2nd Congress of Polish Statistics

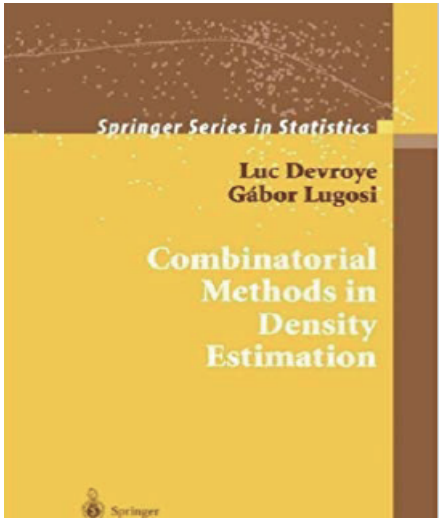
Nearest Neighbor Methods for Nonlinear Regression: Time Series Data Case

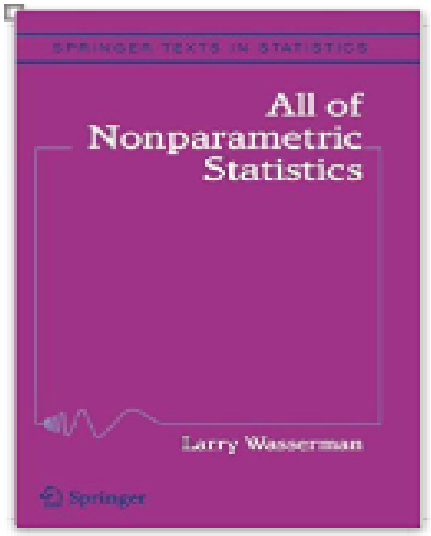
Miroslaw Pawlak

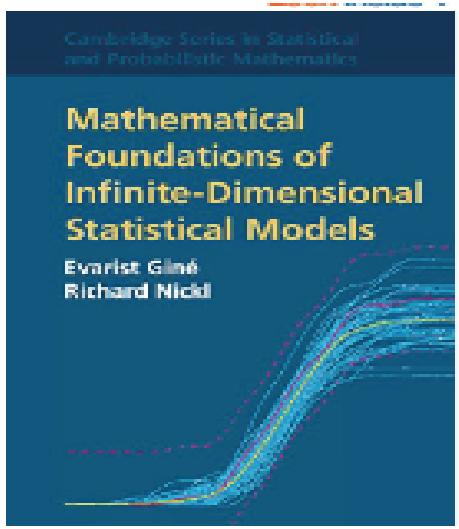
U of M & AGH

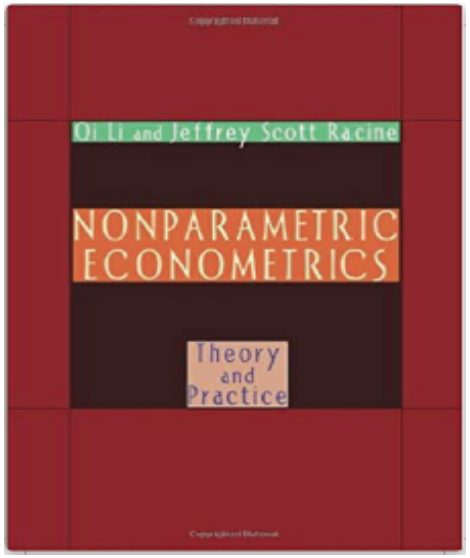
Miroslaw.Pawlak@umanitoba.ca

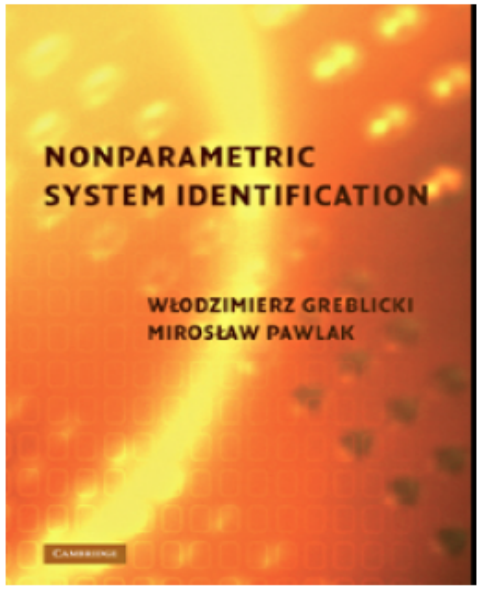
mirek.pawlakk@gmail.com

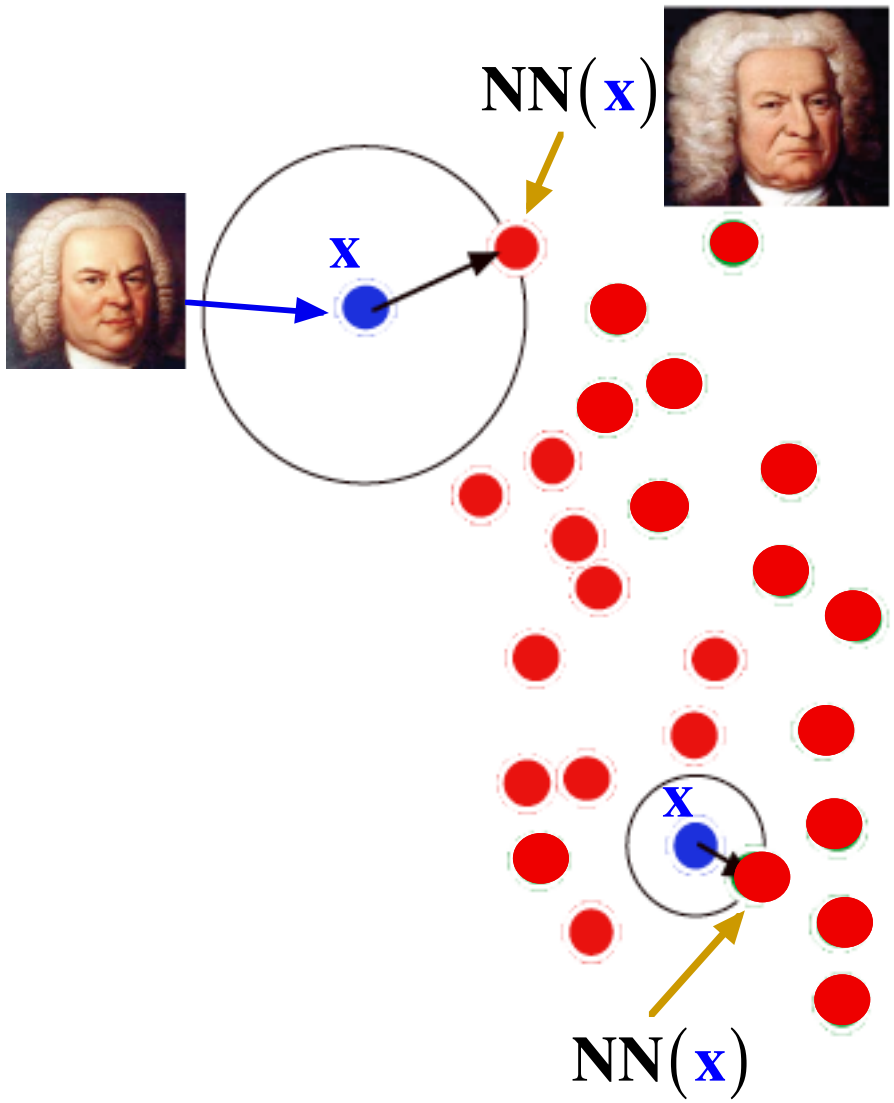

Springer Series in Statistics
Luc Devroye
Gábor Lugosi
Combinatorial Methods in Density Estimation
Springer


SPRINGER TEXTS IN STATISTICS
All of Nonparametric Statistics
Larry Wasserman
Springer

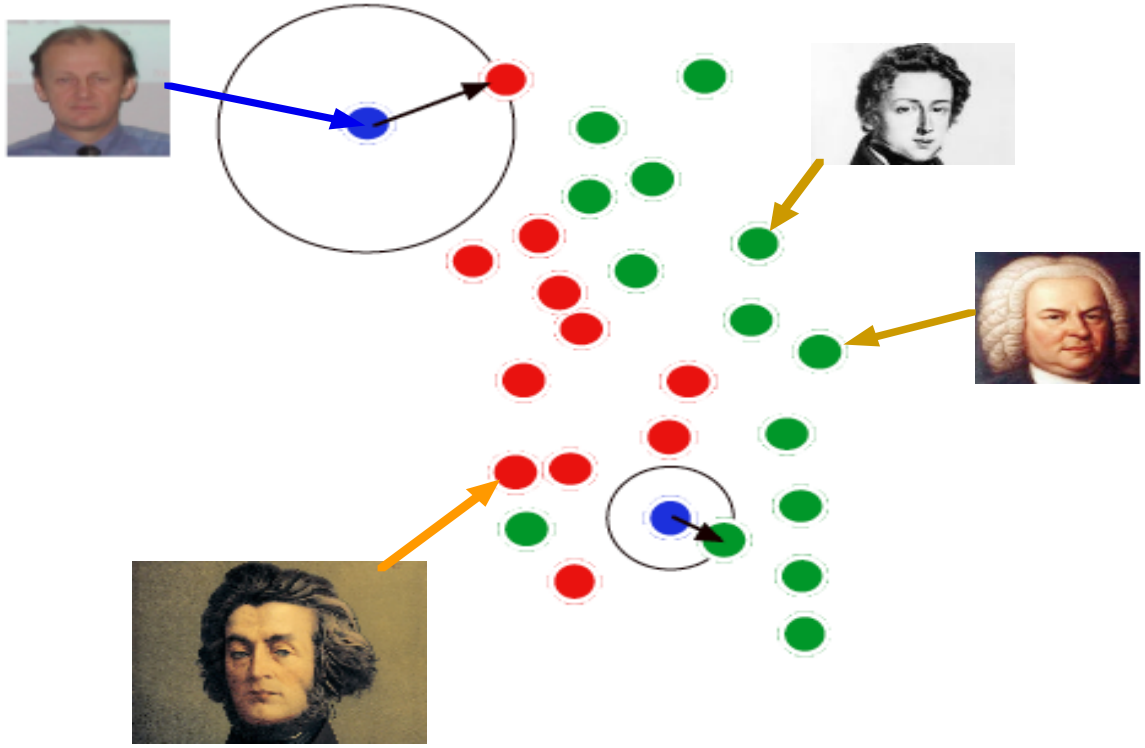

Cambridge Series in Statistical and Probabilistic Mathematics
Mathematical Foundations of Infinite-Dimensional Statistical Models
Evarist Giné
Richard Nickl


Qi Li and Jeffrey Scott Racine
NONPARAMETRIC ECONOMETRICS
Theory and Practice


NONPARAMETRIC
SYSTEM IDENTIFICATION
WŁODZIMIERZ GREBLICKI
MIROSŁAW PAWLAK


NN(x)
x
x
NN(x)

1-NN: Classification Problem



$$R^* \leq \lim_{n\to\infty} P(\text{NN error}) \leq 2R^*$$

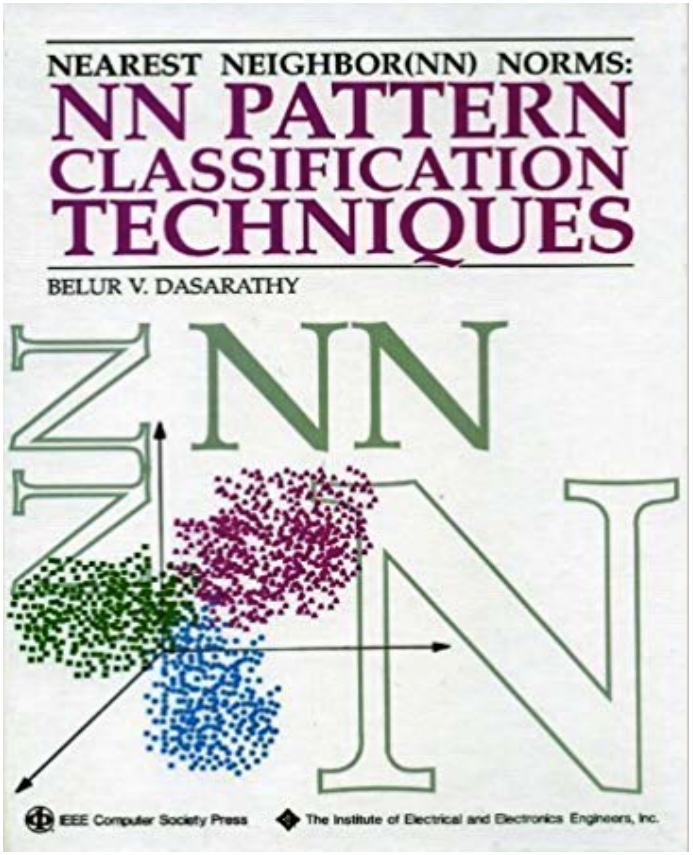
$$\uparrow$$

$$\inf_{\text{all rules}} P(\text{error})$$

Nearest Neighbor Pattern Classification

T. M. COVER, MEMBER, IEEE, AND P. E. HART, MEMBER, IEEE

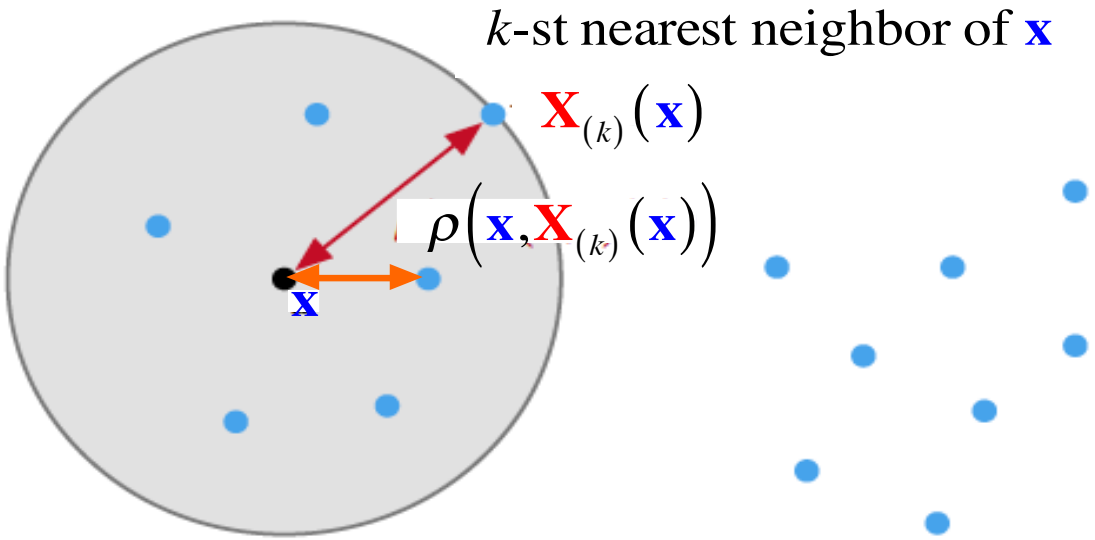
IEEE Inf. Theory, 1967



➤ **E. Fix & J.L. Hodges, Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine , Randolph Field, Texas, 1951.**


INDY/PULSE
HAVING GOOD NEIGHBOURS CAN HELP CUT HEART ATTACK RISK, STUDY SHOWS

Estimation methods based on nearest neighbors are the simplest and most intuitively appealing of all nonparametric techniques



- **Density Estimation**
- **Regression Estimation**
- **Classification**
- **Time Series Analysis**

OUTLINE

I NN Distance

II k-NN Regression Estimates: iid case

III k-NN Regression Estimates: Time Series Systems

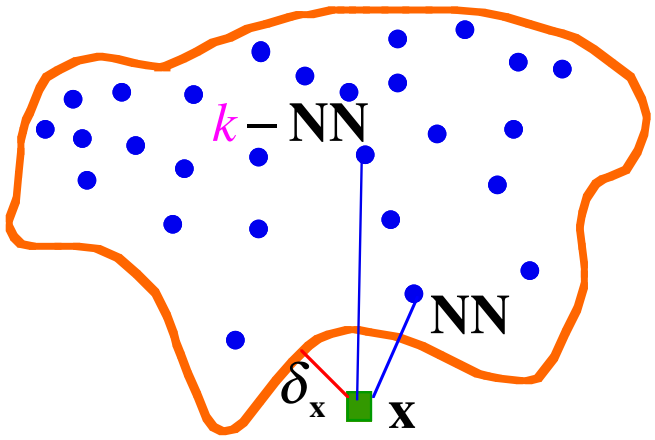
IV k-NN Regression Estimates: Time Series Forecasting

V Concluding Remarks

I NN Distance

Consistency

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbf{R}^d \quad \textbf{iid}$$

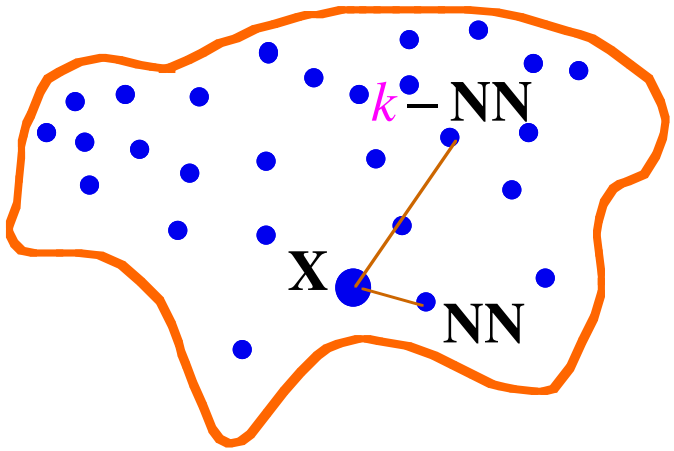


$$\left\|\mathbf{x}-\mathbf{X}_{(1)}(\mathbf{x})\right\| \le \left\|\mathbf{x}-\mathbf{X}_{(2)}(\mathbf{x})\right\| \le \cdots \le \left\|\mathbf{x}-\mathbf{X}_{(n)}(\mathbf{x})\right\|$$

If $\dfrac{k}{n} \to 0 \quad \Rightarrow \quad \left\|\mathbf{x}-\mathbf{X}_{(k)}(\mathbf{x})\right\| \to \delta_{\mathbf{x}}$ **(P, Pr.1)**

- $k = 1$
- $k = \sqrt{n}$
- $k = n/2$

Consistency



$\mathbf{X}$ independent of the data $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$

$$\textbf{If } \frac{k}{n} \to 0 \quad \Rightarrow \quad \left\| \mathbf{X} - \mathbf{X}_{(k)}(\mathbf{X}) \right\| \to 0 \quad (\textbf{P}, \textbf{Pr.1})$$

- $k = 1$
- $k = \sqrt{n}$
- $k = n/2$

Rate $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbf{A} \subset R^d$; $\mathbf{X} \in \mathbf{A}$

↖ **compact set with diameter = 1**

	$\mathbf{E}\left\|\mathbf{X} - \mathbf{X}_{(1)}(\mathbf{X})\right\|^2$	$\mathbf{E}\left\|\mathbf{X} - \mathbf{X}_{(k)}(\mathbf{X})\right\|^2$
$d = 1$	$\frac{2}{n}$	$8\frac{k}{n}$
$d = 2$	$\frac{7.42\ldots}{n}$	$c(2)\frac{k}{n}$
$d \geq 3$	$c(d)\left(\frac{1}{n}\right)^{2/d}$	$c(d)\left(\frac{k}{n}\right)^{2/d}$

General distribution on A

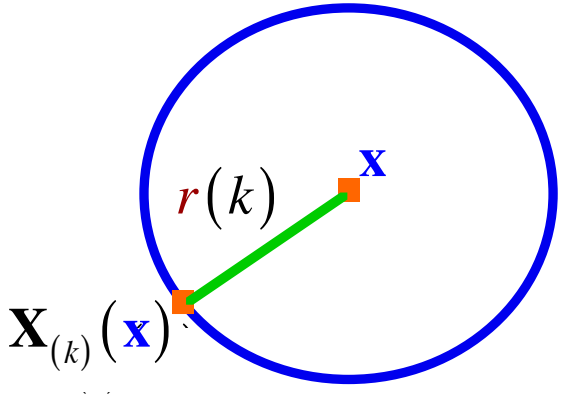
Rate $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbf{A} \subset R^d$; $\mathbf{X} \in \mathbf{A}$ - **compact set**

	$\mathbf{E}\left\|\mathbf{X} - \mathbf{X}_{(1)}(\mathbf{X})\right\|^2$	$\mathbf{E}\left\|\mathbf{X} - \mathbf{X}_{(k)}(\mathbf{X})\right\|^2$
$d = 1$	$c(1)\frac{1}{n^2}$ *	$c(1)\frac{k}{n^2}$ *
$d = 2$	$c(2)\frac{1}{n}$	$c(2)\frac{k}{n}$
$d \geq 3$	$c(d)\left(\frac{1}{n}\right)^{2/d}$	$c(d)\left(\frac{k}{n}\right)^{2/d}$

Distribution with a density defined on $\mathbf{A}$

Coverage Probability - Pr. Integral Transform

$$\int_{\mathbf{B}(\mathbf{x};\, r(k))} f(\mathbf{v})\, d\mathbf{v} \;\propto \mathbf{Beta}(k, n-k+1)$$



$$r(k) = \left\| \mathbf{x} - \mathbf{X}_{(k)}(\mathbf{x}) \right\|$$

- $\mathbf{E}\left[\int_{\mathbf{B}(\mathbf{x};\, r(k))} f(\mathbf{v})\, d\mathbf{v} \right] = \dfrac{k}{n+1}$
- $\mathbf{Var}\left[\int_{\mathbf{B}(\mathbf{x};\, r(k))} f(\mathbf{v})\, d\mathbf{v} \right] = \dfrac{k(n-k+1)}{(n+1)^2 (n+2)}$

LLN $\quad k \to \infty \;\Rightarrow\; \dfrac{\int_{\mathbf{B}(\mathbf{x};\, r(k))} f(\mathbf{v})\, d\mathbf{v}}{k/n} \to 1 \qquad (\mathbf{P})$

II k-NN Regression Estimates: iid data

Regression

- $$\underbrace{\mathbf{E}\left[Y \mid \mathbf{X}=\mathbf{x}\right]}_{m(\mathbf{x})} = \arg\min_{g:\,\mathbf{R}^d \to \mathbf{R}} \mathbf{E}\left[Y - g(\mathbf{X})\right]^2$$

- $$\mathbf{E}\left[\mathbf{Y} - g(\mathbf{X})\right]^2 = \underbrace{\mathbf{E}\left\{\left|\mathbf{Y} - m(\mathbf{X})\right|^2\right\}}_{\textbf{Bayes Risk}} + \mathbf{E}\left\{\left|g(\mathbf{X}) - m(\mathbf{X})\right|^2\right\}$$

- $$Y = m(\mathbf{X}) + \varepsilon \;\Rightarrow\; \mathbf{E}\left[\mathbf{Y} - g(\mathbf{X})\right]^2 = \mathrm{var}\left[\varepsilon\right] + \mathbf{E}\left\{\left|g(\mathbf{X}) - m(\mathbf{X})\right|^2\right\}$$

k-NN Estimates

- **Wish to estimate** $m(\mathbf{x}) = \mathbf{E}[Y \mid \mathbf{X} = \mathbf{x}]$ **from**

$$\mathfrak{I}_n = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\} \; ; \; (\mathbf{X}_i, Y_i) \in \mathbf{R}^d \times \mathbf{R}$$

- **Ordered data set**

$$\mathfrak{I}_n(\mathbf{x}) = \{(\mathbf{X}_{(1)}, Y_{[1]}), ..., (\mathbf{X}_{(n)}, Y_{[n]})\} \leftarrow \mathbf{x} \in \mathbf{R}^d$$

$$\|\mathbf{X}_{(1)} - \mathbf{x}\| \leq \|\mathbf{X}_{(2)} - \mathbf{x}\| \leq \ldots \leq \|\mathbf{X}_{(n)} - \mathbf{x}\|$$

$Y_{[1]}$ $Y_{[2]}$ $Y_{[n]}$ ← **concomitants**

k-NN Estimates: iid data

1-NN Estimate

$$\widetilde{m}_{NN}(\mathbf{x}) = Y_{[1]}(\mathbf{x})$$

$$\$\$ \quad Y = m(\mathbf{X}) + \varepsilon \quad \$\$$$

$$Y_{[1]}(\mathbf{x}) = m\left(\mathbf{X}_{(1)}(\mathbf{x})\right) + \varepsilon_{[1]}$$

$$\$\$ \quad \mathbf{E}\left[m\left(\mathbf{X}_{(1)}(\mathbf{X})\right) - m(\mathbf{X}) \right]^2 \to 0 \quad \$\$$$

1-NN Estimate

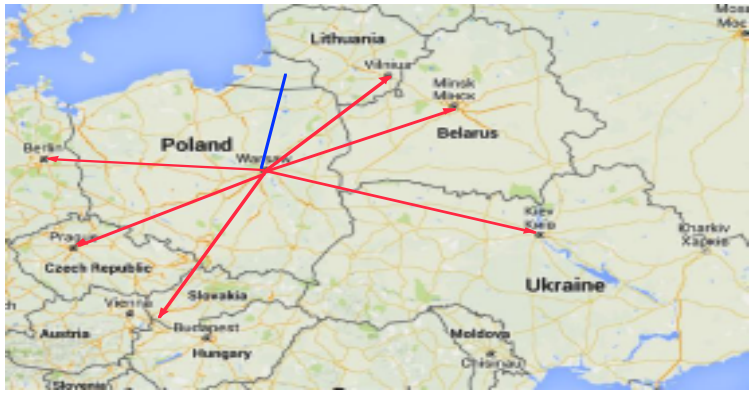
Non-consistent estimate of $m(u)$

$$\widetilde{m}_{NN}(\mathbf{x}) \to m(\mathbf{x}) + \sqrt{\underbrace{\mathbf{E}\left\{\left|\mathbf{Y} - m(\mathbf{X})\right|^2\right\}}_{\text{var of noise}}} \qquad (\mathbf{P})$$

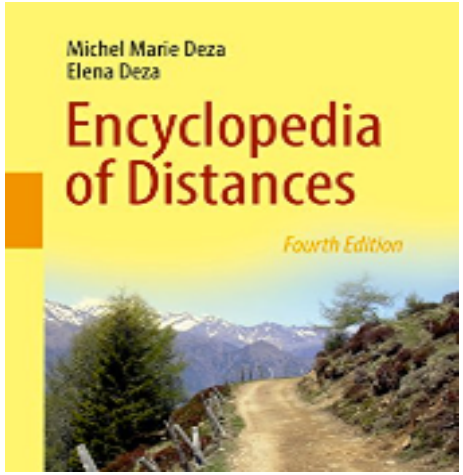
$$\mathbf{E}\left[\widetilde{m}_{NN}(\mathbf{X}) - m(\mathbf{X})\right]^2 \to \underbrace{\mathbf{E}\left[\left|\mathbf{Y} - m(\mathbf{X})\right|^2\right]}_{\sigma_\varepsilon^2}$$

The Standard k-NN Estimate: $\widetilde{m}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} Y_{[i]}$

- k – number of neighbors



- **distance** "metric"



- **The Standard k-NN Estimate:** $\widetilde{m}(\mathbf{x}) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{[i]}$

  - $k_n \to \infty$ - **controls variance**

    $\Leftarrow$ **kernel method** $\left(h_n \to 0, nh_n^d \to \infty\right)$

  - $\frac{k_n}{n} \to 0$ - **controls bias** $\uparrow$

$$\widehat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{h}\right)}$$

The Standard k-NN Estimate: Universal Consistency

Theorem 1 Let $\mathbf{E}|Y|^2 < \infty$. Then

$$\mathbf{E}\left\{\left|\widetilde{m}(\mathbf{X}) - m(\mathbf{X})\right|^2\right\} \to 0 \text{ for all distributions of } (\mathbf{X}, Y)$$

iff

$$k_n \to \infty \text{ and } \frac{k_n}{n} \to 0$$

- Stone: Annals of Stat., 1977
- Distance tie-breaking
- Geometric Proof
  Covering Numbers

The Standard k-NN Estimate: Rate $\Leftarrow d = 1$

Theorem 2 Let $m(\bullet) \in \mathrm{Lip}(\mathbf{R})$ and

$$\tilde{k}_n^* = \left\lceil c_1 n^{2/3} \right\rceil .$$

Then

$$\mathbf{E}\left[\tilde{m}(X) - m(X)\right]^2 = O\left(n^{-2/3}\right)$$

This rate cannot be improved for smoother characteristics

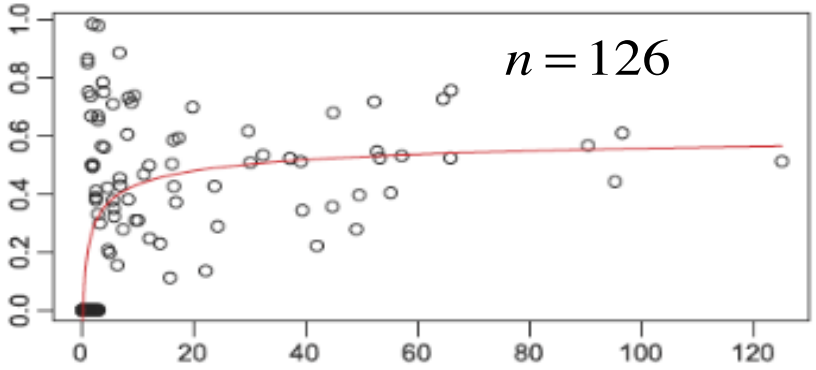
$$O\left(n^{-2/3}\right) \to O\left(n^{-4/5}\right) \to O\left(n^{-4/5}\right) \to \ldots.$$

The Standard k-NN Estimate: Useful Properties

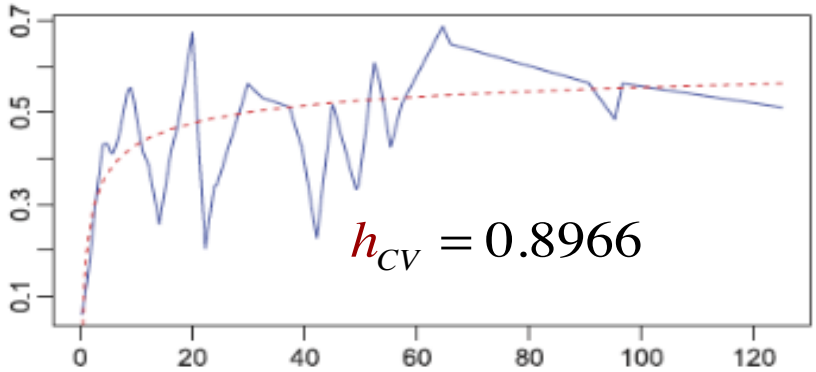
The use of **k** -nearest neighbors weights is convenient in a number of nonparametric estimation applications because they avoid the **random denominator problem** intrinsic to commonly used **kernel weights**.

$$\widetilde{m}(\mathbf{x}) = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{[i]} \;\leftrightarrow\; \widehat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|\mathbf{x}-\mathbf{X}_i\|}{h}\right)}$$

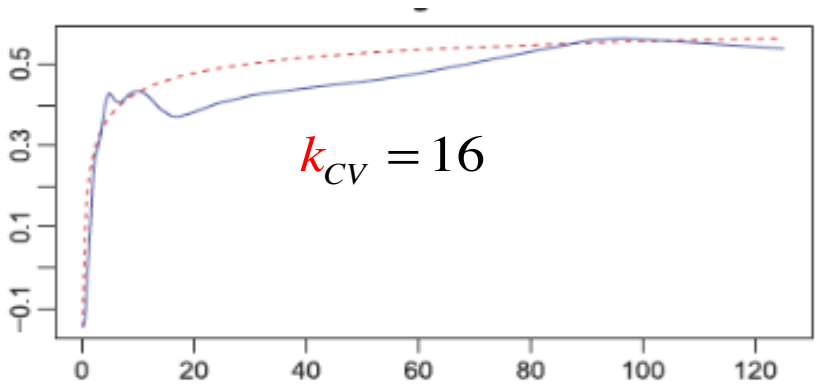
The k -NN estimate adapts to the non-homogeneity of data



LS for $m(x;\theta) = \theta_1 + \theta_2 x^{\theta_3}$



$$\frac{\sum_{i=1}^{n} Y_i K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}$$

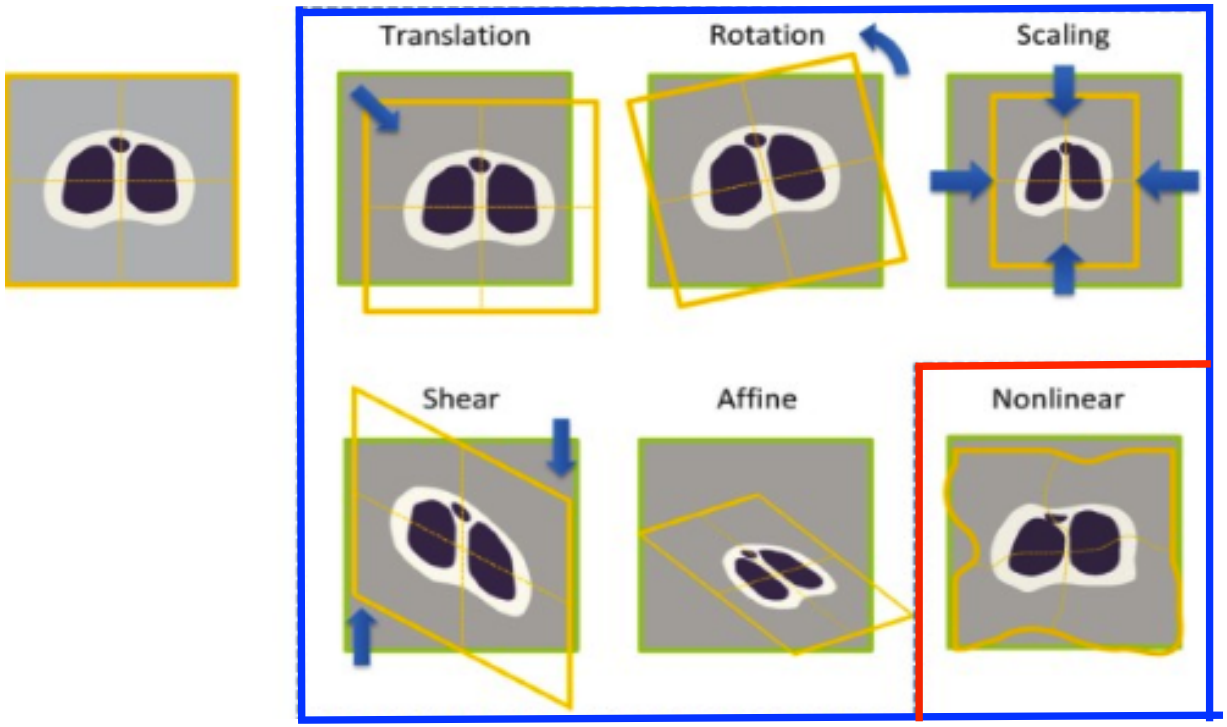


$$\frac{1}{k}\sum_{i=1}^{k} Y_{[i]}$$

- **The k -NN estimate can incorporate invariance**

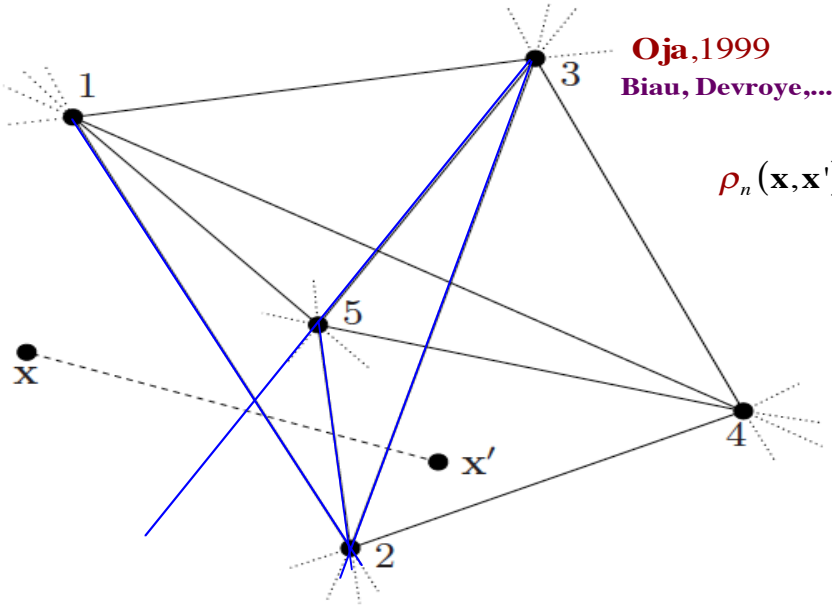
$$\widetilde{m}\left((\mathbf{Ax}+\mathbf{b});(\mathbf{Ax}+\mathbf{b})\Im_n\right) = \widetilde{m}\left(\mathbf{x};\Im_n\right)$$

linear



9 ? **6**

An Affine Invariant Distance



Oja,1999
Biau, Devroye,... JMA, 2012

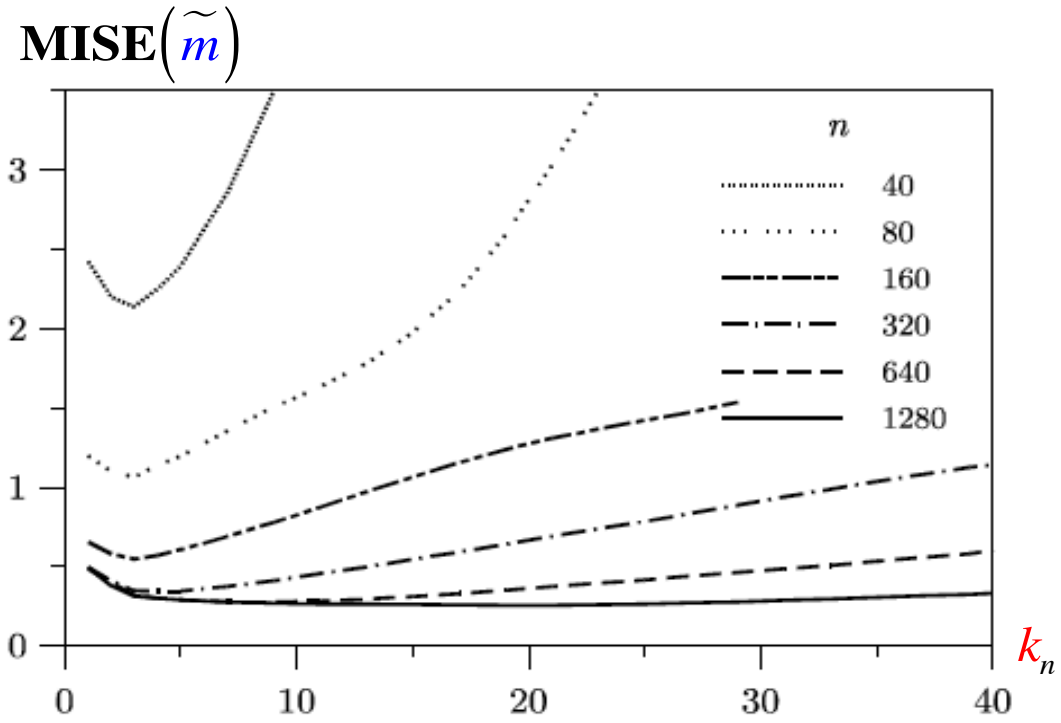
$\rho_n(\mathbf{x},\mathbf{x}') = 0 \Leftrightarrow \mathbf{x} = \mathbf{x}'$

$\rho_n(\mathbf{x},\mathbf{x}')$ = # of hyperplanes passing through 2 out of the points $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ that are separating $\mathbf{x}$ and $\mathbf{x}'$

= # of hyperplanes in $\mathbf{R}^d$ passing through d out of the points $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ that are separating $\mathbf{x}$ and $\mathbf{x}'$

- **The k -NN estimate can be extended to infinite dimensional spaces**
  - **Separable Banach spaces**
  - **Hilbert spaces with a countable orthonormal basis**
  - **Reproducing Kernel Hilbert Space**
    - **Biau,….; IEEE IT, 2010**

- **Choice of** k_n



➤ **CV choice for the k-NN estimate: Li, Annals of Stat. 84**

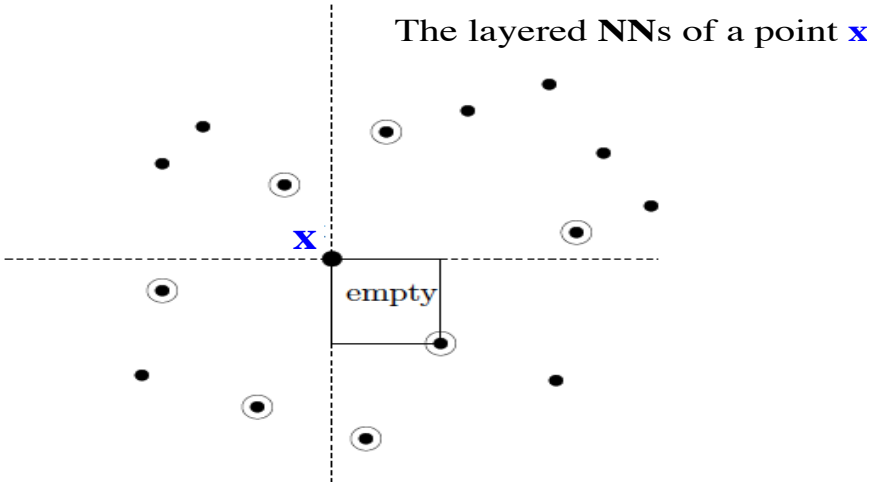
Other k-NN Estimates

Weighted k-NN Estimates

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{k_n} w_{ni} Y_{[i]} \qquad \sum_{i=1}^{k_n} w_{ni} = 1$$

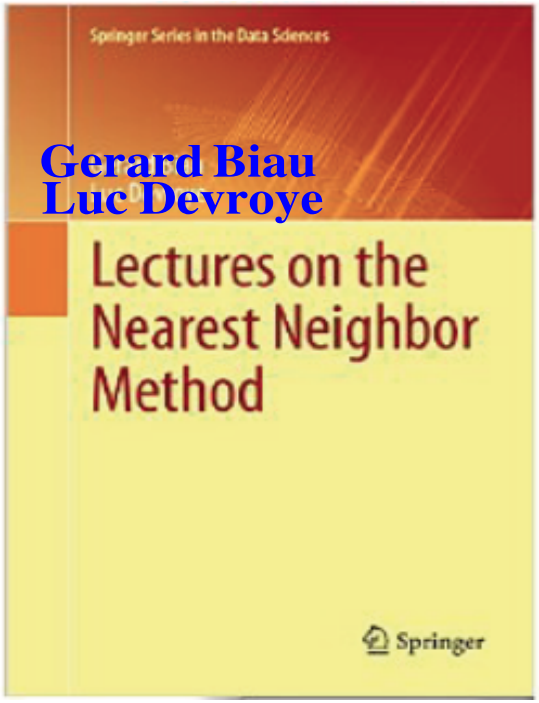
Stone's Theorem (**Stone: Annals of Stat., 1977**) gives general **iff** conditions on the weights in order to obtain the **universal** $\mathbf{L}_p$ - consistency.

Layered k-NN estimate



$$\widetilde{m}(\mathbf{x}) = \frac{1}{\left|l_n(\mathbf{x})\right|} \sum_{\{\mathbf{X}_i \in l_n(\mathbf{x})\}} Y_i$$

- Biau, Devroye, 2010


Springer Series in the Data Sciences
Gerard Biau
Luc Devroye
Lectures on the Nearest Neighbor Method
Springer

← **iid data**

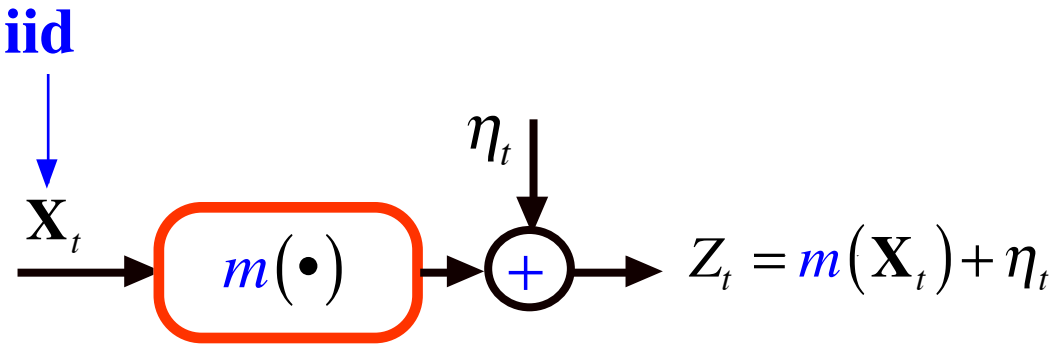
III k-NN Regression Estimates: Dependent Data

S. Yakowitz. Nearest-neighbour methods for time series. J. of Time Series, 1987.

S. Kulkarni and S. Posner. Rates of convergence of nearest neighbour estimation under arbitrary sampling. IEEE Trans. on Inf. Theory, 1995.

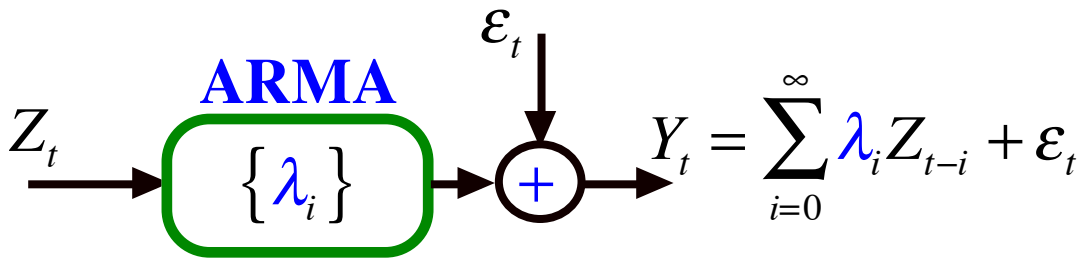
G. Boente and R. Fraiman. Robust nonparametric regression estimation for dependent observations. The Annals of Statistics, 1989.

Nonlinear Regression



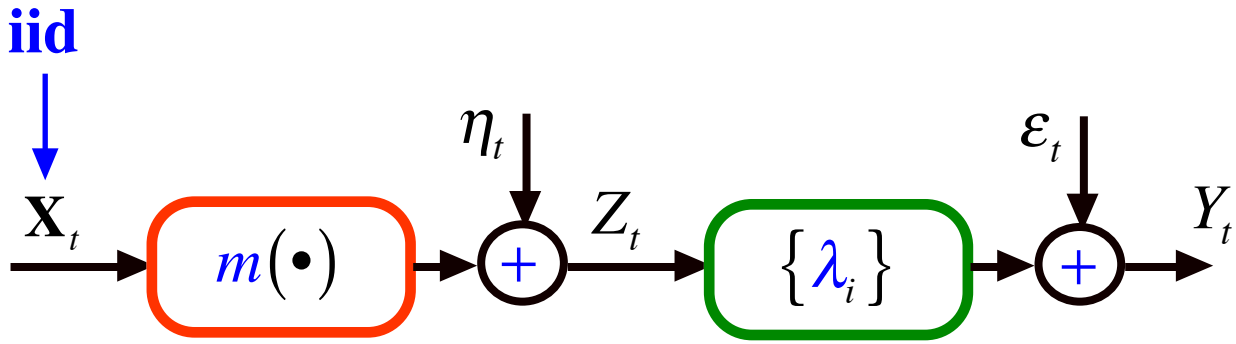
$$m(\mathbf{x}) = \mathbf{E}[Z_t \mid \mathbf{X}_t = \mathbf{x}]$$

Autoregressive (Linear) Process



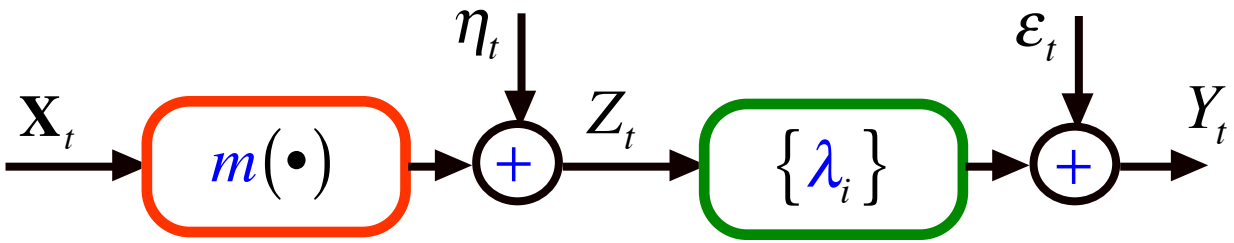
$$\lambda_i = \mathbf{cov}\{Y_t, Z_{t-i}\}$$

Nonlinear Time-Series System: Output Dependence Model



$$Y_t = \sum_{i=0}^{\infty} \lambda_i m\left(\mathbf{X}_{t-i}\right) + \sum_{i=0}^{\infty} \lambda_i \eta_{t-i} + \varepsilon_t$$

Nonlinear System = Hammerstein System



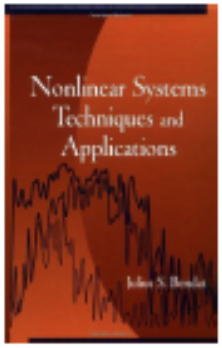
$$x(t)+\int_0^T \underbrace{\Lambda}_{known}(s,t)\,\underbrace{m}_{known}\big(x(s)\big)ds=\underbrace{y}_{known}(t)$$

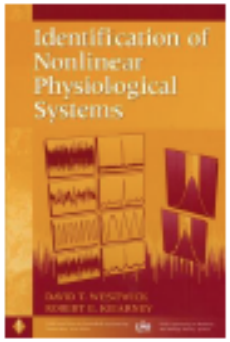
A. Hammerstein: Nichtlineare integralgleichungen nebst anwendungen,

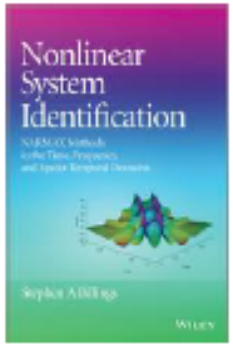
Acta Mathematica, **1930**.

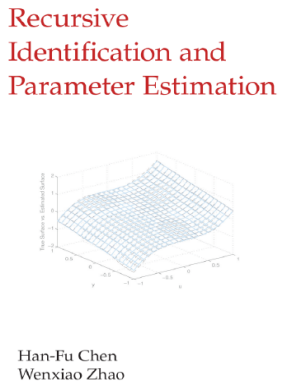
Nonlinear System Identification

- hot research area with numerous applications
- mostly parametric methods
- block-oriented models are the most popular
- lack of the basic statistical theory ←



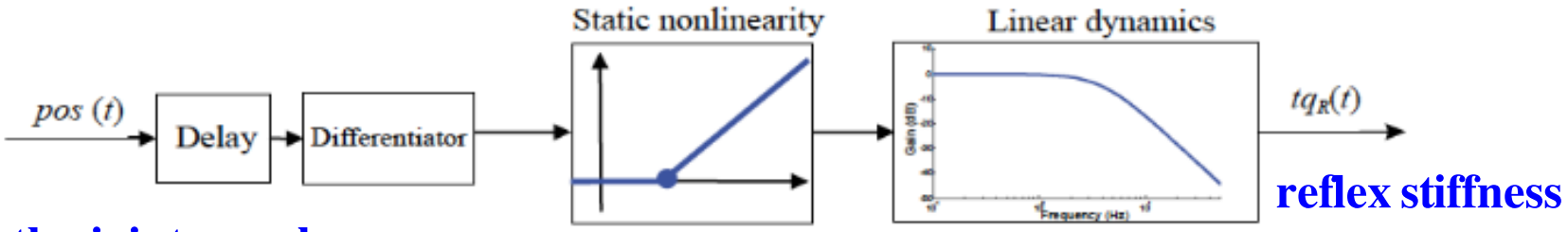






Biomedical Engineering

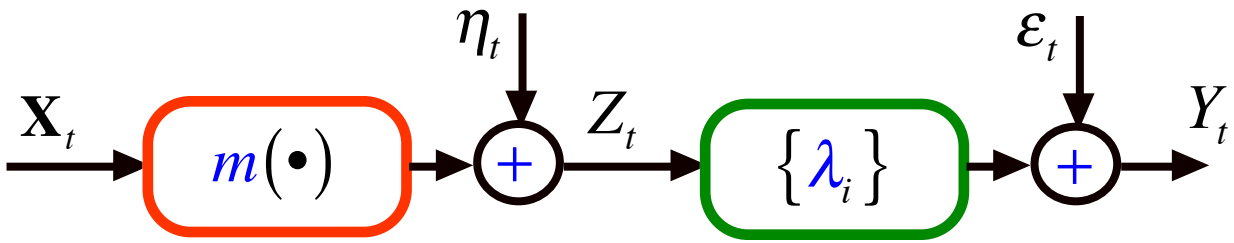
Reflex Stiffness Pathway



- **Kearny, … : Identification of intrinsic and reflex contributions to human ankle stiffness dynamics,"** IEEE Trans. on Biomedical Engineering, 1997

- This **nonlinear model** has been widely used to characterize the **reflex stiffness** in the ankle and trunk joints of both normal and pathological subjects.

Nonlinear System and Regression:



$$Y_t = m(\mathbf{X}_t) + e_t \quad \longleftarrow \quad \lambda_0 = 1$$

$$e_t = \sum_{i=1}^{\infty} \lambda_i m(\mathbf{X}_{t-i}) + \sum_{i=0}^{\infty} \lambda_i \eta_{t-i} + \varepsilon_t$$

$$\operatorname{cov}[e_t; e_{t+k}] = \operatorname{var}[m(\mathbf{X})] \sum_{i=1}^{\infty} \lambda_i \lambda_{i+k} + \sigma_\eta^2 \sum_{i=0}^{\infty} \lambda_i \lambda_{i+k} + \sigma_\varepsilon^2 \mathbf{1}(k=0)$$

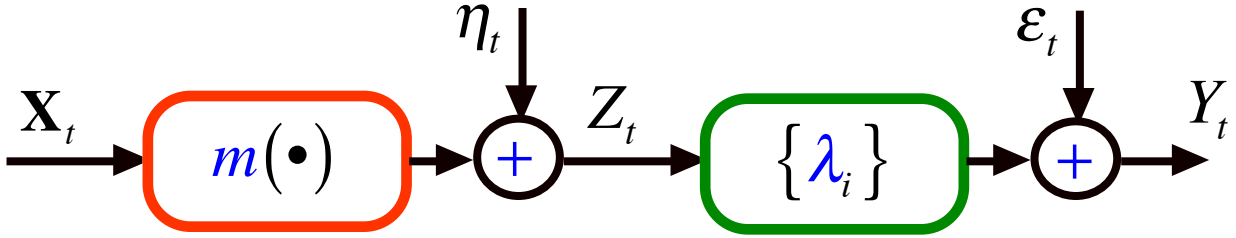
$$\lambda_0 = 1$$

$$E\left[m\left(\mathbf{X}_t\right)\right] = 0 \quad \Rightarrow \quad c = 0$$

$$\Downarrow$$

$$\boxed{E\left\{Y_t \mid \mathbf{X}_t = \mathbf{x}\right\} = m\left(\mathbf{x}\right)}$$

Problem Statement



Given $\Im_n = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$ estimate $m(\mathbf{x})$

A1 $\{\mathbf{X}_t\}$ - **iid** with pdf $f(\bullet)$

A2 • $\mathbf{E}\left[m^2(\mathbf{X}_t)\right] < \infty$, • $\sum_{i=0}^{\infty} |\lambda_i| < \infty$

A3 $\{\eta_t\}, \{\varepsilon_t\}$ - **iid**, zero mean, finite variance

The System $\leftrightarrow$ Nonparametric Inference $\leftrightarrow$ Dependence Structure

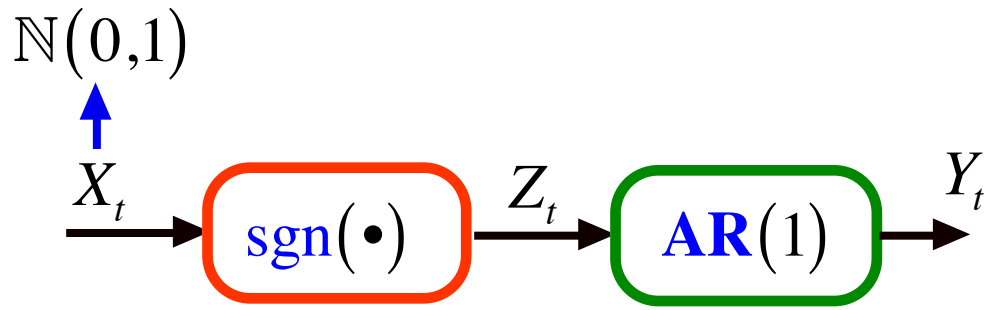
The Annals of Statistics

equivariant estimators are considered: (i) estimators based on kernel methods and (ii) estimators based on k-nearest neighbor kernel methods. Strong consistency of both families is proved under mild conditions.

We obtain strong pointwise consistency of both families of estimates, when the observations (X_t, Y_t) have a φ- or an α-mixing dependence structure [see Billingsley (1968) and Rosenblatt (1956), respectively]. For kernel weights, the

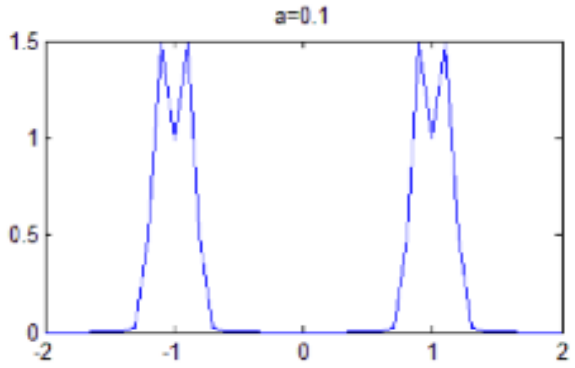
In this paper we go far beyond this classical framework by showing that support vector machines (SVMs) essentially only require that the data-generating process satisfies a certain law of large numbers. We then consider the learnability of SVMs for α -mixing (not necessarily stationary) processes for both classification and regression,….

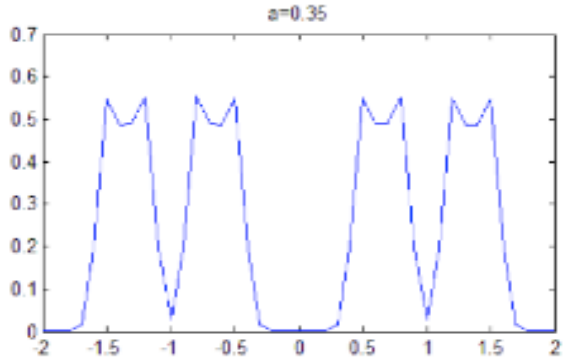
The System $\leftrightarrow$ Dependence Structure

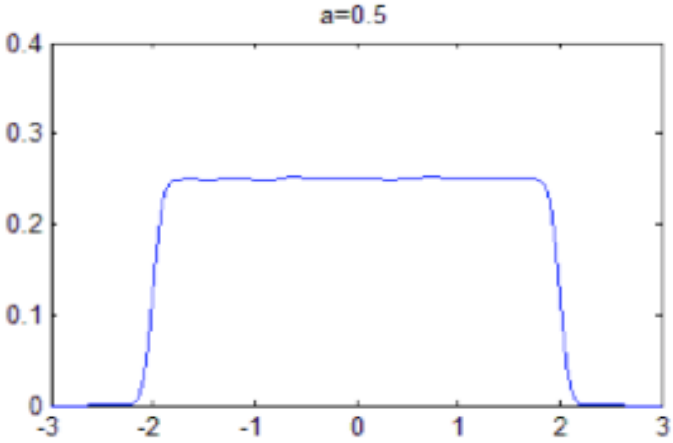


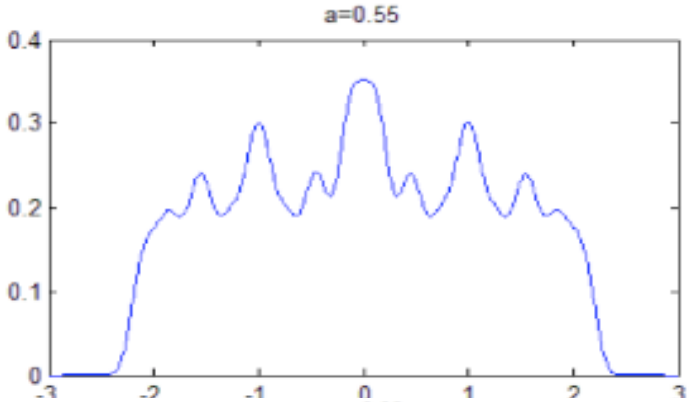
$$Y_t = aY_{t-1} + \underbrace{\mathbf{sgn}\left(X_t\right)}_{\pm 1 \propto \mathbf{pr}\ .1/2}$$

$$Y_t = \sum_{j=0}^{\infty} a^j \left(\pm 1\right)$$

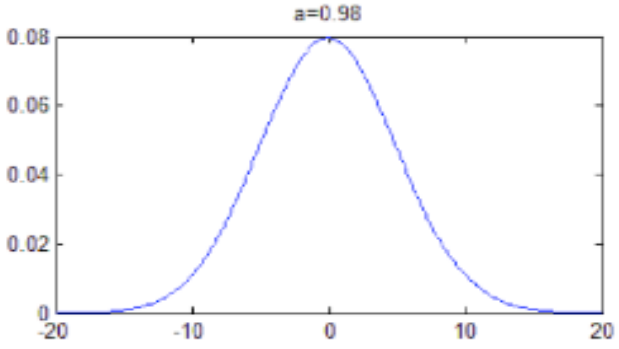








$$\widehat{f}_Y(y) = (nh)^{-1} \sum_{t=1}^{n} \mathbf{K}\left(\frac{y - Y_t}{h} \right)$$



$$Y_t = \sum_{j=0}^{\infty} a^j (\pm 1)$$ - **Bernoulli Convolution**

- **Paul Erdos. On a family of symmetric Bernoulli convolutions. Amer. J. Math., 1939.**

- **Boris Solomyak. On the random series $\sum \pm \lambda^n$ (an Erdos problem). Ann.of Math., 1995.**

- **Peres, Schalg and Solomyak. Sixty years of Bernoulli convolution, 2000.**

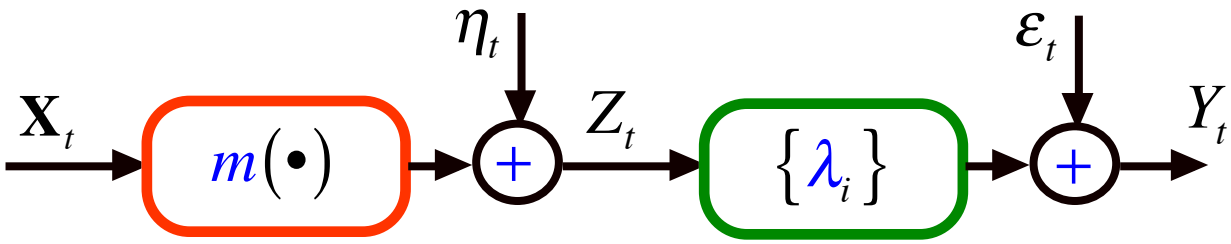
$$Y_t = \sum_{j=0}^{\infty} a^j (\pm 1)$$

- $\{Y_t\}$ **is not mixing for** $a \in (0, 1/2]$

D. W. K. Andrews. Non-strong mixing autoregressive processes. J.Applied Probability, 1984.

- $a \in (0, 1/2) \Rightarrow F_Y$ - **singular**
- $a = 1/2 \Rightarrow F_Y$ - **Uniform**$[-2, 2]$
- $a \in (1/2, 1) \Rightarrow F_Y$ - either **absolutely cont.** or **singular** ?

k-NN Estimates



Wish to estimate

$$m(\mathbf{x}) = E\left[Y_t \mid \mathbf{X}_t = \mathbf{x}\right]$$

from

$$\mathfrak{I}_n = \left\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\right\}$$

Weighted k-NN Estimates

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{k_n} w_{ni} Y_{[i]}$$

$$\sum_{i=1}^{k_n} w_{ni} = 1$$

k-NN Estimates for Dependent Data

S. Yakowitz. Nearest-neighbour methods for time series. J. of Time Series, 1987.

S. Kulkarni and S. Posner. Rates of convergence of nearest neighbour estimation under arbitrary sampling. IEEE Trans. on Inf. Theory, 1995.

G. Boente and R. Fraiman. Robust nonparametric regression estimation for dependent observations. The Annals of Statistics, 1989.

Asymptotic Theory

$$\widehat{m}(\mathbf{x}) = \sum_{i=1}^{k_n} w_{ni} Y_{[i]}$$

$$\sum_{i=1}^{k_n} w_{ni} = 1, \quad w_{ni} \geq 0$$

Theorem 3 (Convergence)

$$k_n \to \infty, \quad \sum_{i=1}^{k_n} w_{ni}^2 \to 0 \quad \Leftarrow \quad \textbf{variance}$$

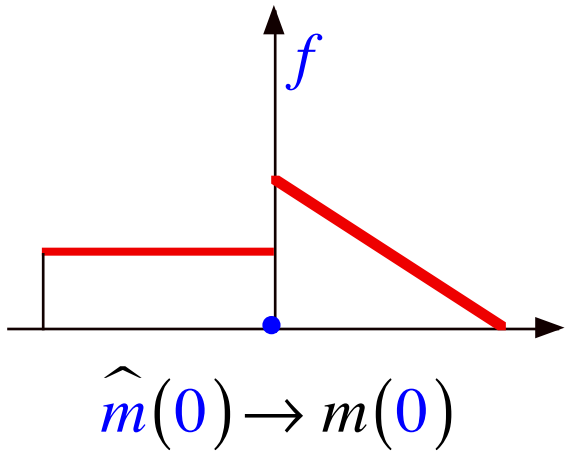
$$\frac{k_n}{n} \to 0 \quad \Leftarrow \quad \textbf{Bias}$$

$$\widehat{m}(\mathbf{x}) \to m(\mathbf{x}) \quad \textbf{(P)}$$

at every point $\mathbf{x} \in \mathbf{S}(f)$ at which $m(\mathbf{x})$ is **continuous**

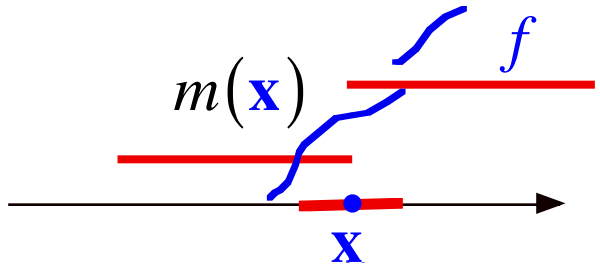
$$\widehat{m}(\mathbf{x}) \to m(\mathbf{x}) \quad (\mathbf{P})$$

at every point $\mathbf{x} \in \mathbf{S}(f)$ at which $m(\mathbf{x})$ is **continuous**



Proof: Ranks

Theorem 4 (MSE)



$$m(\bullet) \in \mathbf{Lip}_{loc}(\mathbf{x})$$

$$\mathbf{E}\left[\hat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2$$

$$\doteq \frac{\mathbf{D}}{n^2}\left[\sum_{i=1}^{k_n} i w_{ni}\right]^2 \qquad \Leftarrow \text{ Bias } (d=1)$$

$$+ \sum_{i=1}^{k_n} w_{ni}^2 \qquad \Leftarrow \text{ Variance}$$

$$+ O\left(n^{-1}\right)$$

Proof:

- $r(k) = \left\| \mathbf{X}_{(k)}(\mathbf{x}) - \mathbf{x} \right\|^2$

- $\mathbf{B}(k) = \int_{\mathbf{B}_{r(k)}(\mathbf{x})} f(\mathbf{v}) d\mathbf{v} \propto \mathbf{Beta}(k, n-k+1)$

- $\sum_{i=1}^{k_n} \sum_{j=1}^{k_n} w_{ni} w_{nj} \mathbf{E}\left[\mathbf{B}(i) \mathbf{B}(j) \right]$

Optimal Estimates ← Optimal Weights + k_n^*

$$\mathbf{E}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2$$

$$\simeq \frac{\mathbf{D}}{n^2}\left[\sum_{i=1}^{k_n} i w_{ni}\right]^2 + \sum_{i=1}^{k_n} w_{ni}^2 \longrightarrow \min_{\{w_{ni}\},\, k_n} \text{ s.t. } \sum_{j=1}^{k_n} w_{ni} = 1$$

$$w_{ni} \geq 0$$

Theorem 5 (Optimal Weights)

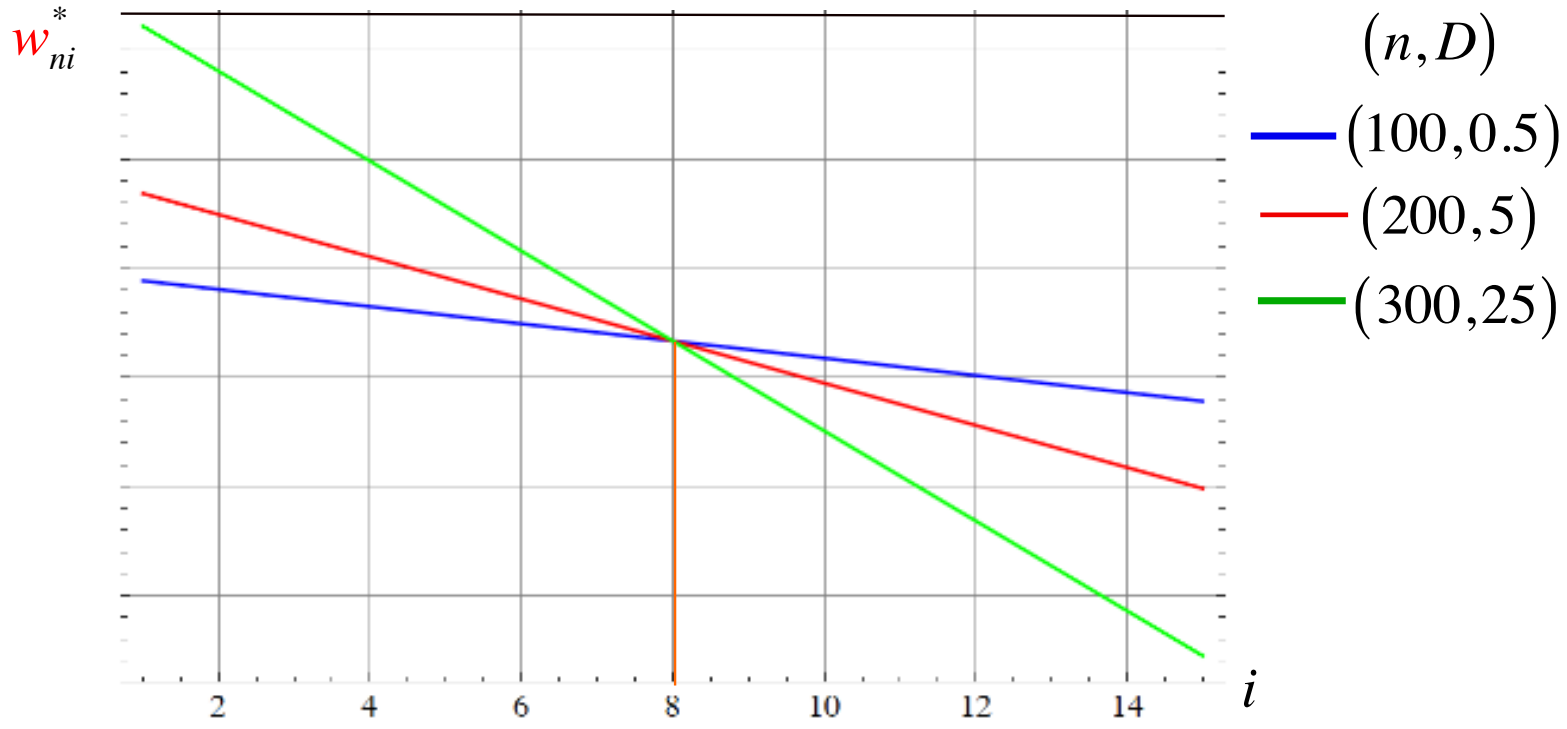
$$w_{ni}^{*} = a_n^{*} - i b_n^{*},\ i = 1,\ldots,k_n^{*}$$

$$a_n^{*} = \frac{1}{k_n^{*}} + \frac{\mathbf{D}}{2}\left(k_n^{*}+1\right)\Delta_n^{*} \quad , \quad b_n^{*} = \mathbf{D}\Delta_n^{*}$$

$$\Delta_n^{*} = \frac{6\left(k_n^{*}+1\right)}{12n^2 + \mathbf{D}\left(k_n^{*}-1\right)k_n^{*}\left(k_n^{*}+1\right)}$$

$$k_n^{*} = \left\lceil \left(\frac{6}{\mathbf{D}}\right)^{1/3} n^{2/3} \right\rceil$$

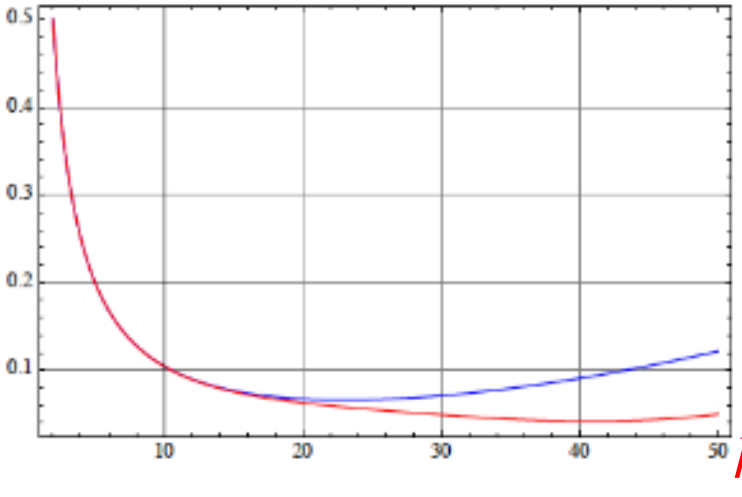
Optimal Linear Weights



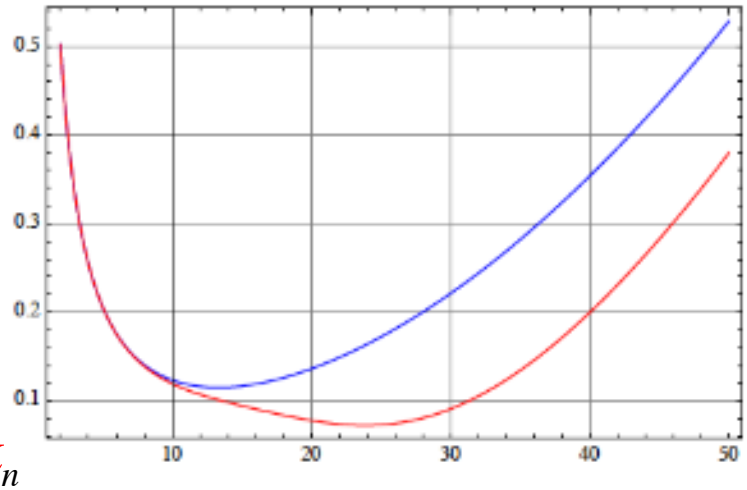
$$i^* = \frac{k+1}{2}$$

Optimal Weights $\leftrightarrow$ Uniform Weights

$\mathbf{D} = 1$



$\mathbf{D} = 5$



$n = 80$

$\mathbf{MSE}\left(w_{ni}^{*}\right)$ ——

$\mathbf{MSE}\left(\frac{1}{k_n}\right)$ ——

Theorem 6 (Rate of Convergence)

$$\text{Let} \quad k_n^* = \left\lceil \left(\frac{6}{\mathbf{D}} \right)^{1/3} n^{2/3} \right\rceil \text{ and } w_{ni}^* \uparrow$$

$$\mathbf{E}\left[\hat{m}(\mathbf{x}) - m(\mathbf{x}) \right]^2 = O\left(n^{-2/3} \right)$$

This rate can be improved for smoother characteristics (weights < 0)

$$O\left(n^{-2/3}\right) \to O\left(n^{-4/5}\right) \to O\left(n^{-6/9}\right) \to .... \to O\left(n^{-\frac{2s}{2s+1}} \right) \to$$

smoothness →

Theorem 6 (Rate of Convergence)

Let $k_n^* = \left\lceil \left(\frac{6}{\mathbf{D}} \right)^{1/3} n^{2/3} \right\rceil$ and w_{ni}^* $\uparrow$

$$\mathbf{E}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x}) \right]^2 = O\left(n^{-2/3} \right)$$

The rate $O\left(n^{-2/3} \right)$ is optimal

$$n^{2/3}\mathbf{E}\left[\widehat{m}(0) - m(0) \right]^2 \to c > 0$$

Proof:

- $\mathbf{E}\left[\sum_{i=1}^{k_n}\mathbf{B}(i)\right]^2$
- $\mathbf{E}\left[\sum_{i=1}^{k_n}\sum_{j=1}^{k_n} j\mathbf{B}(i)\mathbf{B}(j)\right]$
- $\mathbf{E}\left[\sum_{j=1}^{k_n} i\mathbf{B}(i)\right]^2$

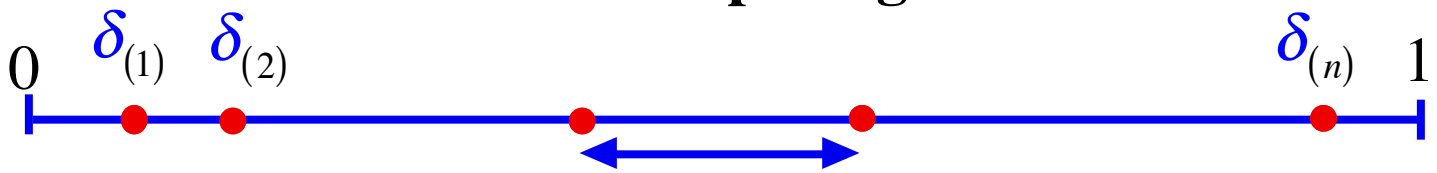
Theorem 7 **(Uniform Rate)** $m(\bullet) \in \mathbf{Lip}(\mathbb{R})$

$$\text{Let} \quad k_n^* \simeq \left\lceil \frac{n^{2/3}}{\log^{2/3} n} \right\rceil \text{ and } \left\{ w_{ni}^* \right\} \quad \uparrow$$

$$\sup_{x \in S(f)} \left| \widehat{m}(x) - m(x) \right| = O_P\left(\frac{\log^{1/3} n}{n^{1/3}} \right)$$

Proof:

Uniform Spacings



$$\mathbf{D}_n = \max_{0 \le i \le n} \left\{ \delta_{(i+1)} - \delta_{(i)} \right\}$$

$$\mathbf{D}_n = \frac{\log n}{n} + O\left(\frac{\log \log n}{n} \right) \quad \textbf{a.s}$$

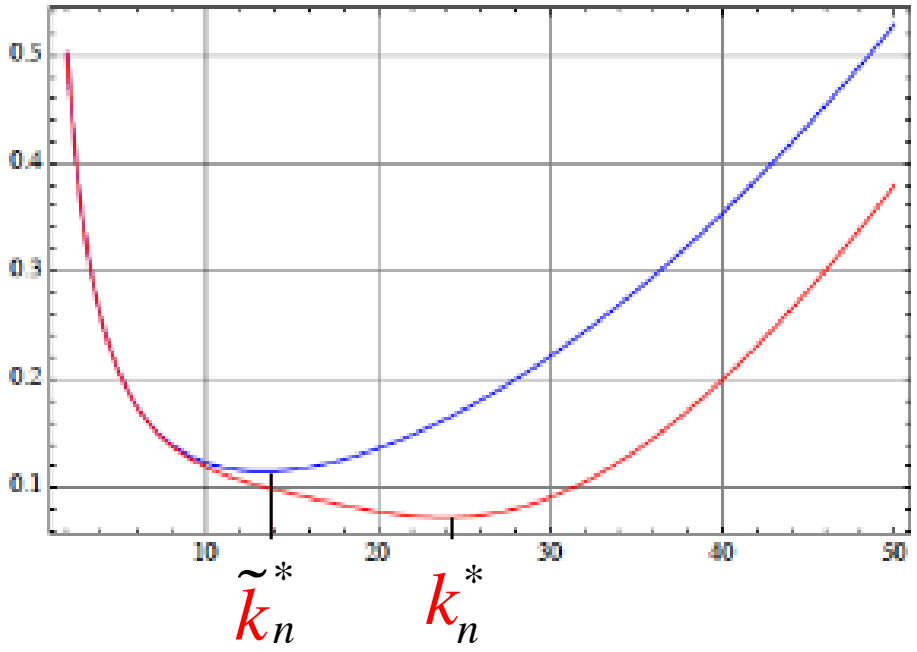
Slud, Devroye

- **Universal relationship for k_n^***

Optimal Linear Weights: $k_n^* = \left\lceil \left(\frac{6}{\mathbf{D}} \right)^{1/3} n^{2/3} \right\rceil$

Uniform Weights: $\tilde{k}_n^* = \left\lceil \left(\frac{2}{\mathbf{D}} \right)^{1/3} n^{2/3} \right\rceil$

$$\frac{k_n^*}{\tilde{k}_n^*} = 3^{1/3} = 1.44...$$



Apply CV for $\widetilde{m}(\mathbf{x}) = \dfrac{1}{k_n}\sum_{i=1}^{k_n} Y_{[i]} \Rightarrow \tilde{k}_{CV} \Rightarrow k_{CV} = 3^{1/3}\tilde{k}_{CV}$

- **Kernel Weights:** $\widehat{m}(x) = \sum_{i=1}^{k_n} v_{ni} Y_{[i]}$

$$v_{ni} = \int_{(i-1)/k_n}^{i/k_n} p(t)\,dt\,,\ i = 1,\ldots.,k_n$$

$\uparrow$ **pdf on** $[0,1]$

$$\mathbf{E}\left[\widehat{m}(\mathbf{x}) - m(\mathbf{x})\right]^2 \simeq \frac{\mathbf{D}}{n^2}\left[\sum_{i=1}^{k_n} i v_{ni}\right]^2 + \sum_{i=1}^{k_n} v_{ni}^2$$

Kernel Weights

- $$\frac{1}{k_n}\sum_{i=1}^{k_n} i v_{ni} = \int_0^1 t p(t)\,dt \left[1 + O\left(\frac{1}{k_n}\right)\right]$$

- $$k_n \sum_{i=1}^{k_n} v_{ni}^2 = \int_0^1 p^2(t)\,dt \left[1 + O\left(\frac{1}{k_n}\right)\right]$$

Kernel Weights

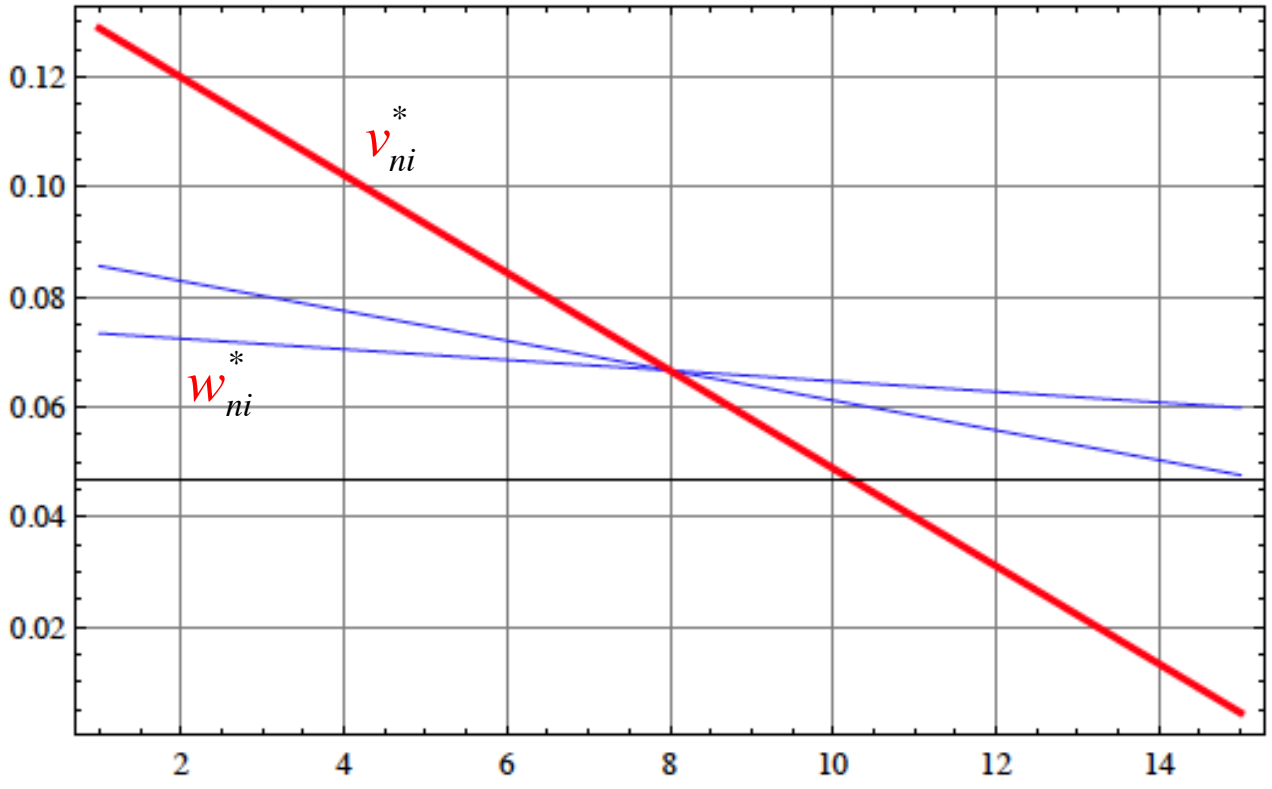
$$\mathbf{E}\left[\widehat{m}(\mathbf{x})-m(\mathbf{x})\right]^2 \simeq \frac{\mathbf{D}}{n^2}\left[\sum_{i=1}^{k_n} i v_{ni}\right]^2 + \sum_{i=1}^{k_n} v_{ni}^2 \longrightarrow \min_{p(\bullet),k_n}$$

$$\mathbf{D}\left[\frac{k_n}{n}\int_0^1 tp(t)dt\right]^2 + \frac{1}{k_n}\int_0^1 p^2(t)dt \longrightarrow \min_{p(\bullet),k_n}$$

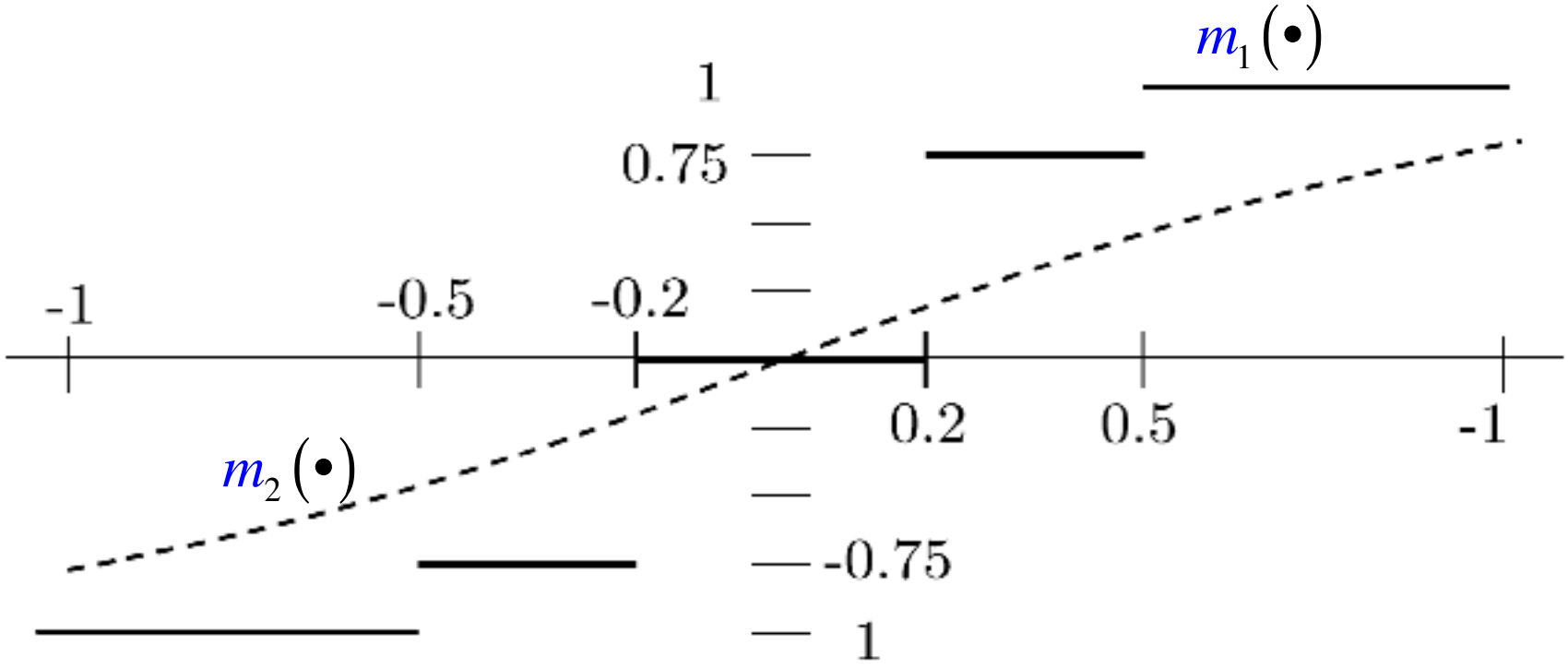
Optimal Kernel Weights

$$v_{ni}^{*} = \frac{1}{k_n^{*}} + \frac{k_n^{*}+1}{k_n^{*2}} - i\frac{2}{k_n^{*2}},$$

$$k_n^{*} = \left\lceil \left(\frac{6}{\mathbf{D}}\right)^{1/3} n^{2/3} \right\rceil \Leftarrow \text{the same as for } w_{ni}^{*}$$


0.12
0.10
0.08
0.06
0.04
0.02
2
4
6
8
10
12
14
v_{ni}^{*}
w_{ni}^{*}

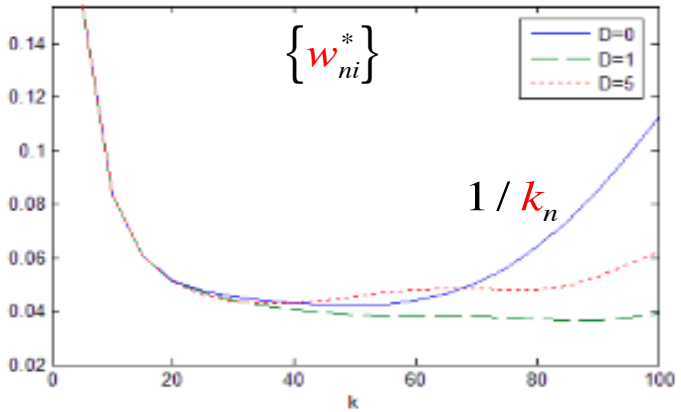
Finite n

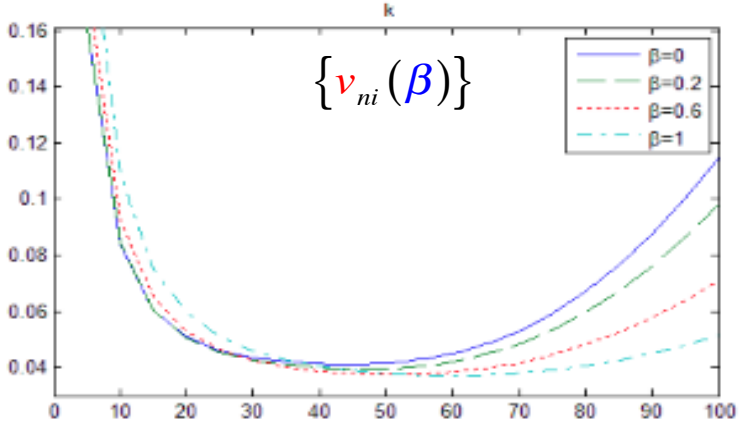


Ex. 1: MISE versus k_n

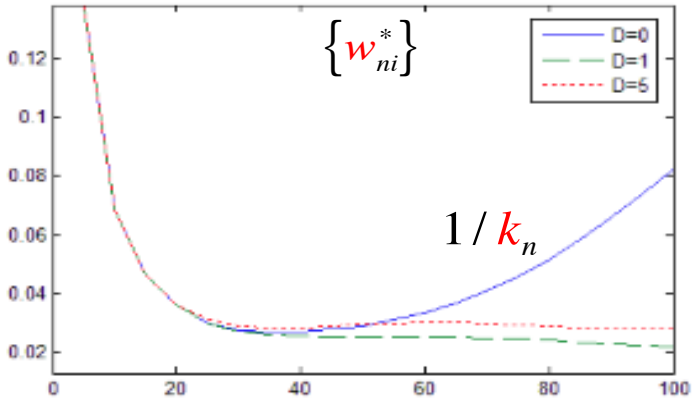
$n = 200$

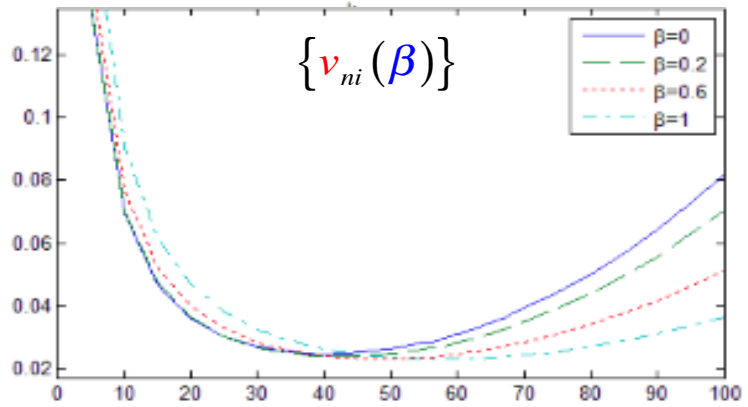
$m_1(\bullet)$





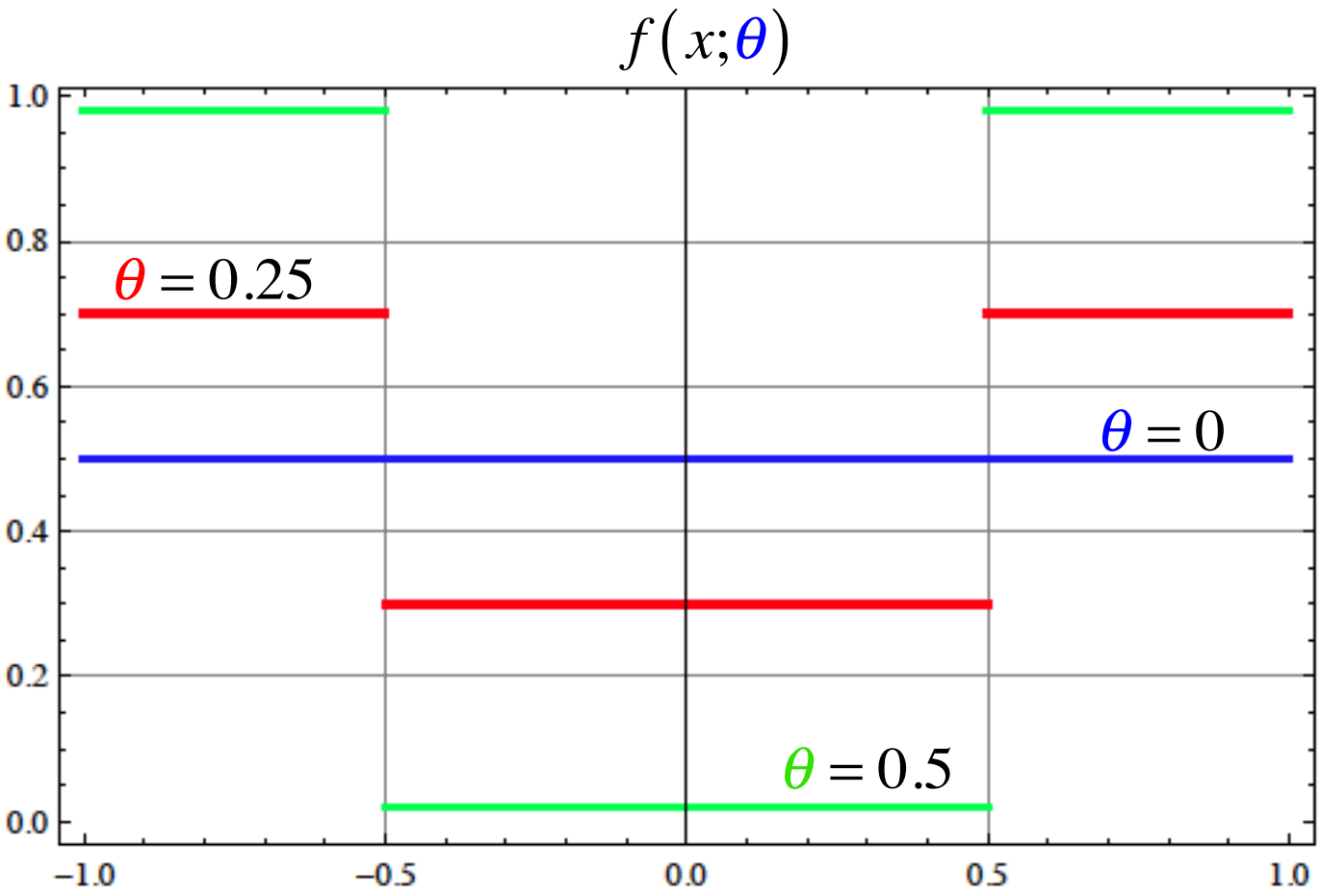
$m_2(\bullet)$



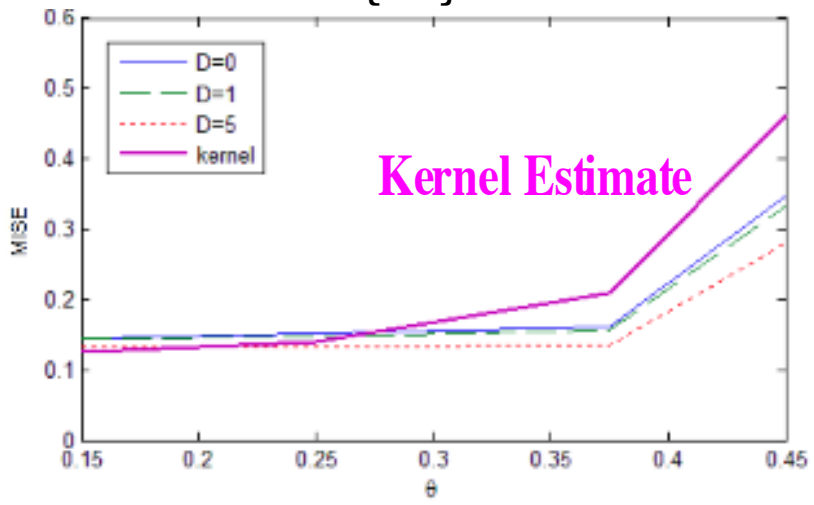


$$v_{ni}(\beta) = \frac{1}{k_n^*} + \beta\frac{k_n^*+1}{k_n^{*2}} - i\frac{2\beta}{k_n^{*2}} \quad \Rightarrow \quad v_{ni}(1) = v_{ni}^* \qquad v_{ni}(0) = 1/k_n$$

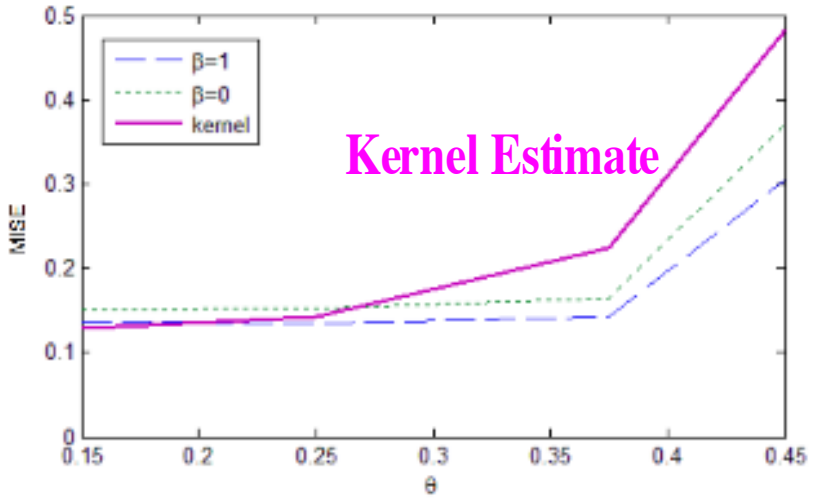
Ex. 2: MISE $\leftrightarrow$ The Input Density



$\{w_{ni}^*\}$



$\{v_{ni}(\beta)\}$

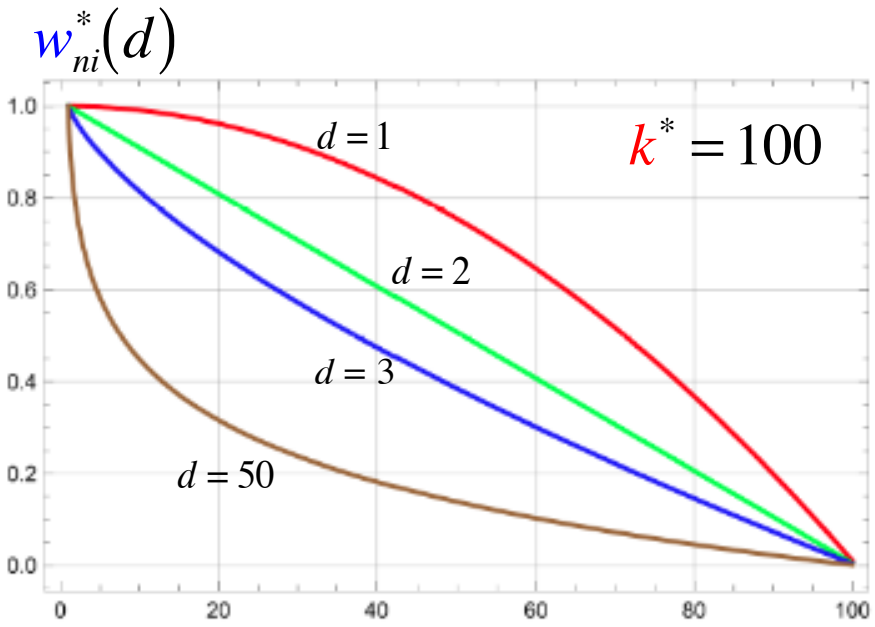


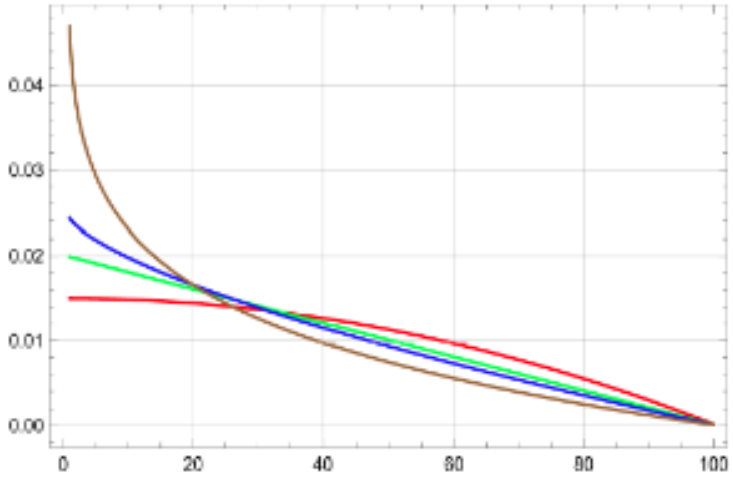
$$v_{ni}(\beta)=\frac{1}{k_n^*}+\beta\frac{k_n^*+1}{k_n^{*2}}-i\frac{2\beta}{k_n^{*2}}$$

$$\Rightarrow \quad v_{ni}(1)=v_{ni}^*$$

$$v_{ni}(0)=1/k_n$$

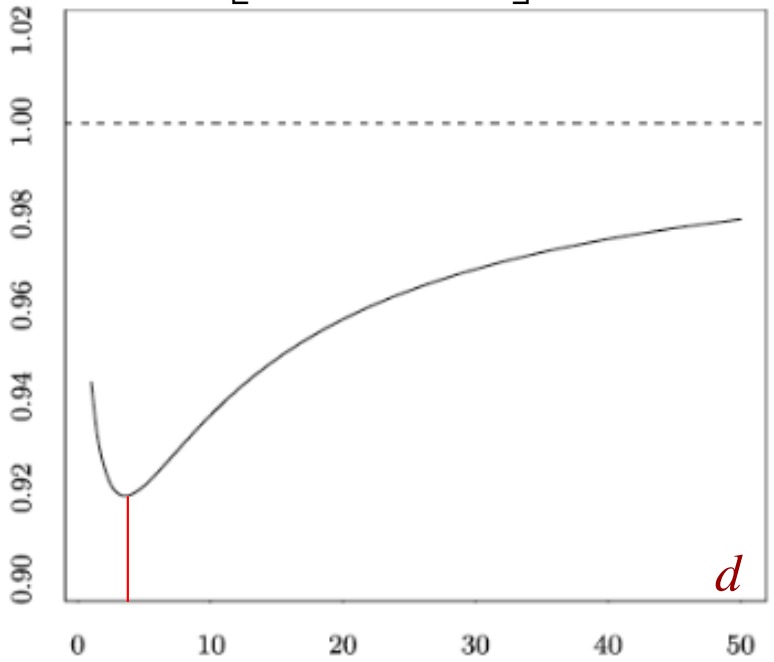
Optimal Positive Weights in $\mathbf{R}^d$: $w_{ni}^*(d) \Leftarrow m(\bullet) \in \mathbf{C}^2(\mathbf{R}^d)$





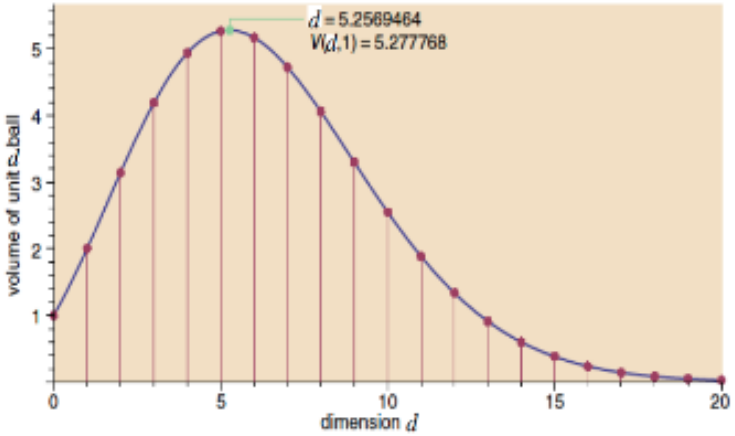
unweighted → $\mathbf{E}\left[\widetilde{m}(\mathbf{X}) - m(\mathbf{X})\right]^2$

weighted → $\mathbf{E}\left[\widehat{m}(\mathbf{X}) - m(\mathbf{X})\right]^2$



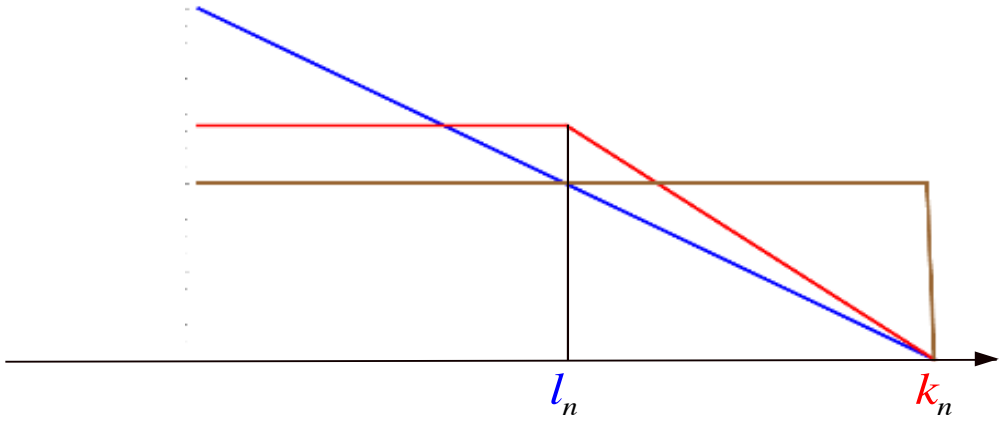
$4 \leftrightarrow d \leftrightarrow 5$

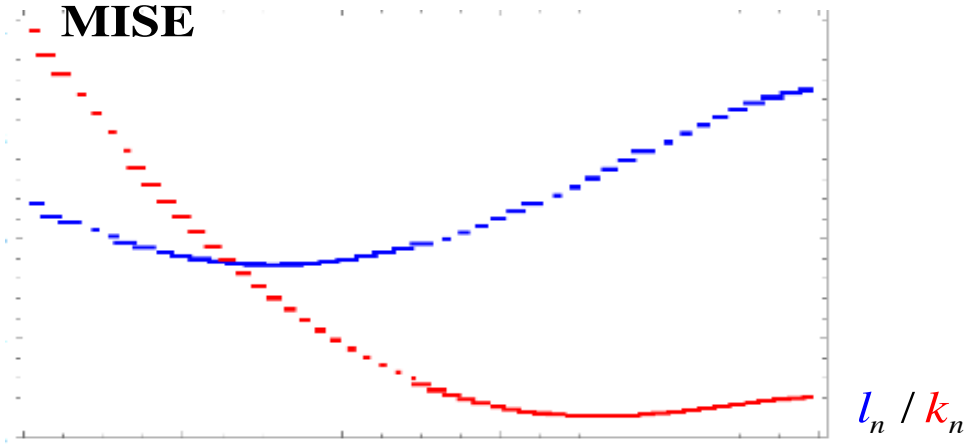
Volume of Unit Ball



➤ **R. J. Samworth. Optimal weighted nearest neighbour classifiers. Ann. of Stat., 2012.**

L_∞ Weights $\longrightarrow$ (k_n, l_n) - NN rules





- **Negative Weights $\Leftarrow$ Smooth $m(\bullet)$**

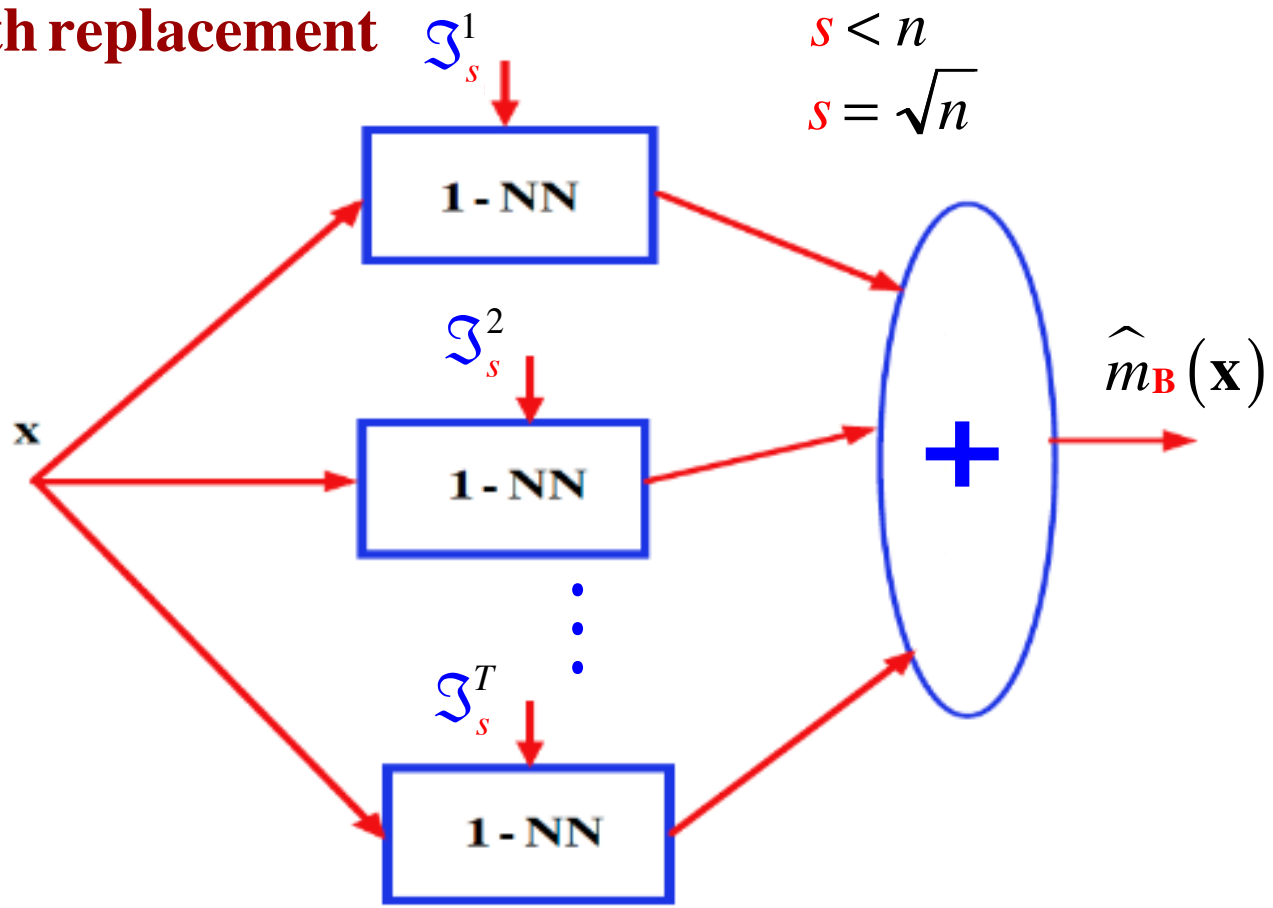
$$\sum_{i=1}^{k_n} i^2 w_{ni} = 0 \quad \longrightarrow \textbf{ better rates}$$

- **The rates can be improved for smoother characteristics**

- **No smoothness of the input density is needed**

k-NN via 1-NN estimate Aggregation: Bagging Weights

resampling with replacement



- $$\widehat{m}_B(\mathbf{x};s)=\frac{m_{NN}\left(\mathbf{x};\Im_s^1\right)+\cdots+m_{NN}\left(\mathbf{x};\Im_s^T\right)}{T}$$

- **For large T the average estimate $\widehat{m}_B(\mathbf{x};s)$ is equivalent to the weighted NN estimate**

$$\widehat{m}_B(\mathbf{x};s)\approx\sum_{i=1}^{n}b_{ni}Y_i$$

$$b_{ni}=\left(1-\frac{i-1}{n}\right)^s-\left(1-\frac{i}{n}\right)^s\approx\tau(1-\tau)^{i-1},\quad\tau=\frac{s}{n}$$

- **If**

$$s_n \to \infty \text{ and } \frac{s_n}{n} \to 0$$

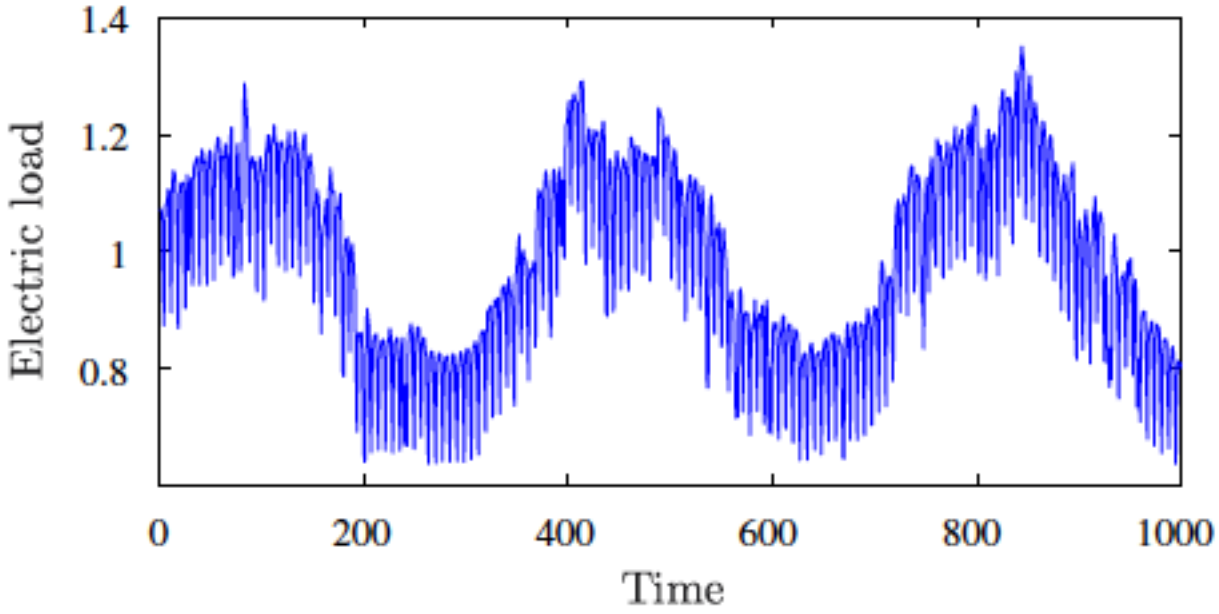
then $\hat{m}_B(\mathbf{x}; s)$ **is a consistent estimate of** $m(\mathbf{x})$

- **Hall & Samworth. Properties of bagged NN classifiers J.Royal. Stat. Soc. 2005.**

- **Biau,…, Journal of Machine Learning Res. , 2010**

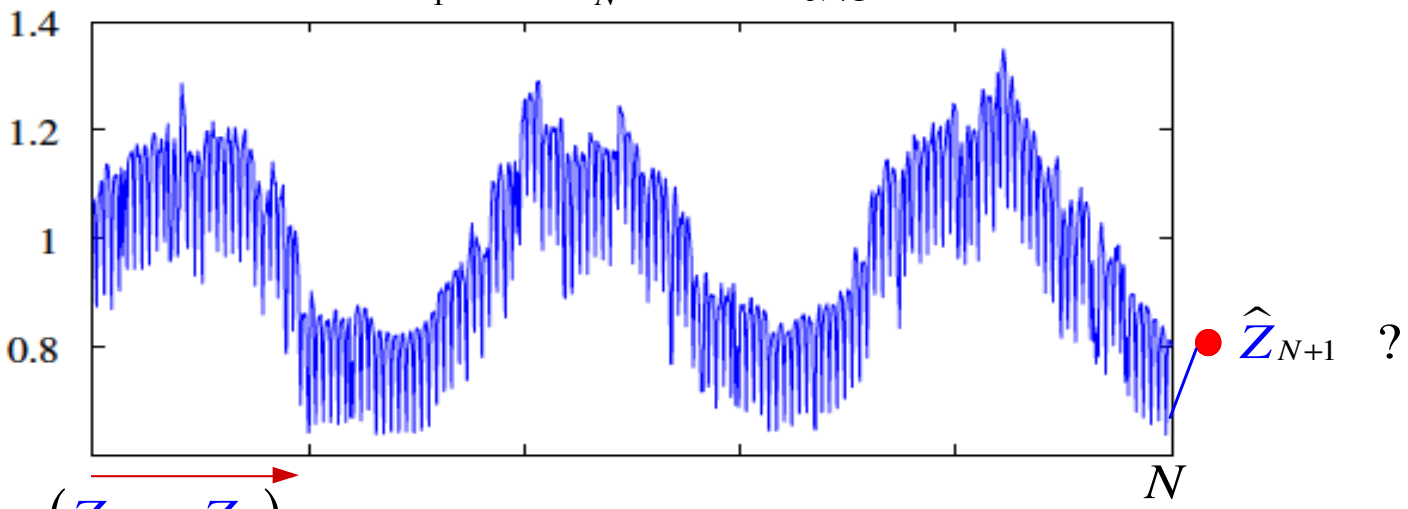
IV k-NN Regression Estimates: Time Series Forecasting

Time Series Forecasting



Poland Electricity Load: 1000 days

$Z_1, \ldots\ldots, Z_N \Rightarrow \hat{Z}_{N+1}$



$\hat{Z}_{N+1}$?

$$1 \begin{cases} \mathbf{X}_1 = (Z_1, \ldots, Z_d) \\ Y_1 = Z_{d+1} \end{cases}$$

$$2 \begin{cases} \mathbf{X}_2 = (Z_2, \ldots, Z_{d+1}) \\ Y_2 = Z_{d+2} \end{cases}$$

$$n = N - d \begin{cases} \mathbf{X}_n = (Z_n, \ldots, Z_{n+d-1}) \\ Y_n = Z_{n+d} \end{cases}$$

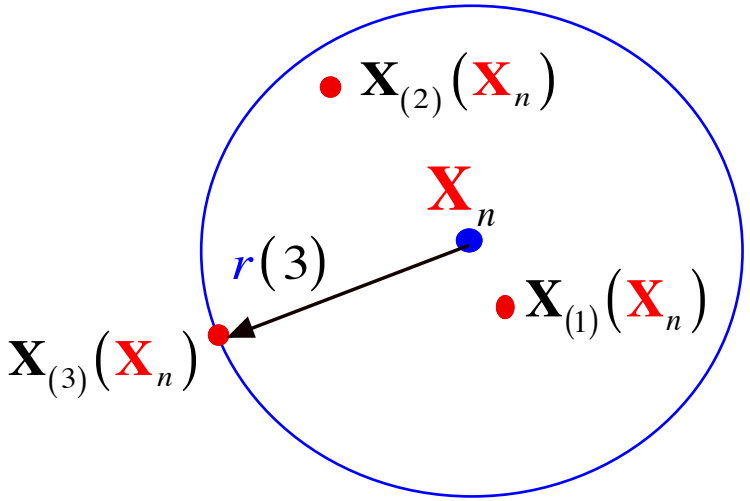
- **The original time series is transformed into the d-dimensional regression type data for prediction of 1-unit of time in the future**

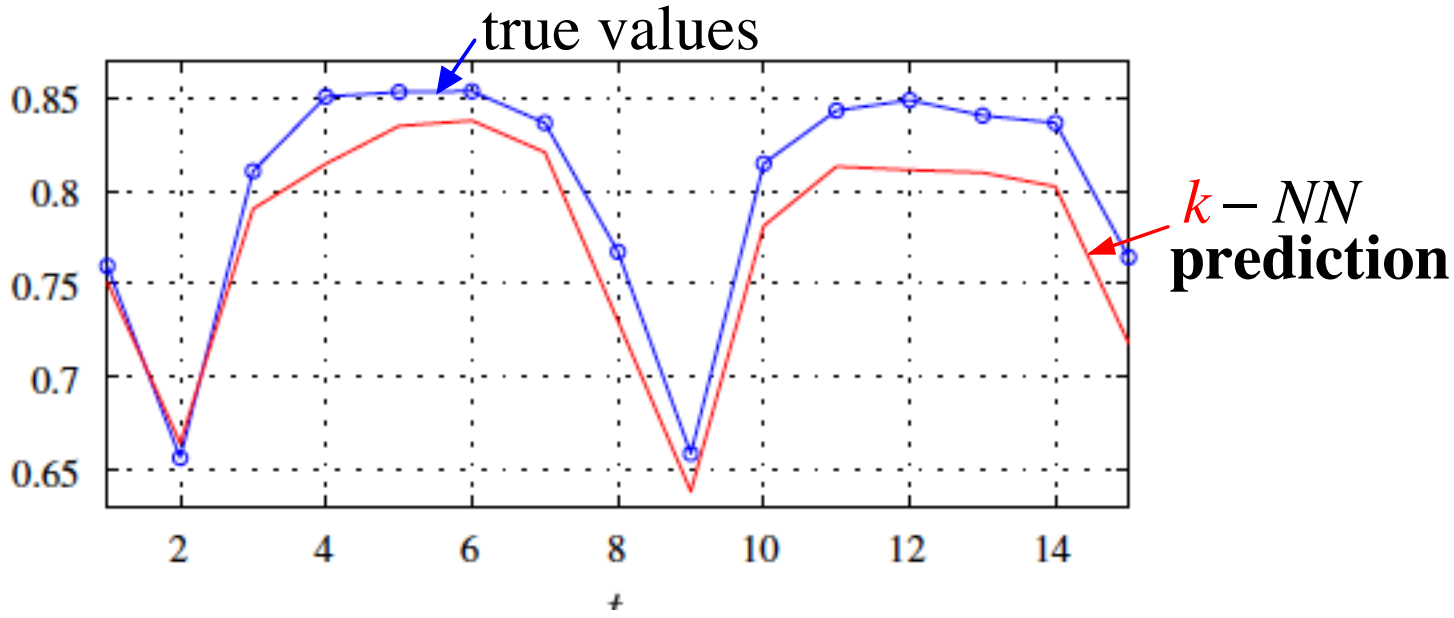
$$Z_1,\ldots,Z_N \quad \rightarrow \quad \{(\mathbf{X}_1,Y_1),\cdots,(\mathbf{X}_n,Y_n)\}$$

$$(\mathbf{X}_t,Y_t)\in \mathbf{R}^d\times\mathbf{R}$$

- **To predict Z_{N+1} we evaluate the k-NN regression estimate at the most recent point $\mathbf{X}_n$**

$$\hat{Z}_{N+1} = \frac{1}{k} \sum_{\{t<n:\mathbf{X}_t \in B_{r(k)}(\mathbf{X}_n)\}} Y_t \quad , \quad \hat{Z}_{N+1} = \sum_{\{t<n:\mathbf{X}_t \in B_{r(k)}(\mathbf{X}_n)\}} w_t Y_t$$




true values
0.85
0.8
0.75
0.7
0.65
2
4
6
8
10
12
14
t
k – NN
prediction

Asymptotic Theory

Let $\{Z_t\}$ be a stationary (α- mixing) stochastic process.

Then for $\hat{Z}_{N+1} = \frac{1}{k} \sum_{\{i:\mathbf{X}_i \in B_{r(k)}(\mathbf{X}_n)\}} Y_i$ with $k_n = cn^{2/(2+d)}$ we have

$$\mathbf{E}\left[\hat{Z}_{N+1} - Z^*_{N+1}\right]^2 = O\left(n^{-\frac{2}{2+d}}\right) \leftarrow \textbf{optimal rate}$$

$$\uparrow$$

$$Z^*_{N+1} = \mathbf{E}\left[Z_{N+1} \mid Z_N, \ldots, Z_{N-d+1}\right] \textbf{- optimal predictor}$$

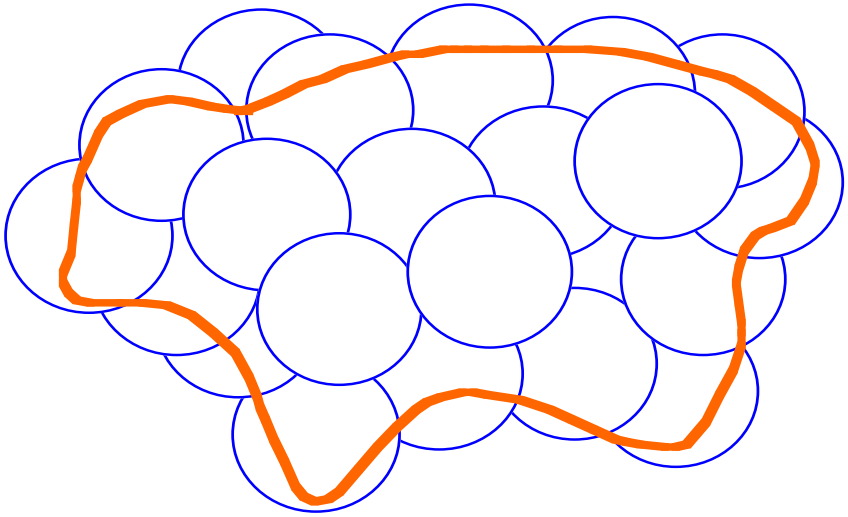
1-NN distance for stationary process

$$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbf{A} \subset R^d; \quad \mathbf{X}_{n+1} \in \mathbf{A}$$

	$\mathbf{E}\left\|\mathbf{X}_{n+1} - \mathbf{X}_{(1)}(\mathbf{X}_{n+1})\right\|^2$	$\mathbf{E}\left\|\mathbf{X}_{n+1} - \mathbf{X}_{(k)}(\mathbf{X}_{n+1})\right\|^2$
$d=1$	$c(1)\dfrac{1+\log(n)}{n}$	$c(1)\dfrac{k}{n}\left[1+\log\left(\dfrac{n}{k}\right)\right]$
$d=2$	$c(2)\dfrac{1}{n}$	$c(2)\dfrac{k}{n}$
$d \geq 3$	$c(d)\left(\dfrac{1}{n}\right)^{2/d}$	$c(d)\left(\dfrac{k}{n}\right)^{2/d}$

Proof by the theory of covering numbers

$\varepsilon_{\min}(r;\mathbf{A}) =$ the smallest radius such there exist r balls of this radius which cover the set $\mathbf{A}$



$$\mathbf{E}\left\|\mathbf{X}_{n+1} - \mathbf{X}_{(1)}(\mathbf{X}_{n+1})\right\|^2 \le \frac{8}{n}\sum_{j=1}^{n}\left(\varepsilon_{\min}(j;\mathbf{A})\right)^2$$

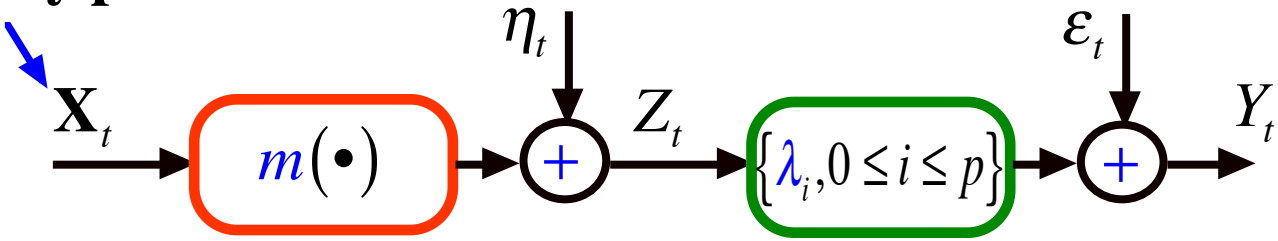
If $\mathbf{A} \subset \mathbf{R}^d$ **then** $\varepsilon_{\min}(r;\mathbf{A}) \le \varepsilon_{\min}(1;\mathbf{A})\, r^{-1/d}$

➤ **Kulkarni & Posner - IT - 1995**

V Extensions and Concluding Remarks

Input-Output Dependent System

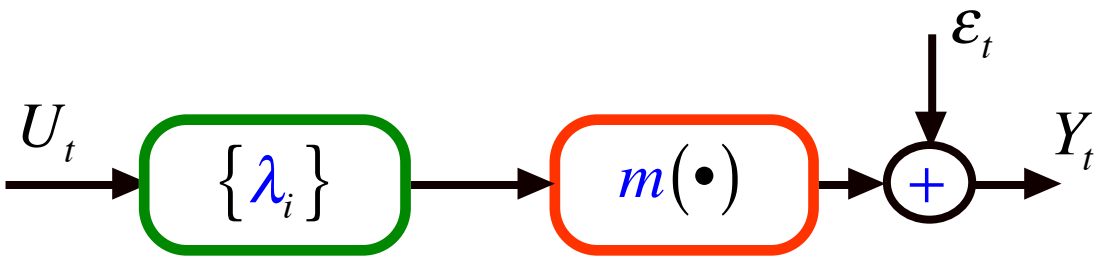
stationary process



$$m(\mathbf{x}) = \sum_{l=0}^{p} a_l(\lambda) \underbrace{r_l(\mathbf{x})}_{\mathbf{E}[Y_t | \mathbf{X}_{t-l} = \mathbf{x}]}$$

Additive Solution in Terms of Lagged Regression Functions

Other Time Series Systems $\Rightarrow$ Wiener System

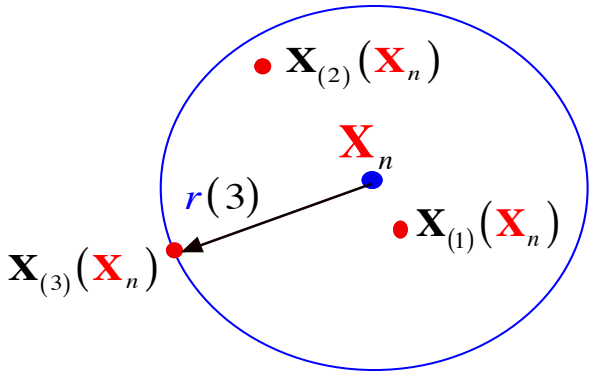


$$Y_t = m\left(\sum_{i=0}^{\infty} \lambda_i U_{t-i}\right) + \varepsilon_t$$

Dynamical Version of Single Index Model

k-NN predictor with a distance correlation approach

$$\hat{Z}_{N+1} = \sum_{\{t<n:\mathbf{X}_t \in B_{r(k)}(\mathbf{X}_n)\}} w_t Y_t$$



$\rho(\mathbf{X}_n;\mathbf{X}_t)$ = **correlation distance between** $\mathbf{X}_n$ **and** $\mathbf{X}_t$

Szekely, Rizzo,…. Measuring and testing independence by correlation of distances; Annals of Stat. 2007.

References

- **W. Greblicki & M.Pawlak: IEEE Trans. Inf. Theory**, 2017

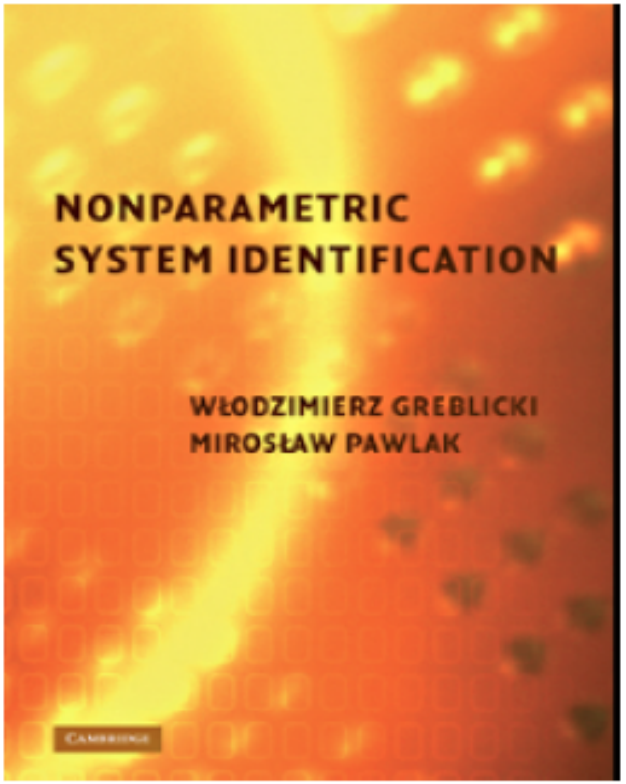
  $\Rightarrow$ **k-NN for nonlinear time series systems**

- **W. Greblicki & M.Pawlak: IEEE Trans. Automatic Control**, 2018

  $\Rightarrow$ **weighted k-NN for nonlinear time series systems**

- **M.Pawlak: IEEE Trans. Inf. Theory**, 2018

$\Rightarrow$ **nonparametric model checks for nonlinear time series systems**


NONPARAMETRIC
SYSTEM IDENTIFICATION
WŁODZIMIERZ GREBLICKI
MIROSŁAW PAWLAK
CAMBRIDGE

Thank you all very much for your time!

"Remember, the other team is using Machine Learning on your games to predict your play. So, kick the ball with your other foot!"