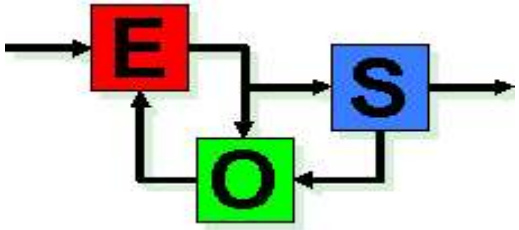
System Identification General Aspects and Structure

M. Deistler
University of Technology, Vienna



Research Unit for Econometrics and System Theory
{deistler@tuwien.ac.at}

November 2007

Contents

1. Introduction

The art of identification is to find a good model from noisy data: Data driven modeling.

This is an important problem in many fields of application.

Systematic approaches: Statistics, System Theory, Econometrics, Inverse Problems.

MAIN STEPS IN IDENTIFICATION

- Specify the model class (i.e. the class of all a priori feasible candidate systems): Incorporation of a priori knowledge.

- Specify class of observations.

- Identification in the narrow sense. An identification procedure is a rule (in the automatic case a function) attaching a system from the model class to the data
  - ⋆ Development of procedures.
  - ⋆ Evaluation of procedures (Statistical and numerical properties).

Here only identification from equally spaced, discrete time data $y_t, t = 1, \ldots T; y_t \in \mathbb{R}^s$ is considered.

MAIN PARTS

- Main stream theory for linear systems (Nonlinear!).
- Alternative approaches to linear system identification.
- Identification of nonlinear systems: parametric, nonparametric

MAINSTREAM THEORY

- The model class consists of linear, time-invariant, finite dimensional, causal and stable systems only. The classification of the variables into inputs and outputs is given a priori.

- Stochastic models for noise are used; in particular noise is assumed to be stationary, ergodic with a rational spectral density.

- The observed inputs are assumed to be free of noise and to be uncorrelated with the noise process.

- Semi-nonparametric approach: A parametric subclass is determined by model selection procedures.
  First step: Estimation of integer valued parameters.
  Then, for the given subclass, the finite dimensional vector of real valued parameters is estimated.

- Emphasis on asymptotic properties (consistency, asymptotic distribution) in evaluation.

3 MODULES IN IDENTIFICATION

- STRUCTURE THEORY: Idealized Problem;
  we commence from the stochastic processes generating the data (or their population moments) rather than from data. Relation between “external behavior” and “internal parameters”.

- ESTIMATION OF REAL VALUED PARAMETERS: Subclass (dynamic specification) is assumed to be given and parameter space is a subset of an Euclidean space and contains a nonvoid open set: M-estimators

- MODEL SELECTION: In general, the orders, the relevant inputs or even the functional forms are not known a priori and have to be determined from data. In many cases, this corresponds to estimating a model subclass within the original model class. This is done, e.g. by estimation of integers, e.g. using information criteria or test sequences.

THE HISTORY OF THE SUBJECT

(i) Early (systematic, methodological) time series analysis dates back to the 18th and 19th century. Main focus was the search for "hidden"periodicities and trends, e.g. in the orbits of planets (Laplace, Euler, Lagrange, Fourier). Periodogram (A. Schuster). Economic time series, e.g. for business cycle data.

(ii) Yule (1921, 1923). Linear stochastic systems (MA and AR systems) used for explaining "almost periodic" cycles: $y_t - a_1 y_{t-1} - a_2 y_{t-2} = \epsilon_t$.

(iii) (Linear) Theory of (weak sense) stationary processes (Cramer, Kolmogoroff, Wiener, Wold). Spectral representation, Wold representation (linear systems), factorization, prediction, filtering and interpolation.

(iv) Early econometrics, in particular, the work of the Cowles Commission (Haavelmo, Koopmans, T. W. Anderson, Rubin, L. Klein). theory of identifiability and of (Gaussian) Maximum Likelihood (ML) - estimation for (finite dimensional) MIMO (multi-input, multi-output) linear systems (vector difference equations) with white noise errors (ARX systems). The maximum lag lengths are assumed to be known a priori. Development of LIML, 2SLS and 3 SLS estimators (T. W. Anderson, Theil, Zellner).

(v) (Nonparametric) Spectral estimation and estimation of transfer functions (Tukey).

(vi) Estimation of AR, ARMA, ARX and ARMAX systems. SISO (single-input,single-output) case. Emphasis on consistency, asymptotic normality and efficiency, in particular, for least squares and ML estimators. (T.W. Anderson, Hannan, Walker).

(vii) Structure theory for (MIMO) state space and ARMA systems (Kalman).

(viii) Box-Jenkins procedure: An “integrated” approach to SISO system identification including order estimation (non automatic), the treatment of certain non-stationarities and numerically efficient ML algorithms. Big impact on applications.

(ix) Automatic procedures for order estimation, in particular, procedures based on information criteria (like AIC, BIC) (Akaike, Rissanen).

(x) Main stream theory for linear system identification (including MIMO systems): Structure theory, order estimation, estimation for “real valued” parameters with emphasis on asymptotic theory (Hannan, Akaike, Caines, Ljung).

(xi) Alternative approaches.

2. Structure Theory

Relation between external behavior and internal parameters. Linear, main stream case: Relations between transfer function and parameters.

Main model classes for linear systems:

- AR(X)
- ARMA(X)
- State Space Models

Here, for simplicity of notation we assume that we have no observed inputs. In many applications AR models still dominate.

Advantages of AR models:

- no problems of non-identifiability, structure theory is simple
- maximum likelihood estimates are of least squares type, i.e. asymptotically efficient and easy to calculate

Disadvantages of AR models:

- less flexible

Here, the focus is on state space models.

State space forms in innovation representation:

$$x_{t+1} = Ax_t + B\varepsilon_t \quad (1)$$

$$y_t = Cx_t + \varepsilon_t \quad (2)$$

where

- y_t: s-dimensional outputs
- x_t: n-dimensional states
- (ε_t) white noise
- $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times s}$, $C \in \mathbb{R}^{s \times n}$: parameter matrices
- n: integer valued parameter

Assumptions:

$$|\lambda_{max}(A)| < 1 \quad (3)$$

$$|\lambda_{max}(A - BC)| \leq 1 \quad (4)$$

$$\mathbb{E}\varepsilon_t \varepsilon_t^{'} = \Sigma > 0 \quad (5)$$

Transfer function:

$$k(z) = \sum_{j=1}^{\infty} K_j z^j + I$$

$$K_j = CA^{j-1}B \quad (6)$$

ARMA forms

$$a(z)y_t = b(z)\varepsilon_t$$

External behavior

$$\begin{aligned} f(\lambda) &= (2\pi)^{-1}k(e^{-i\lambda})\Sigma k^*(e^{-i\lambda}) \\ f &\leftrightarrow (k, \Sigma) \end{aligned}$$

Note that ARMA and state space systems describe the same class of transfer functions.

Relation to internal parameters:

(6) or $a^{-1}b = k$

$U_A = \{k|$ rational, $s \times s$, $k(0) = I$, no poles for $|z| \leq 1$ and no zeros for $|z| < 1\}$

$M(n) \subset U_A$: Set of all transfer functions of order n.

T_A: Set of all A, B, C for fixed s, but n variable, satisfying (4) and (5).

$S(n) \subset T_A$: Subset of all (A, B, C) for fixed n.

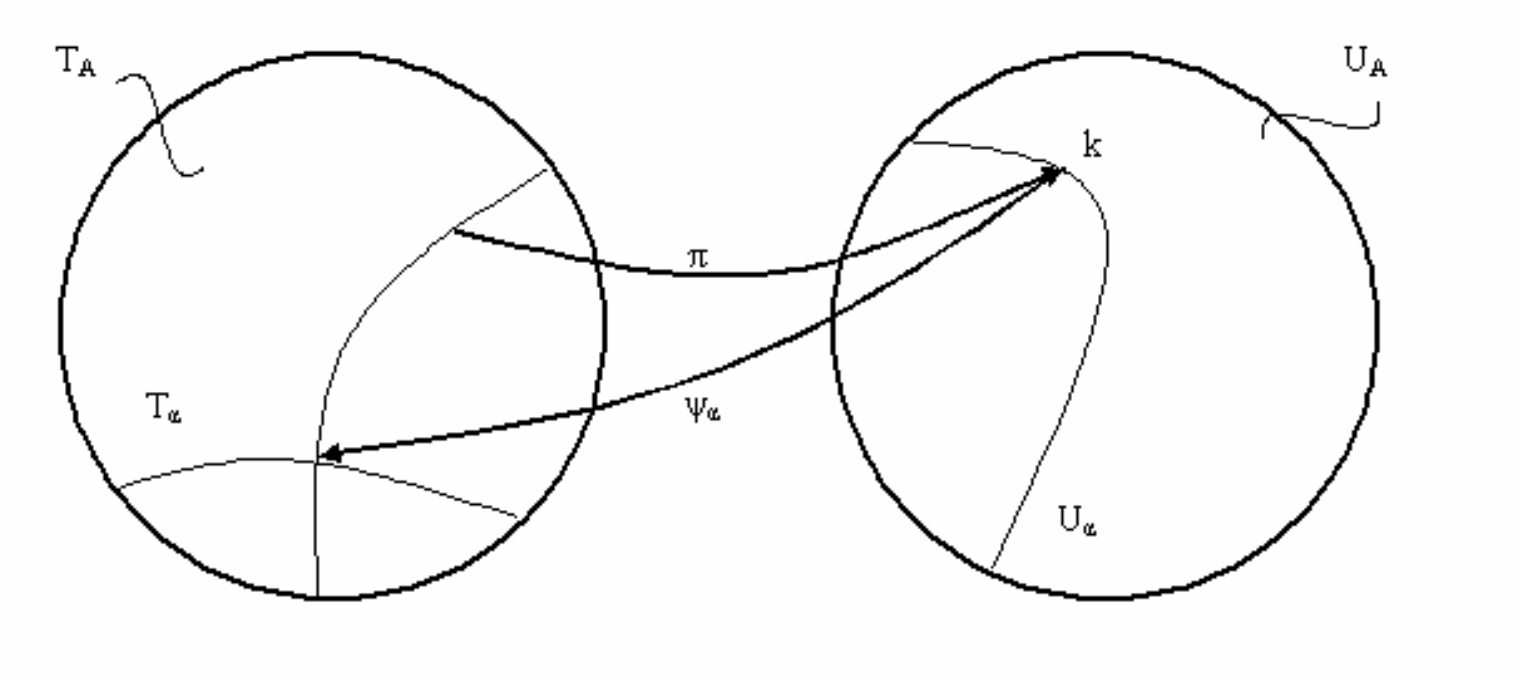
$S_m(n) \subset S(n)$: Subset of all minimal (A, B, C).

$\pi : T_A \rightarrow U_A : \pi(A, B, C) = k = C(Iz^{-1} - A)^{-1}B + I$

π is surjective but not injective

Note: T_A is not a good parameter space because:

- T_A is infinite dimensional
- lack of identifiability
- lack of "well posedness": There exists no continuous selection from the equivalence classes $\pi^{-1}(k)$ for T_α.



Desirable properties of parametrizations:

- U_A and T_A are broken into bits, U_α and $T_\alpha, \alpha \in I$, such that π restricted to T_α: $\pi_{|T_\alpha} : T_\alpha \to U_\alpha$ is bijective. T_α is reparametrized such that it contains an open set in an embeddding $\mathbb{R}^{d_\alpha}$. By $\tau \in T_\alpha$ we denote the vector of free parameters. Then there exists a parametrization $\psi_\alpha : U_\alpha \to T_\alpha$ such that $\psi_\alpha(\pi(A, B, C)) = (A, B, C) \quad \forall (A, B, C) \in T_\alpha$.

- U_α is finite dimensional in the sense that $U_\alpha \subset \cup_{i=1}^{n} M(i)$ for some n.

- Well posedness: The parametrization $\psi_\alpha : U_\alpha \to T_\alpha$ is a homeomorphism (pointwise topology T_{pt} for U_A).

- Differentiability

- U_α is T_{pt}-open in $\bar{U}_\alpha$.

- $\cup_{\alpha \in I} U_\alpha$ is a cover for U_A.

Examples:

- Canonical forms based on $M(n)$, e.g. echelon forms and balanced realizations. Decomposition of $M(n)$ into sets U_α of different dimension. Nice free parameters vs. nice spaces of free parameters.

- "Overlapping description" of the manifold $M(n)$ by local coordinates.

- "Full parametrization" for state space systems. Here $S(n) \subset \mathbb{R}^{n^2+2ns}$ or $S_m(n)$ are used as parameter spaces for $\bar{M}(n)$ or $M(n)$, respectively. Lack of identifiability. The equivalence classes are n^2 dimensional manifolds. The likelihood function is constant along these classes.

- Data driven local coordinates (`DDLC`): Orthonormal coordinates for the $2ns$ dimensional ortho-complement of the tangent space to the equivalence class at an initial estimator. Extensions: `slsDDLC` and `orthoDDLC`

- ARMA systems with prescribed column degrees.

- ARMA parametrizations commencing from writing k as $c^{-1}p$ where c is a least common denominator polynomial for k and where the degrees of c and p serve as integer valued parameters.

In general, state space systems have larger equivalence classes compared to ARMA systems: More freedom in selection of optimal representatives.

Main unanswered question: Optimal tradeoff between "number" and dimension of the pieces U_α.

Problem: Numerical properties of parametrizations

Different parametrizations:

$$\psi_1 : U_1 \to T_1 \subset T_A$$
$$\psi_2 : U_2 \to T_2 \subset T_A$$

STATISTICAL ANALYSIS (*"real world"*):

For the asymptotic analysis, in the case that $U_1 \supset U_2$, U_2 contains a nonvoid open (in U_1) set and $k_0 \in$ int U_2, we have no essential differences in asympotic theory:

- coordinate free consistency:
- different asymptotic distributions, but we know the transformation

NUMERICAL ANALYSIS (*"integer world"*):

- The selection from the equivalence class matters
- Dependence on algorithm

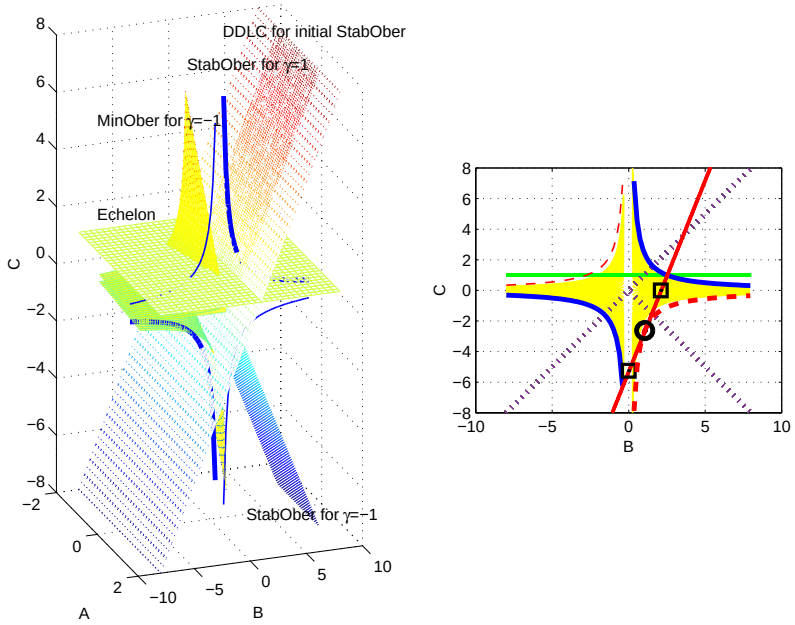
Questions:

- What are appropriate evaluation criteria for numerical properties?
- Which are the optimal parameter spaces (algorithm specific)?

Relation between statistical and numerical precision: curvature of the criterion function:

Consider the case $s = n = 1$ where $(a, b, c) \in \mathbb{R}^3$:

- Minimality: $b \neq 0$ and $c \neq 0$
- Equivalence classes of minimal systems: $\bar{a} = a$, $\bar{b} = tb$, $\bar{c} = ct^{-1}$, $t \in \mathbb{R} \setminus \{0\}$



3. Estimation for a Given Subclass

We here assume that U_α is given, we commence from data.

Identifiable case: $\psi_\alpha : U_\alpha \to T_\alpha$ has the desirable properties.

$\tau \in T_\alpha \subset \mathbb{R}^{d_\alpha}$: vector of free parameters for U_α.

$\sigma \in \underline{\Sigma} \subset \mathbb{R}^{\frac{n(n+1)}{2}}$: free parameters for $\Sigma > 0$.

Overall parameter space: $\Theta = T_\alpha \times \underline{\Sigma}$.

Many identification procedures, at least asymptotically, commence from the sample second moments of the data

GENERAL FEATURES:

$$\hat{\gamma}(s) = T^{-1} \sum_{t=1}^{T-s} y_{t+s} y_t^{'}, \quad s \geq 0$$

Now, $\hat{\gamma}$ can be directly realized as an MA system typically of order Ts; $\hat{\hat{k}}_T$

IDENTIFICATION:

- Projection step (Model reduction). Important for statistical qualities.
- Realization step:

Two types of procedures:

- Optimization based procedures, M-estimators:

$$\hat{\theta}_T = \text{argmin} L_T(\theta; y_1, \ldots, y_T)$$

- Direct procedures: Explicit functions. e. g. instrumental variables methods, subspace methods.

GAUSSIAN MAXIMUM LIKELIHOOD:

$$\hat{L}_T(\theta) = T^{-1}\text{log det}\Gamma_T(\theta) + T^{-1}y'(T)\Gamma_T(\theta)^{-1}y(T)$$

where $y(T) = (y_1', \ldots, y_T')'$, $\Gamma_T(\theta) = \mathbb{E}y(T;\theta)y'(T;\theta)$ and
$\hat{\theta}_T = \text{argmin}_{\theta\in\Theta}\hat{L}_T(\theta)$

- No explicit formula for MLE, in general.
- $\hat{L}_T(k,\Sigma)$ since $\hat{L}_T$ depends on τ only via k: parameter free approach.
- Boundary points are important.

WHITTLE LIKELIHOOD:

$$\begin{aligned}\hat{L}_{W,T}(k,\sigma) &= \text{log det}\Sigma + \\ &\quad (2\pi)^{-1}\int_{-\pi}^{\pi}\text{tr}\left[\left(k(e^{-i\lambda})\Sigma k^*(e^{-i\lambda})\right)^{-1} I(\lambda)\right] d\lambda\end{aligned}$$

where $I(\lambda)$ is the periodogram.

EVALUATION:

- Coordinate free consistency: for $k_0 \in U_\alpha$ and

  $\lim T^{-1} \sum_{t=1}^{T-s} \varepsilon_{t+s} \varepsilon_t^{'} = \delta_{0,s} \Sigma_0$ a.s. for $s \geq 0$ we have $\hat{k}_T \to k_0$ a.s. and $\hat{\Sigma}_T \to \Sigma_0$ a.s.

  Consistency proof: basic idea Wald (1949) for i.i.d. case.

  Noncompact parameter spaces:

$$\lim_{T \to \infty} \hat{L}_T(k, \sigma) = L(k, \sigma) = \text{log det} \Sigma + (2\pi)^{-1}$$
$$\int_{-\pi}^{\pi} \text{tr} \left[\left(k(e^{-i\lambda}) \Sigma k^*(e^{-i\lambda}) \right)^{-1} \left(k_0(e^{-i\lambda}) \Sigma_0 k_0^*(e^{-i\lambda}) \right) \right] d\lambda \qquad (7)$$

  a.s.

  - ⋆ L has a unique minimum at k_0, Σ_0.
  - ⋆ $(\hat{k}_T, \hat{\Sigma}_T)$ enters a compact set, uniform convergence in (7).

- Analogous for $k_0 \in \bar{U}_\alpha$.

- Generalized, coordinate free consistency for $k_0 \notin \bar{U}_\alpha$, $(\hat{k}_T, \hat{\Sigma}_T) \to D$ a.s D: Set of all best approximants to k_0, Σ_0 in $\bar{U}_\alpha \times \underline{\Sigma}$.

- Consistency in coordinates: $\psi_\alpha(\hat{k}_T) = \hat{\tau}_T \to \tau_0 = \psi_\alpha(k_0)$ a.s.

- CLT:
  Under

$$\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$$

  and

$$\mathbb{E}(\varepsilon_t\varepsilon_t'|\mathcal{F}_{t-1}) = \Sigma_0.$$

$$\sqrt{T}(\hat{\tau}_T - \tau_0) \to^d N(0, V)$$

  Idea of proof: Cramer (1946) i.i.d. case: Linearization.

Direct Estimators: IV Methods, subspace methods: Numerically faster, in many cases not asymptotically efficient.

CALCULATIONS OF ESTIMATES

Usual procedure:
consistent initial estimator (e.g. IV or subspace estimator)
+ one Gauss-Newton step gives an asymptotically efficient procedure (e.g. Hannan-Rissanen)

HOWEVER THERE ARE STILL PROBLEMS

- Problem of local minima: “good” initial estimates are required

- Numerical problems: Optimization over a grid
  Statistical accuracy may be higher than numerical accuracy
  Valleys close to equivalence classes corresponding to lower dimensional systems
  “Intelligent” parametrization may help `DDLC`’s and extensions:
  Data driven selection of coordinates from an uncountable number of possibilities
  Only locally homeomorphic

- “Curse of dimensionality”
  lower dimensional parametrizations (e.g. reduced rank models)
  concentration of the likelihood function by a least squares step.

4. Model Selection

Automatic vs. nonautomatic procedures.

Information criteria: Formulate tradeoff between fit and complexity. Based on e.g. Bayesian arguments, coding theory ...

Order estimation (or more general closure nested case): $n_1 < n_2$ implies $\bar{M}(n_1) \subset \bar{M}(n_2)$ and

dim$(M(n_1))$ <dim$(M(n_2))$.

Criteria of the form

$$A(n) = \log \det \hat{\Sigma}_T(n) + 2ns \cdot c(T) \cdot T^{-1}$$

where $\hat{\Sigma}_T(n)$ is the MLE for Σ_0 over $\bar{M}(n) \times \underline{\Sigma}$.

$c(T) = 2$: AIC criterion

$c(T) = c \cdot \log T, c \geq 1$: BIC criterion

Estimator: $\hat{n}_T =$argmin$A(n)$

Statistical evaluation: $\hat{n}_T$ is consistent for

$$\lim_{T\to\infty} \frac{c(T)}{T} = 0, \quad \liminf_{T\to\infty} \frac{c(T)}{\log\log T} > 0$$

Evaluation of uncertainty coming from model selection for estimators of real valued parameters.

Note: Complexity is in the eye of the beholder. Consider e.g. AR models for $s = 1$:

$$y_t + a_1 y_{t-1} + a_2 y_{t-2} = \varepsilon_t$$

Parameter spaces:
$T = \{(a_1, a_2) \in \mathbb{R}^2 | 1 + a_1 z + a_2 z^2 \neq 0 \text{ for } |z| \leq 1\}$

$T_0 = \{(0, 0)\}$

$T_1 = \{(a_1, 0) || a_1| < 1, a_1 \neq 0\}$

$T_2 = T - (T_0 \cup T_1)$

Bayesian justification:

- Positive priors for all classes, otherwise MLE is asymptotically efficient.
- Certain properties of U_α, $\alpha \in I$ are needed, e.g. for BIC to give consistent estimators: closure nestedness, e.g. $n_1 < n_2 \Rightarrow \bar{M}(n_1) \subset \bar{M}(n_2)$ and dim$M(n_1) <$ dim$M(n_2)$.

Main open question:

- Optimal tradeoff between dimension and "number" of pieces.

Problem: Properties of post model selection estimators

- The statistical analysis of the MLE $\hat{\tau}_T$ traditionally does not take into account the additional uncertainty coming from model selection.

- This may result in very misleading conclusions

Consider AR case (nested):

$$y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t$$

where

$$T_p = \left\{ \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} \in \mathbb{R}^p | stability \right\}$$

The estimator (LS) for given p is

$$\hat{\tau}_p = \left(X(p)'X(p)\right)^{-1} X(p)y$$

The post model selection estimator is

$$\tilde{\tau} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} 1_{\{\hat{p}=0\}} + \begin{pmatrix} \hat{a}_1(1) \\ \vdots \\ 0 \end{pmatrix} 1_{\{\hat{p}=1\}} + \ldots + \begin{pmatrix} \hat{a}_1(p) \\ \vdots \\ \hat{a}_p(p) \end{pmatrix} 1_{\{\hat{p}=p\}}$$

Main problem:

- Essential lack of uniformity in convergence of finite sample distributions.

5. Linear Non-Mainstream Cases

- Time varying parameters
- Long memory
- Unstable systems, integration and cointegration
- Symmetric modeling, errors in variables, dynamic factor models, identification in closed loop

TIME VARYING PARAMETERS:

- Slowly varying parameters: Without and with models for time variation. e.g.

$$\begin{aligned} y_t &= a_t y_{t-1} + \varepsilon_t \\ a_t &= a_{t-1} + v_t \end{aligned}$$

Recursive, adaptive estimation. Estimating the speed of variation.

- Structural changes:

$$a_t = \begin{cases} a^{(1)} & \text{for} \quad t \leq T_0 \\ a^{(2)} & \text{for} \quad t > T_0 \end{cases}$$

Change point detection.

- STARX Models (Smooth Transition)

$$y_t = a^{(1)} y_{t-1} + d^{(1)} z_t + (a^{(2)} y_{t-1} + d^{(2)} z_t) G(s_t; \theta) + \varepsilon_t$$

s_t . . . Transition variable
$G(\,.\,;\theta) \;:\quad \mathbb{R} \longrightarrow [0,1]$

LONG MEMORY: e.g. a simple fractionally integrated process of the form

$$(1-z)^d y_t = \varepsilon_t \quad , s=1, \quad d \in (0.5, 0)$$

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} k_j \varepsilon_{t-j}$$

$k_j = (j!)^{-1} d(d+1)\ldots(d+j-1)$
$\sum_j |k_j| = \infty \quad \sum_j k_j^2 < \infty$

non-rational transfer function

$$f(\lambda) \sim \lambda^{-2d} \quad \text{for} \quad \lambda \longrightarrow 0$$

COINTEGRATION:

A stochastic process is called *integrated* (of order 1) if $(1-z)y_t$ is stationary, while (y_t) is not.

An integrated vector process (y_t) is called *cointegrated* if $\exists \alpha \in \mathbb{R}^s$ such that $(\alpha' y_t)$ is stationary.

Motivation:

- Trends in means and variance
- α describes long run equilibrium

STRUCTURE THEORY, AR CASE

$$a(z)y_t = \varepsilon_t$$

SPECIAL REPRESENTATION (Johansen)

$$(1-z)y_t = \Gamma_1(1-z)y_{t-1} + \ldots + \Gamma_{p-1}(1-z)y_{t-p+1} + \Pi y_{t-p} + \varepsilon_t$$

where $\Pi = -a(1)$ and

$$\begin{aligned} rk(\Pi) = s \quad &\Leftrightarrow \quad (y_t) \text{ is stationary} \\ rk(\Pi) = 0 \quad &\Leftrightarrow \quad (y_t) \text{ is integrated, but not cointegrated} \\ \underbrace{rk(\Pi) = r}_{0<r<s} \quad &\Leftrightarrow \quad (y_t) \text{ is cointegrated with } r \text{ l.i. c. v.} \end{aligned}$$

$\Pi = BA'$, $A, B \in \mathbb{R}^{s\times r}$. The rows of A span the cointegrating space.

MLE, AR CASE

- (Gaussian) likelihood $\tilde{L}_T(\Gamma_1, \ldots, \Gamma_{p-1}, \Sigma, B, A)$
- Concentrated likelihood (stepwise) $\hat{L}_T(A, r)$: likelihood ratio tests
  - H_0: at most r ($r < s$) l.i. cointegrating vectors
  - H_1: $r + 1$ l.i. cointegrating vectors
- Nonstandard limiting distributions for the test under H_0
- Asymptotic properties for the MLE's $\hat{A}_T$ and $\hat{B}_T$ under additional normalization: non-normal limiting distributions, different speeds of convergence.

Simplest case: AR(1), scalar

$$y_t = \rho y_{t-1} + \varepsilon_t$$

OLS estimate $\hat{\rho}_T$:

$$\begin{aligned} \sqrt{T}(\hat{\rho}_T - \rho) &\to^d N(0, 1-\rho^2) \text{ for } |\rho| < 1 \\ T(\hat{\rho}_T - 1) &\to^d \frac{\frac{1}{2}(W(1)^2 - 1)}{\int_0^1 W(r)^2 dr} \text{ for } \rho = 1 \end{aligned}$$

where $W(.)$ is the standard Brownian motion; functional CLT's and continuous mapping theorem.

- LR tests and MLE's are available
- Open problems in structure theory and specification

LINEAR DYNAMIC FACTOR MODELS:

Basic model

$$y_t = \Lambda(z)\xi_t + u_t \quad , \mathbb{E}\xi_t u_s^{'} = 0$$

y_t: s-dimensional observations
ξ_t: $r < s$-dimensional factors, in general unobserved
$\Lambda(z) = \sum_i \Lambda_j z^j$: factor loadings
$\hat{y} = \Lambda(z)\xi_t$: latent variables

Spectral densities

$$f_y(\lambda) = \Lambda(e^{-\imath\lambda}) f_\xi(\lambda) \Lambda^*(e^{-\imath\lambda}) + f_u(\lambda)$$

Idea: Dimension reduction in cross section and time, modeling high dimensional time series, using information contained in additional time series.

Tasks:

- Estimation of (low dimensionally parametrized versions of) $\Lambda(z), f_\xi(\lambda), f_u(\lambda)$
- Estimation of factor processes (ξ_t)
- Forecasting

Problems:

- Identifibiality
- Estimation of real valued parameters
- Model selection, in paricular determinig r from data.
- Estimation of factors

Additional a-priori information has to be imposed, otherwise the problem would be ”completely non identifiable”.

Special model classes:

- (Dynamic) principal compenents

$$f_y(\lambda) = O_1(\lambda) f_\xi(\lambda) O_1^*(\lambda) + O_2(\lambda)\Omega_2(\lambda)O_2^*(\lambda)$$

where

$$O_1, O_2 \; : \quad \text{eigenvectors of} \, f_y$$

$$\begin{pmatrix} f_\xi & 0 \\ 0 & \Omega_2 \end{pmatrix} \; : \quad \text{eigenvalues of} \, f_y$$

$$\xi_t = O_1^*(z) y_t \quad , \quad u_t = O_2 O_2^* y_t$$

Interpretation: Best (Frobenius norm) approximation of $f_y(\lambda)$ by rank r matrices, or best approximation of observations (y_t) by latent variables $(\hat{y}_t)$.

- Dynamic factor model with idiosyncratic noise
  Assumption: f_u is diagonal
  Basic idea: Separation of common and of individual components:
  Splitting property: The factors make the component-processes of (y_t) conditionally uncorrelated.
  Identifiability problems: Ledermann bound
  In general the factors are no functions of the observations.
  MLE; tests for the number of factors

- Generalized dynamic factor model: f_u is not necessarily diagonal, but the off-diagonal elements are "small". Identifiability only for $s \to \infty$ (and r const.). Approximation by PCA.

6. Nonlinear Systems

Nonlinear system identification is a word like “non-elephant zoology”

- Asymptotic theory for M-estimation in parametric classes of nonlinear (dynamic) systems. Party analogous to the case of linear models; no general structure theory available.

- Nonparametric estimation for nonlinear time series models, e.g. by kernel methods. Nonlinear autoregressions $y_t = g(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t$. Asymptotic theory, rates of convergence.

- Semi-nonparametric estimation, e.g. by dynamic neural nets. “Universal approximation properties” (nonlinear black box models).

- Special classes of nonlinear systems, e.g. GARCH type models.

- Chaos models: nonlinearity instead of stochasticity.

ARCH and GARCH models

Modeling of volatility clustering:

ARCH:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= c + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 \end{aligned}$$

where z_t is IID with $\mathbb{E}z_t = 0$ and $\mathbb{E}z_t^2 = 1$, $c > 0$, $\alpha_i \geq 0$.
Stationarity condition: $\sum_{i=1}^{p} \alpha_i < 1$.
Note that (ε_t) is white noise but $\mathbb{E}(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = c + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_p \varepsilon_{t-p}^2$

GARCH:

$$\begin{aligned} \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= c + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \end{aligned}$$

where, in addition, $\beta(z) = 1 - \sum_{i=1}^{p} \beta_i z^i \neq 0$ for $|z| \leq 1$ and $\beta_i \geq 0$.
Stationarity condition: $(\alpha_1 + \beta_1) + \ldots (\alpha_p + \beta_p) < 1$

Used for forecasting risk

- Estimation: MLE (ARCH case)

$$\hat{L}_T(c, \alpha_i, \beta_i) = \frac{1}{T}\sum_t \log(c + \sum_i \alpha_i \varepsilon_{t-i}^2) + \frac{1}{T}\sum_t \frac{\varepsilon_t^2}{(c + \sum_i \alpha_i \varepsilon_{t-i}^2)}$$

  The MLE's are obtained by numerical optimization.

- Tests for conditional heteroskedasticity.
  Testing for correlation in ε_t^2.
  Lagrange multiplier test (ARCH case: H_0: $\alpha_1 = \ldots = \alpha_p = 0$)

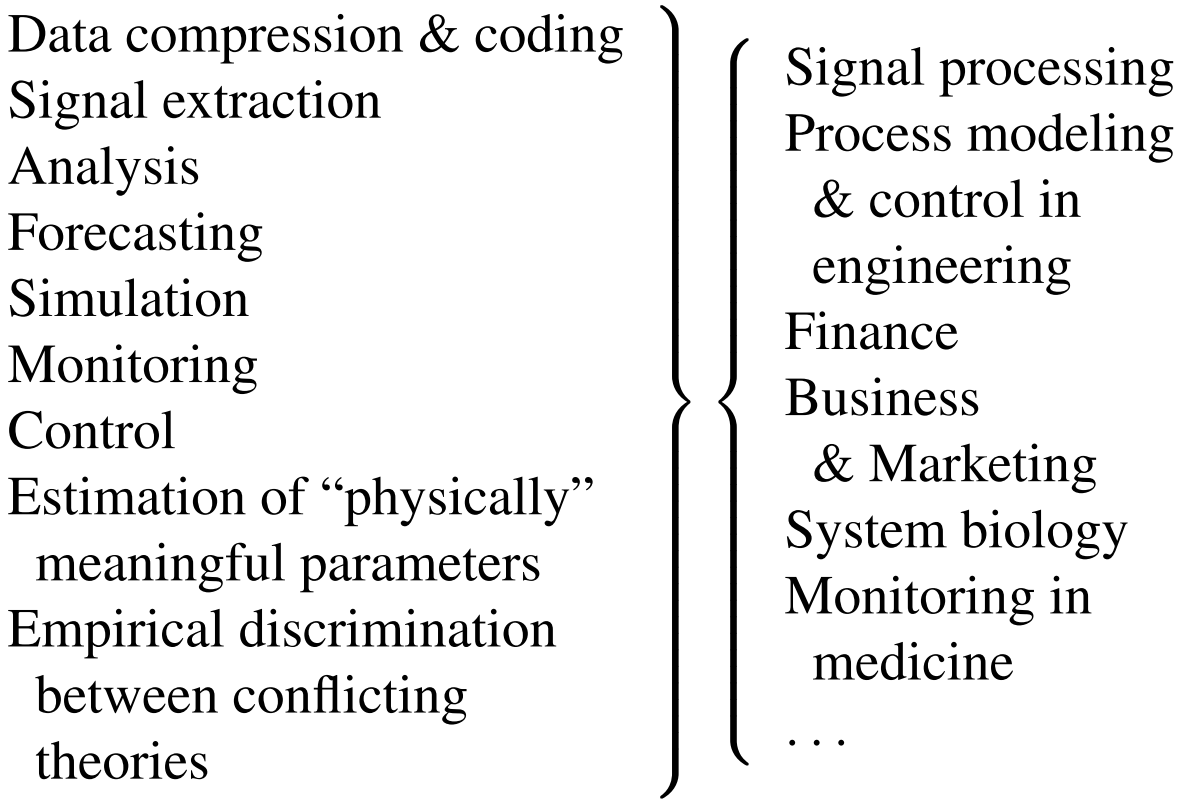
Great number of extensions.

7. Present State and Future Developments

PRESENT STATE:

- Theory and methods have reached a certain state of maturity. Large body of methods and theories available.
  Demand pull rather than theory push.
  Increasing fragmentation corresponding to different fields of application (data structure, model classes, prior knowledge).

- Boom in applications: Number of applications and areas of application are increasing.



"Component manufacturing"
Enabling technology, not very visible

- Different communities and multicultural:

  Econometrics
  Statistics
  Systems & Control
  Signal processing
  Intruders, e.g. neural nets
  Shifting boundaries

IMPORTANT PROBLEMS FOR THE FUTURE

- There are still major open problems in linear systems identification
- Highly structured systems, e.g. compartment models. Additional prior information such as mass-balancing.
- Nonlinear systems
- Spatio-temporal systems, PDE's
- Large data sets, high dimensional time series

- Improved model selection and regularization procedures
- Further automatization
- Hybrid procedures
- Use of symbolic computation

CHANCES AND DANGERS

The increasing number of applications poses challenges.

Will there be still a common body of theory and methods?

Danger of fragmentation and of becoming selfreferential.