Machine Intelligence 2

1.1 Principal Component Analysis

Prof. Dr. Klaus Obermayer

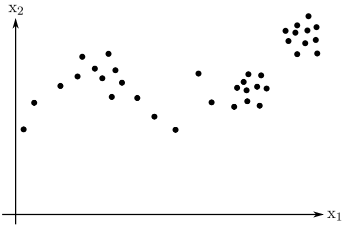
Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Preliminaries

Projection methods & clustering

observations: $\{\underline{\mathbf{x}}^{(\alpha)}\}, \alpha = 1, \ldots, p; \quad \underline{\mathbf{x}} \in \mathbb{R}^N$



What is the relevant "structure"?

$\Rightarrow$ projection methods: search for "interesting" directions in feature space

$\Rightarrow$ clustering methods: grouping & categorization (and prototypes)

The iris data

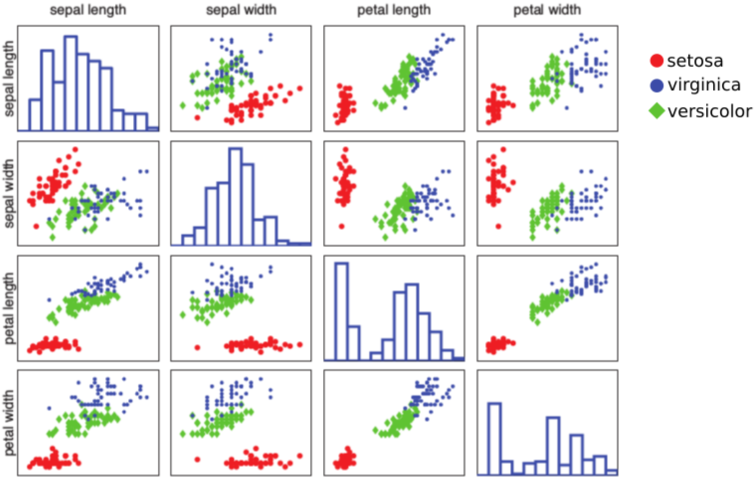
setosa

versicolor

virginica

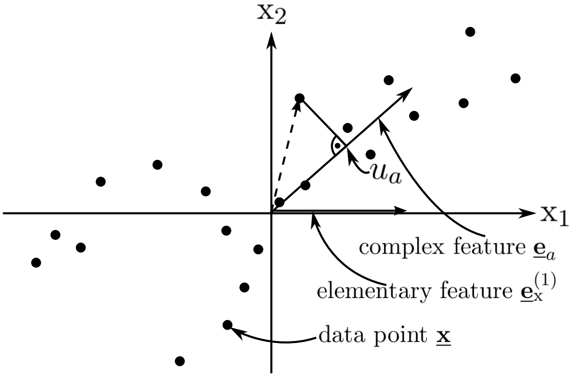
Source: http://www.statlab.uni-heidelberg.de/data/iris/. Used with kind permission of Dennis Kramb and SIGNA.

The iris data: scatter plot



Source: Machine Learning: A Probabilistic Perspective, By Kevin P. Murphy

"Complex" features



- elementary features: vectors $\underline{\mathbf{e}}_x^{(1)}, \underline{\mathbf{e}}_x^{(2)}, \underline{\mathbf{e}}_x^{(3)}, \ldots \underline{\mathbf{e}}_x^{(N)}$ with $\|\underline{\mathbf{e}}_x^{(i)}\|_2 = 1$
- complex feature: $\underline{\mathbf{e}}_a$ (*direction* in feature space) with $\|\underline{\mathbf{e}}_a\|_2 = 1$
- feature value $u_a(\underline{\mathbf{x}}) = \underline{\mathbf{e}}_a^T \cdot \underline{\mathbf{x}}$

Moments of the data: information wrt. location & shape

first moment (sample mean/center of mass):

$$\underline{\mathbf{m}} = \frac{1}{p} \sum_{\alpha=1}^{p} \underline{\mathbf{x}}^{(\alpha)}$$

second moments (covariance matrix):

$$\underline{\mathbf{C}} = \left\{ C_{ij} \right\} \quad \text{with} \quad C_{ij} = \frac{1}{p} \sum_{\alpha=1}^{p} \left(\mathrm{x}_i^{(\alpha)} - m_i \right) \left(\mathrm{x}_j^{(\alpha)} - m_j \right)$$

for "centered" data ($\underline{\mathbf{m}} = \underline{\mathbf{0}}$) this reads:

$$C_{ij} = \frac{1}{p} \sum_{\alpha=1}^{p} \mathrm{x}_i^{(\alpha)} \mathrm{x}_j^{(\alpha)}$$

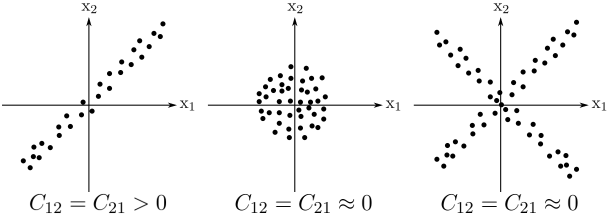
Properties of the covariance matrix

Covariance matrix $\underline{\mathbf{C}} = \{C_{ij}\}$ with $C_{ij} = \frac{1}{p} \sum_{\alpha=1}^{p} \left(\mathrm{x}_i^{(\alpha)} - m_i\right)\left(\mathrm{x}_j^{(\alpha)} - m_j\right)$

$C_{ij} = C_{ji}$ symmetry

$i = j$ $\quad C_{ii} = \frac{1}{p} \sum_{\alpha=1}^{p} \left(\mathrm{x}_i^{(\alpha)} - m_i\right)^2$ $\quad \rightsquigarrow$ variance of variable x_i

$i \neq j$ $\quad C_{ij}: \rightsquigarrow$ covariances



Note: $C_{ij} = 0 \Rightarrow$ variables are uncorrelated BUT might be dependent.

Moments for complex features $\underline{\mathbf{e}}_a$

Mean

$$m_a \quad = \quad \frac{1}{p}\sum_{\alpha=1}^{p} u_a^{(\alpha)} \quad = \quad \frac{1}{p}\sum_{\alpha=1}^{p} \underline{\mathbf{e}}_a^T \cdot \underline{\mathbf{x}}^{(\alpha)} \quad = \underline{\mathbf{e}}_a^T \cdot \underline{\mathbf{m}}$$

Variance

$$\sigma_a^2 \quad = \quad \tfrac{1}{p}\sum_{\alpha=1}^{p} \left(u_a^{(\alpha)} - m_a\right)^2 = \quad \underline{\mathbf{e}}_a^T \underline{\underline{\mathbf{C}}\mathbf{e}}_a$$

See blackboard

$\Rightarrow$ $\mathbf{C}$ determines the variance of the data along every possible direction.

Principal Component Analysis (PCA)

Karhunen-Loève transform

Principal Components (PCs)

"informative" directions

$$\underline{\mathbf{e}}_a^* = \underset{\underline{\mathbf{e}}_a}{\operatorname{argmax}} \left(\sigma_a^2\right) \qquad \text{with} \qquad \|\underline{\mathbf{e}}_a\|_2 = 1$$

Method of Lagrange multipliers λ

$$\underbrace{\underline{\mathbf{e}}_a^T \underline{\mathbf{C}} \underline{\mathbf{e}}_a}_{\text{objective}} - \lambda \underbrace{\left(\underline{\mathbf{e}}_a^T \underline{\mathbf{e}}_a - 1\right)}_{\text{constraints}} \stackrel{!}{=} \max \qquad \text{See blackboard}$$

eigenvalue problem

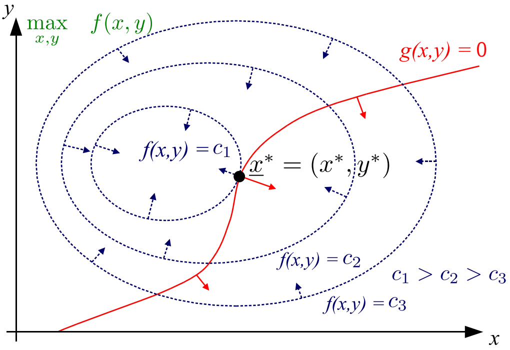
$$\underline{\mathbf{C}} \underline{\mathbf{e}}_a = \lambda \underline{\mathbf{e}}_a$$

$\Rightarrow$ **Principal Components**: normalized eigenvectors $\underline{\mathbf{e}}_a$ of $\underline{\mathbf{C}}$

$\Rightarrow$ The variance along a PC is given by the corresponding eigenvalue

$$\sigma_a^2 = \underline{\mathbf{e}}_a^T \underline{\mathbf{C}} \underline{\mathbf{e}}_a = \lambda \underline{\mathbf{e}}_a^2 = \lambda_a$$

Lagrange multipliers



$$f(x, y) \stackrel{!}{=} \max \qquad \text{and} \qquad g(x, y) = 0$$

at the optimal (x^*, y^*), gradients are (anti)-parallel

$$L_{(x,y;\lambda)} \stackrel{!}{=} f(x, y) + \lambda g(x, y)$$

$$\nabla L = 0 \rightarrow \nabla f = -\lambda \nabla g,$$

Properties of the Principal Components

Covariance matrix $\underline{\mathbf{C}} = \left\{ C_{ij} \right\}$ with $C_{ij} = \frac{1}{p} \sum_{\alpha=1}^{p} \left(\mathrm{x}_i^{(\alpha)} - m_i \right) \left(\mathrm{x}_j^{(\alpha)} - m_j \right)$

$$\underline{\mathbf{C}}\underline{\mathbf{e}}_a = \lambda \underline{\mathbf{e}}_a$$

1. $\underline{\mathbf{C}}_{N \times N}$ is real and symmetric $\Rightarrow$ orthonormal basis of N eigenvectors

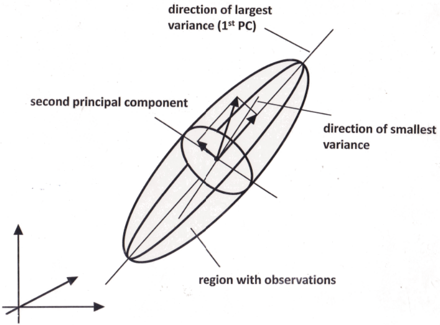
$$\underline{\mathbf{e}}_i^T \cdot \underline{\mathbf{e}}_j = \delta_{ij}$$

2. $\underline{\mathbf{C}}$ is diagonal w.r.t. its eigenbasis, let $\underline{\mathbf{M}} = \left(\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \ldots, \underline{\mathbf{e}}_N \right)$:

$$\underline{\mathbf{M}}^T \underline{\mathbf{C}} \underline{\mathbf{M}} = \widehat{\underline{\mathbf{C}}} = \mathrm{diag}(\underline{\lambda}) = \underline{\mathbf{\Lambda}}$$

$\Rightarrow$ transformation into the eigenbasis yields uncorrelated features
$\Rightarrow$ useful as a preprocessing step ($\rightsquigarrow$ regression, classification)

Properties of the Principal Components



❸ ordering of principal components w.r.t. variance

$$
\begin{array}{ccccccccccc}
\lambda_1 & > & \lambda_2 & > & \lambda_3 & > & \ldots\ldots & > & \lambda_{N-1} & > & \lambda_N \\
\downarrow & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\underline{\mathbf{e}}_1 & & \underline{\mathbf{e}}_2 & & \underline{\mathbf{e}}_3 & & & & \underline{\mathbf{e}}_{N-1} & & \underline{\mathbf{e}}_N
\end{array}
$$

$\boxed{\text{direction of largest variance}} \longrightarrow \boxed{\text{direction of smallest variance}}$

$\underline{\mathbf{e}}_j$: direction of largest variance in the subspace spanned by $\underline{\mathbf{e}}_i, i \geq j$

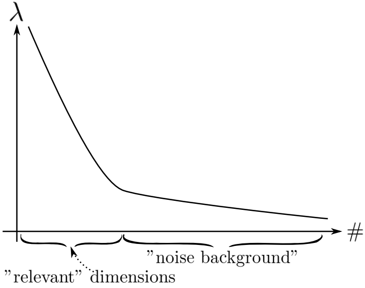
Optimal dimensionality reduction

Representation of $\underline{\mathbf{x}}$ in the basis of Principal Components:

$$\underline{\mathbf{x}} = \underbrace{a_1}_{\underline{\mathbf{e}}_1^T \underline{\mathbf{x}}} \underline{\mathbf{e}}_1 + \underbrace{a_2}_{\underline{\mathbf{e}}_2^T \underline{\mathbf{x}}} \underline{\mathbf{e}}_2 + \ldots + \underbrace{a_N}_{\underline{\mathbf{e}}_N^T \underline{\mathbf{x}}} \underline{\mathbf{e}}_N$$

Reconstruction via projection onto the first M Principal Components

$$\underline{\widetilde{\mathbf{x}}} = a_1 \underline{\mathbf{e}}_1 + a_2 \underline{\mathbf{e}}_2 + \ldots + a_M \underline{\mathbf{e}}_M$$



$\Rightarrow$ compared to other M-dimensional projections, this yields a minimal approximation error E:

$$E = \frac{1}{p} \sum_{\alpha=1}^{p} e^{(\alpha)} \qquad e^{(\alpha)} = (\underline{\mathbf{x}}^{(\alpha)} - \underline{\widetilde{\mathbf{x}}}^{(\alpha)})^2 = \sum_{j=M+1}^{N} (a_j^{(\alpha)})^2$$

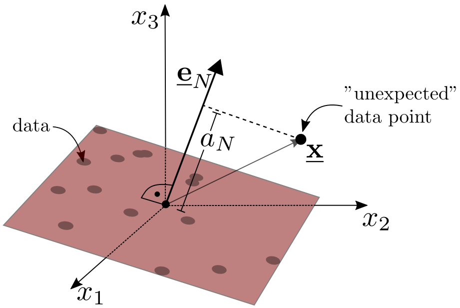
Whitening

- $\rightsquigarrow$ variance is scale sensitive (scaling one dimension can change all PCs)
- $\rightsquigarrow$ analysis of variances criterion only makes sense if scales are "comparable"
- $\rightsquigarrow$ incomparable scales $\rightarrow$ scale variance along all directions to 1 after decorrelation by PCA

$$\underline{\mathbf{v}}^{(\alpha)} = \underline{\mathbf{\Lambda}}^{-\frac{1}{2}} \underline{\mathbf{M}}^T \underline{\mathbf{x}}^{(\alpha)}$$

Outlier detection

Principal Components with smallest eigenvalues (e.g., $\underline{\mathbf{e}}_N$):

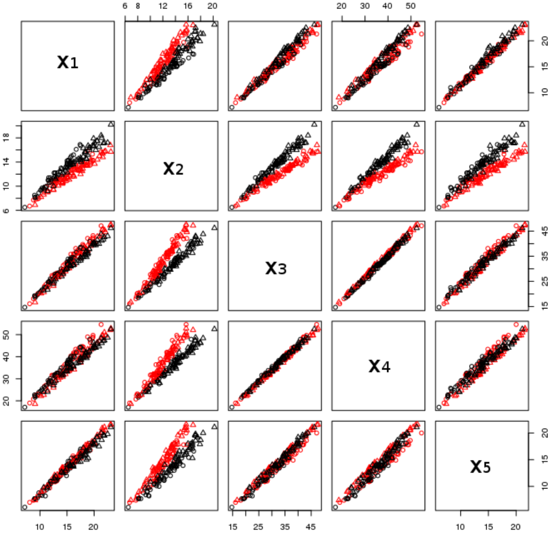


$\leadsto$ outliers / data with novel features can be identified by projecting to last PCs

Leptograpsus variegatus

Source: http://www.seafriends.org.nz/enviro/habitat/rscrust.htm

The Leptograpsus data: scatter plot



sex

black: female

red: male

species

△ : orange

○ : blue

attributes

X_1: frontal lobe size (mm)

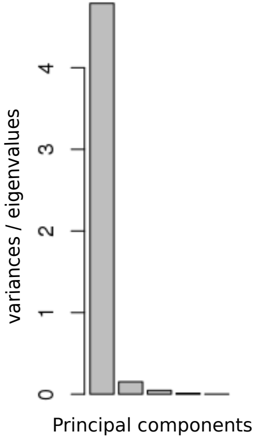
X_2: rear width (mm)

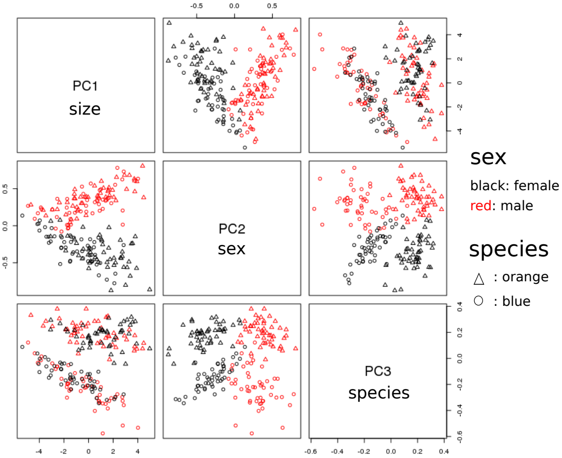
X_3: carapace length (mm)

X_4: carapace width (mm)

X_5: body depth (mm)

Application: Leptograpsus data



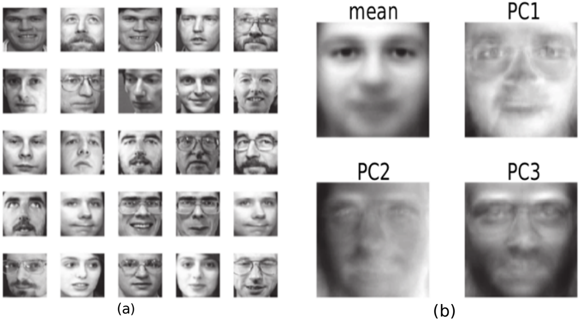


Latent factors

- the data may appear high dimensional, but there may only be a small number of features underlying variability
- dimensionality reduction: projection of the data into a low dimensional subspace which captures the "essence" of the data
- latent factors: remaining PCs with high variance

Application: eigenfaces

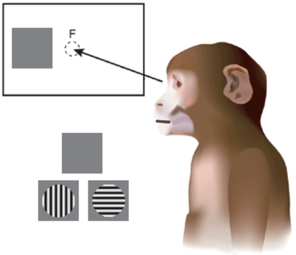
When modeling the appearance of face images, there may only be a few underlying latent factors which describe most of the variability, such as lighting, pose, identity, etc.

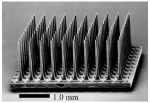


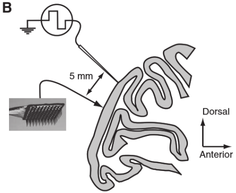
(a) 25 randomly chosen 64 x 64 pixel images from the Olivetti face database. (b) The mean and the first three principal component basis vectors (eigenfaces).

Source: Machine Learning: A Probabilistic Perspective, By Kevin P. Murphy. *Modified captions*.

Application: spiking activity in monkey visual cortex





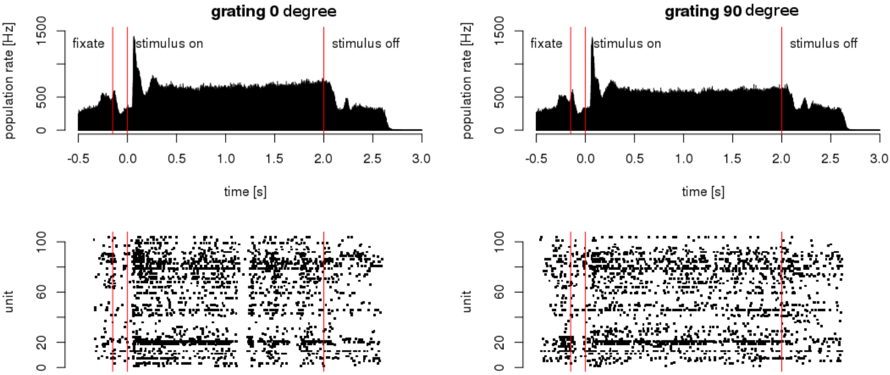


Protocol:

pre-trial → achieve fixation -150ms → stimulus 0-2000ms → post-trial

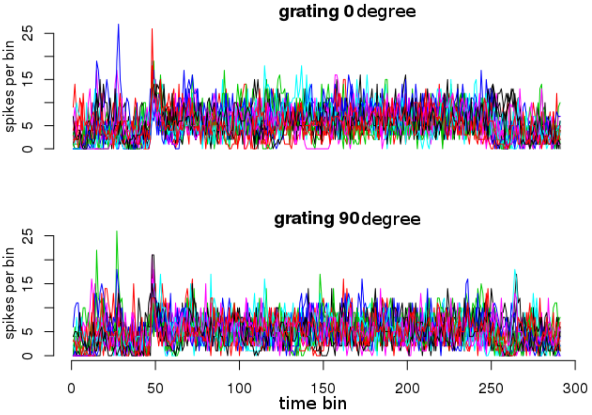
Taken from Kimura et al. 2007 and Smith & Kohn 2008

Application: spiking activity in monkey visual cortex



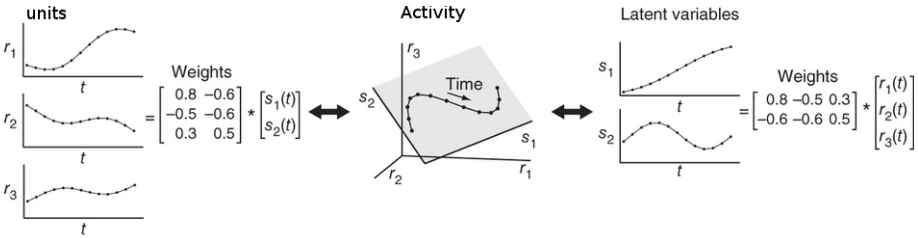
- stimulus driven component (onset & tuning)
- variability across trials
- strong diversity & rich spatiotemporal structure

Application: spiking activity in monkey visual cortex



- post stimulus time histograms
- each color represents one unit

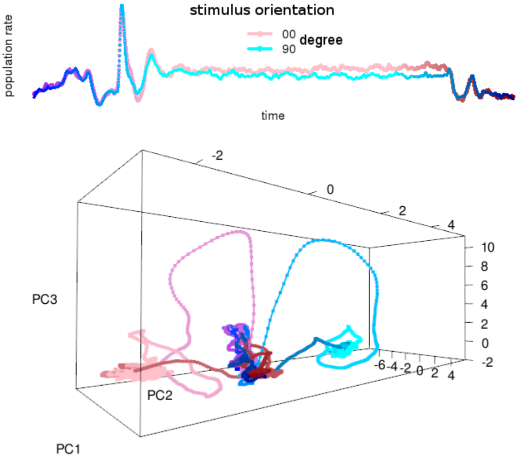
Application: spiking activity in monkey visual cortex

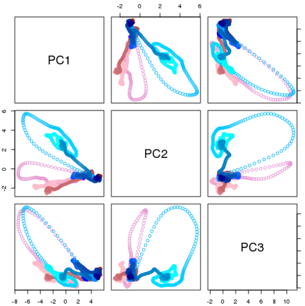


- 3 neurons: 3d space in which each axis represents the firing rate of a unit (r_1 , r_2 , and r_3).
- The rate vectors on a plane (shaded gray).

Taken from Cunningham & Yu. Nat. Neur.2014

Application: spiking activity in monkey visual cortex





Summary of PCA

- linear method for data preprocessing, dimensionality reduction & data compression
- uncorrelated features & whitening
- very large covariance matrices $\Rightarrow$ numerical instabilities
- efficient algorithms for the extraction of PCs with the largest eigenvalues $\Rightarrow$ EM, successive components via *power method*
- biologically inspired methods: Hebbian learning

extensions

- nonlinear features $\Rightarrow$ kernel PCA
- no underlying *generative model* $\Rightarrow$ probabilistic PCA, factor analysis