System Identification, Lecture 11

Kristiaan Pelckmans (IT/UU, 2338)

Course code: 1RT880, Report code: 61800 - Spring 2013
F, FRI Uppsala University, Information Technology

16 Mai 2013

Overview Part II

1. State Space Systems.
2. Subspace Identification.
3. Further Topics.
4. Identification of Nonlinear Models.
5. Wider View.

Overview Identification of Nonlinear Models

1. Taxonomy.
2. Nonlinear Dynamic Models.
3. Nonlinear Approximation.

Nonlinear Dynamic Models

Definition 1. [Linear Superposition Principle (LSP)] *Let $\mathcal{S}: \mathbb{R}^n \to \mathbb{R}^n$ be a system. It satisfies the linear superposition principle iff for any inputs $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^n$, one has that*

$$\mathcal{S}(\mathbf{u}) + \mathcal{S}(\mathbf{u}') = \mathcal{S}(\mathbf{u} + \mathbf{u}')$$

Deviations from LSP ('nonlinear'):

- Time-varying.
- Dynamics depending on inputs (operation regime).
- Saturation, Quantization, Hysteresis, Threshold effects, Limit Cycles.

Nonlinear Models

LTI:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{u}_t \\ \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t \end{cases}$$

Parameter Varying:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}_t\mathbf{x}_t + \mathbf{B}_t\mathbf{u}_t \\ \mathbf{y}_t = \mathbf{C}_t\mathbf{x}_t + \mathbf{D}_t\mathbf{u}_t \end{cases}$$

Bilinear:

$$\begin{cases} \mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}(\mathbf{x}_t \times \mathbf{u}_t) \\ \mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{D}\mathbf{u}_t \end{cases}$$

Nonlinear:

$$\begin{cases} \mathbf{x}_{t+1} = g(\mathbf{x}_t, \mathbf{u}_t) \\ \mathbf{y}_t = h(\mathbf{x}_t, \mathbf{u}_t) \end{cases}$$

Different places to put noise in.

Block-Structured Nonlinear Models

Compromise between Flexibility and Insight.

- Hammerstein models.
- Wiener models.
- Hammerstein-Wiener Models.
- Wiener - Hammerstein Models.
- Volterra Models.

Predictor Models

Optimal Predictor (PEM) for LTI:

- In general model

$$y_{t+1} = H(q^{-1}, \theta_0)u_{t+1} + G(q^{-1}, \theta_0)e_{t+1}$$

where $H(q^{-1}, \theta_0) = 1 + h_1 q^{-1} + \dots$ and $G(q^{-1}, \theta_0) = 1 + g_1 q^{-1} + \dots$.

- Rewrite as optimal predictor

$$\hat{y}_{t+1|t} = L_1(q^{-1}, \theta_0)u_{t+1} + L_2(q^{-1}, \theta_0)y_{t+1}$$

where $L_2(1, \cdot) = 0$.

- Optimal predictor

$$\hat{y}_{t+1|t} = \big(G^{-1}(q^{-1}, \theta_0)H(q^{-1}, \theta_0)\big)u_{t+1} + \big(1 - G^{-1}(q^{-1}, \theta_0)\big)y_{t+}$$

But this does not work in general:

- H, G?
- Monic G?
- Invertible?
- Convolution?

$\Rightarrow$ no general optimal predictor corresponding to nonlinear models.

Trick: formulate *predictor model*:

$$y_{t|t-1} = f_\theta(z_t)$$

with

$$z_t = (u_t, \ldots, u_{t-n}, y_{t-1}, \ldots, y_{t-d'})$$

But description of dynamics?

Model estimation:

$$\min_{\theta \in \Theta} \sum_{t=k'}^{n} \ell\left(y_t - f_\theta(z_t)\right)$$

where

- $'f_\theta'$ is a nonlinear function with unknowns θ.
- $\Theta = \{\theta\}$ set of plausible values.
- Loss function $\ell : \mathbb{R} \to \mathbb{R}$.

Perspectives:

1. Algorithmic.
2. Representation.
3. Convergence.
4. Inference.

$\rightarrow$ Model selection.

Choice of Models

Bias - variance decomposition:

$$\mathbb{E}\|f_* - f_{\hat{\theta}}\|^2 = \|f_* - f_{\theta_*}\|^2 + \mathbb{E}\|f_{\theta_*} - f_{\hat{\theta}}\|^2$$

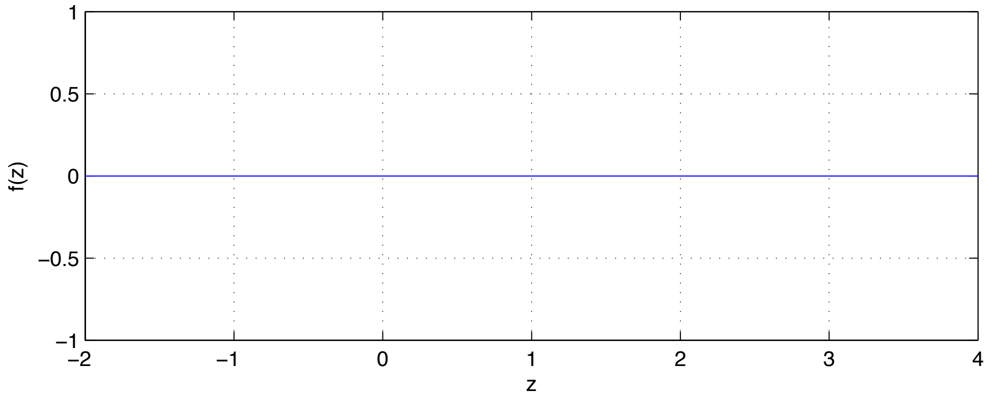
where:

- f_{θ_*} equals best one could do in $\{f_\theta : \theta \in \Theta\}$.
- $\|f_* - f_{\theta_*}\|^2$ equals bias^2, proportional to 'form' Θ.
- $\mathbb{E}\|f_{\theta_*} - f_{\hat{\theta}}\|^2$ variance proportional to size Θ.

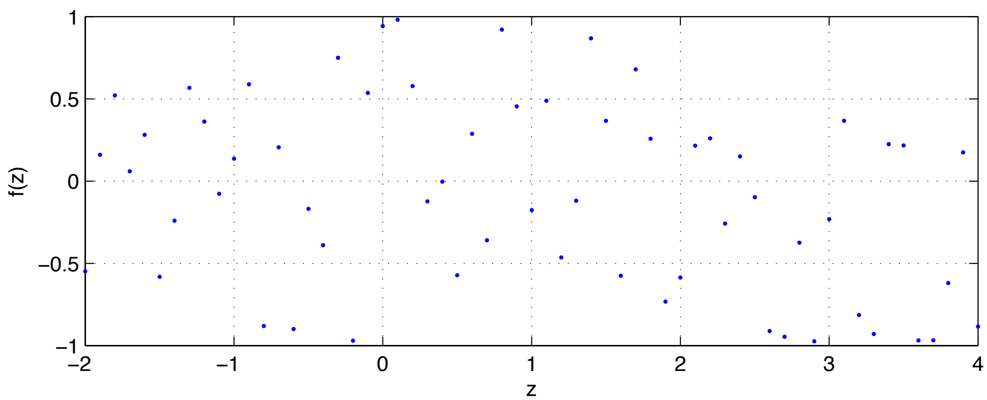
So choose a f_θ and Θ which can be expected to trade bias and variance optimally. Extrema:

- Constant function $f(z) = 0$.
- Lookup table with infinite number of entries $\theta = \{(z, y)\}$.

Constant $f(z) = 0$:



Lookup table with 60 entries $\{(-2, 0.1), \ldots, (4, 0.15)\}$:



General Approximators: Basis Functions

Abstract into 1D: $f_* : \mathbb{R} \to \mathbb{R}$ approximated by $f_\theta : \mathbb{R} \to \mathbb{R}$. Given set $\{\phi_i : \mathbb{R} \to \mathbb{R}\}$, assume the function

$$f_\theta(x) = \sum_{i=1}^{m} \theta_i \phi_i(x)$$

- Linear in the parameters θ! So

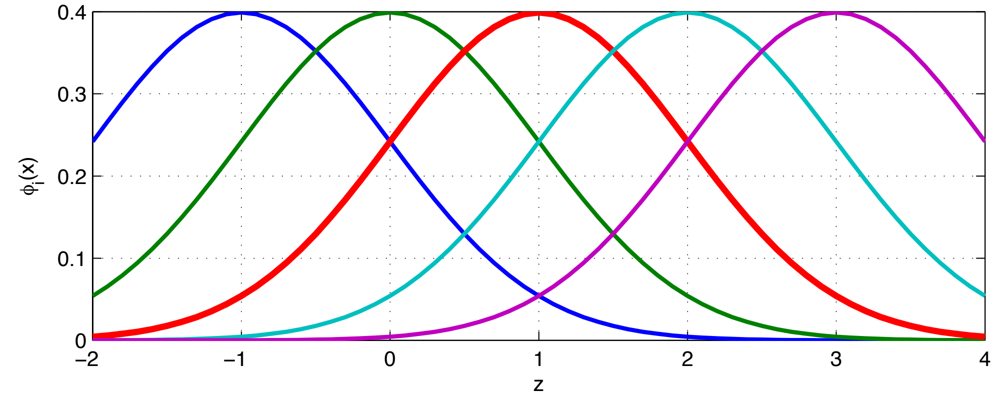
$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^m} \sum_{t=k'}^{n} \left(y_t - f_\theta(z_t)\right)^2$$

Analytical solution as

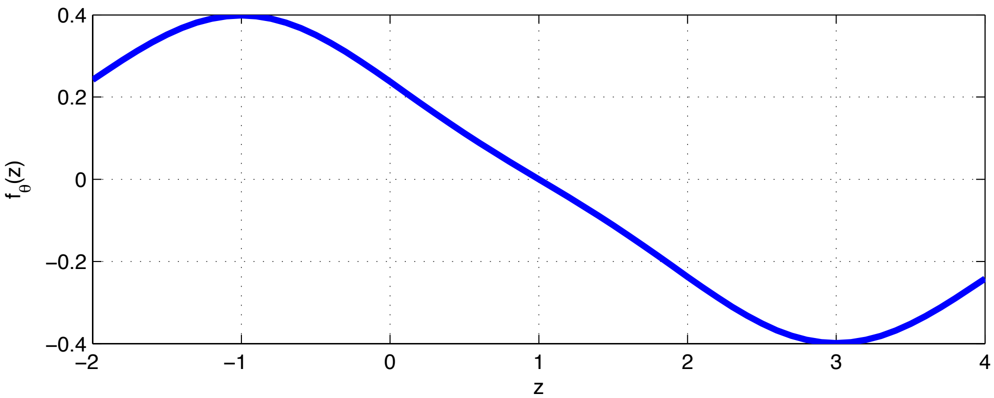
$$\mathbf{R}_m \hat{\theta} = \mathbf{r}_m$$

with covariance matrix $\mathbf{R}_{m,ij} = \sum_{t=1}^{m} \phi_i(z_t)\phi_j(z_t)$ and $\mathbf{r}_{m,i} = \sum_{t=1}^{m} \phi_i(z_t) y_t$.

5 basis functions:



A function f_θ with parameter vector $(1, 0, 0, 0, -1)$:



1. If $m = o(n)$?

2. Choice of Basis Functions.

3. Linear in parameters.

Too much freedom: Regularization

Parametric methods:

$$\min_{\theta \in \mathbb{R}^d} \sum_t \ell(y_t - f_\theta(z_t))$$

When $d \to n$, too high variance (or $\mathbf{R}_d$ ill-conditioned). Better

$$\min_{\theta \in \Theta} \sum_t \ell(y_t - f_\theta(z_t)) + \gamma c(\theta)$$

where

- $c(\theta)$ measures complexity.
- $\gamma > 0$ regularization trade-off.
- If two models give equivalent fit, choose the least complex one.
- $c(\theta) \to c'(f_\theta)$.

General Approximators: Artificial Neural Networks

Model:

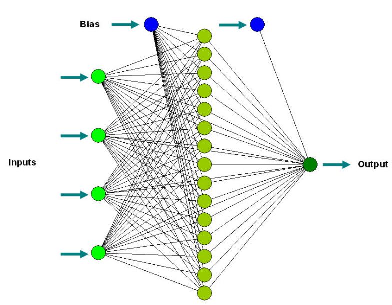
$$f_\theta(x) = \sum_{i=1}^{m} \theta_i \phi_i(x; \theta)$$

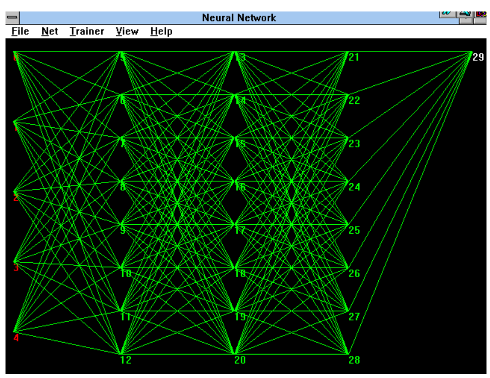
and

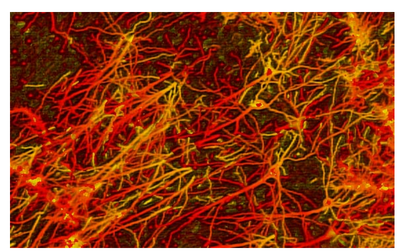
$$\phi_i(x; \theta) = \sum_{i=1}^{m} \theta'_i \phi'_i(x; \theta)$$

and ...

Graphical representation:







But:

- Algorithmic (backpropagation).
- Representation and Design.
- Optimality?

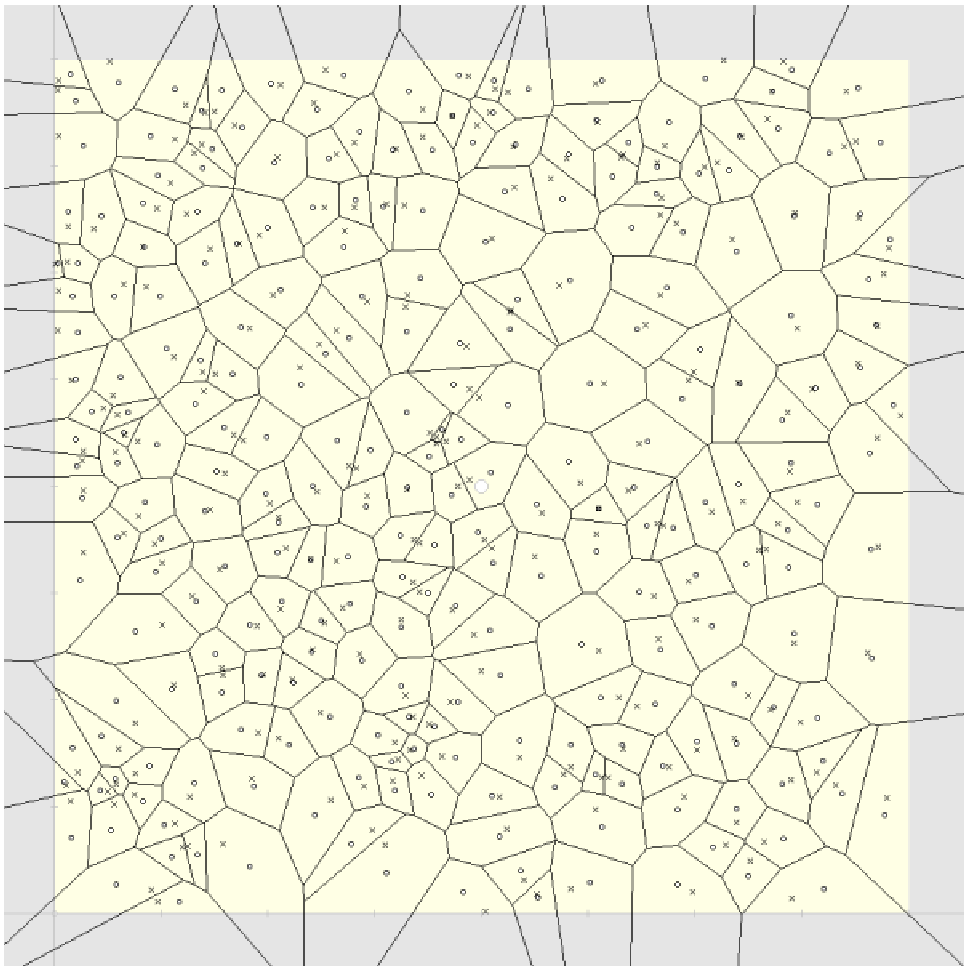
In 90s (70s): when starting from proper optimality function

$$\min_{f \in \mathcal{F}} \sum_t (y_t - f(x_t))^2 + \gamma \|f\|_H$$

then *optimal* representation (network):

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

General Approximators: Nearest Neighbor rules



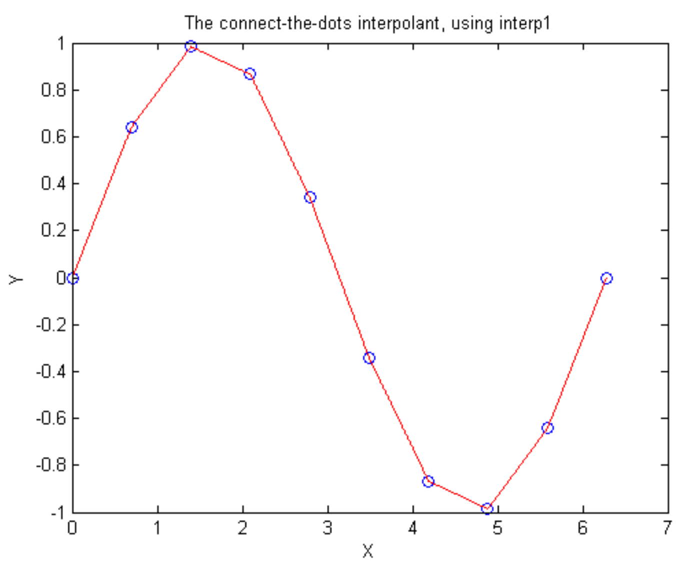
$$f_\theta(x) = y_i I(x \in S_i)$$

General Approximators: Piecewise Linear Systems

Divide domain in disjunct regions $\{S_i\}$

$$f_\theta(x) = \sum_{i=1}^{m} I(x \in S_i)(\theta_i x + b_i)$$

such that joined at knots.



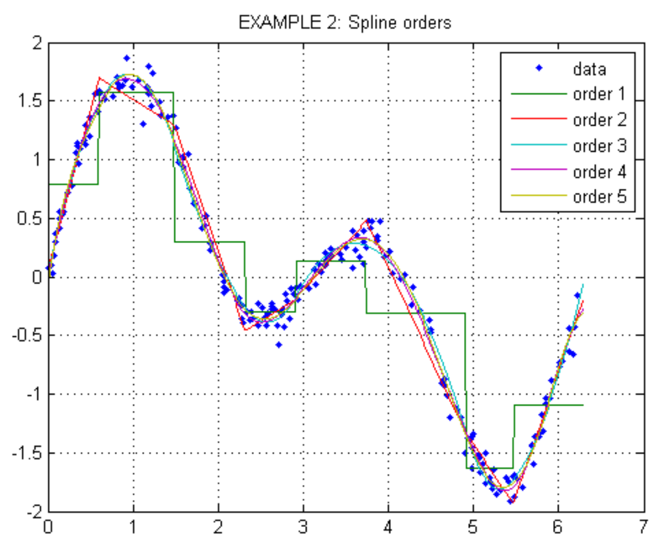
But where to put the knots?

General Approximators: Splines

Divide domain in disjunct regions $\{S_i\}$

$$f_\theta(x) = \sum_{i=1}^{m} I(x \in S_i)(\theta_i x^d + \cdots + b_i)$$

such that joined and differentiable at knots.



But where to put the knots?

- Interpolation $\rightarrow$ B-Splines (numerical).

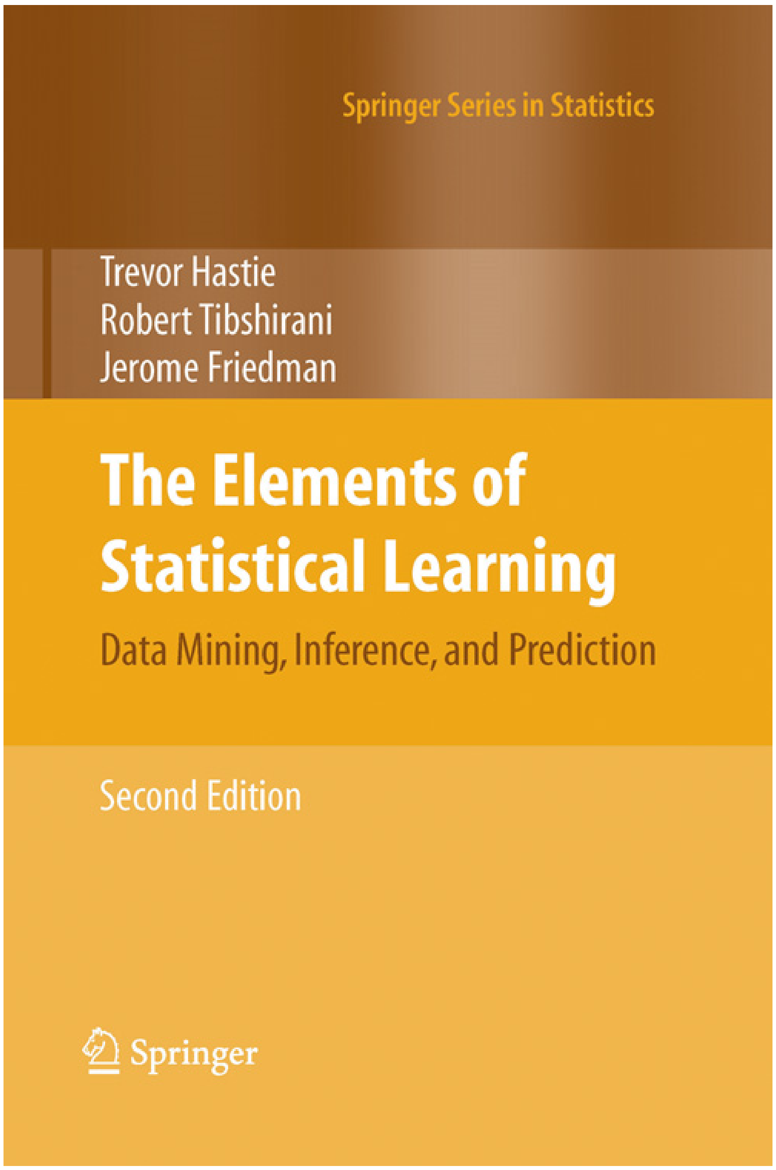
- Noise $\rightarrow$ Smoothing Splines (Bayesian).

General Approximators: Nonparametric Techniques

- Parameter set Θ to function set $\{f\}$.
- No explicit form.
- Algorithmic construction.
- Semi-parameteric $f(z) = f_\theta(z) + g(z)$.
- Fitting noise.
- Prediction and generalization.
- High-dimensional problems.

General Approximators

The Elements of Statistical Learning, Hastie et al. 2002,2009.

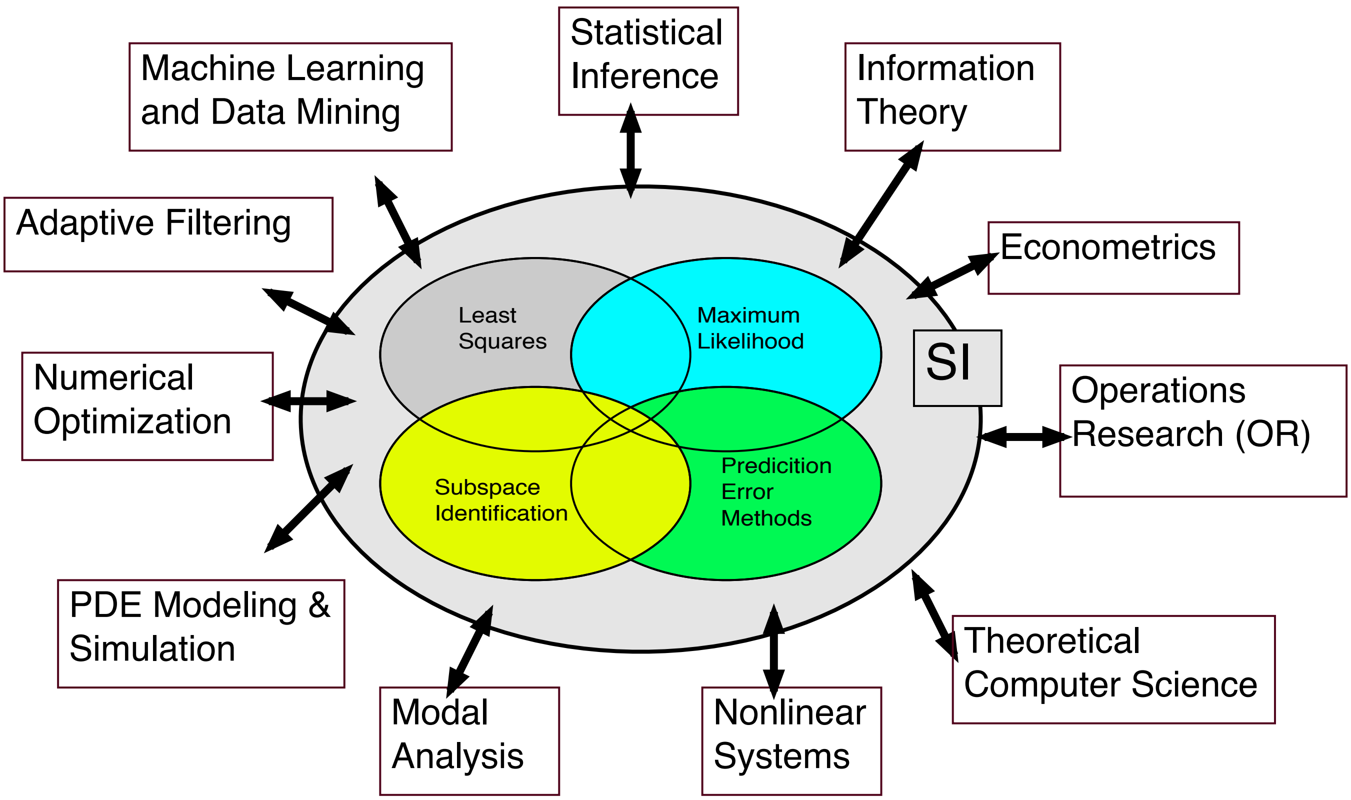


Conclusions

To remember

- Nonlinear Dynamic Models.
- Regularization.
- Toolbox.
- Optimization.

System Identification: A Wider View



1. SI = Recovery/Approximation of Systems from Experiments.
2. Characteristic: Dynamical Nature, Model Structures.
3. Interdisciplinary.

Adaptive Filtering

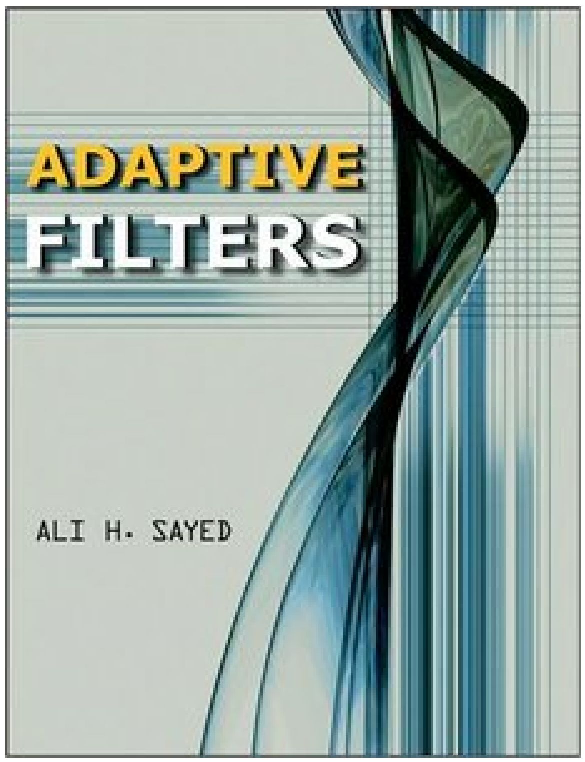
1. What: "Track optimal filter $_t$ which purifies the signals." Ex.:
   (a) Initialize $f_0 = 0_d$, $t = 0$
   (b) Predict $f_{t-1}(\mathbf{x}_t)$ and measure feedback $e_t = (y_t - f_{t-1}(\mathbf{x}_t))$
   (c) Update $f_t = f_{t-1} + g(e_t)$
   (d) Repeat for $t = 1, 2, \ldots$.

2. Why:
   - Communication.
   - Acoustics.
   - Filters.

3. Results:
   - Differential Equation.
   - Algorithmic.
   - Equalization.
   - Efficiency.
   - Time-varying.

4. Relevance 2 SI:
   - D/A and anti-aliasing filters.
   - Equalization and communication.
   - Block-adaptive filters and networks.

5. Text:



Numerical Analysis

1. What: "Numerical analysis is the study of algorithms that use numerical computation (as opposed to general symbolic manipulations) for the problems of mathematical analysis (as distinguished from discrete mathematics)."

2. Why: continuous $\rightarrow$ finite.

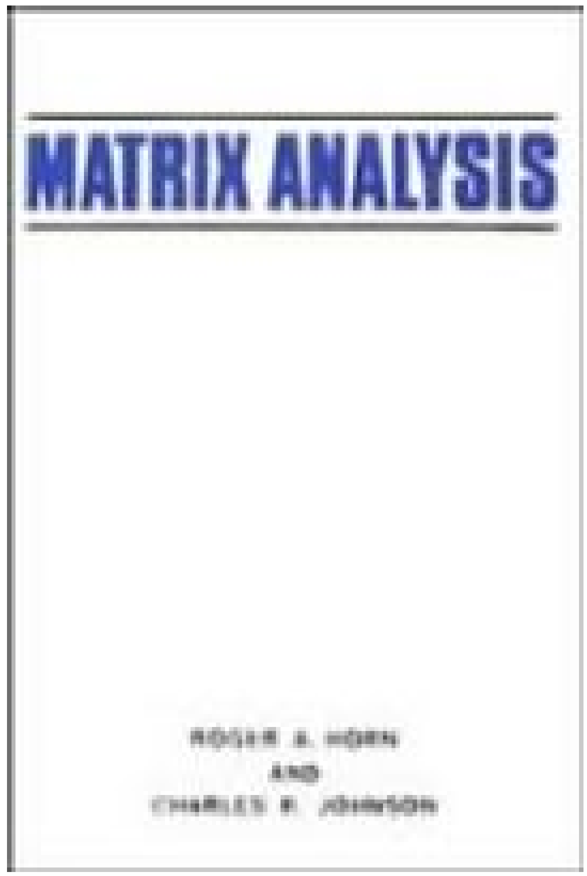
3. Results:

   - Matrix manipulations.
   - Characterizations.
   - Decompositions.

4. Relevance 2 SI:

   - Subspace ID.
   - LAPACK/NUMPACK.
   - Distributed Computation.

5. Text:



Numerical Optimization

1. What:

$$\min_{\theta \in \mathbb{D}} J(\theta) \quad \text{s. t.} \quad \theta \in \Theta$$

2. Why: Local/Global?

3. Results:

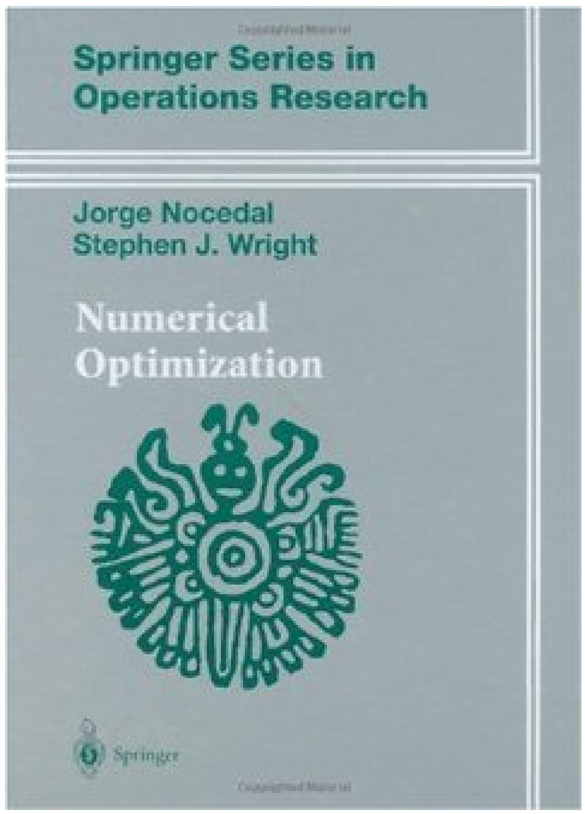
   - LS versus non-LS.
   - Linear versus nonlinear.
   - Convex versus Non-convex.
   - Heuristics.
   - Speed of convergence & Comp. demand.

4. Relevance 2 SI:

   - Toolbox and Embedded Systems.
   - Practical and theoretical efficient algorithms.
   - Differential vs. non-differential.
   - Recursive Identification.

- Motor.
- How to interpret numerical/asymptotic result?

5. Text:

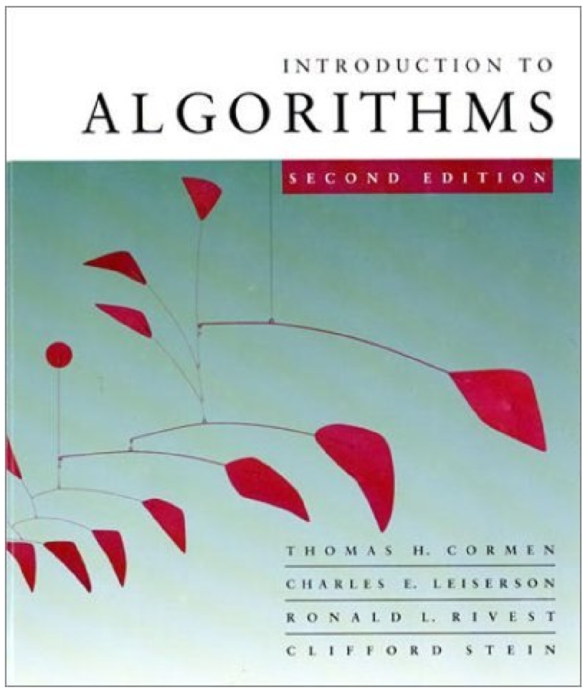


Theoretical Computer Science

1. What: "The design and study of algorithms."
2. Why:
   - Efficient algorithms.
   - Computational and Memory Complexity.
3. Results:
   - Sorting, ..., bin-packing.
   - P versus NP.
   - Randomization.
   - Heuristics.
   - Reduction to numerical analysis.
   - Beyond matrices.
4. Relevance 2 SI:
   - Sequential and Online.

- Nonlinear ID.
- Greedy strategies.

5. Text:



Operations Research

1. What: "Operations research is an interdisciplinary mathematical science that focuses on the effective use of technology by organizations."

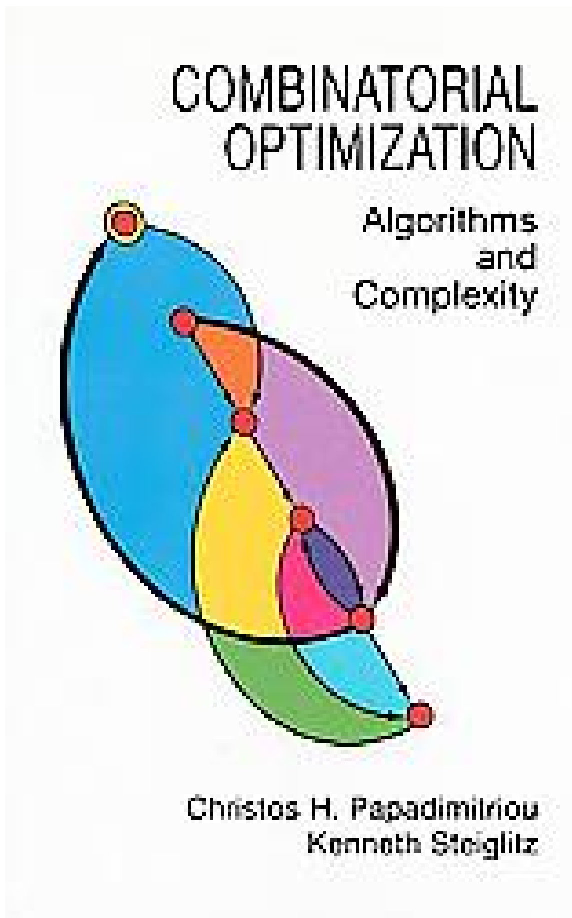
2. Why:
   - WWII.
   - Optimal Strategies.
   - DP.
   - Abstractions (models).

3. Results:
   - MINCUT - MAXFLOW - linear Programming.
   - Combinatorial Optimization.
   - Matching, Allocation, Scheduling, Paths and Routing.
   - Sequential Testing and Quality Control.

4. Relevance 2 SI:

- Combinatorial Models.
- Networked Systems.
- Optimization.

5. Text:



Machine Learning and Data Mining

1. What: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

$$\mathbf{y} \approx f(\mathbf{X})$$

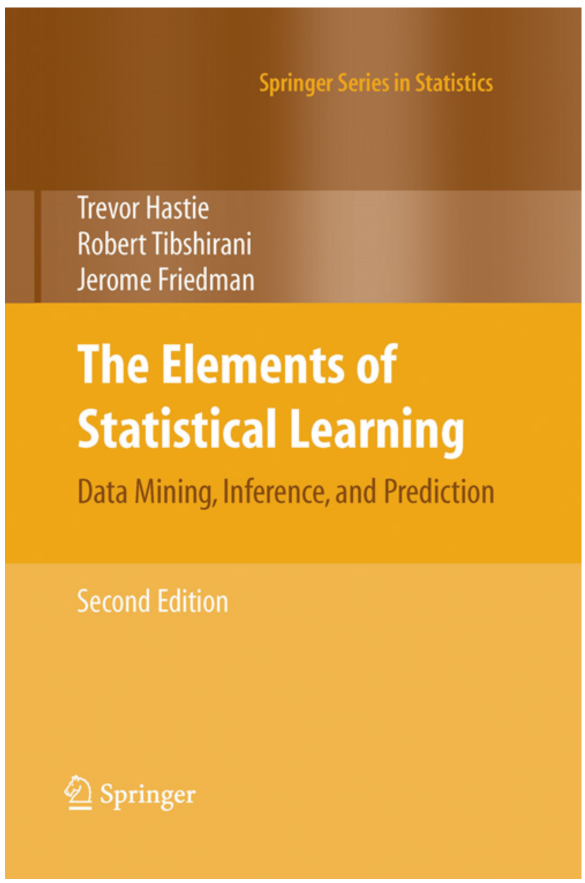
2. Why:
   - Nonlinear models and predictors.
   - How to characterize and relate many different tools?
3. Results:
   - Toolboxes (SVM, splines, Decision trees).
   - ML matured $\rightarrow$ parameters 2 functions.
   - Algorithms.
   - Complexity Control and Generalization.

- Theoretical ML vs. Applications (DARPA).

4. Relevance 2 SI:

   - Off-the-shelve tools.
   - Generalization Analysis.
   - MATLAB, WEKA, Python, ...

5. Text:



Statistical Inference

1. What: "Estimation and inference of Statistical models generating the data, from data." ex.

$$X \sim \mathcal{N}(0_n, \Sigma)$$

   ML:

$$\hat{\Sigma} = \underset{\Sigma}{\operatorname{argmax}}\, L(X_n; \Sigma)$$

2. Why:

   - Stochastics as an abstraction of irrelevant, individual effects.
   - Optimal model $\rightarrow$ Optimal predictor?
   - Averaging behavior.

3. Results:

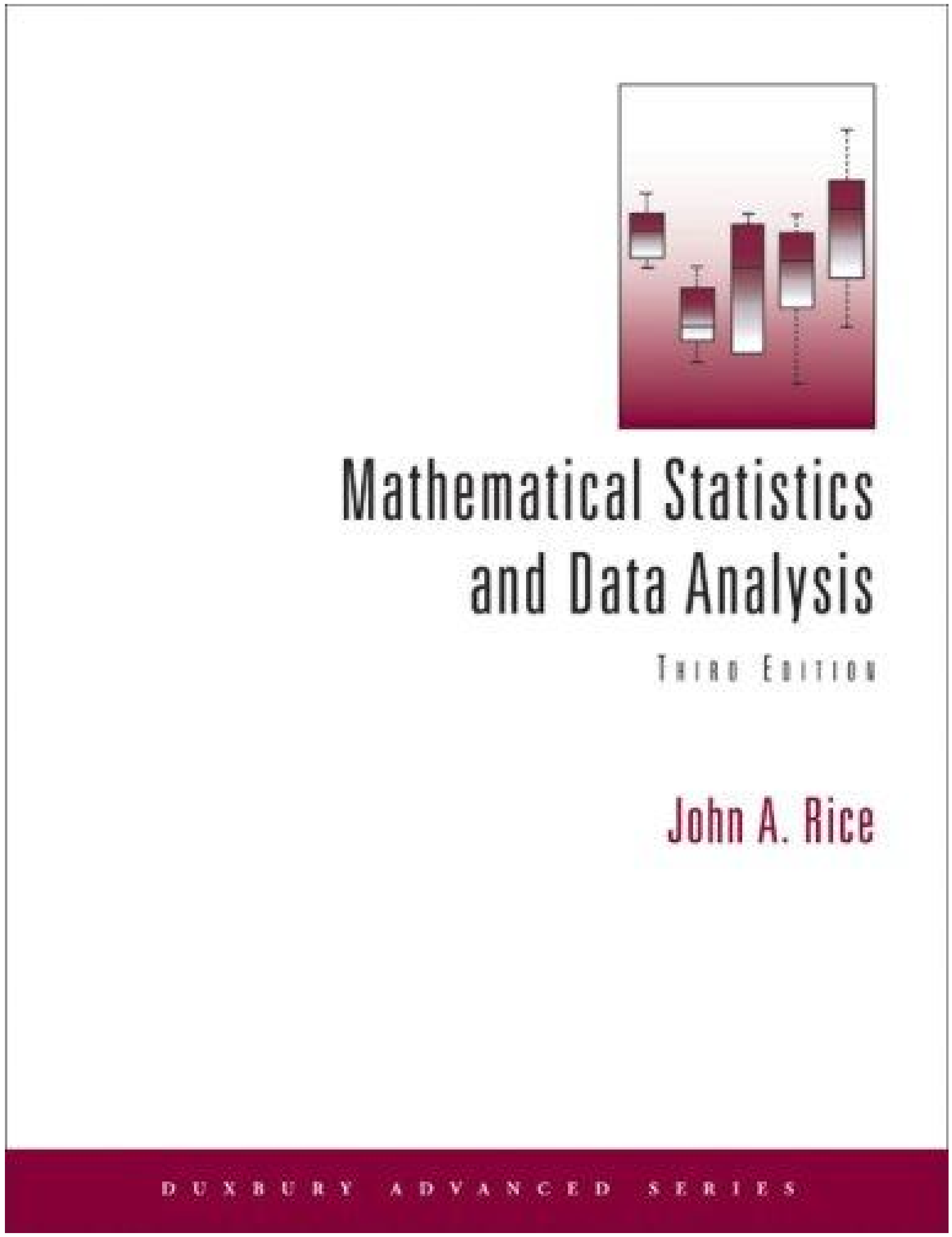
   - Stochastic Processes, IID.

- Statistical Models.
- ML.
- CLT and Cramer-Rao.
- Hypothesis Testing.
- Finite sample results.
- Beyond ML: Penalized ML, U-, L-, M-, V-, R-statistics.

4. Relevance 2 SI:

- Timeseries.
- Often nonlinear in parameters.
- Often Newton-Raphson.
- Inference and covariance.
- R, SAS, Python, stata, SPSS, Matlab, Excel.
- Data visualization tools.

5. Text:


Mathematical Statistics
and Data Analysis
THIRD EDITION
John A. Rice
DUXBURY ADVANCED SERIES

Information Theory

1. What: "Modeling as communication - a model as summary of the data."

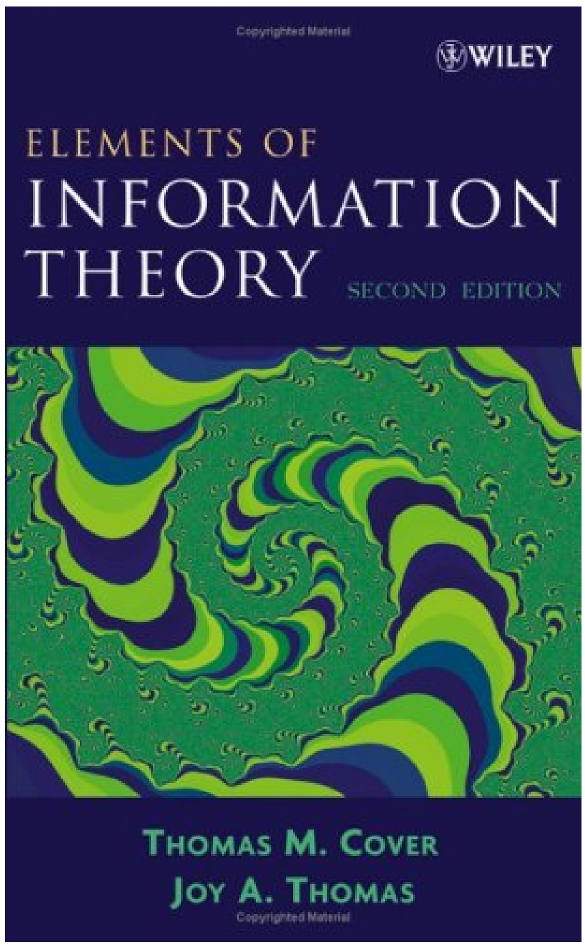
2. Why:
   - Choice of model subjective.
   - Objective guidelines?
   - Fundamental limits.

3. Results:
   - Shannon's source coding theorem $|\text{com}(X)| \geq h(X)$
   - Shannon's noisy source coding theorem $|\text{com}(X)|/|X| \geq \frac{C}{1-h(X)}$
   - Entropy, KL, MI.
   - MDL.
   - rate Distortion theory.

4. Relevance 2 SI:

- Compression.
- Foundation to Stochastic.
- Gambling, Investment and Universal rules.

5. Text:



Econometrics and Financial Matters

1. What: "Econometrics studies statistical properties of econometric procedures"

2. Why:

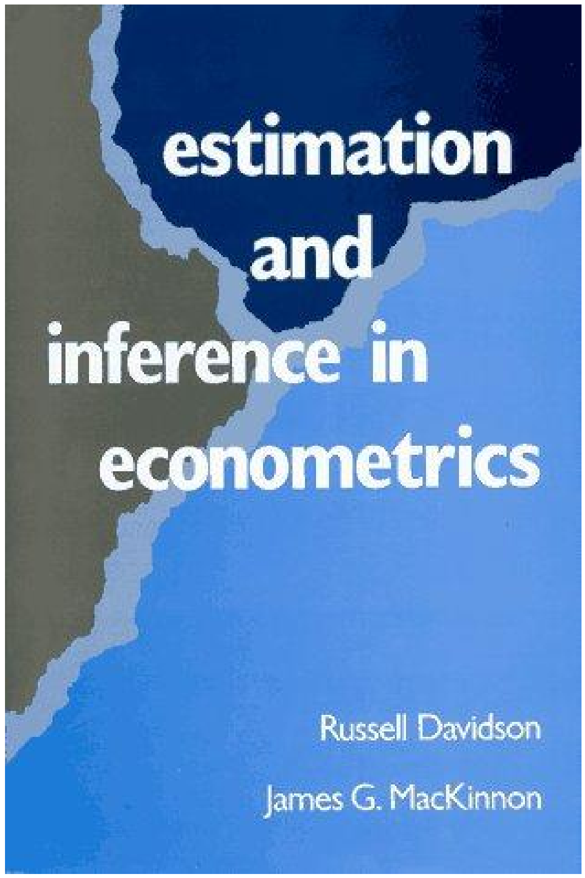
3. Results:

   - Noise and correlations.
   - Jumps and outliers.
   - Variance Stabilizing transformations.
   - Gambling and Maximal profit strategies.
   - Stochastic Calculus ($\hat{I}$to)

4. Relevance SI:

   - Timeseries modeling.
   - Preprocessing.
   - Continuous time.

5. Text:


estimation
and
inference in
econometrics
Russell Davidson
James G. MacKinnon

Nonlinear Systems

1. What: "Study of the dynamics arising from nonlinear systems."

2. Why:

   - Models $\rightarrow$ I/O Behavior.
   - I/O Behavior ? models?

3. Results:

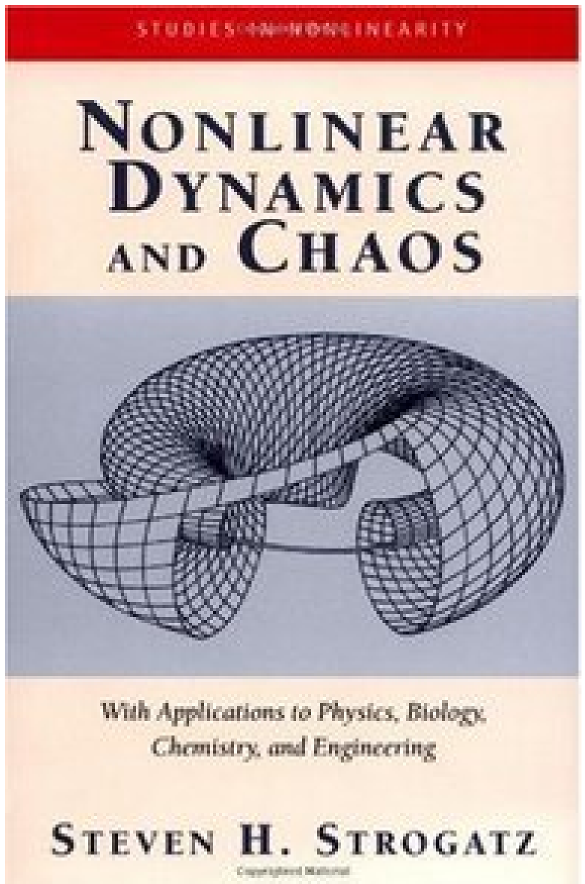
   - Oscillators.
   - Bifurcation Diagrams.
   - Long range prediction.
   - Stability and Limit Cycles.

4. Relevance SI:

   - Diagnostics to Identified Nonlinear model.
   - Observed behavior $\rightarrow$ Model structure?

- Phase

5. Text:


STUDIES IN NONLINEARITY
NONLINEAR DYNAMICS AND CHAOS
With Applications to Physics, Biology, Chemistry, and Engineering
STEVEN H. STROGATZ

Conclusions

To remember

- Least Squares.
- Extensions.
- Toolbox.
- Tuning.