UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

State Space Modeling

MTSA Klagenfurt, September 7th, 2026

Dietmar BAUER

Motivation

What is a state?

- A state refers to the current status of a system.
- It evolves over time, that is, it is dynamic.
- A state summarizes relevant information, condenses what occurred in a system in the past. Thus a state must be something that depends on the past of a system.
- When conveying a state we want to include only information from the past that is relevant for the future evolution.
- It must also be complete in the sense that if we know the state, nothing else from the past is relevant for the future.
- One could say, conditional on the state 'the past and the future are decoupled'.
- For modelling purposes, we need to be able to observe the state or its effects.

One notices that states have a close connection to the **Markov property** (loosely: given the presence, the future is conditionally independent of the past).

Example: Game of chess

- Chess games are sometimes interrupted for breaks.
- If you resume the match, you need to start at the state you ended before.



Source: generated by Gemimi.

Example: Tennis Match

- Tennis matches sometimes are interrupted due to weather conditions.
- If you resume the match, you need to start at the state you ended before.



Source: generated by Gemimi.

Example: Tennis Match (II)

- Here the state is composed of the score plus the information who is serving.
- This depends on how the game has proceeded up to now: If you followed the game closely (observation) you can reconstruct the state.
- The state condenses the information: It is much faster to tell someone the score than to tell the story of the game.
- The state is dynamic: Whenever one player makes a mistake the state is adjusted.
- If you know the state, you can restart and the break is immaterial for the processing of the game.

BUT: This description of the game only is an approximation. The score does not summarize all information. Models are just approximations of a more complex reality.

Basketball game



Is this providing the state of the game?

Formal Definition

In this talk I will deal with the following systems:

Definition (State Space System)

A state space system is defined by the state equation:

$$x_{t+1} = f(x_t, w_t) \tag{1}$$

where $x_t \in \mathbb{R}^n$ denotes the unobserved state, $(w_t)_{t\in\mathbb{Z}}$, $w_t \in \mathbb{R}^m$ is an independent identically distributed (iid) sequence and the measurement equation

$$y_t = h(x_t, v_t) \tag{2}$$

where the process $(y_t)_{t\in\mathbb{Z}}$, $y_t \in \mathbb{R}^s$ is observed and $(v_t)_{t\in\mathbb{Z}}$, $(w_t)_{t\in\mathbb{Z}}$ denote unobserved joint iid sequences.
The functions f and h are smooth (C^∞).
The random variable $x_0 \in \mathbb{R}^n$ is often called *initial state*, if we consider processes started at time $t = 0$.

Examples

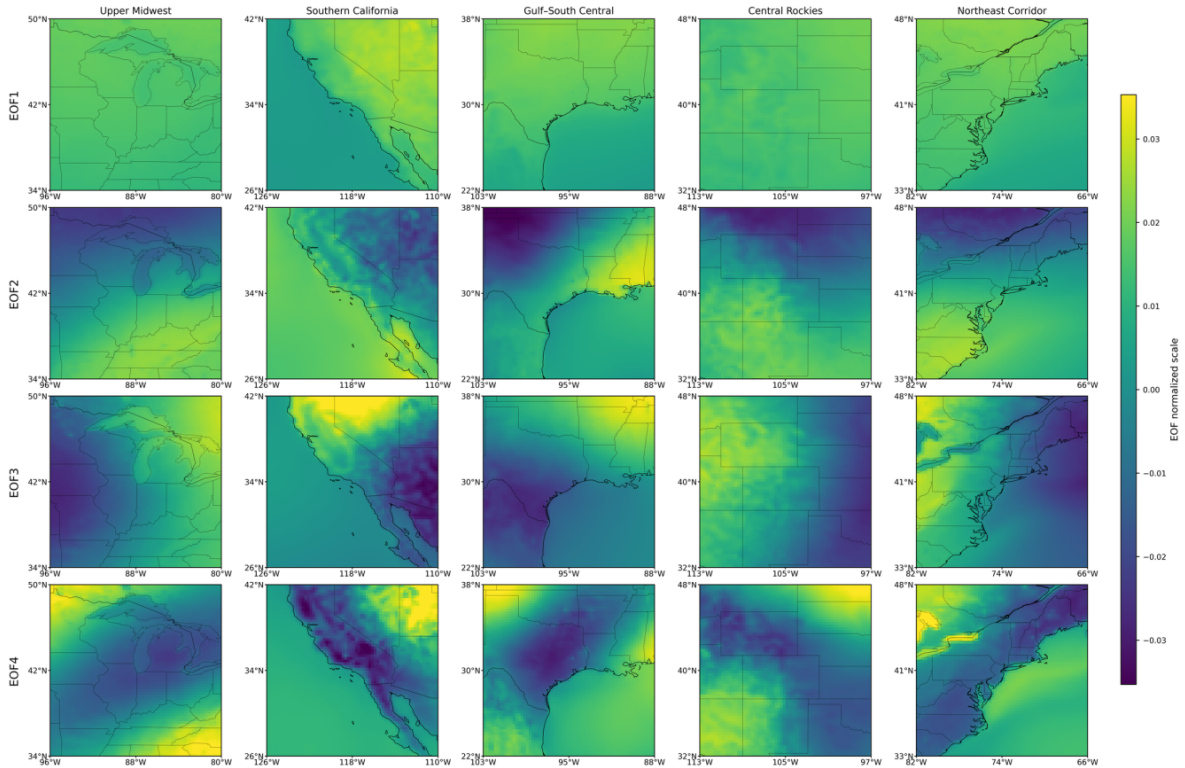
Modelling of spatial data:

- Daily measurements $Y_t \in \mathbb{R}^{64\times 64}$ of temperature (mean hourly 2m temperature measurements) over a long time span (1 January 1940 to 31 December 2025).
- 2-D grids of size 64x64 in five American regions.
- Vectorising one would obtain $y_t \in \mathbb{R}^{4096}$, a 4096 dimensional time series over more than 30 000 days.
- There will be strong spatial and temporal correlations and seasonal effects.
- Models for prediction of mean temperatures may use many different approaches. The paper by Kotharkar and Stein uses a linear state space approach:

$$
\begin{aligned}
y_t &= \Lambda F_t + u_t, \\
F_t &= m(d_t) + s(d_t) y_t^*, \quad d_t : \quad \text{day of year} \\
y_t^* &= C x_t + \varepsilon_t, x_{t+1} = A x_t + B \varepsilon_t.
\end{aligned}
$$

- Spatial dependence introduced by the factor structure; the grid relates to a matrix (2-D tensor) $\Rightarrow$ (*Roberto Casarin, Thursday, PM*).

1. 1. Basics



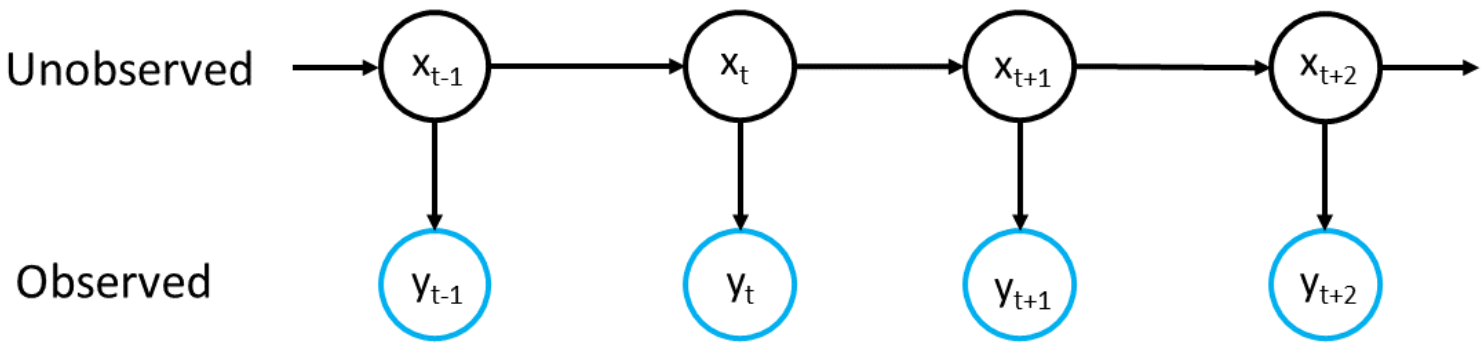
Only four factors explain about 97% of the variation of temperatures.

More on environmental modelling this afternoon: *Eric Hillebrand*

Source: Varun Kotharkar, Michael L. Stein (2026) An Interpretable Low-Rank State-Space Model for Multi-Horizon Simulation of Large-Scale Regional Temperature Fields. Kotharkar and Stein (2026)

Examples (II)

Hidden Markov Models *Roland Langrock, Fri, AM*



- The state is a Markov chain with a finite state space (as in the tennis example).
- The observations y_t (can be uni- or multivariate) are conditionally (on x_t) independent.

Examples (III)

VAR models: defined as the solutions to

$$y_t = A_1 y_{t-1} + ... + A_p y_{t-p} + \varepsilon_t$$

- If we define the state as $x_t = [y_{t-1}, ..., y_{t-p}]'$ then we obtain:

$$x_{t+1} = \begin{pmatrix} y_t \\ \vdots \\ y_{t-p+1} \end{pmatrix} = \begin{pmatrix} A_1 & A_2 & \dots & \dots & A_p \\ I_s & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & I_s & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$= \Pi x_t + B\varepsilon_t = \Pi x_t + B\varepsilon_t.$$

- Additionally $y_t = \begin{pmatrix} A_1 & A_2 & \dots & \dots & A_p \end{pmatrix} x_t + \varepsilon_t$.
- Often $\mathrm{Var}(\varepsilon_t) = \Sigma \in \mathbb{R}^{s \times s}$ is general, while economic theory would rather use uncorrelated 'shocks'. Also there may be relations between the matrices $A_j \in \mathbb{R}^{s \times s}$ that are economically meaningful $\Rightarrow$ *Martin Bruns, Tue, PM*.

Examples (IV)

Non-linear VAR models: defined as the solutions to

$$y_t = h_1(y_{t-1}, ..., y_{t-p}; \theta_h) + \varepsilon_t$$

- Defining the state as $x_t = [y_{t-1}, ..., y_{t-p}]'$ then we obtain:

$$\begin{aligned} y_t &= h_1(x_t; \theta_h) + \varepsilon_t, \\ x_{t+1} &= f(x_t; \theta_h) + B\varepsilon_t. \end{aligned}$$

- Here the function f has more structure: The first component contains the non-linearity. The rest has the same shift structure as for VAR(p).
- More on these: *James Duffy, Wed, AM*.

Examples (V)

VARMA(1,1) processes: defined as the solutions to $y_t = A_1 y_{t-1} + B_1 \varepsilon_{t-1} + \varepsilon_t$.

- If we define the state as $x_t = [y_{t-1}, \varepsilon_{t-1}]'$ then we obtain:

$$\begin{aligned} x_{t+1} = \begin{pmatrix} y_t \\ \varepsilon_t \end{pmatrix} &= \begin{pmatrix} A_1 y_{t-1} + B_1 \varepsilon_{t-1} \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_t \\ \varepsilon_t \end{pmatrix} \\ &= \begin{pmatrix} A_1 & B_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ \varepsilon_{t-1} \end{pmatrix} + \begin{pmatrix} I_s \\ I_s \end{pmatrix} \varepsilon_t = A x_t + B \varepsilon_t. \end{aligned}$$

- Additionally $y_t = \begin{bmatrix} A_1 & B_1 \end{bmatrix} x_t + \varepsilon_t$.
- The same can be done also for higher order VAR(MA) processes.

Hannan and Deistler (2012) Chapter 1 discusses that state space and VARMA are equivalent representation (in a specific sense).

Examples (VI)

Unobserved components models (cf. e.g., Harvey and Proietti (2005)):

- Trending processes show the feature $x_{t+1} = x_t + \mu + w_t$.
- Solutions are of the form $x_{t+1} = x_0 + \mu(t+1) + \sum_{j=0}^{t} w_j$.
- For $\mathrm{Var}(w_t) = 0$ we obtain a linear deterministic trend.
- For $\mathrm{Var}(w_t) > 0$ a random walk with drift (for $\mu \neq 0$).
- Similarly for quarterly data the definition:

$$x_{t+1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} x_t + w_t$$

  defines seasonal terms that either correspond to a fixed seasonal pattern (for zero variance of w_t) or stochastic seasonality.
- Models with different components for data containing trend, level and seasonal components.

1. 2. Road map

In order to work with state space models, we need to

- understand the (stochastic) properties of the generated processes
- understand the structure (related to identifiability)
- develop methods for estimation
- develop understanding on how to work with (inference, prediction) the estimated models

Outline

1 Motivation

2 Properties of general state space processes and models

3 Special Case: LTI State Space Models

4 Properties of linear processes

5 Estimation of Parameters

6 The Kalman Filter

7 Bayesian Estimation

Properties of general state space processes and models

State Transition Equation

General model (with $(v_t)_{t\in\mathbb{Z}}, (w_t)_{t\in\mathbb{Z}}$ being iid white noise):

$$y_t = h(x_t, v_t),$$
$$x_{t+1} = f(x_t, w_t).$$

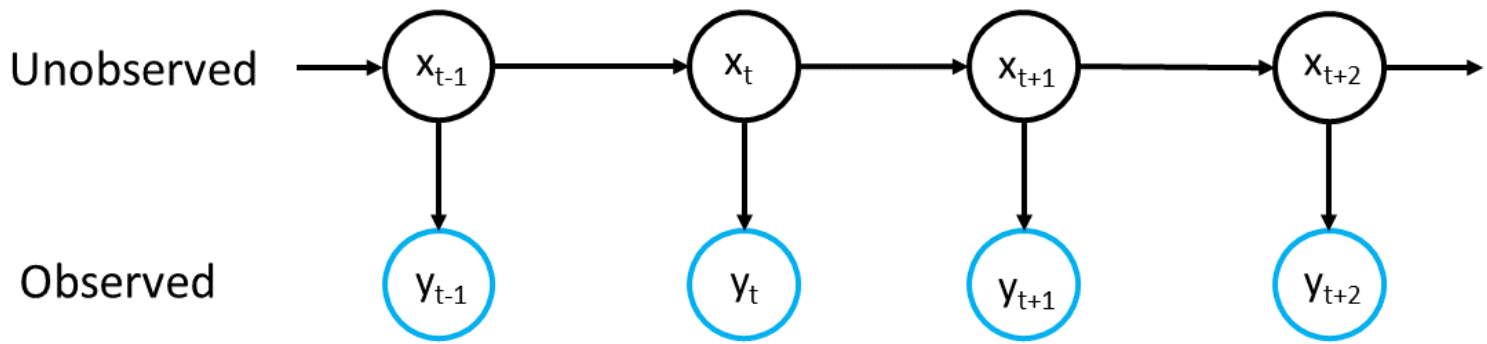
- To understand the properties of such a process, the state (transition) equation is key.
- It has a relatively simple structure, being a Markov chain (for iid process w_t).
- Thus Markov chain theory helps with the understanding.
- Once the state process is understood, the observations y_t are obtained from a static map, which is a simple transformation preserving many properties.

Stationarity and Ergodicity

2. 1. Stationarity and Ergodicity

Evolution of the state

The state process $(x_t)_{t\in\mathbb{Z}}$ is a Markov-chain for white noise $(w_t)_{t\in\mathbb{Z}}$.



The observations y_t are conditionally (on $x_t \in \mathbb{R}^n$) independent.

A special case of this structure are *Hidden Markov Models (HMM) Roland Langrock, Fri, AM*, where the state takes on only a finite number of different values: $x_t \in \{1, ..., K\}$.

Evolution of the state (II)

- The state equation $x_t = f(x_{t-1}, w_{t-1}), t \in \mathbb{Z}$, can be iterated:

$$\begin{aligned} x_t &= f(f(x_{t-2}, w_{t-2}), w_{t-1}) = f(f(f(x_{t-3}, w_{t-3}), w_{t-2}), w_{t-1}) \\ &= \dots \\ &= F^{(t)}(x_0, w_0, w_1, \dots, w_{t-1}) \end{aligned}$$

  where the function $F^{(t)}$ is built recursively $\Rightarrow$ complicated object in general.

- **Example:** For linear state space system $x_{t+1} = Ax_t + Bw_t$ we have:

$$x_t = Ax_{t-1} + Bw_{t-1} = A^t x_0 + A^{t-1} Bw_0 + \dots + Bw_{t-1}.$$

- Often the properties are easier to describe for systems running forever:

$$x_t = \sum_{j=0}^{\infty} A^j Bw_{t-j-1}.$$

  (if this exists).

Key concepts

- wide sense/weak stationarity: first and second moments do not depend on t
  - The expectation is constant over time.
  - The variance is constant over time.
  - The covariances $\mathrm{Cov}(x_t, x_{t-j}) = \gamma_j$ do not depend on t.
- strict stationarity: distribution of finite sections $[x_t', x_{t+1}', \ldots, x_{t+k}']'$ does not depend on t for all $k > 0$.
- ergodicity: implies, for example, that sample moments for just one realisation of the process converge to expectations:

$$\frac{1}{T}\sum_{t=1}^{T} x_t \to \mathbb{E}\, x_0, \quad \frac{1}{T}\sum_{t=j+1}^{T} x_t x_{t-j}' \to \mathbb{E}(x_0 x_{-j}')$$

- Allows inference, for example, in regression settings:

$$y_t = \beta_0 + \beta' x_t + u_t \Rightarrow \hat{\beta} = \widehat{\mathrm{Var}}(x_t)^{-1}\, \widehat{\mathrm{Cov}}(x_t, y_t).$$

2. 1. Stationarity and Ergodicity

- Ergodicity in rough terms means that we can infer all population parameters from just one trajectory.
- In time series this is important, because in most cases we only measure one trajectory.

Example of a non-ergodic behaviour:

- We observe a roulette in a casino and want to infer the probabilities of 'red'.
  - Consecutive turns of the roulette can be modelled as an iid series, that has the ergodicity property: From observing the relative frequency of 'red', we get better knowledge the more often the wheel is spun.



Souce: Google Gemini.

- BUT: If at the start of the day the casino chooses one of a number of different non-identical roulette wheels, our experience from yesterday does not help us for today.

Whether ergodicity holds or not, often is a philosophical question.

Theory on Markov chains

- Meyn and Tweedie (2012) contains a detailed discussion of many properties for Markov chains, also applying to the nonlinear state space system.
- The underlying requirements for stationarity and ergodicity of the Markov chain are stability (in an appropriate sense), irreducibility and aperiodicity.
- All are related to the concept of recurrence.
- If the assumptions hold, then the state process converges to a stationary distribution and is ergodic (empirical moments tend to the true moments) **for every starting value** (which are forgotten along the way).

REMARK: The property, that Markov chains converge to a stationary distribution from every starting point is essential for Bayesian estimation using Markov Chain Monte Carlo sampling.
Gregor Kastner, Thur, AM.

2. 1. Stationarity and Ergodicity

Theorem (adapted from Thm 15.0.1 of Meyn and Tweedie (2012))

Suppose that the chain Φ defined by the Markov transition kernel $P(x, .)$ is ψ-irreducible and aperiodic. Then the following are equivalent:

(i) The chain is positive recurrent with invariant probability measure π [...]

(ii) There exists some petite set $\mathcal{C}$ and $\kappa > 1$ such that

$$\sup_{x \in \mathcal{C}} \mathbb{E}_x[\kappa^{\tau_{\mathcal{C}}}] < \infty.$$

(iii) There exists a petite set $\mathcal{C}$, constants $b < \infty, \beta > 0$ and a function $V \geq 1$ finite at some one $x_0 \in X$ satisfying (drift criterion)

$$\mathbb{E}(V(x_{t+1})|x_t = x) - V(x) \leq -\beta V(x) + b\mathbb{I}_{\mathcal{C}}(x), x \in X.$$

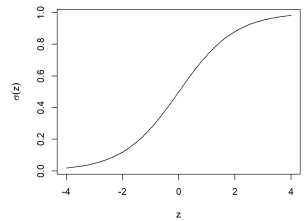
Any of these three conditions imply that the set $S_V = \{x : V(x) < \infty\}$ is absorbing and full, where V is any solution of the inequality above. Then there exists constants $r > 1, R < \infty$ such that for any $x \in S_V$

$$\sum_{n=1}^{\infty} r^n \|P^n(x, .) - \pi\|_V \leq RV(x).$$

Application to a recurrent neural net

Consider for w_t iid with zero mean and variance σ_w^2:

$$x_{t+1} = b_0 + a\sigma(b + cx_t) + w_t \in \mathbb{R}, \sigma(z) = \frac{e^z}{1+e^z} \in [0,1].$$



- Choose the function $V(x) = 1 + x^2$. Then:

$$\begin{aligned}\mathbb{E}(V(x_{t+1})|x_t) &= 1 + \mathbb{E}(x_{t+1}^2|x_t) = 1 + \sigma_w^2 + (b_0 + a\sigma(b + wx_t))^2 \\ &\leq 1 + \sigma_w^2 + (|b_0| + |a|)^2 \leq \mu.\end{aligned}$$

$$\Rightarrow \mathbb{E}(V(x_{t+1})|x_t = x) - V(x) \leq \mu - V(x) \leq \mu\mathbb{I}(x < \sqrt{2\mu}) - \frac{1}{2}V(x).$$

- This can be used to show that RNNs produce ergodic, stationary processes. The main property is the boundedness, hence this also holds for all sorts of NNs.

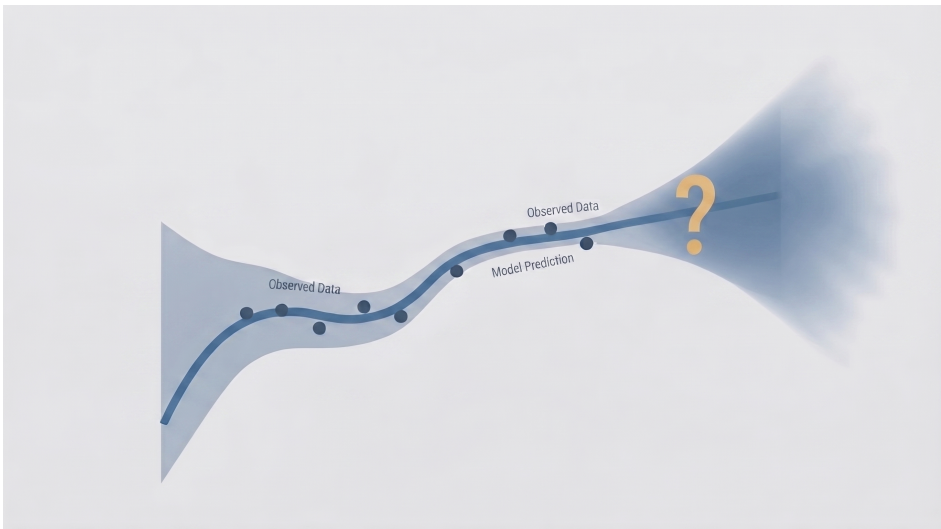
Identification

2. 2. Identification

- Identification (in other words: uniqueness of representation) is an issue in all state space models as the states typically are not directly observed.
- There are two separate issues:
  1. transformations of the state lead to equivalent systems: $\tilde{x}_t = T(x_t)$:

$$y_t = h(x_t, v_t), \quad x_{t+1} = f(x_t, w_t),$$
$$y_t = \tilde{h}(\tilde{x}_t, v_t), \quad \tilde{x}_{t+1} = \tilde{f}(\tilde{x}_t, w_t),$$

  2. State and observation equations are (nonlinear) mappings.



Source: Google Gemini.

2. 2. Identification

In order to work with the processes, we build a model describing the observed process that contains parameters:

$$y_t = h(x_t, v_t; \theta), \quad x_{t+1} = f(x_t, w_t; \theta)$$

There are three terms linked to the issue of identification:

- Controllability: Is the system capable of reaching every region in the state space?
- Observability: Do we have enough measurements to observe the whole system?
- Minimality: Are there redundant states, that could be omitted?



This is very dependent on the model, there are not many general results (at least that I know of).

From now on we will deal with the special case of linear time invariant (LTI) models.

Special Case: LTI State Space Models

Definition

$$y_t = Cx_t + v_t,$$
$$x_{t+1} = Ax_t + Bw_t$$

where $(v_t)_{t\in\mathbb{R}}$ and $(w_t)_{t\in\mathbb{R}}$ are white noise (also jointly).

What about the different noises $(v_t)_{t\in\mathbb{Z}}$ and $(w_t)_{t\in\mathbb{Z}}$?

For LTI state space systems we have (see Hannan and Deistler (2012), p. 17):

Theorem (innovation representation)

Every system $y_t = Cx_t + v_t, \quad x_{t+1} = Ax_t + B_t$ with stable matrix A ($\lambda_{|max|}(A) < 1$) can be represented* such that

$$y_t = C\tilde{x}_t + \varepsilon_t,$$
$$\tilde{x}_{t+1} = A\tilde{x}_t + K\varepsilon_t$$

Here $\lambda_{|max|}(A - KC) \leq 1$, the system is called *minimum-phase*.

* meaning: the stationary solutions are identical.

3. Special Case: LTI State Space Models

Drift criterion: $V(x) = 1 + x'x = 1 + \|x\|_2^2$ leads to (for $\mathrm{Var}(\varepsilon_t) = \Sigma > 0$)

$$\begin{aligned}\mathbb{E}(V(x_{t+1})|x_t = x) &= 1 + \mathrm{tr}\left[x_t' A' A x_t + K\Sigma K'\right] \\ &= 1 + \|Ax_t\|_2^2 + \mathrm{tr}[K\Sigma K'] \leq \lambda_{|max|}(A)^2 \|x_t\|_2^2 + \mu \\ &\leq \lambda_{|max|}(A)^2 V(x_t) + \mu.\end{aligned}$$

For $0 < \lambda_{|max|}(A) < 1$ shows ergodicity and existence of stationary solution.

Iterating the state equation:

$$\begin{aligned}x_t &= Ax_{t-1} + K\varepsilon_{t-1} = A(Ax_{t-2} + K\varepsilon_{t-2}) + K\varepsilon_{t-1} \\ &= A^2 x_{t-2} + AK\varepsilon_{t-2} + K\varepsilon_{t-1} = \ldots \\ &= A^t x_0 + \sum_{j=1}^{t} A^{j-1} K\varepsilon_{t-j}\end{aligned}$$

3. Special Case: LTI State Space Models

- Stability: $\lambda_{|max|}(A) < 1$ ensures that $A^t \to 0$ for $t \to \infty$.
- In this case the distribution of the state converges to a stationary distribution.
- If the system is started in the distant past we obtain

$$x_t = \sum_{j=1}^{\infty} A^{j-1} K \varepsilon_{t-j}$$

- Consequently

$$y_t = \varepsilon_t + \sum_{j=1}^{\infty} C A^{j-1} K \varepsilon_{t-j}.$$

The coefficients $CA^{j-1}K$ are called impulse response sequence: For $\varepsilon_t = \delta_{t,0}$ (impulse: only non-zero for $t = 0$) we get

$$y_t = \delta_{t,0} + \sum_{j=1}^{\infty} C A^{j-1} K \delta_{t-j,0} = C A^{t-1} K, t > 0.$$

3. Special Case: LTI State Space Models

Impulse responses to different systems

- For illustrations I will use the `R` package `modstatespace`, that is currently being developed in my group in Bielefeld. It included i.a. a collection of functions that have been written by Wolfgang Scherrer (TU Wien). It is not finished yet, so be careful with using it.
- It can be downloaded (or directly installed) from the github page `https://github.com/dbauer72/modstatespace`.
- The examples in this lecture run through, I cannot guarantee that the package does not still have flaws.

```
# Step 1: devtools
library(devtools)

# Step 2: install package
install_github("dbauer72/modstatespace")
library(modstatespace)
```

3. Special Case: LTI State Space Models

We define three different systems and investigate the impulse response:

- System 1: AR(1) with parameter $\rho = 0.9$.
- System 2: MA(1) with parameter $\beta = 0.5$
- System 3: ARMA(1,1) with parameter $\alpha_1 = -0.8, \beta = 0.5$.

```
I <- matrix(1,1,1)
AR1 <- list(a = matrix(c(1,-0.9),1,2),b = I, Sigma = diag(1))
syst_AR1 <- arma2StSp(AR1)

# MA(1) system with theta = 0.5.
MA1 <- list(a = I,b = matrix(c(1,0.5),1,2), Sigma = diag(1))
syst_MA1 <- arma2StSp(MA1)

# ARMA(1,1) system with alpha = -0.8, beta = 0.5.
ARMA11 <- list(a = matrix(c(1,0.8),1,2),b = matrix(c(1,0.5),1,2),
    Sigma = diag(1))
syst_ARMA11 <- arma2StSp(ARMA11)

# put all three in a vector
systs <- list(3)
systs[[1]] <- syst_AR1
systs[[2]] <- syst_MA1
systs[[3]] <- syst_ARMA11
class(systs) <- "StSp"
IRs <- impulse(systs,plot=1)
```

3. Special Case: LTI State Space Models

To contrast this, we also use an unobserved components model as System 4:

```
theta_uc <- c(0.1,0,0.5,0.3,0.2,0.4,1,0.7,1)
(syst_uc <- param_syst_uc(theta_uc))
```

Trajectories generated with these different systems looks differently:

```
set.seed(1)
y <- sim_syst(syst_ARMA11,T= 100,stationary=FALSE)
autoplot(y)

set.seed(1)
y <- sim_syst(syst_uc,T= 100,stationary=FALSE)
autoplot(y)
```

We see that LTI systems can produce stationary as well as trending (seasonally trending) non-stationary processes.

Identifying State Space Models

Controllability

$$x_{t+1} = Ax_t + K\varepsilon_t = A^2 x_{t-1} + AK\varepsilon_{t-1} + K\varepsilon_t$$

Therefore:

$$x_t = \underbrace{[K, AK, A^2K, ..., A^{t-1}K]}_{\mathcal{C}_t} \underbrace{\begin{bmatrix} \varepsilon_{t-1} \\ \varepsilon_{t-2} \\ \vdots \\ \varepsilon_0 \end{bmatrix}}_{E_0^{t-1}} + A^t x_0$$

Thus we can achieve every state $x_t \in \mathbb{R}^n$ starting from every other, iff $\mathcal{C}_t$ is of full row rank for one t (and then for all larger ones).

Observability

Consider a string of observed values:

$$\begin{bmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+f-1} \end{bmatrix} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{f-1} \end{bmatrix}}_{\mathcal{O}_f} x_t + \underbrace{\begin{bmatrix} I_s & 0 & \dots & 0 \\ CK & I_s & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ CA^{f-2}K & \dots & CK & I_s \end{bmatrix}}_{\mathcal{E}_f} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t+1} \\ \vdots \\ \varepsilon_{t+f-1} \end{bmatrix}$$

$$Y_{t+f-1}^{t} = \mathcal{O}_f \mathcal{C}_\infty E_{-\infty}^{t-1} + \mathcal{E}_f E_{t+f-1}^{t}$$

- Here $\mathcal{O}_f$ is called *observability matrix*.
- The system is called *observable*, if we can reconstruct the state from all signals (y_t and ε_t) plus the system.
- This works, if and only if the observability matrix (for $f > n$) is of full column rank.

Minimality

Definition (Minimality)

A linear, time invariant state space system is *minimal*, if it is observable and controllable.

Minimality implies that we cannot omit a state component without reducing the dynamics of the system.

- For a minimal system the matrix $\mathcal{H}_{f,\infty} = \mathcal{O}_f \mathcal{C}_\infty$ has rank n for all $f > n$.
- The (i, j)-th block of $\mathcal{H}_{f,\infty}$ equals $CA^{i+j-1}K$.
- Thus $\mathcal{H}_{f,\infty}$ is a function of the impulse response sequence.

3. 1. Identifying State Space Models

Examples of non-minimal systems:

- Not controllable:

$$x_{t+1} = \begin{bmatrix} A_1 & A_{1,2} \\ 0 & A_2 \end{bmatrix} x_t + \begin{bmatrix} K_1 \\ 0 \end{bmatrix} \varepsilon_t$$

  It follows that the second component of the state $x_{t+1,2} = A_2 x_{t,2} = A_2^t x_{0,2}$ is determined from the initial state and not influenced by the noise.

- Not observable:

$$y_t = \begin{bmatrix} C_1 & 0 \end{bmatrix} \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} + \varepsilon_t, \quad \begin{pmatrix} x_{t+1,1} \\ x_{t+1,2} \end{pmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{pmatrix} x_{t,1} \\ x_{t,2} \end{pmatrix} + K\varepsilon_t$$

Here the second component of x_t does neither influence $x_{t+1,1}$ nor y_t. We can do without it and use only

$$y_t = C_1 x_{t,1} + \varepsilon_t, \quad x_{t+1,1} = A_1 x_{t,1} + K_1 \varepsilon_t$$

Reduction to minimality

Theorem (Hannan and Deistler (2012), 2.3.1.)

For every linear time invariant state space system (A, C, K), $A \in \mathbb{R}^{n \times n}$ there exists a state space system $(\tilde{A}, \tilde{C}, \tilde{K})$, $\tilde{A} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, $\tilde{n} \leq n$, that is observable and controllable, such that both systems correspond to the same impulse response CA^jK.
The corresponding integer $\tilde{n}$, the minimal state dimension, is called *order of the system*.

Therefore – when dealing with stationary processes – it is no restriction of generality to confine analysis only to minimal systems in innovation representation for black-box modeling.

Choice of the state space basis

3. 2. Choice of the state space basis

$$Y_{t+f-1}^{t} = \underbrace{\mathcal{O}_f \mathcal{C}_\infty}_{\mathcal{H}_{f,\infty}} E_{-\infty}^{t-1} + \mathcal{E}_f E_{t+f-1}^{t}$$

- The state is not observed, we only observe covariance matrices, that can be related to impulse responses CA^jK, the entries of $\mathcal{H}_{f,\infty} = \mathcal{O}_f \mathcal{C}_\infty$.
- If we choose a different basis for the state space and hence transform x_t to $\tilde{x}_t = \mathcal{T} x_t$ then we obtain:

$$\tilde{x}_{t+1} = \mathcal{T} x_{t+1} = \mathcal{T} A \mathcal{T}^{-1} \tilde{x}_t + \mathcal{T} K \varepsilon_t, \quad y_t = C \mathcal{T}^{-1} \mathcal{T} x_t + \varepsilon_t$$

- Thus (A, C, K) and $(\tilde{A}, \tilde{C}, \tilde{K}) = (\mathcal{T} A \mathcal{T}^{-1}, C \mathcal{T}^{-1}, \mathcal{T} K)$ correspond to the same process and the same impulse response since

$$CA^jK = C\mathcal{T}^{-1}(\mathcal{T} A \mathcal{T}^{-1})^j \mathcal{T} K = \tilde{C} \tilde{A}^j \tilde{K}$$

- Thus we have the freedom to choose n^2 entries to achieve identifiability.
- There is no unique way to do this.

3. 2. Choice of the state space basis

Different restrictions

Two possibilities (among many more):

1. Fixing entries in the observability matrix:
   $S_n\mathcal{O}_n = I_n$ (S_n... selector matrix; related to echelon form)
2. Choosing a special form of the transition equation:
   $A = \text{diag}(\rho_1, ..., \rho_n)$ and the first column of K equals ones (related to Jordan normal form): unobserved component models.

On total we obtain $ns + sn = 2ns$ parameters ($2ns + n^2$ entries in (A, C, K) minus n^2 restrictions) collected in the parameter vector $\theta \in \mathbb{R}^{2sn}$.

REMARKS

- Combinations are possible.
- Both approaches do not work for all minimal systems of order n. But for a generic subset (that is in almost all cases for the Lebegue measure). This may restrict full generality in special situations, but usually does not.

3. 2. Choice of the state space basis

Parameterization

The transfer function and the parameters are related in the following way:

- For $(A(\theta), C(\theta), K(\theta)) : k_j(\theta) = C(\theta)A(\theta)^{j-1}K(\theta)$.
- For given k_j we may calculate $\mathcal{H}_{f,\infty}$
- Factoring this as $\mathcal{H}_{f,\infty} = \mathcal{O}_f\mathcal{C}_\infty$ where $S_n\mathcal{O}_f = I_n$ we obtain unique observability and controllability matrices.
- C is the first block row of $\mathcal{O}_f$, K the first block column of $\mathcal{C}_\infty$.
- A is obtained from $\mathcal{O}_f^{\Uparrow} = \mathcal{O}_{f-1}A$.
- The restriction $S_n\mathcal{O}_f = I_n$ implies that we may use all unrestricted entries of (A, C, K) thus obtained as parameters. There are exactly $2ns$ of these entries, whose place in the matrices is known.

Using these facts we get two different maps (inverse of each other) that we need for numerical estimation:

1. $\phi(\theta) = k(z;\theta) = I_s + zC(\theta)(I_n - zA\theta))^{-1}K(\theta)$ maps parameters to transfer function (and hence impulse response sequence).
2. $\phi^{-1}(k) = \theta$ provides parameters for given transfer function.

3. 2. Choice of the state space basis

Parameterisation

```
# Getting parameters for systems:
(theta <- syst_param(syst_AR1))

(theta <- syst_param(syst_ARMA11))

# Calculating the system based on the parameters:
syst_th <- param_syst(c(-0.8,-0.3,1), n= 1, s=1)

# plot the impulse response
ir_th <- impulse(syst_th,plot=1)

# same for syst_uc
(syst_uc <- param_syst_uc(theta_uc))
ir_uc <- impulse(syst_uc,plot=1)
```

For the UC system we have more than one impulse response, since we have a total of 8 independent noise components (one in the observation equation and 7 in the state equation).

Properties of linear processes

Are LTI systems special?

- The examples showed that LTI state space systems can represent AR, MA and ARMA models.
- Also UC and similar models are of this form.
- Time invariance is often plausible, in particular for processes for which stationarity is plausible.
- Empirically they are used in many contexts successfully.
- Linearity is not as restrictive as it seems on first sight ...

4. Properties of linear processes

Theorem (Wold Decomposition, Hannan and Deistler (2012) Thm 1.3.1.)

Each wide sense stationary stochastic process $(y_t)_{t\in\mathbb{Z}}$ can be uniquely decomposed into the sum of two orthogonal terms:

$$y_t = d_t + z_t,$$

where $(d_t)_{t\in\mathbb{Z}}$ is deterministic* and $(z_t)_{t\in\mathbb{Z}}$ is a MA(∞)-process. Both processes $(d_t)_{t\in\mathbb{Z}}, (z_t)_{t\in\mathbb{Z}}$ can be written as (limits of) functions of $y_s, s \leq t$.
For z_t there exists a representation as:

$$z_t = \varepsilon_t + \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}, \quad \sum_{i=0}^{\infty} \|b_i\|^2 < \infty$$

with white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ such that ε_t can be written as a function of $z_s, s \leq t$.

Thus d_t represents a constant, a linear trend, seasonal dummies and so forth.

We can focus on $z_t = y_t - d_t$ for modeling.

$*$ Deterministic means that d_t can be perfectly predicted based on its past d_s $(s < t)$.

4. Properties of linear processes

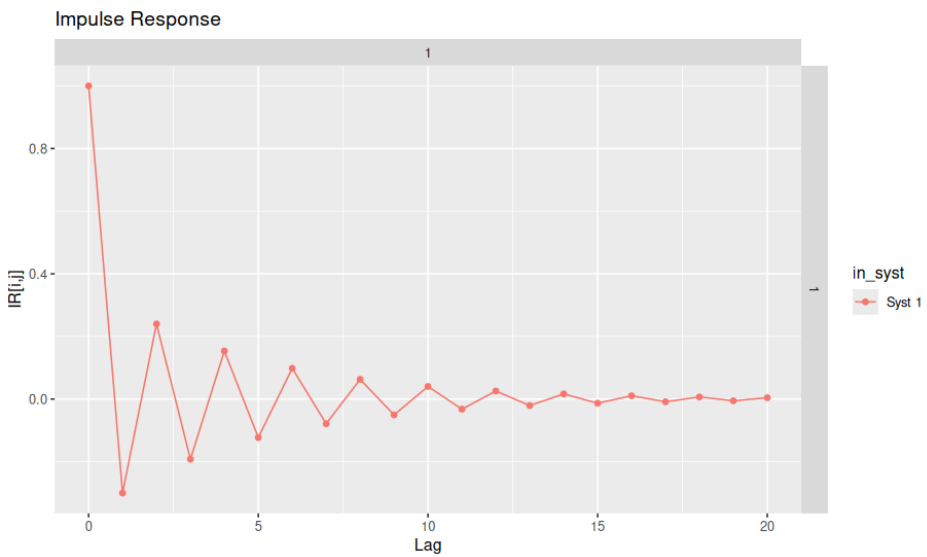
- The noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ has expectation zero and variance Σ.
- Previously we used iid sequences as white noise inputs for the state space system.
- This is appropriate in the context of Markov chains.
- The Wold decomposition is based on projections, which make a different concept more appropriate.
- Here the noise is also called *innovation* sequence and is often modelled as a martingale difference sequence (mds).

Impulse and frequency response

4. 1. Impulse and frequency response

The Wold representation provides a representation in two parts:

- d_t is linearly singular and typically modelled using deterministic components like a trend and seasonal terms (dummies or sin/cosine functions).
- $z_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}$.
- The properties of $(z_t)_{t\in}$ are described using the impulse response and the frequency response.

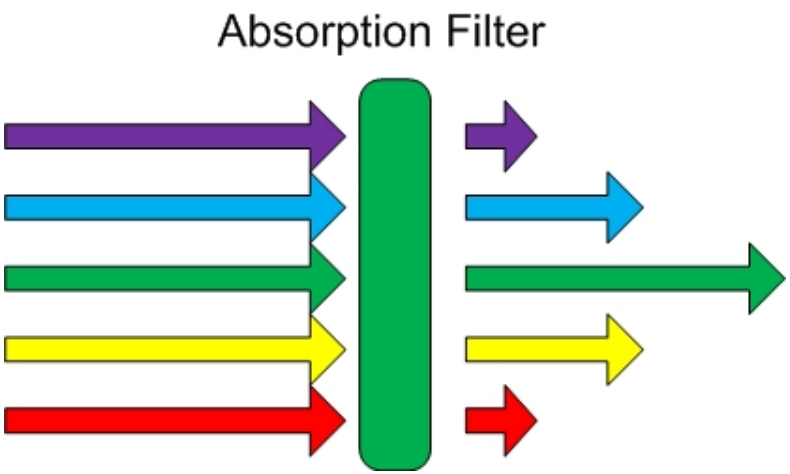


4. 1. Impulse and frequency response

- The impulse response sequence $b_j, j = 0, 1, 2, ...$ can be mapped to the transfer function:

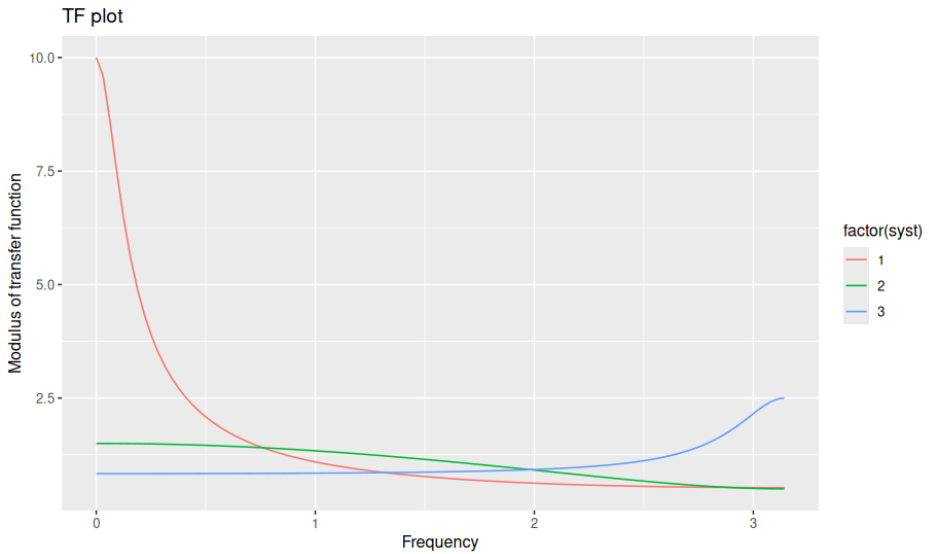
$$b(z) = \sum_{j=0}^{\infty} b_j z^j \quad (z \in \mathbb{C}) \quad \text{or} \quad b(e^{i\omega}) = \sum_{j=0}^{\infty} b_j e^{ij\omega} (\omega \in [0, 2\pi))$$

- The real value ω has the interpretation of a frequency.
- Thus $\tilde{b}(\omega) := b(e^{i\omega})$ is known as the frequency response.
  $\varepsilon_t = e^{-it\omega} \Rightarrow z_t = \varepsilon_t \tilde{b}(\omega)$.



4. 1. Impulse and frequency response

Different transfer functions



- $k_{red}(z) = \frac{1}{1-0.9z}$
- $k_{green}(z) = 1 + 0.5z$
- $k_{blue}(z) = \frac{1+0.5z}{1+0.8z}$

```
1 # frequency response / Bode plot
2 trs <- transfer_function(systs,plot=1)
```

Special case LTI state space system in innovation form

$$y_t = Cx_t + \varepsilon_t, \qquad x_{t+1} = Ax_t + K\varepsilon_t.$$

- Stationary solution for $\lambda_{|max|}(A) << 1$: $y_t = \varepsilon_t + \sum_{i=0}^{\infty} CA^{i-1}K\varepsilon_{t-i}$.
- Impulse response: $b_i = CA^{i-1}K$
- Frequency response (transfer function):

$$\begin{aligned} b(z) &= I_s + \sum_{i=0}^{\infty} CA^i K z^{i+1} \\ &= I_s + zC(\sum_{i=0}^{\infty} A^i z^i)K \\ &= I_s + zC(I_n - zA)^{-1}K. \end{aligned}$$

Covariance Sequence

4. 2. Covariance Sequence

In general: $\gamma_k = \mathrm{Cov}(z_t, z_{t-k}) = \sum_{i=0}^{\infty} b_{i+k} \Sigma b_i'$.

Special Case: LTI system:

$$y_t = Cx_t + v_t, \qquad x_{t+1} = Ax_t + w_t,$$

$$\mathrm{Var}\begin{pmatrix} v_t \\ w_t \end{pmatrix} = \begin{pmatrix} R & S' \\ S & Q \end{pmatrix}$$

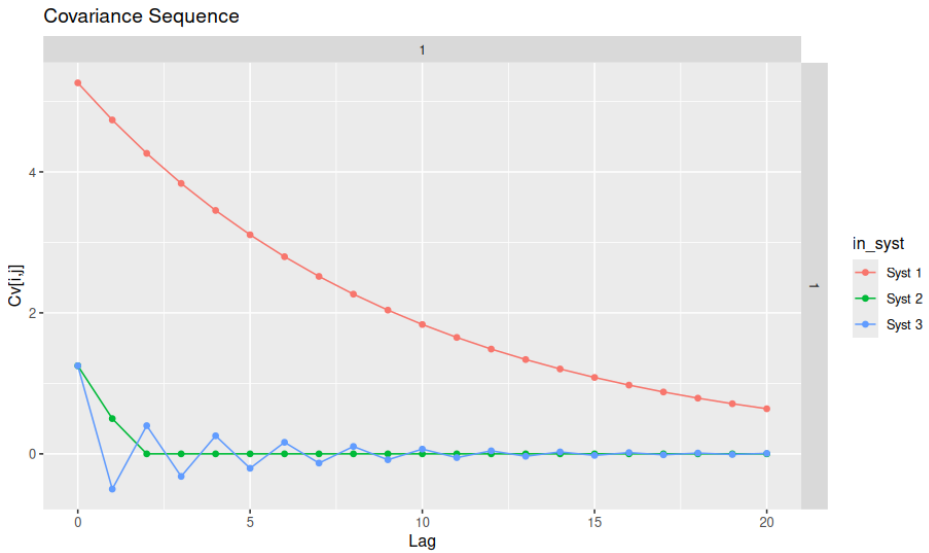
- Variance of the state:

$$P_{t+1} = \mathrm{Var}(x_{t+1}) = \mathrm{Var}(Ax_t + w_t) = AP_tA' + Q,$$
$$P = APA' + Q \qquad \text{(Lyapunov equation)}$$

- Covariances: $k > 0$, stationary case

$$\begin{aligned}\mathrm{Cov}(z_{t+k}, z_t) &= \mathrm{Cov}(Cx_{t+k} + v_{t+k}, z_t) = \mathrm{Cov}(Cx_{t+k}, z_t) \\ &= C\,\mathrm{Cov}(A^{k-1}x_{t+1} + \ldots + A^{k-2}w_{t+1}, z_t) = CA^{k-1}\,\mathrm{Cov}(x_{t+1}, z_t) \\ &= CA^{k-1}\,\mathrm{Cov}(x_{t+1}, Cx_t + v_t) = CA^{k-1}(APC' + S')\end{aligned}$$

4. 2. Covariance Sequence



- $k_{red}(z) = \frac{1}{1-0.9z}$
- $k_{green}(z) = 1 + 0.5z$
- $k_{blue}(z) = \frac{1+0.5z}{1+0.8z}$

```
# covariance function
trs <- cov_fun(systs,plot=1)

# results in a warning and a non-sensical sequence.
tr_uc <- covariance_function(syst_uc,plot=1)
```

Prediction

4. 3. Prediction

For prediction of the output given the state and the past of the observations:

$$\begin{bmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+f} \end{bmatrix} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^f \end{bmatrix}}_{\text{prediction}} x_t + \underbrace{\mathcal{E}_f E_{t+f}^t}_{\text{prediction error}}$$

Therefore (with $Y_{t-1} = [y'_{t-1}, y'_{t-2}, ...]'$)

$$\hat{y}_{t+h|t} = \mathbb{E}(y_{t+h}|x_t, Y_{t-1}) = \mathbb{E}(y_{t+h}|x_t) = CA^h x_t.$$

For predictions (with horizon $h \geq 0$) based on the finite past we obtain:
$\tilde{y}_{t+h|1,..,t-1} = CA^h \tilde{x}_{t|1,...,t-1}$

Estimation of Parameters

Quasi-Maximum Likelihood estimation

- Estimation often uses the Gaussian likelihood, even if the process is not assumed to be normally distributed.
- This is also done for state space systems.
- For parameter vector $\theta \in \mathbb{R}^{2sn}$ we obtain system $(A(\theta), C(\theta), K(\theta))$.
- For system $(A(\theta), C(\theta), K(\theta))$ we have covariances $\gamma_j(\theta)$.
- The process $(y_t)_{t\in\mathbb{Z}}$ then is modelled as

$$y_t = C(\theta)x_t(\theta) + \Phi d_t + \varepsilon_t, \quad x_{t+1}(\theta) = A(\theta)x_t(\theta) + K(\theta)\varepsilon_t$$

- d_t are deterministic terms such as a constant, a linear trend and seasonal components adding to the parameter vector.
- Then the observations (up to deterministics) $\tilde{Y}_T = [(y_1 - \Phi d_1)', (y_2 - \Phi d_2)', ..., (y_T - \Phi d_T)']' \in \mathbb{R}^{sT}$ have expectation zero and covariance matrix denoted as $\Gamma_T(\theta, \Sigma)$.

Quasi-maximum likelihood estimation (II)

- Criterion function proportional to $-2\times$ the log of the Gaussian likelihood (up to constants)

$$L_T(\tilde{Y}_T; \Phi, \theta, \Sigma) = \log\det(\Gamma_T(\theta, \Sigma)) + \tilde{Y}_T' \Gamma_T(\theta, \Sigma)^{-1} \tilde{Y}_T$$

- The parameter vector $\theta \in \mathbb{R}^{2sn}$ does not necessarily correspond to a stable and minimum-phase (invertible) system.
- Thus the parameter space needs to be restricted. Alternatively we can impose penalties for violation of restrictions.
- The covariance matrix $\Gamma_T(\theta)$ also depends on the variance of the state x_0.
- Three options:
  1. Use the stationary distribution.
  2. Set it to zero: $x_0 = 0$.
  3. Estimate the state as additional parameters.

Theorem (Hannan and Deistler (2012) Theorem 4.2.1., 4.3.1., 5.5.1.)

Let $(y_t)_{t\in\mathbb{Z}}$ be generated from a stable and strictly minimum-phase linear state space system of order n in the generic[1] neighborhood of the parameterisation corresponding to the true parameter θ_0. Let the noise process $(\varepsilon_t)_{t\in\mathbb{Z}}$ be an iid sequence[2] with expectation zero and variance $\Sigma_0 > 0$. Assume that we observe $y_t, t = 1, ..., T$, from such a process for $d_t = 1$ and that the parameter vector estimator $\hat{\theta}, \hat{\Phi}, \hat{\Sigma}$ are obtained as minimizing arguments of $L_T(\tilde{Y}_T; \Phi, \theta, \Sigma)$. Then:

(i) The estimators using the stationary distribution for $x_0(\theta)$ are consistent for $T \to \infty$: $\hat{\theta} \to \theta_0, \hat{\Sigma} \to \Sigma_0, \hat{\Phi} \to \Phi_0$.

(ii) The estimators are asymptotically normal: $\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V_\circ)$.

(iii) If the order is estimated using `BIC`, then the order estimate is consistent. If `AIC` is used, then the estimated order is no smaller than the true order.

(iv) If $n > n_0$ the estimated impulse response $\hat{C}\hat{A}^{j-1}\hat{K} \to C_0 A_0^{j-1} K_0$ for all $j \in \mathbb{N}$.

[1] The parameterisation is only defined on a generic subset, not the set of all transfer functions of order n.

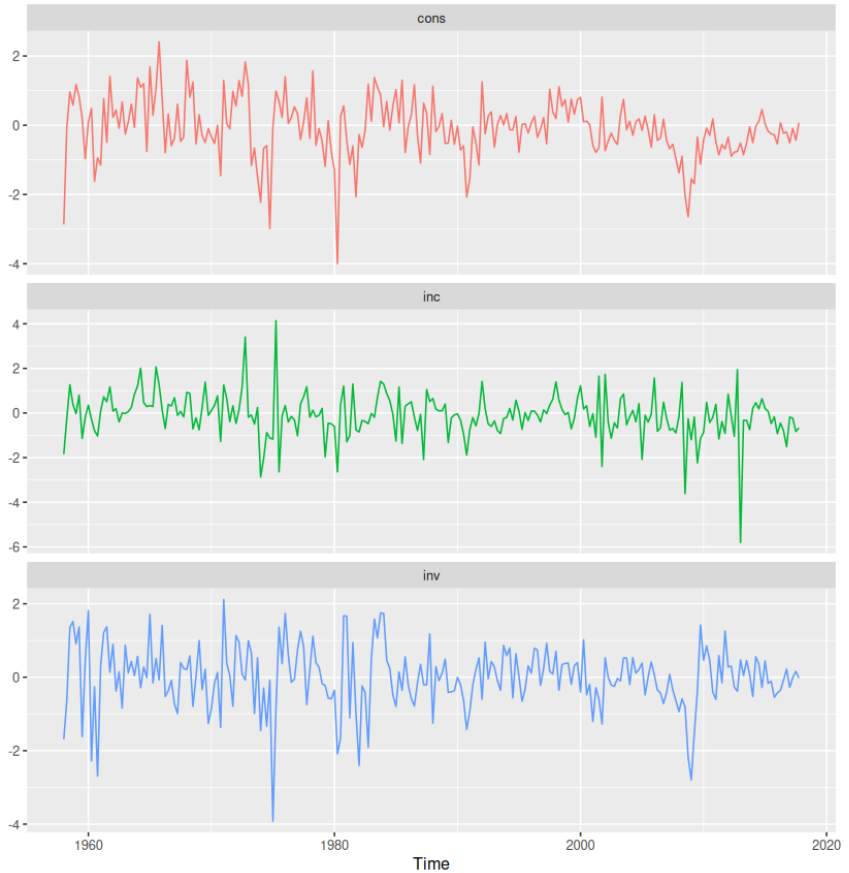
[2] Hannan and Deistler use mds assumptions.

Remarks

- The qMLE needs to be obtained from numerical optimization.
  - ▸ Ingredients for optimization
- The result means that increasing the sample size leads to better estimates. Additionally we can characterise the asymptotic distribution.
- The asymptotic variance matrix V_0 can be estimated consistently, but depends on the assumptions on the noise.
- For small sample sizes using the asymptotic distribution may not provide a good approximation for the finite sample distribution. In such cases the method of bootstrapping can help. *Jens-Peter Kreiß, Tue AM*.
  - ▸ Idea of bootstrapping
- When using the alternative assumptions on the initial values one gets asymptotically equivalent estimators.

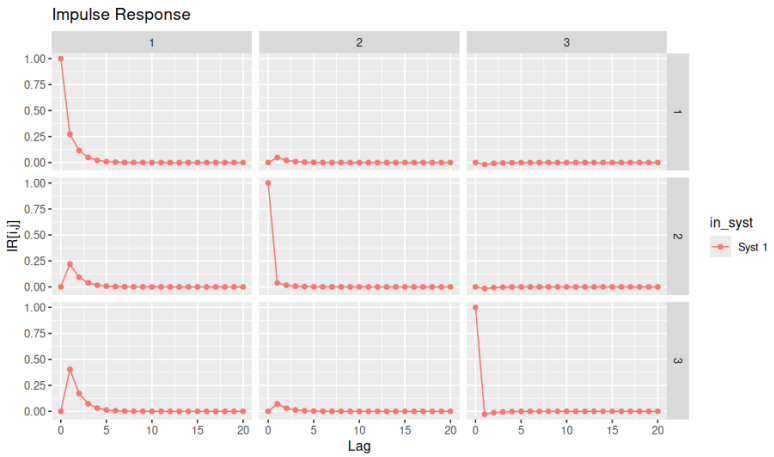
5. 1. Example

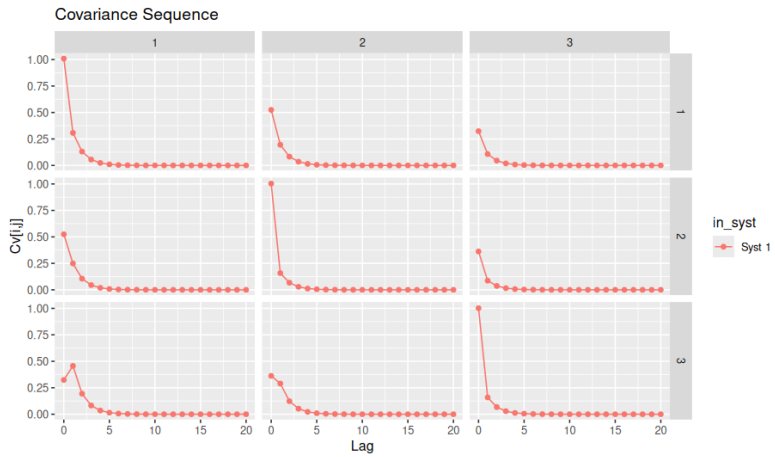
- We use the data and example contained in Scherrer and Deistler (2019).
- Three quarterly measured variables (in billions of chained 2009 US dollars) obtained from FRED:
  - `DPIC96`: Real Disposable Personal Income
  - `GPDIC1`: Real Gross Private Domestic Investment
  - `PCECC96`: Real Personal Consumption Expenditures
- Quarterly growth rates (first difference of logs), demeaned and rescaled to variance equal to 1 (zscores).
- 1985-1991 used for estimation, 1992-2017 for validation.



5. 1. Example

Example






```
out <- est_qml(t(y_mat), 1, 10, pe=FALSE, print.level=2,iterlim =
    1000)
summary_qmle( out,n=1, s=3)

# convert to system
syst_est <- param_syst(out$estimate,n=nest,s=3)

# plot impulse response
impulse(syst_est,plots=1)
covariance_function(syst_est,plots=1)
```

The Kalman Filter

General idea

$$L_T(\tilde{Y}_T; \Phi, \theta, \Sigma) = \log\det(\Gamma_T(\theta, \Sigma)) + \tilde{Y}_T' \Gamma_T(\theta, \Sigma)^{-1} \tilde{Y}_T$$

- The calculation of the criterion L_T is numerically hindered by the inversion of $\Gamma_T(\theta, \Sigma) \in \mathbb{R}^{Ts \times Ts}$ which could be a large matrix.
- Determinant and inverse can be calculated based on an LDL decomposition:

$$\Gamma_T(\theta, \Sigma) = \begin{pmatrix} I_s & 0 & \dots & 0 \\ L_{21} & I_s & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ L_{T,1} & \dots & L_{T,T-1} & I_s \end{pmatrix} \begin{pmatrix} D_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & D_T \end{pmatrix} \begin{pmatrix} I_s & L_{21} & \dots & L_{T,1} \\ 0 & I_s & \ddots & \vdots \\ \vdots & \ddots & \ddots & L_{T,T-1} \\ 0 & \dots & 0 & I_s \end{pmatrix}$$

- Then L^{-1} is lower block diagonal as well with determinant equal to 1.
- Thus the variance of $L^{-1}\tilde{Y}_T$ is block diagonal and the t-th block row reads

$$\tilde{y}_t - \sum_{j=1}^{t-1} \tilde{L}_{tj} \tilde{y}_{t-j}$$

being orthogonal to all values $\tilde{y}_s$, $s < t$ such that $\sum_{j=1}^{t-1} \tilde{L}_{tj}\tilde{y}_{t-j}$ denotes the prediction of $\tilde{y}_t$ based on $\tilde{y}_1, \dots, \tilde{y}_{t-1}$.

Relation to state space systems

$$y_t = Cx_t + v_t, \quad x_{t+1} = Ax_t + w_t$$

REMARK: $\Phi = 0$ (no deterministics) for illustration purposes.

- Regress y_t onto $Y_{t-1} = (y_1', ..., y_{t-1}')'$:

$$\begin{aligned} y_t &= Cx_t + v_t = \beta Y_{t-1} + e_t \\ &= C\hat{x}_{t|t-1} + e_t && (Y_{t-1} \perp v_t), \\ \hat{x}_{t+1|t} &= A\hat{x}_{t|t} + \hat{w}_{t|t} && (Y_{t-1} \perp w_t) \\ &= A\hat{x}_{t|t-1} + K_t e_t. \end{aligned}$$

- This is a recursion in time involving only 'small' dimensional matrices.
- The kernel of the procedure is the calculation of the matrix K_t, the Kalman gain.

6. 1. Filter equations

The **Kalman filter**:

Initial Values	$\hat{x}_{1\|0}$	$= 0,$
	$P_{1\|0}$	$= \mathrm{Var}(x_1),$
Innovation	e_t	$= y_t - C\hat{x}_{t\|t-1},$
Measurement update	$\hat{x}_{t\|t}$	$= \hat{x}_{t\|t-1} + P_{t\|t-1}C'\Sigma_t^{-1}e_t,$
Innovation variance	Σ_t	$= CP_{t\|t-1}C' + \Sigma,$
Est. variance	$P_{t\|t}$	$= P_{t\|t-1} - P_{t\|t-1}C'\Sigma_t^{-1}CP_{t\|t-1}$
Time-update	$\hat{x}_{t+1\|t}$	$= A\hat{x}_{t\|t} + S\Sigma_t^{-1}e_t$
		$= A\hat{x}_{t\|t-1} + K_t e_t,$
Kalman gain	K_t	$= (AP_{t\|t-1}C' + S)\Sigma_t^{-1},$
Est. variance	$P_{t+1\|t}$	$= AP_{t\|t-1}A' + Q - K_t\Sigma_t K_t'$

- Here $\hat{x}_{t|s}$ denotes the LS optimal estimate of x_t based on Y_s.
- $P_{t|t-1}$ ... state estimation error variance $\mathrm{Var}(\hat{x}_{t|t-1} - x_t)$.
- Analogous to the forward algorithm for Hidden Markov Models (HMM) *Roland Langrock, Fri, AM*

6. 2. Remarks

REMARKS:

- One can view the Kalman filter as a means to calculate the likelihood. Or we can view it as a means to estimate the state x_t corresponding to the innovation representation (for prediction, for example).
- The formulas in Hannan and Deistler (2012) can be used starting from a general state space equation with different (also correlated errors) v_t, w_t in both equations.
- The system matrices can be different in each time step (time varying systems are allowed; may also depend on measured signals, if one conditions on them).
- The recursions can be extended to also include smoothers, that is $\hat{x}_{s|t}, s < t$ by adding a backward pass $s = t, t-1, ...$
- Note that Gaussianity is not needed for the filter to return optimal estimates in the same way as the OLS estimator is the BLUE without Gaussianity assumptions.
- More details on Kalman filtering is given in Anderson and Moore (1979) and Durbin and Koopman (2012).

Results

- The Kalman filter draws attention mostly because of its simplicity and optimality property: It provides the best (in LS sense) guess of the state, independent of assumptions on the distribution of the errors (no Gaussianity required!).
- Anderson and Moore (1979) provide conditions so that the Kalman gain $K_t \to K$ and $\Sigma_t \to \Sigma$. For stable matrix A in minimal systems this always is the case.
- This holds also for the general form with different noise v_t and w_t in which case it can be shown that for the resulting $A - KC$ all eigenvalues are in the unit disc.
- Note that this corresponds to the innovation representation!
- Using the prediction error approach leads to the same asymptotic distribution, but results in much faster estimates.

Bayesian Estimation

Main idea

Only a first glimpse, more: *Gregor Kastner, Thu AM*.

- Frequentist approaches start with the assumption that the data is generating from within the model using a 'true' parameter vector $\theta_\circ$.
- Estimation and inference then center on information extracted from the data about this 'true' parameter vector.
- Bayesian methods start by replacing this by the idea that the model including the parameters are a way to express the assessment of the modeller on the data to be expected.
- Before the data is collected, all prior information is encoded into a 'prior density' $p(\theta)$.
- Then the data is collected according to the model using the conditional distribution $p(Y|\theta)$ related to the likelihood.
- After collecting the data we obtain a joint distribution of data and parameters as $p(\theta)p(Y|\theta) = p(\theta, Y)$.
- Bayes theorem allows us to calculate the *posterior* distribution $p(\theta|Y)$.

Estimating the mean of the posterior

Approach A: brute force:

- sample B times from the prior
- calculate the likelihood for each value of the prior and the data
- normalise the likelihood values to sum to one.
- estimate the weighted mean using the normalised likelihood values as weights.

This estimates

$$\bar{\theta} = \frac{\int \theta p(Y_T|\theta)p(\theta)d\theta}{\int p(Y_T|\theta)p(\theta)d\theta} \quad \text{as} \quad \hat{\theta} = \frac{\sum_{b=1}^{B} \theta^{(b)} p(Y_T|\theta^{(b)})}{\sum_{b=1}^{B} p(Y_T|\theta^{(b)})}$$

Estimating the mean of the posterior (II)

Approach B: importance sampling using a proposal density $g(\theta)$:

- sample B times from the proposal $g(\theta)$
- calculate the likelihood for each value in the sample and the data
- normalise the product of the likelihood with the prior divided by the proposal pdf at the samples
- normalise weights to sum to one.
- estimate the weighted mean.

This estimates

$$\bar{\theta} = \frac{\int \frac{\theta p(Y_T|\theta)p(\theta)}{g(\theta)} g(\theta) d\theta}{\int \frac{p(Y_T|\theta)p(\theta)}{g(\theta)} g(\theta) d\theta}$$

Similar estimators can be obtained for other functions of the parameter.

Additional Reading

8. Additional Reading

- A newer textbook containing much of the structure theory for VARMA and state space models (and is probably easier to read for many readers than Hannan and Deistler (2012)) is Deistler and Scherrer (2022).
- An excellent book on state space modelling is Durbin and Koopman (2012).
- A comprehensive book on unobserved component models is Harvey and Proietti (2005).
- Information on stationarity, ergodicity and the related consistency for transformed processes is presented in Pötscher and Prucha (2013).
- Ergodicity in the context of Markov chains is treated in much generality in Meyn and Tweedie (2012).
- The website Hyndman and Killick (2022) collects are large number of R packages that include Kalman filters and other state space modelling tools: CRAN task view time series analysis
- Hyndman also (co-)authored a very nice textbook as an intro to time series modeling using the component model. website of the book.
- Tusell (2011) provides a discussion of `R` packages to calculate the Kalman filter, smoother in various different versions.

References

9. References

Anderson, B. and Moore, J. B. (1979). *Optimal filtering*. Prentice-Hall Information and System Sciences Series, Englewood Cliffs: Prentice-Hall, 1979.

Deistler, M. and Scherrer, W. (2022). *Time Series Models*. Springer International Publishing.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Hannan, E. J. and Deistler, M. (2012). *The statistical theory of linear systems*. SIAM.

Harvey, A. and Proietti, T. (2005). *Readings in unobserved components models*. OUP Oxford.

Hyndman, R. J. and Killick, R. (2022). Cran task view: Time series analysis.

Kotharkar, V. and Stein, M. L. (2026). An interpretable low-rank state-space model for multi-horizon simulation of large-scale regional temperature fields.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Pötscher, B. M. and Prucha, I. R. (2013). *Dynamic nonlinear econometric models: Asymptotic theory*. Springer Science & Business Media.

Scherrer, W. and Deistler, M. (2019). Vector autoregressive moving average models. In *Conceptual Econometrics Using R*, volume 41 of *Handbook of Statistics*, pages 145–191. Elsevier.

Tusell, F. (2011). Kalman filtering in r. *Journal of Statistical Software*, 39(2):1â€“27.

What is needed for estimation?

Parameterization:

- `param_system`: calculating the system corresponding to the parameters.
- `syst_param`: calculating the parameters corresponding to the system.
- important: handling restrictions (transform unrestricted parameters to restricted entries)

Criterion function evaluation:

- `cal_crit` evaluating the criterion function for a combination of data set and parameter vector.
- may return criterion value plus gradient plus Hessian.
- may contain arguments for numerical speed (like analytical gradient calculation) or versions of the criterion

Optimization

- often built in optimization routines (nlm, optim in `R`, fminunc in `MATLAB`, ...).
- often these routines have a number of subvariants (BFGS, BHHHH, ...).
- has to deal with restrictions: $\lambda_{|max|}(A) < 1$ is a 'nasty' one: barrier methods sometimes help.
- requires a starting value: very problem dependent. Should be a fast procedure; random initialization is an option, but does not work for many parameters.

Quantification of Uncertainty:

- asymptotic theory provides formulas, that typically involve the Hessian and the variance of the gradient.
- the criterion function often provides these information; one has to tell the optimizer to return this info.
- The distribution maybe needs to be simulated or bootstrapped.

Back to Presentation

Main idea of bootstrapping

- Bootstrapping is a technique to assess the accuracy of the estimator and/or to adjust possible finite sample biases.
- It is a simulation driven method.
- As such it often requires assumptions on how the data was generated.
- Some versions of it are very flexible.

11. Bootstrap main ideas

Origin of the name



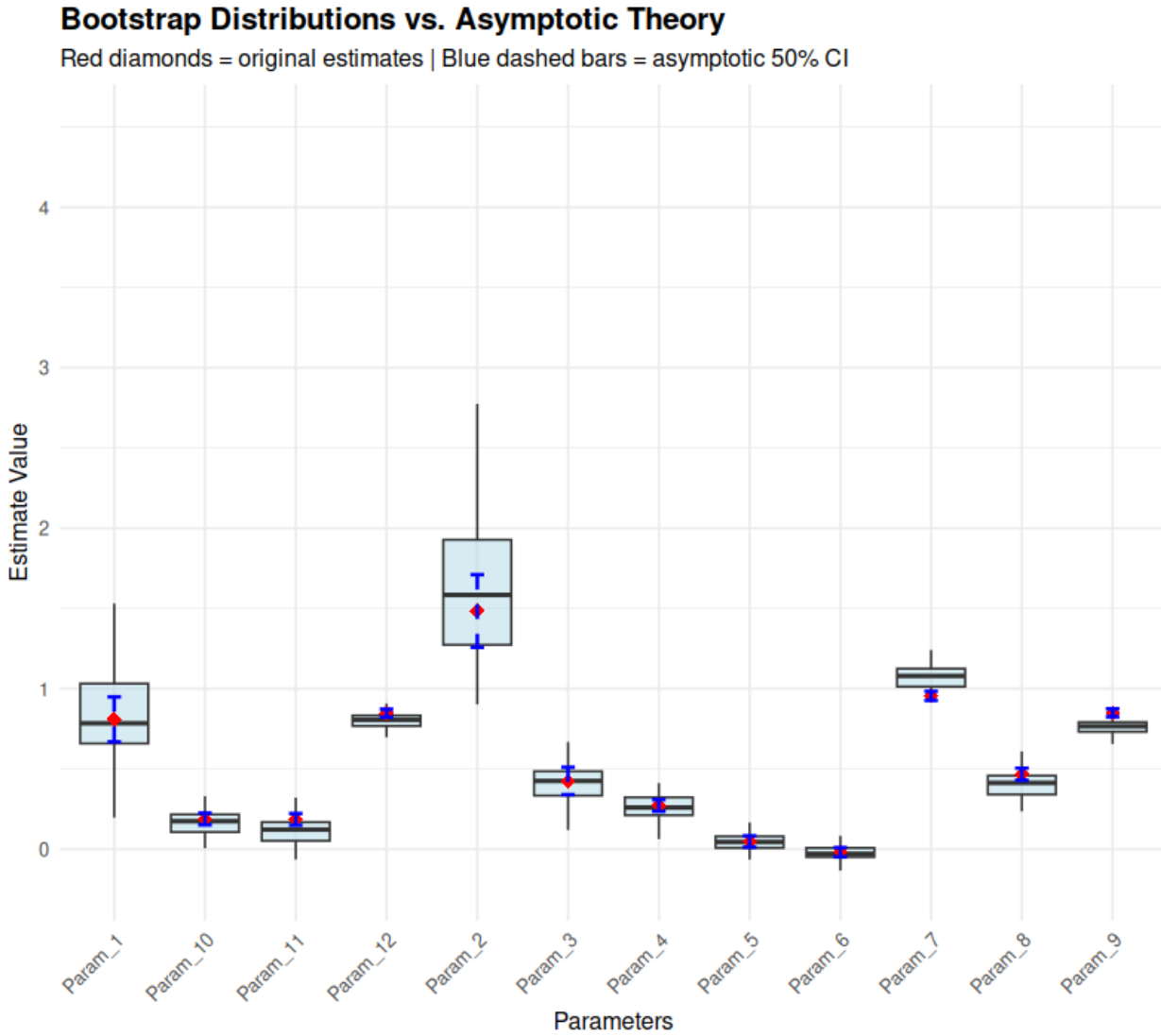
Source: Google Gemini.

Illustration

- LTI system estimated from data.
- Replications are obtained from simulations using new iid (Gaussian) noise terms.

```
M <- 100
boot_est <- matrix(NA,12,M)
pb <- txtProgressBar(min = 1, max = M, style = 3)
for (m in 1:M){
  set.seed(m)
  setTxtProgressBar(pb, m)
  y <- sim_syst(syst_est,T= 136,stationary=TRUE)
  y_matm <- tsibble_to_mat(y)
  outm <- est_qml(t(y_matm), nest, 10, pe=TRUE, print.level=0,
     iterlim = 1000)
  boot_est[,m]<- outm$estimate
}
close(pb)
```

11. Bootstrap main ideas



Back to Presentation