Data driven modeling: Find a good model from noisy data.

- Model class: Set of all a priori feasible candidate systems

- Identification procedure: Attach a system from the model class to time series data $y_t$, $t = 1, \ldots, T$
  
  - Development of procedures
  - Evaluation: Asymptotic properties

Semi-nonparametric approach: Model specification leads to a finite dimensional model (sub)class
Three modules in semi-nonparametric 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 behaviour’ and ‘internal parameters’.

- **Estimation of real valued parameters**: Subclass is assumed to be given; parameter space is a subset of an Euclidean space and contains a nonvoid open set. Estimation e.g. by 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.
Linear systems

- $(\varepsilon_t)$: white noise $(s\text{-dimensional})$, $\mathbb{E}\varepsilon_t\varepsilon_t' = \Sigma$
- $(x_t)$: (observed) inputs $(m\text{-dimensional})$
- $(u_t)$: noise to output $(s\text{-dimensional})$, $\mathbb{E}x_tu_s' = 0$
- $(y_t)$: output $(s\text{-dimensional})$

$l(z)$ is the input-to-output transfer function and $k(z)$ the noise-to-output transfer function.

ECONOMETRIC FORECASTING AND HIGH-FREQUENCY DATA ANALYSIS, Singapore, May 2004
Relation between second moments and $(l, k, \Sigma)$

Relation between second (population) moments of the observations and $(l, k, \Sigma)$:

\[ f_{yx} = l \cdot f_x \]  \hspace{1cm} (1)

\[ f_y = l \cdot f_x \cdot l^* + \frac{1}{2\pi} k \cdot \Sigma \cdot k^* \]  \hspace{1cm} (2)

If $f_x > 0$, then $l = f_{yx} \cdot f_x^{-1}$. 

ECONOMETRIC FORECASTING AND HIGH-FREQUENCY DATA ANALYSIS, Singapore, May 2004
Main model classes for linear systems

AR(X) models: \( a(z)y_t = (d(z)x_t) + \varepsilon_t \)

- \((\varepsilon_t)\): white noise, \( \mathbb{E}\varepsilon_t\varepsilon'_t = \Sigma \), \( \mathbb{E}x_t\varepsilon'_t = 0 \)

- \( a(z) = \sum_{j=0}^{p} a_j z^j \), \( d(z) = \sum_{j=0}^{r} d_j z^j \)

- Integer parameters: \( p, r \)

- Real valued parameters: \( ((a_0, \ldots, a_p, d_0, \ldots, d_r), \Sigma) \); free parameters

Model class:
\[ \{(a_0, \ldots, a_p, d_0, \ldots, d_r) \in \mathbb{R}^{s^2(p+1)+sm(r+1)}|\text{det}(a(z)) \neq 0|z| \leq 1\} \times \Sigma, \]
\[ \Sigma \subset \mathbb{R}^{s(s+1)/2} \]

Relation betw. transfer functions and internal parameters: \( l = a^{-1}d \), \( k = a^{-1} \)

Stability condition: \( \text{det}(a(z)) \neq 0|z| \leq 1 \)

ECONOMETRIC FORECASTING AND HIGH-FREQUENCY DATA ANALYSIS, Singapore, May 2004
Main model classes for linear systems

ARMA(X) models: \( a(z)y_t = (d(z)x_t) + b(z)\varepsilon_t \)

- \((\varepsilon_t)\): white noise, \( \mathbb{E}\varepsilon_t\varepsilon'_t = \Sigma, \mathbb{E}x_t\varepsilon'_s = 0 \)
- \(a(z) = \sum_{j=0}^{p} a_jz^j, d(z) = \sum_{j=0}^{r} d_jz^j, b(z) = \sum_{j=0}^{q} b_jz^j \)
- Integer parameters: e.g. \(p, q, r\)
- Real valued parameters: \(((a_0, \ldots, a_p, b_0, \ldots, b_q, d_0, \ldots, d_r), \Sigma)\); free pars

Relation betw. transfer functions and internal parameters: \( l = a^{-1}d, k = a^{-1}b \)

Stability condition: \( \det(a(z)) \neq 0 |z| \leq 1 \)

Miniphase condition: \( \det(b(z)) \neq 0 |z| \leq (<>1) \)

Left coprimeness of \((a, b, d)\) and non-redundancy of dynamics \(a_0 = b_0\)
Main model classes for linear systems

State Space (StS) models (in innovations form): $s_t$ is the $n$-dimensional state

\begin{align}
    s_{t+1} &= A s_t + B \varepsilon_t (+L x_t) \\
    y_t &= C s_t + \varepsilon_t (+D x_t)
\end{align}

- Integer parameters: e.g. $n$
- Real valued parameters: $((A, B, C, L, D), \Sigma)$

Relation between transfer functions and internal parameters:

\begin{align*}
    l(z) &= D + C(z^{-1}I - A)^{-1}L, \\
    k(z) &= I + C(z^{-1}I - A)^{-1}B
\end{align*}

Stability condition: $|\lambda_{max}(A)| \leq 1$

Miniphase condition: $|\lambda_{max}(A - BC)| < (\leq) 1$

Minimality of $(A, B, C)$ (i.e. $n$ is minimal for given $k(z)$) $\iff$

\[ \text{rk}(B, AB, \ldots, A^{n-1}B) = \text{rk}(C', A'C', \ldots, (A')^{n-1}C') = n \]

The state $s_t$ is obtained by projecting the future of $(y_t)$ onto its past
Applications

In applications, AR(X) models still dominate because of their advantages

- no problems with non-identifiability

- maximum likelihood estimates are of least-squares type, asymptotically efficient and easy to calculate

Their disadvantages are:

- less flexible; more parameters may have to be estimated
Comparison ARMA and state-space systems

ARMA and StS models describe the same class of transfer functions.

Theorem:

- Every ARMA system and every StS system has a rational transfer function $k(z)$ that is causal and stable and satisfies $\det(k(z)) \neq 0$ $|z| \leq 1$.
- Conversely, for every rational, causal and stable transfer function $k(z)$ satisfying $\det(k(z)) \neq 0$ $|z| \leq 1$ there is an ARMA as well as a StS representation.

Relation between second moments and $(k, \Sigma)$: $f_y = \frac{1}{2\pi} k \cdot \Sigma \cdot k^*$

Note:

- Due to the stability and miniphase condition, $k$ corresponds to the Wold representation.
- For $\Sigma > 0$, $k(0) = I$, $k$ and $\Sigma$ are unique for given $f_y$.
- For identifiability it remains to give conditions such that $(a, b)$ or $(A, B, C)$ are unique for given $k$.
2.1 AR(X) identification (1)

This is classical.

Structure theory:

- For $\Sigma > 0$, two AR(X) systems $(\bar{a}, \bar{d})$ and $(a, d)$ are observationally equivalent if and only if there exists a nonsingular matrix $t$ such that $\bar{a} = ta$, $\bar{d} = td$, $\bar{\Sigma} = t\Sigma t'$.

- Thus identifiability is obtained by assuming $a_0 = I$ or by suitable 'structural' restrictions.

- Parameter space in the AR case:
  $$\Theta = \{ (a_1, \ldots, a_p) \in \mathbb{R}^{s^2p} | \text{det}(a(z)) \neq 0, |z| \leq 1 \} \times \Sigma$$
  open subset of $\mathbb{R}^{s^2p}$

- No 'bad' points in parameter space even for e.g. $a_p = 0$. 

ECONOMETRIC FORECASTING AND HIGH-FREQUENCY DATA ANALYSIS, Singapore, May 2004
2.1 AR(X) identification (2)

Estimation of real valued parameters:

- OLS type estimation for the $a_0 = I$ case (or for the just identifiable case).
- Estimators are consistent and asymptotically efficient.
- Simultaneous equations methods (such as TSLS) for the overidentifiable case.

Model selection:

- Estimation of $p$ and selection of inputs (subset selection); see below.
2.2 Structure theory for ARMA and state-space systems

Relation to internal parameters:

\[ k = a^{-1}(z)b(z) \text{ or } k(z) = \sum_{j=0}^{\infty} k_j z^j \text{ where } k_j = CA^{j-1}B \text{ for } j \geq 1, k_0 = I. \]

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

\[ M(n) \subset U_A: \text{ Set of all transfer functions of order } n. \]

\[ T_A: \text{ Set of all } A, B, C \text{ for fixed } s, \text{ but } n \text{ variable, satisfying stability + miniphase assumption.} \]

\[ S(n) \subset T_A: \text{ Subset of all } (A, B, C) \text{ for fixed } n. \]

\[ S_m(n) \subset S(n): \text{ Subset of all minimal } (A, B, C). \]
2.2 Structure theory for ARMA and state-space systems

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

\( \pi \) 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 \( T_\alpha \).
2.2 Structure theory for ARMA and state-space systems
2.2 Structure theory for ARMA and state-space systems

Desirable properties of parametrizations:

- $U_A$ and $T_A$ are broken into bits, $U_\alpha$ and $T_\alpha$, $\alpha \in I$, such that $k$ restricted to $T_\alpha$:
  
  $\pi|T_\alpha : T_\alpha \rightarrow U_\alpha$ is bijective. Then there exists a parametrization $\psi_\alpha : U_\alpha \rightarrow T_\alpha$ such that
  
  $\psi_\alpha(\pi(A, B, C)) = (A, B, C)$  $\forall (A, B, C) \in T_\alpha$.

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

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

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

- $\bigcup_{\alpha \in I} U_\alpha$ 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 $U_\alpha$ of different dimension. Nice free parameters vs. nice spaces of free parameters.

- "Overlapping description" of the manifold $M(n)$ by local coordinates.
2.2 Structure theory for ARMA and state-space systems

- "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 $U_\alpha$.
2.2 Structure theory for ARMA and state-space systems

Problem: Numerical properties of parametrizations

Different parametrizations:

\[ \psi_1 : U_1 \to T_1 \subset T_A, \quad \psi_2 : U_2 \to T_2 \subset T_A \]

For the asymptotic analysis, in the case that \( U_1 \supset U_2 \), \( U_2 \) contains a nonvoid open (in \( U_1 \)) set and \( k_0 \in U_2 \), we have:

STATISTICAL ANALYSIS ("real world"):

- no essential differences: coordinate free consistency
- different asymptotic distributions, but we know the transformation

NUMERICAL ANALYSIS ("integer world"):

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

ECONOMETRIC FORECASTING AND HIGH-FREQUENCY DATA ANALYSIS, Singapore, May 2004
2.2 Structure theory for ARMA and state-space systems

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:
2.2 Structure theory for ARMA and state-space systems

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: $\bar{a} = a$, $\bar{b} = tb$, $\bar{c} = ct^{-1}$, $t \in \mathbb{R} \setminus \{0\}$
2.3 Estimation for a given subclass

We here assume that $U_\alpha$ is given.

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

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

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

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

Many procedures, at least asymptotically, commence from sample $2^{nd}$ 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 $Ts$; $\hat{k_T}$

IDENTIFICATION:

Projection step (model reduction): important for statistical qualities.

Realization step.
2.3 Estimation for a given subclass

- M-estimators:
  \[ \hat{\theta}_T = \arg\min L_T(\theta; y_1, \ldots, y_T) \]

- Direct procedures: Explicit functions.
2.3 Estimation for a given subclass

GAUSSIAN MAXIMUM LIKELIHOOD:

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

where \( y(T) = (y'_1, \ldots, y'_T)' \), \( \Gamma_T(\theta) = \mathbb{E}y(T; \theta)y'(T; \theta) \), \( \hat{\theta}_T = \arg\min_{\theta \in \Theta} \hat{L}_T(\theta) \)

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

Whittle likelihood:

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

where \( I(\lambda) \) is the periodogram.
2.3 Estimation for a given subclass

EVALUATION:

- Coordinate free consistency: for \( k_0 \in U_\alpha \) and 
  \[
  \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T-s} \varepsilon_{t+s} \varepsilon_t' = \delta_{0,s} \Sigma_0 \text{ a.s. for } s \geq 0 \text{ we have } \hat{k}_T \to k_0 \text{ a.s. and } \hat{\Sigma}_T \to \Sigma_0 \text{ a.s.}
  \]

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

Noncompact parameter spaces:

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

- \( L \) has a unique minimum at \( k_0, \Sigma_0 \).
- \( (\hat{k}_T, \hat{\Sigma}_T) \) enters a compact set, uniform convergence in (5).

- Generalized, coordinate free consistency for \( k_0 \not\in \overline{U}_\alpha \), \((\hat{k}_T, \hat{\Sigma}_T) \to D \text{ a.s } D \): Set of all best approximants to \( k_0, \Sigma_0 \) in \( \overline{U}_\alpha \times \Sigma \).

- Consistency in coordinates: \( \psi_\alpha(\hat{k}_T) = \hat{\tau}_T \to \tau_0 = \psi_\alpha(k_0) \text{ a.s.} \)
2.3 Estimation for a given subclass

- CLT:
  Under $\mathbb{E}(\varepsilon_t | \mathcal{F}_{t-1}) = 0$ and $\mathbb{E}(\varepsilon_t \varepsilon'_t | \mathcal{F}_{t-1}) = \Sigma_0$, we have
  \[ \sqrt{T} (\hat{\tau}_T - \tau_0) \overset{d}{\longrightarrow} N(0, V) \]

  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)
2.3 Estimation for a given subclass

HOVERE 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.
2.4 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 $A(n) = \log \det \hat{\Sigma}_T(n) + 2n s \cdot c(T) \cdot T^{-1}$ where $\hat{\Sigma}_T(n)$ is the MLE for $\Sigma_0$ over $\bar{M}(n) \times \Sigma$; $c(T) = 2$: AIC criterion; $c(T) = c \cdot \log T$, $c \geq 1$: BIC criterion

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

Statistical evaluation: $\hat{n}_T$ is consistent for $\lim_{T \to \infty} \frac{\log T}{c(T)} = 0$, $\lim \inf_{T \to \infty} \frac{c(T)}{\log T} > 0$

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

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:

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

\[
T_0 = \{(0, 0)\}
\]

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

\[
T_2 = T - (T_0 \cup T_1)
\]
2.4 Model selection

Bayesian justification:

- Positive priors for all classes, otherwise MLE is asymptotically normal

- Certain properties of $U_\alpha$, $\alpha \in I$ are needed, e.g. for BIC to give consistent estimators: closure nestedness, e.g. $n_1 > n_2 \Rightarrow M(n_1) \supset M(n_2)$

Main open question:

- Optimal tradeoff between dimension and "number" of pieces.
2.4 Model selection

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 $T_p = \{(a_1, \ldots, a_p) \in \mathbb{R}^p | \text{stability}\}$

The estimator (LS) for given $p$ is $\hat{\tau}_p = (X(p)'X(p))^{-1}X(p)y$

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.