

Hypes and Other Important Developments in Statistics

Aad van der Vaart
Vrije Universiteit Amsterdam

May 2009

The Hype

Sparsity

For decades we taught students that to estimate p parameters one needs $n \gg p$ observations.

This was already challenged by nonparametric smoothing techniques.

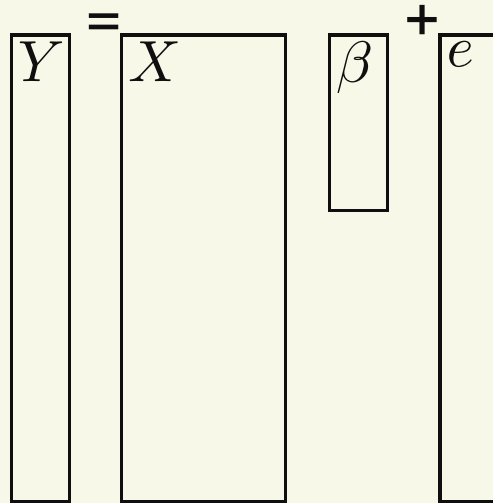
Nowadays we encounter many situations with little data and many unknowns. Therefore, yes we can handle more parameters than observations.

All we need to do is assume sparsity: most parameters are zero. The remaining problem is that we do not know which ones.

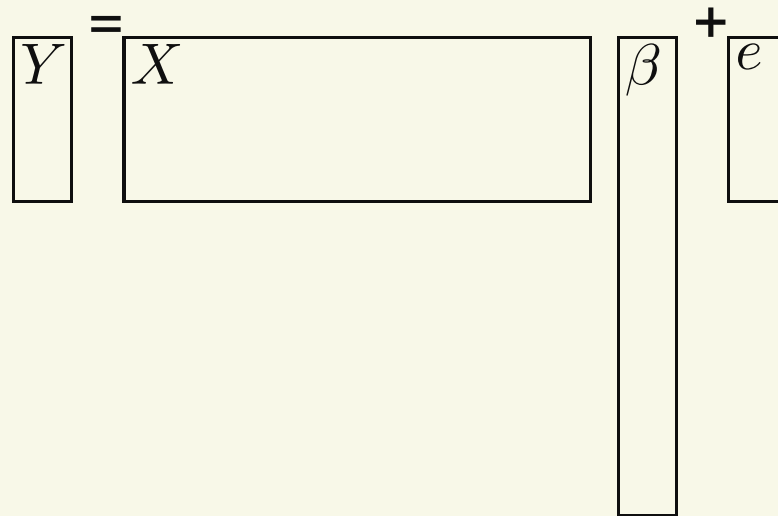
$$p > n$$

Regression $Y = X\beta + e$.

Classical:



Hypic:



A theoretical result

THEOREM

If Y_1, \dots, Y_n are independent, with $Y_i \sim N(\theta_i, 1)$, for unknown $\theta = (\theta_1, \dots, \theta_n)$, then as $p_n/n \rightarrow 0$,

$$\inf_T \sup_{\theta \in \mathbb{R}^n: \#(\theta_i \neq 0) \leq p_n} \mathbb{E}_\theta \|T - \theta\|^2 = 2p_n \log \frac{n}{p_n} (1 + o(1)).$$

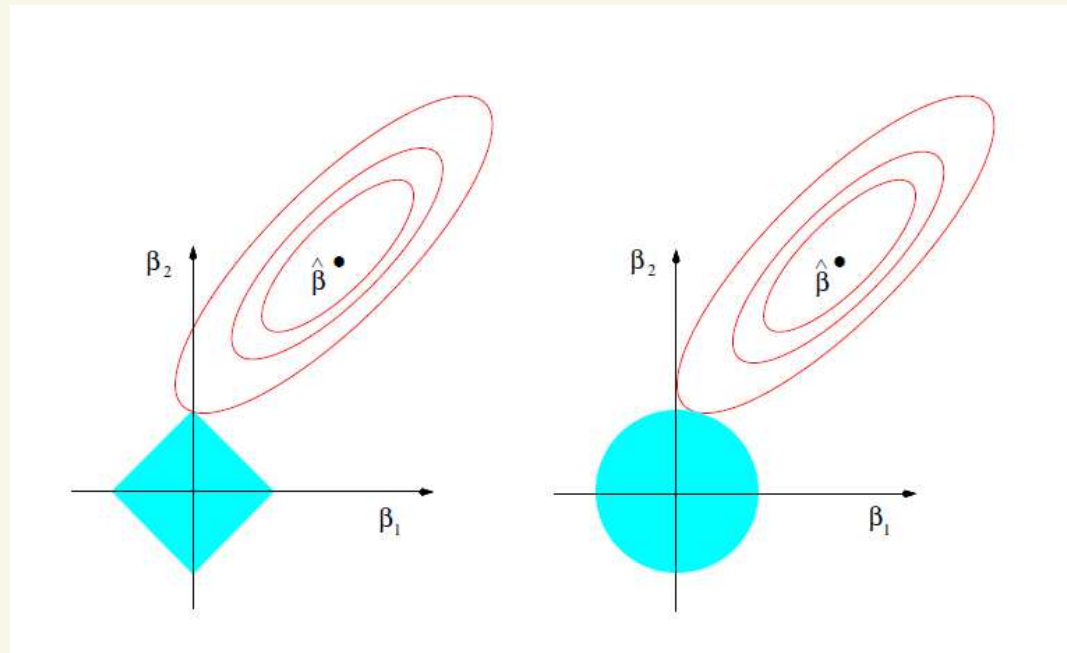
Had we known which $p_n \ll n$ parameters are nonzero, then the left side would have been p_n . We pay a $\log(n/p_n)$ factor only.

There are many procedures that have a risk of minimal order (penalized minimum contrast, truncation with (FDR) substitution, empirical Bayes, full Bayes).

LASSO

The penalized Least squares estimate with ℓ_1 -penalty

$$T_\lambda = \operatorname{argmin}_\theta \|Y - \theta\|^2 + \lambda \|\theta\|_1.$$



The LASSO estimator is very computable and it is sparse, although not sparse enough. Interesting theoretical results relating ℓ_1 and ℓ_0 penalties.

Covariance Matrices

The covariance matrix of p variables X_1, \dots, X_p is the $(p \times p)$ -matrix

$$\Sigma = (\text{cov}(X_i, X_j)).$$

It gives a first impression of dependencies, and it is also the basis of further analysis techniques (principal component analysis, discriminant analysis, graphical models, ...).

The usual estimator based on a random sample X^1, \dots, X^n from the distribution of $X = (X_1, \dots, X_p)^T$ is

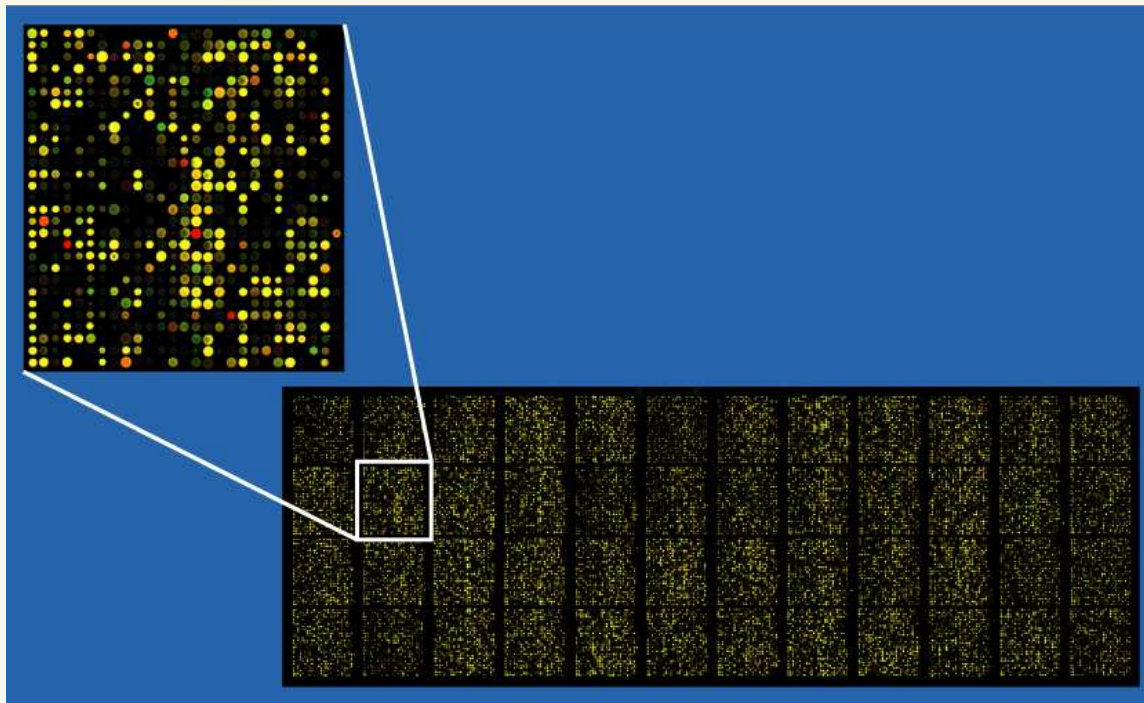
$$S_n = \frac{1}{n} \sum_{i=1}^n (X^i - \bar{X}_n)(X^i - \bar{X}_n)^T.$$

If $p = p_n \rightarrow \infty$, then S_n does not converge to Σ relative to matrix norms, unless $p_n \ll n$. However, sparse matrices can be estimated.

The Previous Hype

Genomics

About 10 years ago 90 % of statistics jumped onto the analysis of micro-arrays. These were a main example of $p \gg n$, for instance 20 000 measurements on every individual in a sample of size 100.

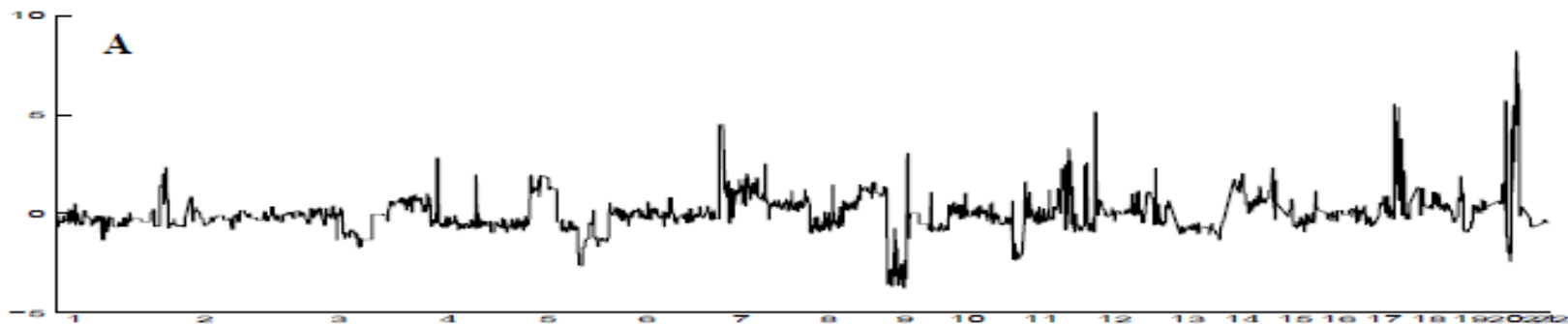


Genomics

About 10 years ago 90 % of statistics jumped onto the analysis of micro-arrays. These were a main example of $p \gg n$, for instance 20 000 measurements on every individual in a sample of size 100.

This massive data has multiplied:

- gene expression arrays
- gene copy number arrays
- proteomics profiles
- single nucleotide polymorphisms
- metabolomic data



Impact

$$\text{impact factor 2009} = \frac{\text{\#citations in 2009 of articles published in 2007-8}}{\text{\#articles published in 2007-8}}$$

Abbreviated Journal Title	ISSN	2007 Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	2007 Articles	Cited Half-life	Eigenfactor™ Score	Article Influence™ Score
ANN PROBAB	0091-1798	2668	1.270	1.349	0.182	77	>10.0	0.01595	1.653
ANN STAT	0090-5364	8107	1.944	2.862	0.192	104	>10.0	0.02818	2.886
BIOINFORMATICS	1367-4803	26676	5.039	6.649	0.531	686	4.4	0.15579	2.406
BIOMETRIKA	0006-3444	9511	1.156	1.706	0.149	74	>10.0	0.01539	1.724
BIostatISTICS	1465-4644	1390	3.058	5.178	0.828	58	4.4	0.01508	3.118
J AM STAT ASSOC	0162-1459	15859	2.086	3.382	0.269	119	>10.0	0.03642	2.978
J R STAT SOC B	1369-7412	8148	2.210	4.630	0.273	44	>10.0	0.02232	4.175
PROBAB THEORY REL	0178-8051	1461	1.295	1.352	0.310	58	9.8	0.01319	1.708
SCAND J STAT	0303-6898	1301	1.118	1.364	0.244	45	>10.0	0.00672	1.332

Multiple Testing

Genomics has rekindled interest in **multiple testing**: for instance for every of 20 000 genes one compares, e.g. by a **t-test**, expression of that gene in samples of healthy and cancer tissues.

What about the overall significance?

We used to teach people that they should not perform many tests at the same time.

False Discovery Rate

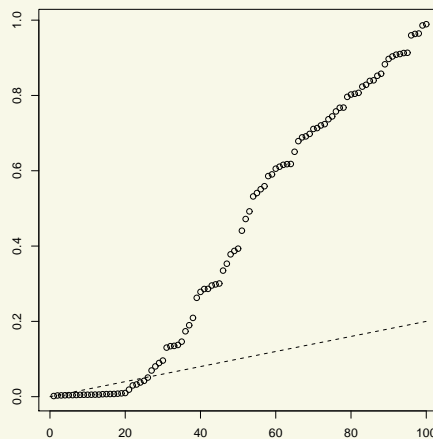
An innovation was the use of the **false discovery rate**

$$FDR = E \frac{\# \text{true hypotheses}}{\# \text{rejected hypotheses}}.$$

THEOREM [Benjamini, Hochberg, JRSSb 1995]

Given N independent p -values, reject all hypotheses i with $p_{(i)} \leq i\alpha/N$.

This has $FDR \leq \alpha\pi_0$, for π_0 the fraction of true hypotheses.



False Discovery Rate

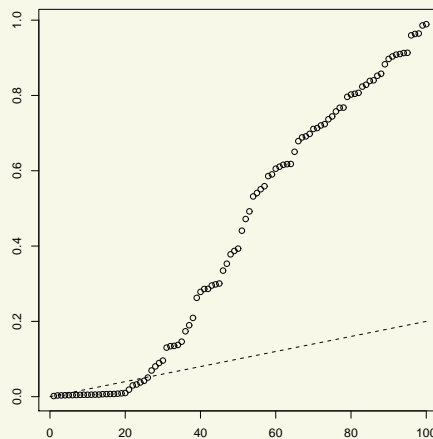
An innovation was the use of the **false discovery rate**

$$FDR = E \frac{\# \text{true hypotheses}}{\# \text{rejected hypotheses}}.$$

THEOREM [Benjamini, Hochberg, JRSSb 1995]

Given N independent p -values, reject all hypotheses i with $p_{(i)} \leq i\alpha/N$.

This has $FDR \leq \alpha\pi_0$, for π_0 the fraction of true hypotheses.



Cited: MathSciNet: 83 times;

Web of Knowledge: 4409 times

False Discovery Rate

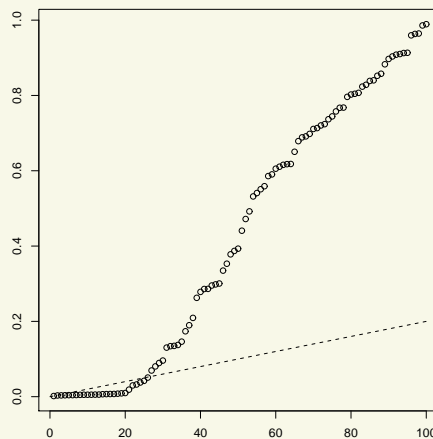
An innovation was the use of the **false discovery rate**

$$FDR = E \frac{\# \text{true hypotheses}}{\# \text{rejected hypotheses}}.$$

THEOREM [Benjamini, Hochberg, JRSSb 1995]

Given N independent p -values, reject all hypotheses i with $p_{(i)} \leq i\alpha/N$.

This has $FDR \leq \alpha\pi_0$, for π_0 the fraction of true hypotheses.



The result has rekindled an interest in **empirical Bayes methods** in order to estimate π_0 .

Higher Criticism

Many competitors and variants of BH have been developed, including **higher criticism** [Donoho and Jin, PNAS 2008.]

$$HC(i; p_{(i)}) = \frac{\sqrt{N}(i/N - p_{(i)})}{\sqrt{i/N(1 - i/N)}}.$$

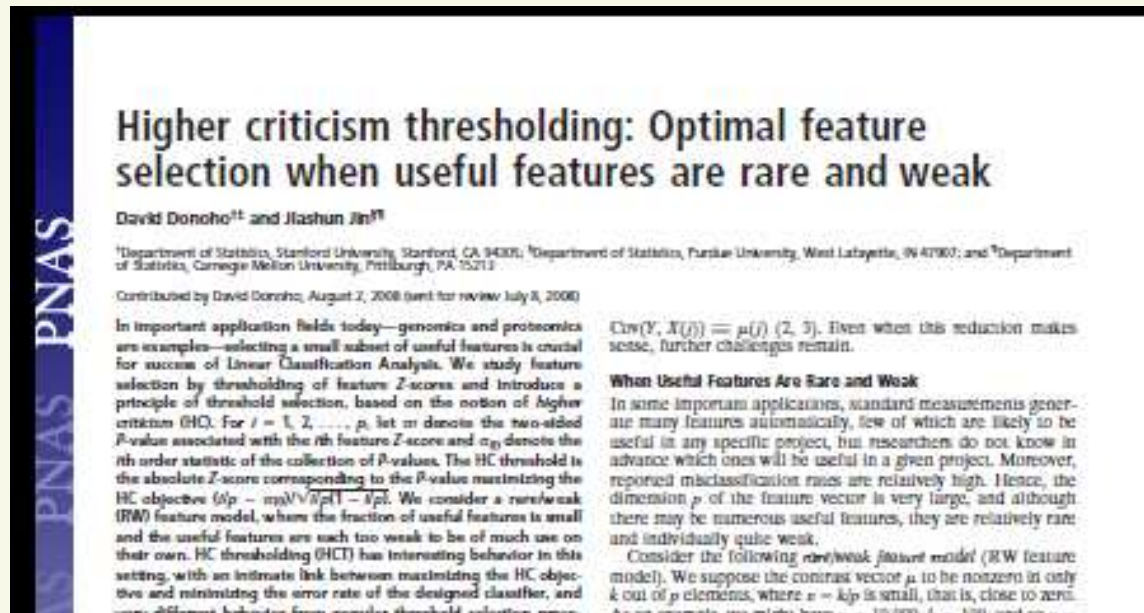
Reject all hypotheses with $i \leq \operatorname{argmax}_i HC(i; p_{(i)})$.

Higher Criticism

Many competitors and variants of BH have been developed, including **higher criticism** [Donoho and Jin, PNAS 2008.]

$$HC(i; p_{(i)}) = \frac{\sqrt{N}(i/N - p_{(i)})}{\sqrt{i/N(1 - i/N)}}.$$

Reject all hypotheses with $i \leq \operatorname{argmax}_i HC(i; p_{(i)})$.



Abbreviated Journal Title	ISSN	2007 Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	2007 Articles	Cited Half-life	Eigenfactor™ Score	Article Influence™ Score
P NATL ACAD SCI USA	0027-8424	394223	9.598	10.369	1.724	3494	7.3	1.74485	4.929

Higher Criticism

Many competitors and variants of BH have been developed, including **higher criticism** [Donoho and Jin, PNAS 2008.]

$$HC(i; p_{(i)}) = \frac{\sqrt{N}(i/N - p_{(i)})}{\sqrt{i/N(1 - i/N)}}.$$

Reject all hypotheses with $i \leq \operatorname{argmax}_i HC(i; p_{(i)})$.

This has theoretical optimality properties in a toy model with independent normal test statistics with means either 0 or μ (**sparse!**), AND:

Table 1. Error rates of standard classifiers on standard examples from Dettling (16)

Method	ALL/reg	Col/reg	Pro/reg	m-reg	R
Bagboo	4.08/0.59	16.10/0.52	7.53/0	0.59	6
Boost	5.67/1	19.14/1	8.71/0.18	1	7.5
RanFor	1.92/0.02	14.86/0.32	9.00/0.22	0.32	2
SVM	1.83/0	15.05/0.35	7.88/0.05	0.35	3
DLDA	2.92/0.28	12.86/0	14.18/1	1	7.5
KNN	3.83/0.52	16.38/0.56	10.59/0.46	0.56	5
PAM	3.55/0.45	13.53/0.11	8.87/0.20	0.45	4
HCT	2.86/0.27	13.77/0.14	9.47/0.29	0.29	1

reg, regret; col, colon; Pro, prostate; m-reg, maximum regret; R, rank based on m-reg.

Higher Criticism



John W. Tukey, 1915–2000

APPENDIX D. SOME JWT WORDS (WITH NEW MEANINGS)
AND PHRASES

alanysis	biweight
alias (in time series)	bland distribution
ANOVA	borrowing strength
badmandments	boxplot
bagplot	cepstrum
batch	coco
bispectrum	complex demodulation
bit	confirmatory data analysis (CDA)

1570

D. R. BRILLINGER

darius	polyspectrum
data analysis	prewhitening
dedomulation	quefreny
deficiency	RadGaussianization
depth	rahmonic
dyadic ANOVA	regressogram
exploratory data analysis (EDA)	reroughing
faceless value	rootogram
family of covers	rough
fences	running median
5-number summary	saphe cracking
flogs	schematic plots
froots	slash distribution
finite character	smear-and-sweep
Garden of Eden	smelting
hamming	smoothing and decimation
(hanging) rootogram	software (first in print)

From Brillinger, AS 2002

The Competitors

Machine Learning

Ten years ago many statisticians expressed fear that they would lose ground to computer scientists, who

- are good in implementations, in particular with large data-sets.
- are better in selling their trade (e.g. **learning** versus **regression**, **machine** versus **regression function**).

The fear was misguided:

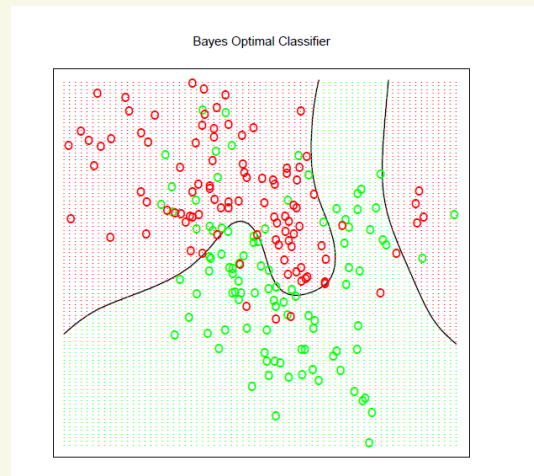
- Many applications need probabilistic models (noise, biased sampling, censoring, dependencies, causality, ...).
- There has been much interaction and cooperation (Bayesian methods, theory for **“statistical learning”**, ...).

Classification

Observe a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of points in $\mathcal{X} \times \{-1, 1\}$.
Construct a function $f: \mathcal{X} \rightarrow \mathbb{R}$ such that $\text{sign } f(X)$ is a good predictor of Y for an additional random point (X, Y) .

The **risk** $f \mapsto \mathbb{P}(Y \neq f(X))$ is minimized by the **Bayes classifier**

$$f(x) = \text{sign}(\mathbb{P}(Y = 1 | X = x) > 1/2).$$

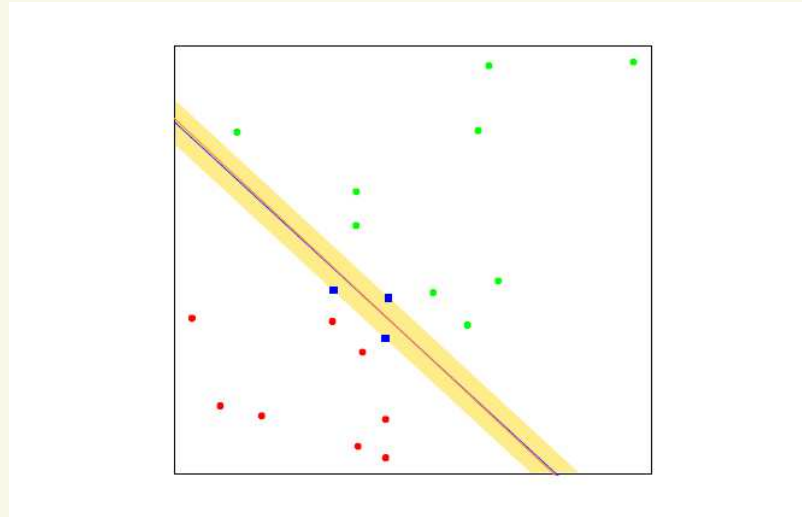


Now use the data to approximate the Bayes classifier.

Many methods (Fisher's discriminant, (structural) empirical risk minimization, neural nets, support vector machines, boosting,...).

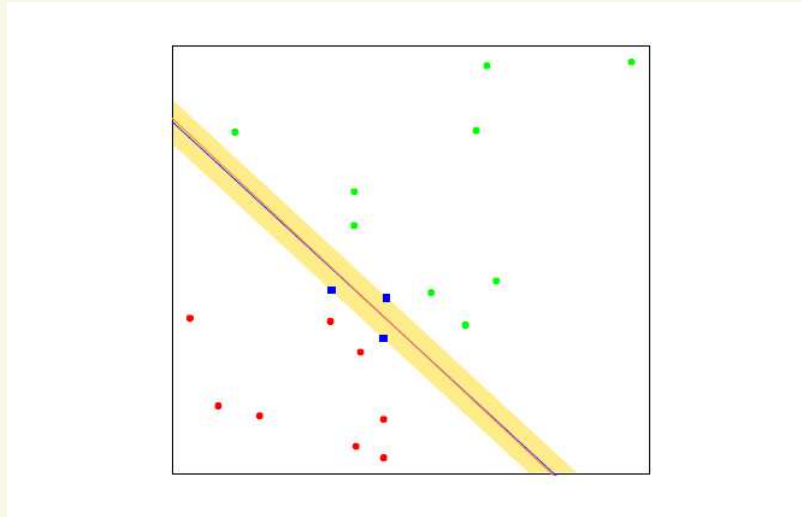
Support Vector Machine

A SVM [Vapnik, 1998] is defined as the hyperplane that maximizes the margin between the two types of input.



Support Vector Machine

A SVM [Vapnik, 1998] is defined as the hyperplane that maximizes the margin between the two types of input.



If the samples are not separable, then one relaxes the numerical optimization problem, and uses the **soft margin SVM**:

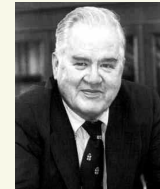
$$\operatorname{argmin}_{(\beta, b) \in \mathbb{R}^d \times \mathbb{R}} \left[\sum_{i=1}^n \left(1 - y_i (\langle x_i, \beta \rangle + b) \right)_+ + \lambda \|\beta\|^2 \right]$$

This is **penalized empirical risk**, with **hinge loss** $\ell(y, x, f) = (1 - yf(x))_+$.

Hinge Loss

$$\ell(y, x, f) = (1 - yf(x))_+$$

darius	polyspectrum
data analysis	prewhitening
dedomulation	quefreny
deficiency	RadGaussianization
depth	rahmonic
dyadic ANOVA	regressogram
exploratory data analysis (EDA)	reroughing
faceless value	rootogram
family of covers	rough
fences	running median
5-number summary	saphe cracking
flogs	schematic plots
froots	slash distribution
finite character	smear-and-sweep
Garden of Eden	smelting
hamming	smoothing and decimation
(hanging) rootogram	software (first in print)
hanning	stem-and-leaf
hat matrix, H	tapering
hinge	toolglass
Huberizing	trimming
jackknife	twicing
linear programming	vacuum cleaner
midmean	vague concept
multihaver	window carpentry
Munkery	winsorizing
polyefficiency	Winsor's principle
polykay	Zorn's Lemma
polysampling	



Statistical Learning

The linear classifier can be generalized to functions in a **Reproducing Kernel Hilbert Space**.

$$\hat{f} = \operatorname{argmin}_{f \in \mathbb{H}} \left[\sum_{i=1}^n (1 - y_i f(x_i))_+ + \lambda \|f\|_{\mathbb{H}}^2 \right].$$

THEOREM [Blanchard, Massart, 2007]

$$\operatorname{Eregret}(\hat{f}, f_{Bayes}) \leq \inf_f \left[\operatorname{regret}(f, f_{Bayes}) + \lambda \|f\|_{\mathbb{H}} \right] + \lambda.$$

This **oracle inequality** is an instance of general results on penalized empirical risk minimization, and is proved using **empirical process theory**. The ultimate result uses multiple RKHSs, i.e. **adaptation**.

Kernel Methods

The connection to RKHSs led to **kernel methods**.

The screenshot shows a Mozilla Firefox browser window displaying the Amazon.com search results for "kernel methods". The search results are listed under the heading "kernel methods" and include the following items:

- Kernel Methods for Pattern Analysis** by John Shawe-Taylor and Nello Cristianini (Hardcover - Jun 28, 2004)
Buy new: ~~\$94.99~~ **\$75.99** 25 Used & new from \$64.99
Get it by **Monday, May 18** if you order in the next 13 hours and choose one-day shipping.
Eligible for **FREE** Super Saver Shipping.
★ ★ ★ ★ ★ (7)
Excerpt - page 17: "... class of pattern analysis algorithms will be referred to as **kernel methods**. 1.3 Exploiting patterns We wish to design pattern analysis algorithms..."
Surprise me! See a random page in this book.
Books: See all 3,412 items
- Kernel Methods in Computational Biology (Computational Molecular Biology)** by Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert (Hardcover - Aug 1, 2004)
Buy new: ~~\$50.00~~ **\$40.46** 25 Used & new from \$34.00
Get it by **Monday, May 18** if you order in the next 14 hours and choose one-day shipping.
Eligible for **FREE** Super Saver Shipping.
★ ★ ★ ★ ★ (2)
Books: See all 3,412 items
- The Kernel Method of Test Equating** by Alina A. von Davier, Paul W. Holland, and Dorothy T. Thayer (Kindle Edition - Oct 1, 2003) - Kindle Book
Buy: **\$53.52**
Auto-delivered wirelessly
Kindle Store: See all 7 items
- An Introduction to Support Vector Machines and Other Kernel-based Learning Methods** by Nello Cristianini and John Shawe-Taylor (Hardcover - Mar 28, 2000)
Buy new: ~~\$60.99~~ **\$61.60** 28 Used & new from \$44.95
Get it by **Monday, May 18** if you order in the next 14 hours and choose one-day shipping.
Eligible for **FREE** Super Saver Shipping.
★ ★ ★ ★ ★ (8)
Excerpt - page 26: "... Another attraction of the **kernel method** is that the learning algorithms and theory can largely be..."
Surprise me! See a random page in this book.
Books: See all 3,412 items
- Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)** by Bernhard Schölkopf and Alexander J. Smola (Hardcover - Dec 15, 2001)
Buy new: ~~\$75.00~~ **\$50.62** 35 Used & new from \$50.62
Get it by **Monday, May 18** if you order in the next 14 hours and choose one-day shipping.
Eligible for **FREE** Super Saver Shipping.
★ ★ ★ ★ ★ (5)
Books: See all 3,412 items
- NEW Patent CD for Method of kernel selection for image interpolation** by BrainDex LLC
1 Used & new from \$39.95
Everything Else: See all 39 items

The page also features a sidebar with "Listmania!" recommendations, a search bar, and a navigation menu. The bottom of the screenshot shows the Windows taskbar with the Start button and several open applications.

The Equal-Minded

Econometrics

Econometricians are statisticians working on economic/social data.

Econometrics

Econometricians are statisticians working on economic/social data.

They share interest in semiparametrics, censoring, causality, ..., and their winners do very related things.



Nobel Prizes Economics 1997 — 2000 — 2003

Econometrics

Econometricians are statisticians working on economic/social data.

Two “recent” innovations:

- Instrumental variable nonparametric regression
- Partial identification

Econometrics — Instrumental Variables, Identification

It is often assumed that e and X are independent in

$$Y = f(X) + e.$$

This is suitable for controlled experiments, but not in **observational studies**.

Econometricians rarely assume more than that $E(e|X) = 0$, thus **identifying** the regression function as $f(X) = E(Y|X)$.

Even this assumption (**X is exogenous**) is often unrealistic. Econometricians use **instruments** Z

$$Y = f(X) + e, \quad \text{and} \quad E(e|Z) = 0.$$

This can identify f , but leads to an inverse problem.

An alternative is to give bounds for f under realistic assumptions.

Biostatistics

Biostatisticians are statisticians.

Biostatistics

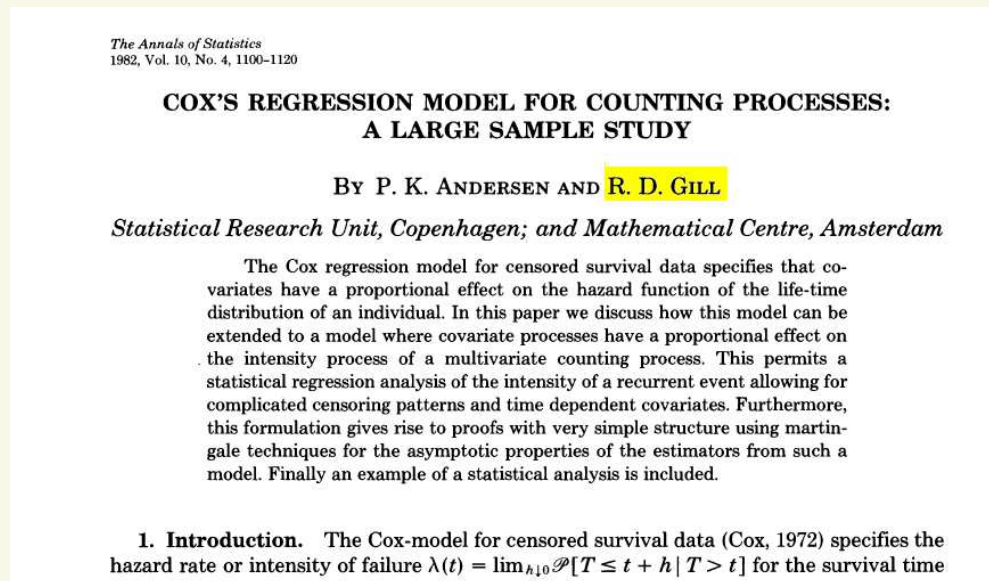
Biostatisticians ought to be mathematical statisticians.

Biostatistics

Biostatisticians are statisticians.

Biostatisticians are statisticians.

We have a certain tradition, and there are still many open questions, concerning methods for **censored data**.



Cited; MathSciNet: 116 times; Web of Knowledge: 942 times.

Biostatisticians often **identify** parameters (e.g. for causal effects) by conditioning on many covariates. This leads to **models with $p > n$** .

The Bayesians

Bayesian statistics



Bayesian inference starts with a **prior** probability distribution Π on the set of parameters θ . The density $x \mapsto p_\theta(x)$ of the data X is viewed as the conditional density given that θ was drawn from Π .

Bayes' rule then gives the **posterior distribution**

$$d\Pi(\theta | X) \propto p_\theta(X) d\Pi(\theta).$$

Bayesian statistics



Bayesian inference starts with a **prior** probability distribution Π on the set of parameters θ . The density $x \mapsto p_\theta(x)$ of the data X is viewed as the conditional density given that θ was drawn from Π .

Bayes' rule then gives the **posterior distribution**

$$d\Pi(\theta | X) \propto p_\theta(X) d\Pi(\theta).$$

Bayesian statistics was a **hype** in the 1990s, because of new computational techniques (**Markov Chain model Carlo**).

Theory for high-dimensional problems is slowly catching up.

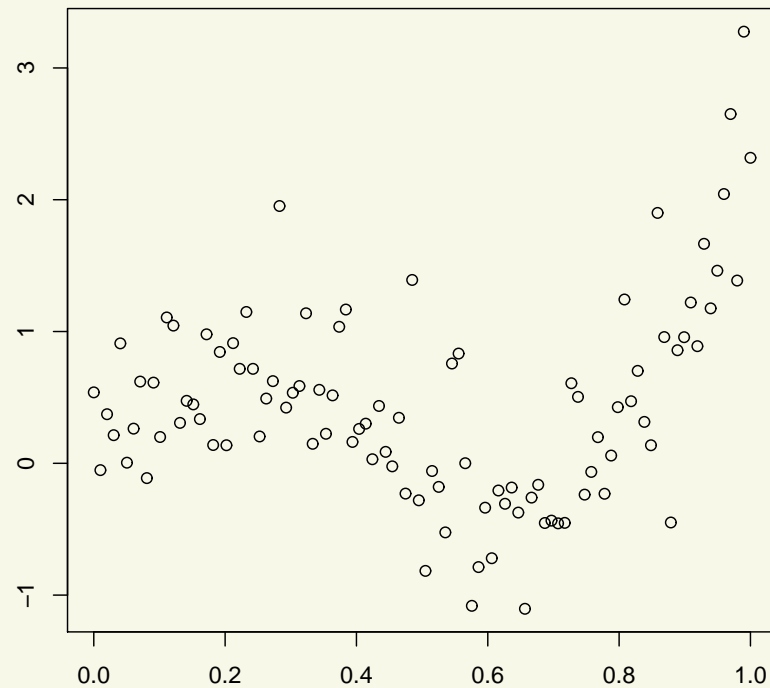
Bayesian methods seem promising for complex and high-dimensional (including sparse) situations (model selection, networks, many covariates,.....).

Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

for e_1, \dots, e_n independent with mean 0.



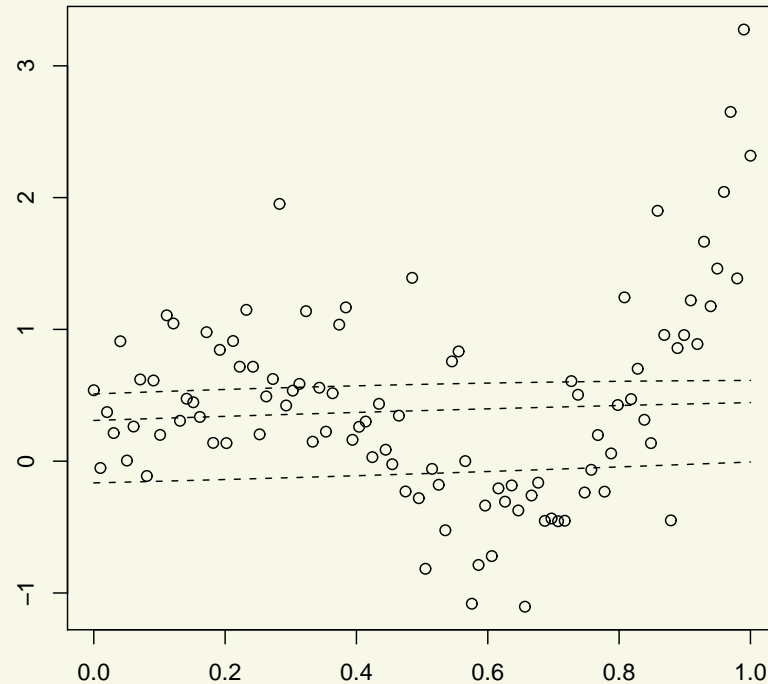
Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

for e_1, \dots, e_n independent with mean 0.

prior, 3 realizations



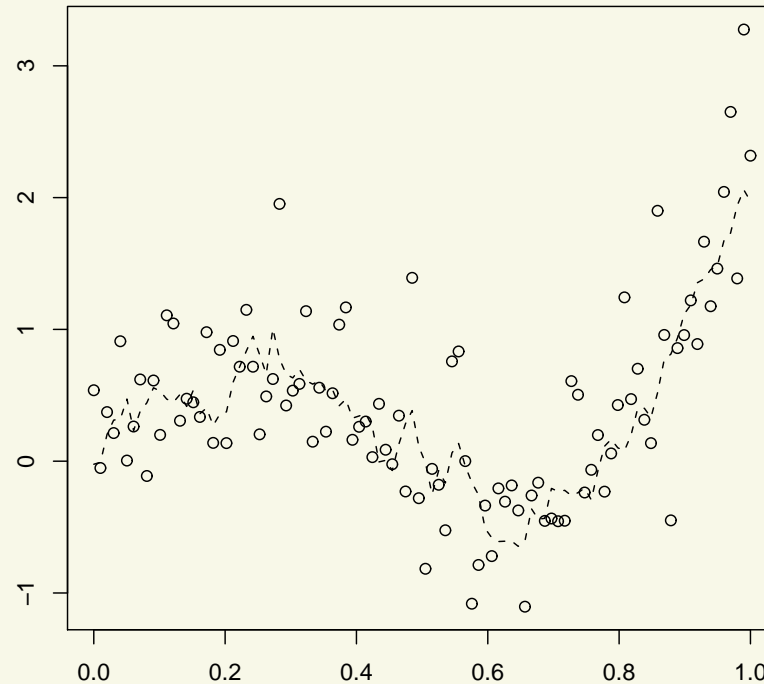
Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

for e_1, \dots, e_n independent with mean 0.

posterior, 1 realization



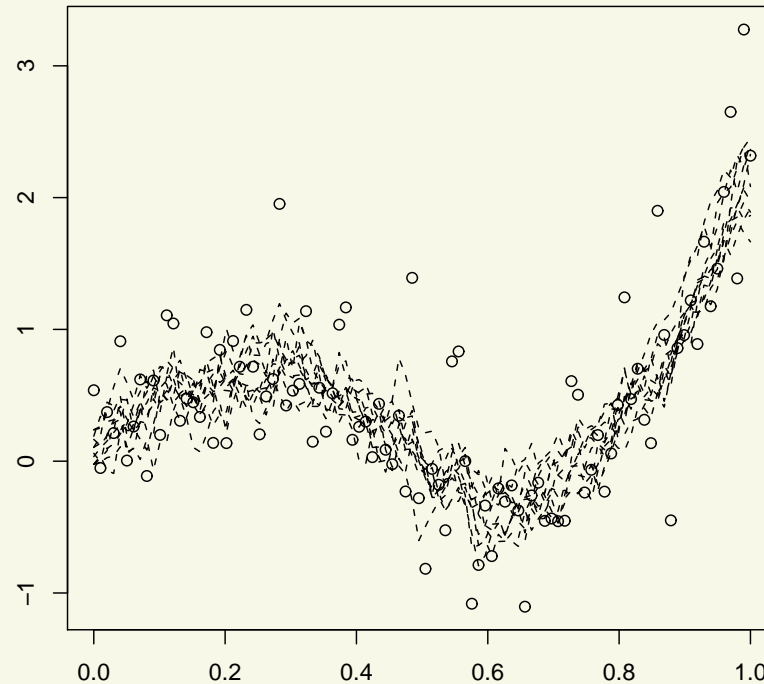
Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

for e_1, \dots, e_n independent with mean 0.

posterior, 20 realizations



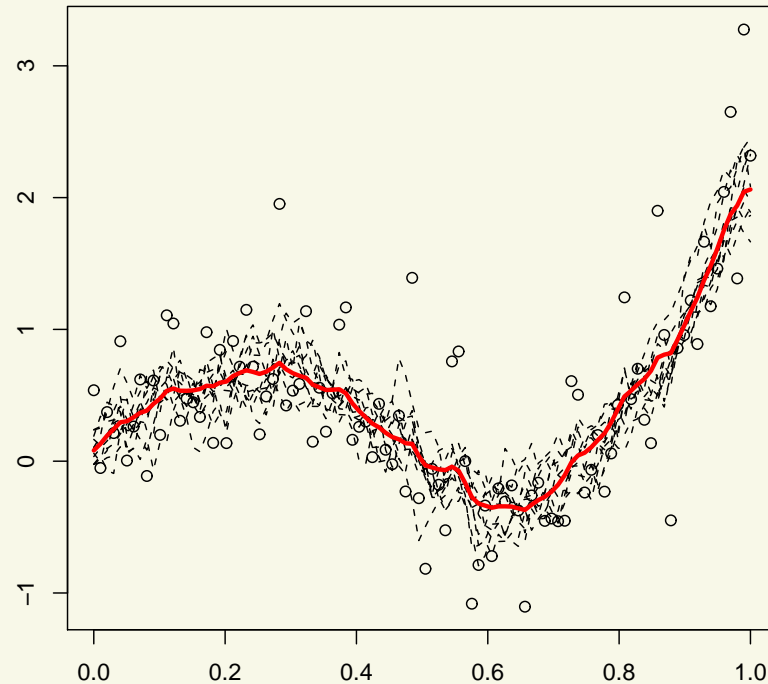
Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

for e_1, \dots, e_n independent with mean 0.

posterior mean



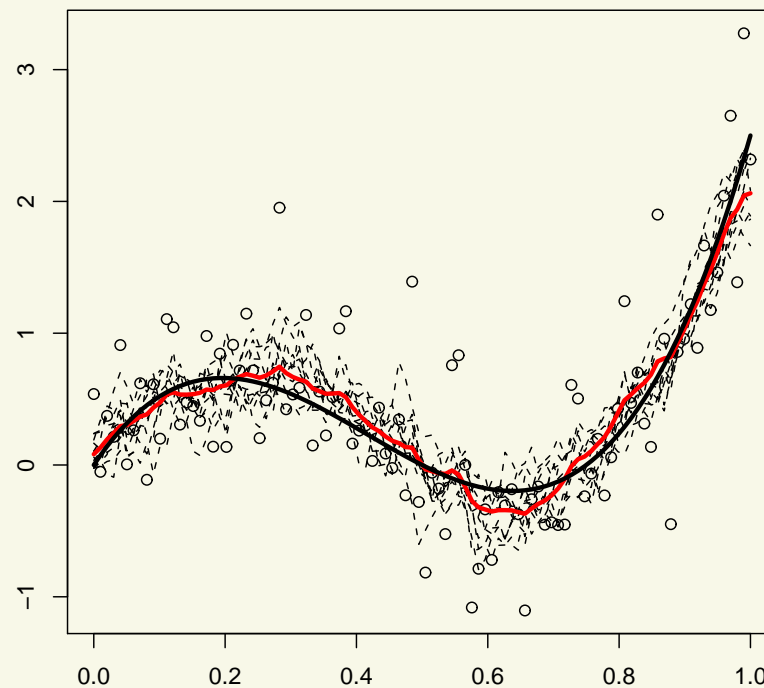
Bayesian Nonparametric Regression

Consider estimating θ based on $(X_1, Y_1), \dots, (X_n, Y_n)$ with

$$Y_i = \theta(X_i) + e_i,$$

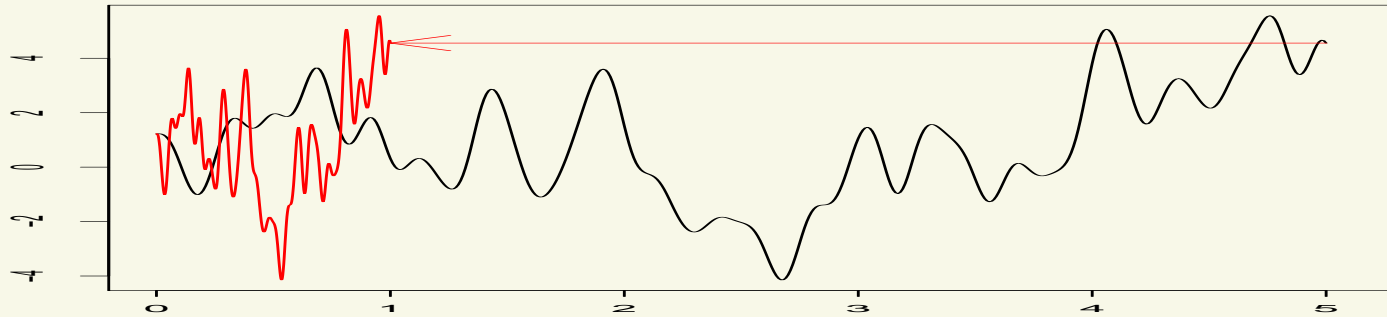
for e_1, \dots, e_n independent with mean 0.

truth



A Theoretical Result

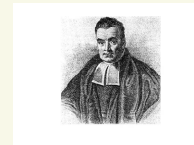
As prior use the law of $t \mapsto G_t/C$ for $t \mapsto G_t$ a centered Gaussian process with $EG_s G_t = \exp(-(s-t)^2)$ and c an independent Gamma variable.



THEOREM

- if $w_0 \in C^\alpha[0, 1]$, then the rate of contraction is nearly $n^{-\alpha/(2\alpha+1)}$.
- if w_0 is supersmooth, then the rate is nearly $n^{-1/2}$.

Reverend Thomas can solve the bandwidth problem!?



The Causes of the Financial Crisis

The Quants

In **derivative pricing** stochastic processes model underlying financial assets (**stocks, interest rates**). Under **no-arbitrage** the price of an **option** is an expected value under some martingale measure.

The Quants

In **derivative pricing** stochastic processes model underlying financial assets (**stocks, interest rates**). Under **no-arbitrage** the price of an **option** is an expected value under some martingale measure.



Derivative pricing has been a driving force to create our current financial crisis (e.g. **CDSs, CDOs**).

Can we contribute better models than Black-Scholes (general semimartingales: stochastic volatility, jump processes, multivariate term structures,...) and develop tools to apply them?

The Risk Managers

Historical analysis of financial processes allows to predict the risk position of a bank. Main challenges:

- highly multivariate dependencies
- rare and extreme events



The Risk Managers

Historical analysis of financial processes allows to predict the risk position of a bank. Main challenges:

- highly multivariate dependencies
- rare and extreme events



Smart statisticians have not been able to prevent the banks from incurring too many risks. (Are there **enough** smart statisticians in our financial institutions?)

The Risk Managers

Historical analysis of financial processes allows to predict the risk position of a bank. Main challenges:

- highly multivariate dependencies
- rare and extreme events

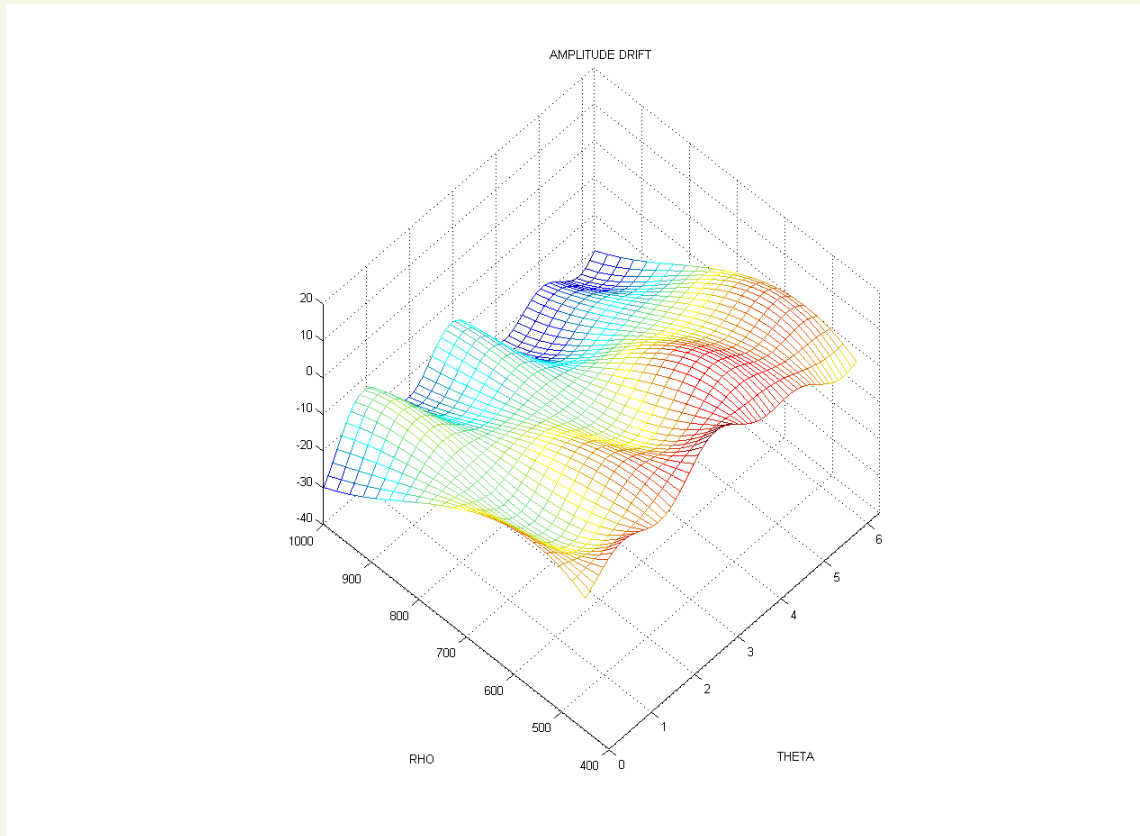


Smart statisticians have not been able to prevent the banks from incurring too many risks. (Are there **enough** smart statisticians in our financial institutions?)

We definitely need more statistics (state space models, high frequency data, multivariate extremes,...)

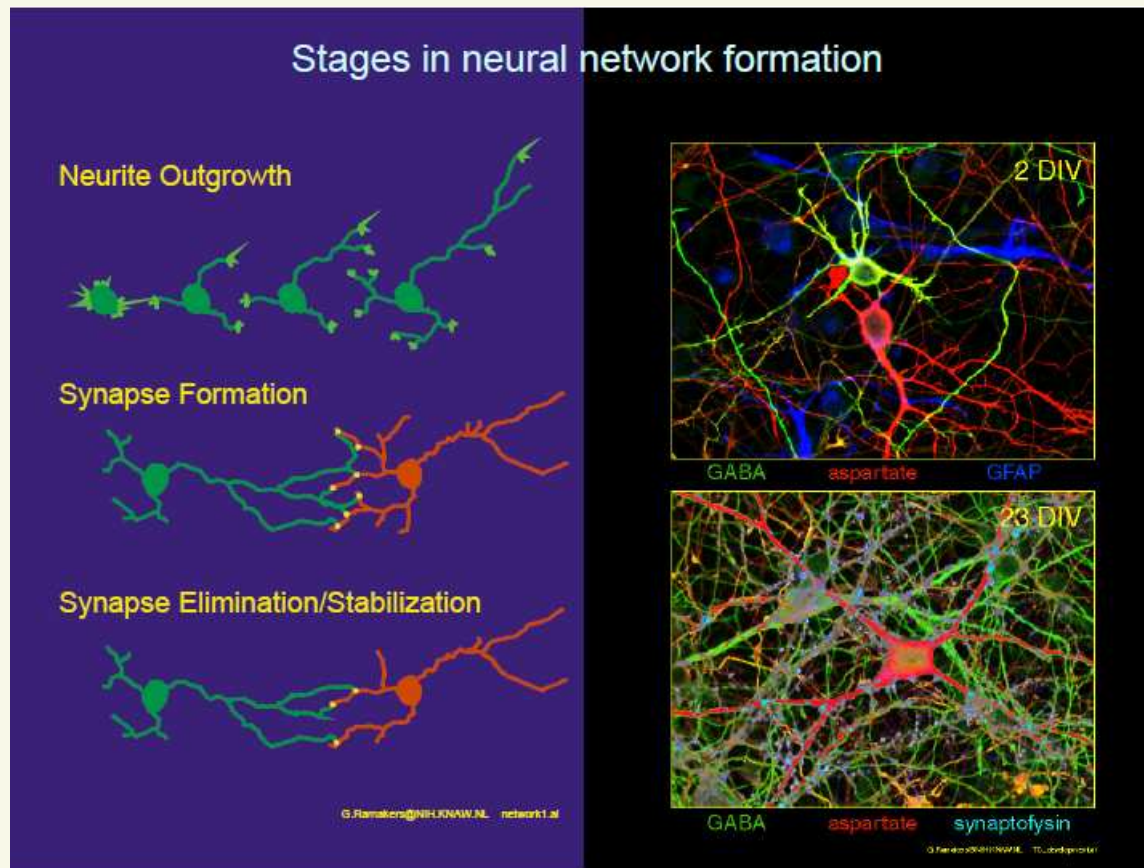
Life Itself

Phase Synchronization

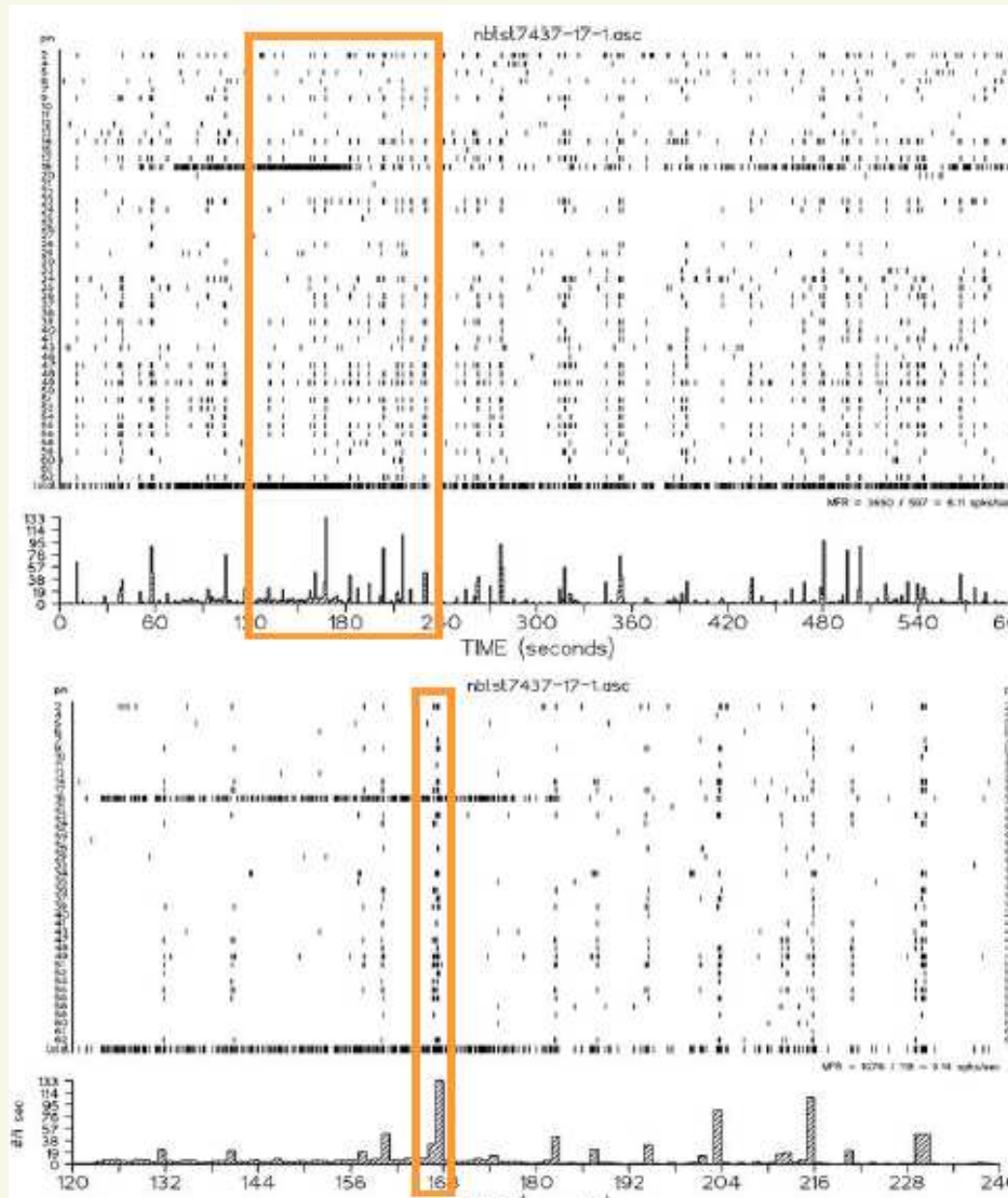


$$d\phi_t^1 = \mu_1(\phi_t) dt + \sigma_1(\phi_t) dW_t^1$$
$$d\phi_t^2 = \mu_2(\phi_t) dt + \sigma_2(\phi_t) dW_t^2.$$

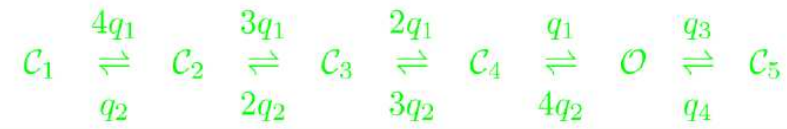
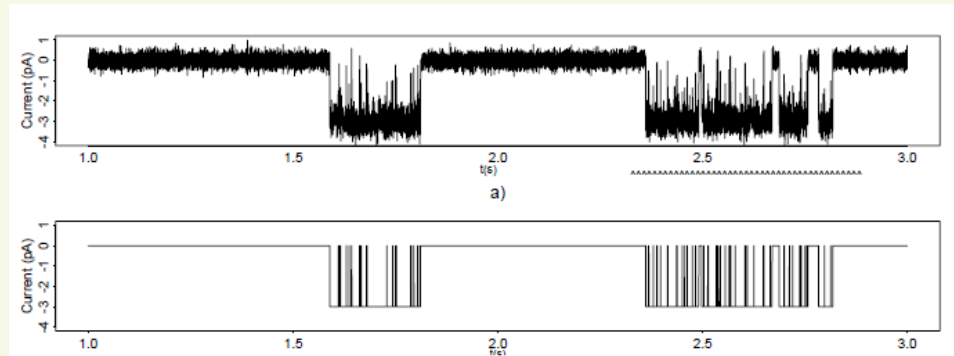
Neural Networks, Spike Trains



Neural Networks, Spike Trains



Ion Channels



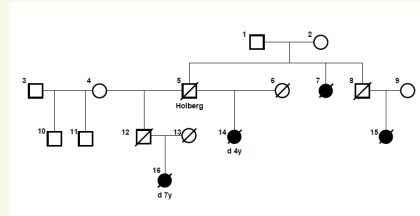
Statistical Genetics



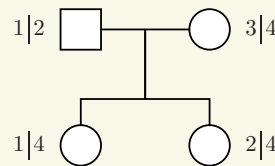
Statistical Genetics

Methods for gene finding have shifted

from **parametric linkage** in big families:



to **nonparametric linkage** in small families:



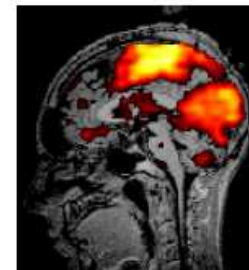
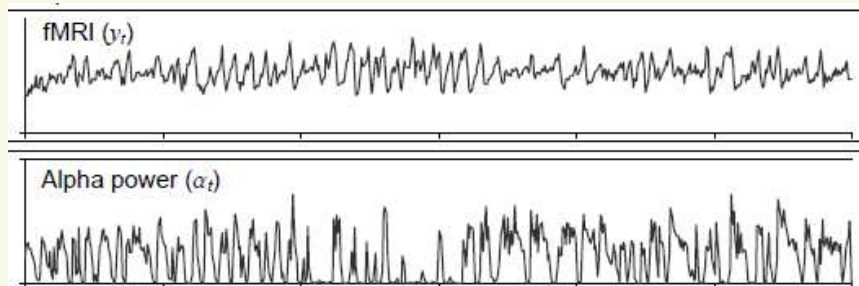
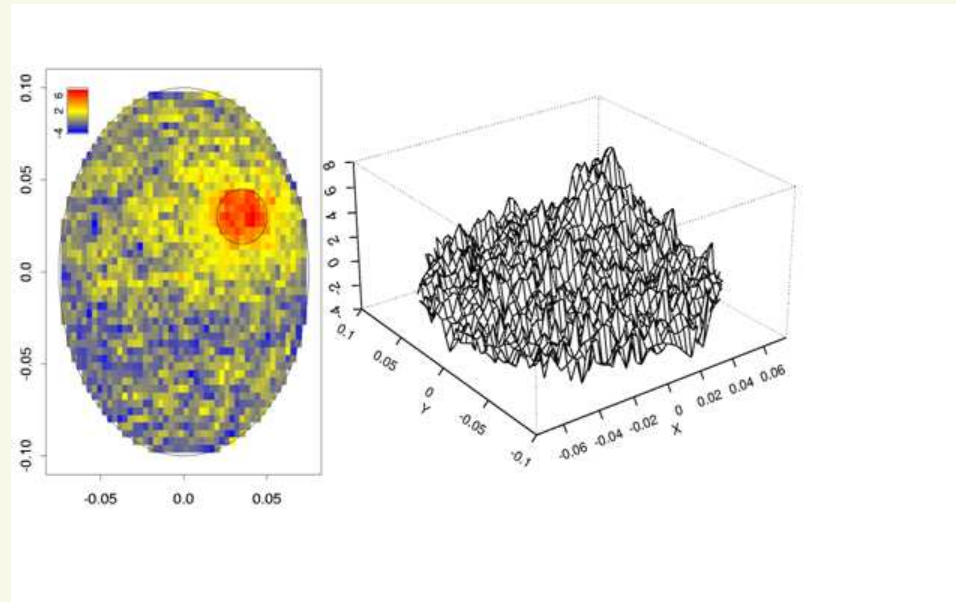
to **association** in unrelated individuals:



Population structure and dynamics have become more important.

Perhaps the biggest problem is $p > n$: how to find the effects of 30000 genes and their **interactions**.

Signals and Pictures



The Future Hypes?

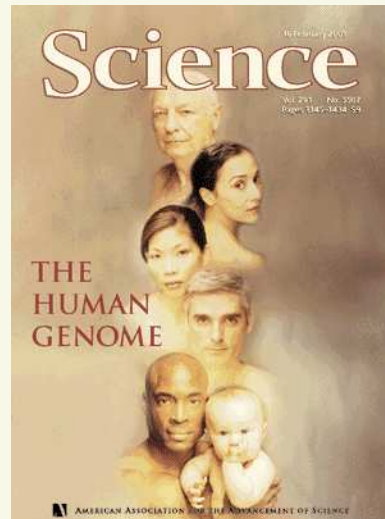
Integromics

Genomics, proteomics, metabolomics.

Many types of data.

Data-bases.

Models.



Structures

Statistics (in particular theory) for:

- graphs (neural networks, causality, genetic networks,...)
- systems of SDEs (neuronal signals, systems biology, finance,...)
- particle systems (systems biology,..)
- spatio-temporal processes (population dynamics, systems biology, neuroscience, ...)
- state space models (finance, neuro-science, ion channels, genes,...)

?????

Statistics has become an extraordinarily broad field.

Separations are blurring (biostatistics, econometrics, machine learning,...).

Computation is nontrivial.

Not easy to connect good applied work to good theory (?)