

# Predictive Learning from Data

## LECTURE SET 4

### Statistical Learning Theory (VC-theory)

Cherkassky, Vladimir, and Filip M. Mulier. *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

Source: Dr. Vladimir Cherkassky (revised by Dr. Hsiang-Han Chen)

---

PLEASE DO NOT DISTRIBUTE WITHOUT AUTHOR'S PERMISSION.

# OUTLINE

- Objectives
- Inductive learning problem setting
- Statistical Learning Theory
- Applications
- Measuring the VC-dimension
- Summary and discussion

# Motivation + Objectives

- Statistical Learning Theory (STL), also known as Vapnik-Chervonenkis (VC) Theory.
- One of the best theory for flexible statistical estimation with **finite** samples.
- STL adopts the goal of **system imitation**, rather than **system identification**.
- **Goal:** math theory for STL
- Three aspects of scientific theory: conceptual, technical, practical

# Keep-It-Direct Principle

- The goal of learning is generalization rather than estimation of true function (system identification)

$$\int Loss(y, f(\mathbf{x}, w)) dP(\mathbf{x}, y) \rightarrow \min$$

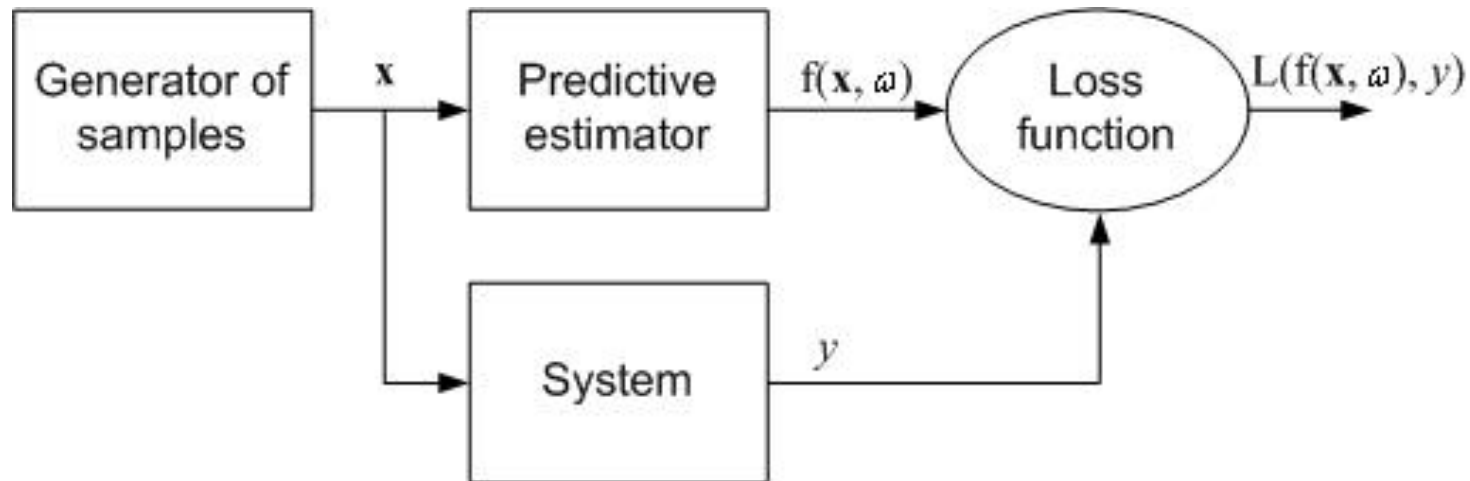
- **Keep-It-Direct Principle** (Vapnik, 1995)

**Do not solve an estimation problem of interest by solving a more general (harder) problem as an intermediate step**

- Good predictive model reflects **certain properties** of unknown distribution  $P(\mathbf{x}, y)$
- Since model estimation with finite data is ill-posed, one should never try to solve a more general problem than required by a given application  
→ Importance of formalizing application requirements as a predictive learning problem.

# Inductive Learning Setting

- The learning machine observes samples  $(\mathbf{x}, y)$ , and returns an estimated response  $\hat{y} = f(\mathbf{x}, w)$
- Two modes of inference: identification vs imitation
- Risk  $\int Loss(y, f(\mathbf{x}, w)) dP(\mathbf{x}, y) \rightarrow \min$



# The Problem of Inductive Learning

- **Given:** finite training samples  $\mathbf{Z}=\{(\mathbf{x}_i, y_i), i=1,2,\dots,n\}$  choose from a given set of functions  $f(\mathbf{x}, \mathbf{w})$  the one that *approximates best* the true output. (in the sense of risk minimization)

## *Concepts and Terminology*

- approximating functions  $f(\mathbf{x}, \mathbf{w})$
- (non-negative) loss function  $L(f(\mathbf{x}, \mathbf{w}), y)$
- expected risk functional  $R(\mathbf{Z}, \mathbf{w})$

**Goal:** find the function  $f(\mathbf{x}, \mathbf{w}_o)$  minimizing  $R(\mathbf{Z}, \mathbf{w})$  when the joint distribution  $P(\mathbf{x}, y)$  is *unknown*.

# Empirical Risk Minimization

- ERM principle in model-based learning
  - Model parameterization:  $f(\mathbf{x}, \mathbf{w})$
  - Loss function:  $L(f(\mathbf{x}, \mathbf{w}), \mathbf{y})$
  - Estimate risk from data:  $R_{emp}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n L(f(\mathbf{x}_i, \mathbf{w}), y_i)$
  - Choose  $\mathbf{w}^*$  that minimizes  $R_{emp}$
- Statistical Learning Theory developed from the theoretical analysis of ERM principle under finite sample settings

# OUTLINE

- Objectives
- Inductive learning problem setting
- **Statistical Learning Theory**
- Applications
- Measuring the VC-dimension
- Summary and discussion



# Statistical Learning Theory

- History and Overview
- Conditions for consistency and convergence of ERM
- VC-dimension
- Generalization bounds
- Structural Risk Minimization (SRM) for learning with finite samples

# History and Overview

- SLT aka **VC-theory** (Vapnik-Chervonenkis)
- Theory for estimating dependencies from finite samples (**predictive learning setting**)
- Based on the ***risk minimization*** approach
- All main results originally developed in 1970's for classification (pattern recognition) but remained largely unknown
- Renewed interest since late 90's due to practical success of **Support Vector Machines (SVM)**

# History and Overview(cont'd)

## MAIN CONTRIBUTIONS

- **Distinction** between problem setting, inductive principle and learning algorithms
- **Direct approach** to estimation with finite data (KID principle)
- **Math analysis of ERM** (inductive setting)
- **Two factors responsible for generalization:**
  - empirical risk (fitting/ training error)
  - complexity(capacity) of approximating functions

# History and Overview(cont'd)

## **VC-theory has 4 parts:**

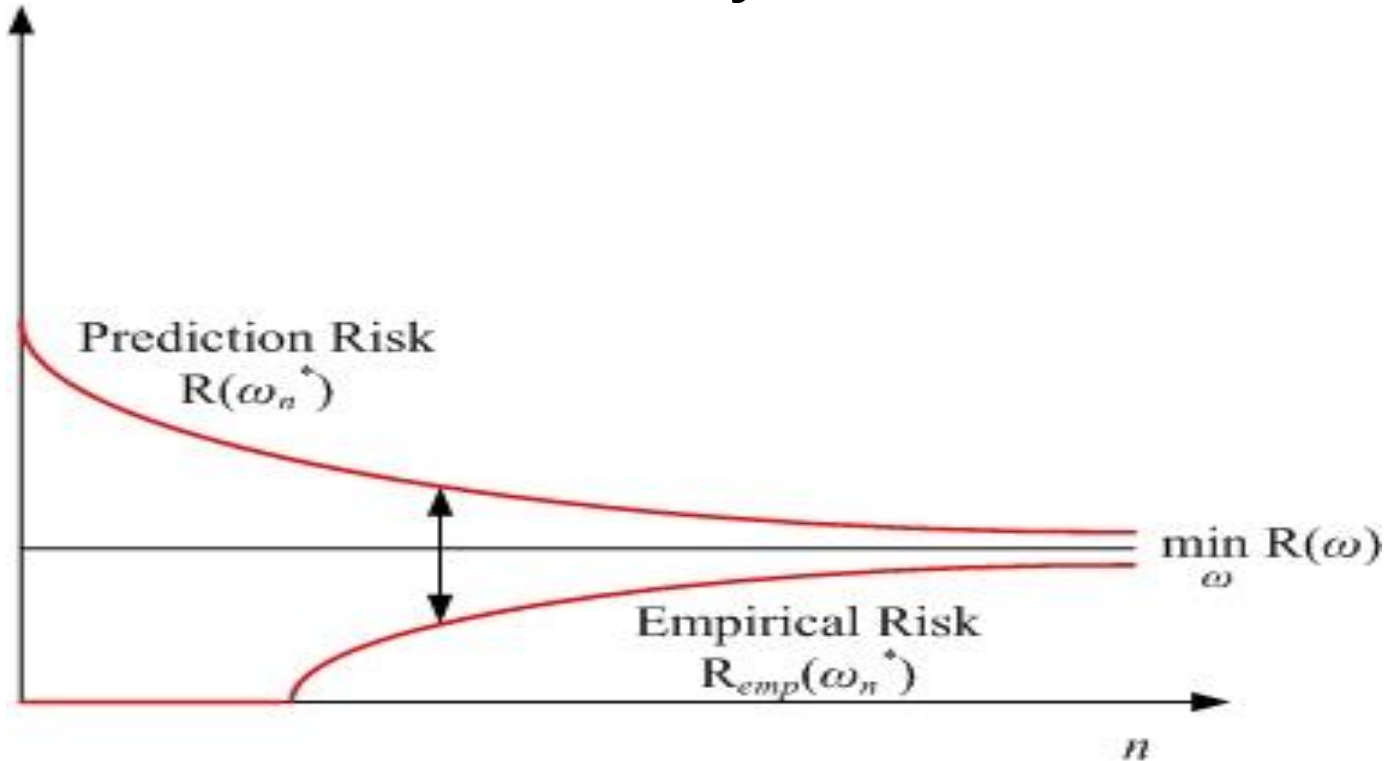
1. Conditions for consistency/ convergence of ERM
2. Generalization bounds
3. Inductive principles (for finite samples)
4. Constructive methods (learning algorithms) for implementing (3)

**NOTE:**  $(1) \rightarrow (2) \rightarrow (3) \rightarrow (4)$

# Consistency/Convergence of ERM

- Empirical Risk **known** but Expected Risk **unknown**
- **Asymptotic consistency requirement:**  
under what (general) conditions models providing *min Empirical Risk* will also provide *min Prediction Risk*, when the number of samples grows large?
- **Why asymptotic analysis** is important?
  - helps to develop useful concepts
  - necessary and sufficient conditions ensure that VC-theory is general and cannot be improved

# Consistency of ERM



- Convergence of empirical risk  $R_{emp}(\omega)$  *does not imply consistency* of ERM
- Models estimated via ERM ( $w^*$ ) are always *biased estimates* of the functions minimizing expected risk:

$$R_{emp}(\omega_n^*) < R(\omega_n^*)$$

# Key Theorem of VC-theory

- For bounded loss functions, the ERM principle is consistent *if and only if* the empirical risk  $R_{emp}(\omega)$  *converges uniformly* to the true risk  $R(\omega)$  in the following sense

$$\lim_{n \rightarrow \infty} P[\sup_{\omega} |R(\omega) - R_{emp}(\omega)| > \varepsilon] = 0, \forall \varepsilon > 0$$

- consistency is determined by the *worst-case approximating function*, providing the largest discrepancy btwn the empirical risk and true risk

**Note:** this condition is *not useful in practice*. Need conditions for consistency in terms of general properties of a set of loss functions (approximating functions)

# Conditions for Consistency of ERM

- **Goal:** to derive conditions for consistency & fast convergence in terms of the **properties of loss functions**
- Indicator 0/1 loss functions (binary classification)

$$Q(z, \omega) = |y - f(x, \omega)|$$

- Each indicator function partitions a given sample  $Z_n = \{z_i\}, i = 1, 2, \dots, n$  into two subsets (two classes).

Each such partitioning is called **dichotomy (二分法)**

- **The diversity**  $N(Z_n)$  of a set of functions  $Q(z, \omega)$  is the number of different dichotomies that can be implemented on a random sample  $Z_n$ 
  - the diversity is **distribution-dependent**



- The Growth Function is the maximum number of dichotomies that can be induced on a sample of size  $n$  (using the indicator function)

$$G(n) = \ln \max_{Z_n} N(Z_n)$$

since the max possible number  $N$  is  $2^n$

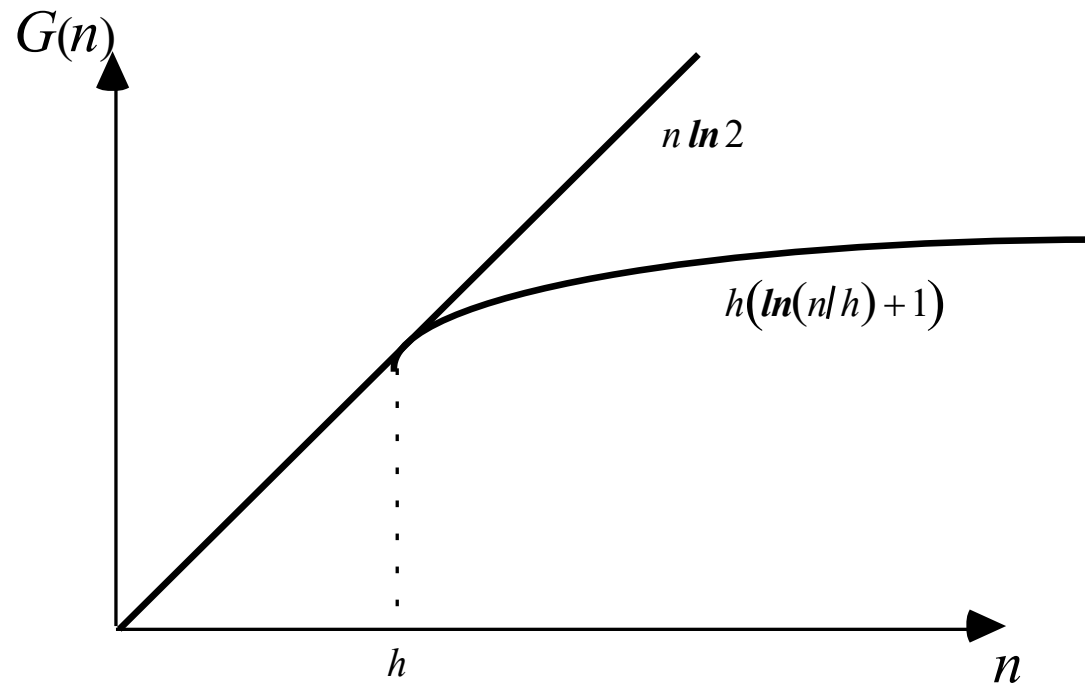
$$G(n) \leq n \ln 2$$

- The Growth Function is **distribution-independent**, and it **depends only** on the properties of a set of functions  $Q(z, \omega)$

**Necessary & sufficient condition for consistency of ERM**

$\lim G(n)/n = 0, n \rightarrow \infty$  (also for **fast rate of convergence**)

# Properties of the Growth Function (GF)



- **Theorem** (Vapnik & Chervonenkis, 1968): The Growth Function is **either linear** or **bounded by a logarithmic function** of the number of samples  $n$ .
- The point  $n=h$  where the GF starts to slow down is called **the VC-dimension** (of a set of indicator functions)

# VC-dimension

- If the bound on the GF stays linear for any  $n$ :  
 $G(n) = n \ln 2$  then the **VC-dimension is infinite**  
 $\rightarrow \lim G(n)/n = \ln 2$  and the ERM is **inconsistent**. That is,  
*any sample of size  $n$  can be split in  $2^n$  possible ways;  
hence no valid generalization is possible*
- **Necessary and sufficient condition** for consistency of ERM:  
Vapnik-Chervonenkis (VC) dimension is **finite**  
(this condition is **distribution-independent**)

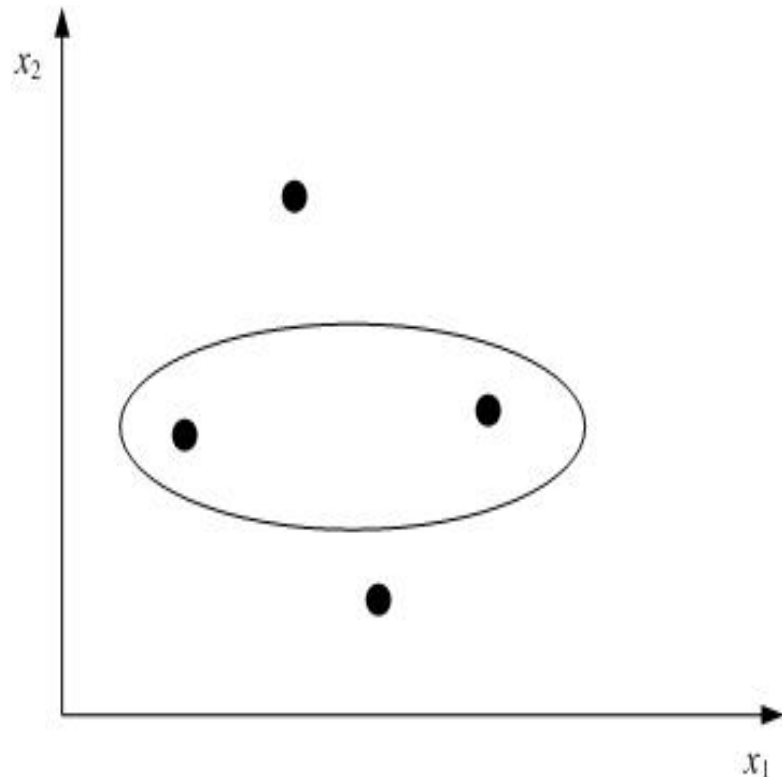
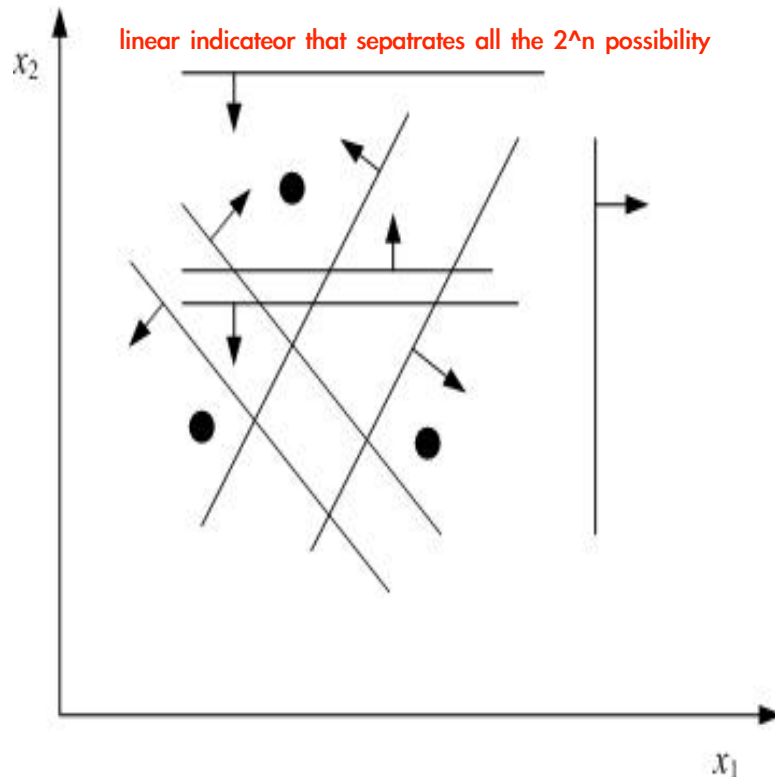
- VC-dimension measures the ability (of a set of functions) to fit or 'explain' available finite data.
- VC-dimension of a set of indicator functions:
  - **Shattering**: if  $n$  samples can be separated by a set of indicator functions in all  $2^n$  possible ways, then these samples can be shattered by this set of functions.
  - A set of functions has VC-dimension  $h$  if there exist  $h$  samples that *can be shattered* by this set of functions, but there *does not exist  $h+1$  samples* that can be shattered.
- **Examples** of analytic calculation of VC-dimension for *simple sets of functions* are shown next

- Linear Indicator Functions ( $d=2$ ):

$d = 2$   
is the  
param  
off the  
func

there exist 3 points that can be shattered, but 4 cannot.

→ VC-dim. = 3. In general,  $h=d+1$

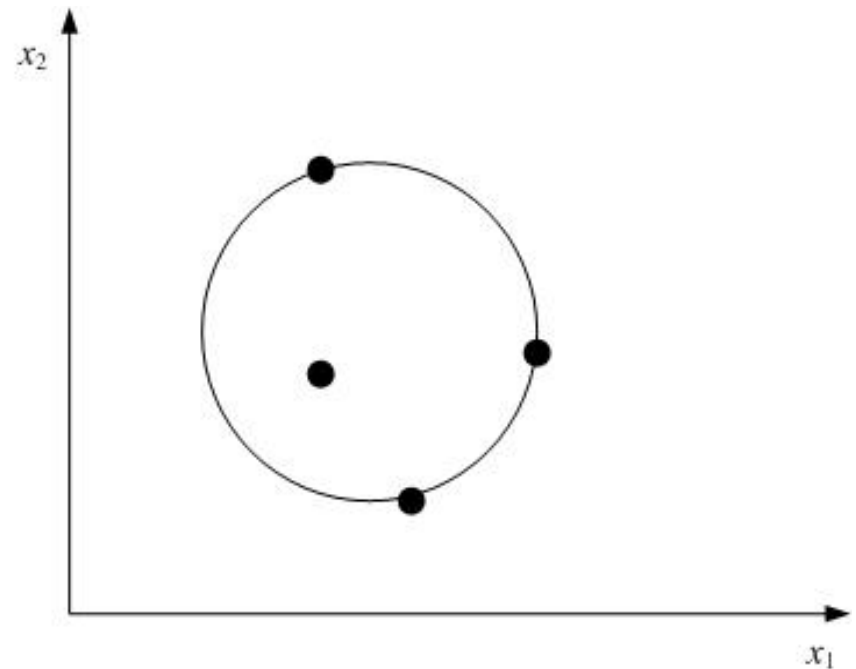
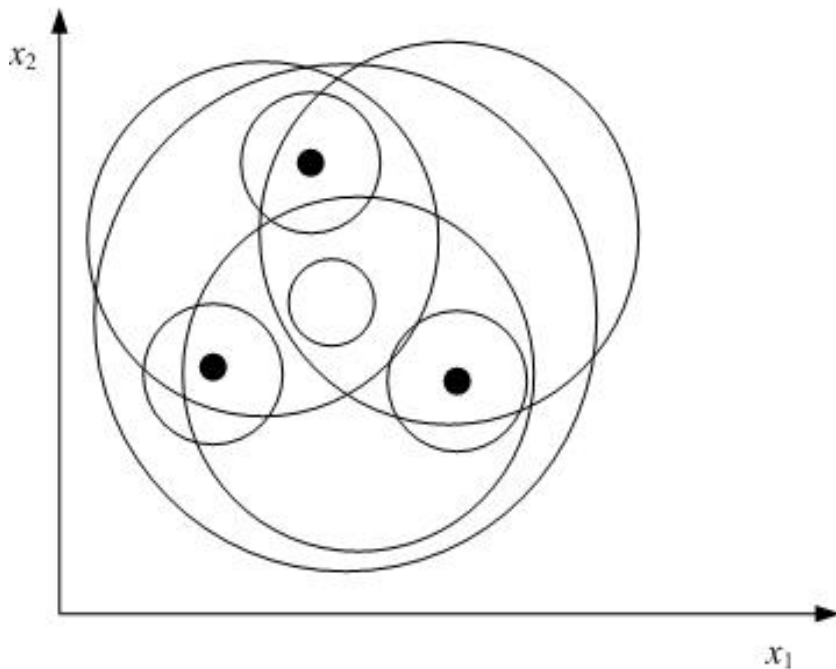


- Spherical (local) functions ( $d=2$ ):

there exist 3 points that can be shattered, but 4 cannot.

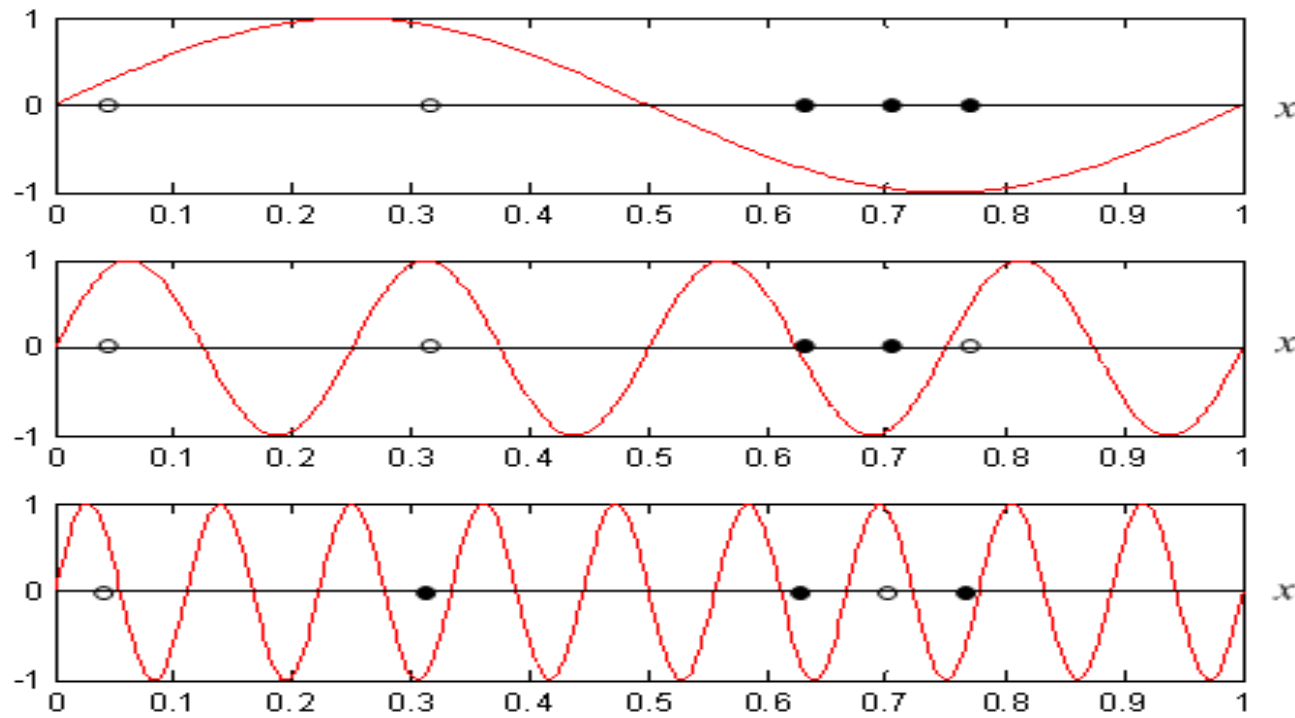
→ VC-dim. = 3. In general,  $h=d+1$

*Note:* in these examples, VC-dim = DoF (# of parameters)



- Example of infinite VC-dimension

A set of indicator functions  $y = I(\sin(\omega x))$  has infinite VC dimension



- Linear combination of fixed basis functions

$$f(\mathbf{x}, \mathbf{w}) = I \left( \sum_{i=1}^m w_i g_i(\mathbf{x}) + w_0 > 0 \right)$$

is equivalent to linear functions in m-dimensional space

→ VC-dimension = m + 1

(this assumes linear independence of basis functions)

- In general, analytic estimation of VC-dimension is hard
- VC-dimension can be
  - equal to DoF
  - larger than DoF
  - smaller than DoF

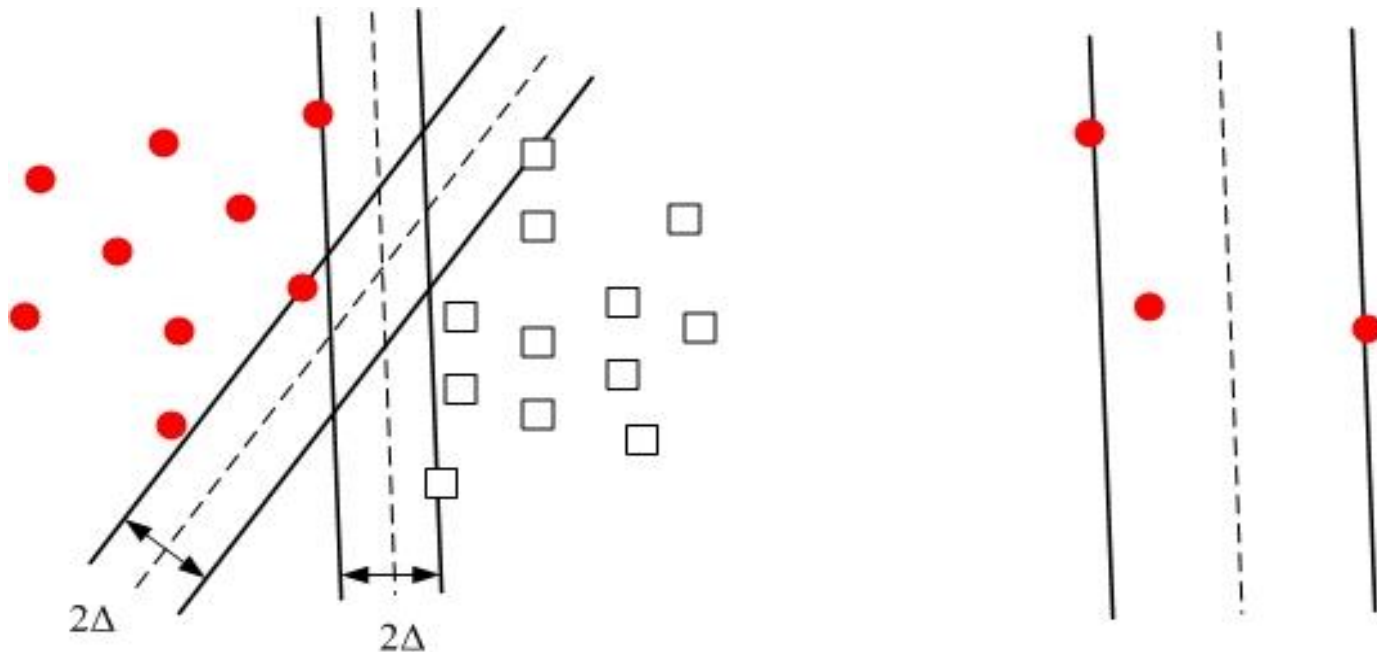


- Delta-margin hyperplanes:

consider linear slabs  $D(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b$  separating samples from two classes, such that the distance btwn  $D(\mathbf{x})$  and the closest data point is larger than some positive value  $\Delta$

later used in svm for model complexity control, e.g higher delta is not that good at separating->overfiring, adjust the delta

- For large  $\Delta$ , VC-dimension **may be smaller** than  $d+1$ :



# VC-dimension and Falsifiability

- A set of functions has VC-dimension  $h$  if
  - (a) It **can explain** (shatter) a set of  $h$  samples  
~ there exists  $h$  samples that **cannot falsify** it  
*and*
  - (b) It **can not shatter**  $h+1$  samples  
~ any  $h+1$  samples **falsify** this set
- *Finiteness of VC-dim* is **necessary and sufficient** condition for generalization  
(for *any* learning method based on ERM)

# Philosophical interpretation: VC-falsifiability

- **Occam's Razor**: Select the model that explains available data *and* has the smallest number of free parameters (entities)
  - **VC theory**: Select the model that explains available data *and* has low VC-dimension (i.e. can be **easily falsified**)
- **New principle of VC falsifiability**

# Generalization Bounds

- **Bounds for learning machines** (implementing ERM) evaluate the difference btwn (unknown) risk and known empirical risk, as a function of sample size  $n$  and the properties of the loss functions (approximating fcts).
- **Classification:** the following bound holds with probability of  $1 - \eta$  for **all approximating functions**

$$R(\omega) < R_{emp}(\omega) + \Phi(R_{emp}(\omega), n/h, -\ln \eta / n)$$

where  $\Phi$  is called the **confidence interval**

- **Regression:** the following bound holds with probability of  $1 - \eta$  for **all approximating functions**

$$R(\omega) < R_{emp}(\omega) / \left(1 - c\sqrt{\varepsilon}\right)_+$$

where 
$$\varepsilon = \varepsilon\left(\frac{n}{h}, \frac{-\ln \eta}{n}\right) = a_1 \frac{h\left(\ln \frac{a_2 n}{h} + 1\right) - \ln(\eta/4)}{n}$$

# COMMENTS ON VC-BOUNDS

- **Useful** for conceptual understanding of general properties & limitations of all learning methods
- **Not appropriate** for practical use
- Properties of VC-bounds:
  - the confidence interval *monotonically* decreases to zero (*with  $n$* )
  - bounds *depend strongly* on  $n/h$ ;
  - bounds *do not depend* on dimensionality

# Practical VC Bound for regression

- **Practical regression bound** can be obtained by setting the confidence level  $\eta = \min(4 / \sqrt{n}, 1)$  and theoretical constants  $c=1$ ,  $a_1=a_2=1$  :

$$R(h) \leq R_{emp}(h) \left( 1 - \sqrt{\frac{h}{n} - \frac{h}{n} \ln \frac{h}{n} + \frac{\ln n}{2n}} \right)_+^{-1}$$

- **Compare** to analytic bounds (SC, FPE) in Lecture Set 2
- **Analysis** (of denominator) shows that

**$h < 0.8 n$  for any estimator**

In practice:

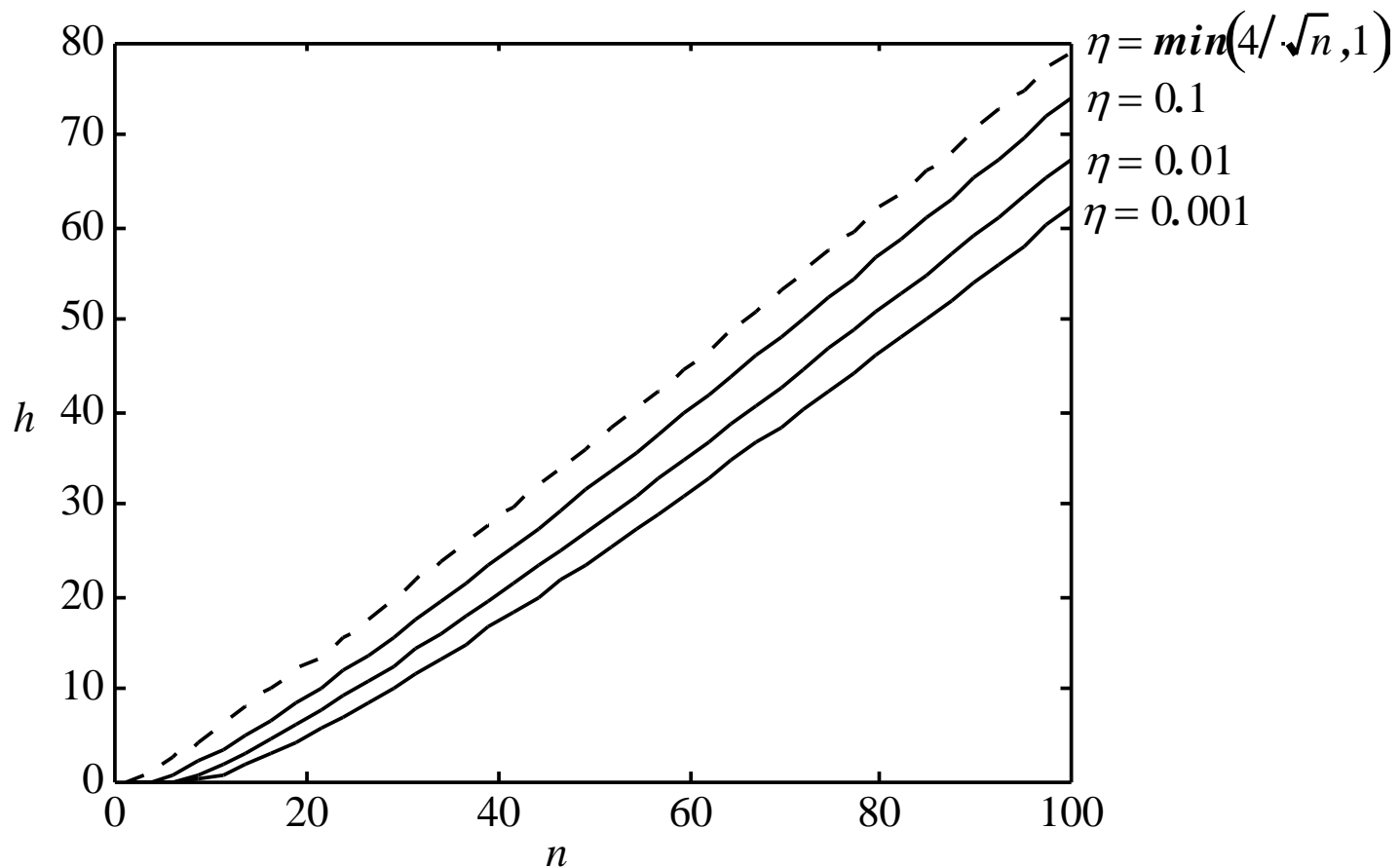
**$h < 0.5 n$  for any estimator**

**NOTE:** generalization bounds *do not depend on dimension!*

# Visual Illustration

- **Max possible  $h$  for given sample size  $n$**

for different confidence level values:



# Structural Risk Minimization

- **Analysis of generalization bounds**

$$R(\omega) < R_{emp}(\omega) + \Phi(R_{emp}(\omega), n/h, -\ln \eta / n)$$

suggests that when  $n/h$  is **large**, the term  $\Phi$  is **small**

$$\rightarrow R(\omega) \sim R_{emp}(\omega)$$

This leads to parametric modeling approach (ERM)

- When  $n/h$  is **not large** (say, less than 20), **both terms** in the right-hand side of VC- bound need to be **minimized**

$\rightarrow$  make the VC-dimension a controlling variable

- SRM = formal mechanism for controlling model complexity

Set of loss functions has **nested structure**  $S_k = \{Q(\mathbf{z}, \omega), \omega \in \Omega_k\}$

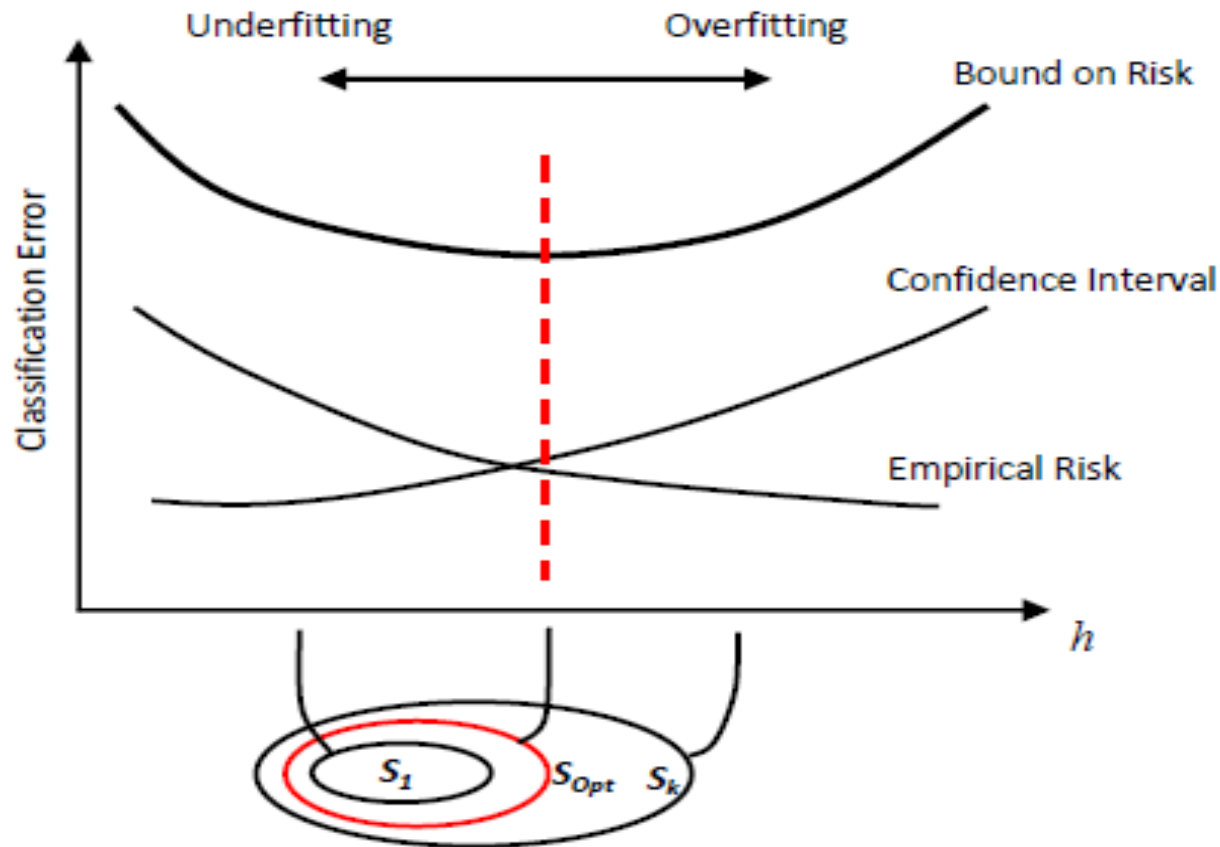
$$S_1 \subset S_2 \subset \dots \subset S_k \subset \dots \text{ such that } h_1 \leq h_2 \leq h_k \leq \dots$$

$\rightarrow$  **structure**  $\sim$  **complexity ordering**



# Structural Risk Minimization

- An upper bound on the true risk and the empirical risk, as a function of VC-dimension  $h$  (for fixed sample size  $n$ )



# Structural Risk Minimization

- Contrast SRM vs ERM approach to data modeling

Given training examples  $(\mathbf{x}, y)$  sampled from unknown  $P(\mathbf{x}, y)$

## Empirical Risk Minimization Modeling

1. Make assumptions about parameterization of admissible decision functions  $f(\mathbf{x}, \omega)$ .
2. For each admissible model, estimate empirical risk (classification error) for the training data.
3. Select the classifier (decision function) providing smallest empirical risk.

## Structural Risk Minimization Modeling

1. Introduce nested structure on a set of functions  $f(\mathbf{x}, \omega)$ .
2. Minimize empirical risk for each element of a structure  $S_k$ .
3. Estimate guaranteed risk for each element  $S_k$  using VC- bounds.
4. Choose the best element  $S_0$  providing smallest risk estimate, and select the function minimizing empirical risk on  $S_0$  as the best model.

# SRM Approach

- Use **VC-dimension** as a controlling parameter for minimizing VC bound:

$$R(\omega) < R_{emp}(\omega) + \Phi(n/h)$$

- **Two general strategies for implementing SRM:**
  1. Keep  $\Phi(n/h)$  fixed and minimize  $R_{emp}(\omega)$   
(most statistical and neural network methods)
  2. Keep  $R_{emp}(\omega)$  fixed and minimize  $\Phi(n/h)$   
(Support Vector Machines)

# Common SRM structures

- **Dictionary structure**

A set of algebraic polynomials  $f_m(x, \mathbf{w}) = \sum_{i=0}^m w_i x^i$   
is a structure since  $f_1 \subset f_2 \subset \dots \subset f_k \subset \dots$

More generally  $f_m(\mathbf{x}, \mathbf{w}, \mathbf{V}) = \sum_{i=0}^m w_i g(\mathbf{x}, \mathbf{v}_i)$   
where  $g(\mathbf{x}, \mathbf{v}_i)$  is a set of basis functions (**dictionary**).

The number of terms (basis functions)  $m$  specifies an element of a structure.

For fixed basis fcts, VC-dim = number of parameters  $w_i$

# Common SRM structures (cont'd)

- **Feature selection** (aka subset selection)

Consider **sparse polynomials**  $f_m(x, \mathbf{w}) = \sum_{i=0}^m w_i x^{k_i}$   
where  $k_i$  is a (positive) integer

Each monomial is a feature  $\rightarrow$  the goal is to select a set of  $m$  features providing min. empirical risk (MSE)

This is a structure since  $f_1 \subset f_2 \subset \dots \subset f_m \subset \dots$

**More generally**, representation  $f_m(\mathbf{x}, \mathbf{w}, \mathbf{V}) = \sum_{i=0}^m w_i g(\mathbf{x}, \mathbf{v}_i)$

where  $m$  basis functions are selected from a large (given) set of  $M$  functions

**Note:** nonlinear optimization, VC-dimension is **unknown**

# Common SRM structures (cont'd)

- **Penalization**

Consider algebraic polynomial of **fixed degree**

$$f(x, \mathbf{w}) = \sum_{i=0}^{10} w_i x^i \quad \text{where} \quad \|\mathbf{w}\|^2 \leq c_k \quad c_1 < c_2 < c_3 \dots$$

For each (positive) value  $c$  this set of functions specifies an element of a structure  $S_k = \{ f(\mathbf{x}, \mathbf{w}), \|\mathbf{w}\|^2 \leq c_k \}$

Minimization of empirical risk (MSE) on each element  $S_k$  of a structure is a **constrained minimization problem**

This optimization problem can be equivalently stated as minimization of the **penalized empirical risk functional**:

$$R_{pen}(\omega, \lambda_k) = R_{emp}(\omega) + \lambda_k \|\mathbf{w}\|^2 \quad \text{where the choice of } \lambda_k \sim c_k$$

**Note:** VC-dimension is unknown

# Common SRM structures (cont'd)

- **Initialization structure**

The structure is defined for nonlinear optimization algorithm A fitting training data using a set of functions  $f(\mathbf{x}, \mathbf{w})$  with initial conditions  $\mathbf{w}_0$

$S_k = [A: f(\mathbf{x}, \mathbf{w}), \|\mathbf{w}_0\| < C_k]$  where  $C_1 < C_2 < \dots$

- **Early stopping rules**

The structure is defined for nonlinear optimization algorithm A fitting training data. The structure is index by the *number of iterative steps* of algorithm A, starting with initial (small) values of parameters  $\mathbf{w}_0$

# Common SRM structures (cont'd)

- **Margin-based structure**

Recall large-margin separating hyperplanes for classification.

*Larger-margin* hyperplanes form a subset of all *smaller-margin* hyperplanes:

$$S_{\Delta_1} \subset S_{\Delta_2} \subset \dots \subset S_{\Delta_k} \subset \dots$$

for  $\Delta_1 > \Delta_2 > \dots$

→ VC-dimension controlled by margin size  $\Delta$



# Example of SRM for Regression

- Polynomial regression using different structures

Regression problem:  $y = 0.8\sin(2\pi\sqrt{x}) + 0.2x^2 - 0.5\sqrt{x} + \xi$

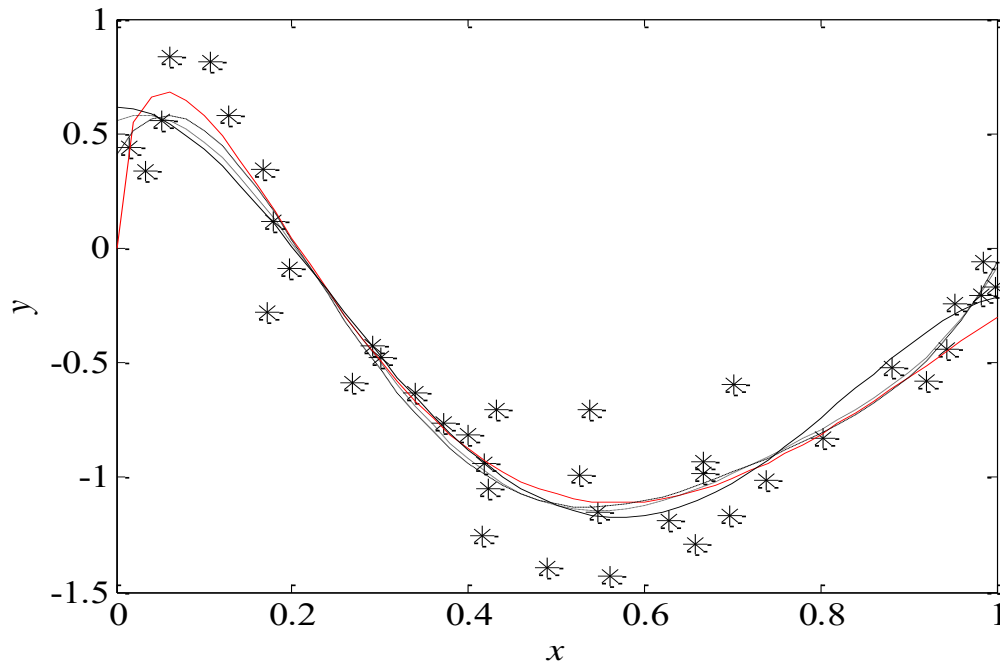
x-values uniform on  $[0,1]$ , Gaussian noise  $\sigma = 0.05$

training set  $\sim 40$  samples; validation  $\sim 40$  samples

- Different structures

- dictionary (degrees 1 to 10)
- penalization (degree 10 with ridge penalty)
- sparse polynomials (of degree 1 to 6)

- Dictionary  $\hat{y} = 0.4078 + 6.4198x - 68.2162x^2 + 163.7679x^3 - 158.3952x^4 + 55.9565x^5$
- Penalization  $\lambda = 1.013e-005$
- Feature selection  $\hat{y} = 0.6186 - 22.7337x^2 + 41.1772x^3 - 19.2736x^4$



target function ~ **red line**; dictionary ~ black dashed;  
 penalization ~ black dotted; feature selection ~ black solid

# Aside: model interpretation

- **Outside the scope of VC-theory**
- **In the last example:**  
**different structures** on the *same set of functions*  
lead to **different interpretations**
- Often interpretation assumes function identification setting
- In practice, interpretation always requires application domain knowledge

# Practical Implementation of SRM

- **Need to**
  - estimate VC-dim for an element of a structure  $S_k$
  - minimize empirical risk for each element  $S_k$
- **Both possible** for linear approximating functions
- **Both difficult** for nonlinear parameterization
  - many heuristic learning algorithms

# SRM structures: summary

- **SRM structure** ~ complexity ordering on a set of admissible models (approximating functions)
- **Many different structures on the same set of approximating functions**

Which one is the 'best'?

- depends on the properties of application data
- can be decided via empirical comparisons
- **SRM = mechanism for complexity control**
  - selecting optimal complexity for a given data set
  - new measure of complexity: **VC-dimension**
  - model selection using **analytic VC-bounds**

# OUTLINE

- Objectives
- Inductive learning problem setting
- Statistical Learning Theory
- **Applications**
  - **model selection (for regression)**
  - **market timing of international funds**
  - **signal denoising**
- Measuring the VC-dimension
- Summary and discussion

# Application: VC-based model selection

see <http://www.mitpressjournals.org/toc/neco/15/7>

- Standard regression setting  $y = g(\mathbf{x}) + \textit{noise}$
- Statistical criteria
  - Akaike Information Criterion (**AIC**)

$$AIC(d) = R_{emp}(d) + \frac{2d}{n} \hat{\sigma}^2$$

- Bayesian Information Criterion (**BIC**)

$$BIC(d) = R_{emp}(d) + (\ln n) \frac{d}{n} \hat{\sigma}^2$$

where  $d \sim$  (effective) DoF,  $n \sim$  sample size

- Require noise estimation  $\hat{\sigma}^2 = \frac{n}{n-d} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$

- Many methods require **noise estimation**

$$\hat{\sigma}^2 = \frac{n}{n-h} \cdot \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{n}{n-h} R_{emp}$$

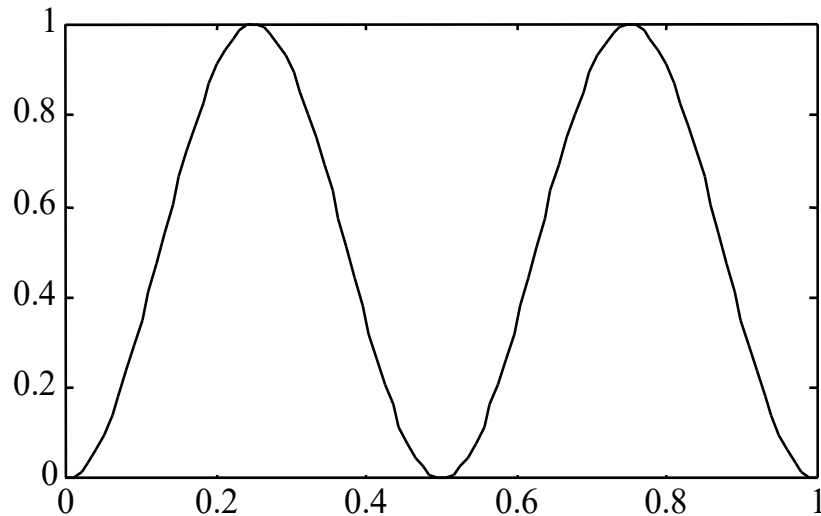
- One approach is to estimate noise for each (fixed) model complexity  $\rightarrow$  **multiplicative criteria**
- Another approach is to estimate noise first and then use it in the **additive AIC or BIC criteria**
- This study uses **additive AIC/BIC** assuming **known noise variance**
- VC-based model selection (aka VM)**

$$R(h) \leq R_{emp}(h) \left( 1 - \sqrt{\frac{h}{n} - \frac{h}{n} \ln \frac{h}{n} + \frac{\ln n}{2n}} \right)_+^{-1}$$

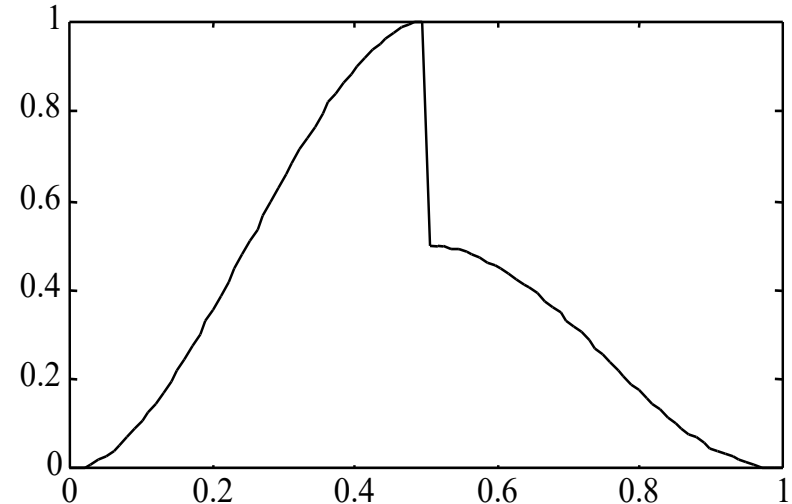


# Comparison for univariate regression

(a) Sine squared function  $\sin^2(2\pi x)$

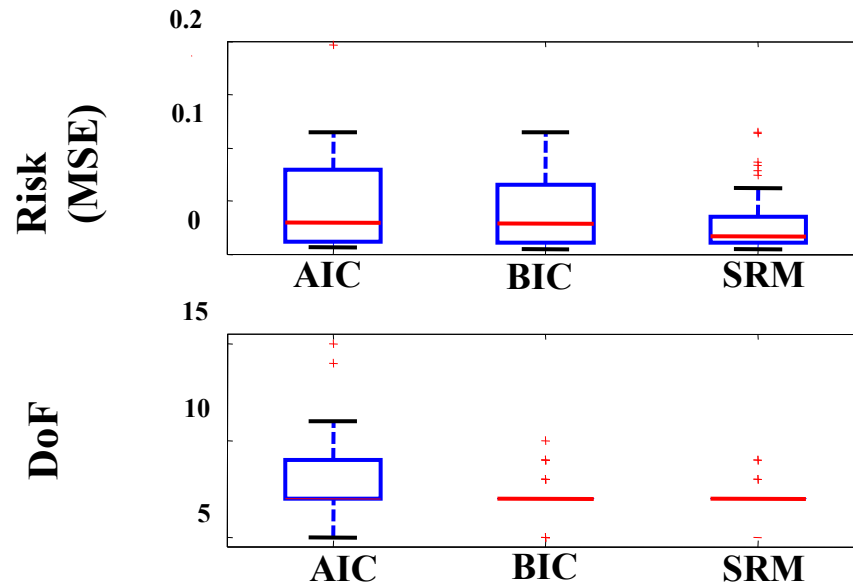


(b) Piecewise polynomial



- Target functions: continuous + discontinuous
- Approximating functions: algebraic polynomials  
Fourier basis

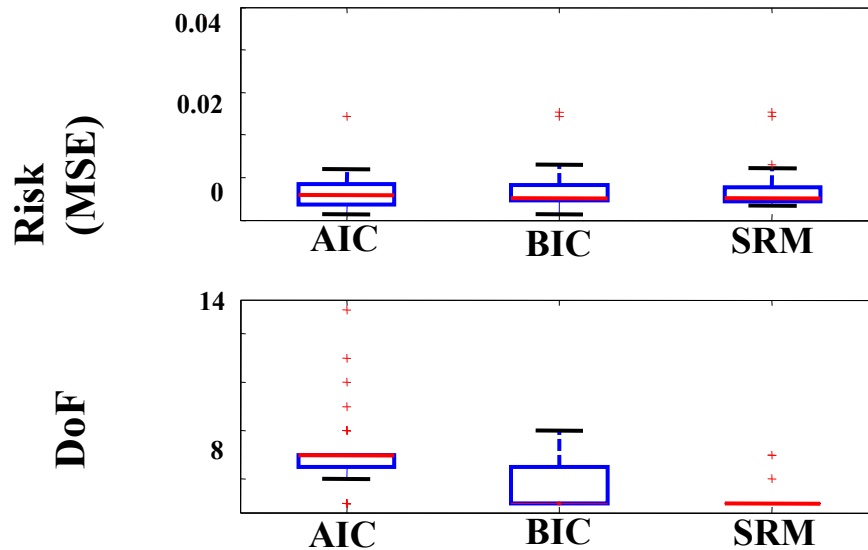
Sine-squared target fct, polynomial regression  
sample size  $n=30$ , noise  $\sigma = 0.2$



Comparison of AIC, BIC & SRM (or VM):

- prediction risk (MSE)
- selected DoF ( $\sim h$ )

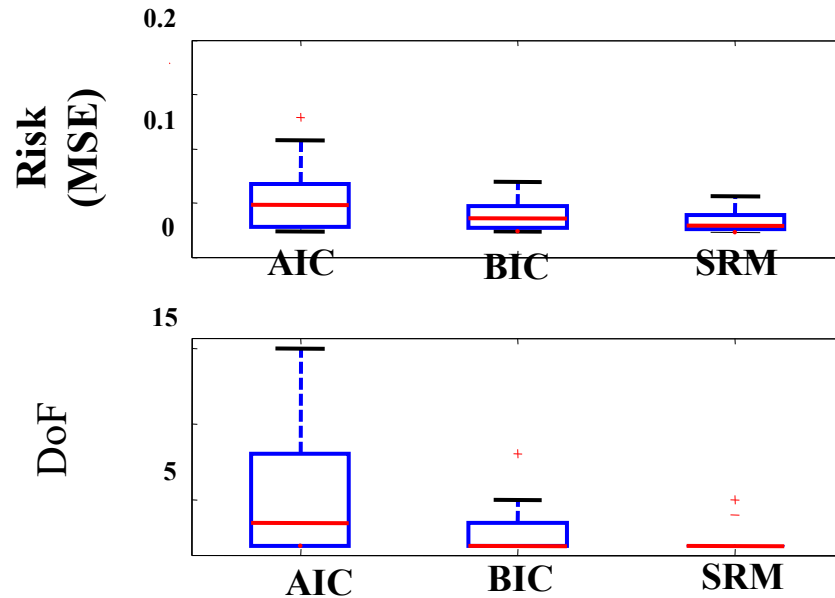
# Sine-squared target fct, polynomial regression sample size $n=100$ , noise $\sigma = 0.2$



Better model selection approaches select models providing **lowest guaranteed prediction risk** (i.e. with lowest risk at the 95 percent mark) and also **smallest variation of the risk** (i.e., narrow box plots).

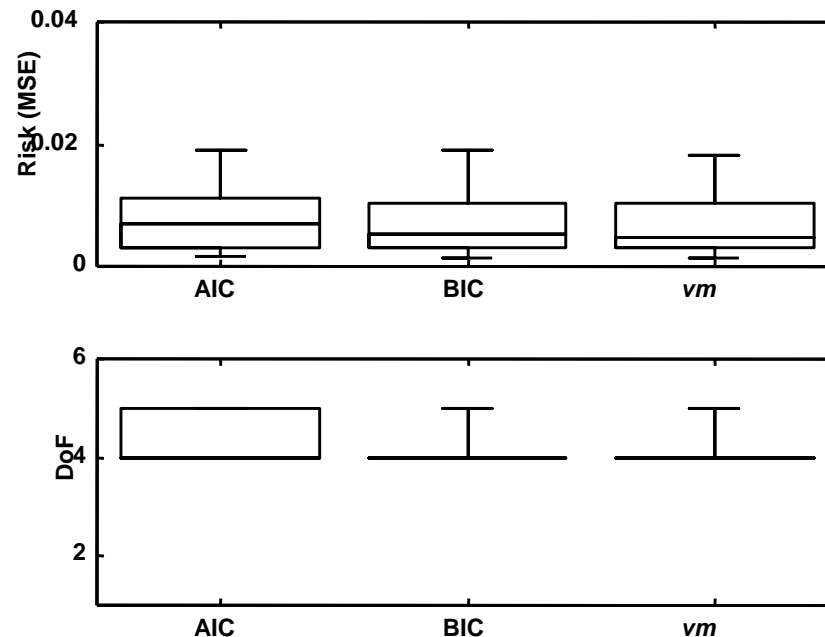
# Piecewise polynomial, Fourier estimation

sample size  $n=30$ , noise  $\sigma = 0.2$



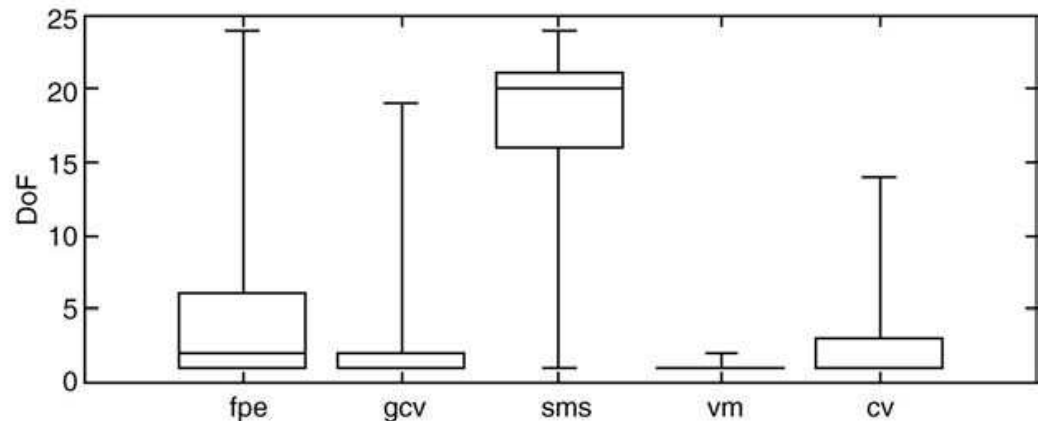
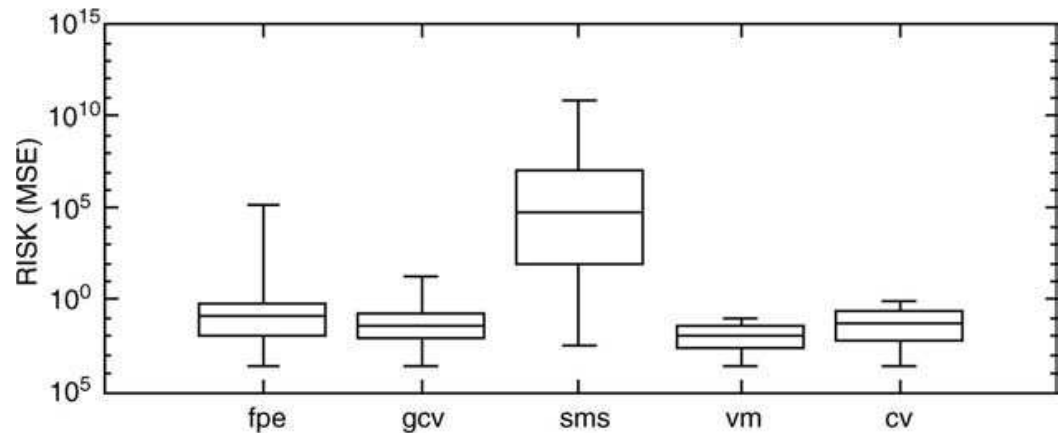
# Comparison for linear subset selection

- Target function:  $y = x_1 + 2x_2 + x_3 + 0 * x_4 + 0 * x_5 + \xi$
- x-values uniform in  $[0,1]$ ,  $n=30$  samples  
gaussian noise  $\sigma = 0.2$
- Approximating functions: **linear subset selection**



# Unknown target function with pure Gaussian noise.

- Model selection results for sample size 30.
- Using algebraic polynomial estimators.
- The true model is the mean of training samples (DoF = 1).
- VC based model selection typically selects the “correct” model (DoF = 1).



# Lots of confusion about model selection

- Cherkassky et al (1999): **VC-analytic bound works very well for (univariate) regression problems**

Hastie et al (2001): **SRM (using VC-bound) performs poorly overall**

See <http://www.mitpressjournals.org/toc/neco/15/7>

- What is the cause for disagreement?
  - technical sloppiness + lack of common sense
- More confusing studies are generated each year

See <http://www.springerlink.com/content/kq4g614j3xhdupuu/fulltext.pdf>

Discussion question: read the paper Sugiyama & Ogawa(2002) and try to understand the deficiencies in their experimental procedure and/or assumptions used in this paper.

# Market Timing of International Funds: A Decade after the Scandal

Vladimir Cherkassky and Sauprik Dhar  
University of Minnesota  
[cherk001@umn.edu](mailto:cherk001@umn.edu)

Presented at CIFEr, March 30, 2012



# OUTLINE

- **Motivation + Background**
  - mutual fund basics
  - scandals in early 2000's
  - regulations on frequent trading
- Predictive (VC-theoretical) methodology
- Empirical Results: market timing of TWIEX
- Conclusions and policy implications

# Timing of International Funds

- **International mutual funds**
  - priced at 4 pm EST
  - reflect price of foreign securities traded at European/ Asian markets
  - Foreign markets close earlier than US market
- **Possibility of inefficient pricing**
- Market timing exploits this inefficiency.**
- **Logical solution:** implement efficient pricing [ Zitzewitz 2003]
- **Solution adopted:** restrictions on trading

# Data Analytic Approach

- **Timing of international mutual funds**

Can it be consistently profitable under

- **past market conditions ?** (2004 ~ 2005)
- **current market conditions ?** (2009 ~ 2012)

- **Predictive data modeling:**

- estimate trading model (using past data)
- apply this model for prediction (trading)

- **Diversified international fund** (TWIEX)

American Century Int'l Growth Fund

# Binary Classification Setting

- TWIEX ~ American Century Int'l Growth
- **Input indicators** (for trading) ~ **today**
  - SP 500 index (daily % change) ~  $x_1$
  - Euro-to-dollar exchange rate (% change) ~  $x_2$
- **Output**: TWIEX NAV (% change) ~ **next day**
- **Model parameterization** (fixed):
  - linear  $g(\mathbf{x}, \mathbf{w}) = w_1 x_1 + w_2 x_2 + w_0$
  - quadratic  $g(\mathbf{x}, \mathbf{w}) = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + w_0$
- **Decision rule** (estimated from training data):
$$D(\mathbf{x}) = \text{Sign}(g(\mathbf{x}, \mathbf{w}^*))$$

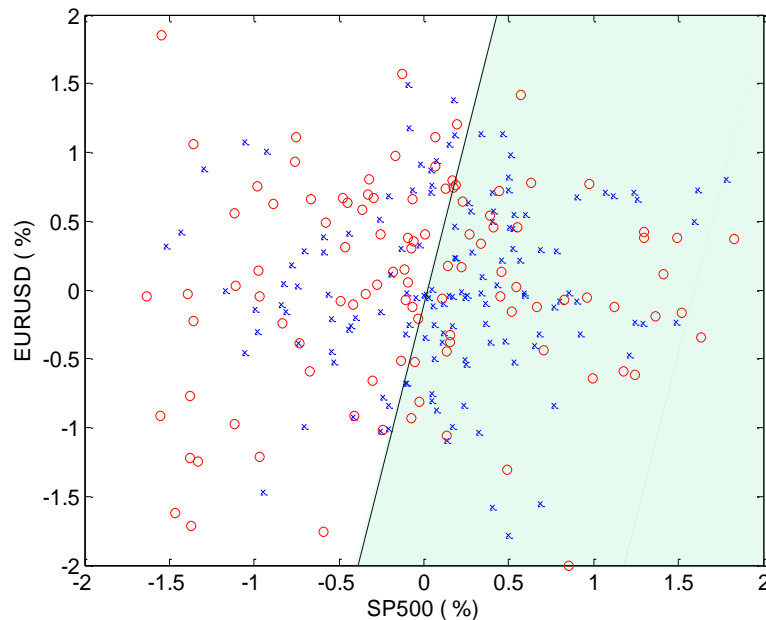
# VC theoretical Methodology

- **When a trained model can predict well?**
  - (1) Future/test data is similar to training data**  
i.e., use 2004 period for training, and 2005 for testing
  - (2) Estimated model is ‘simple’ *and* provides good performance during training period**  
i.e., trading strategy is *consistently better* than *buy-and-hold* during training period

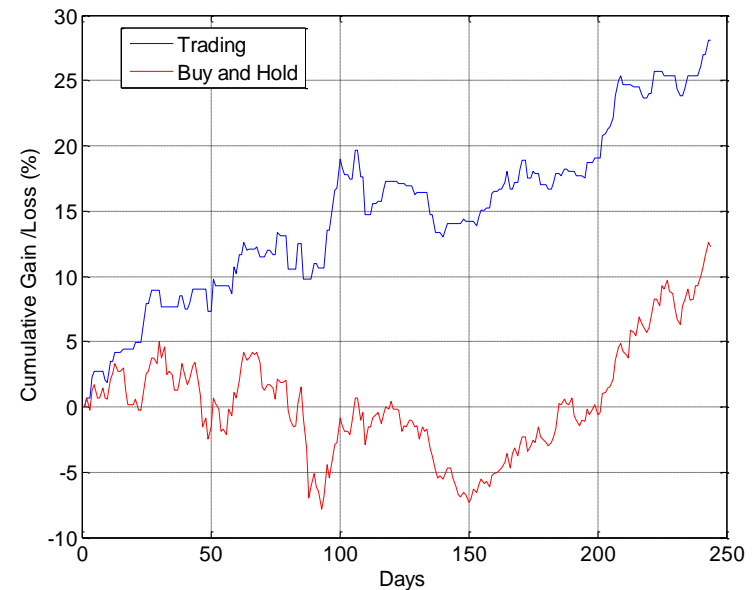
# Empirical Results: 2004 -2005 data

## *Linear model*

Training data 2004



Training period 2004

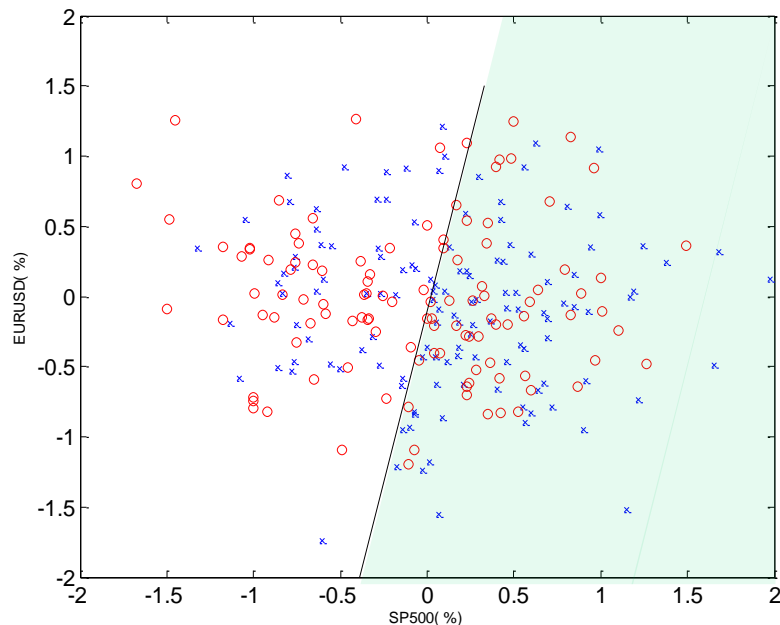


→ can expect good performance with test data

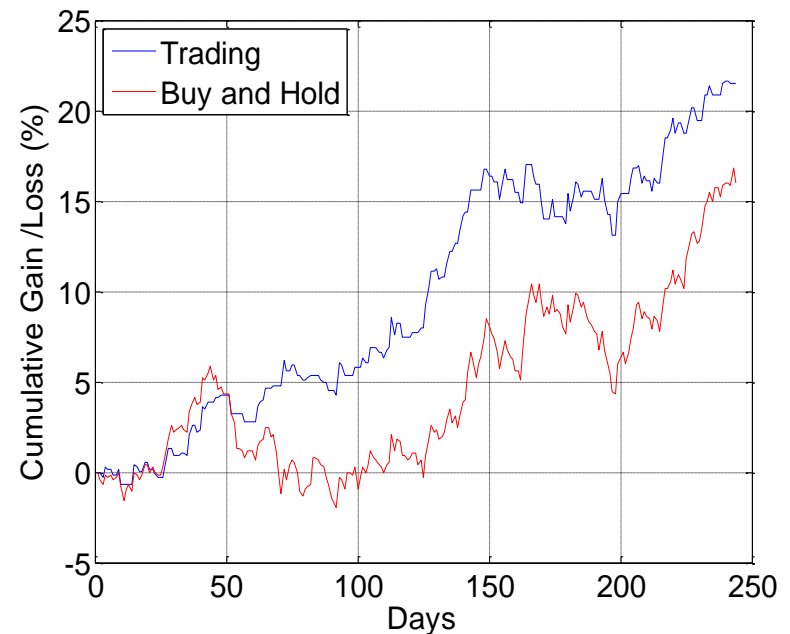
# Empirical Results: 2004 -2005 data

## *Linear model*

Test data 2005



Test period 2005



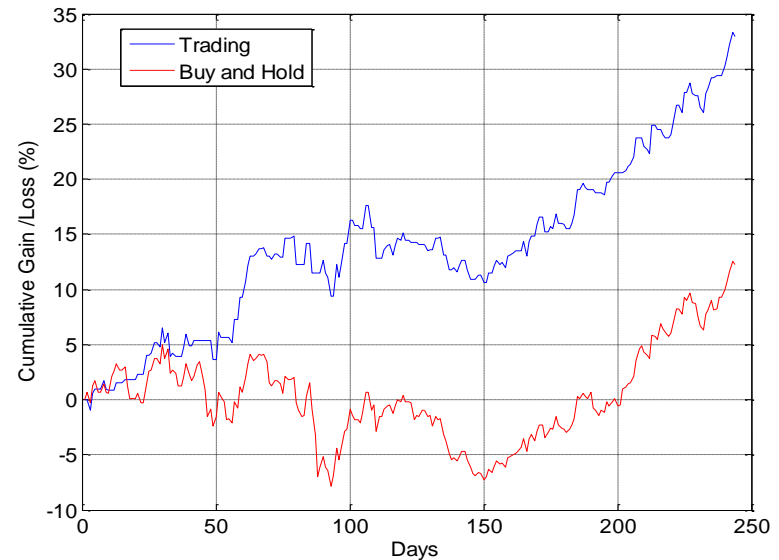
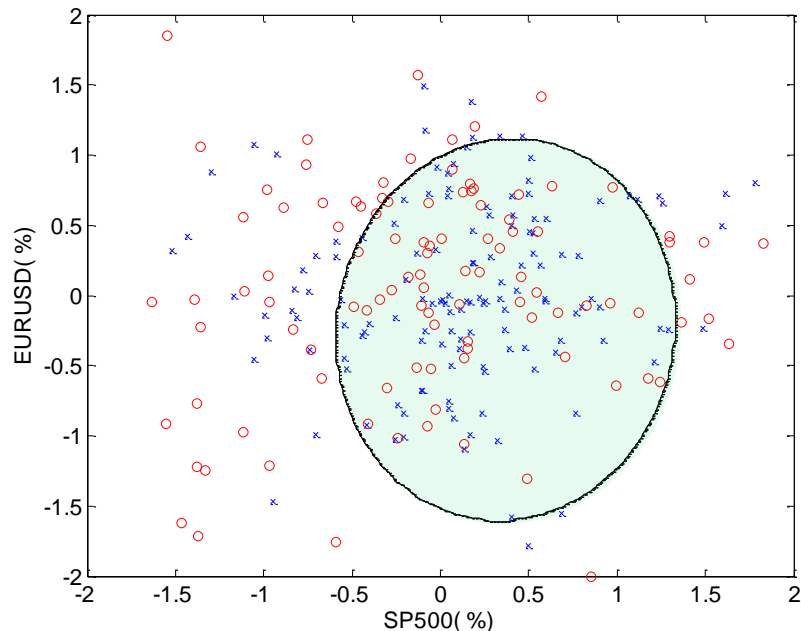
confirmed good prediction performance

# Empirical Results: 2004 -2005 data

## *Quadratic model*

Training data 2004

Training period 2004



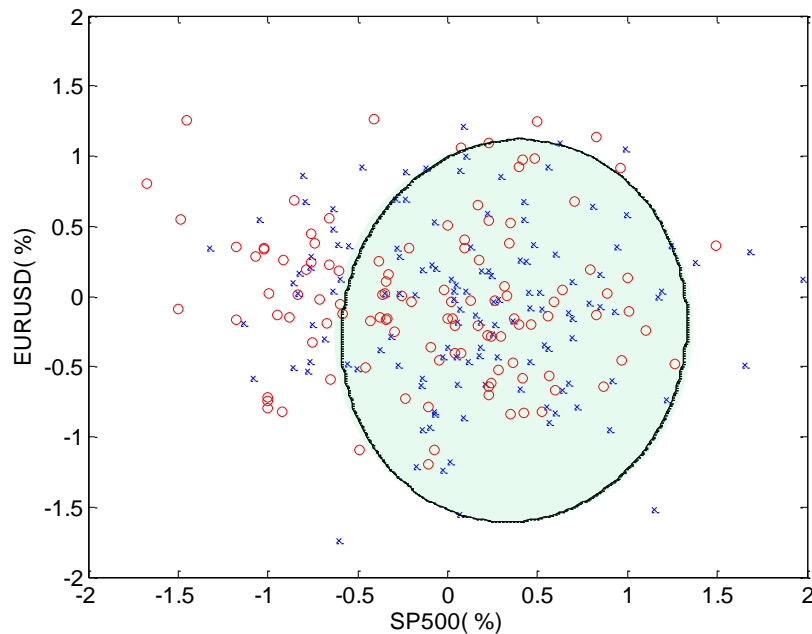
→ can expect good performance with test data



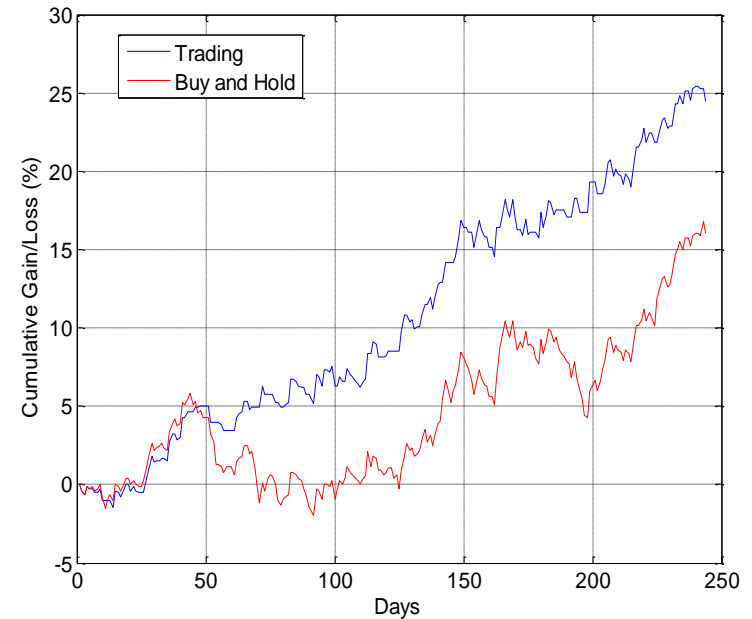
# Empirical Results: 2004 -2005 data

## *Quadratic model*

Test data 2005



Test period 2005

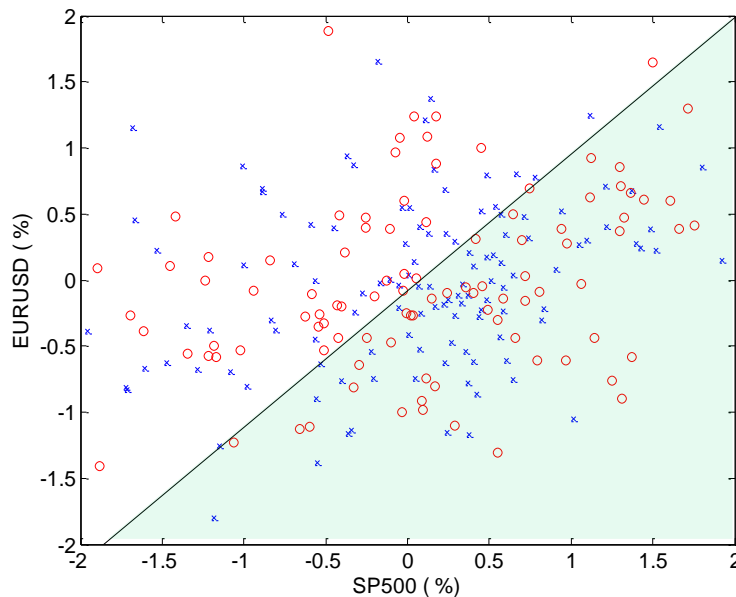


confirmed good test performance

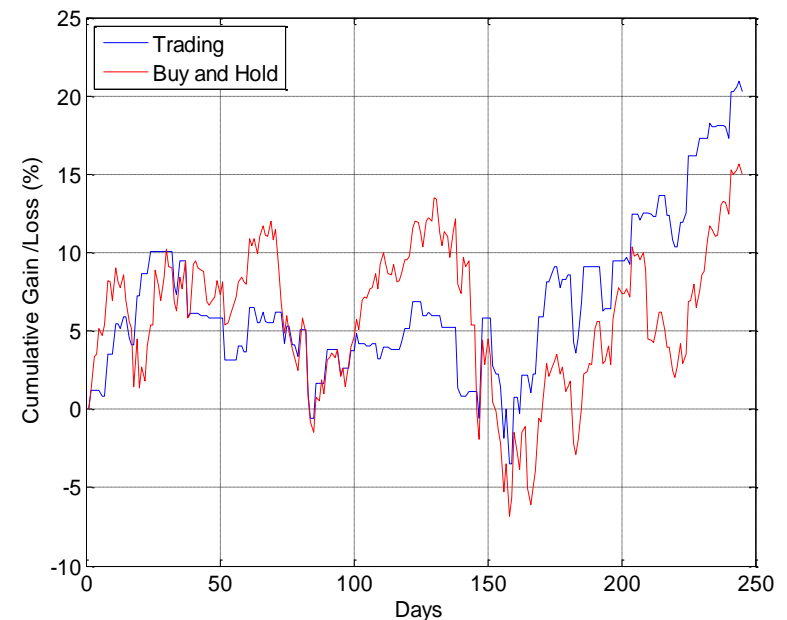
# Empirical Results: 2010 -2011 data

## *Linear model*

Training data 2010



Training period 2010

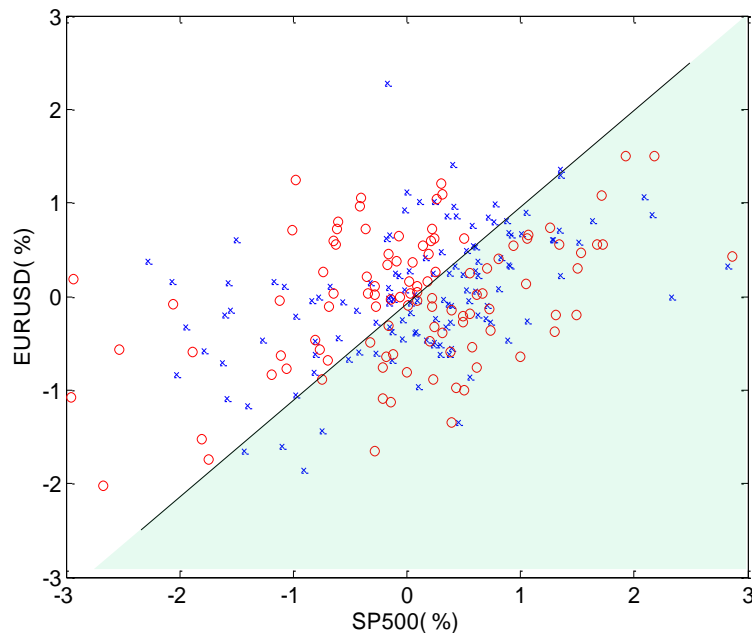


→ performance is *no better* than buy-and-hold

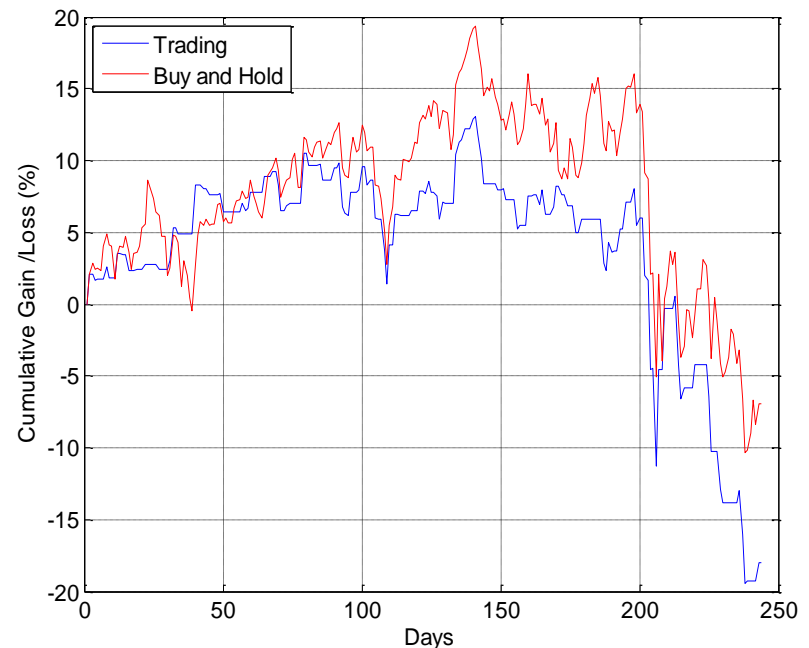
# Empirical Results: 2010 -2011 data

## *Linear model*

Test data 2011



Test period 2011

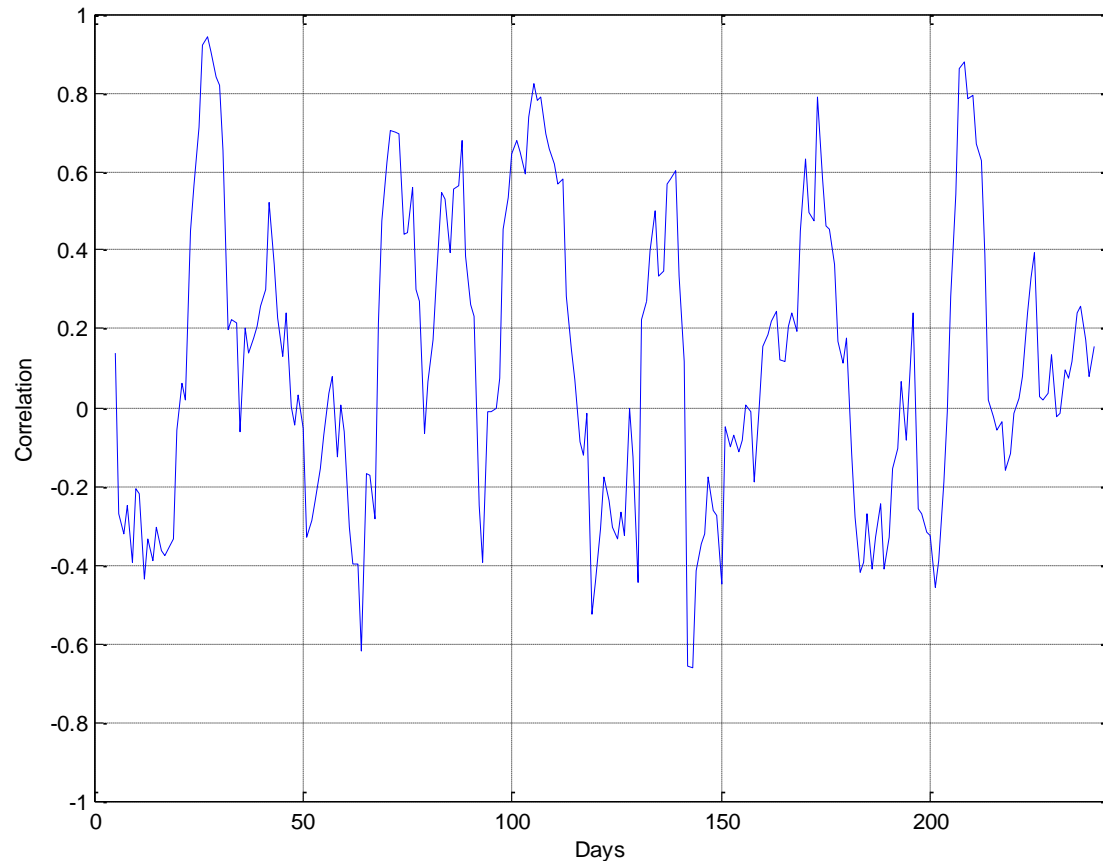


Confirmed poor test performance

# Summary of Results for TWIEX

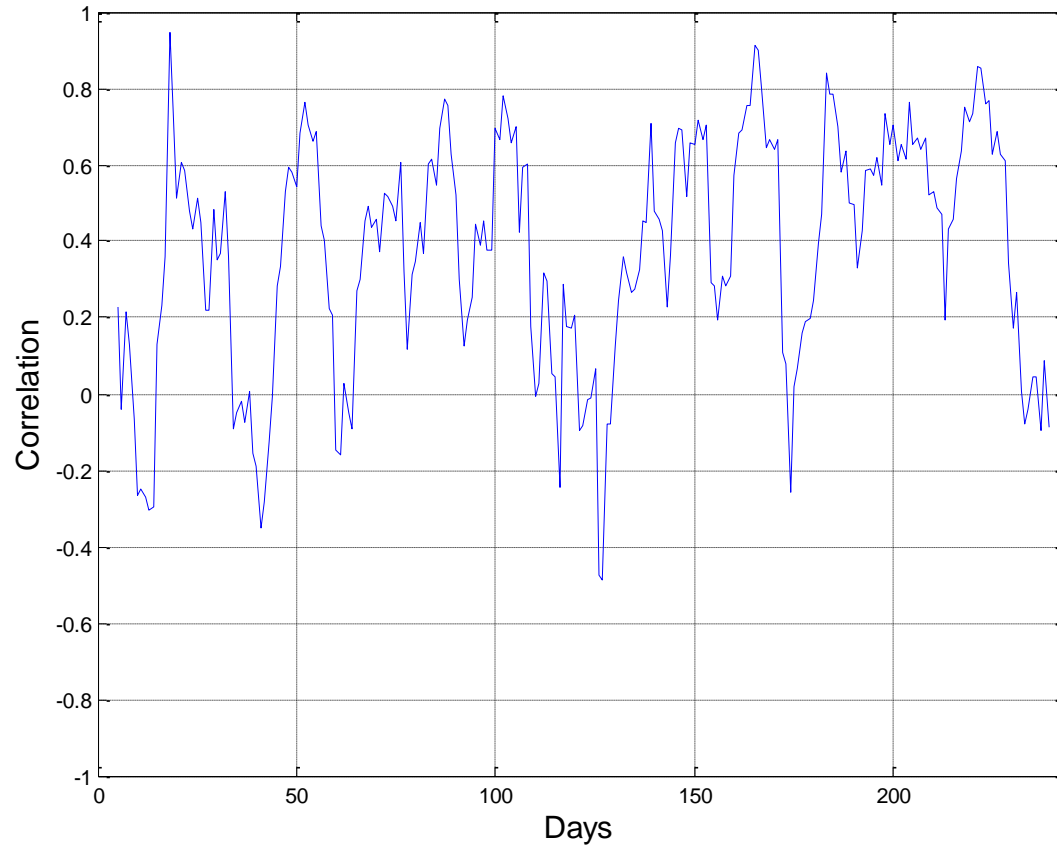
- *Market timing worked well* during past market conditions (i.e., 2004 – 2005 period)
- *Market timing does not work* under current market conditions (i.e., 2008 – 2012 period)
- Similar conclusions hold for
  - other international mutual funds
  - other time periods (say prior to 2004)
- **Explanation:** statistical characteristics of the stock market *have changed*

# Correlation SP500-EURUSD year 2004 = 0.090



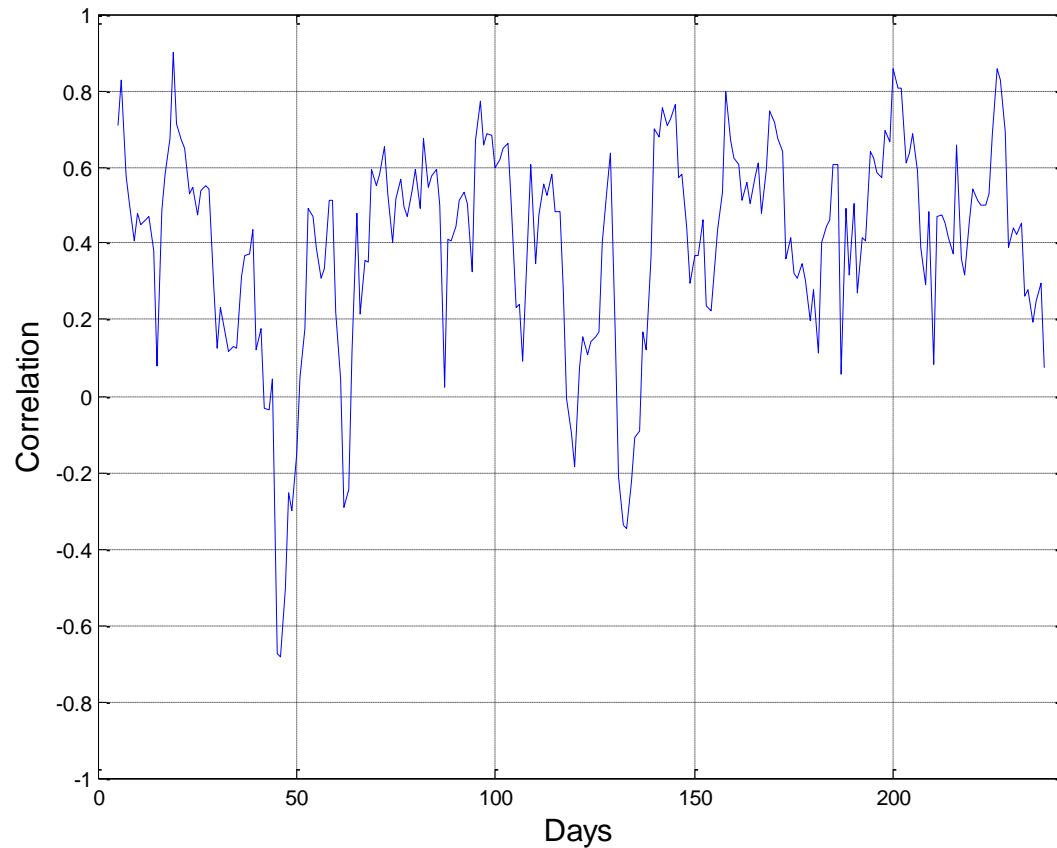
Correlation trend between SP500 and EURUSD for 2004 using a window of 9 days

# Correlation SP500-EURUSD year 2010=0.394



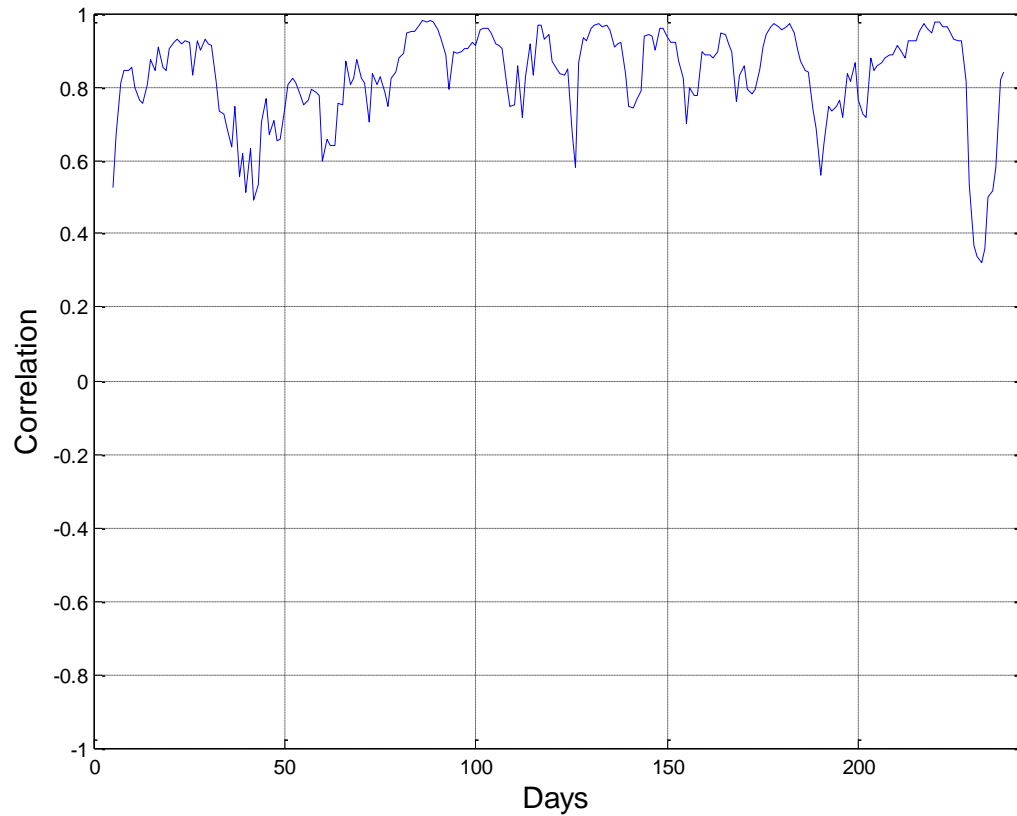
Correlation trend between SP500 and EURUSD for 2010 using a window of 9 days

# Correlation SP500-TWIEX year 2005 = 0.429



Correlation trend between SP500 and TWIEX for 2005 using a window of 9 days

# Correlation SP500-TWIEX year 2010 = 0.877



Correlation trend between SP500 and TWIEX for 2010 using a window of 9 days



# Some possible interpretations

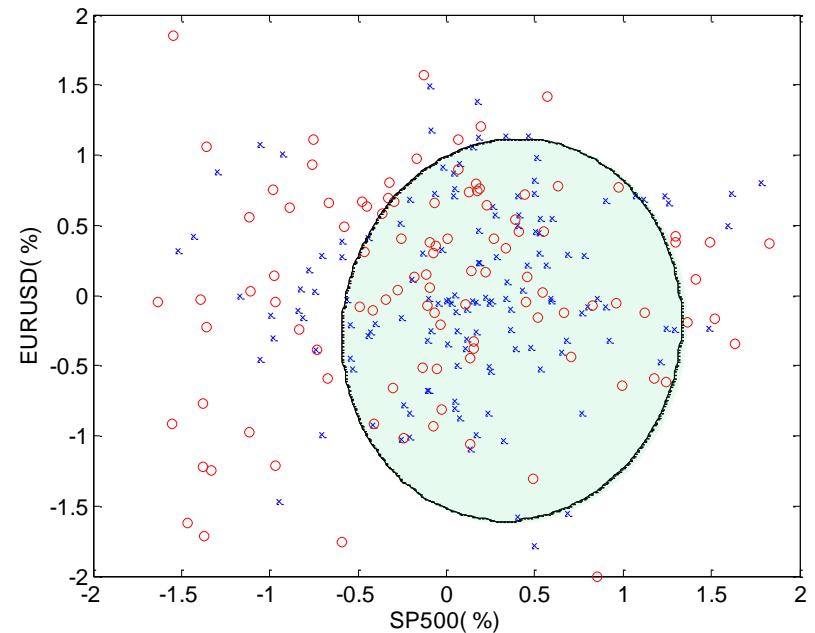
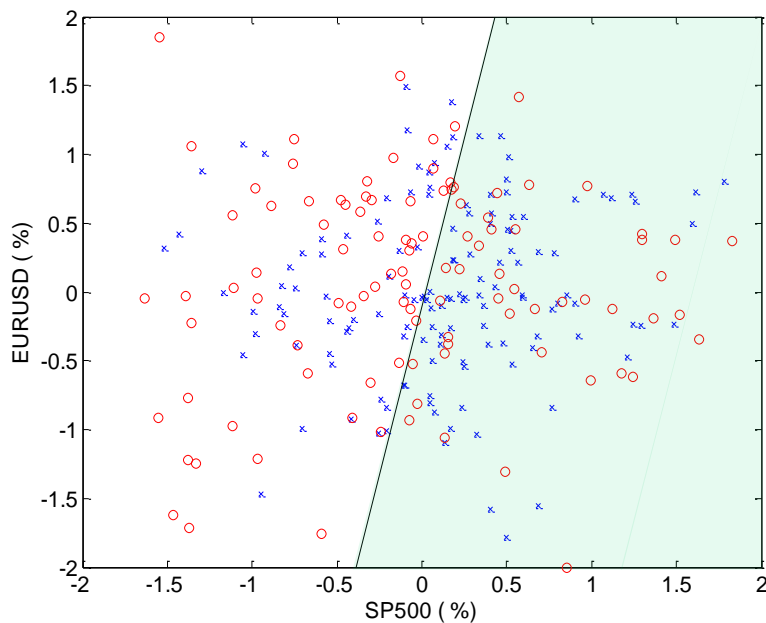
- International markets are more tightly linked (correlated) to US stock market, due to globalization and electronic trading.
- During 2009 – 2011 period, it is the US stock market that follows European markets (and not vice versa). – a large correlation btw SP500 & TWIEX.
- The procedure for calculating the daily NAV value of TWIEX has changed, in order to reflect more accurately the daily changes of the US stock market.

# Aside: Interpretation vs Prediction

- Interpretation: outside the scope of VC-theory
- Only prediction can be *objectively evaluated*
- **Multiplicity of good predictive models**, which reflect different aspects of the data
- *Which model is true?*
- Model interpretation should reflect application-domain knowledge, rather than data-analytic modeling alone

# Interpretation vs Prediction

- Two good trading strategies estimated from 2004 training data

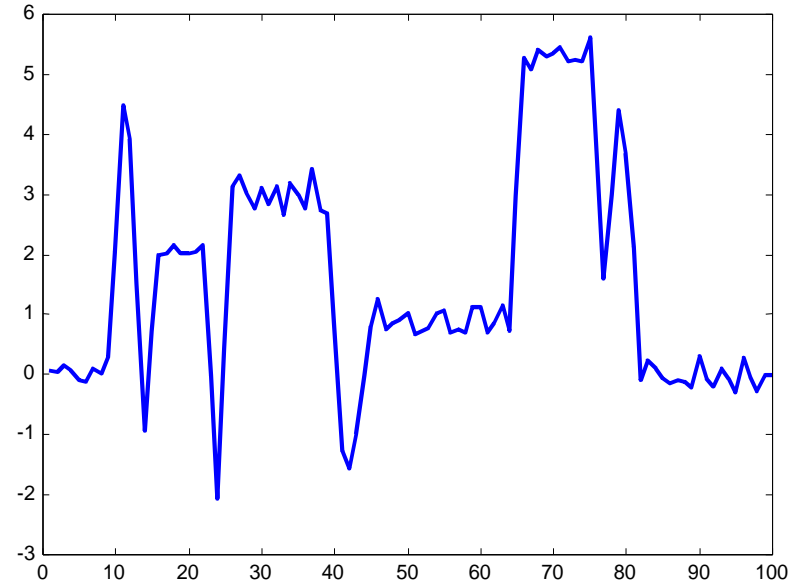
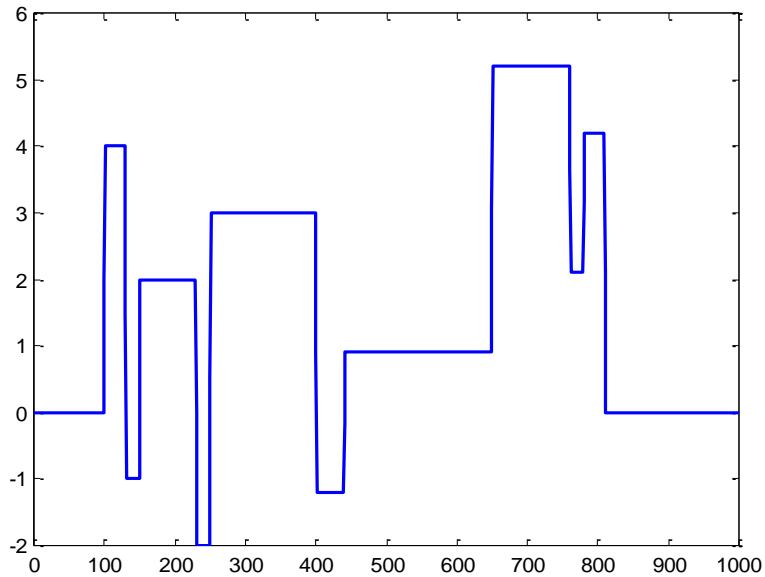


- Both models predict well for test period 2005
- Which model is true?*

# Conclusions and Policy Implications

- **Market timing of international funds**
  - has been indeed profitable **in the past**
  - **does not work** under present market conditions
- **Restrictions on frequent trading**
  - reflect past market conditions
  - constrains risk management by small investors
- **Philosophical/policy question:**  
can these trading restrictions be really justified?

# Application: signal denoising



# Signal denoising problem statement

- **Regression formulation** ~ real-valued function estimation (with squared loss)
- **Signal representation**: linear combination of orthogonal basis functions (harmonic, wavelets)

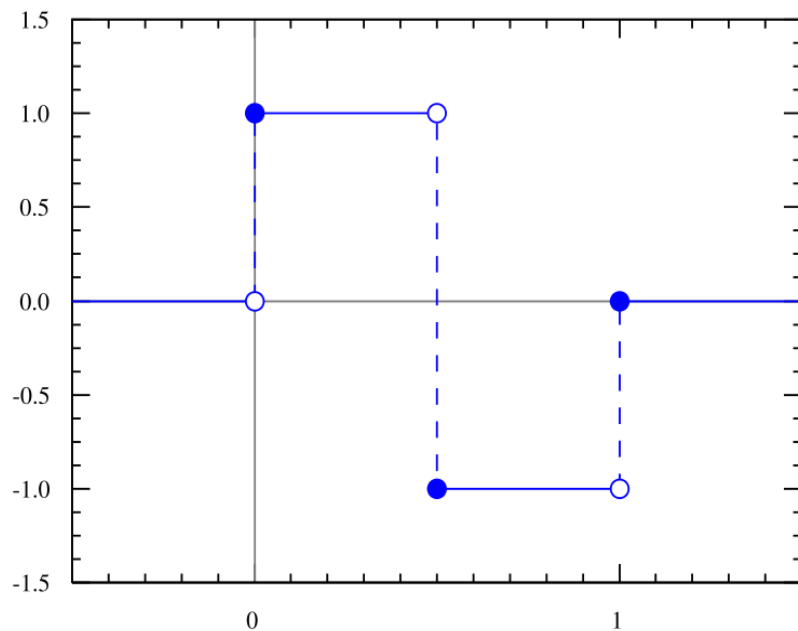
$$y = \sum_i w_i g_i(x)$$

- **Differences** (from standard formulation)
    - fixed sampling rate
    - training data x-values = test data x-values
- Computationally efficient orthogonal estimators:  
Discrete Fourier/Wavelet Transform (DFT / DWT)

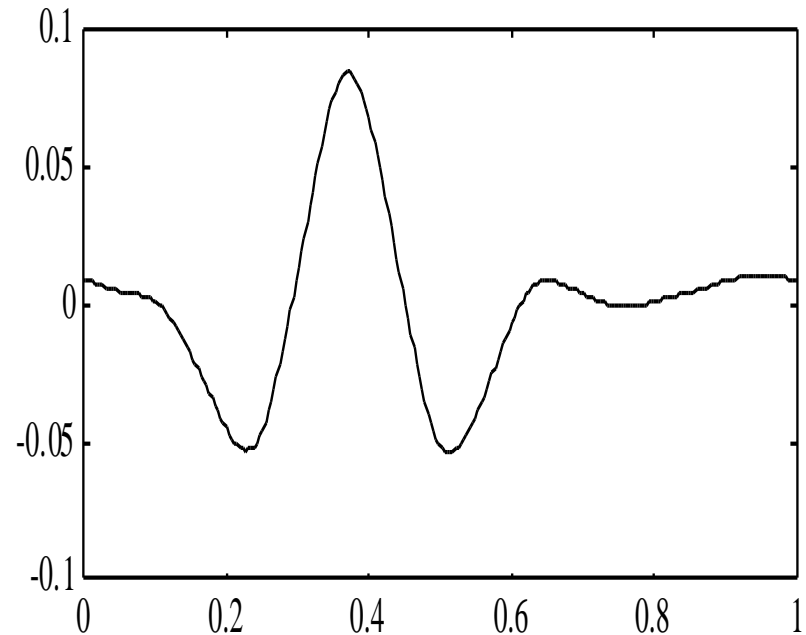
# Examples of wavelets

see <http://en.wikipedia.org/wiki/Wavelet>

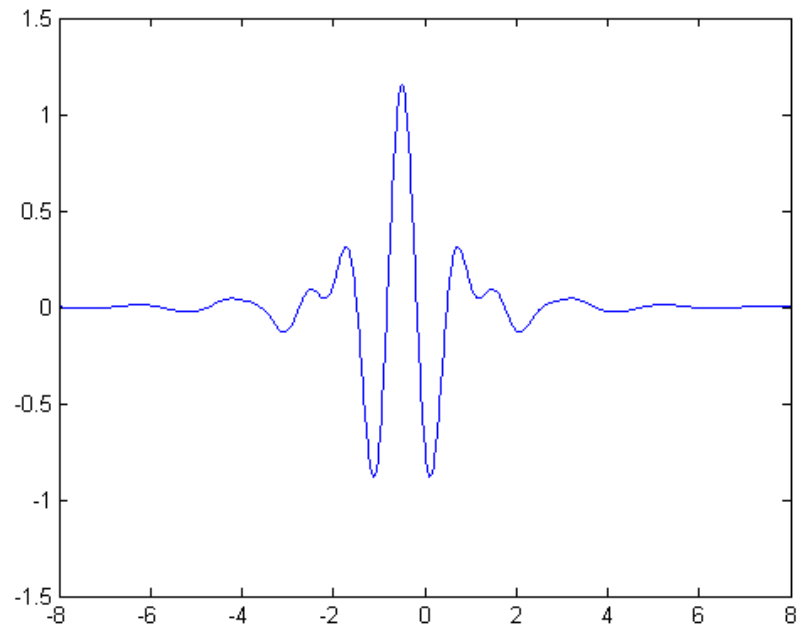
## Haar wavelet



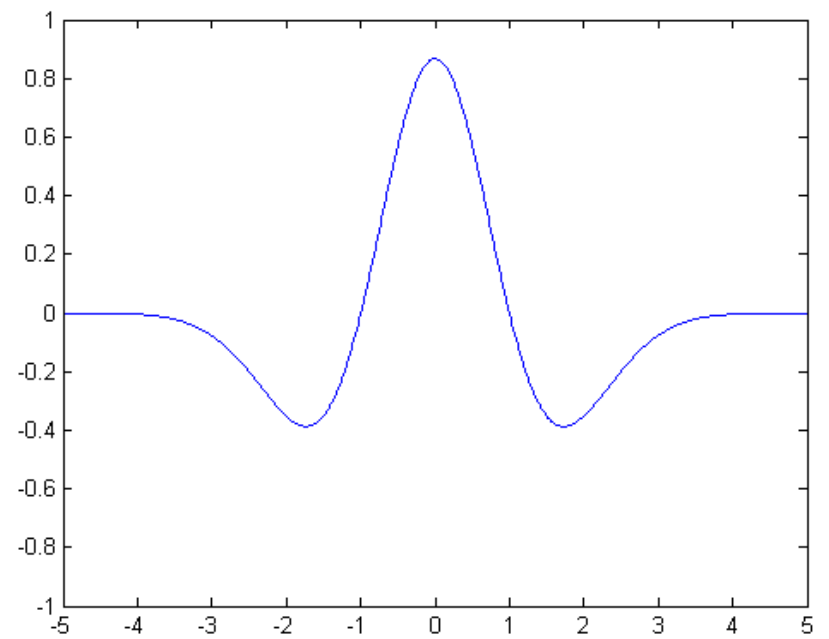
## Symmlet



# Meyer



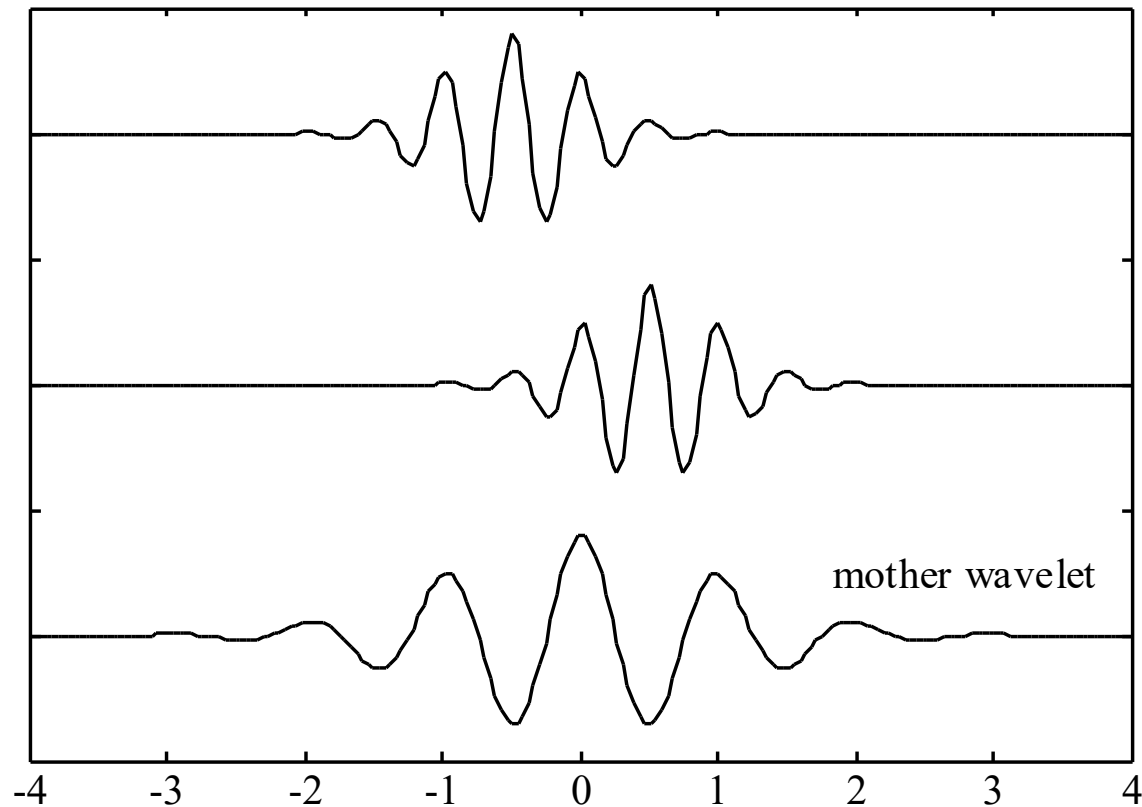
# Mexican Hat





# Wavelets (cont'd)

Example of **translated and dilated** wavelet basis functions:



# Issues for signal denoising

- **Denoising via (wavelet) thresholding**
  - wavelet thresholding = **sparse feature selection**
  - **nonlinear estimator** suitable for ERM
- **Main factors for signal denoising**  $y = \sum w_i g_i(x)$ 
  - Representation** (choice of basis functions)
  - Ordering (of basis functions)** ~ SRM structure
  - Thresholding** (~ model selection)
- **Large-sample setting**: representation
- **Finite-sample setting**: thresholding + ordering

# VC framework for signal denoising

- **Ordering of (wavelet) thresholding** =  
= structure on orthogonal basis functions

Traditional ordering  $|w_{k1}| \geq |w_{k2}| \geq \dots |w_{km}| \geq \dots$

Better ordering  $\frac{|w_{k1}|}{freq_{k1}} \geq \frac{|w_{k2}|}{freq_{k2}} \geq \dots \frac{|w_{km}|}{freq_{km}} \geq \dots$

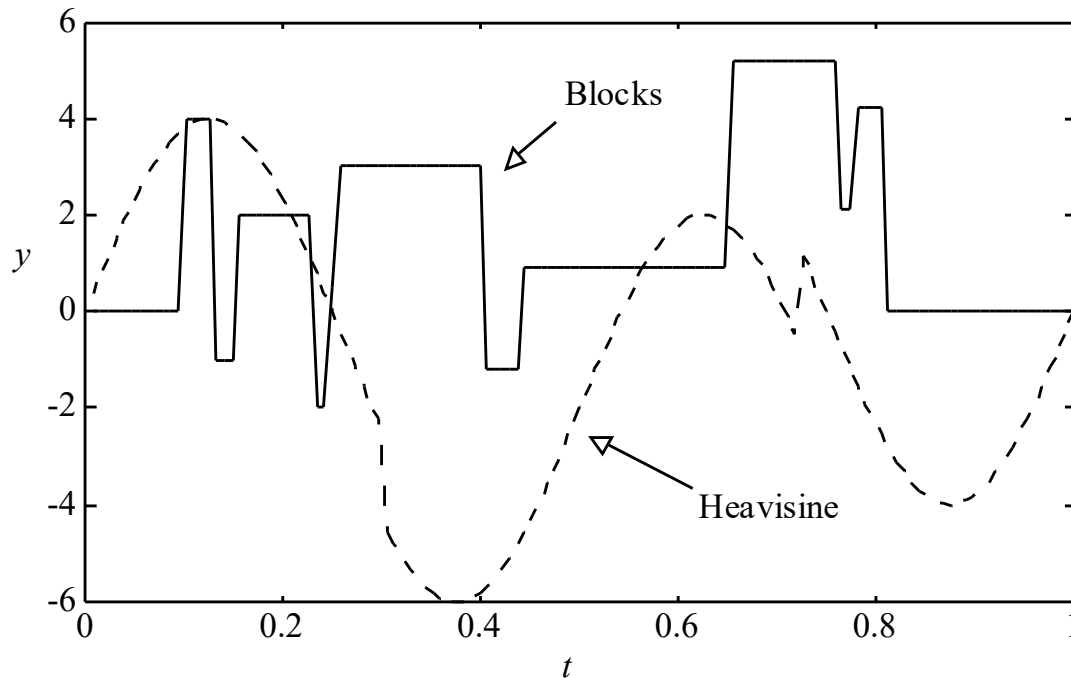
- **VC- thresholding**

Opt number of wavelets  $\sim$  min of VC-bound

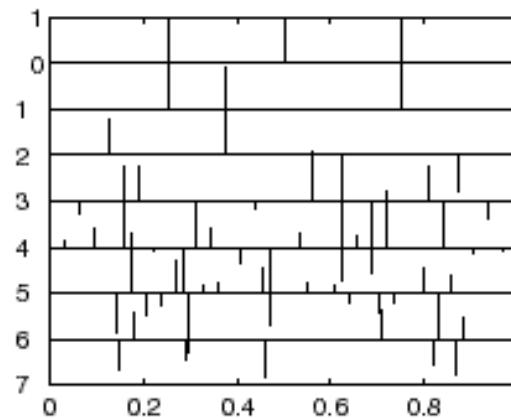
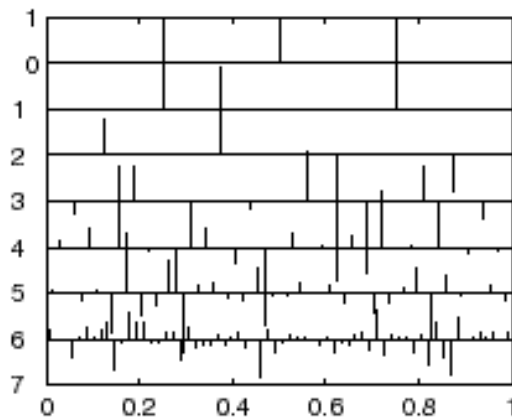
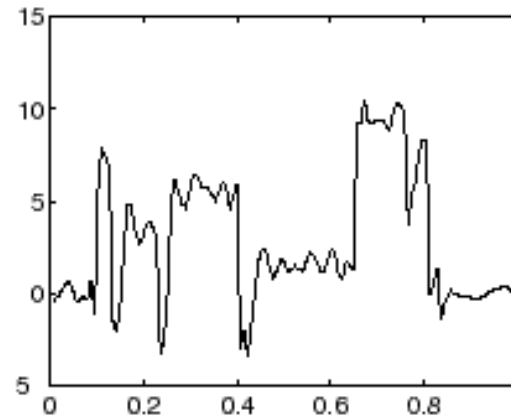
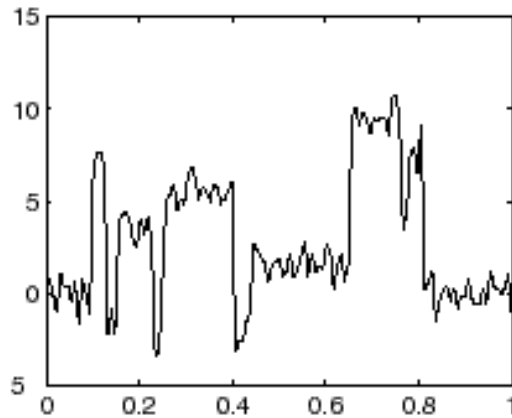
Usually take VC-dim.  $h=m$  (number of wavelets or DoF)

# Empirical Results: signal denoising

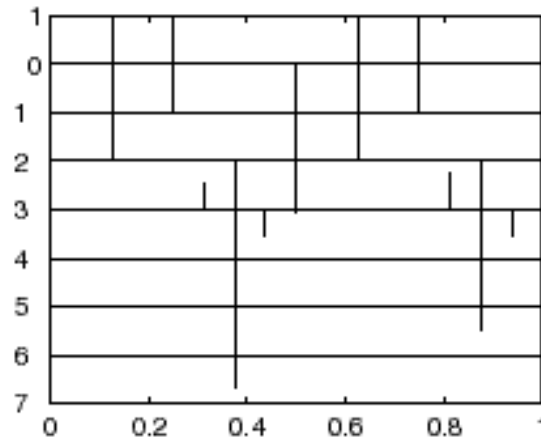
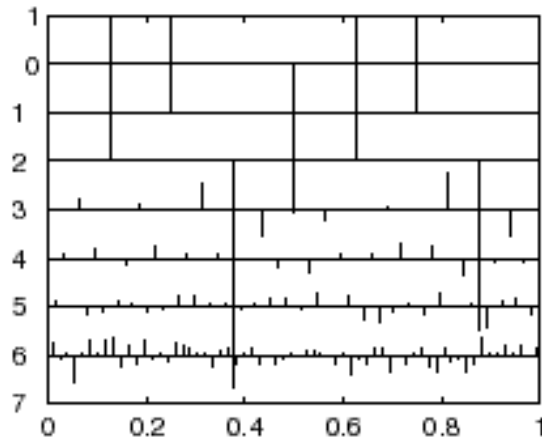
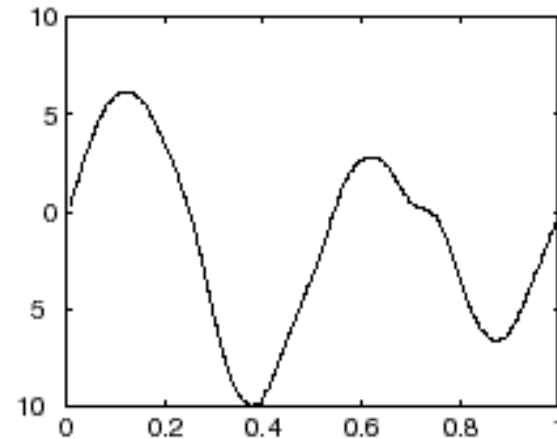
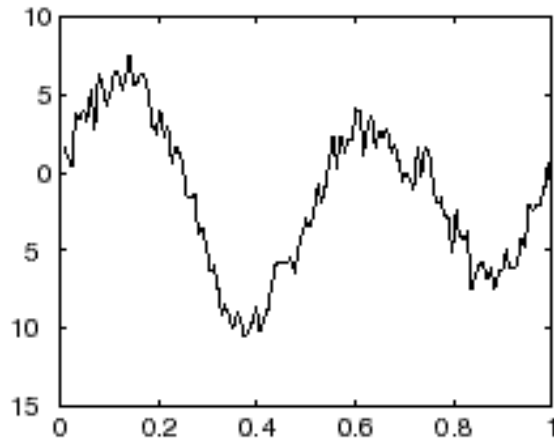
- **Two target functions**
- **Data set:** 128 noisy samples, SNR = 2.5



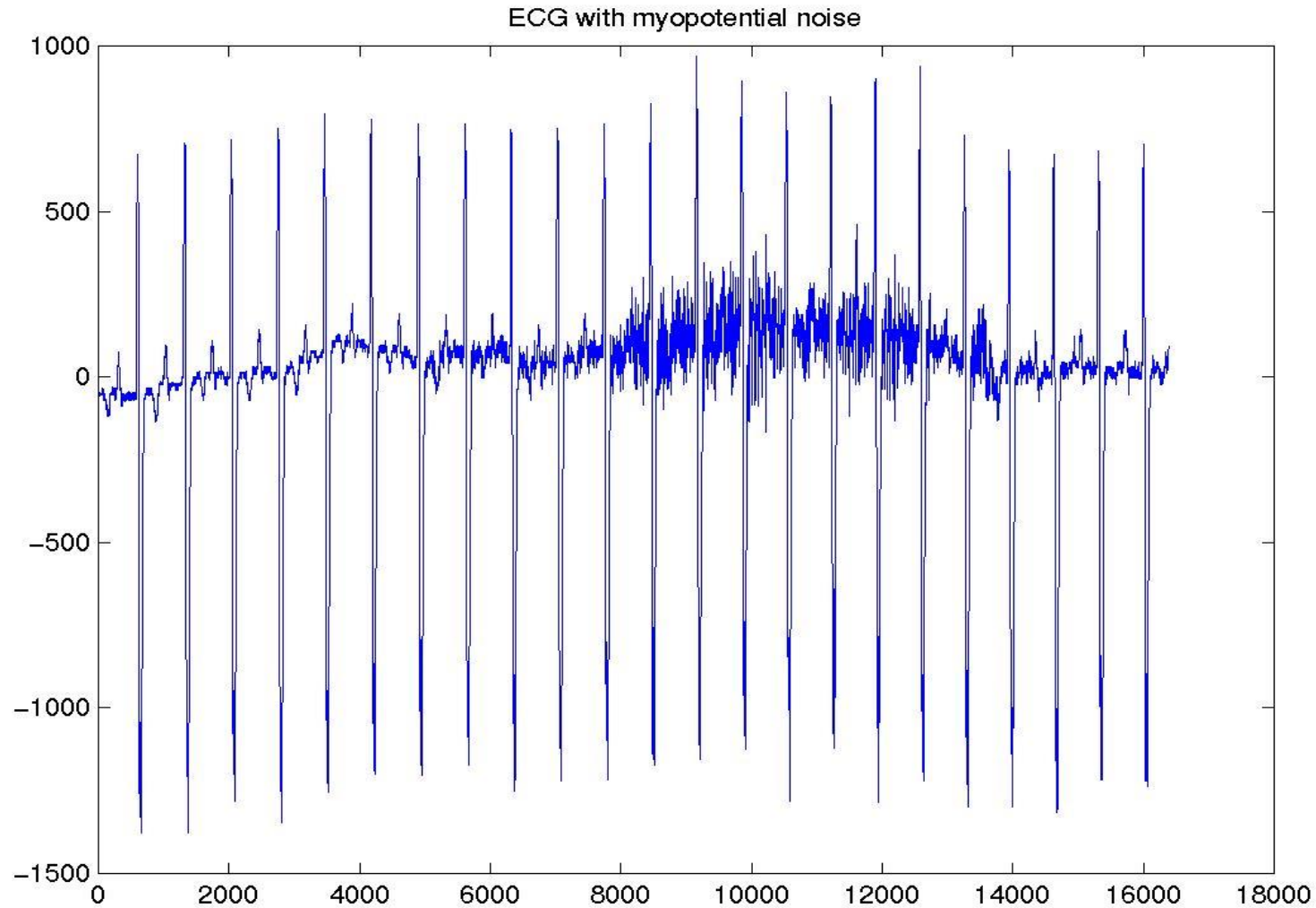
# Empirical Results: Blocks signal estimated by VC-based denoising



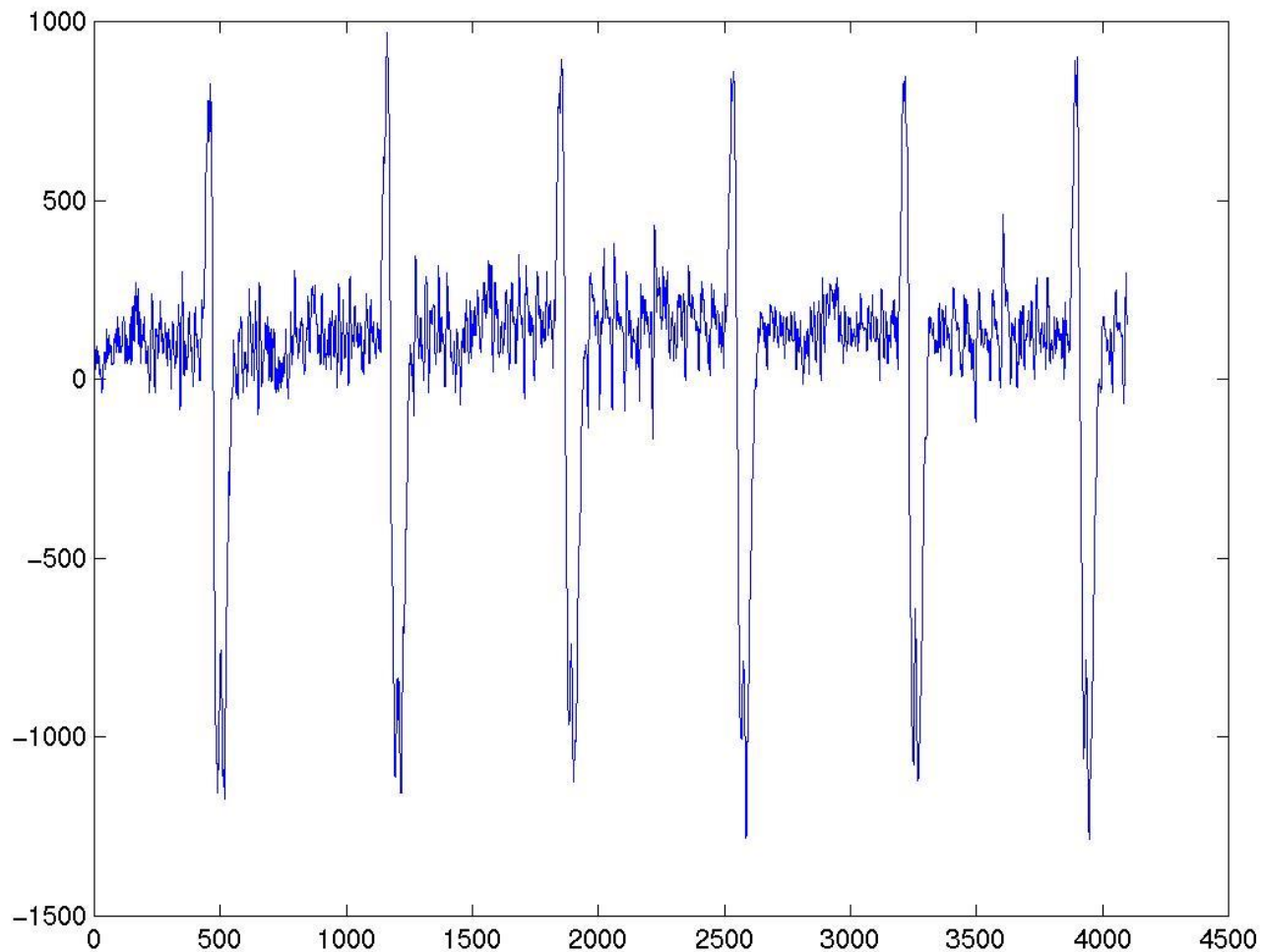
# Empirical Results: Heavisine estimated by VC-based denoising



# Application Study: ECG Denoising



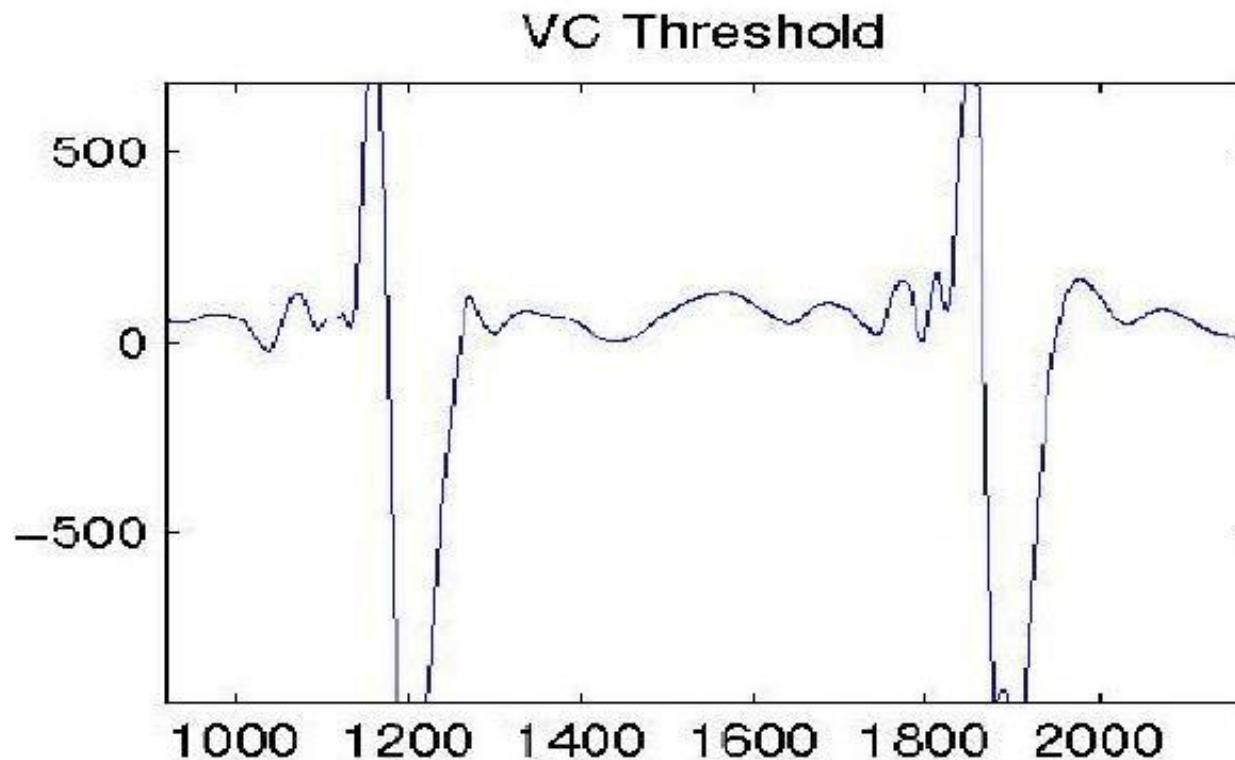
# A closer look of a noisy segment





# Denoised ECG signal

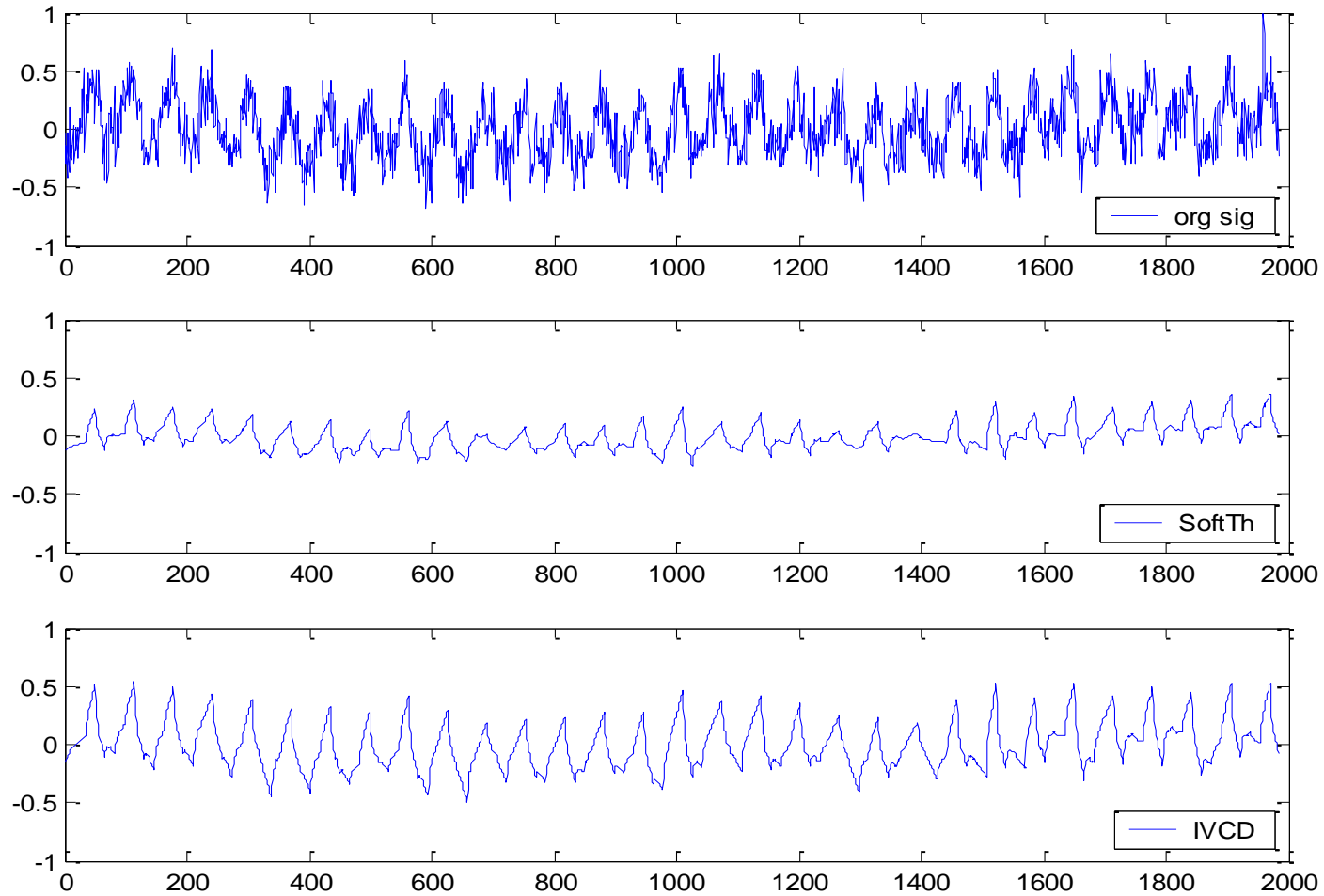
VC denoising applied to 4,096 noisy samples.  
The final model (below) has 76 wavelets



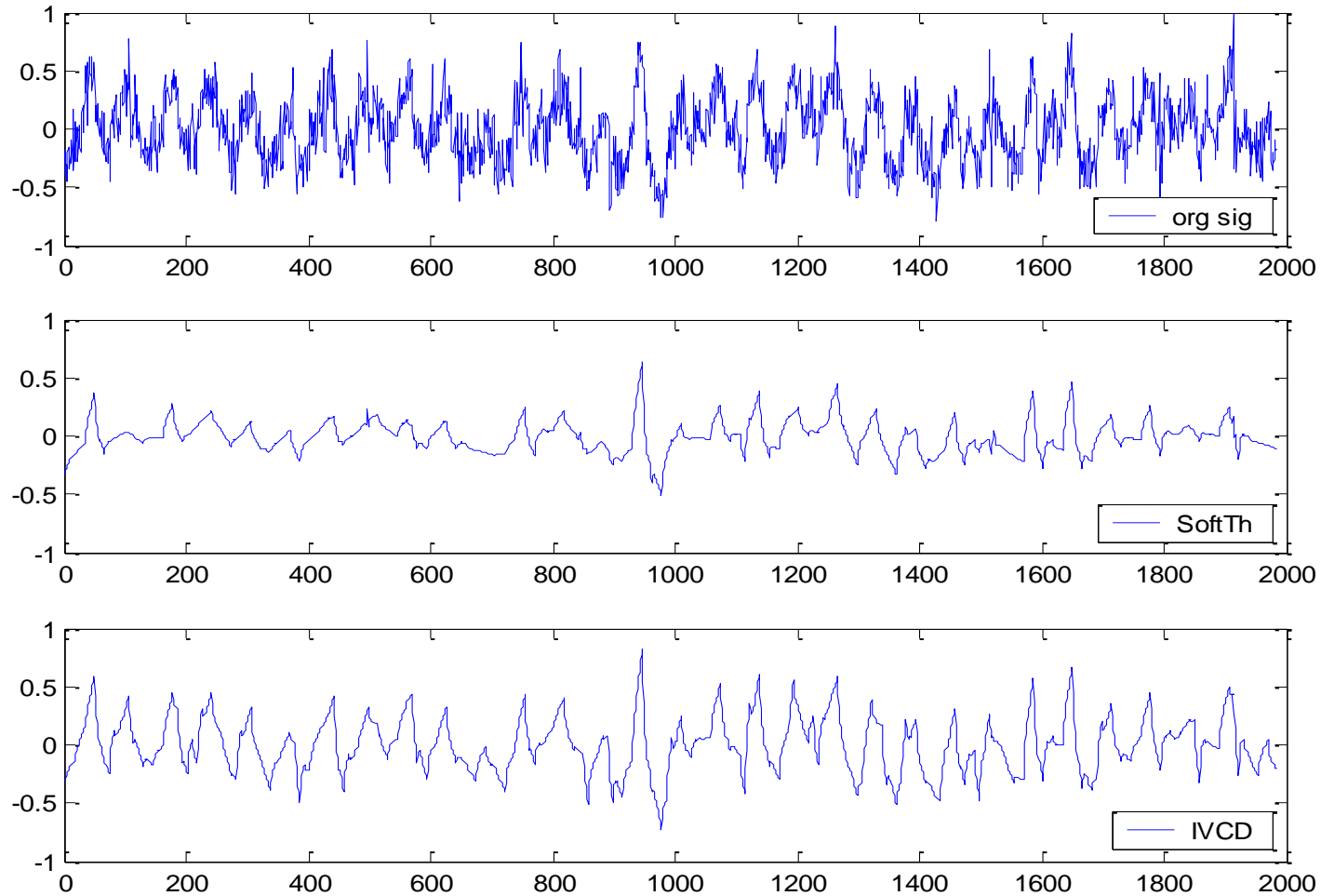
# Application: fMRI signal denoising

- **MOTIVATION:** understanding brain function via functional MRI
- **DATA SET:** provided by CMRR at UMN
- Signals (~ waveforms) recorded at a certain brain location response to external stimulus applied 32 times
- **Data Set 1** ~ signals at the *visual cortex* in response to visual input light blinking 32 times
- **Data Set 2** ~ signals at the *motor cortex* recorded when human subject moves his finger 32 times
- FMRI denoising = obtaining a better version of noisy signal

# Visual Cortex Data + Denoising



# Motor Cortex Data + Denoising



# Discussion

- **Application of VC-theory to signal denoising**
  - orthogonal basis functions
  - **nonlinear estimator**: sparse feature selection
- **Finite sample setting: importance of**
  - **Ordering (of basis functions)** ~ SRM structure
  - **Model selection** (thresholding)
- **Large-sample setting:**
  - **type of basis functions** (representation)

# OUTLINE

- Objectives
- Inductive learning problem setting
- Statistical Learning Theory
- Applications
- **Measuring the VC-dimension**
- Summary and discussion

# Measuring the VC-Dimension

- VC-dimension is difficult to estimate (for most practical learning methods)
- **Experimental estimation** (Vapnik et al 1994) for binary classification problems
- **Main idea:**
  - apply learning method to randomly labeled data and measure the training error
  - the deviation of the training error from 0.5 depends on the flexibility (VC-dimension) of an estimator.

# Measuring the VC-Dimension

- **Experimental estimation** (Vapnik et al 1994)  
for binary classification problems:
  - Based on theoretic analysis of the **maximum deviation of error rates** between two independently labeled data sets
  - Perform repeated random experiments with different data sets and sample sizes
  - Estimate VC-dim by fitting the theoretic function (which depends only on  $n/h$ )



# Measuring the VC-dim: Theory

- Consider binary classification, and denote  $n$  labeled training samples  $\mathbf{Z}_n = \{\mathbf{z}_i, i = 1, \dots, n\}$
- Apply an estimator (learning method) to training data and measure the **max deviation** of error rates observed on two independently labeled data sets (of size  $n$ ):

$$\xi(n) = \max_{\omega} (| \text{Error}(\mathbf{Z}_n^1) - \text{Error}(\mathbf{Z}_n^2) |)$$

- According to VC-theory, this deviation is bounded by

$$\xi(n) \leq \Phi(n/h) \quad \text{or} \quad \xi(n) \approx \Phi(n/h)$$

where

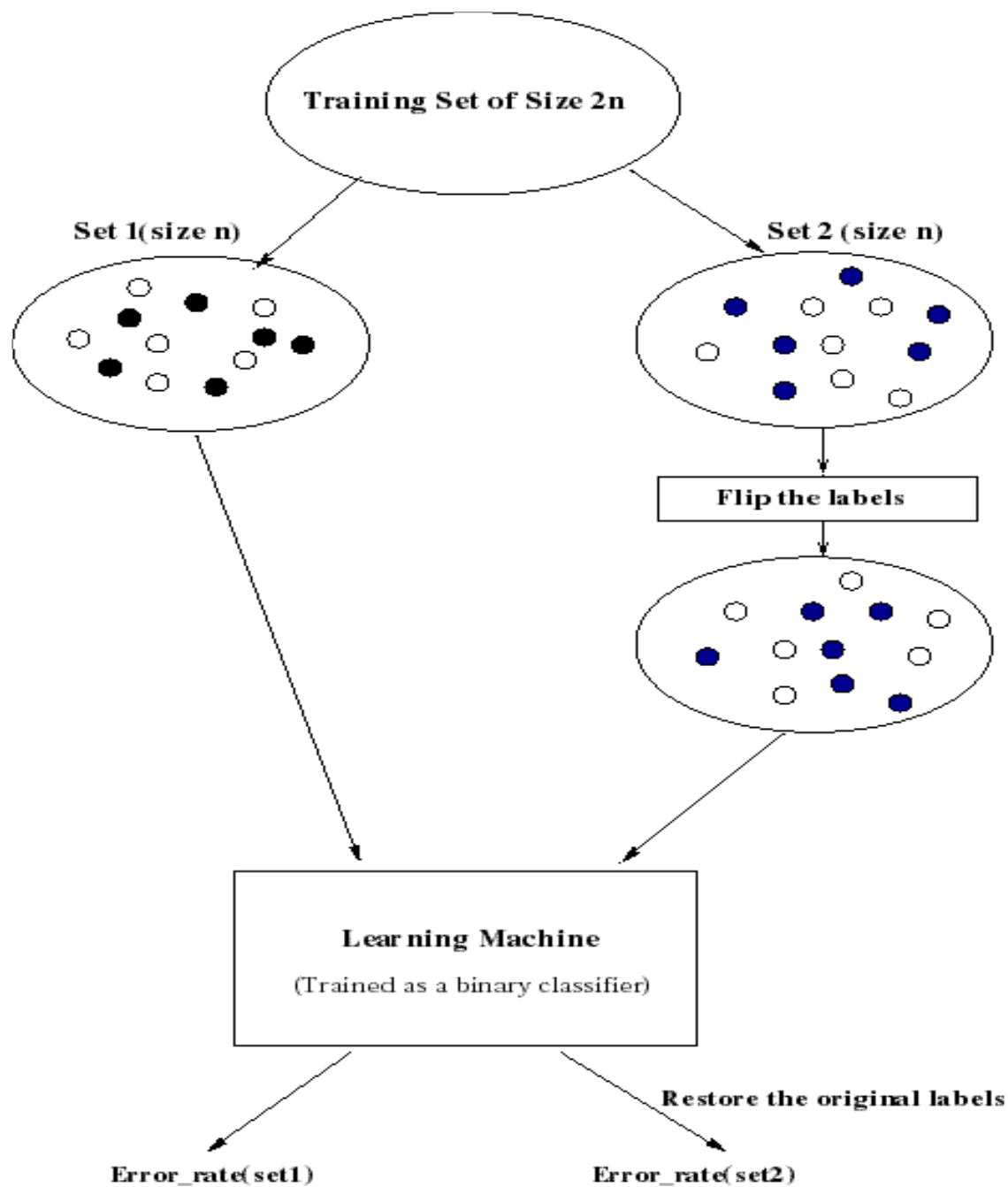
$$\Phi(\tau) = \begin{cases} 1 & \text{if } (\tau < 0.5) \\ a \frac{\ln(2\tau) + 1}{\tau - k} \left( \sqrt{1 + \frac{b(\tau - k)}{\ln(2\tau) + 1}} + 1 \right) & \text{otherwise} \end{cases}$$

# Measuring the VC-Dimension

Using experimental measurements of  $\xi(n)$  we can fit these measurements to analytic function  $\Phi(n/h)$  that depends only on VC-dim.

## Experimental procedure (for one measurement)

- Generate a randomly labeled set of size  $2n$
- Split it into two sets of equal size:  $Z1$  and  $Z2$
- Flip the class labels for the second set  $Z2$
- Merge the two sets and train binary classifier
- Separate the sets and flip the labels on the second set back again
- Measure the difference between the error rates on the two sets:  $\xi(n) = |Error(Z1) - Error(Z2)|$ .



# Measuring the VC-Dimension

- Single measurements of  $\xi(n)$  is affected by random variability of random sample
- To reduce variability:
  - the experiment is repeated for different data sets with varying sample sizes  $n_1, n_2, \dots, n_k$ , in the range  $0.5 \leq n_i/h \leq 30$
  - several ( $m_i$ ) repeated experiments are performed for each sample size  $n_i$
- The effective VC-dimension provides the best fit between  $\Phi(n/h)$  and measured values  $\bar{\xi}(n_i)$

$$h^* = \arg \min_h \sum_{i=1}^k [\bar{\xi}(n_i) - \Phi(n_i / h)]^2$$

# Measuring the VC-Dimension

- Can be applied to estimating the VC-dimension of penalized estimators (ridge regression), i.e. finding dependency  $h=h(\lambda)$ 
  - see *Example 7.1* on p.266 (in the textbook)
- This dependency can be then used for *analytic model selection* using practical VC-bound
- Empirical results show that estimated VC-dimension works better than using ‘effective’ DoF for ridge regression, in conjunction with analytic model selection criteria.

# OUTLINE

- Objectives
- Inductive learning problem setting
- Statistical Learning Theory
- Applications
- Measuring the VC-dimension
- **Summary and discussion**

# Summary and Discussion: VC-theory

- **Methodology**
  - learning problem setting (KID principle)
  - concepts (risk minimization, VC-dim., structure)
- **Interpretation/ evaluation of existing methods**
- **Model selection** using VC-bounds
- **Basis for new types of inference** (TBD later)
- **Clear limitations/constraints** for all learning methods based on the idea of ERM
- **What a theory cannot do:**
  - provide formalization (for a given application)
  - select 'good' structure (for a given application)
  - always a gap between theory and applications

# References

- **General references on VC-theory**

V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995

V. Vapnik, *Statistical Learning Theory*, Wiley, 1998

- **Model selection using VC-bounds**

V. Cherkassky, X. Shao, F. Mulier and V. Vapnik, Model Complexity Control for regression using VC generalization bounds, *IEEE Trans. on Neural Networks*, 10, 5, 1075-1089, 1999

V. Cherkassky and Y. Ma, Comparison of model selection for regression, *Neural Computation*, MIT Press 15 (7), 1691-1714, 2003

- **Market timing of international mutual funds**

V. Cherkassky and S. Dhar, Market Timing of International Funds: A Decade after the Scandal, *Proc CIFE* 2012

- **Measuring the VC-dimension**

V. Vapnik, E. Levin and Y Le Cun, Measuring the VC-dimension of a learning machine, *Neural Computation*, 6, 851-876, 1994

X. Shao, V. Cherkassky and W. Li, Measuring the VC-dimension using optimized experimental design, *Neural Computation*, 12, 1969-1986, 2000

- **Signal denoising using VC-theory**

V. Cherkassky and X. Shao, Signal estimation and denoising using VC-theory, *Neural Networks*, Pergamon, 14, 37-52, 2001

V. Cherkassky and S. Kilts, Myopotential denoising of ECG signals using wavelet thresholding methods, *Neural Networks*, Pergamon, 14, 1129-1137, 2001