Principal Component Analysis
Application to images

Václav Hlaváč

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics
Center for Machine Perception
http://cmp.felk.cvut.cz/~hlavac, hlavac@fel.cvut.cz

Outline of the lecture:

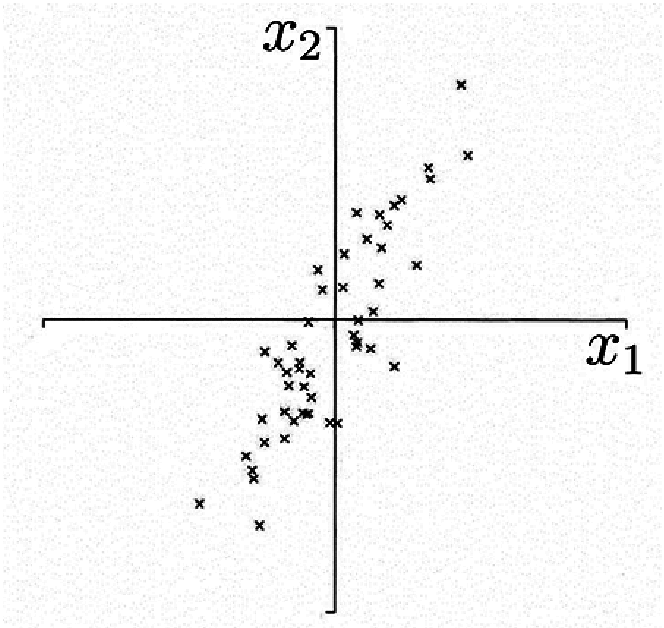
- Principal components, informal idea.
- Needed linear algebra.
- Least-squares approximation.
- PCA derivation, PCA for images.
- Drawbacks. Interesting behaviors live in manifolds.
- Subspace methods, LDA, CCA, ...

Eigen-analysis

- Seeks to represent observations, signals, images and general data in a form that enhances the mutual independence of contributory components.
- Very appropriate tools are provided by linear algebra.
- One observation or measurement is assumed to be a point in a linear space.
- This linear space has some ‘natural’ orthogonal basis vectors which allow data to be expressed as a linear combination with regards to the coordinated system induced by this base.
- These basis vectors are the eigen-vectors.

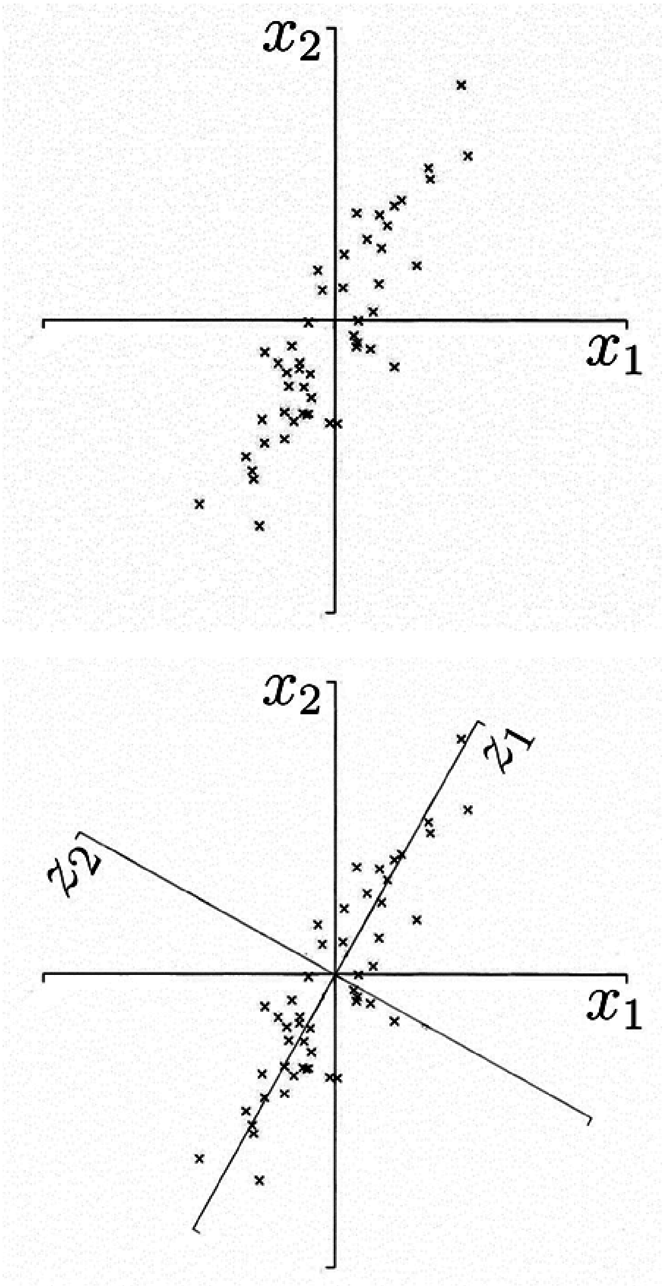
Geometric motivation, principal components (1)

- Two-dimensional vector space of observations, (x_1, x_2).
- Each observation corresponds to a single point in the vector space.
- Goal:
  Find another basis of the vector space which treats variations of data better.



Geometric motivation, principal components (2)

- Assume a single straight line approximating best the observation in the least-square sense, i.e. by minimizing the sum of distances between data points and the line.
- The first principal direction (component) is the direction of this line. Let it be a new basis vector z_1.
- The second principal direction (component, basis vector) z_2 is a direction perpendicular to z_1 and minimizing the distances to data points to a corresponding straight line.
- For higher dimensional observation spaces, this construction is repeated.



Eigen-numbers, eigen-vectors

- Assume a square $n \times n$ regular matrix A.
- Eigen-vectors are solutions of the eigen-equation

$$A\,\mathbf{x} = \lambda\,\mathbf{x}\,,$$

  where λ is called the eigen-value (which may be complex).
- We start reviewing eigen-analysis from a deterministic, linear algebra standpoint.
- Later, we will develop a statistical view based on covariance matrices and principal component analysis.

A system of linear equations, a reminder

- A system of linear equations can be expressed in a matrix form as $A\mathbf{x} = \mathbf{b}$, where A is the matrix of the system.

  Example:

$$\left.\begin{array}{rcrcrcr} x & + & 3y & - & 2z & = & 5 \\ 3x & + & 5y & + & 6z & = & 7 \\ 2x & + & 4y & + & 3z & = & 8 \end{array}\right\} \Longrightarrow A = \begin{bmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 5 \\ 7 \\ 8 \end{bmatrix}.$$

- The augmented matrix of the system is created by concatenating a column vector $\mathbf{b}$ to the matrix A, i.e., [A$|\mathbf{b}$].

$$\text{Example:} \qquad [A|\mathbf{b}] = \left[\begin{array}{rrr|r} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array}\right].$$

- This system has a unique solution if and only if the rank of the matrix A is equal to the rank of the extended matrix [A$|\mathbf{b}$].

Similar transformations of a matrix

- Let A be a regular matrix.
- Matrices A and B with real or complex entries are called similar if there exists an invertible square matrix P such that $P^{-1}A\,P = B$.
- Useful properties of similar matrices: they have the same rank, determinant, trace, characteristic polynomial, minimal polynomial and eigen-values (but not necessarily the same eigen-vectors).
- Similarity transformations allow us to express regular matrices in several useful forms, e.g. Jordan canonical form, Frobenius normal form.

Characteristic polynomial, eigen-values

- Let I be the unitary matrix having values 1 only on the main diagonal and zeros elsewhere.
- The polynomial of degree n given as $\det(A - \lambda I)$ is called the characteristic polynomial.
- Then the eigen-equation $A\,\mathbf{x} = \lambda\,\mathbf{x}$ holds if $\det(A - \lambda I) = 0$.
- The roots of the characteristic polynomial are the eigen-values λ.
- Consequently, A has n eigen-values which are not necessarily distinct—multiple eigen-values arise from multiple roots of the polynomial.

Jordan canonical form of a matrix

- Any complex square matrix is similar to a matrix in the Jordan canonical form

$$\begin{bmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_p \end{bmatrix}, \text{ where } J_i \text{ are Jordan blocks } \begin{bmatrix} \lambda_i & 1 & & 0 \\ 0 & \lambda_i & \ddots & 0 \\ 0 & 0 & \ddots & 1 \\ 0 & 0 & & \lambda_i \end{bmatrix},$$

  in which λ_i are the multiple eigen-values.
- The multiplicity of the eigen-value gives the size of the Jordan block.
- If the eigen-value is not multiple then the Jordan block degenerates to the eigen-value itself.

Least-square approximation

- Assume that abundant data comes from many observations or measurements. This case is very common in practice.
- We intent to approximate the data by a linear model - a system of linear equations, e.g. a straight line in particular.
- Strictly speaking, the observations are likely to be in a contradiction with respect to the system of linear equations.
- In the deterministic world, the conclusion would be that the system of linear equations has no solution.
- There is an interest in finding the solution to the system which is in some sense 'closest' to the observations, perhaps compensating for noise in observations.
- We will usually adopt a statistical approach by minimizing the least square error.

Principal component analysis, introduction

- PCA is a powerful and widely used linear technique in statistics, signal processing, image processing, and elsewhere.
- Several names: the (discrete) Karhunen-Loève transform (KLT, after Kari Karhunen and Michael Loève) or the Hotelling transform (after Harold Hotelling).
- In statistics, PCA is a method for simplifying a multidimensional dataset to lower dimensions for analysis, visualization or data compression.
- PCA represents the data in a new coordinate system in which basis vectors follow modes of greatest variance in the data.
- Thus, new basis vectors are calculated for the particular data set.
- The price to be paid for PCA's flexibility is in higher computational requirements as compared to, e.g., the fast Fourier transform.

Derivation, M-dimensional case (1)

- Suppose a data set comprising N observations, each of M variables (dimensions). Usually $N \gg M$.
- The aim: to reduce the dimensionality of the data so that each observation can be usefully represented with only L variables, $1 \leq L < M$.
- Data are arranged as a set of N column data vectors, each representing a single observation of M variables: the n-th observations is a column vector $\mathbf{x}_n = (x_1, \ldots, x_M)^\top$, $n = 1, \ldots, N$.
- We thus have an $M \times N$ data matrix X. Such matrices are often huge because N may be very large: this is in fact good, since many observations imply better statistics.

Data normalization is needed first

- This procedure is not applied to the raw data R but to normalized data X as follows.
- The raw observed data is arranged in a matrix R and the empirical mean is calculated along each row of R. The result is stored in a vector $\mathbf{u}$ the elements of which are scalars

$$u(m) = \frac{1}{N} \sum_{n=1}^{N} R(m,n) \,, \quad \text{where } m = 1, \ldots, M \,.$$

- The empirical mean is subtracted from each column of R: if $\mathbf{e}$ is a unitary vector of size N (consisting of ones only), we will write

$$X = R - \mathbf{u}\,\mathbf{e} \,.$$

Derivation, M-dimensional case (2)

If we approximate X in a lower dimensional space M by the lower dimensional matrix Y (of dimension L), then the mean square error ε^2 of this approximation is given by

$$\varepsilon^2 = \frac{1}{N} \sum_{n=1}^{N} |\mathbf{x}_n|^2 - \sum_{i=1}^{L} \mathbf{b}_i^\top \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \, \mathbf{x}_n^\top \right) \mathbf{b}_i \ ,$$

where $\mathbf{b}_i$, $i = 1, \ldots, L$ are basis vector of the linear space of dimension L. If ε^2 is to be minimized then the following term has to be maximized

$$\sum_{i=1}^{L} \mathbf{b}_i^\top \operatorname{cov}(\mathbf{x}) \, \mathbf{b}_i \ , \quad \text{where} \ \operatorname{cov}(\mathbf{x}) = \sum_{n=1}^{N} \mathbf{x}_n \, \mathbf{x}_n^\top \ ,$$

is the covariance matrix.

Approximation error

- The covariance matrix $\operatorname{cov}(\mathbf{x})$ has special properties: it is real, symmetric and positive semi-definite.
- So the covariance matrix can be guaranteed to have real eigen-values.
- Matrix theory tells us that these eigen-values may be sorted (largest to smallest) and the associated eigen-vectors taken as the basis vectors that provide the maximum we seek.
- In the data approximation, dimensions corresponding to the smallest eigen-values are omitted. The mean square error ε^2 is given by

$$\varepsilon^2 = \operatorname{trace}\big(\operatorname{cov}(\mathbf{x})\big) - \sum_{i=1}^{L} \lambda_i = \sum_{i=L+1}^{M} \lambda_i \,,$$

where $\operatorname{trace}(A)$ is the *trace*—sum of the diagonal elements—of the matrix A. The trace equals the sum of all eigenvalues.

Can we use PCA for images?

- It took a while to realize (Turk, Pentland, 1991), but yes.
- Let us consider a 321×261 image.

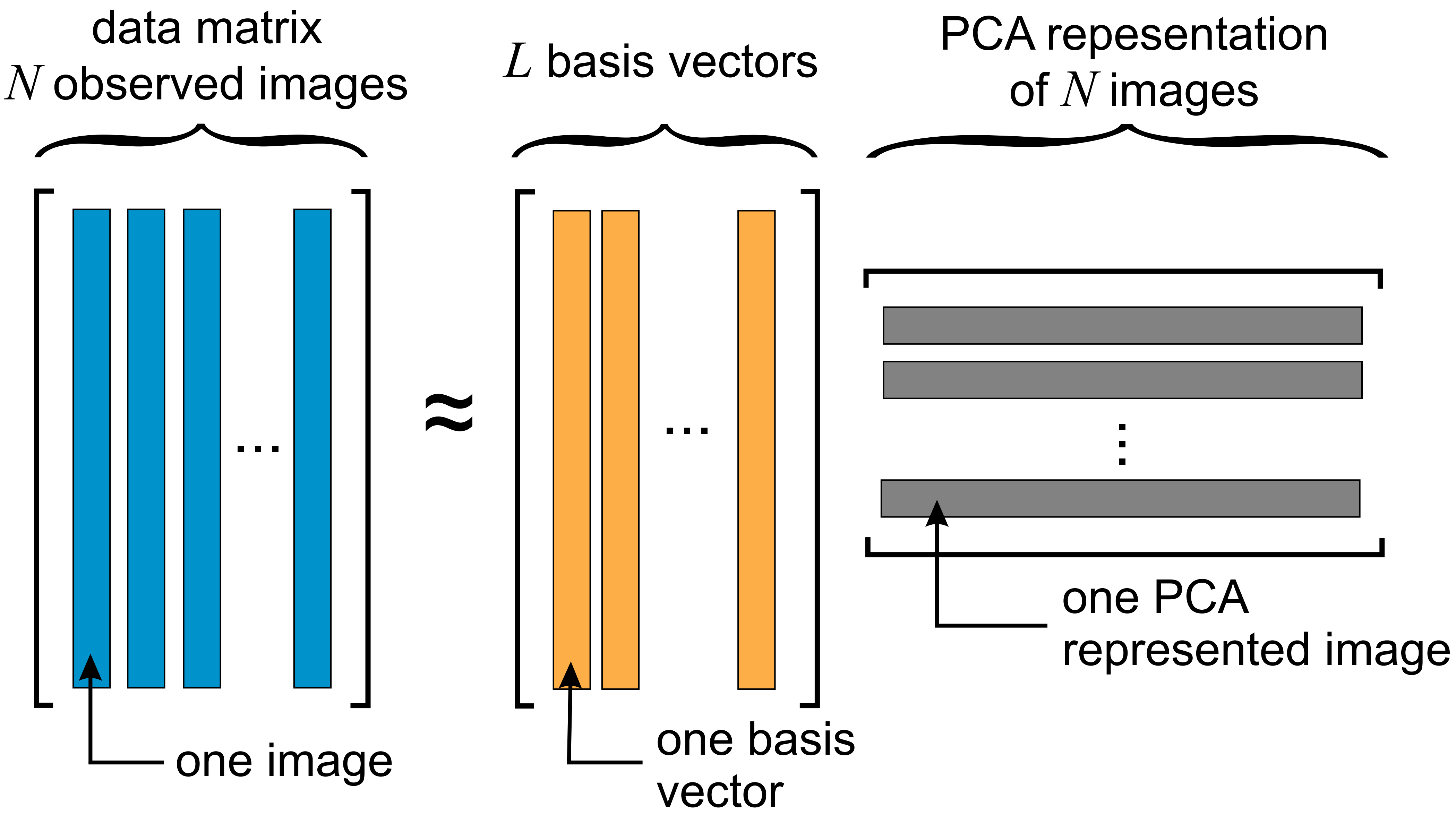
- The image is considered as a very long 1D vector by concatenating image pixels column by column (or alternatively row by row), i.e. $321 \times 261 = 83781$.
- The huge number 83781 is the dimensionality of our vector space.
- The intensity variation is assumed in each pixel of the image.

What if we have 32 instances of images?

Fewer observations than unknowns, and what?

- We have only 32 observations and 83781 unknowns in our example!
- The induced system of linear equations is not over-constrained but under-constrained.
- PCA is still applicable.
- The number of principle components is less than or equal to the number of observations available (32 in our particular case). This is because the (square) covariance matrix has a size corresponding to the number of observations.
- The eigen-vectors we derive are called eigen-images, after rearranging back from the 1D vector to a rectangular image.
- Let us perform the dimensionality reduction from 32 to 4 in our example.

PCA, graphical illustration



Approximation by 4 principal components only

- Reconstruction of the image from four basis vectors $\mathbf{b}_i$, $i = 1, \ldots, 4$ which can be displayed as images.
- The linear combination was computed as $q_1 \mathbf{b}_1 + q_2 \mathbf{b}_2 + q_3 \mathbf{b}_3 + q_4 \mathbf{b}_4 = 0.078\,\mathbf{b}_1 + 0.062\,\mathbf{b}_2 - 0.182\,\mathbf{b}_3 + 0.179\,\mathbf{b}_4$.

Reconstruction fidelity, 4 components

Reconstruction fidelity, original

PCA drawbacks, the images case

- By rearranging pixels column by column to a 1D vector, relations of a given pixel to pixels in neighboring rows are not taken into account.
- Another disadvantage is in the global nature of the representation; small change or error in the input images influences the whole eigen-representation. However, this property is inherent in all linear integral transforms.

Data (images) representations

Reconstructive (also generative) representation

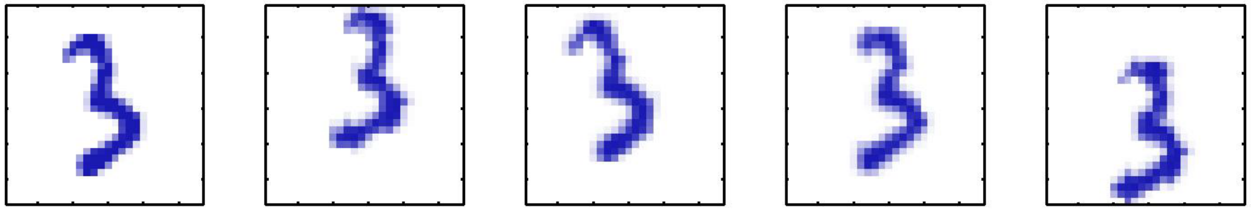
- Enables (partial) reconstruction of input images (hallucinations).
- It is general. It is not tuned for a specific task.
- Enables closing the feedback loop, i.e. bidirectional processing.

Discriminative representation

- Does not allow partial reconstruction.
- Less general. A particular task specific.
- Stores only information needed for the decision task.

Dimensionality issues, low-dimensional manifolds

- Images, as we saw, lead to enormous dimensionality.
- The data of interest often live in a much lower-dimensional subspace called the manifold.
- Example (courtesy Thomas Brox):
  The 100×100 image of the number 3 shifted and rotated, i.e. there are only 3 degrees of variations.



All data points live in a 3-dimensional manifold of the 10,000-dimensional observation space.

- The difficulty of the task is to find out empirically from the data in which manifold the data vary.

Subspace methods

Subspace methods explore the fact that data (images) can be represented in a subspace of the original vector space in which data live.

Different methods examples:

Method (abbreviation)	***Key property***
Principal Component Analysis (PCA)	reconstructive, unsupervised, optimal reconstruction, minimizes squared reconstruction error, maximizes variance of projected input vectors
Linear Discriminative Analysis (LDA)	discriminative, supervised, optimal separation, maximizes distance between projection vectors
Canonical Correlation Analysis (CCA)	supervised, optimal correlation, motivated by regression task, e.g. robot localization
Independent Component Analysis (ICA)	independent factors
Non-negative matrix factorization (NMF)	non-negative factors
Kernel methods for nonlinear extension	local straightening by kernel functions