ELEC2870 - Machine learning: regression and dimensionality reduction

Feature selection

Michel Verleysen

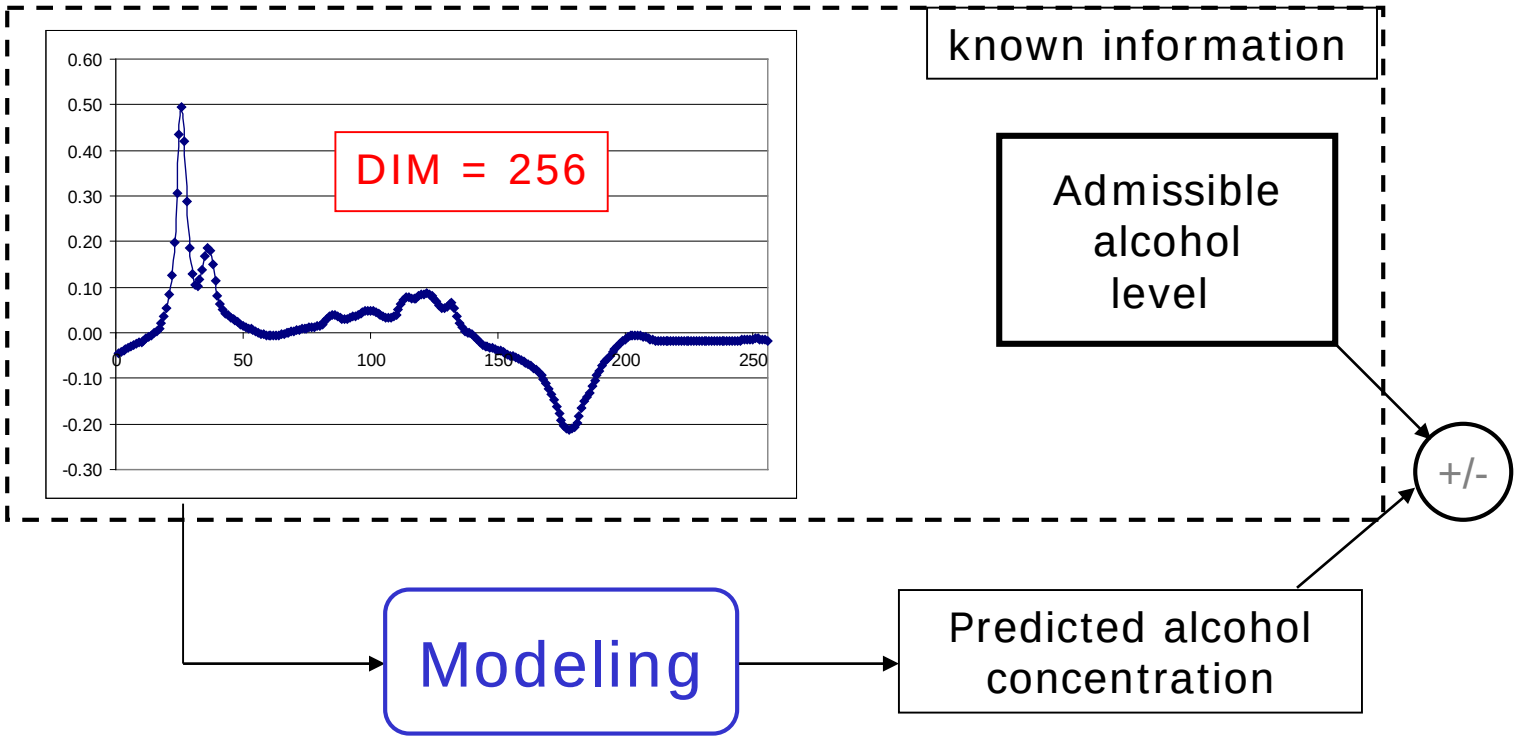
Machine Learning Group
Université catholique de Louvain
Louvain-la-Neuve, Belgium
michel.verleysen@uclouvain.be

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

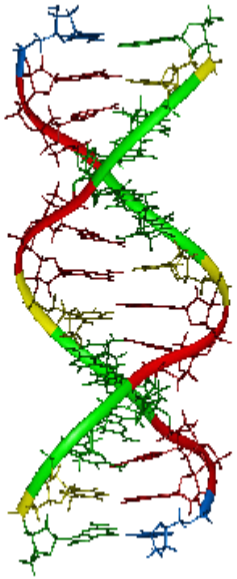
HD data exist !

- Enhanced data acquisition possibilities
  → many HD data!
  classification - clustering - regression



HD data exist !

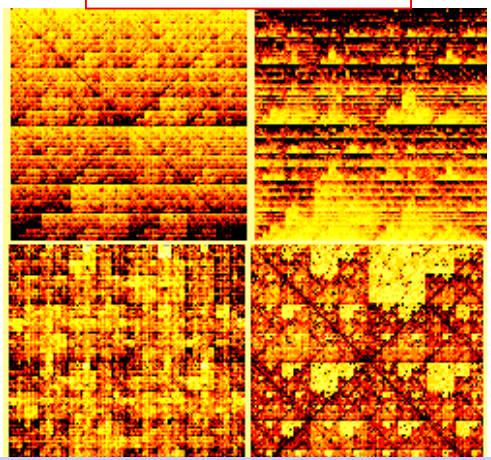
- Enhanced data acquisition possibilities
  → many HD data!
  classification - clustering - regression

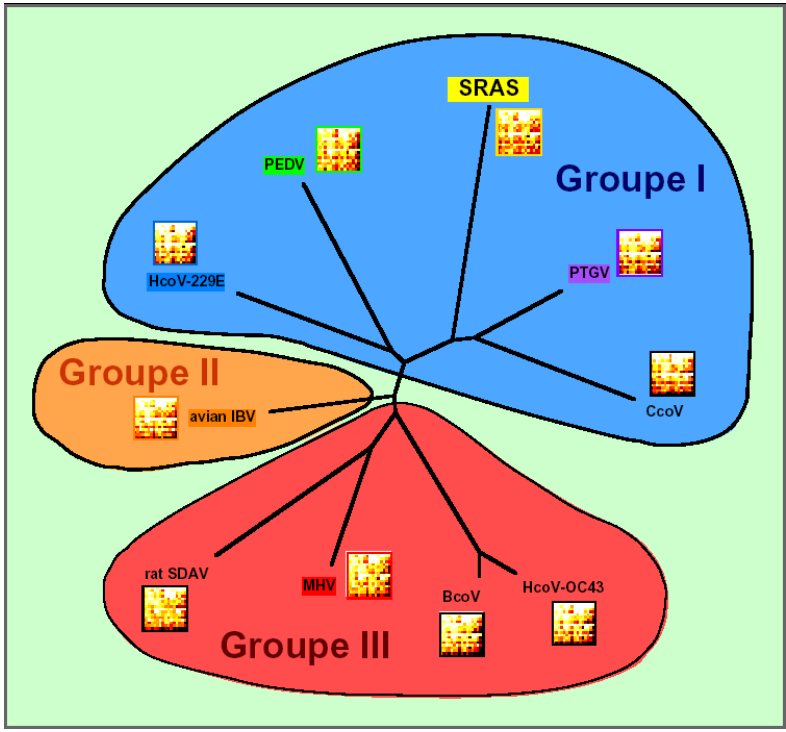


feature extraction

clustering

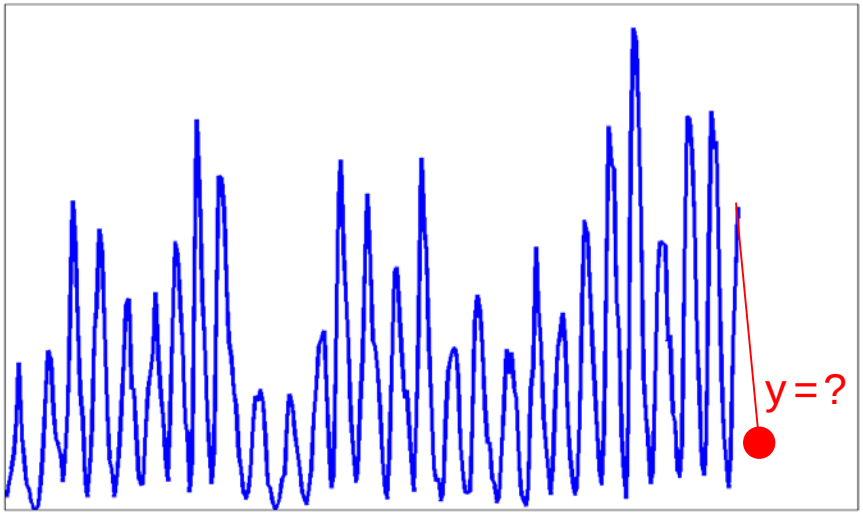
DIM = 16384





HD data exist !

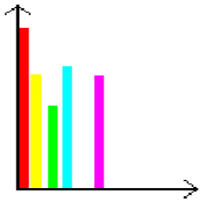
- Enhanced data acquisition possibilities
  → many HD data!
  classification - clustering - regression



$$y = f\left(x_{t-\mathrm{DIM}+1},\ldots x_{t-1},x_t\right)$$

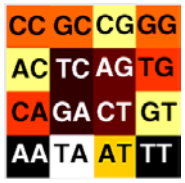
Generic data analysis

- Traditional data



When I find myself in times of trouble Mother Mary comes to me Speaking words of wisdom, let it be. …

When	1
Times	1
Trouble	1
Let	65
wisdom	1



number of Variables or features

Data matrix

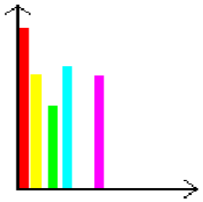
number of observations

Analysis

Models

High-dimensional data analysis

- High-dimensional data



number of
variables or features

When I find myself
in times of trouble
Mother Mary
comes to me
Speaking words of
wisdom, let it be.
…

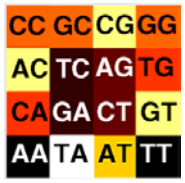
When	1
Times	1
Trouble	1
Let	65
wisdom	1

Data
matrix

Analysis

Models

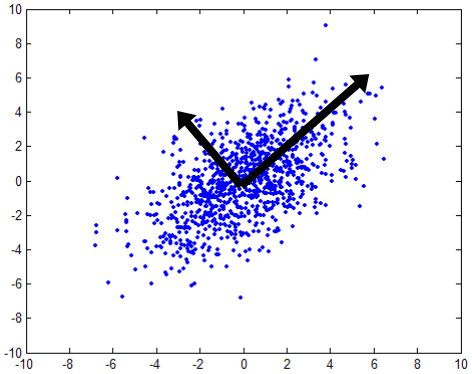
number of
observations



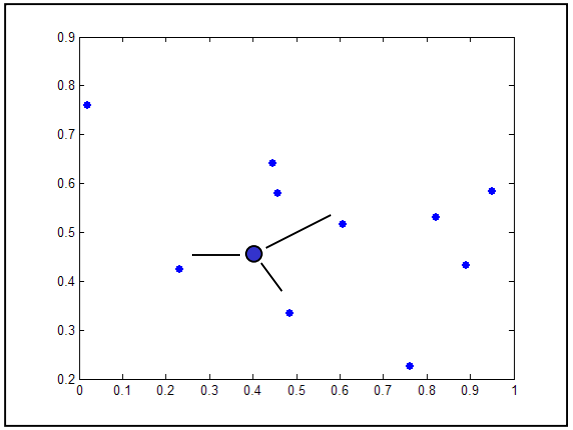
High-dimensional data

$$\text{data} = \begin{pmatrix} 0.32 & 2.5 & -0.01 & -3.7 & \dots & 12.1 \end{pmatrix}^T \in \Re^d$$

- Data are described in a normed vector space (Euclidean space)
- Tools derived from algebra & geometry

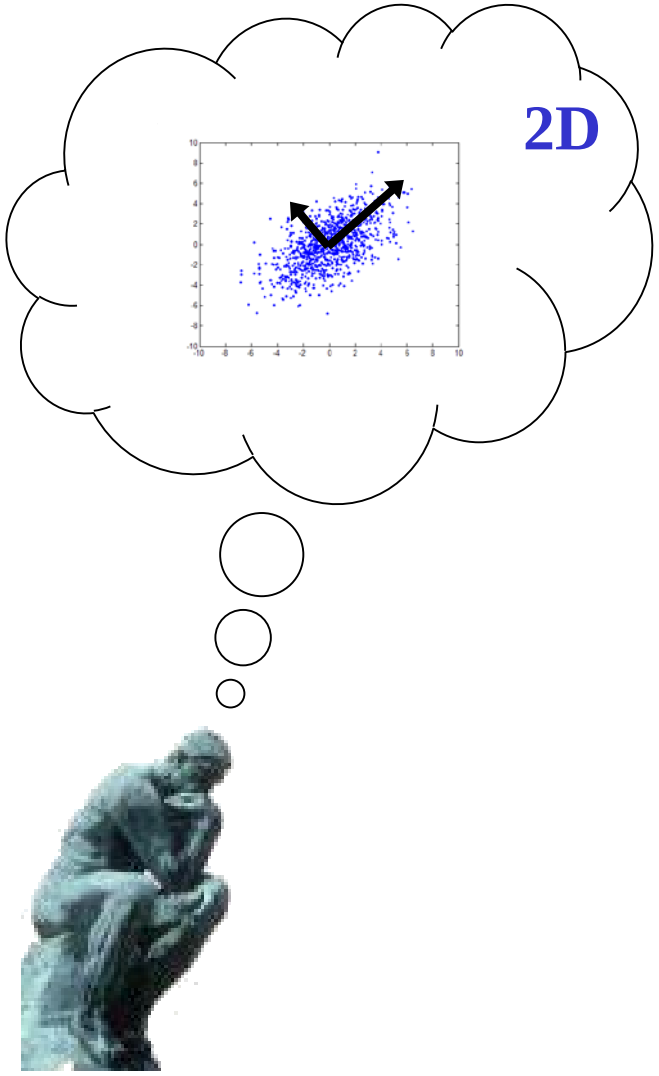


Principal Component Analysis



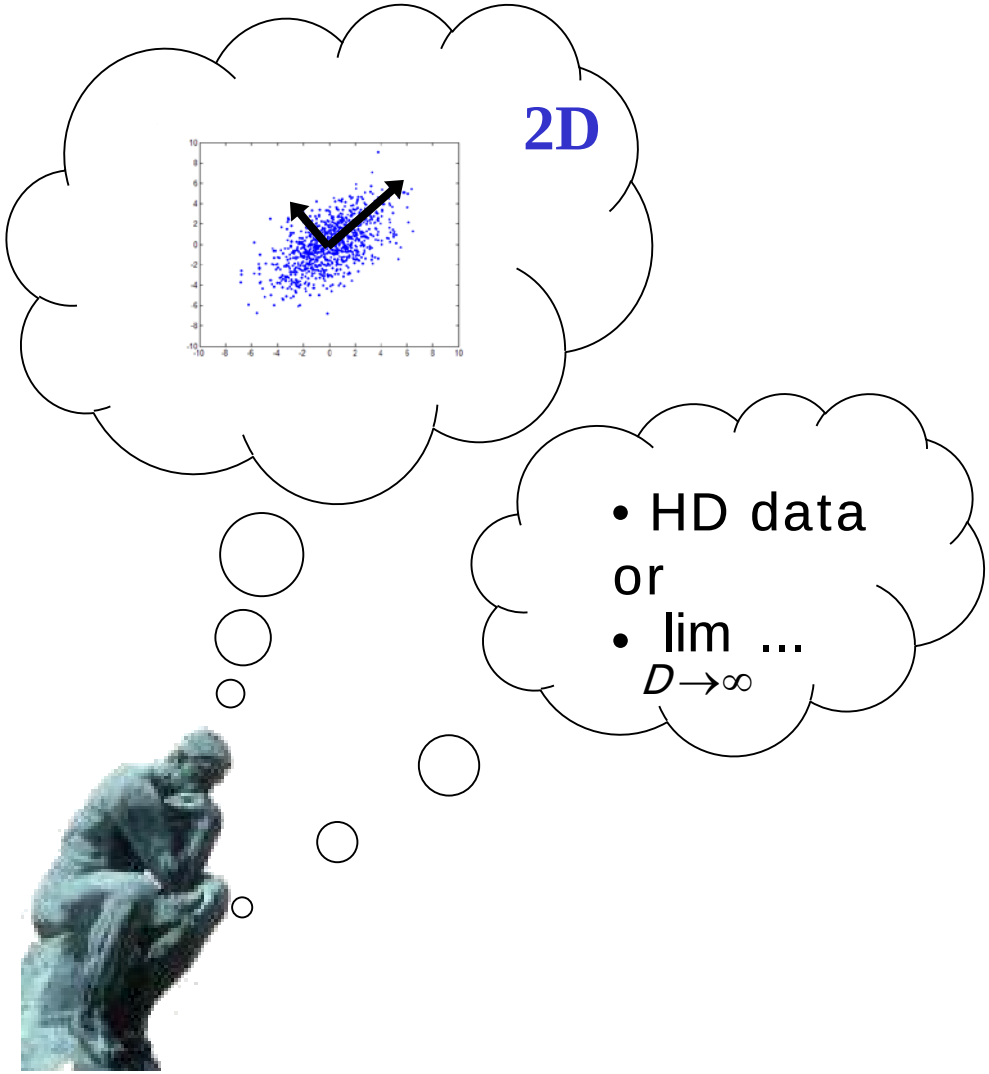
k-Nearest Neighbours

High-dimensional data



- Situations we can imagine, represent, draw
- Strong intuition of how the tools behave
- Consider cases where #observations >> *d*

High-dimensional data



- ~~Situations we can imagine, represent, draw~~
- No representation

- ~~Strong intuition of how the tools behave~~
- No intuition

- ~~Consider cases where #observations >> e^D~~
- Often #observations << e^D

Linear tools

- Principal component analysis (PCA):
  based on covariance matrix
  - huge (DIM x DIM)
  - poorly estimated with finite number of data

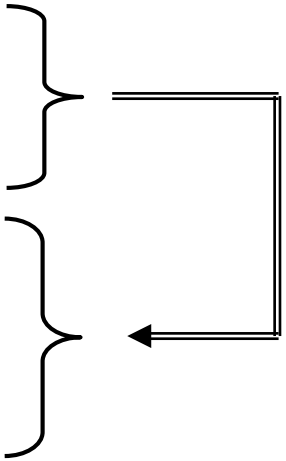
- Other methods:
  - Linear discriminant analysis (LDA)
  - Partial least squares (PLS)
  - ...

  Similar problems!

Nonlinear tools

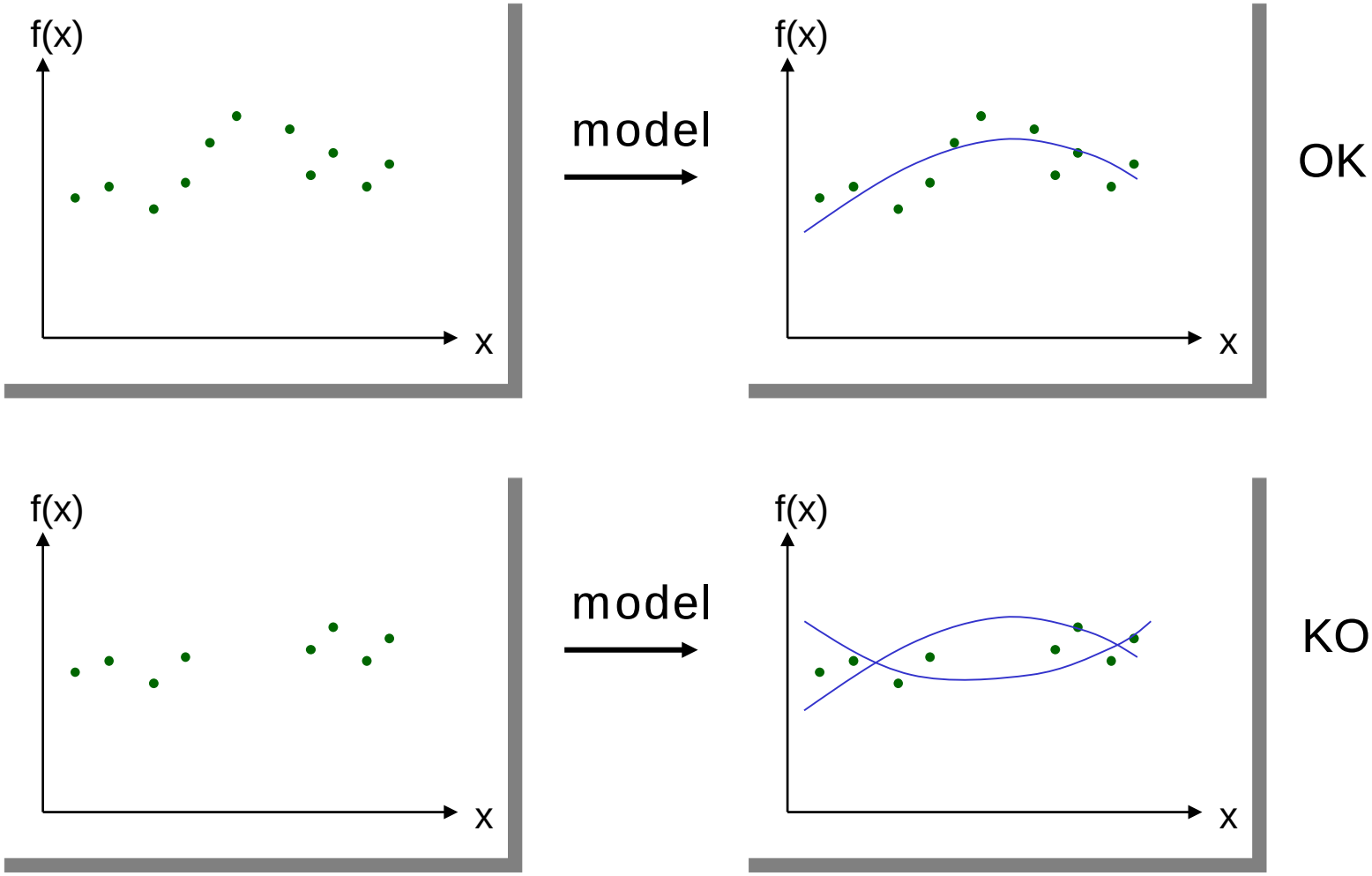
$$y = f(x_1, x_2, \ldots, x_d, \theta)$$

- If d ↗↗, size(θ) ↗↗
- θ results from the minimization of a non-convex cost function
  - local minima
  - numerical problems (flats, high slopes)
  - convergence
  - etc



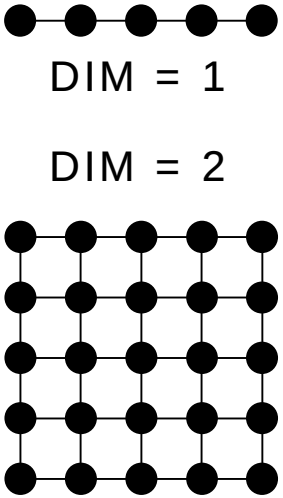
- Ex: Multi-layer perceptrons, Gaussian mixtures (RBF), self-organizing maps, etc.

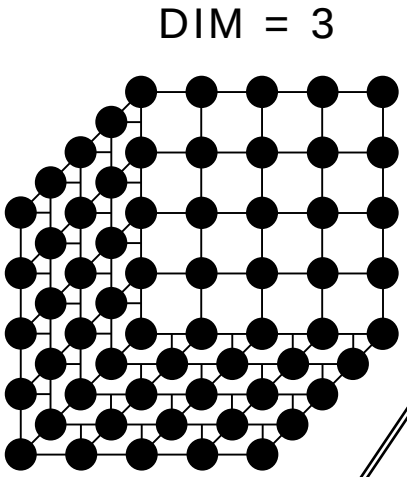
Curse of dimensionality



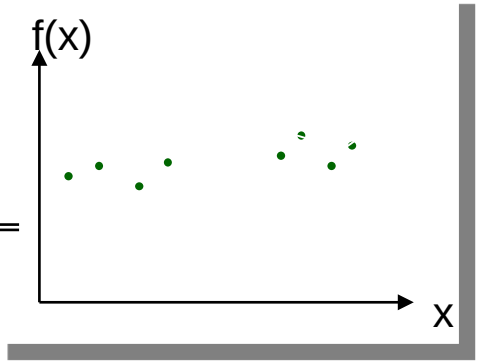
Curse of dimensionality

- Number of points on a grid increases exponentially with DIM



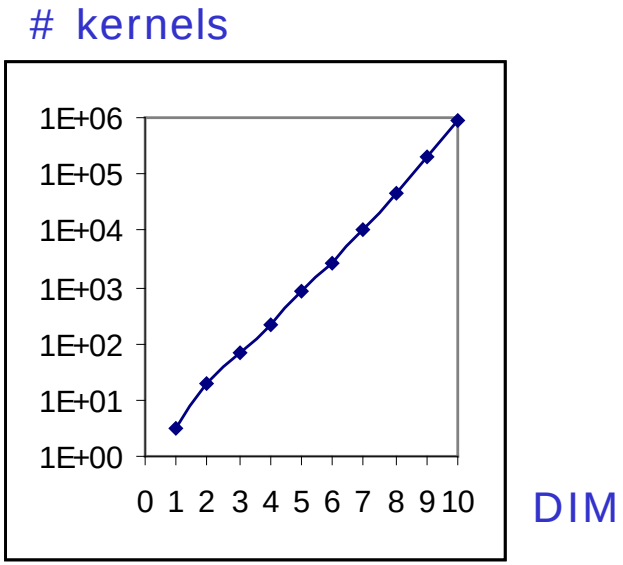


- In high DIM:
  - never enough data
  - never sure to *interpolate*



Curse of dimensionality

- Example: Silverman (1986)
  Number of Gaussian kernels necessary to approximate a (Gaussian) distribution in DIM



Empty space phenomenon

- "Statistical" view of the same problem:

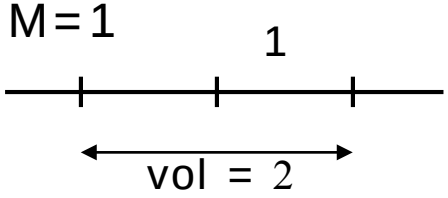
  To estimate parameters
  ex: covariance matrix of Gaussian distribution in DIM ↗↗

  one needs (too…) many data

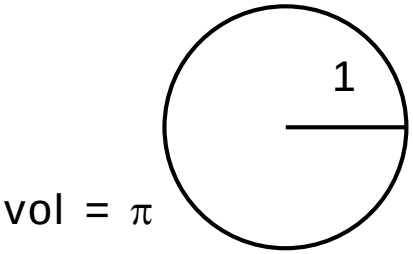
- Example: histograms in DIM ↗↗
  - compromize between accuracy and # of data in each bin

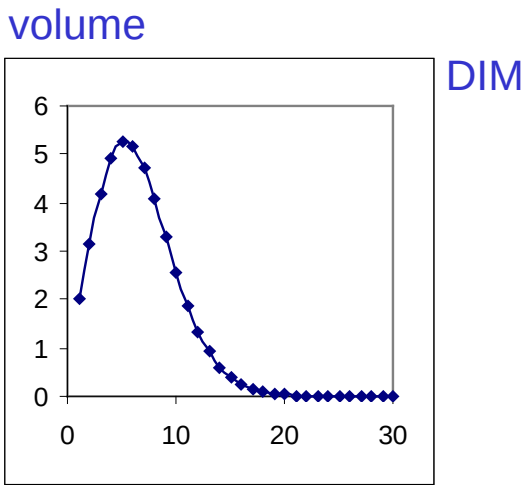
Surprizing facts: sphere

- Volume of sphere of constant radius (=1) in dimension DIM
- DIM=1



DIM=2

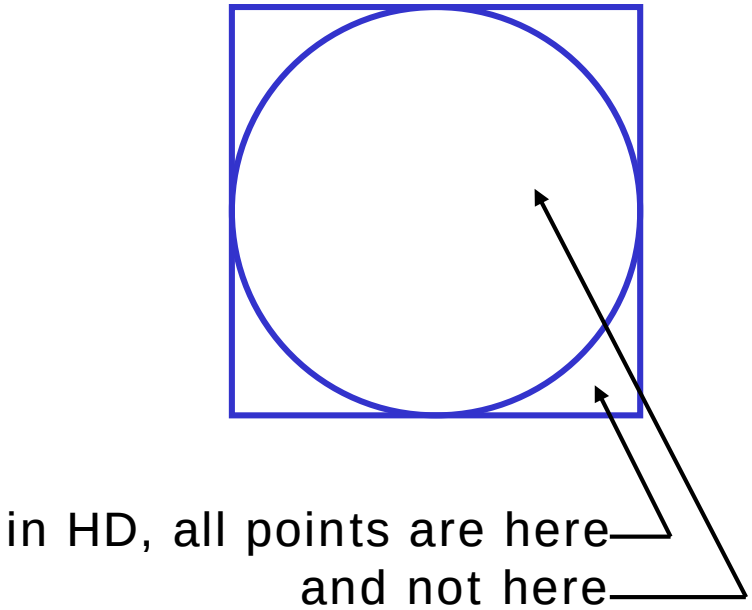


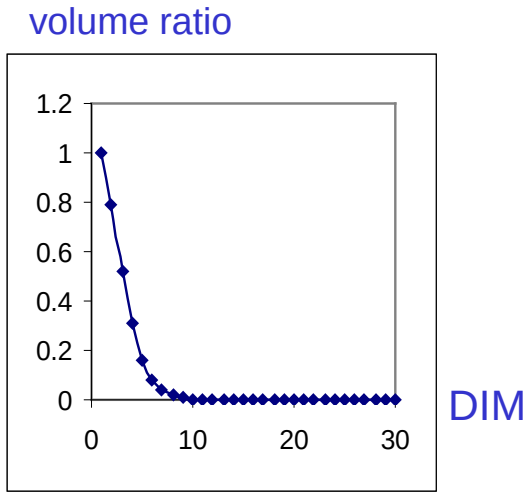


$$V(d) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d$$

Surprizing facts: corners

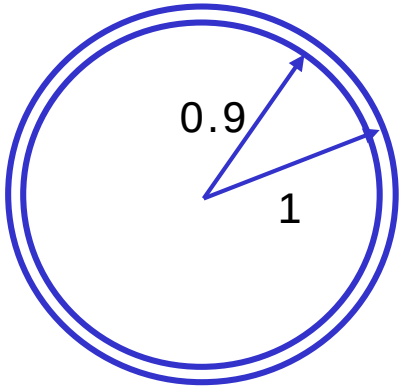
- Ratio volume sphere / cube

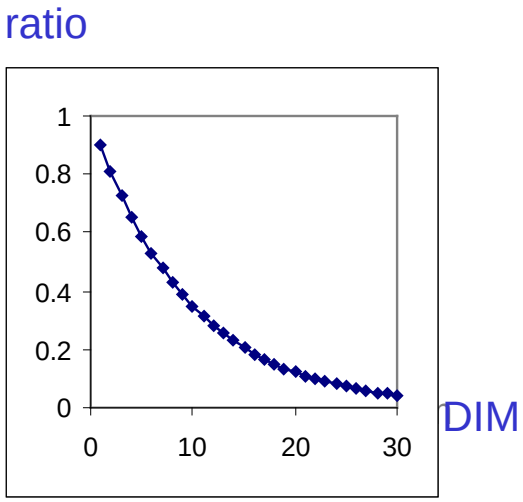




Surprizing facts: spheres

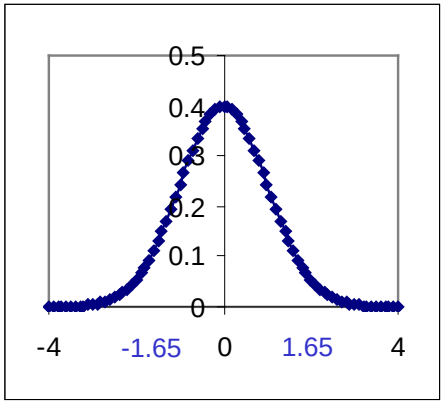
- Volume ratio of embedded spheres





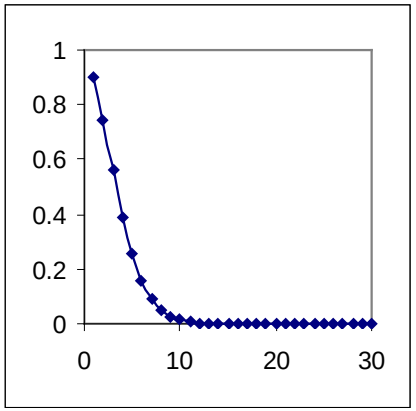
Surprizing facts: Gaussians

- Multi-DIM Gaussian distributions



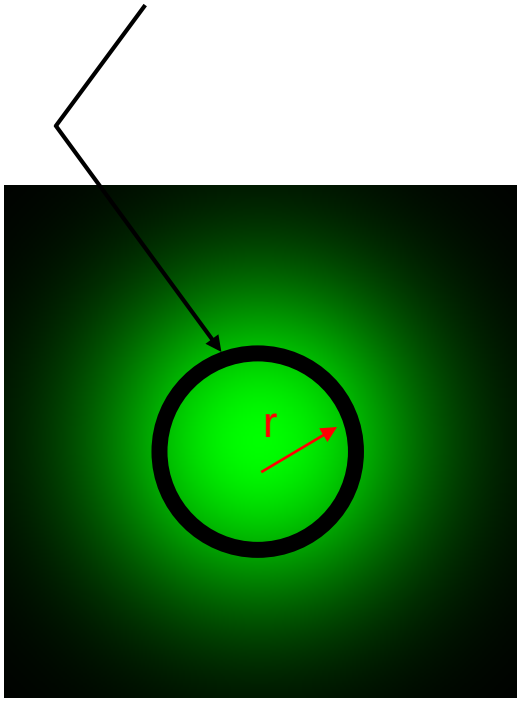
DIM=1

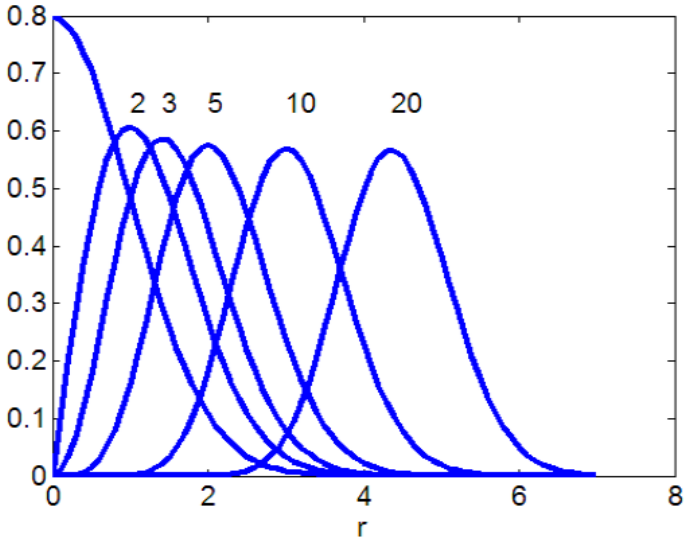
- % points inside a sphere of radius 1.65



Surprizing facts: Gaussians

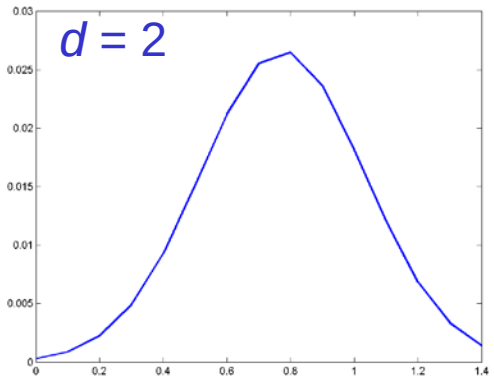
- Another view of high-DIM Gaussian distributions:
  - Probability to find a point at distance *r* from the center of a DIM-dimensional multinormal distribution

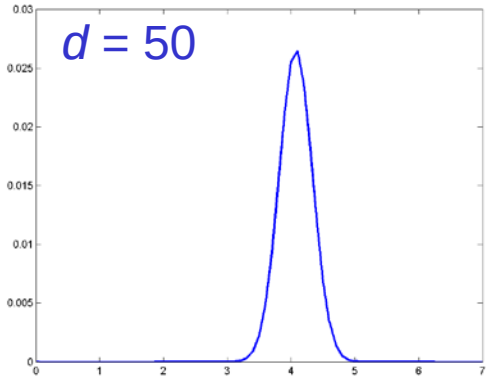




Concentration of the Euclidean norm

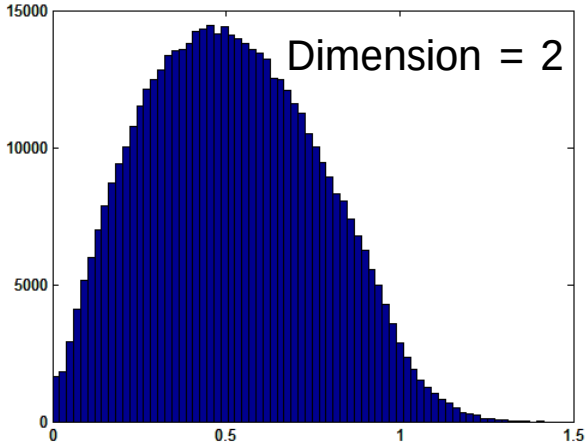
- Distribution of the norm of random vectors
  - i.i.d. components in [0,1]
  - norms in $[0, \sqrt{d}]$ as

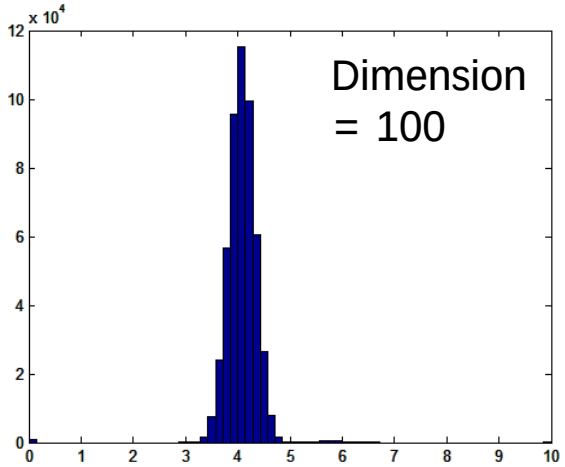




- Norms concentrate around their expectation
- They don't discriminate anymore !

Distances also concentrate





Pairwise distances seem nearly equal for all points

Relative contrast vanishes as the dimension increases

If $\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}\left(\|X\|_2\right)}}{\mathrm{E}\left(\|X\|_2\right)} = 0$ then $\frac{DMAX_d - DMIN_d}{DMIN_d} \to_p 0$

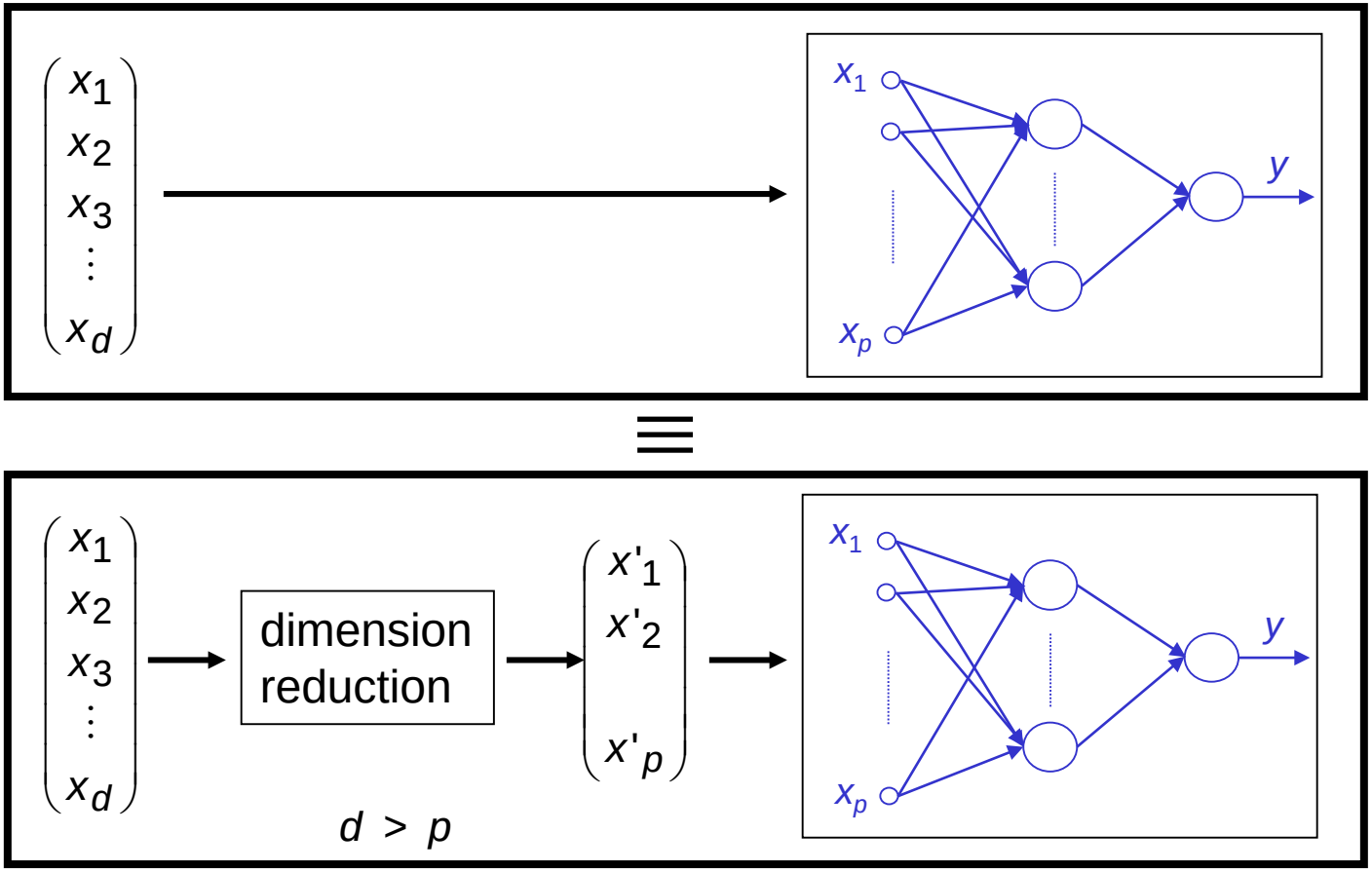
when $d \to \infty$

[Beyer]

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Reducing (the curse of) dimensionality



Why reducing the dimensionality ?

- Theoretically not useful :
  - More information means easier task
  - Models can ignore irrelevant features
    (e.g. set weights to zero)

« In theory, practice and theory are the same. But in practice, they're not »

- Lot of inputs means …
  - Lots of parameters & Large input space

➔ Curse of dimensionality and risks of overfitting !

Why feature selection?

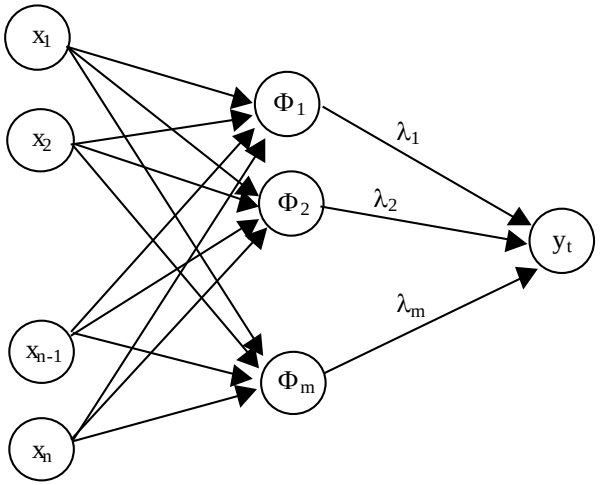
- Feature selection is often as important as the model itself !
- Industrial example

Road Temperature

Air Temperature

Wind Direction

Dew Temperature

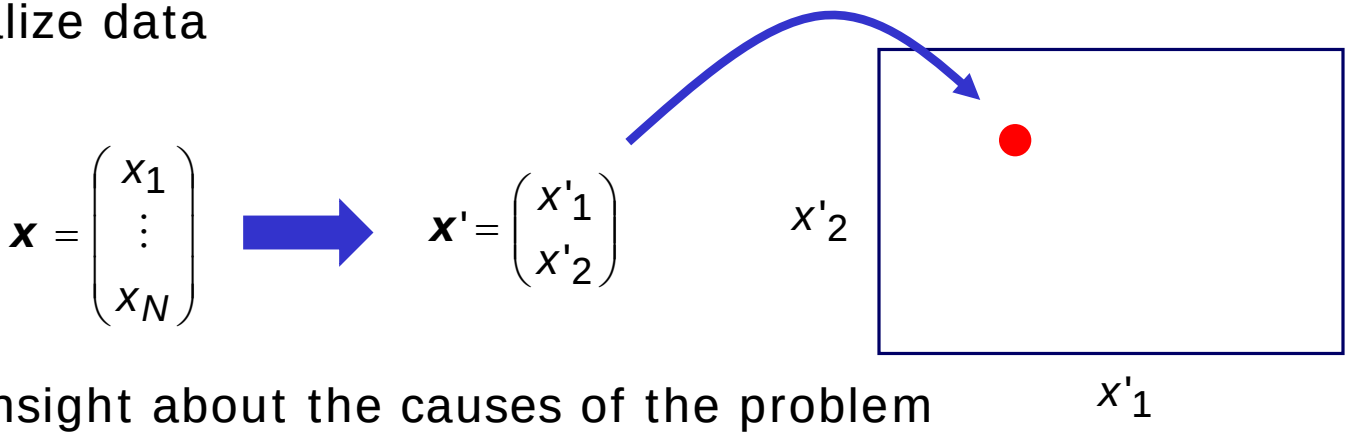


Presence of ice on the road

- which feature (input, variable, attribute, predictor, ...) should we feed our model with ??

Why feature selection?

- To visualize data



- To get insight about the causes of the problem

  If air temperature is selected and road temperature is not

  Then water temperature is closer to air temperature than to road temperature.

- Reduce data collection time/cost, computation time, etc.

Feature selection reduces dimensionality

- Unsupervised

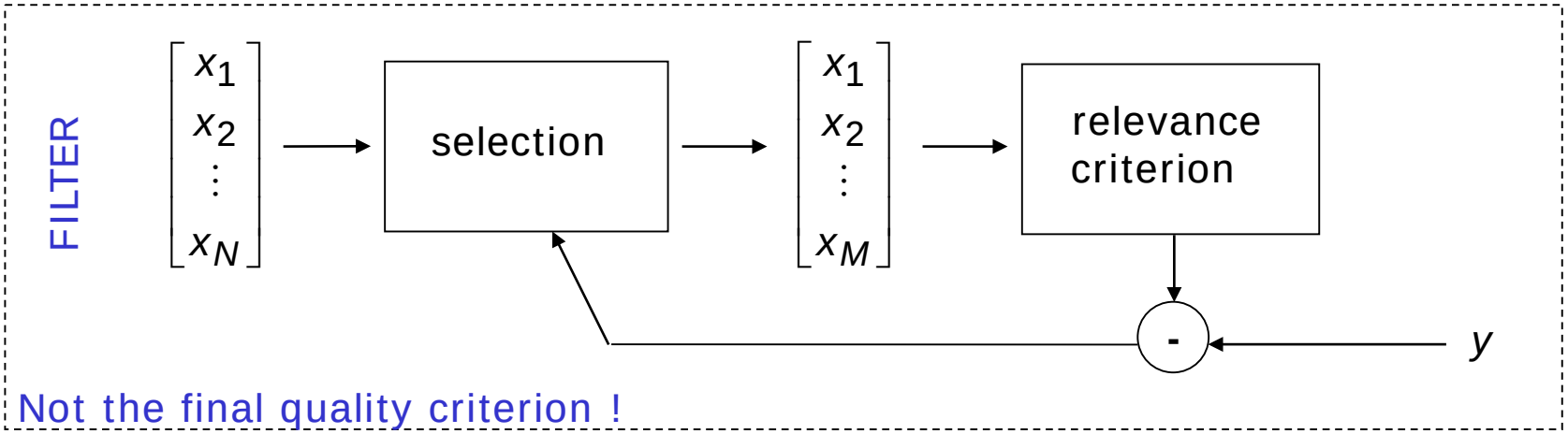
	Linear	**Nonlinear**
Selection	Correlation between inputs	Mutual information between inputs
Projection	Principal Component Analysis	Sammon's Mapping, Kohonen maps

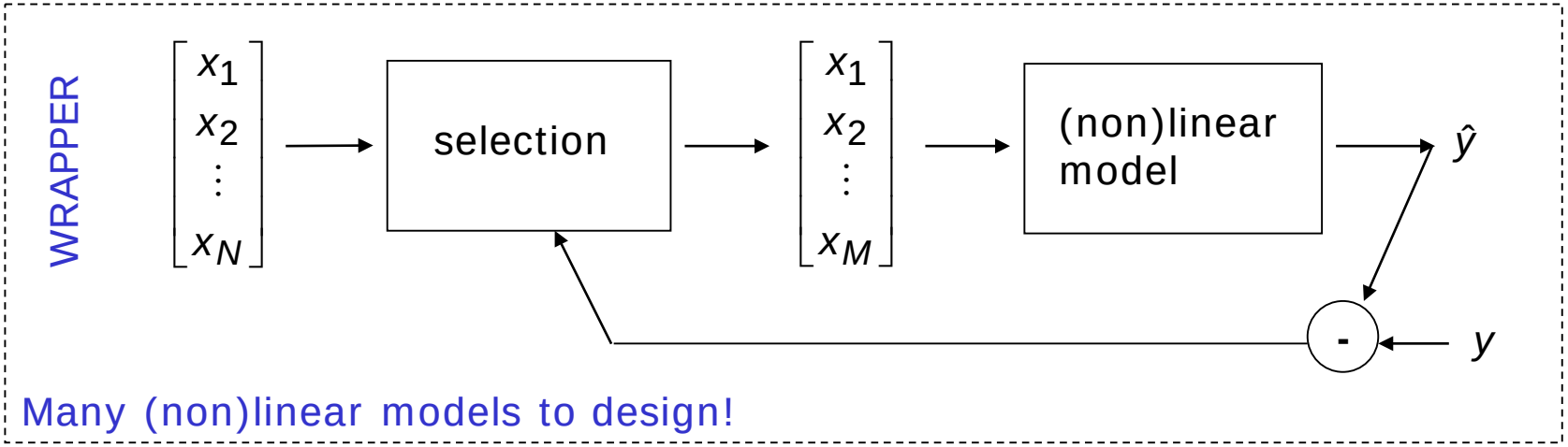
- Supervised

	Linear	**Nonlinear**
Selection	Correlation between inputs and output	Mutual information between inputs and outputs, Greedy algorithms, Genetic algorithms
Projection	Linear Discriminant Analysis, Partial Least Squares	Projection pursuit, Model (wrapper)

Supervised selection: filter versus wrapper

- Supervising does not necessarily mean to use the model!





The ingredients of feature selection

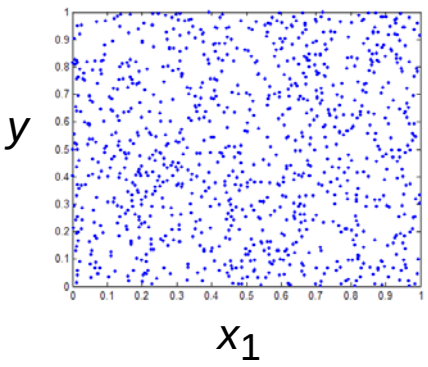
- Key Element 1 : Subset relevance assessment
  - Among all 2^d-1 possible susbsets, which is the best one?
- Key Element 2 : Optimal Subset search
  - How not to consider all 2^d-1 possible susbsets?
- Filters
  - Correlation
  - Mutual information
- Wrappers
  - Greedy search
  - Genetic algorithms

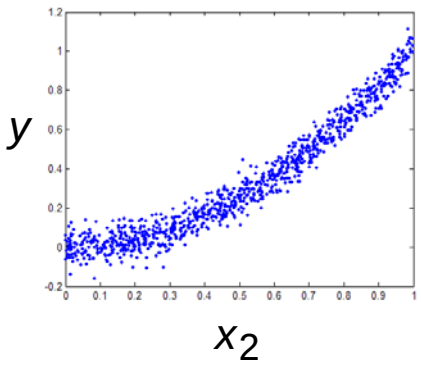
Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Subset relevance assessment

- Is x_1 relevant to predict y? What about x_2?





- Relevance is difficult to define

- Filter approach (model free): $P(y \mid x_i) \neq P(y)$
  - a variable (or set of) is relevant if it is statistically dependent on y

- Wrapper approach (uses model f): $\min_f (y - f(x_i))^2 \approx 0$
  - a variable (or set of) is relevant if the model built on it shows good performances

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Correlation, a linear filter

- Definition : correlation between random variable X and random variable Y (E[.] is the expectation operator) :

$$\rho_{xy} = \frac{E[(x - E[x]) \cdot (y - E[y])]}{\sqrt{E[(x - E[x])^2] \cdot E[(y - E[y])^2]}}$$

- Estimation : when one has a dataset $\{x^j, y^j\}$
  ($\overline{x}$ means the average of x_i)

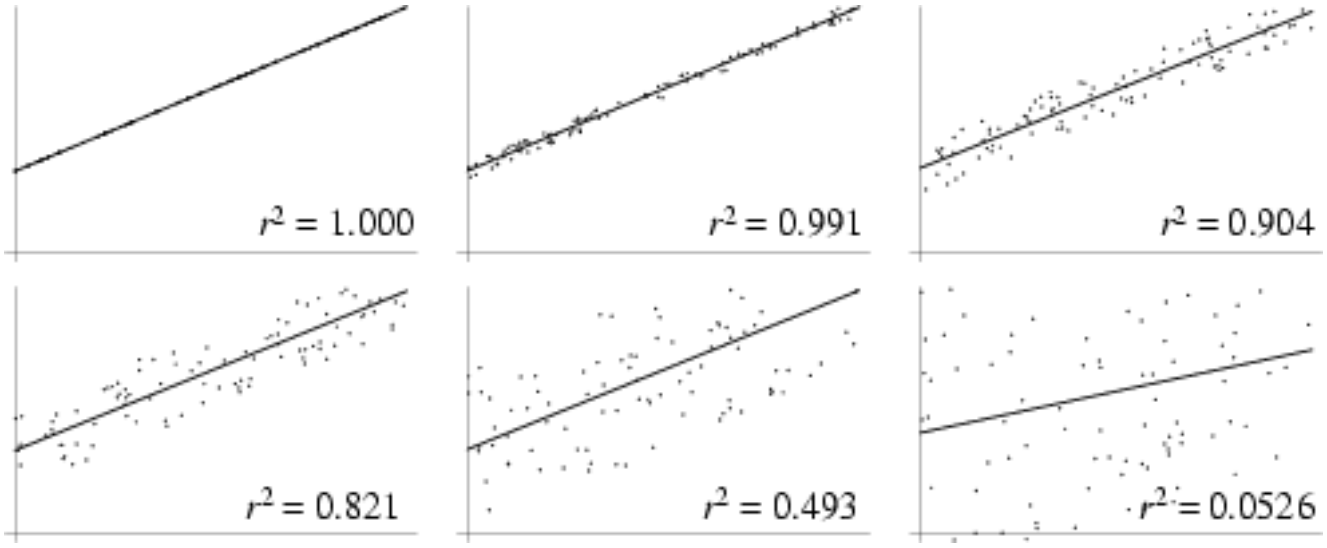
$$r = \frac{\sum_{j=1}^{N} \left((x^j - \overline{x}) \cdot (y^j - \overline{y}) \right)}{\sqrt{\sum_{j=1}^{N} \left((x^j - \overline{x})^2 \right) \cdot \sum_{j=1}^{N} \left((y^j - \overline{y})^2 \right)}}$$

- Measures linear dependencies
  - Always comprised between -1 and +1
  - 0 indicates decorrelation (no linear relation)

Correlation, a linear filter

- Examples

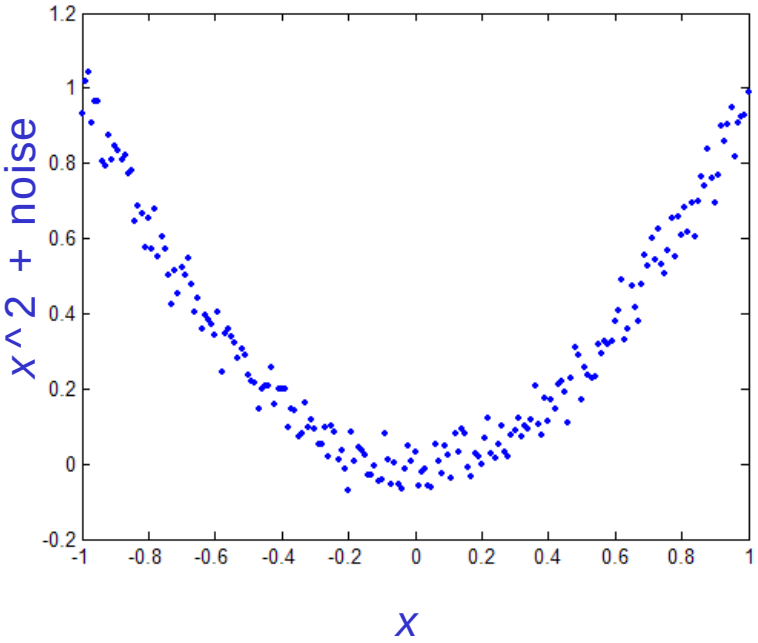
Strong correlation



Weak correlation

Correlation does not measure nonlinear relations

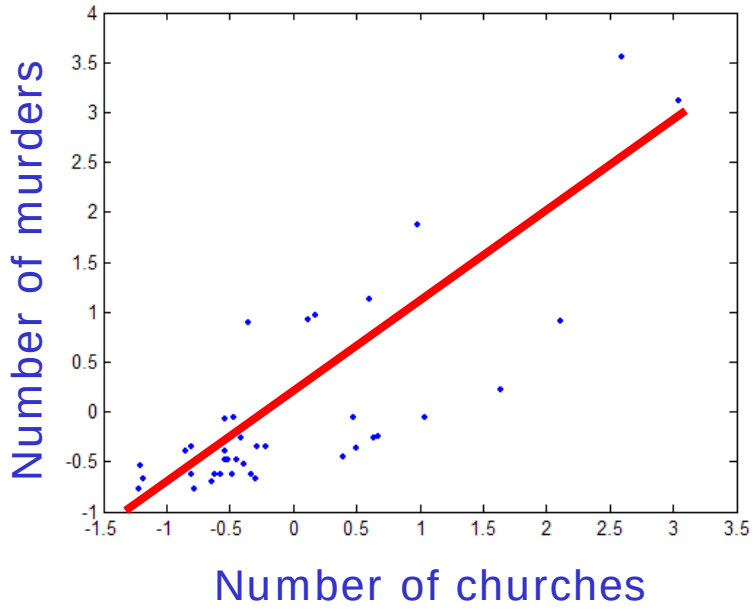
- Low correlation does not mean absence of relationship



$\rho_{xx^2} \approx 0$

Correlation does not mean causality

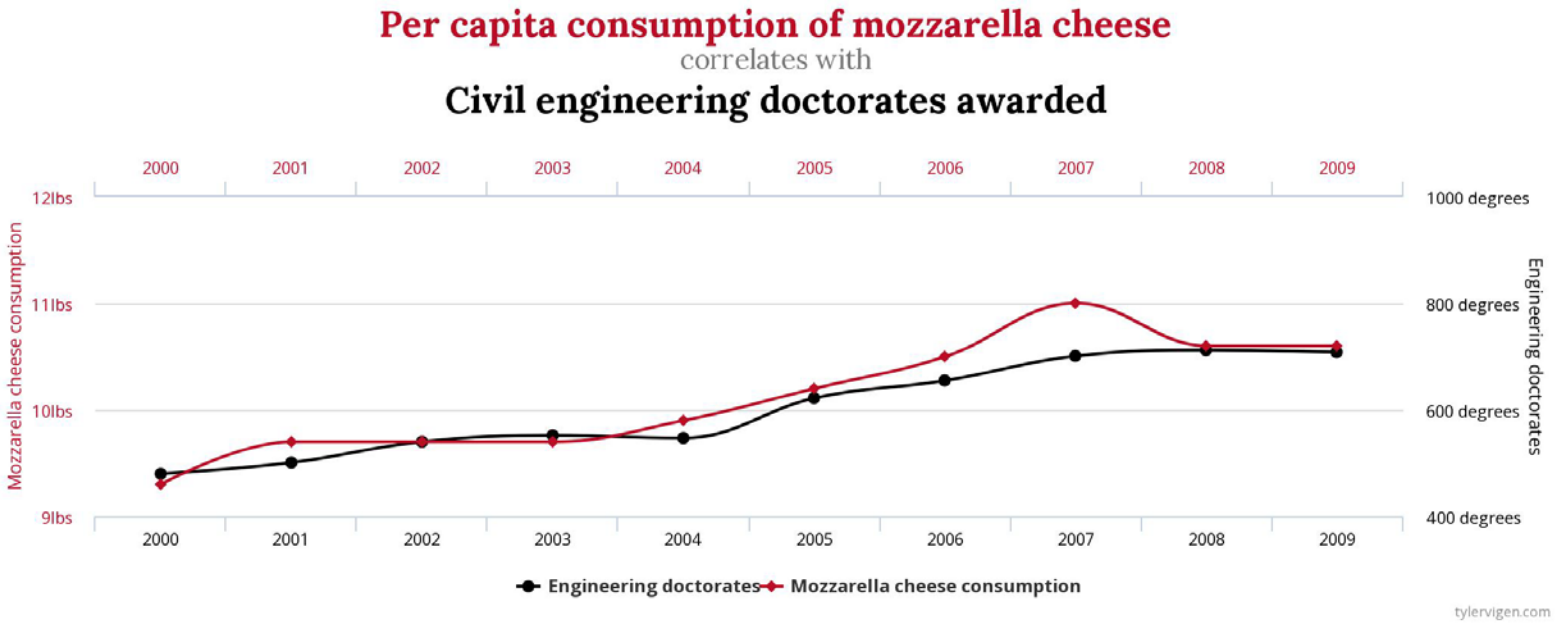
- High correlation does not mean causality
  - Number of murders in a city highly correlated (0.80) with number of churches
  - Simply because both murders and number of churches increase with population density



	christian chruches	murders 2002
Albuquerque	211	61
Atlanta	1500	152
Austin	353	25
Baltimore	466	253
Boston	370	60
Charlotte	505	67
Cleveland	980	80
Colorado Springs	400	25
Colombus	436	81
Denver	859	51
Detroit	1165	402
El paso	320	14
Fresno	450	42
Honolulu	39	18
Houston	1750	256
Indianapolis	1191	112
Jacksonville	21	3
Kansas city	1001	83
Long beach	236	67
Los Angeles	2000	654
Miami	911	65
milwaukee	411	111
Minneapolis	419	47
New Orleans	712	258
New York	2233	587
oakland	374	108
Oklahoma City	25	38
Omaha	236	26
philadelphia	963	288
Portland	498	20
St Louis	900	111
San Diego	373	47
San Francisco	540	68
San Jose	403	26
Seattle	482	26
Tucson	382	47
Tulsa	330	26
Virginia Beach	248	3
Washington	742	264

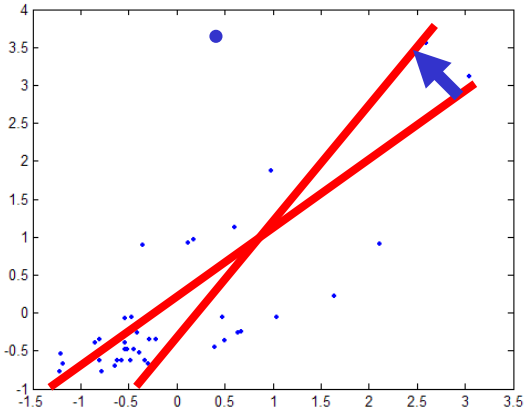
Correlation may be spurious

- 95.85% correlation:



Limitations of correlation

- Correlation
  - is linear
  - is parametric (it makes the hypothesis of a ...linear model)
  - does not *explain*
  - is almost impossible to define between more than 2 variables
  - is sensitive to outliers (R^2 = 1 - NMSE)



Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Mutual information

- Relevance of a subset X_S : mutual information $I(X_S;y)$ between this subset and the target variable Y
- What is the mutual information?
- Mutual information between random variable x and random variable y measures how the uncertainty on y is reduced when x is known. (and vice versa)
- Let's begin by the entropy...

Entropy

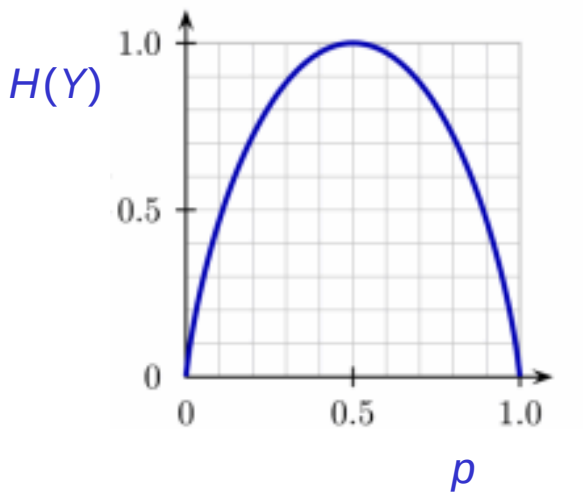
- The entropy of a random variable is a measure on its uncertainty

$$
\begin{aligned}
H(y) &= -E[\log(P[y])] \\
&= -\sum_{y \in \Omega} \log(P[y]) P[y] && \text{when } Y \text{ is discrete} \\
&= -\int \log(P[y]) \mathrm{d}y && \text{when } Y \text{ is continous}
\end{aligned}
$$

- Can be interpreted as the average number of bits needed to describe y

Example of entropy

- Entropy of a discrete variable y taking values in $\Omega = \{0,1\}$, with $P[y=1] = p$ and $P[Y=0] = 1-p$



$$H(y) = -\sum_{y \in \Omega} P[y] \log(P[y])$$

- Uncertainty is maximal when both events have the same probability

Conditional entropy

- Conditional entropy H(y |x) measures the uncertainty on y when x is known

$$H(y \mid x) = H(y, x) - H(x)$$

- If Y and X are independent,

$$H(y \mid x) = H(y)$$

the uncertainty on Y is the same if we know X as if we don't !

Mutual information

- Mutual information between x and y

$$I(y;x) = H(y) - H(y \mid x) = H(x) - H(x \mid y)$$

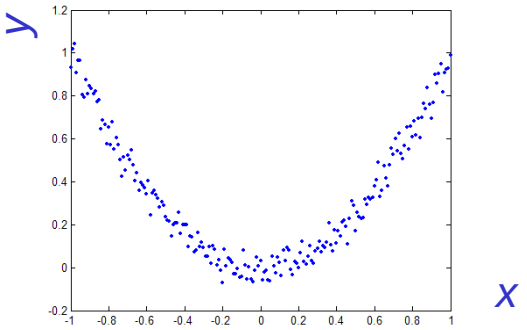
- Difference between entropy of y and entropy of y when x is known

- Some properties:
  - If x and y are independent, $I(y;x) = 0$
  - $I(y;y) = H(y)$
  - $I(y;x)$ is always non negative and less than $\min(H(y), H(x))$

Nonlinear dependencies with MI

- Mutual information identifies nonlinear relationships between variables

- Examples:
  - x uniformly distributed over [-1 1]
  - $y = x^2$ + noise
  - z uniformly distributed over [-1 1]
  - z and x are independent



- Results:

1000 samples	y,y	x,y	z,y
Correlation	1	0.0460	0.0522
Mutual information	2.2582	1.1996	0.0030

High-dimensional mutual information

- What about the relevance of a set of features?
- Reminder:

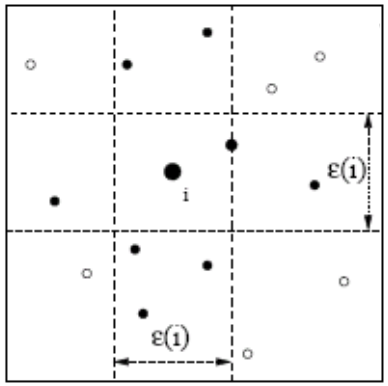
$$I(x,y) = H(y) - H(y \mid x) = H(x) - H(x \mid y)$$

- x and y may be vectors!
- If X is a subset of features, its relevance may still be evaluated
- Evaluating subsets is the right issue!

- The difficulty is in the *estimation* of $I(y;x)$:
  - histograms, kernels and splines suffer from the curse of dimensionality!
  - k-NN based estimators are the (almost only) solution

Estimators for mutual information

- All estimators suffer from the curse of dimensionality !
- Histograms, kernels (Parzen windows,etc.) are the worst...
- *k*-NN based estimators (Kraskov) are more robust



Relevance and redundancy

- In practice: many methods based on the estimation of the mutual information on a limited subset of features (2, 3, few...)
- Relevance: $I(x_i, y)$
- Redundancy: $I(x_i, x_j)$
- Principle when deadling with bivariate mutual information only: Max. relevance and min. redundancy together!
  - this is a multi-objetcive criterion
  - depends on the weighting between the two criteria
- $\exists$ extensions to tri-variate mutual information (interaction measure), etc.

Stopping criterion

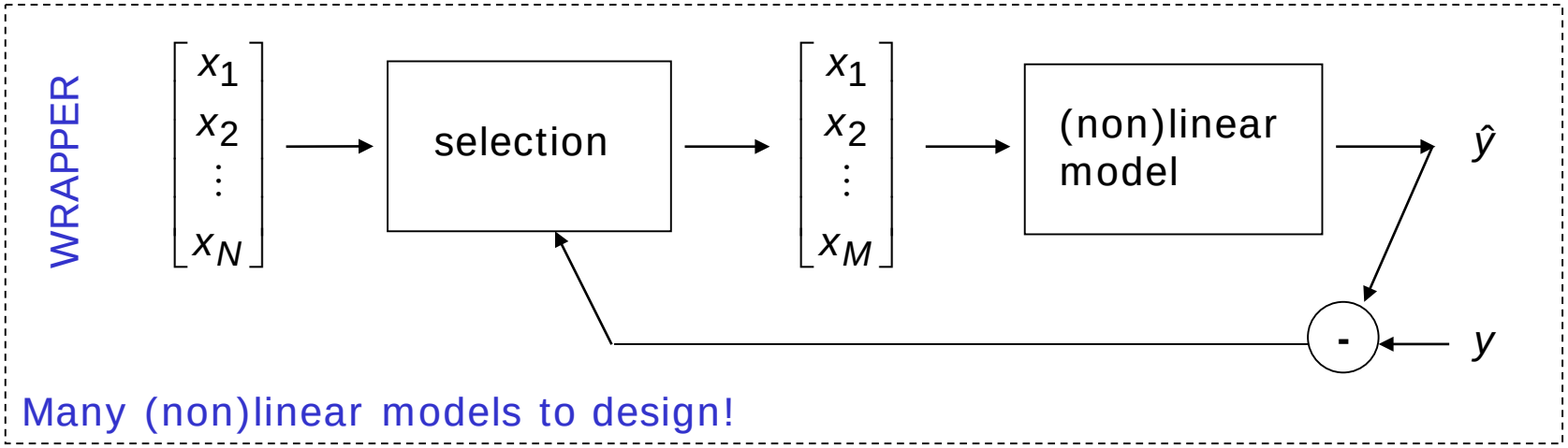
- In theory, the mutual information (of a set) can only increase when adding variables

- When to stop then?
  - Often: when the estimation of the MI decreases (bad idea!)
  - Better: evaluate the statistical relevance (hypothesis test) of the addition of a new variable (ex: permutation test)
  - For external reasons: fix the max. number of variables

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Wrappers



- Just build the models, and evaluate them...
- Problems come when the models are "computationally intensive"
  - need to train each of them
  - need to set the hyper-parameters (cross-validation, etc.)
  - need to evaluate the performances (double cross-validation, etc.)

Wrappers versus filters

FILTERS	WRAPPERS
+ Fast : build only one model **+ Intuitive** : identifies statistical dependency	**+** Relevance criterion **easy to estimate** **+ Model-aware** : identifies optimal subset to build optimal model
- Relevance criterion **hard to estimate** **- Model-ignorant** : most relevant subset might not be optimal for subsequent model	**- Slow** : must build lots of models **- Not intuitive** : features for best model might not actually be most explanatory variables

Outline

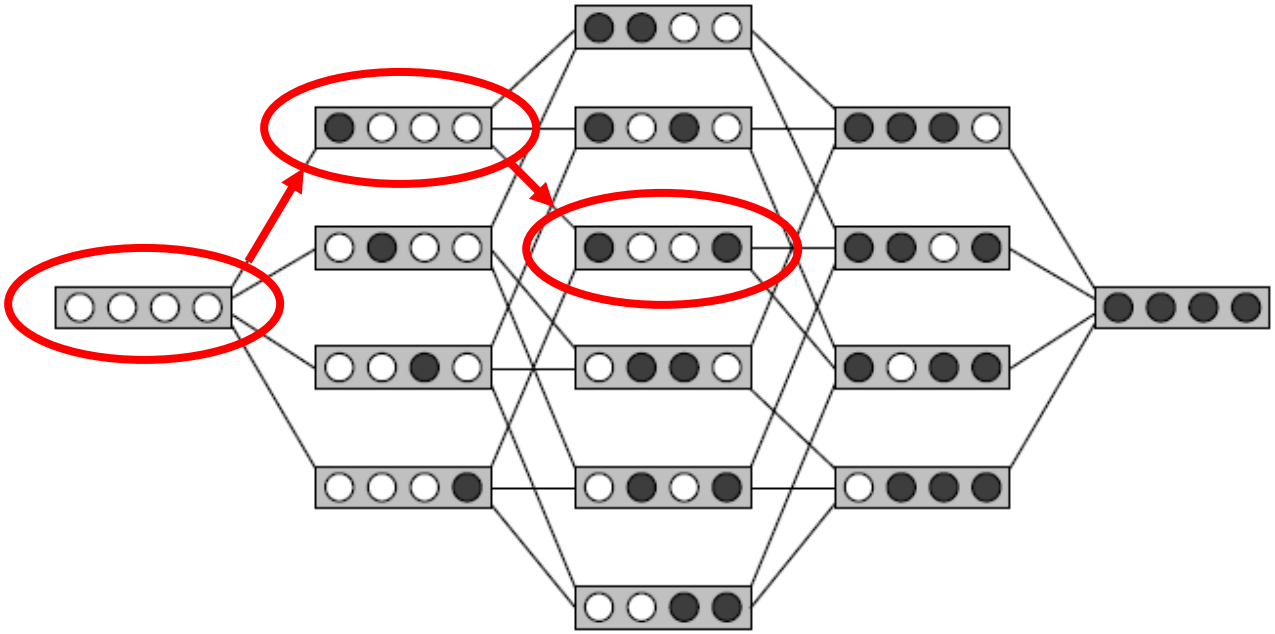
- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Optimal subset search

- In theory: just try 2^d subsets and evaluate them...
  - Evaluation: mutual information (filters), or model itself (wrapper)
  - just imagine for d=200 ☺
- In practice: greedy procedure
  - Define an intial subset
  - Choose a strategy to update subset
  - Decide when to stop

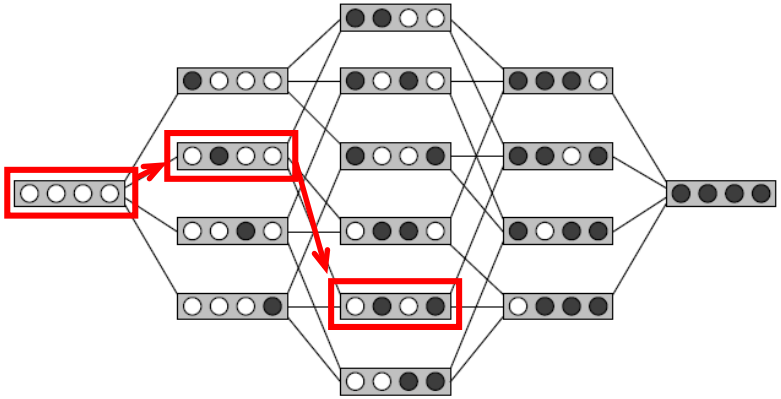
Greedy subset search

- Example with 4 features



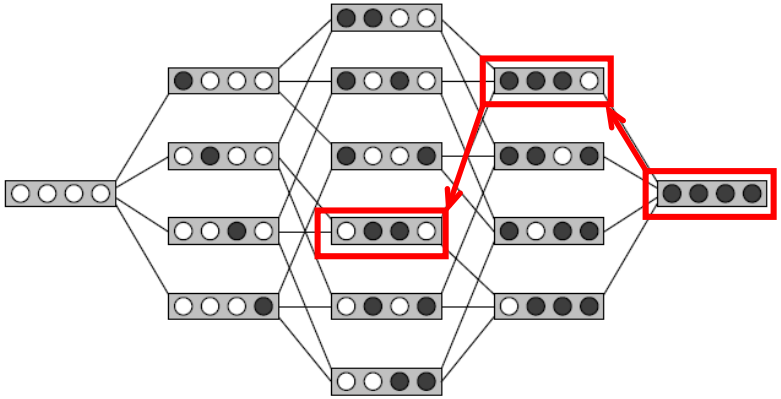
- Makes typothesis that best subset can be constructed iteratively

Forward search



- Define an intial subset
  - begin with empty set
- Choose a strategy to update subset
  - Filter: add feature that increases the most the relevance/redundancy compromize
  - Wrapper: add feature that increases the most the performances of the model
- Decide when to stop
  - Filter: needs a supplementary criterion
  - Wrapper: stop when adding a feature increases the generalization error

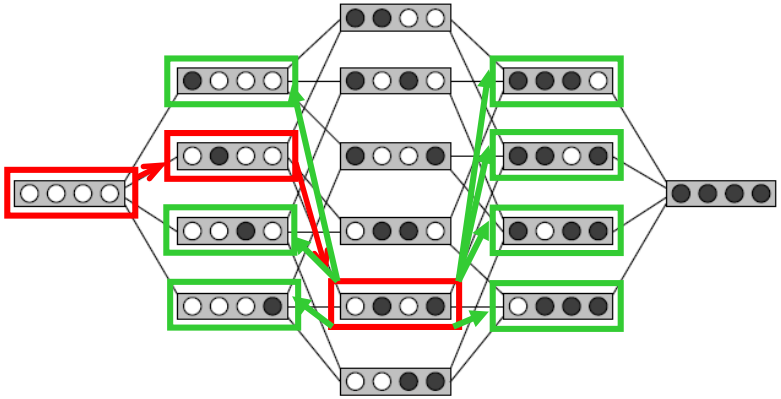
Backward search



- Define an intial subset
  - begin with the full set
- Choose a strategy to update subset
  - Filter: remove feature that increases most the relevance/redundancy compromize
  - Wrapper: remove feature that increases the most the performances of the model
- Decide when to stop
  - Filter: needs a supplementary criterion
  - Wrapper: stop when removing a feature increases the generalization error

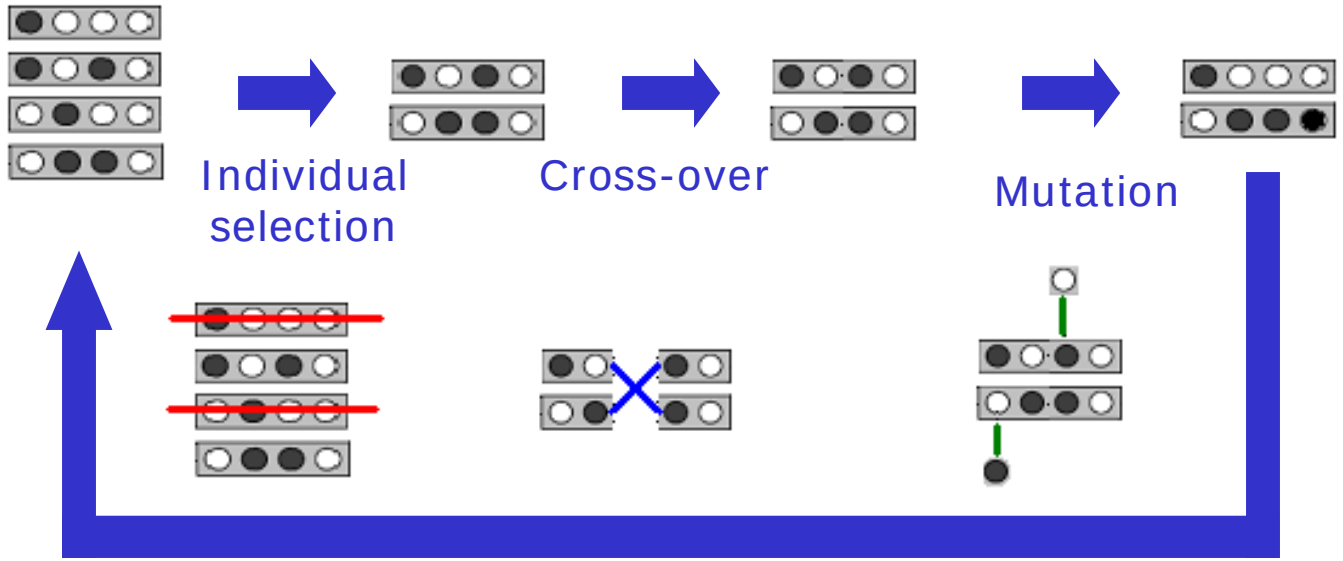
Variants

- Forward-backward
  - At each step, consider all additions and removals of one variable, and select the best result
  - Wrapper: this makes sense
  - Filter: in theory, the mutual information cannot increase with less variables



Genetic algorithms

- "clever" random exploration of the space of subsets
  1. Draw initial population (candidate subsets)
  2. Select individuals
  3. Apply cross-overs and mutations on individuals
  4. Repeat from 2 until a new population is generated
  5. Select best individuals and repeat from 1



Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Embedded methods

- Idea: to build the model and restrict the number of features used by the model together
- Appealing idea! Optimizing the whole can only be better than separating into 2 parts (feature selection and model)

- Example: LASSO
  - Linear model $y = W.X$
  - Criterion: $\min_w \underbrace{\frac{1}{N}\sum_{j=1}^{N}\left(y^j - Wx^j\right)^2}_{\text{training error}} + \underbrace{\mu\|W\|_1}_{\text{regularization}}$
  - Usually (ridge regression): regularization term is a 2-norm
  - Here a 1-norm (this reduces the effective number of features used by the model, both in theory and in practice)

- Many interesting directions for embedded methods !

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Illustrative examples

- Housing (a traditional benchmark)
  - to show that it can work
- Business plan classification
  - to understand
- IR Spectra
  - to analyze data
- Time series
  - how to choose the regressor?

Outline

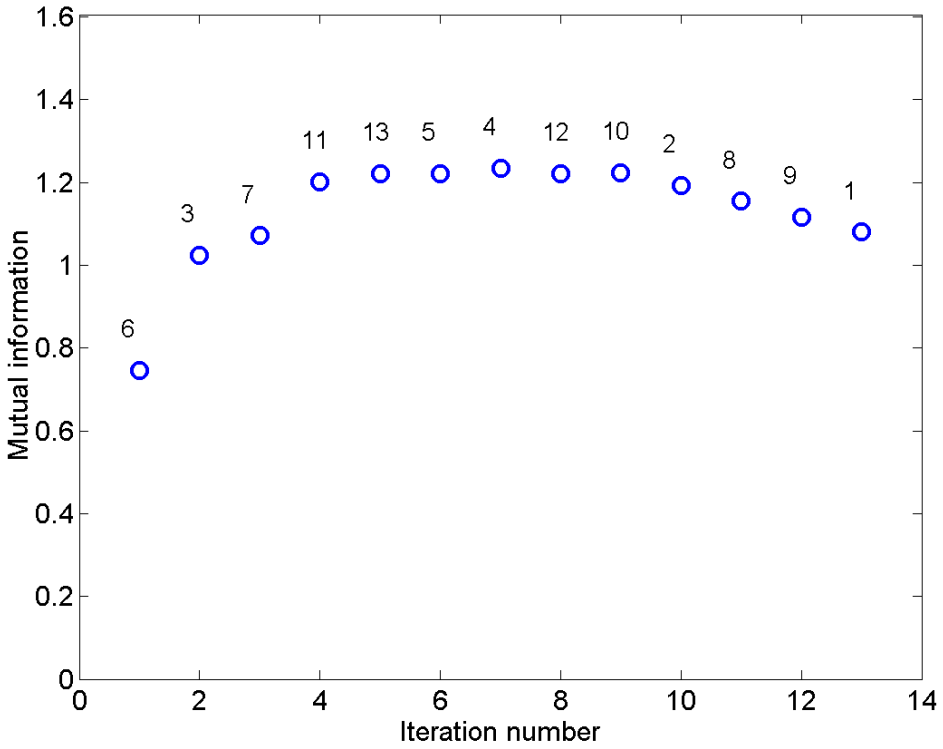
- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Housing

- Dataset origin: StatLib library, Carnegie Mellon Univ.
- Concerns housing values in suburbs of Boston
- Attributes:
  1. CRIM per capita crime rate by town
  2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS proportion of non-retail business acres per town
  4. CHAS Charles River dummy variable (= 1 if tract bounds river, 0 otherw.)
  5. NOX nitric oxides concentration (parts per 10 million)
  6. RM average number of rooms per dwelling
  7. AGE proportion of owner-occupied units built prior to 1940
  8. DIS weighted distances to five Boston employment centres
  9. RAD index of accessibility to radial highways
  10. TAX full-value property-tax rate per $10,000
  11. PTRATIO pupil-teacher ratio by town
  12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT % lower status of the population
  14. MEDV Median value of owner-occupied homes in $1000's

Housing

- Forward selection by mutual information between output and set of attributes



Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Business plan classification

- Context: a competition of business plans evaluated by experts
- The question: is the success of a company in relation with the scores given by the experts?
- 161 business plans
- 7 criteria:
  - interest for investor
  - content of the business plan
  - usefulness for clients
  - differentiation with other products
  - size of market
  - competitors
  - global rating
- To predict: success of company after two years

Business plan classification

- Important variables are evaluated by:
  - correlation
  - difference of means
  - mutual information

- Results

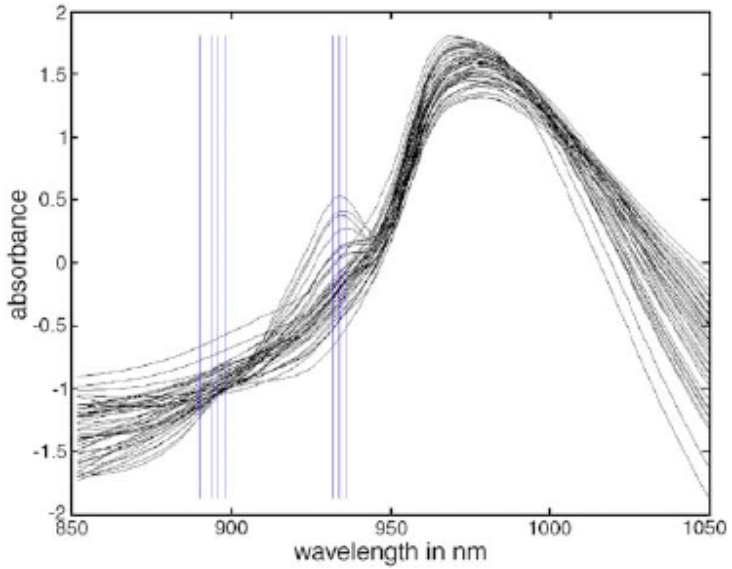
	high	low
correlation	3,4	7,1,2,5,6,7
difference of means	3,4,7	1,2,5,6
mutual information	3,4,5	7,1,2,6,7

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Infrared spectra (Tecator)

- Meat spectra (Tecator)



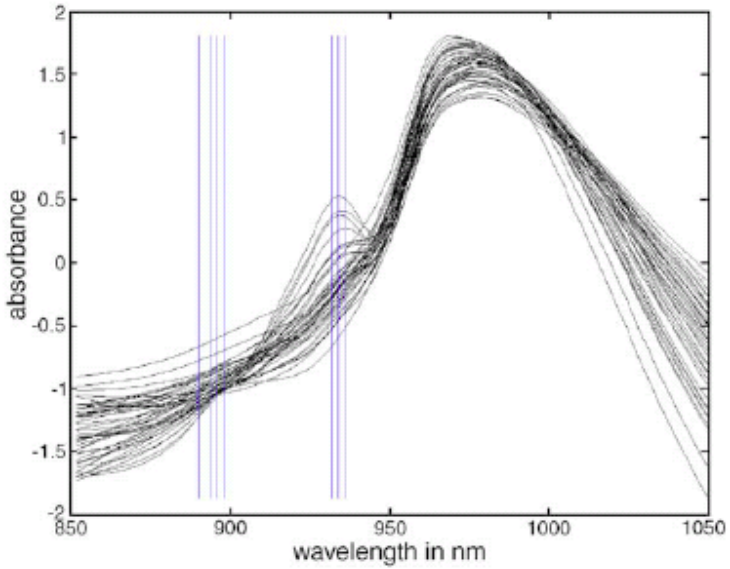
- Prediction of fat content
- 100 wavelengths
- 172 training samples, 43 test samples

Infrared spectra (Tecator)

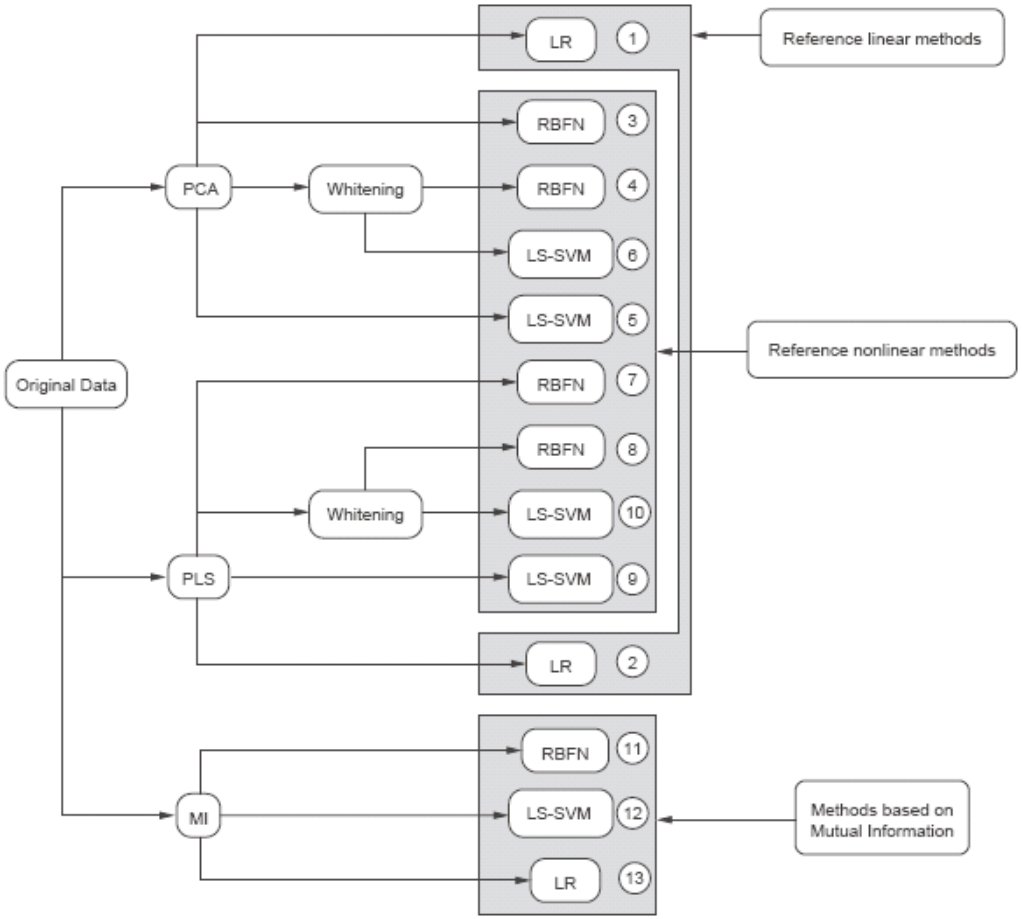
- Original methodology (as an example, not a rule…)
- 100 wavelengths (sometimes 1000, 2000…) -> very high-dimensional data
- estimation of MI becomes hard
- Two solutions:
  - ranking: $I(x_i;i)$
    does not take relations bewteen Xi into account
  - forward selection: $I((x_i, x_j, x_k, \dots);y)$
    curse of dimensionality in the estimation
- Take both!
  - set of features selected by forward (until decrease)
  - set of features selected by ranking (until the union of both sets reaches N elements)
  - try the 2^N IM evaluations (filter) or the 2^N models (wrapper)

Infrared spectra (Tecator)

- Only 7 variables are selected
- In two ranges
- Probably information about first and second derivatives



Infrared spectra (Tecator)

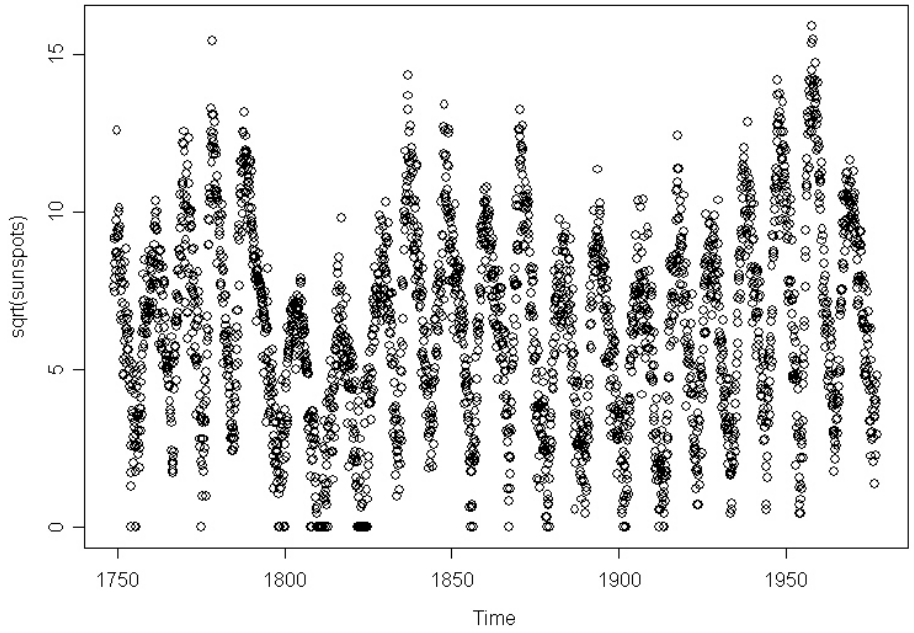


model	variables	NMSEt
1	42	1.64 e-2
2	20	1.36 e-2
3	42	4.6 e-3
4	42	1.78 e-2
5	42	7.8 e-2
6	42	6.08 e-2
7	20	8.27 e-3
8	20	5.2 e-3
9	20	1.35 e-2
10	20	1.12 e-1
11	7	6.60 e-3
12	7	2.70 e-3
13	7	2.93 e-2

Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Time series



- Model: $x(t+1) = f(x(t), x(t-1), x(t-2), x(t-3), x(t-4), \ldots)$

Time series

- Three questions
  - sampling frequency

$$x(t+1) = f(x(t), x(t-1), x(t-2), x(t-3), x(t-4), \ldots)$$
$$\text{or} \quad x(t+1) = f(x(t), x(t-1), x(t-4), x(t-7), x(t-10), \ldots)$$

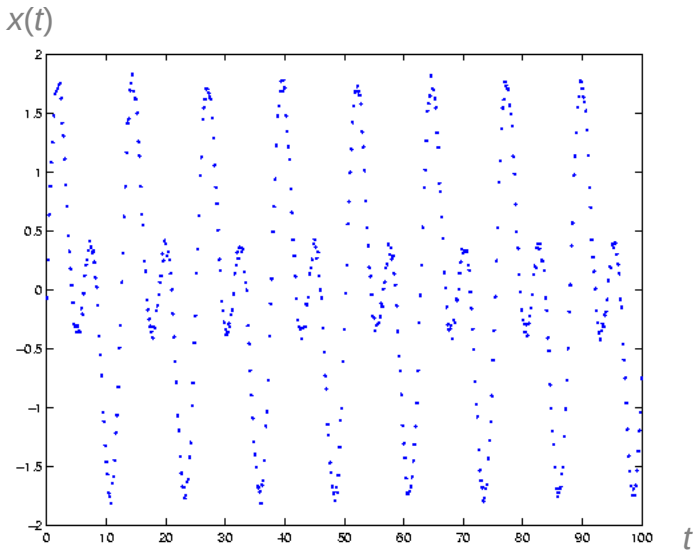
  - size of regressor

$$x(t+1) = f(x(t), x(t-1), x(t-2), x(t-3), x(t-4), \ldots)$$
$$\text{or} \quad x(t+1) = f(x(t), x(t-1), x(t-2), x(t-3))$$

  - non-contiguous values

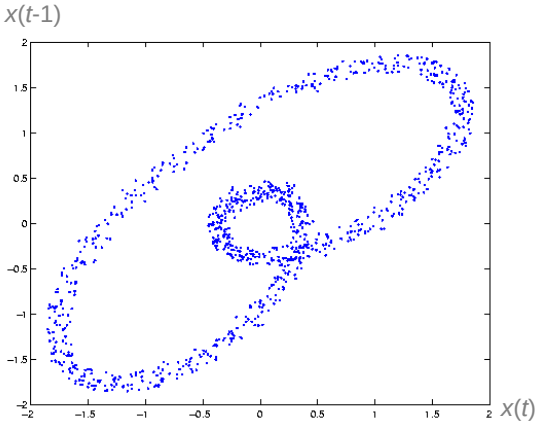
$$x(t+1) = f(x(t), x(t-1), x(t-2), x(t-3), x(t-4), \ldots)$$
$$\text{or} \quad x(t+1) = f(x(t), x(t-1), x(t-7), x(t-8), x(t-14), \ldots)$$

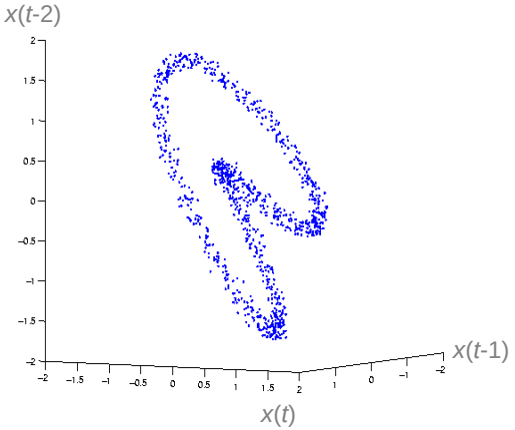
A digression about the size of regressors



time series

intrinsic dimension (q) = 1





- Takens' theorem:
  $q \leq$ size of regressor $\leq 2q+1$
  (AR model)

A digression about the size of regressors

- Forecasting: Taken's theorem

  $q \leq$ size of regressor $\leq 2q+1$

- In the $2q+1$ space, there exists a q-surface without intersection points

- Projection from $2q+1$ to q possible !

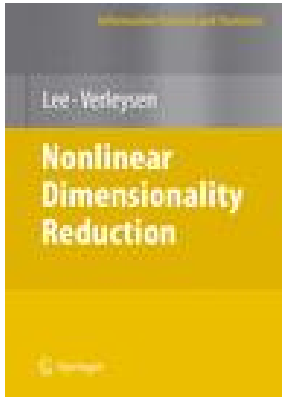
- In practice
  - Take many input variables
  - Select $2q+1$ by feature selection
  - If needed, project them on q new variables

Selection / projection

- Selection
  (choosing among the original features which ones to keep)
  - + easy
  - + interpretability of the features

- Projection
  (creating new features from the original ones)
  - + more general → possibly more efficient
  - - more difficult
  - - features not interpretable

  – The book:
    - Nonlinear Dimensionality Reduction
      Springer, Series: Information Science and Statistics
      John A. Lee, Michel Verleysen, 2007
      300 pp., ISBN: 978-0-387-39350-6



Conclusions

- Feature selection = two ingredients
  - Subset evaluation criterion
  - Greedy search in the space of subsets
- Mutual information is a good “filter” criterion
  - But difficult to evaluate
  - Many options to make a compromise between relevance and redundancy
- Wrapper approach is better (for prediction), but computationally (too) intensive

Some good questions (and not so good answers...)

- Is most relevant feature necessarily optimal for prediction ?
  - No for example if the model is not powerful enough
- Is most useful feature for prediction necessarily relevant ?
  - No for example bias terms are irrelevant but can help for prediction
- Are two redundant features really unuseful together ?
  - Perfectly redundant features are useless
  - Highly correlated features can be useful if the model is able to use the 'information' contained in their differences
- Can two features be useless alone and useful together ?
  - Yes, think of the XOR problem