Learning high-dimensional data

Michel Verleysen
Université catholique de Louvain (Louvain-la-Neuve, Belgium)
Electricity department

January 2002

Acknowledgements

- Part of this work has been realized in collaboration with PhD students:
  - Philippe Thissen
  - Jean-Luc Voz
  - Amaury Lendasse
  - John Lee
- Some ideas and figures come from
  - D. L. Donoho, High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. Lecture on August 8, 2000, to the American Mathematical Society "Math Challenges of the 21st Century". Available from http://www-stat.stanford.edu/~donoho/.

Data mining

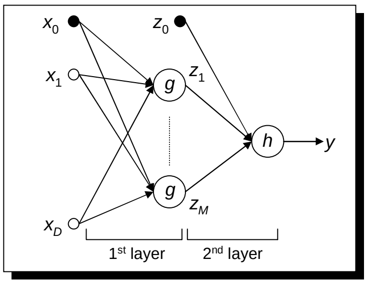
N lines (number of observations)

$$\overbrace{\begin{pmatrix} 1.2 & 7.5 & -1.9 & 2 & \cdots & 1.9 \\ -7.6 & 12 & 17.2 & 2.4 & \cdots & 1.5 \\ -8.5 & 13 & 14 & 8.5 & \cdots & -1.9 \\ 9 & 5.4 & -5.2 & 8.2 & \cdots & 9.4 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 25.1 & 5.2 & -9.1 & -8.5 & \cdots & 5.4 \end{pmatrix}}^{D \text{ colums (dimension of space)}}$$

- Data mining: find information in large databases (large = *D* and/or *N*)

High-Dimensional spaces

- Inputs = **H**igh-**D**imension (HD) vectors (*D* is large)
  - many parameters in the model
  - local minima
  - slow convergence



- The questions

> - Learning algorithms in HD spaces ?
> - Local learning or not ?

- The arguments
  - real data are seldom HD (concept of intrinsic dimension)
  - local learning is not worse than global learning…

Contents

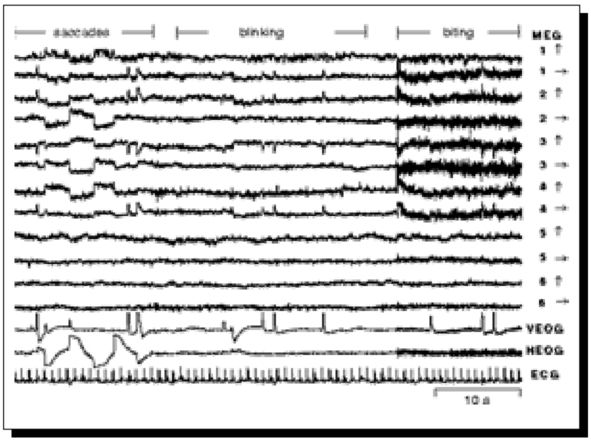
- High-dimensional data
  - Surprizing results
  - Intrinsic dimension
- Local learning
  - Use of distance measures
- Dimension reduction
  - non-linear projection
  - application to time-series forecasting

Contents

- High-dimensional data
  - Surprizing results
  - Intrinsic dimension
- Local learning
  - Use of distance measures
- Dimension reduction
  - non-linear projection
  - application to time-series forecasting

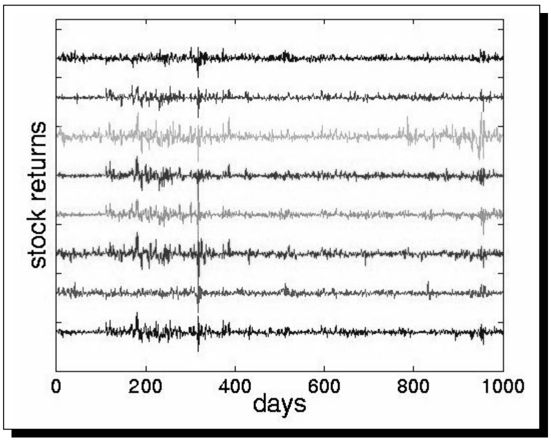
Data mining: large databases

- (biomedical) signals



Data mining: large databases

- financial data



Data mining: large databases

- imagery

Data mining: large databases

- recordings of consumers' habits (credit cards)
- bio- data (human genome, etc.)
- satellite images
- hyperspectral images
- …

John Wilder Tukey

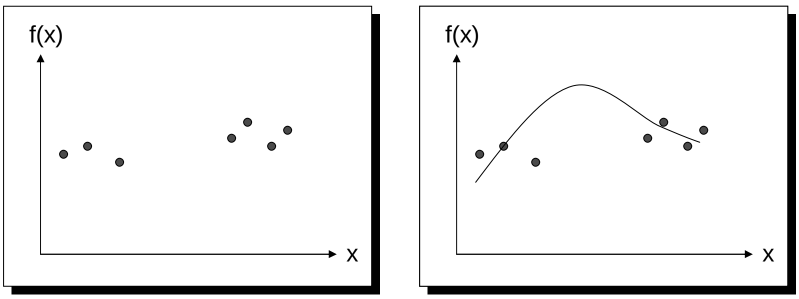
- The Future of Data Analysis, *Ann. Math. Statis.*, 33, 1-67, 1962.

« Analyze data rather than prove theorems… »

- In other words:
  - data are here
  - they will be coming more and more in the future
  - we must analyze them
  - with very humble means
  - insistence on mathematics will distract us from fundamental points

From D.L. Donoho, op. cit.

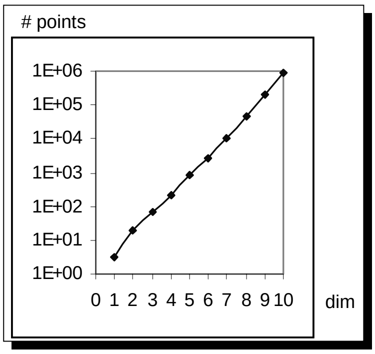
Empty space phenomenon



- Necessity fo *fill* space with learning points
- # learning points exponential with dimension

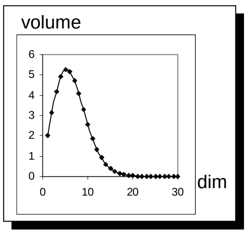
Example: Silvermann's result

- How to approximate a Gaussian distribution with Gaussian kernels
- Desired accuracy: 90%

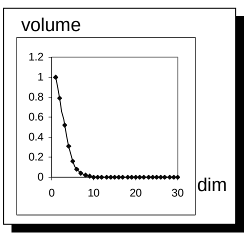


Surprizing phenomena in HD spaces

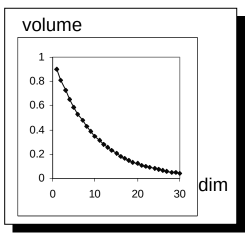
- Sphere volume



- Sphere volume / cube volume

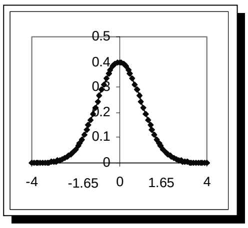


- Embedded spheres (radius ratio = 0.9)

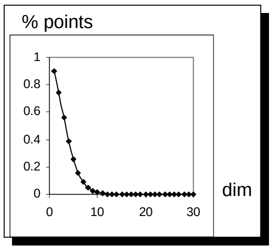


Gaussian kernels

- 1-D Gaussian



- % points inside a sphere of radius 1.65



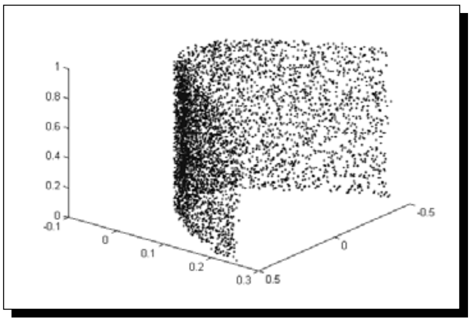
Concentration of measure phenomenon

- Take all pairwise distances in random data
- Compute the average *A* and the variance *V* of these distances
- If *D* increases then
  - *V* remains fixed
  - *A* increases
- All distances seem to concentrate !!!

- Example: Eucliden norm of samples
  - average *A* increases with $(D)^{0.5}$
  - variance *V* remains fixed
  - → samples seem to be normalized !

Intrinsic dimension

- No definition
- Example:

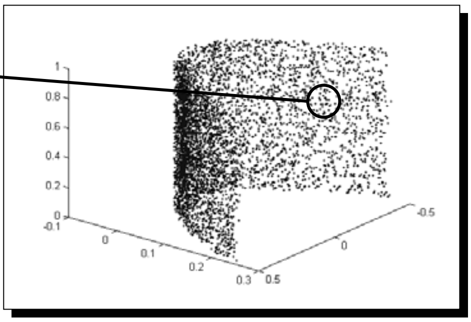


From "Analyse de données par réseaux de neurones auto-organisés », P. Demartines, Ph.D. thesis, Institut National Polytechnique de Grenoble (France), 1994.

Estimation of intrinsic dimensions 1/4

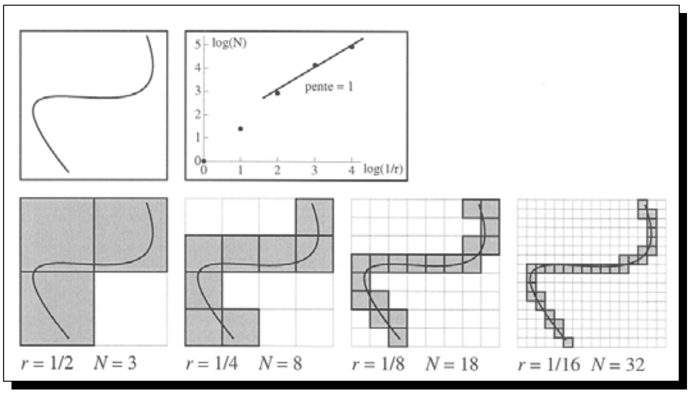
local PCA	a posteriori estimation
box counting	Grassberger-Proccacia

- small region → approximately plane
- PCA applied on small regions



Estimation of intrinsic dimensions 2/4

local PCA	a posteriori estimation
box counting	Grassberger-Proccacia



From P. Demartines, op. cit.

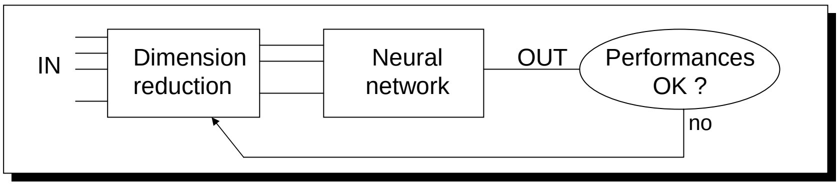
Estimation of intrinsic dimensions 3/4

local PCA	a posteriori estimation
box counting	Grassberger-Proccacia

- Similar to box counting
- Mutual distances between pairs of points
- Advantage: N(N-1)/2 distances for N points
- log *N* –log *r* graph: number *N* of pairs of points closer than *r*

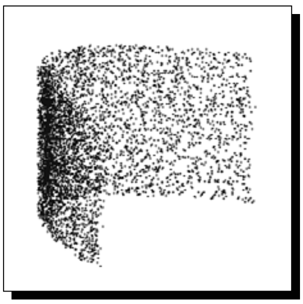
Estimation of intrinsic dimensions 4/4

local PCA	a posteriori estimation
box counting	Grassberger-Proccacia

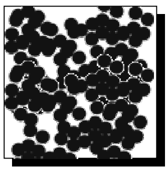


- example: forecasting problem

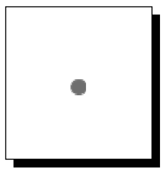
Intrinsic dimension: limitation of the concept



- Seen from very close



- Seen from very far



Contents

- High-dimensional data
  - Surprizing results
  - Intrinsic dimension
- Local learning
  - Use of distance measures
- Dimension reduction
  - non-linear projection
  - application to time-series forecasting

Local learning

- « Local »: by means of local functions
- Typical example: Gaussian kernels
- Radial Basis Function Networks

$$f(\mathbf{x}) = \sum_{j=1}^{P} w_j K_j(\mathbf{x}) \qquad K_j(\mathbf{x}) = exp\left(-\frac{\left\|\mathbf{x}-\mathbf{c}_j\right\|^2}{h_j^2}\right)$$

- Local **and** global: ANN are interpolators, and do not extrapolate to regions without learning data !

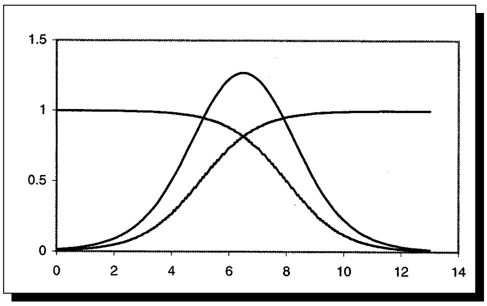
Radial-Basis Function Networks (RBFN) for approximation

$$f(\boldsymbol{x}) = \sum_{j=1}^{P} w_j K_j(\boldsymbol{x}) \qquad K_j(\boldsymbol{x}) = exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{c}_j\right\|^2}{h_j^2}\right)$$

- Advantages (over MLP, …):
  - splitted computation of - centres c_j
    - widths h_j
    - multiplying factors w_j
  - easier learning

Local learning 2/2

- Sum of sigmoids = Gaussian !



Problem with (local ?) learning

- Most ANN use *distances* between input and weight vectors:
  - RBFN: as argument to radial kernels
  - VQ and SOM: to choose the « winner »
  - first layer in MLP: ***x w***
- In high-dimensional spaces: all these distances seem identical !
  (concentration of measure phenomenon)
- $\rightarrow$ need for
  - other neural networks
  - same neural networks, with other distance measures

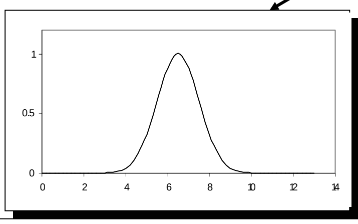
Using new distance measures

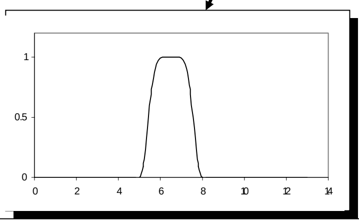
- Example: RBF

$$f(\mathbf{x}) = \sum_{j=1}^{P} w_j K_j(\mathbf{x}) \qquad K_j(\mathbf{x}) = exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{h_j^2}\right)$$

- Use of « super-Gaussian » kernels:

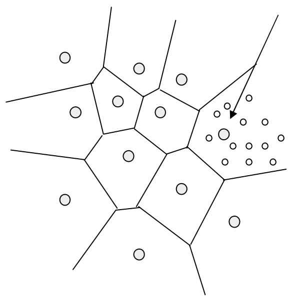
$$K_j(\mathbf{x}) = exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_j\|^r}{h_j^r}\right)$$





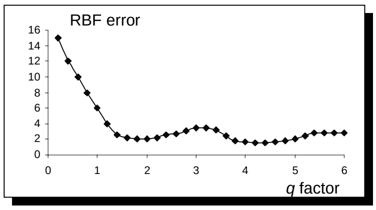
Not assuming evidence…

- widths of kernels in RBF are not necessarily equal to the STDV of samples in the Voronoi zone!



$$\sigma_j = \mathrm{STDV}(\text{points in Voronoi zone})$$

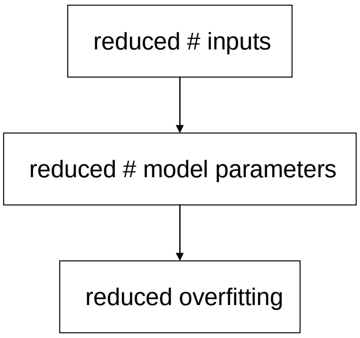
$$K_j(\boldsymbol{x}) = exp\left(-\frac{\left\|\boldsymbol{x}-\boldsymbol{c}_j\right\|^2}{q\sigma_j^2}\right)$$



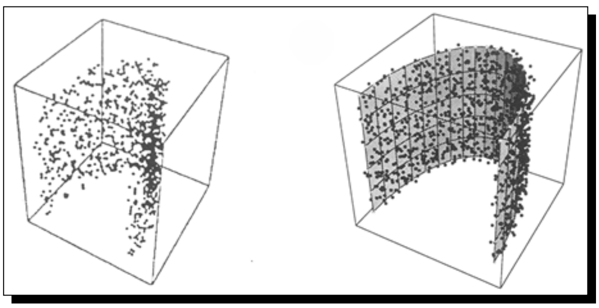
Contents

- High-dimensional data
  - Surprizing results
  - Intrinsic dimension
- Local learning
  - Use of distance measures
- Dimension reduction
  - non-linear projection
  - application to time-series forecasting

Dimension reduction



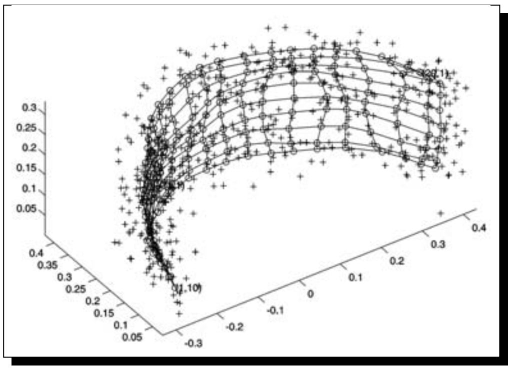
Dimension reduction & intrinsic dimension



- Questions:
  - intrinsic dimension is unknown
  - non-linear submanifolds

Kohonen maps

- Based on ***topology*** preservation



Michel Verleysen 33

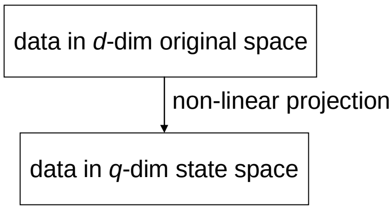
Distance-preservation methods

- Many non-linear projection methods:
  based on ***distance*** preservation between input and output pairs
- Examples
  - Multi-dimensional scaling (MDS)
  - Sammon's mapping
  - Curvilinear Component Analysis
- Principle: place points in the output space, so that pairwise distances are as equal as possible to the corresponding pairwise distances in the input space
  - Impossible to respect *all* pairwise distances → insist on small ones (*local* pairs)

Michel Verleysen 34

Curvilinear Component Analysis

- Principle

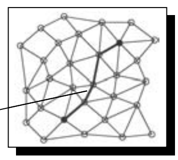


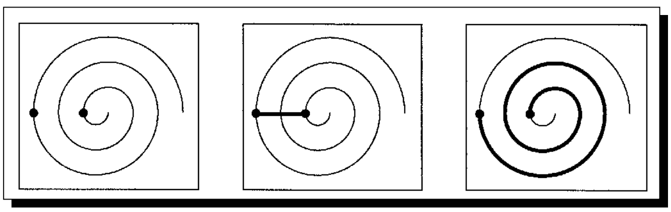
- $q < d$
- respect of mutual distance between pairs of points

$$E = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(X_{ij} - Y_{ij}\right)^2 F\left(Y_{ij}, \lambda\right)$$

Adjusting parameters in CCA

- Need for curvilinear distance
  - VQ $\rightarrow$ centroids
  - linking centroids
  - measuring the distances via the links

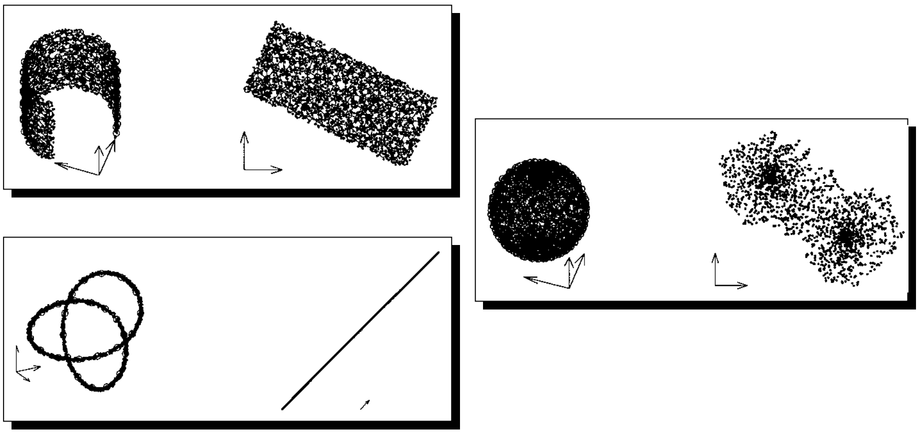




- In practice: generalized distance

$$\Delta_{ij} = \left(1 - \omega(t)\right) X_{ij} + \omega(t) \delta_{ij}$$

CCA: examples



Application of dimension reduction to forecasting tasks

- Regressor: - past values $x(t\text{-}i)$
  - exogenous data $in(j)$
- Forecasting

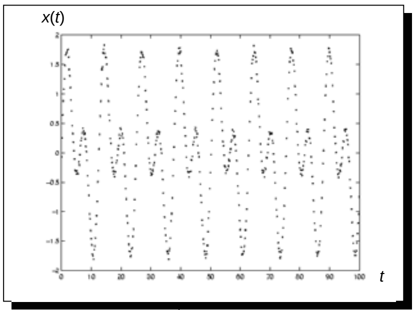
$$x(t+1) = f\left(x(t), x(t-1), \ldots, x(t-k), in(1), in(2), \ldots, in(l)\right)$$

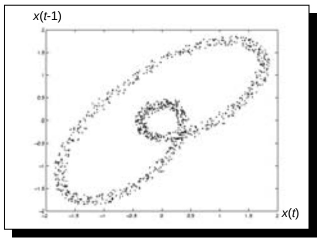
- Non-linear forecasting:
  - 1. optimise regressor on linear predictor
    2. use the same regressor with non-linear predictor *f*
  - trials and errors (computational load !)

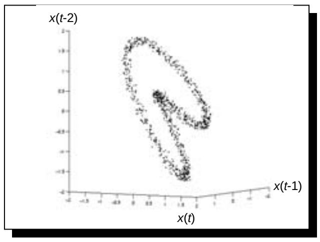
Forecasting: selection of input variables

- Starting with many input variables, then reduce their number
- Two options:
  1. selection of input variables
     - ↗ interpretability
     - ↘ limited to existing variables
  2. projection of input variables
     - linear: PCA
     - non-linear: CCA, Kohonen, etc.
       based on **Taken's theorem**

Forecasting: Taken's theorem 1/2







time series

intrinsic dimension (q) = 1

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

Forecasting: Taken’s theorem 2/2

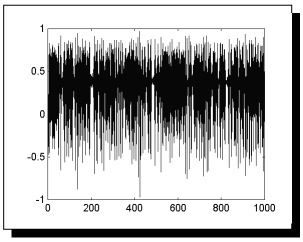
- Takens’ theorem:
  $q \leq$ size of regressor $\leq 2q+1$
- In the $2q+1$ space, there exists a q-surface without intersection points
- Projection from $2q+1$ to q possible !

Forecasting: 1st example 1/2

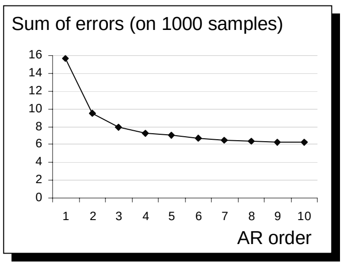
- Artificial series

$$x(t+1) = ax(t)^2 + bx(t-2) + \varepsilon(t)$$

Two past values!

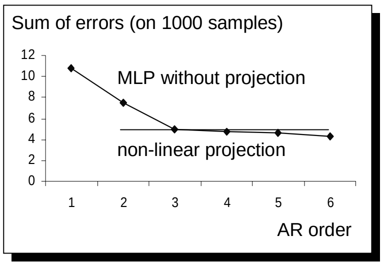


- Linear AR model



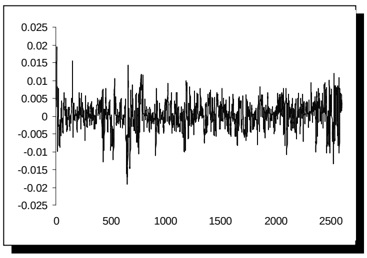
Forecasting: 1st example 2/2

- Non-linear AR model
  - initial regressor: size=6
  - intrinsic dimension: 2
  - CCA from dim=6 to dim=2
  - MLP on 2-dim data



Forecasting: 2nd example 1/2

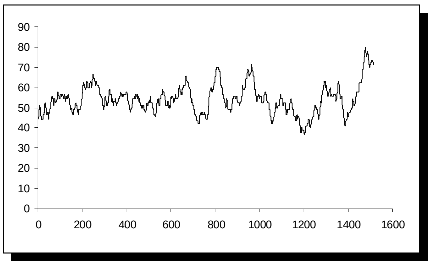
- Daily returns of BEL20 index



- 42 indicators from inputs and exogenous variables:
  - returns: x_t, x_{t-10}, x_{t-20}, x_{t-40}, …, y_t, y_{t-10}, …
  - differences of returns: x_t-x_{t-5}, x_{t-5}-x_{t-10}, …, y_t-y_{t-5}
  - oscillators: *K*(20), *K*(40), …
  - moving averages: *MM*(10), *MM*(50), …
  - exponential moving averages: *MME*(10), *MME*(50), …
  - etc

Forecasting: 2nd example 2/2

- Method:
  - 42 indicators
  - PCA → 25 variables
  - CCA → 9 variables ← Grassberger-Proccacia: intrinsic dimension = 9
  - RBF → forecasting
- Result: % of correct approximations of sign (90-days average)
- In average: 57.2% on test set



Conclusion

- High dimensions: > 3 !
- NN in general: also difficulties in high dimensions
- Common problems:
  - Euclidean distance
  - empty space phenomenon
- Towards solutions…
  1. Local NN (RBFN, etc.): easier learning
  2. Generic methods for non-linear projections: reduce dimension
- Open perspectives:
  - using dimension reduction techniques based on topology (and not on distances)
  - study the possibility of non-Euclidean distances and non-Gaussian kernels