 = \mu^*.$$

Therefore, the maximal variance is the largest eigenvalue λ_1 of C, and the first principal component should be the corresponding eigenvector $\boldsymbol{v}_1$.

Second principal component

What about the second principal component?

- This should maximise the "remaining variance" not captured by $\boldsymbol{v}_1$.
- This should be orthogonal to $\boldsymbol{v}_1$.

Hence, finding $\boldsymbol{v}_2$ corresponds to

$$\max_{\boldsymbol{v}} \mathrm{Var}(\boldsymbol{v}) = \boldsymbol{v} \cdot C\boldsymbol{v} \quad \text{subject to } \|v\|_2 = 1 \text{ and } \boldsymbol{v} \cdot \boldsymbol{v}_1 = 0.$$

Introduce Lagrange multipliers $\alpha, \beta \in \mathbb{R}$ and consider

$$L(\boldsymbol{v}, \alpha, \beta) = \boldsymbol{v} \cdot C\boldsymbol{v} - \alpha(\boldsymbol{v}^\top \boldsymbol{v} - 1) - \beta \boldsymbol{v} \cdot \boldsymbol{v}_1.$$

At a critical point $(\boldsymbol{v}^*, \alpha^*, \beta^*)$ we have all partial derivatives of L vanishing:

$$\frac{\partial L}{\partial \boldsymbol{v}} = 2(C\boldsymbol{v}^* - \alpha^* \boldsymbol{v}^*) - \beta^* \boldsymbol{v}_1 = 0, \quad \frac{\partial L}{\partial \alpha} = \|\boldsymbol{v}^*\|_2^2 - 1 = 0, \quad \frac{\partial L}{\partial \beta} = \boldsymbol{v}^* \cdot \boldsymbol{v}_1 = 0.$$

Solving for β^*:

$$\beta^* = \boldsymbol{v}_1^\top (\beta^* \boldsymbol{v}_1) = 2(C\boldsymbol{v}^* - \alpha^* \boldsymbol{v}^*) \cdot \boldsymbol{v}_1 = 2(\lambda_1 \boldsymbol{v}_1 - \alpha^* \boldsymbol{v}_1) \cdot \boldsymbol{v}^* = 0.$$

Then, as before, we have $C\boldsymbol{v}^* = \alpha^* \boldsymbol{v}^*$. I.e., $(\alpha^*, \boldsymbol{v}^*)$ is an eigenpair of C, but which pair? Plugging back into L shows that α^* should be the second largest eigenvalue λ_2 of C with the second principal component as the corresponding eigenvector $\boldsymbol{v}_2$.

PCA in general

From variance maximisation, we find that for an adjusted (i.e., mean zero) data matrix $A \in \mathbb{R}^{m \times n}$ with covariance matrix $C = \frac{1}{n-1} AA^\top \in \mathbb{R}^{m \times m}$,

- the first principal component, i.e., the direction of maximal variance, is the eigenvector $\boldsymbol{v}_1$ of C corresponding to the largest eigenvalue λ_1;
- the second principal component, i.e., the direction of maximal variance orthogonal to $\boldsymbol{v}_1$, is the eigenvector $\boldsymbol{v}_2$ of C corresponding to the second largest eigenvalue λ_2;
- the kth principal component, i.e., the direction of maximal variance orthogonal to $\text{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}\}$, is the eigenvector $\boldsymbol{v}_k$ of C corresponding to the kth largest eigenvalue λ_k.

Since C has m eigenvalues, there will be m principal components.

Remaining issues:

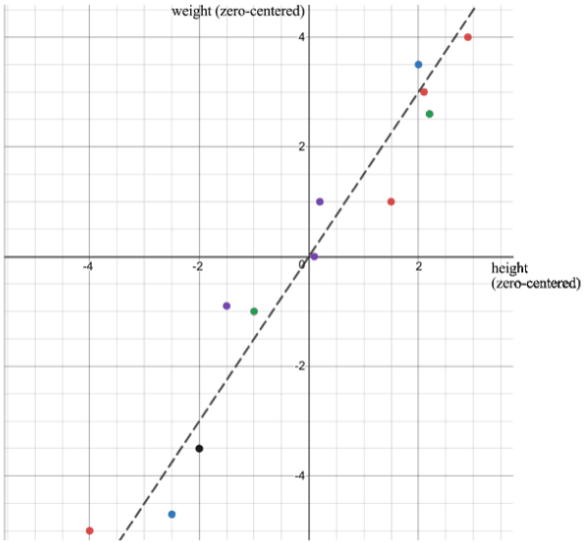
- A procedure to "rank" the eigenvalues in decreasing order, so that we can extract the principal components more easily.
- A criterion to choose how many principal components to use for a "good" summary of the data.

PCA motivating example VI

Returning to the example with "height" and "weight" data. The eigenvalue decomposition of the covariance matrix C gives

$$C = Q\widehat{C}Q^\top \quad \text{with} \quad \widehat{C} = \begin{pmatrix} 0.1940 & 0 \\ 0 & 14.4078 \end{pmatrix} \text{ and } Q = \begin{pmatrix} -0.8196 & 0.5729 \\ 0.5729 & 0.8196 \end{pmatrix}$$

From this we see that the first principal component is $\boldsymbol{v}_1 = (0.5729, 0.8196)^\top$, i.e., the direction of the dotted line, and the second principal component is $\boldsymbol{v}_2 = (-0.8196, 0.5729)^\top$.



Just how much of the variance is explained by the first principal component?

When can we discard the second principal component to obtain a simple predictive model?

Class exercise

1. How can we ensure that for a symmetric matrix C (such as the covariance matrix $C = \frac{1}{n-1}AA^\top$), there is a complete set of orthonormal eigenvectors? I.e., the matrix C is non-defective. [Hint: look back at the Schur factorization slides on Part 1].

2. Show that the kth principal component should be taken as the eigenvector of C corresponding to the kth largest eigenvalue by solving the variance maximisation problem

   $$\max_{\boldsymbol{v}} \mathrm{Var}(\boldsymbol{v}) = \boldsymbol{v} \cdot C\boldsymbol{v} \quad \text{subject to } \|\boldsymbol{v}\|_2^2 = 1, \quad \boldsymbol{v} \cdot \boldsymbol{v}_i = 0 \text{ for } 1 \leq i \leq k-1$$

   with the help of the Lagrangian

   $$L(\boldsymbol{v}, \alpha, \mu_1, \ldots, \mu_{k-1}) = \boldsymbol{v} \cdot C\boldsymbol{v} - \alpha(\|\boldsymbol{v}\|_2^2 - 1) - \sum_{i=1}^{k-1} \mu_i \boldsymbol{v} \cdot \boldsymbol{v}_i,$$

   where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ are the first $k-1$ principal components.

Alternate formulation

A problem with the eigenvalue decomposition $C = Q\widehat{C}Q^\top$ is that there is no ordering of the eigenvalues in $\widehat{C}$, e.g. the "height" vs "weight" example. What is a decomposition of C that provides a ranking of the eigenvalues in decreasing value? Ans: The singular value decomposition.

Recall: The (full) singular value decomposition of a matrix $B \in \mathbb{R}^{k\times l}$ is

$$B = U\Sigma V^\top$$

with $U \in \mathbb{R}^{k\times k}$, $V \in \mathbb{R}^{l\times l}$ orthogonal, and $\Sigma \in \mathbb{R}^{k\times l}$ is diagonal.

In particular, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_l)$ where the singular values

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l \geq 0,$$

are the (positive) square root of the eigenvalues of $B^\top B \in \mathbb{R}^{l\times l}$.

The columns of V are the eigenvectors corresponding to the eigenvalue $\{\sigma_i^2\}$ of $B^\top B$. Moreover, the SVD can be written as a sum of rank-one matrices:

$$B = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

where $\mathrm{rank}(B) = r \leq \min(k, l)$ and u_i, v_i are the ith columns of U and V.

Formulation with SVD

Therefore, performing the SVD of the scaled data matrix transpose $B = \frac{1}{\sqrt{n-1}} A^\top \in \mathbb{R}^{n \times m}$, we obtain orthogonal matrices $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{m \times m}$ and diagonal matrix $\Sigma \in \mathbb{R}^{n \times m}$ such that

$$B = \frac{1}{\sqrt{n-1}} A^\top = U \Sigma V^\top.$$

How does this help?

- The set of singular values $\sigma_1, \ldots, \sigma_m$ contained on the diagonal of Σ coincide with the square root of the eigenvalues $\lambda_1, \ldots, \lambda_m$ of $B^\top B = \frac{1}{n-1} A A^\top = C$. Moreover, they are arranged so that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m.$$

- The columns of V are the eigenvectors of $B^\top B = C$, i.e., the matrix V contains the principal components.

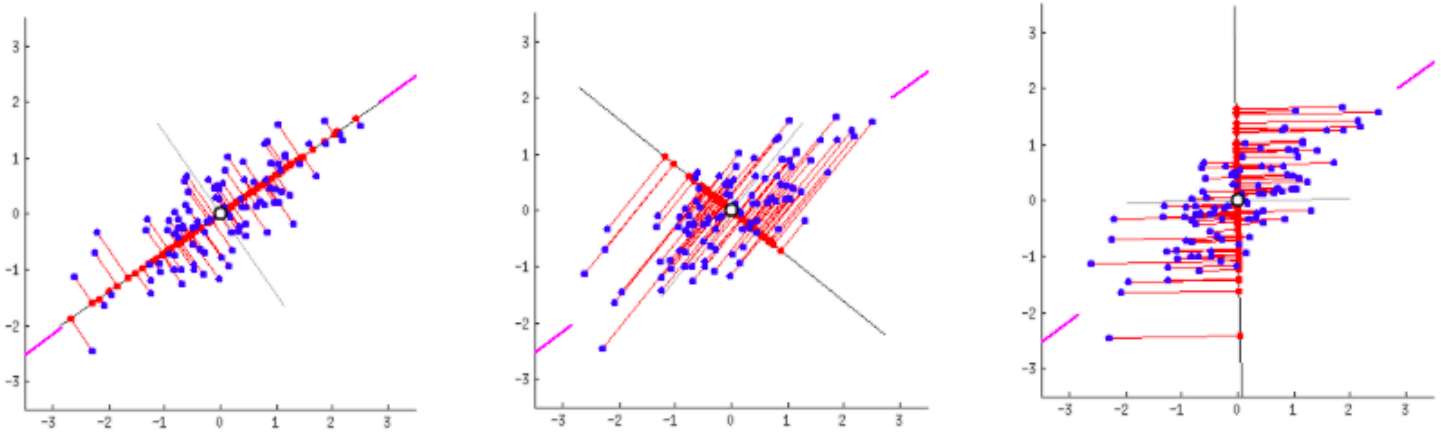
Thus, computing the SVD of $\frac{1}{\sqrt{n-1}} A^\top$ provides an ordering of the eigenvalues, as well as the principal components in one fell swoop! This answers the first issue about an efficient decomposition that allows us to rank the eigenvalues in order.

Variance and eigenvalues

Before discussing how to choose the number of principal components, we first relate the notion of variance and eigenvalues.

The correlation of variables in the data sample (e.g. "height" and "weight") are neatly summarized in the covariance matrix C.

The variance is an important quantity, as a larger variance means a larger dispersion of data points along a line (principal component), and a larger dispersion means more information of the data points are contained on the line.



How do we quantify the variance of the data along a principal component?

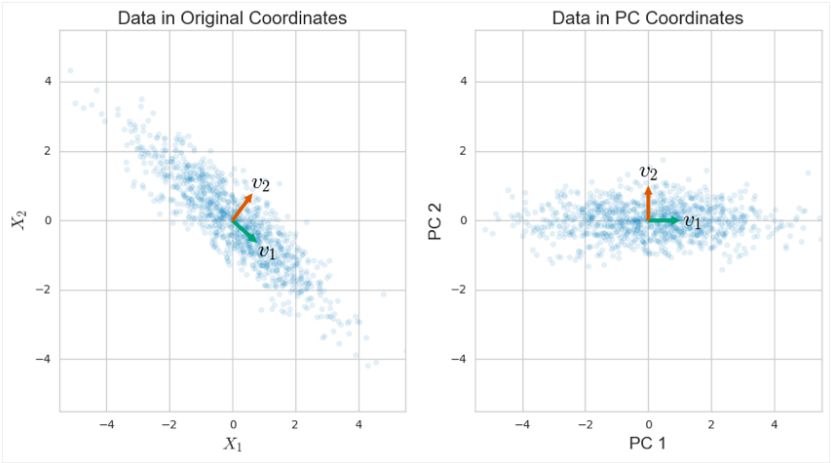
Variance and eigenvalues II

If the covariance matrix $C \in \mathbb{R}^{m \times m}$ is diagonal, i.e., $C = \text{diag}(\lambda_1, \ldots, \lambda_m)$, then the m variables $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ that compose the data matrix A are uncorrelated (as $C_{ij} = 0$ for $i \neq j$), and the (sample) variance of $\boldsymbol{x}_i$ is λ_i.

From the SVD $\frac{1}{\sqrt{n-1}} A^\top = U \Sigma V^\top$, we have the eigenvalue decomposition

$$C = V \Sigma^2 V^\top.$$

This also gives a change of basis, transforming from the standard basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m\}$ to a new basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ of principal components, so that the covariance matrix in the new basis is diagonal.

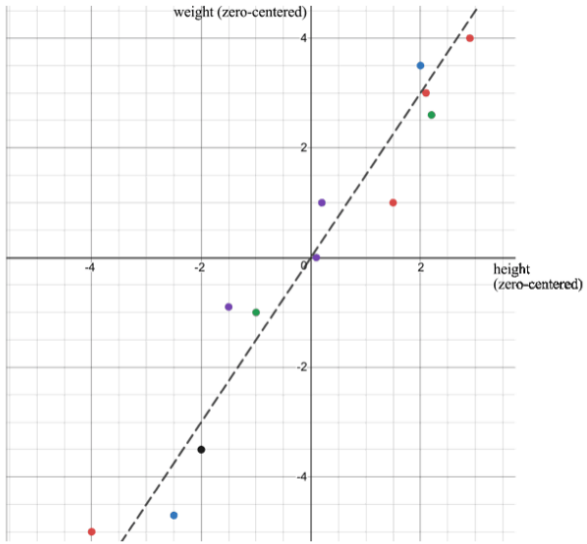


Then, the variance along the kth principal component is just λ_k, i.e., the kth largest eigenvalue of C.

Selection of principal components

In applications, PCA can be viewed as a form of dimension reduction. For an adjusted data matrix $A \in \mathbb{R}^{m \times n}$ with large $m, n \gg 1$, PCA is used to extract out a smaller number K of principal components that can capture the essence of the data.

In the previous example with "height" and "weight" data, the plot shows that it is enough to obtain the first principal component for the line of "best" fit.



A more concrete way of selecting the number K of principal component is to look at their contribution to the total variance.

Total and explained variance I

Let us consider an example data set[3]

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
2	1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
3	1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
4	1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735

The 13 columns denote the different features, and each row (5 out of 124 shown here) is a particular sample of the data. In our notation, this corresponds to the transpose data matrix $A^\top \in \mathbb{R}^{13\times 124}$.

The procedure is:

1. Adjust the above data matrix $A^\top$ by subtracting the mean of each column (aka standardise the data set).
2. Construct the covariance matrix $C \in \mathbb{R}^{13\times 13}$.
3. Compute the eigenvalues and eigenvectors of C.

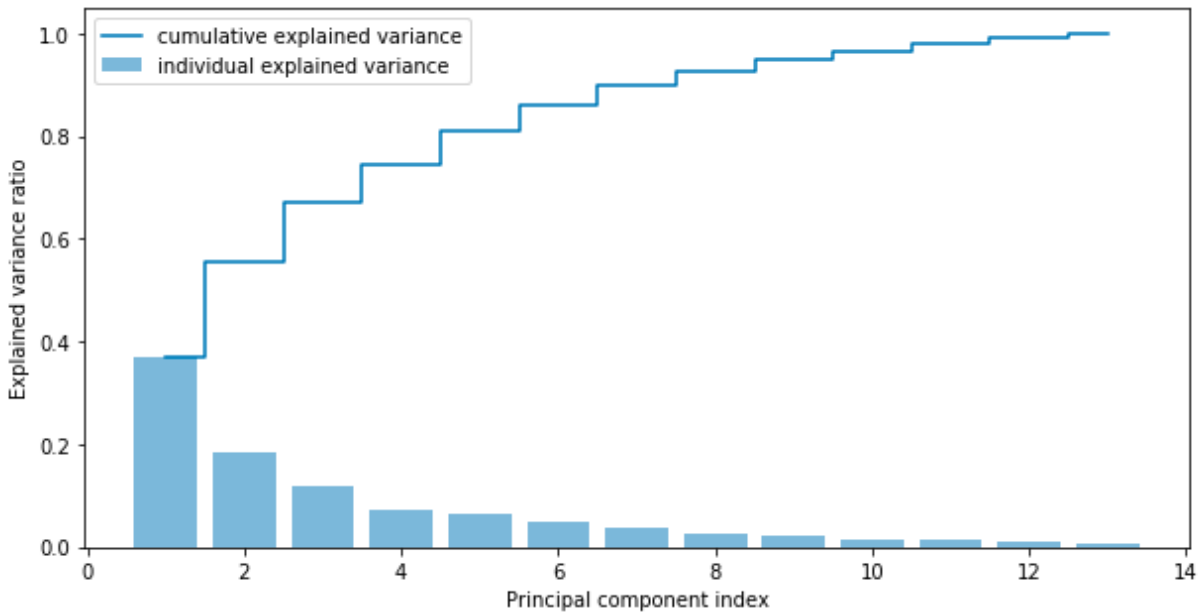
Before deciding how many principal components to keep we plot the variance explained ratios of the eigenvalues, which is the fraction

$$\frac{\lambda_j}{\sum_{i=1}^{13} \lambda_i}$$

[3] Figures taken from `https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad`

Total and explained variance II

The variance explained ratio plot is



From this we see that

- the first principal component accounts for 40% of the variance,
- the first and second components account for 60% of the variance,
- of course, 100% of the variance is accounted for by using all principal components.

So, a suitable number of principal components depends on how much variance you want to capture. There is always a trade-off between computational efficiency/storage (smaller K) and performance (larger captured variance).

Measuring accuracy I

Suppose k principal components are selected for an adjusted set of n samples $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Then, the projection of $\boldsymbol{x}_i$ to the k-dimensional subspace spanned by the principal components $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$ is

$$\boldsymbol{p}_i = (\boldsymbol{x}_i \cdot \boldsymbol{v}_1)\boldsymbol{v}_1 + \cdots + (\boldsymbol{x}_i \cdot \boldsymbol{v}_k)\boldsymbol{v}_k = \sum_{j=1}^{k} (\boldsymbol{v}_j \cdot \boldsymbol{x}_i)\boldsymbol{v}_j.$$

The reconstruction error for the ith data point $\boldsymbol{x}_i$ is therefore

$$\boldsymbol{e}_i = \boldsymbol{x}_i - \boldsymbol{p}_i = \sum_{j=k+1}^{m} (\boldsymbol{x}_i \cdot \boldsymbol{v}_j)\boldsymbol{v}_j,$$

since $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ forms an orthonormal basis of $\mathbb{R}^m$. Then,

$$\|\boldsymbol{e}_i\|_2^2 = \sum_{j=k+1}^{m} (\boldsymbol{x}_i \cdot \boldsymbol{v}_j)^2.$$

The relative reconstruction error E_k using k principal components is

$$E_k = \left(\frac{\sum_{i=1}^{n} \|\boldsymbol{e}_i\|_2^2}{\sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2} \right)^{1/2}$$

Measuring accuracy II

Using the orthonormality of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$, and the fact that they are eigenvectors of C, we can show (class exercise) that

$$\sum_{i=1}^{n} \|\boldsymbol{e}_i\|_2^2 = (n-1) \sum_{j=k+1}^{m} \lambda_j, \quad \sum_{i=1}^{m} \|\boldsymbol{x}_i\|_2^2 = (n-1) \sum_{j=1}^{m} \lambda_j.$$

Hence, the relative reconstruction error can be expressed as

$$E_k = \left(\frac{\sum_{j=k+1}^{m} \lambda_j}{\sum_{j=1}^{m} \lambda_j} \right)^{1/2}.$$

Due to the equivalence between variance maximisation and minimising reconstruction error, E_k provides a measure of the loss of variance by choosing k out of m principal components.

At present there is no defining criterion to pick the best k. The heuristics is "pick the smallest k that captures at least 85/90/95% of the variance". While this is fine for low dimensions, for high dimensions $m, n \gg 1$ it is questionable if high variance = high importance.

Projected data

Suppose we choose k out of m principal components after the analysis of an adjusted data matrix $A \in \mathbb{R}^{m \times n}$.

Let us construct the projection matrix $W \in \mathbb{R}^{m \times k}$ whose columns are the k principal components $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$. We define the PCA subspace as the vector space spanned by the columns of W. Then, the data in the PCA subspace is summarised by the matrix $Y := W^\top A \in \mathbb{R}^{k \times n}$.

In particular,

$$Y = \begin{pmatrix} — & \boldsymbol{v}_1 & — \\ & \vdots & \\ — & \boldsymbol{v}_k & — \end{pmatrix} \begin{pmatrix} | & | & & | \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & & | \\ \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_n \\ | & | & & | \end{pmatrix},$$

where $\boldsymbol{y}_j$ gives the coordinate of the jth sample in the PCA subspace:

$$\boldsymbol{y}_j = (\boldsymbol{v}_1 \cdot \boldsymbol{x}_j, \boldsymbol{v}_2 \cdot \boldsymbol{x}_j, \ldots, \boldsymbol{v}_k \cdot \boldsymbol{x}_j)^\top.$$

Furthermore, if we calculate the covariance matrix C_Y of the data Y, we see

$$C_Y = \frac{1}{n-1} YY^\top = \frac{1}{n-1} W^\top AA^\top W = W^\top CW = W^\top V\Sigma^2 V^\top W.$$

Since $V^\top W \in \mathbb{R}^{m \times k}$ is diagonal with entries 1, we have

$$C_Y = W^\top V\Sigma^2 V^\top W = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_k^2).$$

Hence, in the PCA subspace the features/principal components are uncorrelated!

Application - Facial recognition

Goal: To build a computational model of facial recognition, i.e., an algorithm that determines whether a facial image belongs to some individual we know.

Difficulties: Faces are complex and multidimensional. Many things can complicate the recognition algorithm, e.g. lighting, pose, background, foreground, smiling, frowning etc.

Approach: Use PCA to decompose a training set of facial images into a small set of characteristic feature images called eigenfaces (developed by Turk and Pentland 1991). The linear span of these eigenfaces is denoted as the face space.

Recognition is performed by projecting a new facial image into the face space. Then, it can be classified by comparing its position with the positions of known individuals in the face space.

The Eigenface approach I

Every two-dimensional image I with $N \times N$ array of pixels can be considered as a vector of dimension N^2. A typical 256-by-256 pixel image lies in a 65,536 dimensional space (Huge!)

The key assumption is that the images of faces will not be randomly distributed in this huge image space, but can be described by a relatively low dimensional subspace, which we call the face space.

Suppose we have M faces $\Gamma_1, \ldots, \Gamma_M$, each can be interpreted as a vector of dimension N^2. The averaged face is $\Psi := \frac{1}{M} \sum_{i=1}^{M} \Gamma_i$.

Figure: Taken from Turk and Pentland (91). Left is the data of face images. Right is the averaged face

The Eigenface approach II

We define the deviations $\Phi_i := \Gamma_i - \Psi$ for $1 \leq i \leq M$, and apply PCA to the set of data $\{\Phi_1, \ldots, \Phi_M\}$.

Construct the data matrix $A = [\Phi_1 | \Phi_2 | \cdots | \Phi_M] \in \mathbb{R}^{N^2 \times M}$, and the covariance matrix

$$C = \frac{1}{M-1} A A^\top \in \mathbb{R}^{N^2 \times N^2}.$$

The task is to extract of the first k principal components from the eigenvectors of C to build the face space.

Practical issue: Determining the eigenvalues and eigenvectors of a N^2-by-N^2 matrix is an intractable task!

Does this means we give up?

The Eigenface approach III

Solution: Consider the matrix

$$B := \frac{1}{M-1} A^\top A \in \mathbb{R}^{M \times M}.$$

If $M < N^2$, i.e., the number of data points in image space is less than the dimension of the image space. Then, it might be more feasible to find the eigenvalues/eigenvectors of B.

But what's the point? Let v be an eigenvector of B with eigenvalue μ. Then,

$$Bv = \frac{1}{N-1} A^\top Av = \mu v \quad \Longrightarrow \quad ABv = \frac{1}{N-1} AA^\top Av = C(Av) = \mu Av.$$

I.e., Av is an eigenvector of C with eigenvalue μ.

Are they also orthogonal? If u and v are two orthogonal eigenvectors of B, then

$$Au \cdot Av = u^\top A^\top Av = (M-1)u^\top Bv = \mu_v (M-1) u^\top v = 0.$$

Consequences:

- There are only M meaningful eigenvectors from the covariance matrix C, as the rest are associated with the zero eigenvalue.
- The calculations are greatly reduced!

The Eigenface approach IV

Let $\{v_1, \ldots, v_M\}$ be the orthonormal eigenvectors of $B = \frac{1}{M-1} A^\top A$, ranked with decreasing values of corresponding eigenvalues, and set

$$w_i = Av_i \quad \text{for } 1 \leq i \leq M.$$

Then, $\{w_1, \ldots, w_M\}$ are the eigenvectors of C with non-zero eigenvalues.

Figure: The first seven eigenfaces calculated from the data set.

Exercise. If μ is the eigenvalue corresponding to the orthonormal eigenvector v for B, what is the corresponding eigenvalue to the eigenvector Av for C?

Recognising new faces

Fix $k < M$ (can be chosen arbitrarily or by looking at the variance explained ratio plot), then the face space $\mathcal{S}$ is

$$\mathcal{S} = \text{span}\{w_1, \ldots, w_k\}.$$

Given a new face Γ, its coordinate in the face space $\mathcal{S}$ is given by the vector $(\gamma_1, \ldots, \gamma_k)$, where

$$\gamma_i = w_i \cdot (\Gamma - \Psi) \quad \text{for } 1 \leq i \leq k.$$

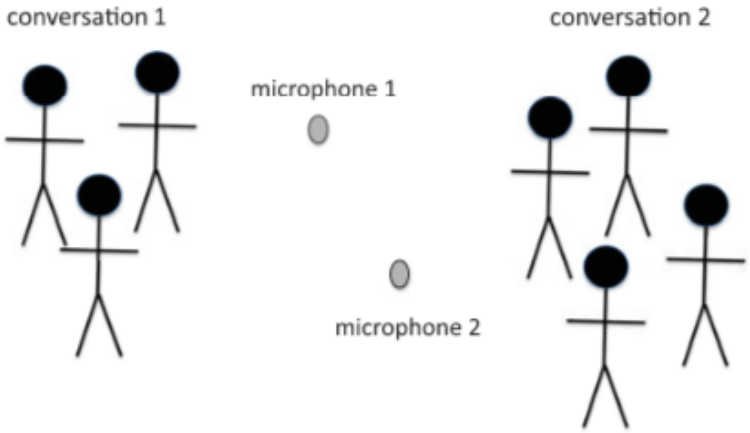
Its approximation $\hat{\Gamma}$ can be expressed as

$$\hat{\Gamma} = \sum_{i=1}^{k} \gamma_i w_i$$

Figure: A new face image (left) and its projection to the face space spanned by the 7 eigenfaces.

§8 - Independent component analysis (ICA)

The Cocktail party problem



- Two conversations happening simultaneously.
- Two microphones placed at different locations and receive a mixture of signals (the conservations + other noise).
- How can we separate out the signals to reconstruct each conversation?

Mathematically: let $s_1(t)$ and $s_2(t)$ be the signals from the two conservations. We measure the mixed recorded signals

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t),$$
$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t),$$

at microphones 1 and 2, where a_{ij} are the mixing parameters.

The mathematical problem: Given $(x_1(t), x_2(t))$, find $(s_1(t), s_2(t))$.

Blind source reconstruction

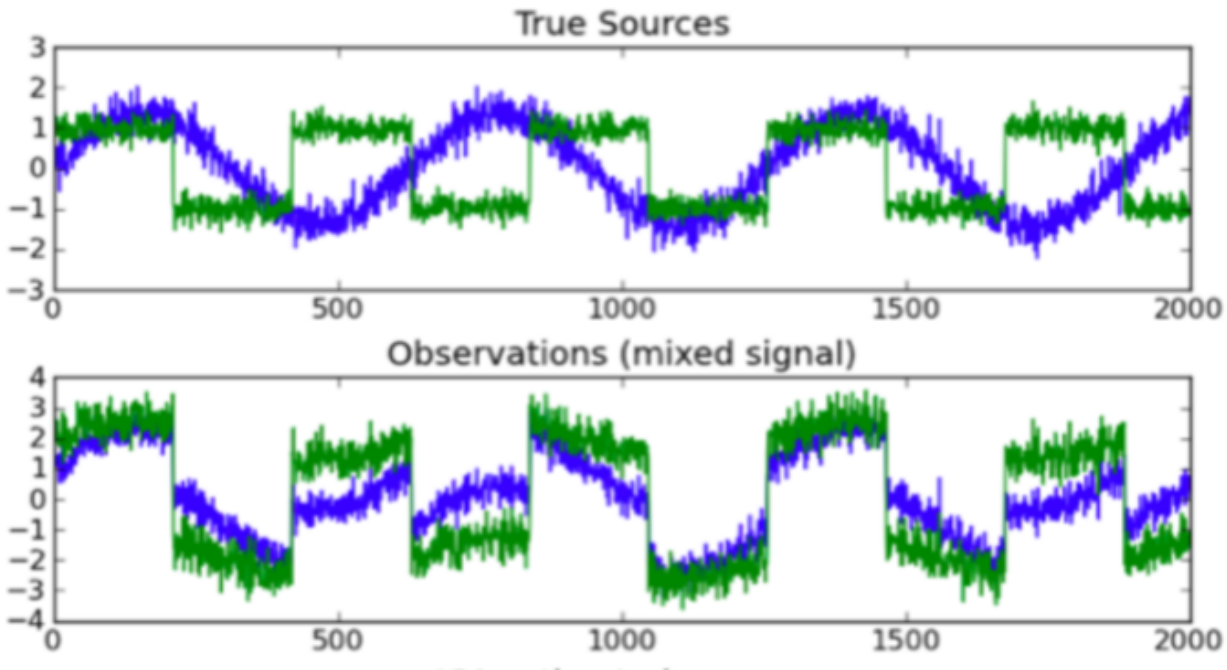


Figure: Taken from `https://www.cs.ubc.ca/~jnutini/documents/mlrg_pca.pdf`

From the observed mixed signals $x_1(t)$ and $x_2(t)$, recover the two original signals $s_1(t)$ and $s_2(t)$.

Setting of a more general problem

Suppose we record a multi-dimensional data $\boldsymbol{x}$, each sample is a random draw from an unknown probability distribution $P(\boldsymbol{x})$.

We assume there exists some underlying sources $\boldsymbol{s}$ where each source s_i is statistically independent of all other sources s_j, $j \neq i$.

The key assumption of the independent component analysis (ICA) is that the observed data $\boldsymbol{x}$ is a linear mixture of the underlying source $\boldsymbol{s}$, i.e., there is an unknown invertible square matrix A such that

$$\boldsymbol{x} = A\boldsymbol{s}.$$

The goal of ICA is find the unknown mixing matrix A, or more specifically, an approximation W to its inverse A^{-1}, so that

$$\widehat{\boldsymbol{s}} := W\boldsymbol{x}$$

is a good approximation of the true underlying source $\boldsymbol{s} = A^{-1}\boldsymbol{x}$.

Challenges and strategies

Since the mixing matrix A and the underlying source $\boldsymbol{s}$ unknown, it appears impossible to infer both A and $\boldsymbol{s}$ from the equation

$$\boldsymbol{x} = A\boldsymbol{s}.$$

One strategy (divide and conquer) is to just find the mixing matrix A, as oppose to finding both A and $\boldsymbol{s}$ simultaneously. We will again use the singular value decomposition, namely if

$$A = U\Sigma V^\top,$$

then we will find ways to get approximations $\tilde{U}$, $\tilde{\Sigma}$ and $\tilde{V}$ just from the data $\boldsymbol{x}$ so that

$$W := \tilde{V}\tilde{\Sigma}^{-1}\tilde{U}^\top$$

is a good approximation of A^{-1}.

Probability recap

In these slides we will always assume X is a real continuous random variables with values in $(-\infty, \infty)$. Associated to X is its probability distribution function f_X, which gives

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

I.e., the probability that the random variable X realises a value in the interval $[a, b]$ is given by the integral of f_X over $[a, b]$.

Note that for continuous random variables, it does not make sense to find $\mathbb{P}(X = c)$.

The Expectation/Mean of a random variable X with probability distribution p_X is

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} x p_X(x)dx =: \mu,$$

and the variance of X is

$$\mathbb{E}((X - \mu)^2) = \int_{-\infty}^{\infty} (x - \mu)^2 p_X(x)dx =: \sigma^2.$$

Joint distribution and independence

Two random variables X and Y can be assigned a joint probability distribution $p_{X,Y}$, if for any subset $D \in \mathbb{R}^2$ it holds

$$\mathbb{P}((X, Y) \in D) = \int_D p_{X,Y}(x, y)dx\, dy.$$

E.g.,

$$\mathbb{P}((X, Y) \in (a, b) \times (c, d)) = \int_c^d \int_a^b p_{X,Y}(x, y)dx\, dy.$$

We define the marginal distribution p_X of X by

$$p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dy$$

and the marginal distribution p_Y of Y by

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X,Y}(x, y)dx.$$

Then, X and Y are independent if the joint distribution can be factorised as

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

Ambiguities of ICA

The ICA problem involves solving

$$\boldsymbol{x} = A\boldsymbol{s}$$

for unknown A and random variable $\boldsymbol{s}$. Immediately, we see that

- it is not possible to determine the variance of $\boldsymbol{s}$, since a scalar multiple of a component s_j can be cancelled by dividing columns of A by the same scalar. I.e.,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} a_{11}/\alpha & a_{12}/\beta \\ a_{21}/\alpha & a_{22}/\beta \end{pmatrix} \begin{pmatrix} \alpha s_1 \\ \beta s_2 \end{pmatrix}.$$

  To deal with this, we fix the variance of each signal s_i to be 1, i.e., $\mathrm{Var}(s_i) = \mathbb{E}((s_i - \mathbb{E}(s_i))^2) = \mathbb{E}(s_i^2) = 1$. [This is called Whitening] But note that there is still the ambiguity of the sign, since $-s_i$ is also a solution with variance 1.

- there is no natural ordering of the signal components $\boldsymbol{s}$, since for any permutation matrix P, it holds $\boldsymbol{x} = AP^{-1}P\boldsymbol{s}$ and AP^{-1} is a new unknown mixing matrix. I.e.,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{pmatrix} \begin{pmatrix} s_2 \\ s_1 \end{pmatrix}.$$

  But in practice, this is not a big problem.

Illustration I

Consider two independent random variables s_1 and s_2, whose probability distributions are the uniform distribution

$$p(s_i) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{if } -\sqrt{3} \leq s_i \leq \sqrt{3}, \\ 0 & \text{otherwise.} \end{cases}$$

Then, the mean is zero and the variance is 1. The joint probability density is the product $p(s_1, s_2) = p(s_1)p(s_2)$ due to the independence, which is again a uniform distribution on the square $[-\sqrt{3}, \sqrt{3}]^2$.

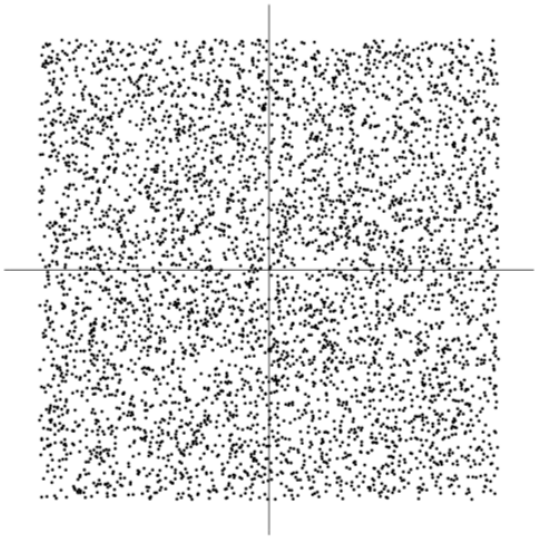
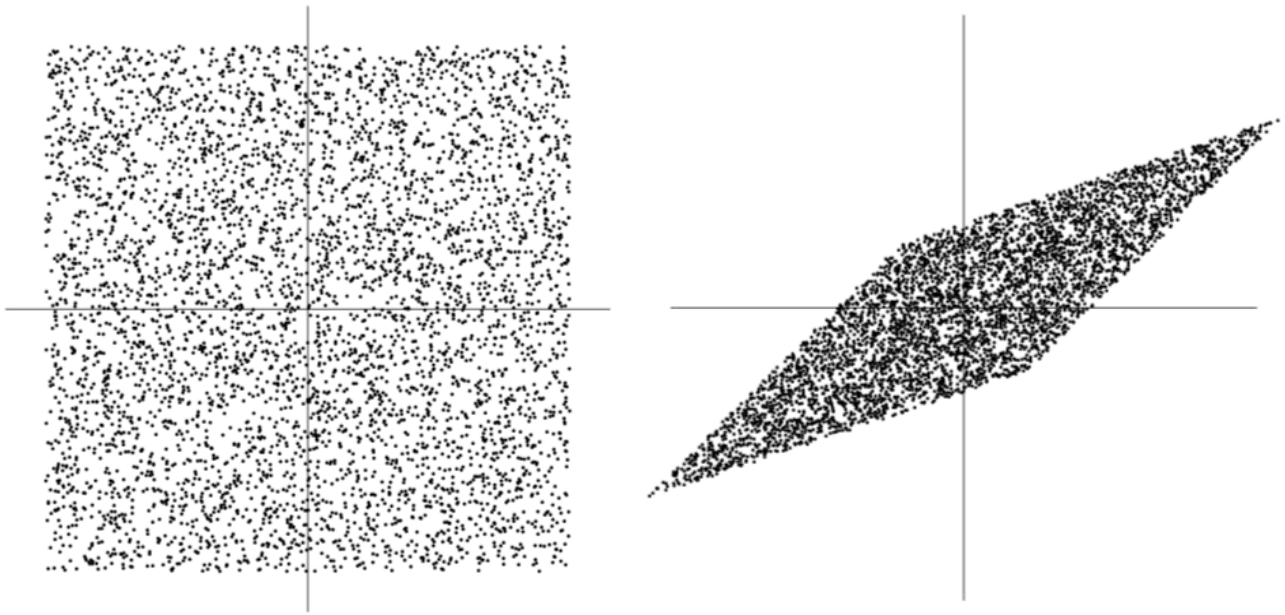


Illustration II

Consider a mixing matrix

$$A = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$$

applied to the sample for the signals $\boldsymbol{s} = (s_1, s_2)$ on the left, leading to the sample for the mixture $\boldsymbol{x} = (x_1, x_2)$ on the right.



The new random variables $\boldsymbol{x} = A\boldsymbol{s}$ are not independent, since if x_1 attains the maximum value, then we can infer the value of x_2.

Uncorrelated

Two random variables X and Y are uncorrelated if their covariance is zero:

$$\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

Lemma: Independence implies uncorrelated.

Proof:

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp_{X,Y}(x,y)dxdy \\ &= \Big(\int_{-\infty}^{\infty} xp_X(x)dx\Big)\Big(\int_{-\infty}^{\infty} yp_Y(y)dy\Big) = \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

□

On the other hand, uncorrelated does not imply independence!

Non-Gaussian condition

In any ICA algorithm, there is a fundamental restriction that the underlying signals $s_1, \ldots, s_N$ cannot be Gaussian random variables, i.e., their probability distribution is not the normal/Gaussian distribution $N(\mu_i, \sigma_i)$.

Why? Suppose the mixing matrix A is orthogonal, i.e., a rotation. For two independent random variable s_1 and $s_2 \sim N(0,1)$, their joint probability distribution is

$$p(s_1, s_2) = p(s_1)p(s_2) = \frac{1}{2\pi} \exp\Big(- \frac{s_1^2 + s_2^2}{2} \Big) = \frac{1}{2\pi} \exp\Big(- \frac{\|\boldsymbol{s}\|^2}{2} \Big).$$

Under the orthogonal matrix A, we have $\|A\boldsymbol{s}\| = \|\boldsymbol{s}\|$, which means the probability distribution does not change under A.

This implies that there is no information on the directions of the columns of A, and hence we cannot estimate A.

Therefore, within the ICA algorithm, we should make it so that the computed output $\widehat{\boldsymbol{s}} = W\boldsymbol{x}$ should have a non-Gaussian distribution by maximising certain measures of non-Gaussianity.

Standardised moments

Let X be a random variable with a probability distribution p and mean $\mu = \mathbb{E}(X)$. The standardised moment of degree k is the ratio

$$\tilde{\mu}_k = \frac{\mu_k}{\sigma^k},$$

where the kth moment about the mean is

$$\mu_k := \mathbb{E}((X - \mu)^k) = \int_{-\infty}^{\infty} (x - \mu)^k p(x) dx,$$

and the kth power of the standard deviation is

$$\sigma^k := \sqrt{\mathbb{E}((X - \mu)^2)}^k.$$

Note: $\tilde{\mu}_1 = 0$ (1st standardised moment is always zero), while $\tilde{\mu}_2 = 1$ (2nd standardised moment is 1).

Definitions: We call $\tilde{\mu}_3$ the skewness and $\tilde{\mu}_4$ the kurtosis.

Note: by definition of the mean and variance, $\mu_1 = 0$ and $\mu_2 = \text{Var}(X) = \sigma^2$.

Class exercise

Consider a random variable X with the normal probability distribution $p_X(x)$

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big).$$

Show that

- the skewness $\tilde{\mu}_3 = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$ is zero.
- the kurtosis $\tilde{\mu}_4 = \frac{\mathbb{E}((X-\mu)^4)}{\sigma^4}$ is 3.

Hint: Use a suitable substitution, integration by parts, and the following fact:

$$\int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi}.$$

Kurtosis I

The Kurtosis is the 4th standardised moment

$$\text{Kurt}(X) = \frac{\mathbb{E}((X-\mu)^4)}{(\mathbb{E}((X-\mu)^2))^2} = \mathbb{E}\Big[\big(\frac{X-\mu}{\sigma}\big)^4\Big]$$

is a measure of the "tailedness" of the probability distribution p_X of X. Equivalently, it is a measure of outliers in the distribution.

Since Gaussian distribution has kurtosis 3, it is common to define the excess kurtosis

$$\text{EKurt}(X) = \text{Kurt}(X) - 3 = \frac{\mu_4}{\sigma^4} - 3.$$

Then, we say a probability distribution p_X is

- mesokurtic if $\text{EKurt}(X) = 0$.
- leptokurtic if $\text{EKurt}(X) > 0$ ("Lepto-" means slender and so distributions have "fatter tails", aka supergaussian).
- platykurtic if $\text{EKurt}(X) < 0$ ("Platy-" means broader and so distributions have "thinner tails", aka subgaussian).

Note: there is an inconsistency in the literature where the word "kurtosis" is often associated to the formula $\mathbb{E}(X^4) - 3(\mathbb{E}(X^2))$, which is the excess kurtosis in this course!

Kurtosis II

Graphically:

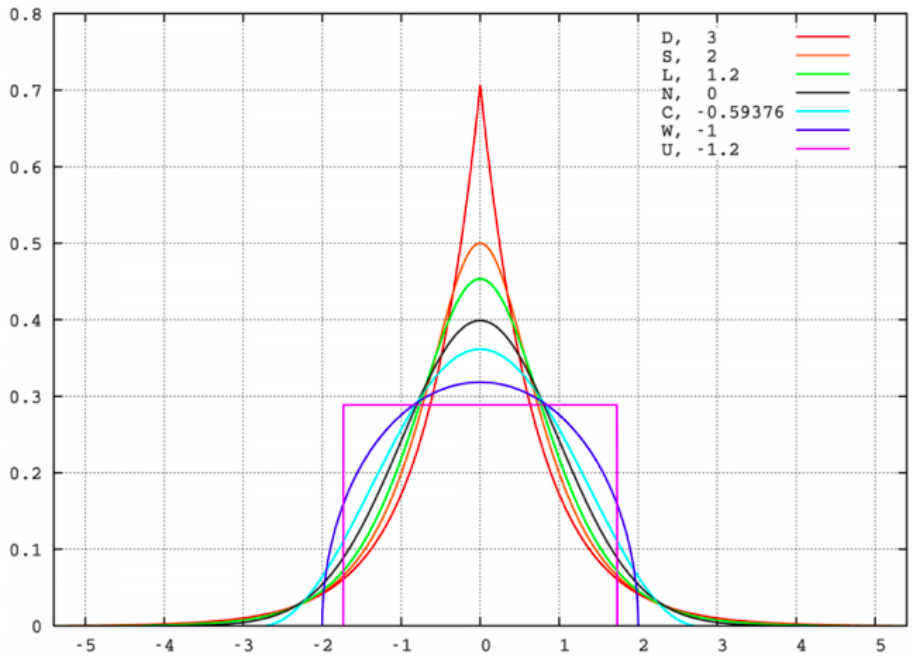
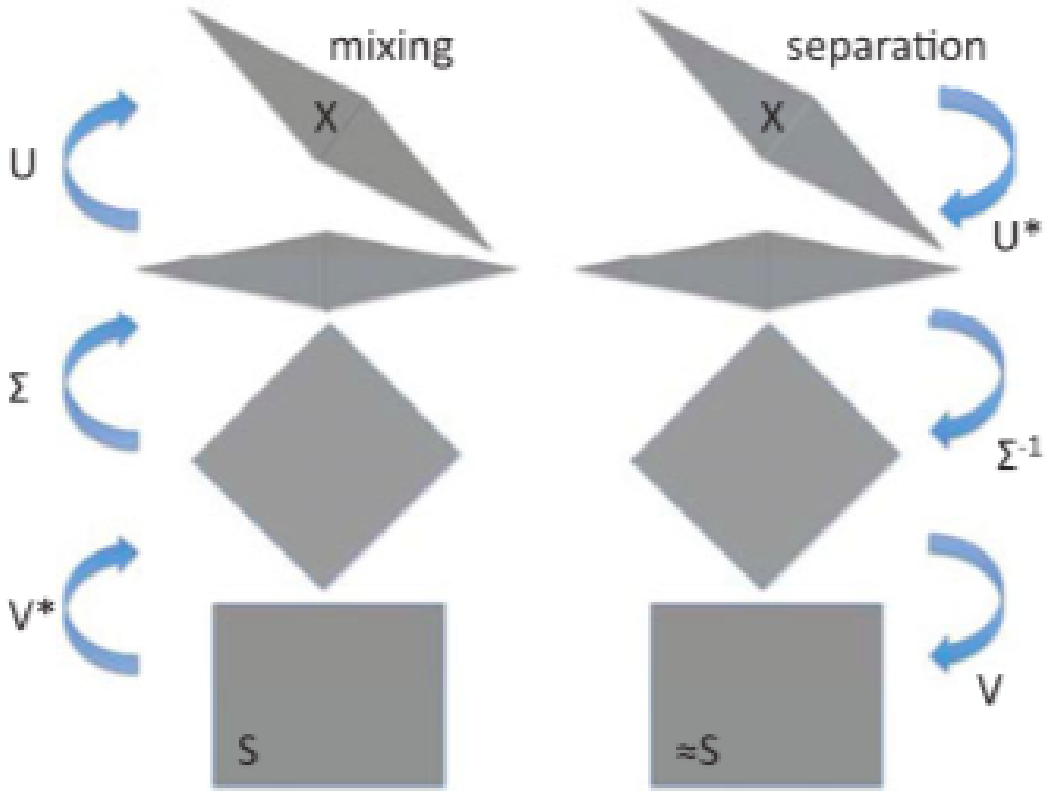


Figure: Taken from `https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/statistics-definitions/kurtosis-leptokurtic-platykurtic/`.

- Platykurtic distributions (cyan, blue, purple) have tails that are "thinner" compared to the normal distribution, or in some cases, non-existent.
- Leptokurtic distributions (red, orange, green) have tails that are "fatter" than the normal distribution.

Two-by-two case SVD-based ICA (Farid and Adelson)

Suppose we have two signals $\boldsymbol{x} = (x_1, x_2)$ and two sources $\boldsymbol{s} = (s_1, s_2)$. Assuming the mixing matrix A is invertible/full rank, visually the mixing and separation process can be summarised with the help of the SVD of A as



The mixing matrix A first rotates the signals S with $V^\top$, then stretches to a parallelogram with Σ, and then rotate again with U to get the data X.

Step 1 - Recovering $U^\top$

Geometrically, we want to align the axes of the parallelogram with the standard axes. The orthogonal/rotation matrix U is of the form

$$U = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for some angle θ.

We assume the data X is composed of M points $\{(x_1(j), x_2(j))\}_{j=1}^{M}$, and we can extract the long and short axes of the parallelogram (as they correspond to the direction of maximal and minima variance, aka the principal components).

Let θ be the angle between the the long axis and the horizontal axis. For each data point $(x_1(j), x_2(j))$, under the action of $U^\top$ they become

$$\begin{pmatrix} z_1(j) \\ z_2(j) \end{pmatrix} = U^\top \begin{pmatrix} x_1(j) \\ x_2(j) \end{pmatrix} = \begin{pmatrix} x_1(j)\cos\theta + x_2(j)\sin\theta \\ -x_1(j)\sin\theta + x_2(j)\cos\theta \end{pmatrix}.$$

Recovering $U^\top$ II

The variance of the projected data $\{(z_1(j), 0)\}_{j=1}^M$ onto the horizontal axis is

$$\mathrm{Var}(\theta) = \sum_{j=1}^{M} |z_1(j)|^2 = \sum_{j=1}^{M} x_1(j)^2 \cos^2\theta + 2x_1(j)x_2(j)\cos\theta\sin\theta + x_2^2(j)\sin^2\theta.$$

The direction of maximal variance is given by the angle θ_* maximising this function, and the direction of minimal variance is given by the angle $\theta_* - \frac{\pi}{2}$.

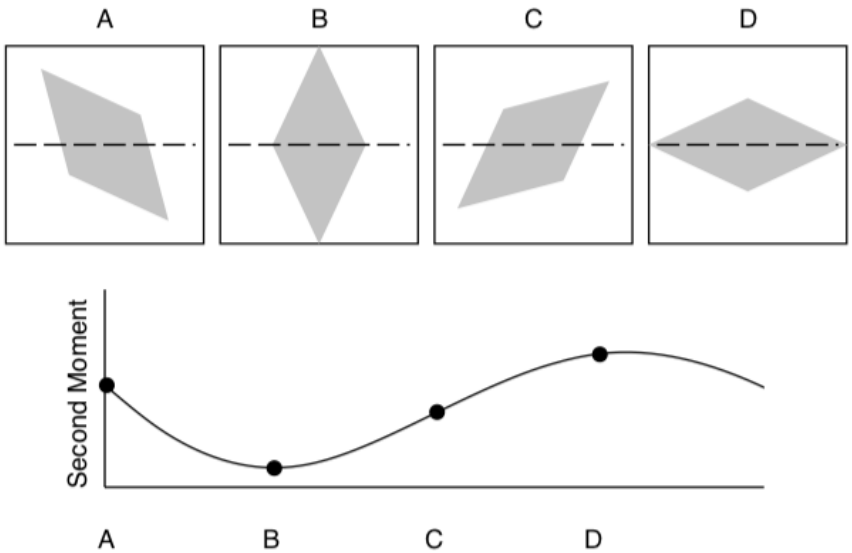


Figure: Variance projected onto horizontal axis. Taken from H. Farid and E.H. Adelson. J. Optical. Soc. America (1999)

Step 1 - Recovering $U^\top$ III

Differentiating $\mathrm{Var}(\theta)$ gives (class exercise)

$$\frac{d}{d\theta}\mathrm{Var}(\theta) = \sum_{j=1}^{M}\Big([x_2^2(j) - x_1^2(j)]\sin 2\theta + 2x_1(j)x_2(j)\cos 2\theta\Big).$$

Then,

$$\begin{aligned}\frac{d}{d\theta}\mathrm{Var}(\theta_*) = 0 \quad &\Leftrightarrow \quad \frac{\sin 2\theta_*}{\cos 2\theta_*} = -\frac{2\sum_{j=1}^{M} x_1(j)x_2(j)}{\sum_{j=1}^{M} x_2^2(j) - x_1^2(j)} \\ &\Leftrightarrow \quad \theta_* = \frac{1}{2}\tan^{-1}\left(-\frac{2\sum_{j=1}^{M} x_1(j)x_2(j)}{\sum_{j=1}^{M} x_2^2(j) - x_1^2(j)}\right).\end{aligned}$$

The orthogonal matrix $U^\top$, associated to the rotation of the parallelogram back to its aligned position, is given by

$$U^\top = \begin{pmatrix} \cos\theta_* & \sin\theta_* \\ -\sin\theta_* & \cos\theta_* \end{pmatrix}.$$

Step 2 - Recovering Σ^{-1}

Since the variance is associated to the singular values, we can estimate the singular values along the two directions as

$$\sigma_1^2 = \mathrm{Var}(\theta_*) = \sum_{j=1}^{N} \left[\begin{pmatrix} x_1(j) & x_2(j) \end{pmatrix} \begin{pmatrix} \cos\theta_* \\ \sin\theta_* \end{pmatrix} \right]^2,$$

$$\sigma_2^2 = \mathrm{Var}(\theta_* - \tfrac{\pi}{2}) = \sum_{j=1}^{M} \left[\begin{pmatrix} x_1(j) & x_2(j) \end{pmatrix} \begin{pmatrix} \cos(\theta_* - \frac{\pi}{2}) \\ \sin(\theta_* - \frac{\pi}{2}) \end{pmatrix} \right]^2.$$

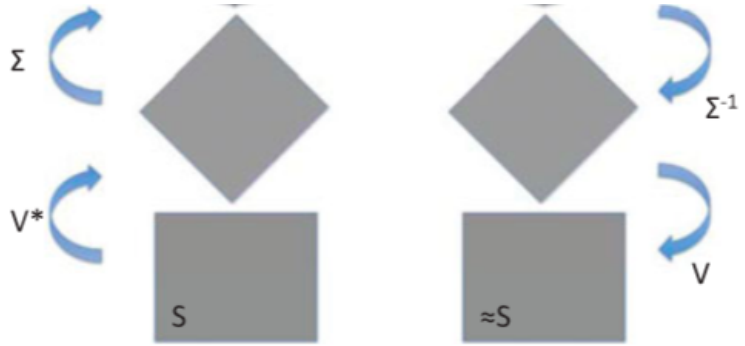
This gives us the diagonal elements of $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2)$. To undo this scaling, the inverse Σ^{-1} is

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix}.$$

This is well-defined, as A is assumed to be full-rank, and so σ_1 and σ_2 are positive!

Part 3 - Recovering V

The last step is to obtain the rotation matrix V, which is more subtle, as we need to produce nearly independent non-Gaussian probability distributions for s_1 and s_2.



For this we will use the kurtosis. From the data $\{(x_1(j), x_2(j))\}_{j=1}^{M}$, we denote the transformed data

$$\begin{pmatrix} y_1(j) \\ y_2(j) \end{pmatrix} = \Sigma^{-1} U^{\top} \begin{pmatrix} x_1(j) \\ x_2(j) \end{pmatrix}.$$

Suppose the rotation matrix V is of the form

$$V = \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix}$$

for some angle ψ. Under the action of V, we have

$$\begin{pmatrix} s_1(j) \\ s_2(j) \end{pmatrix} = V \begin{pmatrix} y_1(j) \\ y_2(j) \end{pmatrix} = \begin{pmatrix} y_1(j)\cos\psi + y_2(j)\sin\psi \\ -y_1(j)\sin\psi + y_2(j)\cos\psi \end{pmatrix}.$$

Recovering V II

We choose ψ_* as the angle maximising both the variance and the excess kurtosis of the first signal s_1. Graphically:

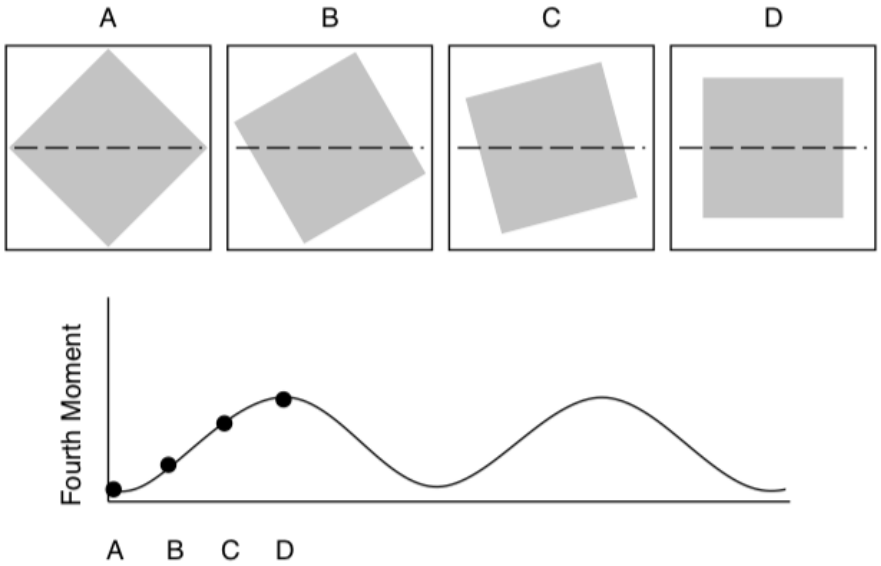


Figure: Fourth moment projected onto horizontal axis. Taken from H. Farid and E.H. Adelson. J. Optical. Soc. America (1999)

From the calculation of $U^\top$, maximising variance means that

$$\sum_{j=1}^{M} \left([y_2^2(j) - y_1^2(j)] \sin 2\psi_* + 2y_1(j)y_2(j) \cos 2\psi_* \right) = 0.$$

Keep this in mind!

Recovering V III

Since we assumed the variance of s_1 to be equal to 1, the excess kurtosis, as a function of ψ, is

$$\text{EKurt}(\psi) = \sum_{j=1}^{M} \Big(y_1(j)\cos\psi + y_2(j)\sin\psi \Big)^4 - 3.$$

Then, (see H. Farid, E.H. Adelson Separating Reflections from Images Using Independent Component Analysis)

$$\begin{aligned}\frac{d}{d\psi}\text{EKurt}(\psi) = \sum_{j=1}^{M} & -\tfrac{1}{8}y_1^4(j)(8\sin 2\psi + 4\sin 4\psi) + y_1^3(j)y_2(j)(2\cos 2\psi + 2\cos 4\psi) \\ & + 3y_1^2(j)y_2^2(j)\sin 4\psi + y_1(j)y_2^3(j)(2\cos 2\psi - 2\cos 4\psi) \\ & + \tfrac{1}{8}y_1^2(j)(8\sin 2\psi - 4\sin 4\psi).\end{aligned}$$

However, there is no analytical solution to $\frac{d}{d\psi}\text{EKurt}(\psi) = 0$! I.e., we cannot find a formula for ψ_*.

Recovering V III

To obtain an analytical solution, we instead maximise the normalised kurtosis

$$K(\psi) = \sum_{j=1}^{M} \frac{1}{y_1^2(j) + y_2^2(j)} \Big(y_1(j)\cos\psi + y_2(j)\sin\psi \Big)^4.$$

A shortish calculation shows

$$\begin{aligned}
\frac{dK}{d\psi} &= \sum_{j=1}^{M} [y_2^2(j) - y_1^2(j)] \sin 2\psi + 2y_1(j)y_2(j)\cos 2\psi \\
&\quad + \sum_{j=1}^{M} \frac{1}{y_1^2(j) + y_2^2(j)} \Big(\underbrace{[2y_1^3(j)y_2(j) - 2y_1(j)y_2^3(j)]}_{=:A(j)} \cos 4\psi \Big) \\
&\quad + \sum_{j=1}^{M} \frac{1}{y_1^2(j) + y_2^2(j)} \Big(\underbrace{[3y_1^2(j)y_2^2(j) - \tfrac{1}{2}y_1^4(j) - \tfrac{1}{2}y_2^4(j)]}_{=:B(j)} \sin 4\psi \Big).
\end{aligned}$$

By the variance maximisation, the red term vanishes! So, the angle ψ_* maximising $K(\psi)$ and $\mathrm{Var}(\psi)$ is given by

$$\psi_* = \tfrac{1}{4}\tan^{-1}\left[\frac{-\sum_{j=1}^{M} A(j)/(y_1^2(j) + y_2^2(j))}{\sum_{j=1}^{M} B(j)/(y_1^2(j) + y_2^2(j))} \right].$$

SVD-based ICA – Summary for $N = 2$ case

Given the data $\boldsymbol{x}$, the reconstructed signal $\boldsymbol{s}$ is

$$\begin{aligned}\boldsymbol{s} &= A^{-1}\boldsymbol{x} = V\Sigma^{-1}U^{\top}\boldsymbol{x} \\ &= \begin{pmatrix} \cos\psi_* & \sin\psi_* \\ -\sin\psi_* & \cos\psi_* \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_1} & 0 \\ 0 & \frac{1}{\sigma_2} \end{pmatrix} \begin{pmatrix} \cos\theta_* & \sin\theta_* \\ -\sin\theta_* & \cos\theta_* \end{pmatrix} \boldsymbol{x}\end{aligned}$$

where

$$\theta_* = \frac{1}{2}\tan^{-1}\left(-\frac{2\sum_{j=1}^{M} x_1(j)x_2(j)}{\sum_{j=1}^{N} x_2^2(j) - x_1^2(j)}\right),$$

$$\sigma_1 = \left(\sum_{j=1}^{M}\left[x_1(j)\cos\theta_* + x_2(j)\sin\theta_*\right]^2\right)^{1/2},$$

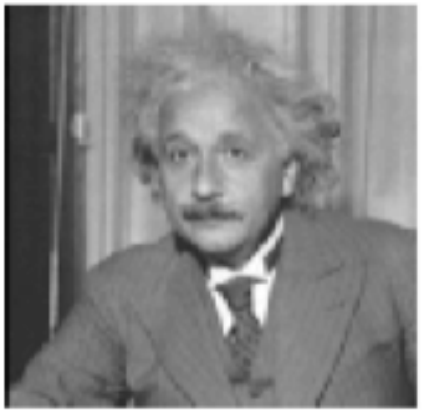
$$\sigma_2 = \left(\sum_{j=1}^{M}\left[x_1(j)\cos(\theta_* - \tfrac{\pi}{2}) + x_2(j)\sin(\theta_* - \tfrac{\pi}{2})\right]^2\right)^{1/2},$$

$$\psi_* = \frac{1}{4}\tan^{-1}\left(\frac{-\sum_{j=1}^{M} A(j)/(y_1^2(j) + y_2^2(j))}{\sum_{j=1}^{M} B(j)/(y_1^2(j) + y_2^2(j))}\right),$$

and $A(j), B(j), y_i(j)$ can be found in previous slides.

Numerical experiment (Farid and Adelson)

- Original signal: x_1 - image of Einstein, x_2 - image of Mandrill (a species of primate).

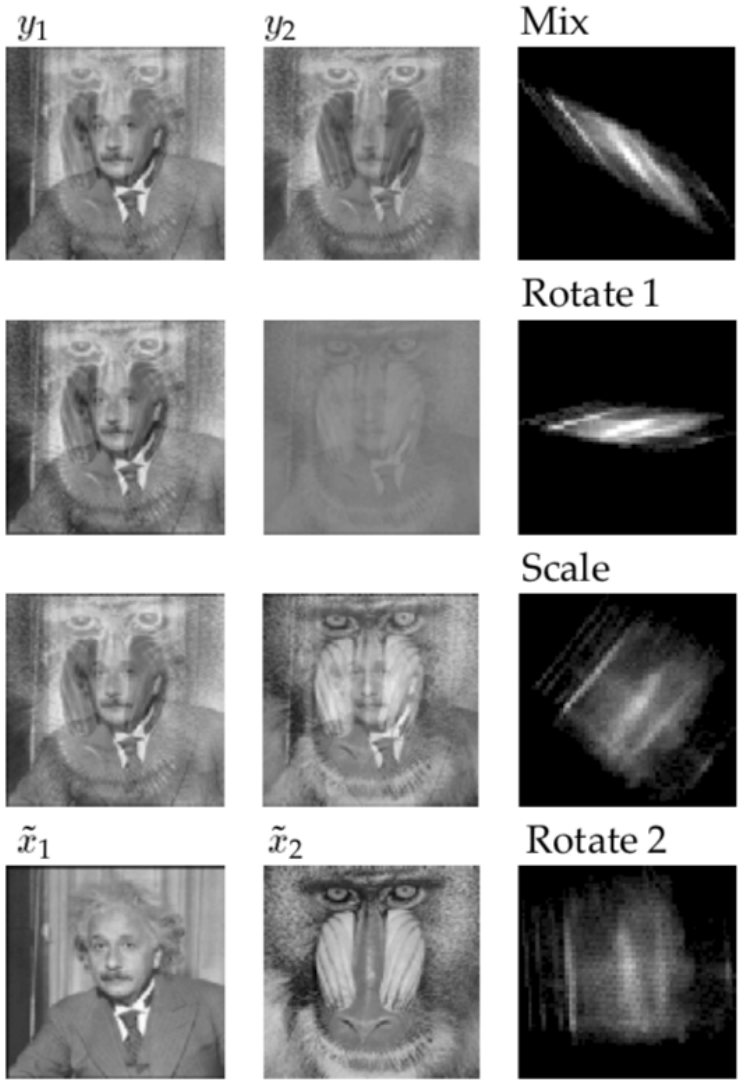



- Mixing matrix

$$A = \begin{pmatrix} 1.00 & -0.49 \\ 0.50 & -0.66 \end{pmatrix}$$

- Output observations: y_1 and y_2


Numerical experiment (Farid and Adelson)



Third column is the sampled joint probability distribution.

	Actual	Estimated
A	$\begin{pmatrix} 1.00 & -0.49 \\ 0.50 & -0.66 \end{pmatrix}$	$\begin{pmatrix} 1.00 & -0.63 \\ 0.49 & -0.79 \end{pmatrix}$
θ	35.7°	37.4°
σ_1/σ_2	4.41	4.55
ψ	35.4°	41.4°