Computational Pattern Analysis and Statistical Learning

Lecture 4: Getting a grip on high-dimensional data

Tijl De Bie, Konstantin Tretyakov
(Largely based on joint work with Nello Cristianini and John Shawe-Taylor)

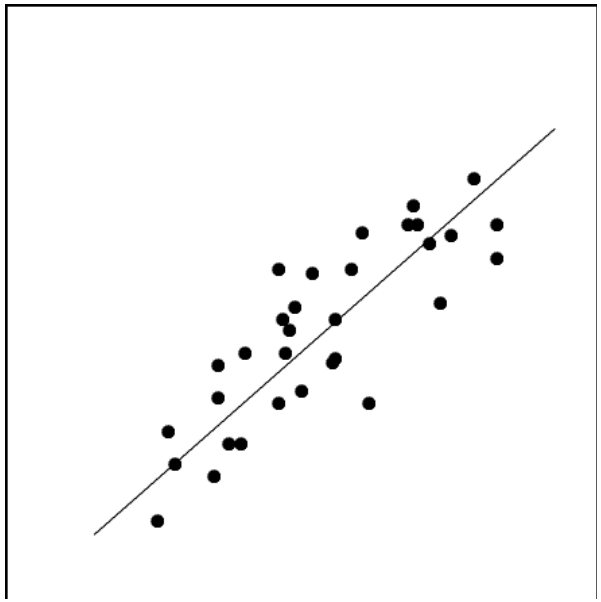
Tartu, Estonia

November 2006

- A framework for pattern analysis, emphasizing
  - pattern function – pattern space – pattern magnitude – capacity functional
  - pattern visualisation – pattern matching – pattern discovery
  - stability and significance analysis
  - fast search techniques
- Examples from patterns in discrete spaces
- Importance of patterns in vector spaces
  The remainder of this course will be about this subject (aka *machine learning*, *statistical learning theory*)

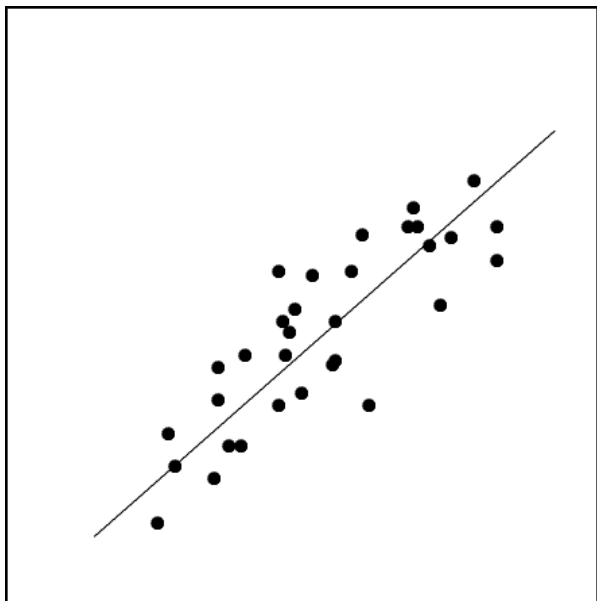
- This Lecture: unsupervised learning, in particular:
  - Clustering: explain data by defining a limited number of prototypes
  - Dimensionality reduction: explain data by defining a restricted subspace
- Start with dimensionality reduction...

Principal Component Analysis



- Often, high-dimensional data can be explained by few 'factors'
- A factor = an 'explanation' for the data
- Rest is 'noise', somehow irrelevant fluctuations
- How to capture these factors?

Principal Component Analysis

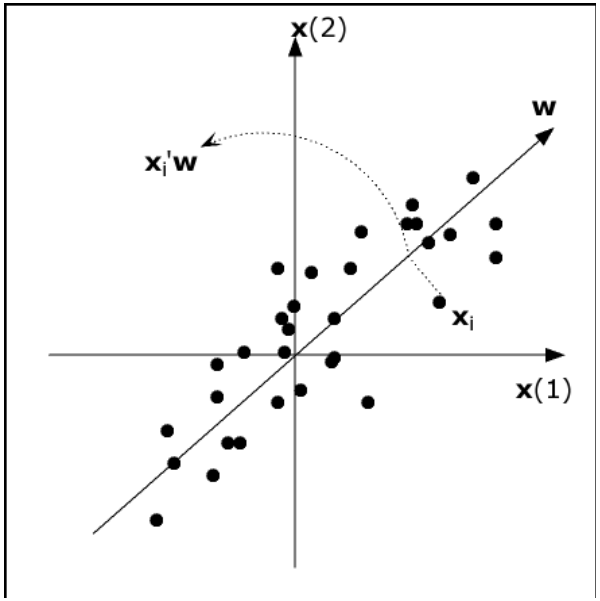


- Assume that the factors account for the largest part of the variance
- Search for directions in the data space, along which the data has a large variance

Principal Component Analysis

- Projections:

$$\mathbf{Xw} = \begin{pmatrix} \mathbf{x}_1'\mathbf{w} \\ \mathbf{x}_2'\mathbf{w} \\ \vdots \\ \mathbf{x}_n'\mathbf{w} \end{pmatrix}$$



- 'Variance along a direction $\mathbf{w}$':

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 = \frac{1}{n}\left\|\mathbf{Xw}\right\|^2 = \frac{1}{n}\left(\mathbf{Xw}\right)'\left(\mathbf{Xw}\right) = \frac{1}{n}\mathbf{w}'\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w}$$

Principal Component Analysis

- Define pattern function as the variance of the projection (inner product) with $\mathbf{w}$:

$$\pi_{\mathbf{w}}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 = \frac{1}{n}\mathbf{w}'\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w}$$

- Pattern discovery problem: maximise this over $\mathbf{w}$?
- Note: can grow large by multiplying $\mathbf{w}$ with a factor... not very meaningful
- Capacity constraint: define capacity functional

$$C(\pi_{\mathbf{w}}) = \|\mathbf{w}\|^2 = \mathbf{w}'\mathbf{w}$$

- Require $C(\pi_{\mathbf{w}}) \leq 1$ (for example)

Principal Component Analysis

- Hence, the pattern recognition problem becomes:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \frac{1}{n}\mathbf{w}'\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}'\mathbf{w} \leq 1 \end{aligned}$$

- (Result will be that $\mathbf{w}'\mathbf{w} = 1$)
- To solve: Lagrange multiplier $\lambda \geq 0$

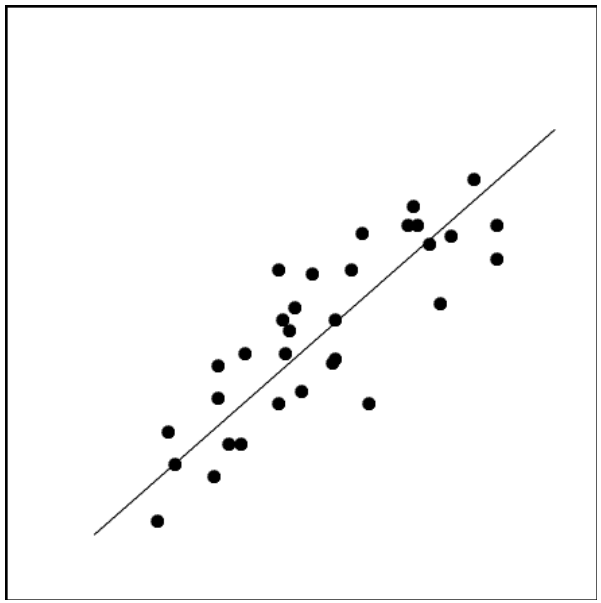
$$\max_{\mathbf{w}} \frac{1}{n}\mathbf{w}'\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w} - \lambda\left(\mathbf{w}'\mathbf{w} - 1\right)$$

- Take gradient w.r.t $\mathbf{w}$ and equate to $\mathbf{0}$:

$$\frac{1}{n}\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w} = \lambda\mathbf{w}$$

- This is an eigenvalue problem! (Followed by normalising $\mathbf{w}$, such that $\mathbf{w}'\mathbf{w} = 1$) Easy to solve...

Principal Component Analysis



- $\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w} = \lambda\mathbf{w}$
- Eigenvector corresponding to the largest eigenvalue maximises the variance:

$$\frac{1}{n}\frac{\mathbf{w}\left(\mathbf{X}'\mathbf{X}\right)\mathbf{w}}{\mathbf{w}'\mathbf{w}} = \lambda$$

- Other eigenvectors are orthogonal, and represent increasingly less important directions in the data

Principal Component Analysis

- Dimensionality reduction:
  - Find dominant eigenvalue/eigenvector pairs $(\mathbf{w}_k, \lambda_k)$, normalise $\mathbf{w}_k$
  - Project the data on these eigenvectors: $v_i(k) = \mathbf{x}_i' \mathbf{w}_k$
  - These $v_i(k)$ is the strength of the kth 'factor' in $\mathbf{x}_i$
  - In matrix notation with $\mathbf{W} = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_{d'} \end{pmatrix}$, and $\mathbf{V}(i,:) = \mathbf{v}_i'$:

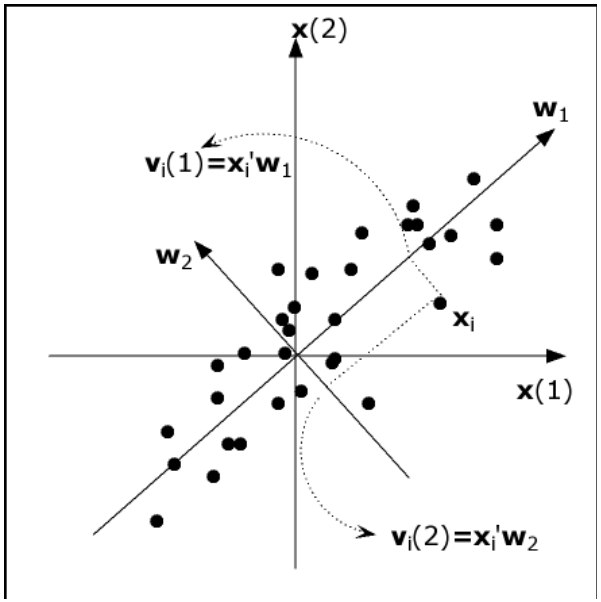
$$\mathbf{V} = \mathbf{X}\mathbf{W}$$

- Reconstruction based on this, equivalent to projecting $\mathbf{X}$ on the column space of $\mathbf{W}$:

$$\mathbf{X} \approx \mathbf{X}\mathbf{W}\mathbf{W}'$$

Principal Component Analysis

- For example, in this 2-D data set 2 factors can be identified (which fully explain the data)

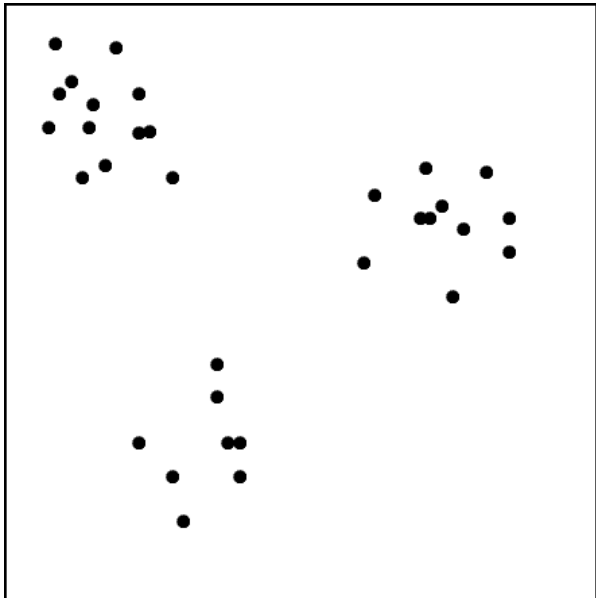


- In practice, usually used for hugely dimensional data, to visualise the structure in 2-D...

K-means clustering

- We have now identified several factors underlying the data
- Each factor explains one way of variation

- Here:
  - identify 'prototype' vectors $\boldsymbol{\mu}_k$
  - such that each data point $\mathbf{x}_i$ is close to one $\boldsymbol{\mu}_k$



K-means clustering

- Pattern function:

$$\pi_{\{\boldsymbol{\mu}_k\}}(\mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n}\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

- Note: if $\{\boldsymbol{\mu}_k\}$ would contain n elements (as many as the data set size), the pattern function can be made $= 0$
- This would be meaningless...
- So introduce capacity functional:

$$C\left(\pi_{\{\boldsymbol{\mu}_k\}}\right) = |\{\boldsymbol{\mu}_k\}|$$

K-means clustering

- Pattern recognition problem:

$$\begin{aligned} \min_{\{\boldsymbol{\mu}_k\}} \quad & \frac{1}{n}\sum_{i=1}^{n} \min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \\ \text{s.t.} \quad & |\{\boldsymbol{\mu}_k\}| \leq d' \end{aligned}$$

- (Result will be that $|\{\boldsymbol{\mu}_k\}| = d'$)
- Note so easy to solve...
- Combinatorial problem
- Still, we can do something

K-means clustering

- Start with initial guess for $\{\boldsymbol{\mu}_k\}$
- Iterative procedure
  - Assign data points $\mathbf{x}_i$ to centres $\boldsymbol{\mu}_k$ by solving $\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ for each:

$$k(i) = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

  - Solve $\min_{\{\boldsymbol{\mu}_k\}} \frac{1}{n} \sum_{i=1}^{n} \left\|\mathbf{x}_i - \boldsymbol{\mu}_{k(i)}\right\|^2$ with fixed assignments of points $\mathbf{x}_i$ to centres $\boldsymbol{\mu}_k$
- Guaranteed to converge
- Not necessarily globally...

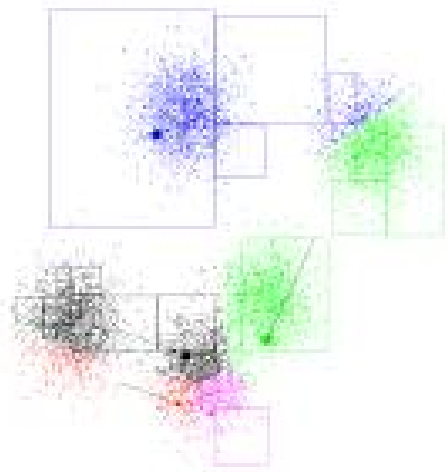
K-means clustering

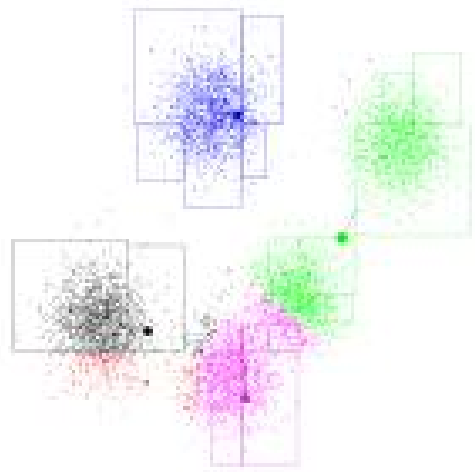
- Assign data points $\mathbf{x}_i$ to centres $\boldsymbol{\mu}_k$ by solving $\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$ for each:

$$k(i) = \arg\min_k \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

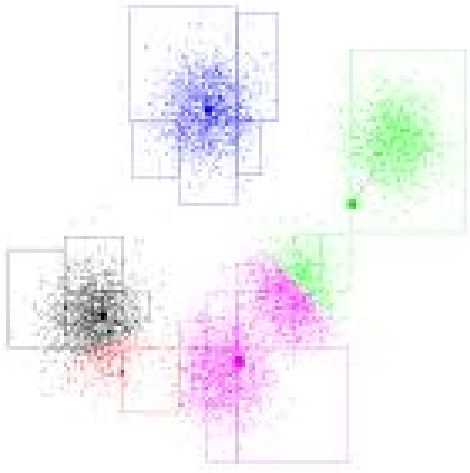
- Solve $\min_{\{\boldsymbol{\mu}_k\}} \frac{1}{n} \sum_{i=1}^{n} \left\|\mathbf{x}_i - \boldsymbol{\mu}_{k(i)}\right\|^2$ with fixed assignments of points $\mathbf{x}_i$ to centres $\boldsymbol{\mu}_k$
- First step is trivial:
  - Compute distance of $\mathbf{x}_i$ to each $\boldsymbol{\mu}_k$
  - Pick closest $\boldsymbol{\mu}_k$
- Second step is easy too:
  - $\boldsymbol{\mu}_k \leftarrow \frac{1}{\sum_{i:k(i)=k} 1} \sum_{i:k(i)=k} \mathbf{x}_i$
  - (exercise!)

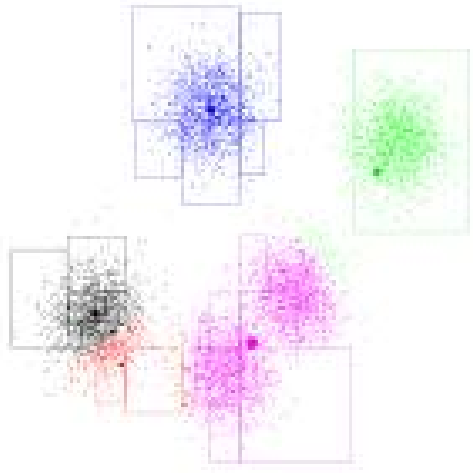
K-means clustering



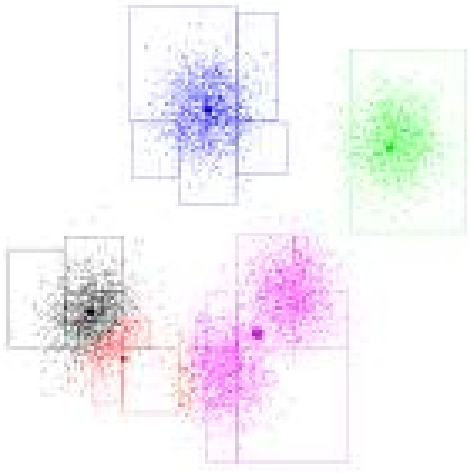


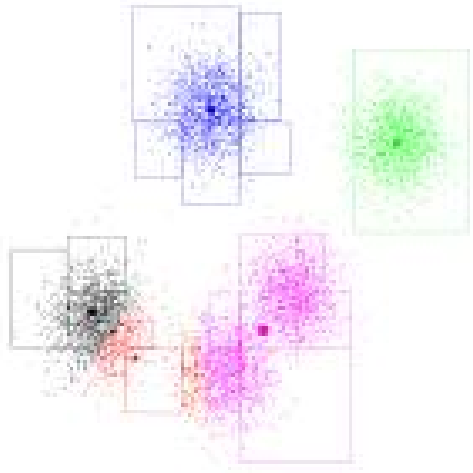
K-means clustering



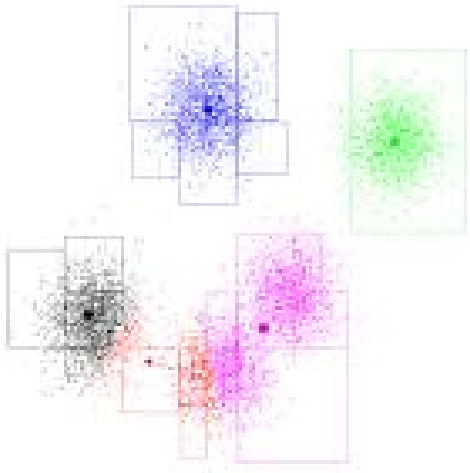


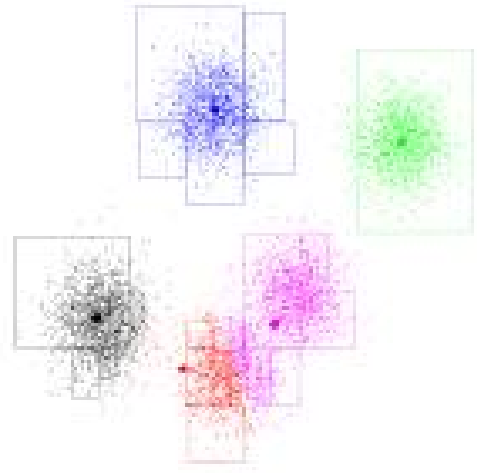
K-means clustering



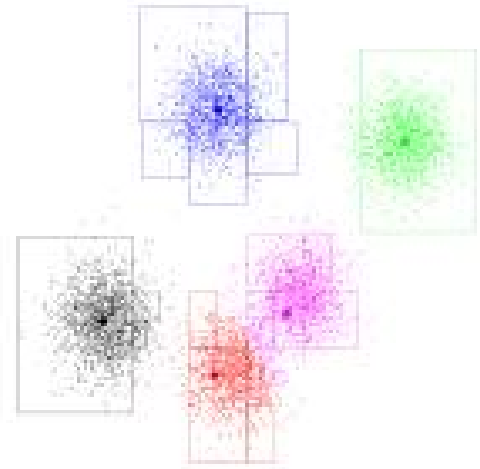


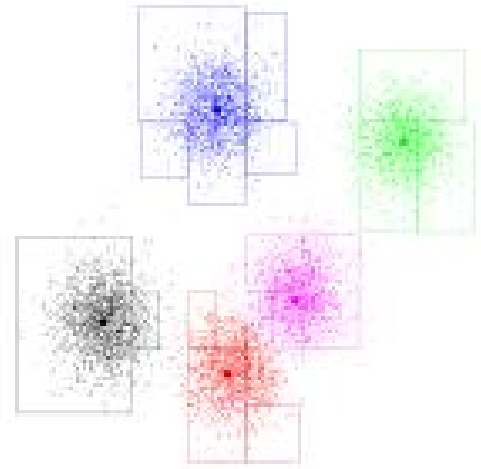
K-means clustering





K-means clustering





Summary

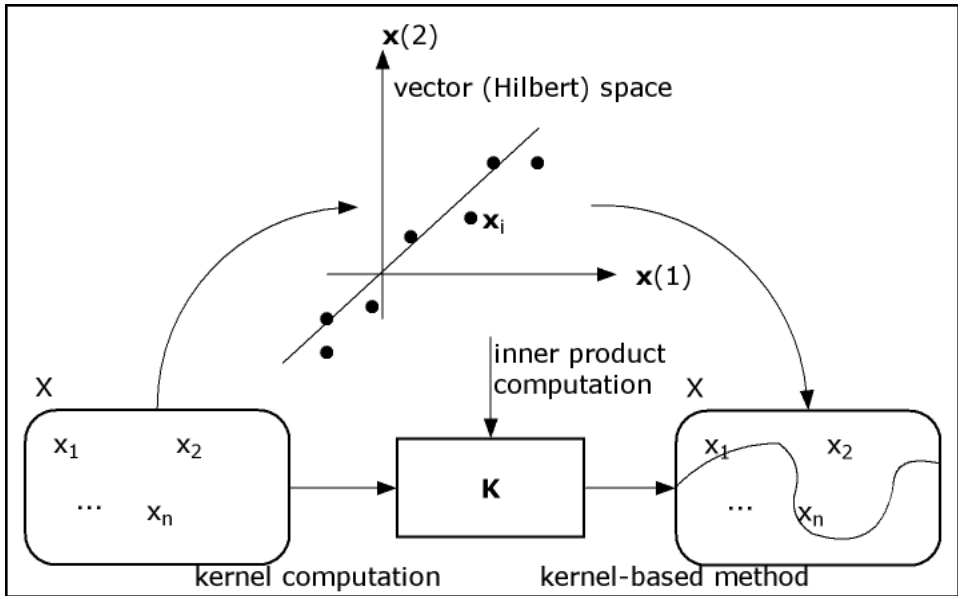
- We have discussed 2 techniques to get a grip on high-dimensional data
- Choose which one is most suitable...
- All this is when we can specify the vectors explicitly
- What if we cannot, or do not want to do that?
- In the extreme, what if the data is infinite dimensional?
- We will see that after the break...

Recapitulation

- We have described two methods for *exploratory data analysis / unsupervised learning*
- Both are useful in large-dimensional spaces
- But: as we have seen with Ridge Regression, often these spaces are hard to represent explicitly
- Their dimension may even be infinite...

Recapitulation

- In the case of Ridge Regression:
  - We could express the algorithm in terms of inner products (kernel functions)
  - We could express the evaluation of the regression function in terms of inner products (kernel functions)



Recapitulation

- The goal of all capacity constraints (and we nearly always use one) is to improve stability of the discovered patterns
- How can we quantify this?
- This is the second part of Lecture 4B... (at least for PCA)
- Now first: kernel versions!

Basic quantities using kernels

- Reminder: a kernel function is an inner product between vector representations:

$$k(x_i, x_j) = \mathbf{x}_i' \mathbf{x}_j$$

- We can express a lot in terms of inner products...

Basic quantities using kernels

- The norm of a vector:

$$\|\mathbf{x}\|^2 = \mathbf{x}'\mathbf{x} = k(x, x)$$

- The distance between two vectors:

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j)$$

- The norm of a linear combination:

$$\|\mathbf{X}'\boldsymbol{\alpha}\|^2 = \boldsymbol{\alpha}'\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$$

- Now, all quantities needed for K-means and for PCA can be expressed in such terms!

Kernel K-means

- First observation: the $\boldsymbol{\mu}_k$ are a linear combination of the $\mathbf{x}_i$

$$\boldsymbol{\mu}_k = \sum_{i=1}^{n} \mathbf{x}_i \alpha_k(i) = \mathbf{X}'\boldsymbol{\alpha}_k$$

- (Indeed: it is the 'mean' of a subset of them – *representer theorem!*)
- First step in the iteration: find minimum distance cluster mean – compute distance between $\boldsymbol{\mu}_j$ and $\mathbf{x}_j$

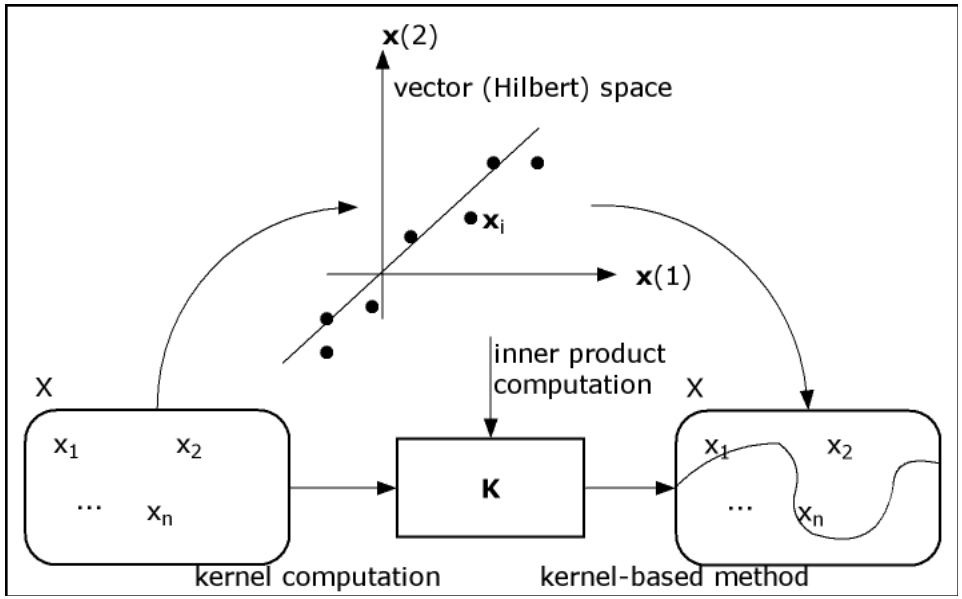
$$\begin{aligned}
\|\mathbf{x}_j - \boldsymbol{\mu}_k\|^2 &= \left\| x_j - \sum_{i=1}^{n} \mathbf{x}_i \alpha_k(i) \right\|^2 \\
&= (\boldsymbol{\alpha}_k - \mathbf{e}(j))' \mathbf{X}\mathbf{X}' (\boldsymbol{\alpha}_k - \mathbf{e}(j)) \\
&= k(x_j, x_j) + \boldsymbol{\alpha}_k{}'\mathbf{K}\boldsymbol{\alpha}_k - 2\boldsymbol{\alpha}_k{}'\mathbf{K}\mathbf{e}(j)
\end{aligned}$$

Kernel K-means

- The second step requires the computation of $\boldsymbol{\mu}_j$
- Can be done implicitly by equating $\alpha_k(i)$ to 1 if $\mathbf{x}_i$ has $\boldsymbol{\mu}_k$ as closest cluster centre (i.e. $\alpha_{k(i)} = 1$)

Kernel K-means

- Hence, also for kernel K-means:
  - We could express the algorithm in terms of inner products (kernel functions)
  - We could express the evaluation of the distance computation (and hence cluster assignment) in terms of kernel functions



Kernel PCA

- Here the story is even simpler
- The eigenvalue problem of PCA:

$$\mathbf{X}'\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

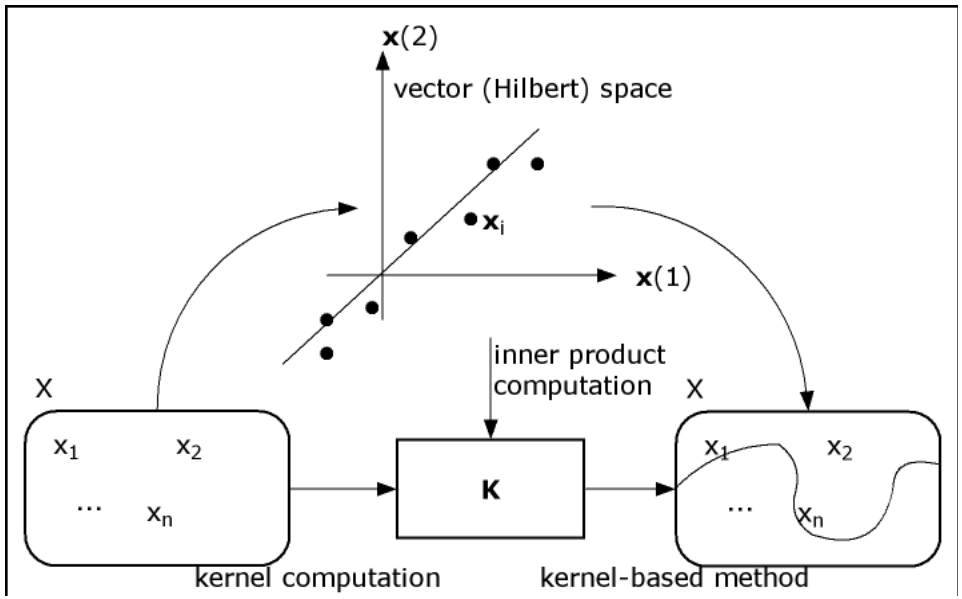
- Note (*representer theorem!*): $\mathbf{w} = \mathbf{X}'\left(\frac{1}{\lambda}\mathbf{X}\mathbf{w}\right) = \mathbf{X}'\boldsymbol{\alpha}$ for some alpha
- Plug this in, and multiply left and right hand side by $\mathbf{X}$:

$$\begin{aligned}\mathbf{X}\mathbf{X}'\mathbf{X}\mathbf{X}'\boldsymbol{\alpha} &= \lambda\mathbf{X}\mathbf{X}'\boldsymbol{\alpha}\\ \mathbf{K}^2\boldsymbol{\alpha} &= \lambda\mathbf{K}\boldsymbol{\alpha}\\ \mathbf{K}\boldsymbol{\alpha} &= \lambda\boldsymbol{\alpha}\end{aligned}$$

- To project a new data point on $\mathbf{w}$, it suffices to know the corresponding $\boldsymbol{\alpha}$, indeed: $\mathbf{x}'\mathbf{w} = \mathbf{x}'\mathbf{X}\boldsymbol{\alpha} = \sum_{i=1}^{n}\alpha_i\mathbf{x}'\mathbf{x}_i$

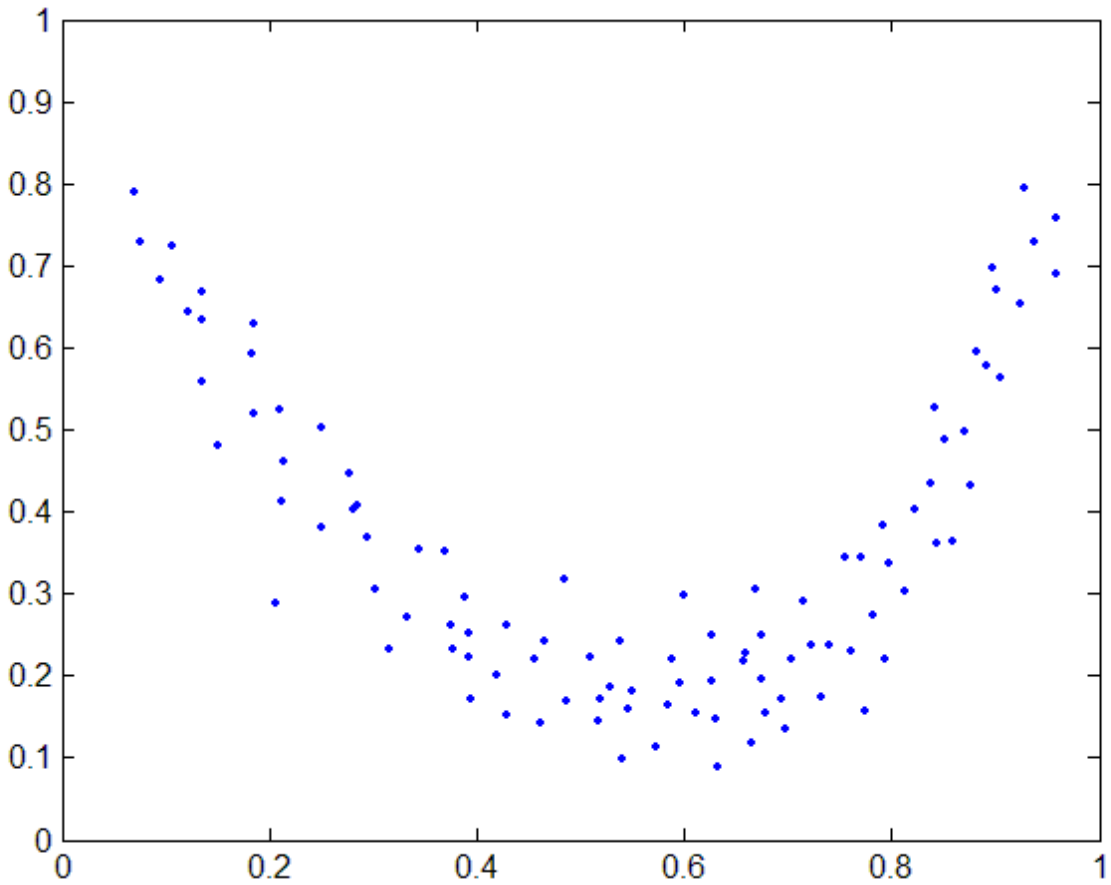
Kernel PCA

- Hence, also for kernel PCA:
  - We could express the algorithm in terms of kernel functions
  - We could express the projection in terms of kernel functions



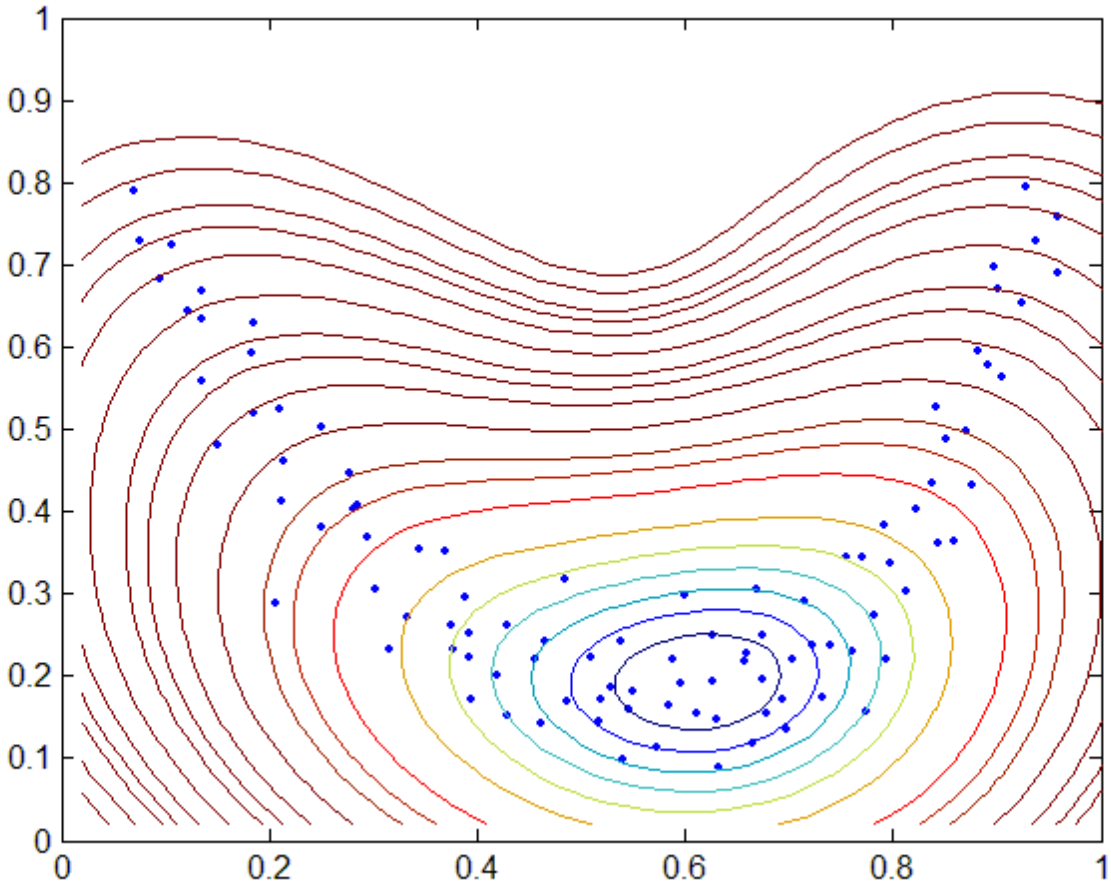
Kernel PCA

- Example:



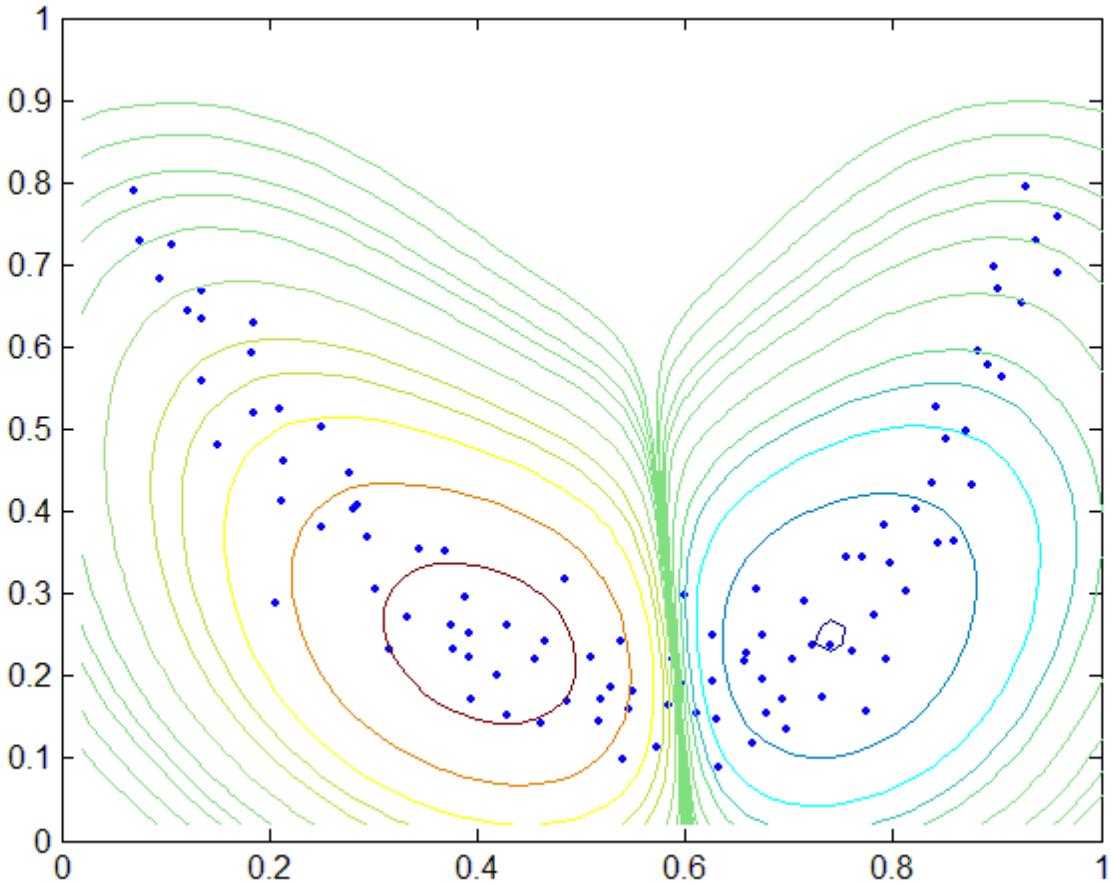
Kernel PCA

- Example:



Kernel PCA

- Example:



Why stability?

- Using PCA, we find directions along which the data has a large variance
- What about the distribution the data is sampled from (i.i.d.)?
- Is the variance large in expectation?
- I.e., what is the expectation of the pattern function?

$\rightarrow$ make stability intervals for $E\left\{\pi_{\mathbf{w}}\left(\underline{X}\right)\right\}$:
with probability $1-\delta$, there holds that
$\pi_{\mathbf{w}}\left(X\right) - E\left\{\pi_{\mathbf{w}}\left(\underline{X}\right)\right\} < \epsilon$

Stability of averaging pattern functions

- Remember, $\pi_{\mathbf{w}}(\mathbf{X})$ is an averaging pattern function:

$$\pi_{\mathbf{w}}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} g_{\pi_{\mathbf{w}}}(\mathbf{x}_i)$$

- In general for averaging pattern functions, thanks to linearity of the expectation operator, and for i.i.d. x_i:

$$E\{\pi(\underline{X})\} = E\left\{\frac{1}{n}\sum_{i=1}^{n} g_{\pi}(\underline{\mathbf{x}}_i)\right\} = E\{g_{\pi}(\underline{\mathbf{x}})\}$$

- Hence, here:

$$E\{\pi(\underline{\mathbf{X}})\} = E\left\{(\underline{\mathbf{x}}'\mathbf{w})^2\right\}$$

- $\Rightarrow$ a confidence interval for $E\{\pi(\underline{\mathbf{X}})\}$ is automatically one for $E\left\{(\underline{\mathbf{x}}'\mathbf{w})^2\right\}$

Stability of PCA

- Two problems imaginable:
  - Stability of a fixed pattern function $\pi_{\mathbf{w}}$, with fixed $\mathbf{w}$
  - Stability of a discovered pattern function $\pi_{\mathbf{w}}$, with $\mathbf{w}$ direction of maximal variance for the data
- Important distinction!
- Second one requires stability of all pattern functions in the pattern space, and hence something like a union bound
- However, pattern space is infinite here (parameterised by all vectors $\mathbf{w}$ of norm at most 1, in a high-dimensional space)... so we will need something more sophisticated

Stability of PCA: a given pattern function

- Stability question: for which ϵ and δ holds that $\pi_{\mathbf{w}}(X) - E\{\pi_{\mathbf{w}}(\underline{X})\} < \epsilon$ with probability $1-\delta$? Use McDiarmid's theorem:

Theorem (McDiarmid's inequality)

For f a function of $X = \{x_1, x_2, ..., x_i, ..., x_n\} \in \mathcal{X}$ and $\underline{x_i}$ iid, if $f(X)$ has bounded differences c_i, meaning that $|f(X) - f(X^i)| \leq c_i$, we have that

$$P(f(\underline{X}) - E\{f(\underline{X})\} < \epsilon) \geq 1 - \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Stability of PCA: a given pattern function

- We need to assume that the norm of $\mathbf{x}$ is bounded: $\|\mathbf{x}\|^2 = k(x, x) \leq R^2$
- Then, for $\|\mathbf{w}\|^2 \leq 1$, using the Cauchy Schwarz inequality: $0 \leq (\mathbf{x}'\mathbf{w})^2 \leq \|\mathbf{x}\|^2 \cdot \|\mathbf{w}\|^2 = R^2$
- Hence: bounded differences: $|\pi(X) - \pi(X^i)| \leq c_i$ with $c_i = \frac{R^2}{n}$
- (Intuition: the pattern function is an average of bounded numbers... so it should be concentrated/stable)

- So, with $c_i = \frac{R^2}{n}$:

$$\begin{aligned} P(\pi_{\mathbf{w}}(\underline{X}) - E\{\pi_{\mathbf{w}}(\underline{X})\} < \epsilon) &\geq 1 - \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right) \\ &= 1 - \exp\left(\frac{-2n\epsilon^2}{R^4}\right). \end{aligned}$$

- Rework this by introducing δ with $\epsilon = R^2\sqrt{\frac{\ln(1/\delta)}{2n}}$
- For any fixed $\mathbf{w}$, and with a probability of at least $1-\delta$,

$$\begin{aligned} \pi_{\mathbf{w}}(X) - E\{\pi_{\mathbf{w}}(\underline{X})\} &< R^2\sqrt{\frac{\ln(1/\delta)}{2n}} \\ \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i'\mathbf{w})^2 - E\left\{(\underline{\mathbf{x}}'\mathbf{w})^2\right\} &< R^2\sqrt{\frac{\ln(1/\delta)}{2n}} \end{aligned}$$

Stability of PCA: a discovered pattern function

- As mentioned, in practice **w** is *discovered*
- Then we need to show that this inequality holds for all **w** for complete pattern space explored
- Assume cardinality N
- Then the same bound would hold with probability $1 - N\delta$ (union bound)
- However, N is infinitely large...
- So we need another technology: Rademacher bounds

Stability of PCA: a discovered pattern function

- Using McDiarmid's inequality again (and the same bounds on the differences c_i):

$$\begin{aligned}
& \pi_{\mathbf{w}}(X) - E_{\underline{\mathbf{X}}}\{\pi_{\mathbf{w}}(\underline{\mathbf{X}})\} \\
\leq & \max_{\pi_{\mathbf{w}}: C(\pi_{\mathbf{w}}) \leq c} \left(\pi_{\mathbf{w}}(X) - E_{\underline{\mathbf{X}}}\{\pi_{\mathbf{w}}(\underline{\mathbf{X}})\}\right) \\
\leq & E_{\underline{\mathbf{Z}}}\left\{\max_{\mathbf{w}}\left(\pi_{\mathbf{w}}(\underline{\mathbf{Z}}) - E_{\underline{\mathbf{X}}}\{\pi_{\mathbf{w}}(\underline{\mathbf{X}})\}\right)\right\} + R^2\sqrt{\frac{\ln(1/\delta)}{2n}} \\
\leq & E_{\underline{\mathbf{XZ}}}\left\{\max_{\mathbf{w}}\left(\pi_{\mathbf{w}}(\underline{\mathbf{Z}}) - \pi_{\mathbf{w}}(\underline{\mathbf{X}})\right)\right\} + R^2\sqrt{\frac{\ln(1/\delta)}{2n}}
\end{aligned}$$

- What remains to be bounded is the additional term $E_{\underline{\mathbf{XZ}}}\{\max_{\mathbf{w}}(\pi_{\mathbf{w}}(\underline{\mathbf{Z}}) - \pi_{\mathbf{w}}(\underline{\mathbf{X}}))\}$ – the *complexity term* of the pattern space

Stability of PCA: a discovered pattern function

- Applied to the PCA problem:

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 - E_{\underline{\mathbf{X}}}\left\{\left(\underline{\mathbf{x}}'\mathbf{w}\right)^2\right\} \\
\leq \quad & \max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 - E_{\underline{\mathbf{X}}}\left\{\left(\underline{\mathbf{x}}'\mathbf{w}\right)^2\right\}\right) \\
\leq \quad & E_{\underline{\mathbf{Z}}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2 - E_{\underline{\mathbf{X}}}\left\{\left(\underline{\mathbf{x}}'\mathbf{w}\right)^2\right\}\right)\right\} + R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}} \\
\leq \quad & E_{\underline{\mathbf{X}}\underline{\mathbf{Z}}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right)\right\} + R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}
\end{aligned}
$$

- The second term we had before as well!

Stability of PCA: a discovered pattern function

- What remains to be bounded is the additional term $E_{\underline{\mathbf{XZ}}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right)\right\}$ – the *complexity term* of the pattern space
- Bounded by the *Rademacher complexity*:

$$\begin{aligned}
& E_{\underline{\mathbf{XZ}}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right)\right\} \\
= \quad & E_{\underline{\mathbf{XZ}}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2-\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right]\right)\right\} \\
= \quad & E_{\underline{\mathbf{XZ}}\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\frac{1}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left[\left(\underline{\mathbf{z}}_i'\mathbf{w}\right)^2-\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right]\right)\right\} \\
\leq \quad & E_{\underline{\mathbf{X}}\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right|\right)\right\}
\end{aligned}$$

Stability of PCA: a discovered pattern function

- Furthermore, the Rademacher complexity can be upper bounded with high probability by the *empirical Rademacher complexity*: with probability $1-\delta$:

$$\begin{aligned} & E_{\underline{\mathbf{X}}\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\underline{\mathbf{x}}_i'\mathbf{w}\right)^2\right|\right)\right\} \\ \leq & \quad E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\mathbf{x}_i'\mathbf{w}\right)^2\right|\right)\right\}+2R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}} \\ = & \quad \widehat{\mathcal{R}}_{\mathbf{X}}\left(\Pi\right)+2R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}} \end{aligned}$$

Stability of PCA: a discovered pattern function

- In summary, with a probability of at least $1 - 2\delta$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 - E_{\underline{\mathbf{x}}}\left\{\left(\underline{\mathbf{x}}'\mathbf{w}\right)^2\right\} \leq \widehat{\mathcal{R}}_{\mathbf{X}}\left(\Pi\right) + 3R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}$$

  where

$$\widehat{\mathcal{R}}_{\mathbf{X}}\left(\Pi\right) = E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\mathbf{x}_i'\mathbf{w}\right)^2\right|\right)\right\}$$

- Can we compute, or bound, the empirical Rademacher complexity?

Bounding the empirical Rademacher complexity

$$
\begin{aligned}
\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) &= E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left(\mathbf{x}_i'\mathbf{w}\right)^2\right|\right)\right\} \\
&= E_{\underline{\sigma}}\left\{\max_{\mathbf{w}}\left(\left|\frac{2}{n}\sum_{i=1}^{n}\underline{\sigma_i}\left\langle\mathbf{x}_i\mathbf{x}_i', \mathbf{w}\mathbf{w}'\right\rangle\right|\right)\right\} \\
&= \frac{2}{n}E_{\sigma}\left\{\max_{\mathbf{w}}\left|\left\langle\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\mathbf{x}_i', \mathbf{w}\mathbf{w}'\right\rangle\right|\right\} \\
&\leq \frac{2}{n}E_{\sigma}\left\{\max_{\mathbf{w}}\left(\sqrt{\left\langle\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\mathbf{x}_i', \sum_{i=1}^{n}\sigma_i\mathbf{x}_i\mathbf{x}_i'\right\rangle}\sqrt{\left\langle\mathbf{w}\mathbf{w}', \mathbf{w}\mathbf{w}'\right\rangle}\right)\right\}
\end{aligned}
$$

(Last inequality: Cauchy-Schwarz)

$$
\begin{aligned}
\widehat{\mathcal{R}}_{\mathbf{X}}(\Pi) &\leq \frac{2}{n} E_{\sigma}\left\{\sqrt{\left\langle \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \mathbf{x}_i', \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \mathbf{x}_i' \right\rangle}\right\} \\
&\leq \frac{2}{n} \sqrt{E_{\sigma}\left\{\left\langle \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \mathbf{x}_i', \sum_{i=1}^{n} \sigma_i \mathbf{x}_i \mathbf{x}_i' \right\rangle\right\}} \\
&= \frac{2}{n} \sqrt{E_{\sigma}\left\{\sum_{i,j=1}^{n} \sigma_i \sigma_j \left\langle \mathbf{x}_i \mathbf{x}_i', \mathbf{x}_j \mathbf{x}_j' \right\rangle\right\}} \\
&= \frac{2}{n} \sqrt{\sum_{i=1}^{n} \|\mathbf{x}_i\|^4}.
\end{aligned}
$$

(Second inequality: Jensen)

The complete bound

- The complete bound becomes:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i'\mathbf{w}\right)^2 - E_{\underline{\mathbf{x}}}\left\{\left(\underline{\mathbf{x}}'\mathbf{w}\right)^2\right\} \leq \frac{2}{n}\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|^4} + 3R^2\sqrt{\frac{\ln\left(1/\delta\right)}{2n}}$$

- All terms on the right hand side decrease with $\sqrt{n}$
- Hence, the pattern is stable, and more stable for larger n

Wrap-up

- Introduced Exploratory data analysis methods
  - K-means / clustering
  - Principal Component Analysis
- Derived kernel versions – allow to apply these methods in a nonlinear way, or on diverse data types by means of the kernel trick
- Derived a bound showing stability of the PCA result