Big Data: The curse of dimensionality and variable selection in identification for a high dimensional nonlinear non-parametric system

Er-wei Bai

University of Iowa, Iowa, USA

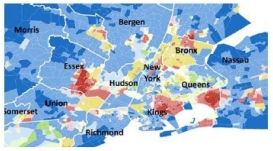
Queen's University, Belfast, UK


CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
Piccolo
demotivateur.fr

Big Data?

Deriving meaning from, and acting upon data sets that are large, high dimensional, heterogeneous, incomplete, contradictory, noisy, fast varying,..., with many different forms and formats.

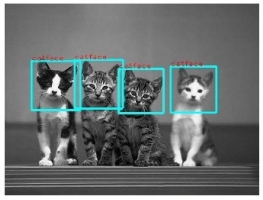
Survey/voting data and blogs for sentiment analysis



NETFLIX

Social media

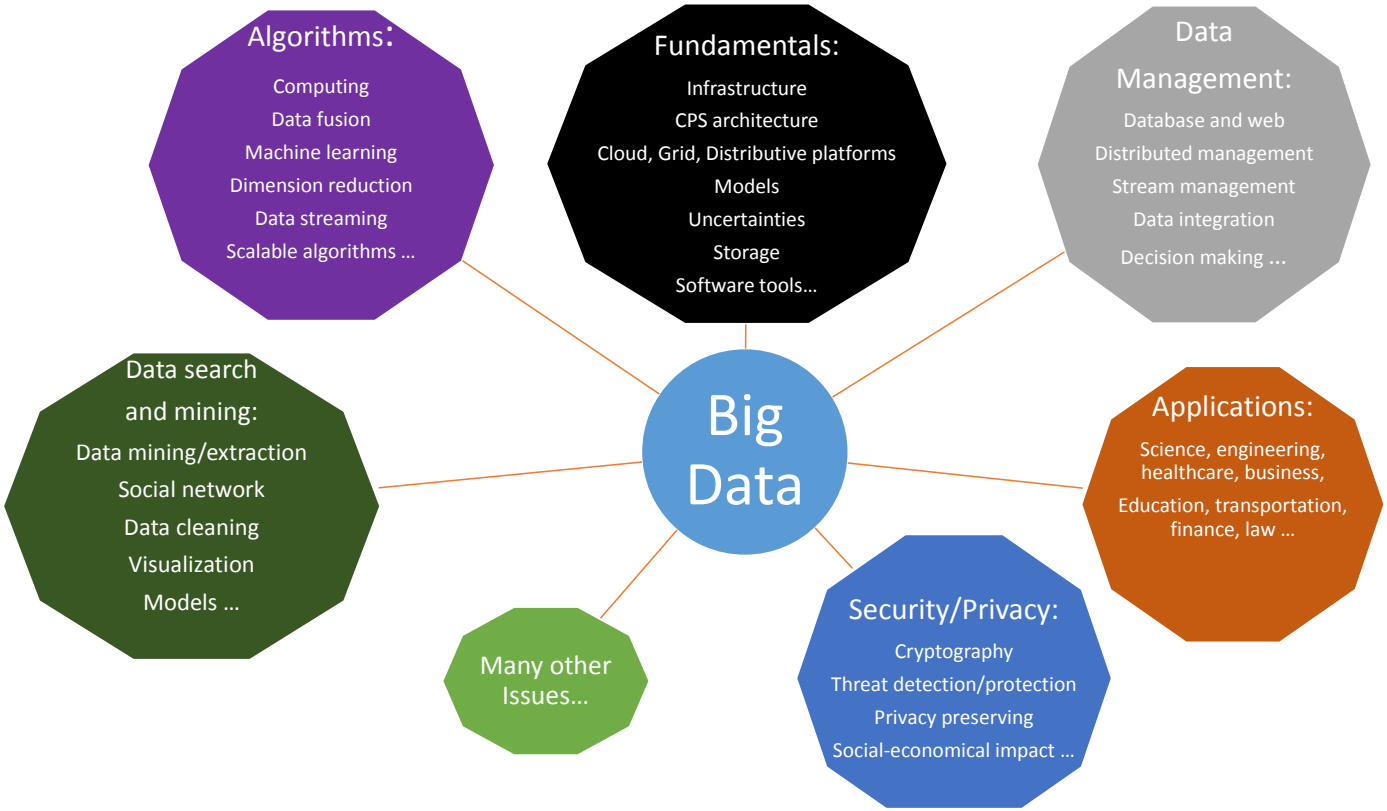
Object detection, classification, tag correlation.



Consumer choice data

Sports/game data

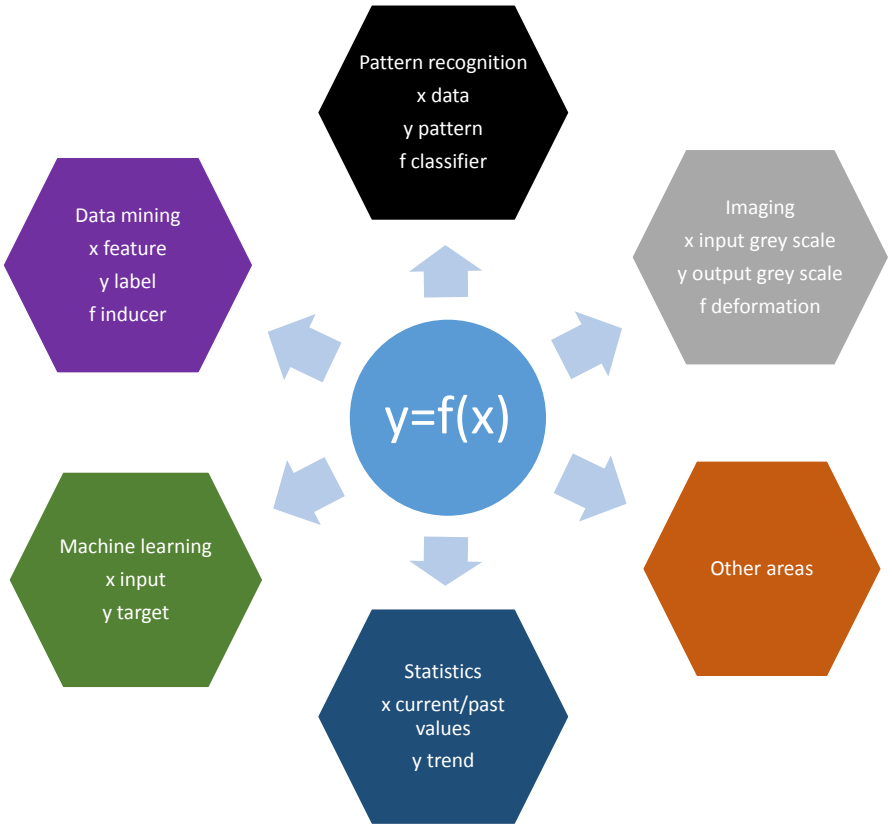
Health data


Big Data
Algorithms:
Computing
Data fusion
Machine learning
Dimension reduction
Data streaming
Scalable algorithms …
Fundamentals:
Infrastructure
CPS architecture
Cloud, Grid, Distributive platforms
Models
Uncertainties
Storage
Software tools…
Data Management:
Database and web
Distributed management
Stream management
Data integration
Decision making …
Data search and mining:
Data mining/extraction
Social network
Data cleaning
Visualization
Models …
Applications:
Science, engineering, healthcare, business,
Education, transportation, finance, law …
Many other Issues…
Security/Privacy:
Cryptography
Threat detection/protection
Privacy preserving
Social-economical impact …

What we are interested: for identification purpose

$$y(\cdot) = f(x(\cdot)) + noise$$

To estimate the unknown nonlinear non-parametric $f(\cdot)$ when the dimension is high.


Pattern recognition
x data
y pattern
f classifier
Imaging
x input grey scale
y output grey scale
f deformation
Other areas
Statistics
x current/past values
y trend
Machine learning
x input
y target
Data mining
x feature
y label
f inducer
y=f(x)

Non-parametric nonlinear system identification is pattern recognition or data mining or ...

Given x^* and want to estimate $y^* = f(x^*)$.

Step 1: Similar patterns or neighborhoods $\|x^* - x(k)\| \leq h$. $\Rightarrow$ $k_1, ..., k_l$ with $y(k_i) = f(x(k_i))$.

Step 2: y^* is a convex combination of $y(k_i)$ with the weights $\frac{K(\frac{x^*-x(k_i)}{h})}{\sum_j K(\frac{x^*-x(k_j)}{h})}$,

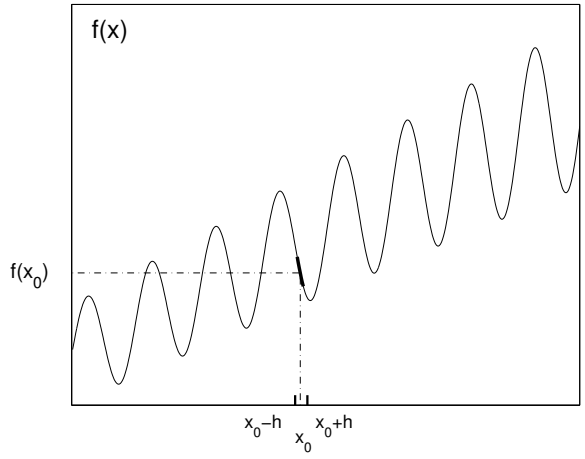
$$y^* = \sum_i \frac{K(\frac{x^*-x(k_i)}{h})}{\sum_j K(\frac{x^*-x(k_j)}{h})} y(k_i)$$

Supervised learning with infinitely many patterns $y(k_i)$.

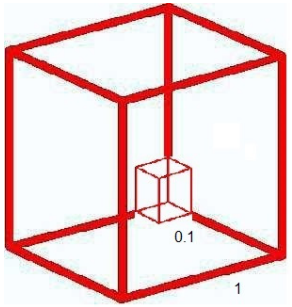
"Parametric" or basis approaches: polynomial (Volterra), splines, neural networks, RKHS, ...

Local or Point by point: DWO, local polynomial, kernel,...

f is estimated point by point. If $f(x_0)$ is of interests, only the data $\{x \ : \ \|x - x_0\| \leq h\}$ is used for some $h > 0$ and data far away from x_0 are not very useful.



Curse of Dimensionality: Empty space. $x \in R^n$, $x_0 = (0, ..., 0)^T$ and $C = \{x \ : \ |x_k| \leq 0.1\}$.



Randomly sample a point x, $Prob\{x \in C\} = 0.1^n$.

On average, the number of points in C is $N \cdot 0.1^n$.

To have one point in C

$N \geq 10^n \Rightarrow$ 1 billion when $n = 9$.

The curse of dimensionality The colorful phrase the "curse of dimensionality" was coined by Richard Bellman in 1961.

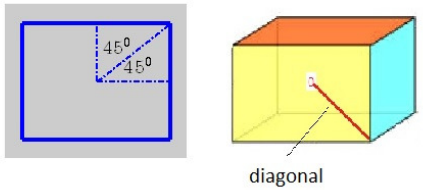
High Dimensional Data Analysis: the curse and blessing of dimensionality, D. Donoho, 2000

The Curse of Dimensionality: A Blessing to Personalized Medicine, *J of Clinical Oncology*, Vol 28, 2010

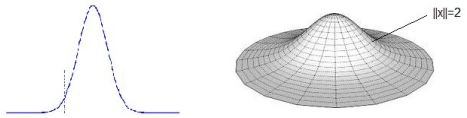
The diagonal is almost orthogonal to all coordinate axes.

A unit cube $[-1,1]^n$ centered at the origin. The angel between its diagonal $v = (\pm 1, ..., \pm 1)^T$. and any axis e_i is

$$cos(\theta) =< \frac{v}{\|v\|}, e_i >= \frac{\pm 1}{\sqrt{n}} \to 0$$



High dimensional Gaussian. Let $p(x) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{\|x\|^2}{2}}$.



Calculate $Prob\{\|x\| \geq 2\}$

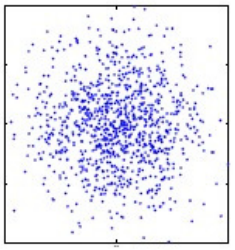
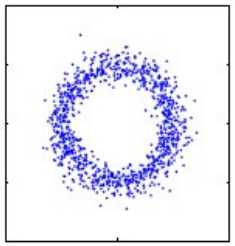
n	1	2	5	10	20	100
Prob	0.04550	0.13534	0.54942	0.94734	0.99995	1.0000

For a high dimensional Gaussian, the entire samples are almost in the tails.

Concentration. $x = (x_1, ..., x_n)^T$ and x_k's iid Gaussian of zero mean. Then,

$$Prob\{\|x\|^2 \geq (1+\epsilon)\mu_{\|x\|^2}\} \leq e^{-\epsilon^2 n/6},$$

$$Prob\{\|x\|^2 \leq (1-\epsilon)\mu_{\|x\|^2}\} \leq e^{-\epsilon^2 n/4},$$

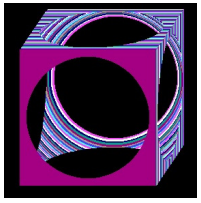
$\mathbb{R}^2$:  $\mathbb{R}^n$: 

High dimensional iid vectors are distributed close to the surface of a sphere of radius $\mu_{\|x\|^2}$.

Volumes of cubes and balls.

$$V_{ball} = \frac{\pi^{n/2} r^n}{\Gamma(n/2+1)}, \; V_{cube} = (2r)^n$$

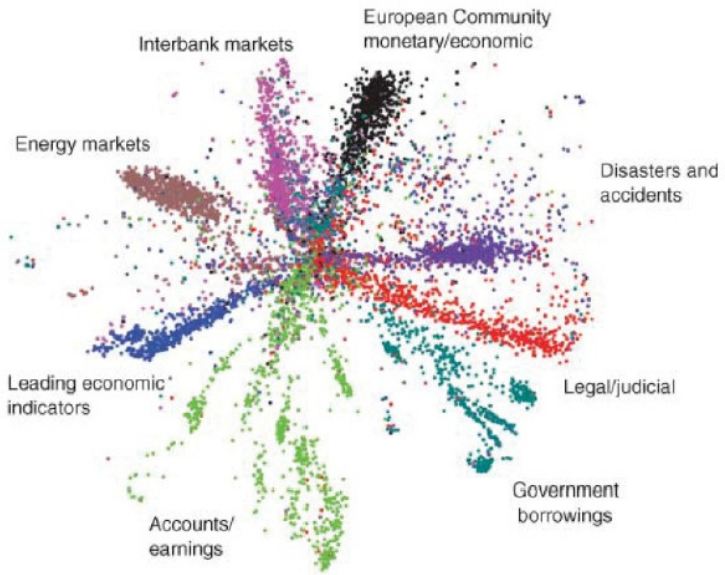
$$\lim_{n \to \infty} V_{ball}/V_{cube} = 0$$



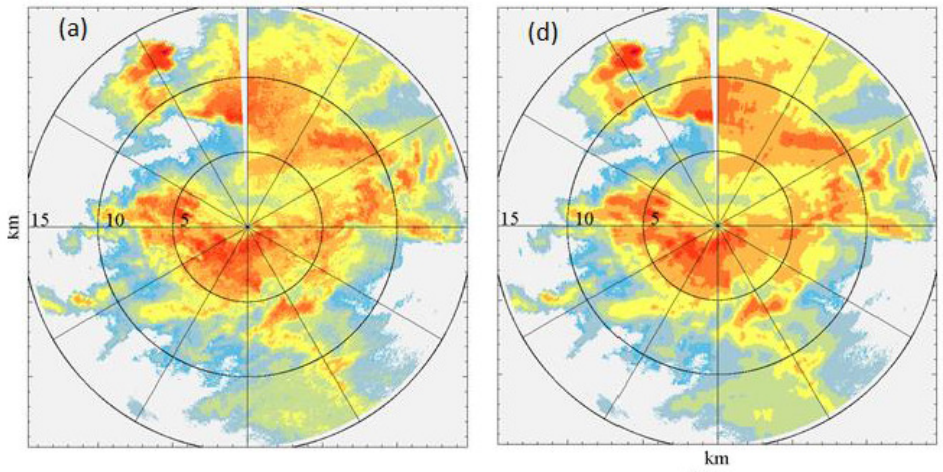
The volume of a cube concentrates more on its corners as n increases.

Extreme dimension reduction is possible, Science 2006

804,414 Reuters newswire stories, each article is represented as a vector containing the counts of the most frequently used 2000 words and further dimensionality reduction to R^2 (Boltzmann neural network). Text to a 2-dimensional code

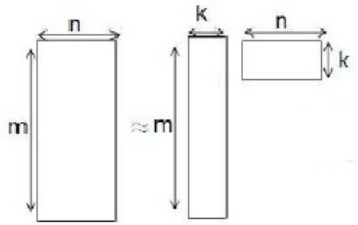


Taken from Hinton, 2006.



Observed on Iowa XPOL-2 radar.

Approximation by taking 5% of significant singular values.



m=1930, n=413, k=20

What we do in high dimensional nonlinear nonparametric system identification:

Underlying models: Additive model, Block oriented nonlinear systems...

Dimension asymptotic: The limit distributions...

Variable selection and dimension reduction

Global (nonlinear nonparametric) models

...

Dimension reduction and variable selection: a well studied topic in the linear setting, e.g.,

Principal component analysis, Low rank matrix approximation, Factor analysis, Independent component analysis, Fisher discriminator, Forward/backward stepwise(stage),...

Penalized (regularized) optimization and its variants

$$\min \|Y - X\widehat{\beta}\|^2 + \lambda \sum |\widehat{\beta}_i|^\alpha, \ 0 \leq \alpha < \infty$$

$$\text{or} \quad \min \|Y - X\widehat{\beta}\|^2, \quad s.t, \quad \sum |\widehat{\beta}_i|^\alpha \leq t$$

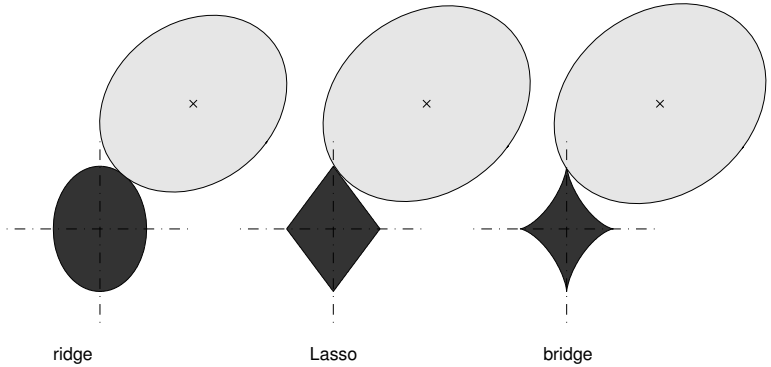
Nonlinear(support vector regression): (Model Fits) + λ (Model Complexity)

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B., Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso, Journal of Machine Learning Research, irrepresentable condition.

Geometric interpretation

$$arg\min_{\hat{\beta}} \|Y - X\hat{\beta}\|^2 = arg\min_{\hat{\beta}} (\hat{\beta} - \beta_0)^T X^T X (\hat{\beta} - \beta_0)$$

$$s.t, \quad \sum |\hat{\beta}_i|^\alpha \leq t$$



$\alpha = 1$(Lasso), $\alpha = 2$(ridge), $\alpha < 1$(bridge)

Compressive sensing

$$Y = X\beta, \quad \beta \text{ is sparse.}$$

$$\min_{\widehat{\beta}} \|Y - X\widehat{\beta}\|_1 \Rightarrow \widehat{\beta} = \beta$$

provided that X satisfies the restricted isometry property.

E. Candes and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?", IEEE Trans. Inform. Theory, Dec. 2006.

In a (FIR) system identification setting

$$Y = \begin{pmatrix} u_1 & \cdots & u_{-n+2} \\ \vdots & \ddots & \vdots \\ u_N & \cdots & u_{-n+N+1} \end{pmatrix} \beta + e + w$$

e outliers and w random noise.

$$\widehat{\beta} = arg \min_{\xi} \|Y - X\xi\|_1$$

Under some weak assumptions, iid on w and e has $k \leq \beta N$ non-zero elements for some $\beta < 1$, then in probability as $N \to \infty$,

$$\|\widehat{\beta} - \beta\|_2 \to 0$$

W. Xu, EW Bai and M Cho (2014) Automatica, "System Identification in the Presence of Outliers and Random Noises: a Compressed Sensing Approach".

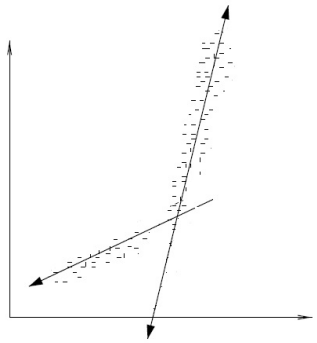
Big difference between linear and nonlinear settings

Local vs global dimension.

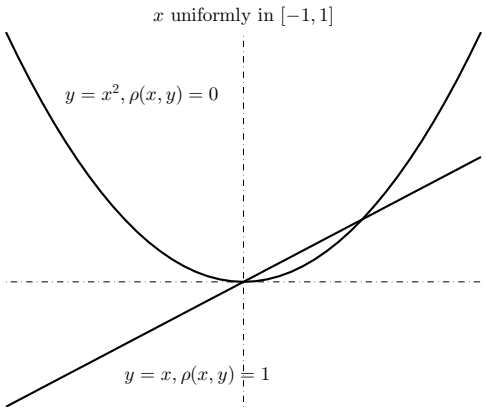
$$y(k) = f(u(k-1), u(k-2), u(k-3), u(k-4))$$

$$= \begin{cases} u(k-4) & u(k-1) \geq 0 \\ u(k-4)u(k-2) & u(k-4) < 0, u(k-2) \geq 0 \\ u(k-4)u(k-3) & u(k-4) < 0, u(k-2) < 0 \\ u(k-1) & otherwise \end{cases}$$

Linear (global) PCA may not work.



Methods work in a linear setting may not work in a nonlinear setting, e.g., correlation methods



Linear: $y(k) = x(k) \cdot 1 \Rightarrow \frac{cov(y,x)}{\sqrt{cov(y)}\sqrt{cov(x)}} = 1.$

Nonlinear: $y(k) = x(k)^2 \cdot 1$

$$cov(y, x) = E(yx) - E(y)E(x) = \int_{-1}^{1} a^3 \cdot \frac{1}{2} da = 0$$

y depends on x nonlinearly and the correlation is zero.

For a high dimensional nonlinear problem, approximation is a key. The ambient dimension is very high, yet its desired property is embedded in a low dimensional structure. The goal is to design efficient algorithms that reveal dominate variables for which one can have some theoretical guarantees.

Manifold Embedding (Nonlinear PCA): Eliminate redundant/dependent variables.

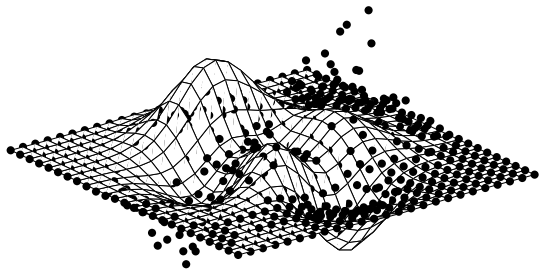
- If $x = (x_1, x_2)$ and $x_1 = g(x_2)$,

$$\Longrightarrow y = f(x) = f(g(x_2), x_2) = h(x_2).$$

- If $x = (x_1, x_2) = (g_1(z), g_2(z))$

$$\Longrightarrow y = f(x) = f(g_1(z), g_2(z)) = h(z).$$

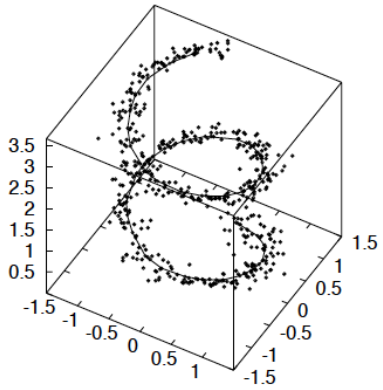
More realistically, $y \approx h(x_2)$ $(h(z))$.



One dimensional: principal curve: Let $x = (x_1, ..., x_n)^T$ and $f(t) = (f_1(t), ..., f_n(t))$ be a curve in R^n. Define $s_f(x)$ to be the value of t corresponding to the point of $f(t)$ that is closest to x. The principal curve is

$$f(t) = \mathbf{E}(x \mid s_f(x) = t)$$

$f(t)$ is a curve that passes through the middle of the distribution of x.



Principle surface is not easy and computationally prohibitive!

Local (linear/nonlinear) embedding: Find a set of lower dimensional data that resembles the local structure of the original high dimensional data. The key is "distance preservation" or "topology preservation".

Multidimensional scaling, Science, 2000

Given high dimensional $\vec{x}_i$'s, define

$$d_{ij} = \|\vec{x}_i - \vec{x}_j\|$$

Find lower dimensional $\vec{z}_i$'s that minimizes the pairwise distance

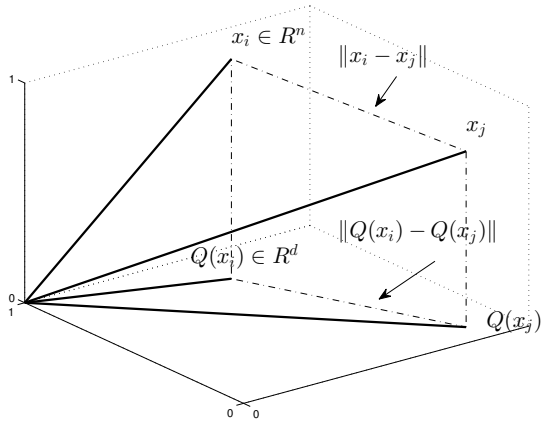
$$\min_{\vec{z}_i} \sum_{i<j} (\|\vec{z}_i - \vec{z}_j\| - d_{ij})^2$$

In a linear setting, $\|\cdot\|$ is the Euclidean norm and in a nonlinear setting, $\|\cdot\|$ is usually the distance along some manifold (Isomap).

Euclidean norm embedding: (Johnson-Lindenstrauss).

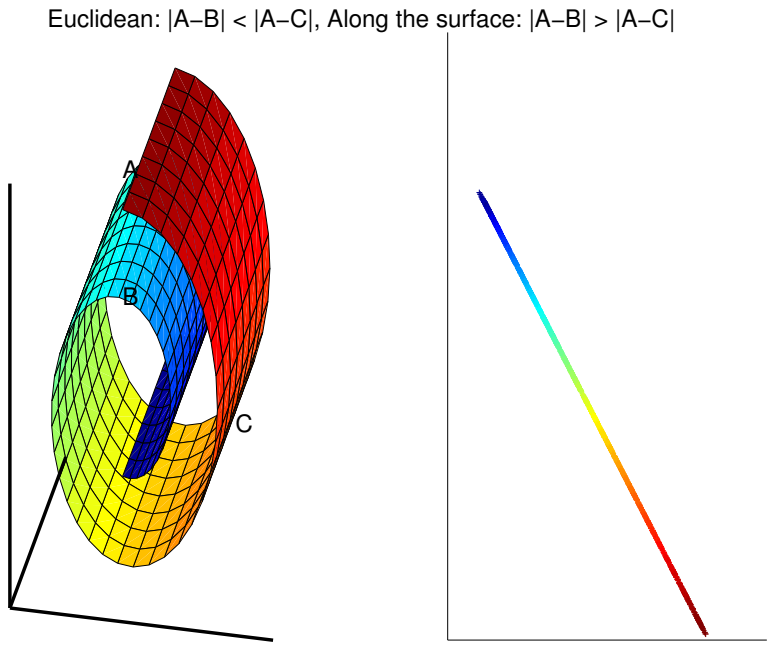
Theorem: Let $x_1, ..., x_n \in R^n$, $\epsilon \in (0, 1/2)$ and $d \geq O(\epsilon^{-2} \log n)$. There exists a matrix $Q : R^n \to R^d$ and with a high probability,

$$(1 - \epsilon)\|x_i - x_j\|^2 \leq \|Q(x_i) - Q(x_j)\|^2 \leq (1 + \epsilon)\|x_i - x_j\|^2$$



The space is "approximately" d dimensional not n dimensional if pairwise distance is an interest. What if on a manifold?

Isomap distance:



Local Linear Embedding

Find the local structure of the data

$$\min_{w_{ij}} \sum_i \|\vec{x_i} - \sum_j w_{ij}\vec{x}_j\|^2$$

Find a lower dimensional data z that resembles the local structure of x, under some regularization

$$\min_{z_i} \sum_i \|\vec{z_i} - \sum_j w_{ij}\vec{z}_i\|^2$$

Estimation is carried out in a lower dimensional space

$$y = f(z)$$

Example: Science, 2000

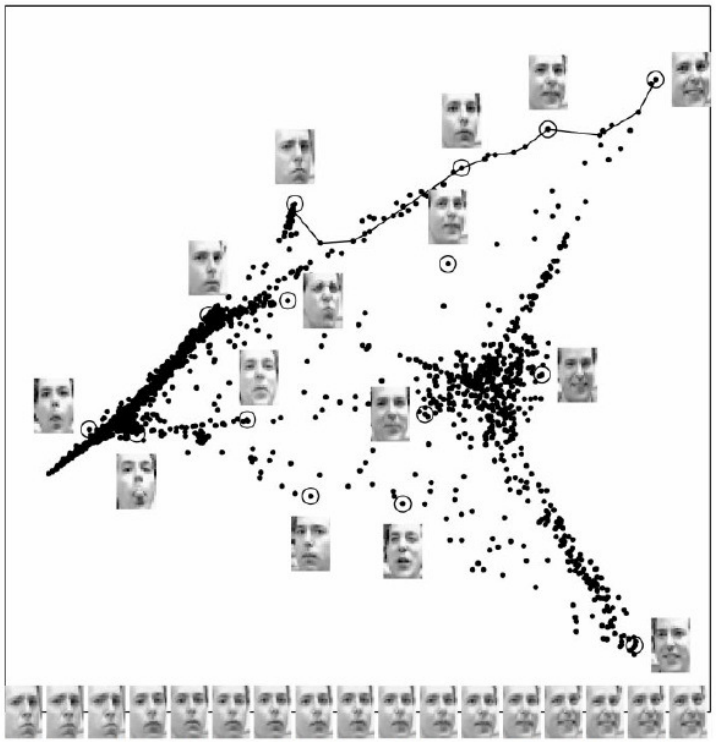
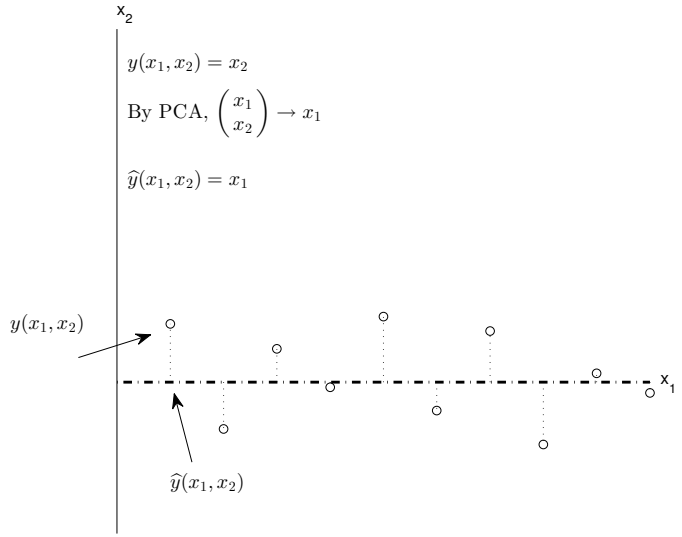


Fig. 3. Images of faces (*11*) mapped into the embedding space described by the first two coordinates of LLE. Representative faces are shown next to circled points in different parts of the space. The bottom images correspond to points along the top-right path (linked by solid line), illustrating one particular mode of variability in pose and expression.

Unsupervised: dimension reduction is based on x alone without considering output y:

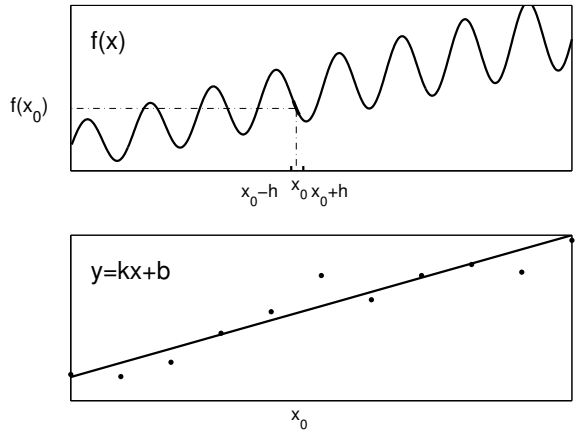


Add a penalty term,

$$\min_{z_i} \lambda \sum_i \|\vec{z}_i - \sum_j w_{ij}\vec{z}_i\|^2 + (1-\lambda)(\text{output error term})$$

An active research area. Henrik Ohlsson, Jacob Roll and Lennart Ljung, "Manifold-Constrained Regressors in System Identification", 2008

Global Model The problem of the curse of dimensionality is the lack of a global model. As a consequence, only data in (x_0-h, x_0+h) can be used and the majority of data is discarded. As in a linear case, every data is used $\min \| \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} - \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix} \begin{pmatrix} k \\ b \end{pmatrix} \|^2$.



Bayesian in particular Gaussian process model is one of the global models. Rasmussen CE and C. Williams (2006) Gaussian Pro- cess for Machine Learning, The MIT Press, Cambridge, MA

Gaussian process regression model

Consider a scalar nonlinear system

$$y(k) = f(x(k)) + v(k)$$

where $v(\cdot)$ is an iid radome sequence of zero mean and finite variance σ_v^2.

In a Gaussian Process setting, $f(x_0), f(x(1)), ..., f(x(N))$ are assumed to follow a joint Gaussian distribution with zero mean (not necessary though) and a covariance matrix Σ_p.

Example:

Let $y_0 = f(x_0), Y = (y(1), y(2), ..., y(N))'$. Since $(y_0, Y)'$ follows a joint Gaussian

$$\begin{pmatrix} y_0 \\ Y \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} c & B \\ B' & A \end{pmatrix})$$

Given x_0, what is y_0?. The conditional density of y_0 conditioned on Y is also Gaussian

$$y_0 \sim \mathcal{N}(BA^{-1}Y, c - BA^{-1}B')$$

The minimum variance estimate $\widehat{y}_0 = \widehat{f}(x_0)$ of $f(x_0)$ is the conditional mean given by

$$\widehat{f}(x_0) = BA^{-1}Y$$

Huntington Disease Example: Neurological cognitive test

Totally 60 patients. A cognitive test data for 21 patients were missing.

	project	external_id	session	Neurological	Cognitive_0(	Cognitive_0(	Cognitive_0(	Cognitive_0(	Cognitive_0(	Cognititive_(	MotorEval_0(
2	PHD_048	PHD-AS1-52	11197	2	52	80	58	98	24	48	5
3	PHD_048	PHD-AS1-50	34395	3	???	90	55	100	24	72	11
4	PHD_048	PHD-AS1-51	27205	0	66	117	60	130	15	57	0
5	PHD_048	PHD-AS1-50	46230	3	49	64	30	80	39	74	17
6	PHD_048	PHD-AS1-50	67069	2	32	73	35	86	32	89	8
7	PHD_048	PHD-AS1-52	52987	2	37	45	26	62	36	83	13
8	PHD_048	PHD-AS1-50	52629	1	???	76	49	100	25	92	5
9	PHD_048	PHD-AS1-51	52712	2	45	83	42	110	27	42	9
10	PHD_048	PHD-AS1-51	80975	2	47	74	32	91	25	64	5
11	PHD_048	PHD-AS1-50	74292	4	50	78	40	110	37	63	24
12	PHD_048	PHD-AS1-52	71011	1	???	31	45	104	23	82	1
13	PHD_054	PHD-AS1-51	60635	0	NA	NA	NA	NA	NA	NA	1
14	PHD_054	PHD-AS1-50	88270	1	NA	NA	NA	NA	NA	NA	9
15	PHD_054	PHD-AS1-50	20715	0	78	102	70	118	16	24	4
16	PHD_054	PHD-AS1-52	17288	0	50	76	46	80	30	49	0
17	PHD_054	PHD-AS1-50	85937	1	43	53	21	79	23	69	7
18	PHD_144	PHD-AS1-51	15979	0	53	90	68	94	24	73	1
19	PHD_054	PHD-AS1-52	86324	1	51	96	56	113	23	69	5
20	PHD_054	PHD-AS1-51	91752	1	54	84	58	111	37	51	5
21	PHD_054	PHD-AS1-50	76328	1	???	78	50	85	14	36	5
22	PHD_144	PHD-AS1-50	82635	1	33	73	41	85	30	97	3
23	PHD_144	PHD-AS1-50	42463		36	64	31	66	35	128	
24	PHD_144	PHD-AS1-51	49568	3	48	66	43	94	19	58	7
25	PHD_144	PHD-AS1-52	84953	0	50	77	31	83	22	75	0
26	PHD_144	PHD-AS1-52	41642	0	52	76	50	106	25	NA	0
27	PHD_144	PHD-AS1-51	24795	4	26	53	10	61	56	300	51
28	PHD_144	PHD-AS1-51	54608	2	???	84	58	91	32	46	3
29	PHD_144	PHD-AS1-52	53108	0	50	83	37	107	31	54	3
30	PHD_144	PHD-AS1-51	32375	NA	NA	NA	NA	NA	NA		
31	PHD_144	PHD-AS1-52	94863	0	56	103	72	94	19	46	0
32	PHD_048	PHD-AS1-52	16905	2	NA	NA	NA	NA	NA	NA	9

Modified Gaussian Model: The first row are predicted and the second "true but missing" data.

$$\begin{pmatrix} 43.9 & 43.9 & 44.6 & 44.6 & 45.2 & 45.2 & 45.5 & 45.6 & 46.0 & 46.3 & 46.6 \\ 45 & 45 & 45 & 45 & 46 & 46 & 46 & 46 & 46 & 46 & 46 \end{pmatrix}$$

$$\begin{pmatrix} 46.6 & 46.6 & 46.8 & 46.8 & 47.2 & 47.6 & 48.0 & 48.0 & 48.0 & 48.2 \\ 47 & 48 & 48 & 48 & 48 & 49 & 49 & 49 & 49 & 49 \end{pmatrix}$$

Bai et al, 2014

Some discussions:

- Hard vs soft approaches.
- Top down vs bottom up approaches.
- Simplified but reasonable models.
- ...

Top down approach:

$$y(k) = f(x_1(k), ..., x_n(k)) + noise$$

$x_i(\cdot)$ is irrelevant $\Rightarrow \frac{\partial f}{\partial x_i} = 0$. Local linear estimator

$$f(x(k)) = f(x^0) + (x(k) - x^0)^T \frac{\partial f}{\partial x}|_{x^0} + h.o.t$$

$$\min_{\widehat{f(x^0)}, \hat{\beta}} \sum_{k=1}^{N} \{y(k) - [1, (x(k) - x^0)^T] \begin{pmatrix} \widehat{f(x^0)} \\ \hat{\beta} \end{pmatrix}\}^2 \cdot K_Q(x(k) - x^0)$$

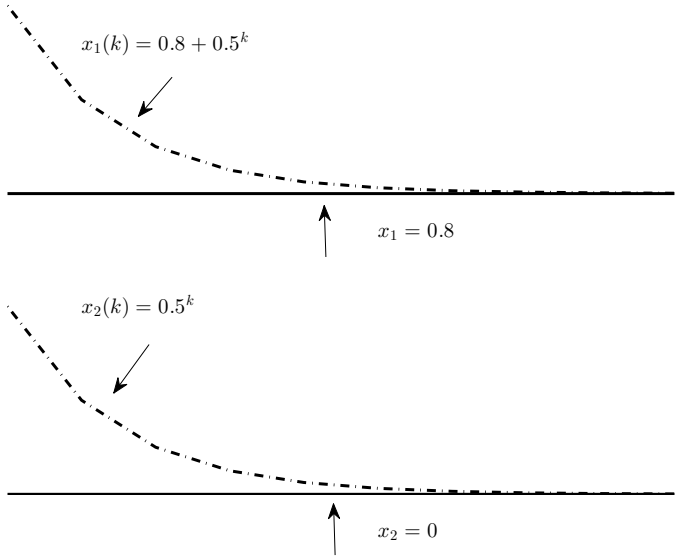
where

$$\beta^* = \begin{pmatrix} \beta_1^* \\ \vdots \\ \beta_d^* \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad |\beta_1^*| > 0, ..., |\beta_d^*| > 0, \beta_{d+1}^* = ... = \beta_n^* = 0$$

$$A^* = \{j \ : \ |\beta_j^*| > 0\} = \{1, 2, 3, ..., d\}$$

Goal is both parameter and set convergence.

Parameter and set convergence: Example:



$\widehat{x}_{jk} \to x_j, \quad j = 1, 2$

$A = \{j \ : \ |x_j| > 0\} = \{1\}.$

$\widehat{A}_k = \{j \ : \ |x_{jk}| > 0\} = \{1, 2\} \not\to A$

Top down approach: The approach is an adaptive convex penalized optimization,

$$J(\beta) = \min_{\beta} \|Z - \Phi\beta\|^2 + \lambda \sum_{j=1}^{n} w_j |\beta_j|,$$

$w_j = \frac{1}{|\hat{\beta}_j|}$ calculated from the local linear estimator. Then under some technical conditions, in probability as $N \to \infty$,

$$\widehat{\frac{\partial f}{\partial x_i}}|_{x^0} \to \frac{\partial f}{\partial x_i}|_{x^0} \quad \text{if} \quad |\frac{\partial f}{\partial x_i}|_{x^0} > 0 \quad \text{parameter convergence}$$

$$Prob\{\widehat{\frac{\partial f}{\partial x_i}}|_{x^0} = 0\} = 1, \quad \text{if} \quad |\frac{\partial f}{\partial x_i}|_{x^0} = 0 \quad \text{set convergence}$$

The approach works if the the available data is long.

E Bai et al, "Kernel based approaches to local nonlinear non-parametric variable selection", Automatica, 2014, K Bertin, "Selection of variables and dimension reduction in high-dimensional non-parametric regression", EJS, 2008

Bottom up approach:Forward/Backward

First proposed by Billings, 1989, "Identification of MIMO non-linear systems using a forward-regression orthogonal estimator". Extremely popular in statistics and in practice. For a long time was considered to be Ad Hoc but recent research showed otherwise,

T Zhang, "Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations", IEEE TIT,2011, "On the Consistency of Feature Selection using Greedy Least Squares Regression", J of Machine learning Res. 2009, Tropp, "Greed is Good: Algorithmic Results for Sparse Approximation", IEEE TIT, 2004...

Forward selection What if the dimension is high and the available data set is limited.

To illustrate, consider a linear case

$$Y = (\, x_1, x_2, ..., x_n \,) \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$\text{First step:} \quad i_1^* = arg \min_{i \in [1,...,n]} \|Y - a_i x_i\|^2$$

$$\text{kth step:} \quad i_k^* = arg \min_{i \in [1,...,n]/[i_1^*,...,i_{k-1}^*]} \|Y - (x_{i_1^*}, ..., x_{i_{k-1}^*}) \begin{pmatrix} a_1^* \\ \vdots \\ a_{k-1}^* \end{pmatrix} - a_i x_i\|^2$$

Bottom up approach: Backward selection

Start with the chosen $(i_1, ..., i_p)$ and $a_1, ..., a_p$.

$$i_1^* = arg \min_{i \in n/[i_2,...,i_p]} \|Y - (x_{i_2}, ..., x_{i_p}) \begin{pmatrix} a_2 \\ \vdots \\ a_p \end{pmatrix} - a_i x_i\|^2,$$

Repeated for every i_j.

Forward/Backward for nonlinear system:

The minimum set of unfalsified variables

$$|f(x_1, ..., x_n) - g(x_{i1}, ..., x_{ip})| \approx 0$$

Low dimensional neighborhood

One dimensional neighborhood of $x_i(k)$,

$$\{x(j) \in R^n \mid (x(k) - x(j))_i = \sqrt{(x_i(k) - x_i(k))^2} \leq h\}$$

p-dimensional neighborhood of $(x_{i1}(k), ..., x_{ip}(k))$,

$$\{x(j) \in R^n \mid (x(k) - x(j))_{i1,...,ip}$$

$$= \sqrt{(x_{i1}(k) - x_{i1}(k))^2 + ... + (x_{ip}(k) - x_{ip}(j))^2} \leq h\}$$

Algorithm:

Step 1: Determine the bandwidth.

Step 2: The number of variables are determined by the modified Box-Pierce test.

Step 3: Forward selection.

Step 4: Backward selection.

Step 5: Terminate.

Example: the actual system lies approximately on a unknown dimensional manifold:

$$x_2(k) \approx g(x_1(k)) \Rightarrow y(k) \approx f(x_1(k), g(x_1(k))) = f_1(x_1(k))$$

which is lower dimensional.

$$y(k) = 10sin(x_1(k)x_2(k))+20(x_3(k)-0.5)^2+10x_4(k)+5x_5(k)$$

$$+x_6(k)x_7(k)+x_7(k)^2+5cos(x_6(k)x_8(k))+exp(-|x_8(k)|)$$

$$+0.5\eta(k), k = 1, 2, ..., 500$$

$\eta(k)$ is i.i.d. Gaussian noise of zero mean and unit variance.

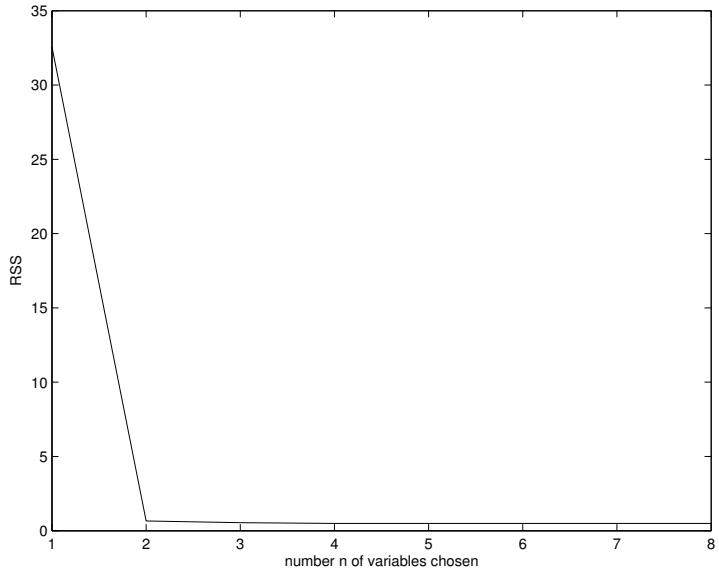
$x_3(k), x_5(k)$ are independent and uniformly in $[-1, 1]$ and

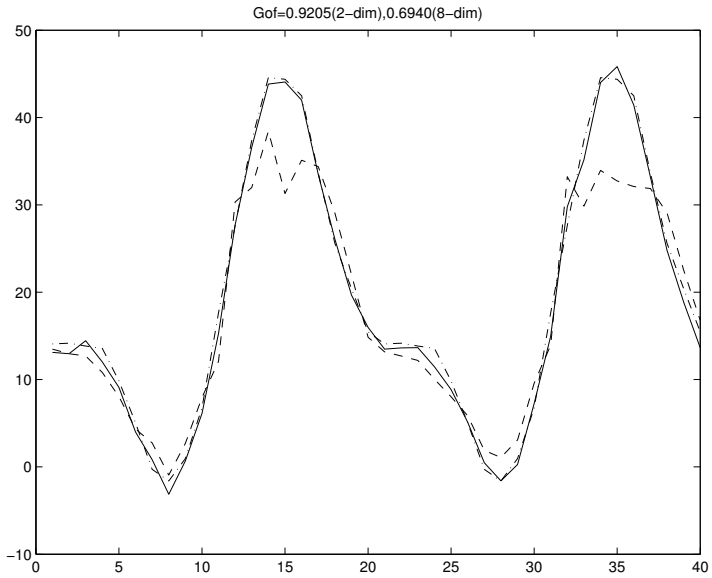
$$\begin{aligned} x_4(k) &= x_3(k) \cdot x_5(k) + 0.1 \cdot \eta(k) \\ x_1(k) &= x_3(k)^2 \cdot x_5(k) + 0.1 \cdot \eta(k) \\ x_2(k) &= x_3(k) \cdot x_5(k)^2 + 0.1 \cdot \eta(k) \\ x_6(k) &= x_1(k) - x_4(k) + 0.1 \cdot \eta(k) \\ x_7(k) &= x_3(k)^3 \cdot x_5(k) + 0.1 \cdot \eta(k) \\ x_8(k) &= x_2(k) \cdot x_5(k) + 0.1 \cdot \eta(k) \end{aligned}$$

$x_1(\cdot), ..., x_8(\cdot)$ are not exactly but approximately on an two dimensional manifold. $h = 0.2$ was chosen by the 5-fold cross validation. Test signal

$$x_3(k) = 0.9 * sin(\frac{2\pi k}{20}),\ x_5(k) = 0.9 * cos(\frac{2\pi k}{20}),\ k = 1, ..., 40$$

Results:





Bai et al, 2013, "On Variable Selection of a Nonlinear Non-parametric System with a Limited Data Set"

Nonlinear systems with short term memory and low degree of interactions: Additive systems

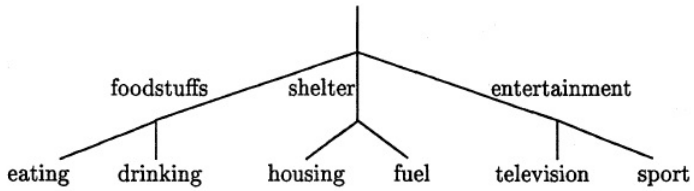
$$y(k) = f(x_1(k), ..., x_n(k)) + noise$$

$$= \sum_{i=1}^{n} f_i(x_i(k)) + \sum_{j>i} f_{ij}(x_i(k), x_j(k)) + noise$$

An n-dimensional system becomes a number of 1 and 2-dimensional systems. Under some technical conditions, e.g., iid or Galois on the inputs, f_i and f_{ij} can be identified independently...

Bai et al TAC(2008), Automatica(2009), Automatica(2010), TAC(2010)...

Why? Consider a household expense tree:



The most widely used nonlinear model in practice (and also in the statistics literature.)

A simple example:

$$y(k) = \bar{c} + \bar{f}_1(x_1(k)) + \bar{f}_2(x_2(k)) + \bar{f}_{12}(x_1(k), x_2(k)) + v(k)$$

$$= c + f_1(x_1(k)) + f_2(x_2(k)) + f_{12}(x_1(k), x_2(k)) + v(k)$$

Each term can be identified separately and low dimension.

$$\begin{array}{lcl} c & = & Ey(k) \\ f_1(x_1^0) & = & E(y(k) \mid x_1(k) = x_1^0) - c \\ f_2(x_2^0) & = & E(y(k) \mid x_2(k) = x_2^0) - c \\ f_{12}(x_1^0, x_2^0) & = & E(y(k) \mid x_1(k) = x_1^0, x_2(k) = x_2^0) - c \end{array}$$

Quick boring derivation:

$$
\begin{array}{lcl}
g_{1,12}(x_1^0) & = & E(\bar{f}_{12}(x_1(k), x_2(k)) \mid x_1(k) = x_1^0) \\
g_{2,12}(x_2^0) & = & E(\bar{f}_{12}(x_1(k), x_2(k)) \mid x_2(k) = x_2^0) \\
c_{12} & = & E(\bar{f}_{12}(x_1(k), x_2(k))) \\
c_1 & = & E(\bar{f}_1(x_1(k)) + g_{1,12}(x_1(k))) \\
c_2 & = & E(\bar{f}_2(x_2(k)) + g_{2,12}(x_2(k))) \\
f_{12}(x_1(k), x_2(k)) & = & \bar{f}_{12}(x_1(k), x_2(k)) - g_{2,12}(x_2(k)) - g_{1,12}(x_1(k)) + c_{12} \\
f_1(x_1(k)) & = & \bar{f}_1(x_1(k)) + g_{1,12}(x_1(k)) - c_1 \\
f_2(x_2(k)) & = & \bar{f}_2(x_2(k)) + g_{2,12}(x_2(k)) - c_2 \\
c & = & \bar{c} - c_{12} = c_1 + c_2
\end{array}
$$

$\Rightarrow$

$$
\begin{array}{l}
Ef_1(x_1(k)) = Ef_2(x_2(k)) = Ef_{12}(x_1(k), x_2(k)) = 0 \\
Ef_1(x_1(k)) \cdot f_2(x_2(k)) = Ef_i(x_i(k)) \cdot f_{12}(x_1(k), x_2(k)) = 0, \;\; i = 1, 2,
\end{array}
$$

Convergence and sparsity

Consider

$$J = \|Y - (\sum_i f_i + \sum_{j<k} f_{jk})\|^2 + \lambda_1 \sum_i (\|f_i\|^2 + \sum_{j \neq k} \|f_{jk}\|^2)^{1/2} + \lambda_2 \sum_{j<k} \|f_{jk}\|$$

Then, under some technical conditions, the true f_i, f_{jk}'s can be identified and sparse.

Rachenko and James, J of ASA, 2012

Thanks!

"I liked it better before big data when we just had good old regular data."