Estimating linear dynamical models using subspace methods

Dietmar Bauer

Institut f. Econometrics, Operations Research and System Theory, TU Wien, Austria
currently: Cowles Foundation, Yale University
Dietmar.Bauer@tuwien.ac.at

General setup: We are given a dataset of (multivariable) input ($u_t \in \mathbb{R}^m$) and (multivariable) output ($y_t \in \mathbb{R}^s$) measurements (classified) at discrete time instants $t = 1, \ldots, T$.

We want to estimate a linear dynamical model describing the data set and obtain accuracy measures for the purposes of

- prediction
- simulation
- validation of Theory

We do not have any knowledge on process, noise, ...

General setup: We are given a dataset of (multivariable) input ($u_t \in \mathbb{R}^m$) and (multivariable) output ($y_t \in \mathbb{R}^s$) measurements (classified) at discrete time instants $t = 1, \ldots, T$.

We want to estimate a linear dynamical model describing the data set and obtain accuracy measures for the purposes of

- prediction
- simulation
- validation of Theory

We do not have any knowledge on process, noise, ...

Simplest linear dynamic model: ARX model

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + b_0 u_t + \cdots + b_q u_{t-q} + v_t$$

Estimation: least squares fitting, explicit solution.

General setup: We are given a dataset of (multivariable) input ($u_t \in \mathbb{R}^m$) and (multivariable) output ($y_t \in \mathbb{R}^s$) measurements (classified) at discrete time instants $t = 1, \ldots, T$.

We want to estimate a linear dynamical model describing the data set and obtain accuracy measures for the purposes of

- prediction
- simulation
- validation of Theory

We do not have any knowledge on process, noise, ...

Simplest linear dynamic model: ARX model

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + b_0 u_t + \cdots + b_q u_{t-q} + v_t$$

Estimation: least squares fitting, explicit solution.

Next level of complexity: ARMAX

$$v_t = \varepsilon_t + c_1 \varepsilon_{t-1} + \cdots + c_r \varepsilon_{t-r}$$

Assumptions:

- ▷ $(\varepsilon_t)_{t\in\mathbb{Z}}$ is white noise process (square integrable, mean zero, constant variance, uncorrelated for different time instants).
- ▷ $(u_t)_{t\in\mathbb{Z}}$ stationary
- ▷ for $a(z) = I - a_1 z - \cdots - a_p z^p$ (z complex variable) holds $\det a(z) \neq 0, |z| \leq 1$ (stability assumption)

Then for $b(z) = b_0 + \cdots + b_q z^q, c(z) = I + c_1 z + \cdots + c_r z^r$

$$a(z)^{-1}b(z) = l(z) = \sum_{j=0}^{\infty} L_j z^j, \quad a(z)^{-1}c(z) = k(z) = \sum_{j=0}^{\infty} K_j z^j$$

are power series expansions converging on $|z| \leq 1$.

Define: $y_t = \sum_{j=0}^{\infty} L_j u_{t-j} + \sum_{j=0}^{\infty} K_j \varepsilon_{t-j}$.
Convergence in mean square guaranteed.

Assumptions:

- ▷ $(\varepsilon_t)_{t\in\mathbb{Z}}$ is white noise process (square integrable, mean zero, constant variance, uncorrelated for different time instants).
- ▷ $(u_t)_{t\in\mathbb{Z}}$ stationary
- ▷ for $a(z) = I - a_1 z - \cdots - a_p z^p$ (z complex variable) holds $\det a(z) \neq 0, |z| \leq 1$ (stability assumption)

Then for $b(z) = b_0 + \cdots + b_q z^q, c(z) = I + c_1 z + \cdots + c_r z^r$

$$a(z)^{-1}b(z) = l(z) = \sum_{j=0}^{\infty} L_j z^j, \quad a(z)^{-1}c(z) = k(z) = \sum_{j=0}^{\infty} K_j z^j$$

are power series expansions converging on $|z| \leq 1$.

Define: $y_t = \sum_{j=0}^{\infty} L_j u_{t-j} + \sum_{j=0}^{\infty} K_j \varepsilon_{t-j}$.
Convergence in mean square guaranteed.

Then

- $(y_t)_{t\in\mathbb{Z}}$ is stationary
- $(y_t)_{t\in\mathbb{Z}}$ fulfills the ARMAX VDEs for $t \in \mathbb{Z}$.

$$y_t - a_1 y_{t-1} - \cdots - a_p y_{t-p}$$

$$= \sum_{j=0}^{\infty} L_j u_{t-j} + K_j \varepsilon_{t-j} - a_1 (L_j u_{t-j-1} + K_j \varepsilon_{t-j-1}) -$$
$$\cdots - a_p (L_j u_{t-j-p} - K_j \varepsilon_{t-j-p})$$
$$= L_0 u_t + K_0 \varepsilon_t + (L_1 - a_1 L_0) u_{t-1} + (K_1 - a_1 K_0) \varepsilon_{t-1} +$$
$$\cdots + (L_j - a_1 L_{j-1} - \ldots a_p L_{j-p}) u_{t-j}$$
$$+ (K_j - a_1 K_{j-1} - \ldots a_p K_{j-p}) \varepsilon_{t-j} + \ldots$$

Coefficients are exactly the coefficients obtained in a comparison of coefficients in

$$a(z)k(z) = c(z), \quad a(z)l(z) = b(z)$$

$(k(z), l(z))$ pair of transfer functions describes the input/output mapping as well as the dynamics of the noise in the system.

$$y_t - a_1 y_{t-1} - \cdots - a_p y_{t-p}$$

$$= \sum_{j=0}^{\infty} L_j u_{t-j} + K_j \varepsilon_{t-j} - a_1 (L_j u_{t-j-1} + K_j \varepsilon_{t-j-1}) -$$
$$\cdots - a_p (L_j u_{t-j-p} - K_j \varepsilon_{t-j-p})$$
$$= L_0 u_t + K_0 \varepsilon_t + (L_1 - a_1 L_0) u_{t-1} + (K_1 - a_1 K_0) \varepsilon_{t-1} +$$
$$\cdots + (L_j - a_1 L_{j-1} - \ldots a_p L_{j-p}) u_{t-j}$$
$$+ (K_j - a_1 K_{j-1} - \ldots a_p K_{j-p}) \varepsilon_{t-j} + \ldots$$

Coefficients are exactly the coefficients obtained in a comparison of coefficients in

$$a(z)k(z) = c(z), \quad a(z)l(z) = b(z)$$

$(k(z), l(z))$ pair of transfer functions describes the input/output mapping as well as the dynamics of the noise in the system.

W.r.o.g. assume $\det c(z) \neq 0, |z| < 1$ (minimum-phase).

Example: scalar MA(1): $v_t = \varepsilon_t + c_1 \varepsilon_{t-1}$.

$$\mathbb{E} v_t v_{t-j} = \begin{cases} \sigma^2 (1 + c_1^2) & , \quad j = 0, \\ \sigma^2 c_1 & , \quad j = \pm 1, \\ 0 & , \quad \text{else} \end{cases}$$

Get the same covariance for $\eta_t + 1/c_1 \eta_{t-1}$, where η_t has variance $\sigma^2 c_1^2$.

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + b_0 u_t + \ldots b_q u_{t-q} + \varepsilon_t + \cdots + c_r \varepsilon_{t-r}$$

$$x_t' = [y_{t-1}', \ldots, y_{t-p}', u_{t-1}', \ldots, u_{t-q}', \varepsilon_{t-1}', \ldots, \varepsilon_{t-r}']'$$

Define:

$$A = \left[\begin{array}{cccc|cccc|ccc} a_1 & \ldots & \ldots & a_p & b_1 & \ldots & \ldots & b_q & c_1 & \ldots & c_r \\ I & 0 & & & & & & & & & 0 \\ & \ddots & & & & & & & & & \\ 0 & \ddots & 0 & & & & & & & & \\ & & I & 0 & & & & & & & \\ \hline & & & & 0 & 0 & & & & & \\ & & & & I & 0 & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & I & 0 & & & \\ \hline & & & & & & & & 0 & & \\ & & & & & & & & I & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & I & 0 \end{array}\right]$$

$$B' = [b_0', \ldots, 0, I, 0, \ldots], K' = [I, 0, \ldots, 0, I, 0, \ldots],$$

$$C = [a_1, \ldots, a_p, b_1, \ldots, b_q, c_1, \ldots, c_r], D = b_0$$

Follows:

$$\begin{aligned} y_t &= Cx_t + Du_t + \varepsilon_t \\ x_{t+1} &= Ax_t + Bu_t + K\varepsilon_t \end{aligned}$$

LINEAR, TIME INVARIANT, DISCRETE TIME, FINITE DIMENSIONAL, STATE SPACE SYSTEMS.

$$
\begin{aligned}
y_t &= Cx_t + Du_t + \varepsilon_t \\
&= C(Ax_{t-1} + Bu_{t-1} + K\varepsilon_{t-1}) + Du_t + \varepsilon_t \\
&= CA(Ax_{t-2} + Bu_{t-2} + K\varepsilon_{t-2}) + CBu_{t-1} + \\
&\quad CK\varepsilon_{t-1} + Du_t + \varepsilon_t \\
&= \dots \\
&= Du_t + \sum_{j=1}^{\infty} CA^{j-1}Bu_{t-j} + \varepsilon_t + \sum_{j=1}^{\infty} CA^{j-1}K\varepsilon_{t-j} \\
&= \sum_{j=0}^{\infty} L_j u_{t-j} + \sum_{j=0}^{\infty} K_j \varepsilon_{t-j}
\end{aligned}
$$

Therefore

- $K_0 = I, K_j = CA^{j-1}K, j > 0$
- $L_0 = D, L_j = CA^{j-1}B, j > 0$

<u>Transferfunction:</u>

$$k(z) = \sum_{j=0}^{\infty} K_j z^j = I + \sum_{j=0}^{\infty} CA^j z^{j+1} K = I + zC(I - zA)^{-1}K$$
$$l(z) = \sum_{j=0}^{\infty} L_j z^j = D + zC(I - zA)^{-1}B$$

K_j, L_j are called *impulse response coefficients*.

Exactly the rational transfer functions correspond to ARMAX systems.
For each ARMAX system there exists a state space representation.
For each (finite dimensional) state space system the corresponding transfer functions are rational.

ARMAX and state space systems are two different representations of the same mathematical object, the transfer function!

Kronecker Theorem

A pair of transfer functions $(k(z), l(z))$ is rational in z if and only if the rank of the Hankel matrix $\mathscr{H}$ of its power series coefficients is of finite rank n.

Here

$$\mathscr{H} = \begin{bmatrix} [K_1, L_1] & [K_2, L_2] & [K_3, L_3] & \dots \\ [K_2, L_2] & [K_3, L_3] & & \\ [K_3, L_3] & & & \\ \vdots & & & \end{bmatrix}$$

Uniqueness of State Space Representations

Question: Given a pair of rational transfer functions $(k(z), l(z))$. Can we describe the state space systems that correspond to this pair?

Minimality:

- ▷ Observability: $\mathscr{O} = [C', A'C', (A^2)'C', \cdots]'$
- ▷ Controllability: $\mathscr{C} = [[B,K], A[B,K], A^2[B,K], \cdots]$

(A,B,C,D,K) minimal $\Leftrightarrow$ rank $\mathscr{O}$ = rank $\mathscr{C}$ = dim(x_t).
Note: $\mathscr{H} = \mathscr{O}\mathscr{C}$!
If a system is not minimal, the state dimension can be reduced without changing the corresponding pair of transfer functions.

Uniqueness of State Space Systems: Under the minimality constraint, unique up to the choice of the basis:

Two minimal state space systems (A,B,C,D,K) and $(\tilde{A},\tilde{B},\tilde{C},\tilde{D},\tilde{K})$ are *observationally equivalent,* if and only if there exists a nonsingular transformation matrix S, such that

$$\tilde{A} = SAS^{-1}, \tilde{B} = SB, \tilde{C} = CS^{-1}, \tilde{D} = D, \tilde{K} = SK$$

Stability assumption: $|\lambda_{max}(A)| < 1$ for a minimal system (A,B,C,D,K).
Minimum-phase assumption: $|\lambda_{max}(A-KC)| \leq 1$.

Let $S_n = \{(A,B,C,D,K)$, minimal, stable, minimum-phase, order $n\}$.
Let $\pi : S_n \rightarrow M_n$:

$$(A,B,C,D,K) \mapsto (I+zC(I-zA)^{-1}K, D+zC(I-zA)^{-1}B)$$

Stability assumption: $|\lambda_{max}(A)| < 1$ for a minimal system (A,B,C,D,K).
Minimum-phase assumption: $|\lambda_{max}(A-KC)| \leq 1$.

Let $S_n = \{(A,B,C,D,K)$, minimal, stable, minimum-phase, order $n\}$.
Let $\pi : S_n \rightarrow M_n$:

$$(A,B,C,D,K) \mapsto (I+zC(I-zA)^{-1}K, D+zC(I-zA)^{-1}B)$$

Question: How can we describe M_n parsimoniously?

Answer: Canonical form:
Bijective mapping $M_n \rightarrow S_n$ that assigns one state space system to each pair of transfer function $(k(z),l(z))$.

Essentially selects one particular system from each class of observationally equivalent systems.

A canonical form hence can be seen as a subset of S_n.

Canonical forms are used to construct

Parameterization: Mapping from some real parameter set to M_n i.e. $\forall \theta \in \Theta$

$$\begin{aligned} \phi(\theta) &= (A(\theta), B(\theta), C(\theta), D(\theta), K(\theta)) \\ (k(z;\theta), l(z;\theta)) &= \pi(\phi(\theta)) \end{aligned}$$

where usually $\phi(\theta)$ lies in the image of the canonical form.

- Highly nonlinear mapping.
- for multi-output systems there exists no continuous parameterization of M_n: Usually pieces of M_n are defined, which are parameterized.

Canonical forms are used to construct

Parameterization: Mapping from some real parameter set to M_n i.e. $\forall \theta \in \Theta$

$$\begin{aligned} \phi(\theta) &= (A(\theta), B(\theta), C(\theta), D(\theta), K(\theta)) \\ (k(z;\theta), l(z;\theta)) &= \pi(\phi(\theta)) \end{aligned}$$

where usually $\phi(\theta)$ lies in the image of the canonical form.

- Highly nonlinear mapping.
- for multi-output systems there exists no continuous parameterization of M_n: Usually pieces of M_n are defined, which are parameterized.

Number of parameters needed, heuristic argument:
counting entries in (A,B,C,D,K): $n^2+nm+sn+sm+ns$
- degree of freedom in choice of state basis (n^2):
$2ns+m(n+s)$

Assume the probabilistic model is the state space model with *Gaussian* innovation sequence ε_t with mean zero and variance $\Omega > 0$ (additional $s(s+1)/2$ parameters).

Parameter set: $\theta \in \Theta \subset \mathbb{R}^{ns+sm+nm+ns+s(s+1)/2}$.

Let $Y_{1,T}^+ = [y_1', \cdots, y_T']'$ and $U_{1,T}^+ = [u_1', \cdots, u_T']'$

-2 times the Log - Likelihood of $Y_{1,T}^+$:

$$L(Y_{1,T}^+; \theta, U_{1,T}^+) = \log\det\Gamma_T(\theta) + \left[\tilde{Y}_{1,T}^+(\theta)\right]' \Gamma_T(\theta)^{-1} \left[\tilde{Y}_{1,T}^+(\theta)\right]$$

$\tilde{Y}_{1,T}^+(\theta) = Y_{1,T}^+ - \mathscr{L}_T(\theta) U_{1,T}^+$
$\mathscr{L}_T(\theta)$ has
$[C(\theta)A(\theta)^{i-2}B(\theta), \cdots, C(\theta)B(\theta), D(\theta), 0, \cdots]$ as its i-th row.
$\Gamma_T(\theta)$ ... covariance of $\tilde{Y}_{1,T}^+(\theta)$ according to θ.

ML estimate:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} L(Y_{1,T}^+; \theta, U_{1,T}^+)$$

Notes:

- ▷ Assumptions on the parameter space have to be imposed in order to ensure the feasibility of this definition.
- ▷ In order to find the maximum of the likelihood function, all pieces of M_n have to be searched in

principle. However, there is one generic piece, which usually suffices.

- ▷ Asymptotic properties well studied: Hannan & Deistler (1988) contains consistency, asymptotic normality, LIL, order estimation methods and their consistency, procedures to obtain initial values inclusively analysis, approximation properties, ...
- ▷ Actual calculation: Using the Kalman filter recursions for the recursive evaluation of the inverse helps in numerical calculation of the criterion function.
- ▷ Approximations: Neglecting the effect of initial values one arrives at a least squares type of criterion:

$$L(Y_{1,T}^{+};\theta,U_{1,T}^{+}) \approx -T\log\det(\Omega) - \sum_{t=1}^{T} \varepsilon_t(\theta)'\Omega^{-1}\varepsilon_t(\theta)$$

  Known as prediction error methods.
- ▷ Nonlinear optimization problem: Starting values? Local optima? Numerical problems for high dimensional parameter spaces

This is motivation to search for different estimation schemes!

Center of attention in subspace methods is the state:

Inserting $\varepsilon_t = y_t - Cx_t - Du_t$ one obtains for given (A,B,C,D,K), that (strict minimum-phase assumption!)

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_{t-1} + K\varepsilon_{t-1} \\ &= Ax_{t-1} + Bu_{t-1} + K(y_{t-1} - Cx_{t-1} - Du_{t-1}) \\ &= (A-KC)x_{t-1} + (B-KD)u_{t-1} + Ky_{t-1} \\ &= \Sigma_{j=1}^{\infty} \bar{A}^{j-1}\bar{B}u_{t-j} + \bar{A}^{j-1}Ky_{t-j} \\ &= \mathscr{K} Z_t^- \end{aligned}$$

where $\bar{A} = (A-KC), \bar{B} = B - KD.$

$$\mathscr{K} = [[K,\bar{B}], \bar{A}[K,\bar{B}], \cdots]$$

$$Z_t^- = [y_{t-1}', u_{t-1}', y_{t-2}', u_{t-2}', \cdots]'$$

denotes the vector of the whole past of $z_t = [y_t', u_t']'$.

Therefore: The state x_t at time t is contained in the past of the joint process z_t, i.e. $x_t \in \sigma\{z_{t-1}, z_{t-2}, \cdots\}$.

Furthermore:

$$x_t = \mathscr{K}_p Z_{t,p}^- + \bar{A}^p x_{t-p}$$

where

$$\mathscr{K}_p = [[K,\bar{B}], \cdots, \bar{A}^{p-1}[K,\bar{B}]], Z_{t,p}^- = [y_{t-1}', u_{t-1}', \cdots, u_{t-p}]'$$

Prediction of y_t: in mean square sense

$$\begin{aligned} y_{t+j} &= Cx_{t+j} + Du_{t+j} + \varepsilon_{t+j} \\ &= CAx_{t+j-1} + CBu_{t+j-1} + Du_{t+j} + CK\varepsilon_{t+j-1} + \varepsilon_{t+j} \\ &= \cdots \\ &= CA^j x_t + \sum_{l=0}^{j} L_l u_{t+j-l} + \sum_{l=0}^{j} K_l \varepsilon_{t+j-l} \end{aligned}$$

Meaning: Let $Z_t^- = [z'_{t-1}, z'_{t-2}, \cdots]', U_t^+ = [u'_t, u'_{t+1}, \cdots]'$. Then prediction of y_{t+j} on the basis of Z_t^- and U_t^+ is equal to

$$y_{t+j|t} = CA^j x_t + \sum_{l=0}^{j} L_l u_{t+j-l}$$

Concluding:

- ▷ The state lies in the past of the joint process.
- ▷ Given the whole input series, the state is sufficient statistic for prediction of the future of y_t based on the input process u_t and the past of y_t.

Some people say: The state is an interface between the past and the future.

Central equation for subspace algorithms:

$$\begin{aligned} Y_{t,f}^{+} &= \mathscr{O}_f \mathscr{K}_p Z_{t,p}^{-} + \mathscr{U}_f U_{t,f}^{+} + \mathscr{O}_f \bar{A}^p x_{t-p} + \mathscr{E}_f E_{t,f}^{+} \\ &= \beta_z Z_{t,p}^{-} + \beta_u U_{t,f}^{+} + N_{t,f}^{+} \end{aligned}$$

$$[Y_{t,f}^{+}, U_{t,f}^{+}, E_{t,f}^{+}] = \left[\begin{array}{c|c|c} y_t & u_t & \varepsilon_t \\ \vdots & \vdots & \vdots \\ y_{t+f-1} & u_{t+f-1} & \varepsilon_{t+f-1} \end{array}\right],$$

$$Z_{t,p}^{-} = \begin{bmatrix} z_{t-1}' & \cdots & z_{t-p}' \end{bmatrix}', z_t = [y_t', u_t']',$$

$$\mathscr{O}_f = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{bmatrix}, \mathscr{E}_f = \begin{bmatrix} I & 0 & & \\ CK & I & \ddots & \\ \vdots & \ddots & \ddots & 0 \\ CA^{f-2}K & \cdots & CK & I \end{bmatrix}$$

$$\mathscr{U}_f = \begin{bmatrix} D & 0 & & \\ CB & D & \ddots & \\ \vdots & \ddots & \ddots & 0 \\ CA^{f-2}B & \cdots & CB & D \end{bmatrix}$$

$$\mathscr{K}_p = [[K, B-KD], \cdots, \bar{A}^{p-1}[K, B-KD]], \bar{A} = A - KC.$$

Most subspace algorithms for linear discrete time systems share some common structure:

Central equation suggests to use LS in

$$Y_{t,f}^{+} = \beta_z Z_{t,p}^{-} + \beta_u U_{t,f}^{+} + \text{residuals} \qquad (*)$$

Basic Outline:

1. Estimate $[\mathscr{O}_f \mathscr{K}_p, \mathscr{U}_f]$ using the above equation $(*)$. This results in an estimate $\hat{\beta}_z \in \mathbb{R}^{fs \times p(m+s)}$ (and an estimate $\hat{\beta}_u$).
2. Approximate the typically full rank matrix $\hat{\beta}_z$ by a rank n matrix $\hat{\mathscr{O}}_f \hat{\mathscr{K}}_p$.
3. Estimate the state as $\hat{x}_t = \hat{\mathscr{K}}_p Z_{t,p}^{-}$ and use LS in the system equations to estimate the system matrices.

Different algorithms use various ways to perform the three basic steps. Most popular methods: `MOESP` (Verhaegen), `N4SID` (DeMoor & VanOverschee), `CCA` (Larimore). Some of these use different Step 3.

Rest of the talk: Present the Larimore algorithm as a prototype and give some of the known results on the asymptotic properties.

Step 1: LS fit, autoregressive structure.

$$Y^+_{t,f} = \beta_z Z^-_{t,p} + \beta_u U^+_{t,f} + N^+_{t,f}$$

Data available only for $t = p+1, \ldots, T-f$.

Choosing initial and end effects:

- set initial effects to zero: LS fit can be expressed as a nonlinear function of the estimated covariance sequence of z_t. Immediate to obtain
  - robust estimators
  - estimators for data sets with missing data
  - (recursive estimators)
- use only available data: avoids bias, LS fit via QR decomposition.

Both approaches can be implemented to handle large data sets in reasonable time.

Step 2: Rank n approximation:

Let $\hat{W}_f^+ \in \mathbb{R}^{fs \times fs}$ and $\hat{W}_p^- \in \mathbb{R}^{(s+m)p \times (s+m)p}$ be two nonsingular (a.s.) matrices. Then consider the SVD

$$\hat{W}_f^+ \hat{\beta}_z \hat{W}_p^- = \hat{U}\hat{\Sigma}\hat{V}^T = \hat{U}_n \hat{\Sigma}_n \hat{V}_n^T + \hat{R}$$

Results in approximation

$$\hat{\mathscr{O}}_f \hat{\mathscr{K}}_p = [(\hat{W}_f^+)^{-1} \hat{U}_n \hat{\Sigma}_n^{1/2}][\hat{\Sigma}_n^{1/2} \hat{V}_n^T (\hat{W}_p^-)^{-1}]$$

Commonly used weighting matrices (comp. Peternell, 1995, Van Overschee 1995)

Alg.	$\hat{W}_f^+$	$\hat{W}_p^-$
CCA	$(\hat{\Gamma}_f^{+,\Pi})^{-1/2}$	$(\hat{\Gamma}_p^{-,\Pi})^{1/2}$
N4SID	I_{fs}	$(\hat{\Gamma}_p^-)^{1/2}$
MOESP	I_{fs}	$(\hat{\Gamma}_p^{-,\Pi})^{1/2}$
Freq. W.	$(K_W(i-j))_{i,j}$	$(\hat{\Gamma}_p^{-,\Pi})^{1/2}$

$\hat{\Gamma}_f^{+,\Pi} = \hat{\Gamma}_{y,y} - \hat{\Gamma}_{y,u}\hat{\Gamma}_{u,u}^{-1}\hat{\Gamma}_{u,y}, \hat{\Gamma}_p^{-,\Pi} = \hat{\Gamma}_{z,z} - \hat{\Gamma}_{z,u}\hat{\Gamma}_{u,u}^{-1}\hat{\Gamma}_{u,z}$, where u stands for $U_{t,f}^+$, y for $Y_{t,f}^+$ and z for $Z_{t,p}^-$. $K_W(i)$ ...impulse response sequence of some transfer function $k_W(z) = \sum_{j=0}^{\infty} K_W(j) z^j$ having a nonsingular constant term $K_W(0)$ and $K_W(i) = 0, i < 0$.

CCA in the no inputs case:

SVD on

$$\left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,f}^{+}, Y_{t,f}^{+}\right)^{-1/2} \left(\frac{1}{T}\sum_{t=1}^{T} Y_{t,f}^{+}, Z_{t,p}^{-}\right) \left(\frac{1}{T}\sum_{t=1}^{T} Z_{t,p}^{-}, Z_{t,p}^{-}\right)^{-1/2}$$

Estimates the canonical correlations between $Y_{t,f}^{+}$ and $Z_{t,p}^{-}$, hence the name.

Remark: The name 'subspace methods' also is due to this step of estimating the subspace spanned by the columns of the observability matrix $\mathscr{O}_f$.

Notation is not unified. Be careful, if you read papers. Labels like CCA and N4SID are used for different algorithms.

In this step the order has to be specified.

Basically problem of determining the rank of a perturbed matrix $\hat{\beta}_z$.

Nonstandard, since the perturbation is also singular.

Can be done based on the estimated singular values $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \ldots$.

Define:

$$NIC(n) = \|\hat{R}_n\|^2 + \frac{d(n)C_T}{T}$$

where $d(n)$ denotes number of parameters,
$C_T > 0, C_T/T \to 0$ a penalty term.

- Peternell (1995): Frobenius norm $\sum_{j=n+1}^{M} \hat{\sigma}_j^2$
- Bauer (1998): Two norm $\hat{\sigma}_{n+1}^2$
- Camba-Mendez and Kapetanios (2001):
  $\|\hat{R}_n\|^2 = -\sum_{j=n+1}^{M} \log(1-\hat{\sigma}_j^2)$

By choosing the penalty C_T large enough, consistency can be shown.

Also rank testing procedures exist (sequential testing approaches).

Underresearched area!

Intuitive idea: Use rank restricted regression techniques.

Impose rank restriction on $\beta_z = \mathscr{O}_f \mathscr{K}_p$:

1. Least squares estimation: Estimation problem:
   $N_{t,f}^+ = Y_{t,f}^+ - \beta_z Z_{t,p}^- - \beta_u U_{t,f}^+$

   $$[\tilde{\beta}_z, \hat{\beta}_u] = \arg\min \operatorname{tr}\left[W \frac{1}{T} \sum_{t=1}^{T} N_{t,f}^+ (N_{t,f}^+)' \right]$$

   under $\operatorname{rank}[\beta_z] = n$.

2. ML estimation under Gaussian i.i.d. distributed noise: Under $\operatorname{rank}[\beta_z] = n$

   $$[\tilde{\beta}_z, \hat{\beta}_u] = \arg\min \quad \log\det\left[\frac{1}{T} \sum_{t=1}^{T} N_{t,f}^+ (N_{t,f}^+)' \right]$$

Solution to ML problem is equivalent to solving the LS problem for

$$W = \left(T^{-1} \sum_{t=1}^{T} Y_{t,f}^{+,\Pi} (Y_{t,f}^{+,\Pi})' \right)^{-1}$$

where $Y_{t,f}^{+,\Pi}$ denotes the residuals from the regression of $Y_{t,f}^+$ on $U_{t,f}^+$.

Solution to the reduced rank regression problem is found via SVD: unrestricted estimate:

$$\hat{\beta}_z = \left(\sum_{t=1}^{T} Y_{t,f}^{+,\Pi}(Z_{t,p}^{-,\Pi})'\right)\left(\sum_{t=1}^{T} Z_{t,p}^{-,\Pi}(Z_{t,p}^{-,\Pi})'\right)^{-1}$$

SVD: $$W^{1/2}\hat{\beta}_z\hat{W}_p^- = \hat{U}\hat{\Sigma}\hat{V}' = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}_n$$

where

- $Z_{t,p}^{-,\Pi}$ denotes the residuals from the regression of $Z_{t,p}^-$ on $U_{t,f}^+$.
- $\hat{W}_p^- = (T^{-1}\sum_{t=1}^{T} Z_{t,p}^{-,\Pi}(Z_{t,p}^{-,\Pi})')^{1/2}$.
- $\hat{U}_n \in \mathbb{R}^{fs\times n}, \hat{V}_n \in \mathbb{R}^{(s+m)p\times n}$
- $\hat{\Sigma}_n = \mathrm{diag}(\hat{\sigma}_1,\ldots,\hat{\sigma}_n) \in \mathbb{R}^{n\times n}$: $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \hat{\sigma}_n > 0$ are the dominating n singular values.

Reduced rank estimator

$$\boxed{\tilde{\beta}_z = \hat{\mathscr{O}}_f\hat{\mathscr{K}}_p = W^{-1/2}\hat{U}_n\hat{\Sigma}_n\hat{V}_n'(\hat{W}_p^-)^{-1}}$$

This is exactly what is done in CCA!

Step 3: Estimation of the System Matrices: The main idea of this procedure is due to W. Larimore (1983).

From step 2 of the subspace algorithm we use $\hat{\mathscr{K}}_p = \hat{\Sigma}_n^{1/2} \hat{V}_n^T (\hat{W_p^-})^{-1}$ to estimate the state as $\hat{x}_t = \hat{\mathscr{K}}_p Z_{t,p}^-$. Knowing the state, the system matrices can be estimated using linear regressions in the system equations.

Notation: $\langle a_t, b_t \rangle = \frac{1}{T} \sum_{t=1}^{T} a_t b_t^T$. Then:

$$[\hat{C}, \hat{D}] = \langle y_t, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix} \rangle \langle \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix} \rangle^{-1}$$

$$[\hat{A}, \hat{B}] = \langle \hat{x}_{t+1}, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix} \rangle \langle \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix}, \begin{bmatrix} \hat{x}_t \\ u_t \end{bmatrix} \rangle^{-1}$$

Estimation of K:

Let $\hat{\varepsilon}_t = y_t - \hat{C}\hat{x}_t - \hat{D}u_t$. Then estimate K using least squares in $\hat{x}_{t+1} = K\hat{\varepsilon}_t + r_t$ (assuming a direct feedthrough term D is also estimated).

If no direct feedthrough term is estimated (i.e. $D = 0$): $\hat{\varepsilon}_t = y_t - \hat{C}\hat{x}_t$ is included in the second equation above.

Assumptions on the noise: ε_t is a strictly stationary martingale difference sequence adapted to the sequence of sigma algebras $\mathscr{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \cdots\}$ (No Gaussianity required!) fulfilling:

$$\begin{aligned} \mathbb{E}\{\varepsilon_t | \mathscr{F}_{t-1}\} &= 0 \\ \mathbb{E}\{\varepsilon_t \varepsilon_t' | \mathscr{F}_{t-1}\} &= \Omega = \mathbb{E}\varepsilon_t \varepsilon_t' \\ \mathbb{E}\{\varepsilon_{t,a} \varepsilon_{t,b} \varepsilon_{t,c} | \mathscr{F}_{t-1}\} &= \omega_{a,b,c} \\ \mathbb{E}\{\varepsilon_{t,a}^4\} &< \infty \end{aligned}$$

The input is generated by a stable and strictly minimum-phase state space system where the noise fulfills the above assumptions. The true order n is known.

Assumptions on Inputs: ARMA, generated by white noise (mean zero martingale difference sequence fulfilling the properties stated before), strictly minimum-phase and stable

Assumptions on weighting matrices:

- f constant not depending on T: $\hat{W}_f^+ \to W_f > 0$ a.s.
- $f \to \infty$: either $\hat{W}_f^+ = I$ or CCA choice.

$\hat{W}_p^-$: CCA choice. No influence on asymptotic variance.

Under assumption of known order:

Known asymptotic properties

- Strong consistency of transfer function, if

$$f \geq n, p = p(T) \to \infty, p(T) = o(\log T^a)$$

for some $0 < a < \infty$. I.e. for fixed $z = e^{i\omega}$

$$\hat{k}(z) = \pi(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{K}) \to k_0(z) \quad \text{a.s.}$$

- Strong consistency for system matrices on a generic subset of M_n:
$\hat{A} \to A_0, \hat{B} \to B_0, \hat{C} \to C_0, \hat{D} \to D_0, \hat{K} \to K_0$

- Asymptotic normality for a generic subset of M_n, if additionally $\liminf_T -2p(T)\log\rho_0/\log T > 1$:

$$\sqrt{T}\,\mathrm{vec}(\hat{A}-A_0, \hat{B}-B_0, \hat{C}-C_0, \hat{D}-D_0, \hat{K}-K_0) \xrightarrow{d} \mathcal{N}(0, V)$$

- $\hat{p}$ that fulfills the above assumption can be estimated using `AIC` in an ARX approximation of y_t.

Special case: No inputs, $m = 0$.

- ▷ Explicit expression for asymptotic variance V for fixed f:

$$M_1 M_1' + M_2 \left[\Gamma_z \otimes \left\{ \mathscr{O}^\dagger [\mathscr{E}_f (I_f \otimes \Omega) \mathscr{E}_f'] (\mathscr{O}^\dagger)' \right\} \right] M_2'$$

  where $W = (W_f^+)' W_f^+$.
  $\mathscr{O}^\dagger = (\mathscr{O}_f' W \mathscr{O}_f)^{-1} \mathscr{O}_f' W$, M_1 and M_2 do not depend on f or W_f^+. $\Gamma_z = \mathbb{E} Z_{t,\infty}^- (Z_{t,\infty}^-)'$.

- ▷ $\hat{W}_p^-$ does not appear in the expression, W_f^+ enters explicitely.

- ▷ For each fixed f the optimal choice is according to CCA, i.e. $\hat{W}_f^+ = (T^{-1} \sum_{t=1}^T Y_{t,f}^+ (Y_{t,f}^+)')^{-1/2}$ with minimal variance

$$M_1 M_1' + M_2 \left[\Gamma_z \otimes (\mathscr{O}_f' W \mathscr{O}_f)^{-1} \right] M_2'$$

- ▷ Optimal variance is monotonically decreasing in f!

- ▷ Procedure using $\min(f, p) \geq \frac{-d \log T}{2 \log \rho_0}, d > 1$, $\max(f, p) = o((\log T)^a)$ is asymptotically equivalent to prediction error methods (for correctly specified system order n).

General case:

- White inputs u_t: same results as in the no inputs case.
- Coloured inputs u_t:
  - No simple expressions for V exist.
  - Influence of f not known.
  - Effects of W_f^+ not known.
  - Not equivalent to pseudo ML.

Approximation results for misspecified order (smaller than the true order):

Explicit expressions can be found. Interpretation is not known.