Modeling and Identification of Stationary Stochastic Processes

ALESSANDRO CHIUSO

Department of Information Engineering,
Università di Padova, Italy

June 2012

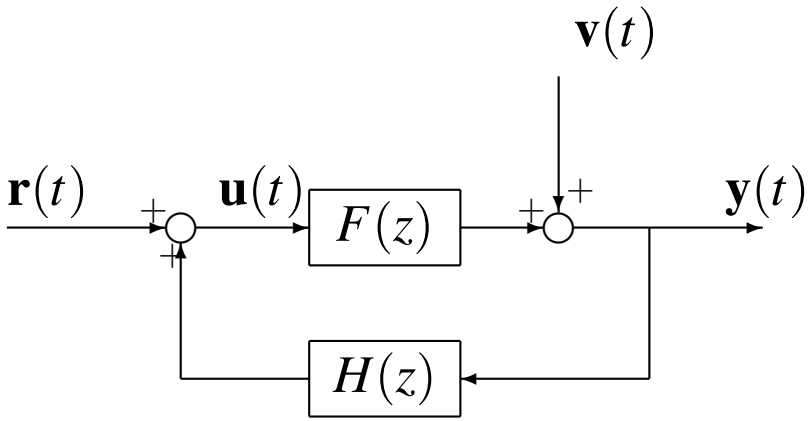
Main Objectives of this lectures

To provide the background on modeling stationary stochastic processes with linear stochastic models as well as to introduce the main issues which arise in their identification from measured data.

OUTLINE

1. Motivation: Modeling and Identification
   - Feedback Interconnections
   - Main Ingredients: Data, Models and Criteria
2. Review of linear models of stationary processes
   - General facts about stationary processes
   - Spectral factorization and ARMA Models
3. Interconnected Stochastic Systems
   - Modeling with Inputs
   - Feedback Interconnections and Granger Causality
   - Dynamic Networks
4. System Identification
   - Data: trajectories from stationary and ergodic processes
   - Models: Linear models (finitely parametrized?)
   - Criteria: Prediction Error Methods and Model Complexity Selection

Motivation: Modeling and Identification of (interconnected) Stochastic Systems



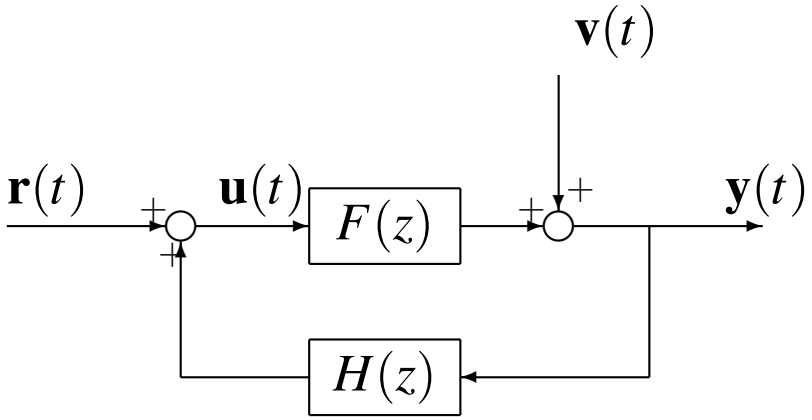
Joint model for the "processes" $\mathbf{y}$ e $\mathbf{u}$ (we haven't defined what the are, yet...)

1. The signals $\mathbf{v}$ e $\mathbf{r}$ are to be seen as "exogenous" (references, noises etc.)

2. Feedback may be intrinsic and does not necessarily correspond to "physical" control loops..

3. This scheme can be more "complicated" resulting in a "network" of interconnections; this interconnection has to be *well posed* and *internally stable*.

Motivation: Modeling and Identification of (interconnected) Stochastic Systems



PROBLEMS:

1. Which kind of "stochastic processes" $\mathbf{y}$ e $\mathbf{u}$ (not defined, yet...) does this model describe?

2. How do I get an estimate of $F(z)$ and of the "statistical" properties of the "disturbance" $\mathbf{v}$ from measured data $\{u(1), y(1), ..., u(T), y(T)\}$?

INGREDIENTS OF AN IDENTIFICATION PROBLEM

- DATA: $\{u(1), y(1), ..., u(T), y(T)\}$
- MODELS: linear/nonlinear, parametric/non parametric etc.
- CRITERIA: how do we choose a good model? how do we choose the model class and its complexity (order for linear parametric models)?

Essential ingredients #1: The Data

We need to introduce a *probablistic description of the data*. The data at our disposal at some fixed time instant represent only partial evidence about the behavior of the system as we do not know the future continuation of the input and output time series. Yet,
all possible continuations of our present data must carry information about the same physical phenomenon we are about to model, and hence the possible continuations of the data cannot be "totally random" and must be related to what we have observed so far. So, data must have a "memory"; i.e. their own dynamics, and in order to discover models of systems, we have to first understand models of uncertain signals.

Data will be modeled as **stochastic processes** in fact discrete-time stochastic processes. Since the underlying phenomenon (system) that we want to describe is assumed to be time invariant the stochastic processes which model the observed data will be stationary. There are important exceptions where data need not be stationary but only are there some statistical properties (e.g. some model parameters) which are stationary.

In addition our data have to be “rich enough” to reveal the underlying statistical properties. In practice ensemble properties (e.g. means, covariances etc.) are to be estimated from *ONE* sample trajectory:
this requires **ERGODICITY**.

Essential ingredients #2: The model class

In real systems, there are always many other variables besides the preselected inputs and outputs which influence the time evolution of the system. These variables represent the unavoidable interaction of the system with its environment. For this reason, even in the presence of a true causal relation between inputs and outputs there always are some *unpredictable* fluctuations of the values taken by the measured output $\mathbf{y}(t)$ which are not explainable in terms of past input (and/or output) history.

We cannot (and do not want to) take into account these variables explicitely in the model as some of them may be inaccessible to measurement and in any case this would lead to complicated models with too many variables. We need to work with models of small complexity and treat the unpredictable fluctuations in some simple *aggregate* manner.

Models (however accurate) are *always mathematical idealizations of nature*. No physical phenomenon, even if the experiments were conducted in an ideal interactions-free environment can be described **exactly** by a set of differential or difference equations and even more so if the equations are a priori restricted to be linear, finite-dimensional and time-invariant. So the observables, even in an ideal "disturbance-free" situation cannot be expected to obey *exactly* any linear time-invariant model.

A realistic formulation of the problem requires a satisfactory notion of *non-rigid*, i.e. *flexible or approximate*, notion of mathematical model of the observed data.
A model should be able to accept as legitimate, data sets (time series) which may possibly differ slightly from each another.

Imposing rigid ”exact” descriptions of the type $F(u,y) = 0$ to experimental data has been criticized since the early beginnings of experimental science. Particularly illuminating is Gauss' general philosophical discussion in *Theoria motus corporum caelestium* sect. III, p. 236.

Example: there has been a widespread belief in the early years of control science that identification was merely a matter of solving (exactly) for h a linear convolution equation

$$y(t) = \sum_{t_0}^{t} h(t-\tau)u(\tau) \tag{1}$$

or, equivalently, by matching exactly pointwise harmonic response data with linear transfer function models. Results have always been extremely sensitive even to small perturbations in the data.

New incoming data tend to change the model drastically, which means that a model determined in this way has very poor predictive capabilities.

The reason is that data obey exactly rigid relations of this kind “with probability zero”. If in addition the model class is restricted to be finite-dimensional, which is what is necessary for control applications, imposing the integral equation model (1) on real data normally leads to disastrous results. This is by now very well-known and documented in the early literature. In the language of numerical analysis, fitting rigid models to measured data invariably leads to very *ill-conditioned problems*.

We shall follow Gauss idea of describing data by a *distribution function*; i.e. work in a **probabilistic setting**. Models will then be probabilistic objects.

Other alternatives are possible, say using deterministic model classes consisting of a rigid “exact” model as a “nominal” object, plus an uncertainty ball around it. In this case, besides a nominal model, the identification procedure is required to provide at least bounds on the magnitude of the relative “uncertainty region” around the nominal model.

Here one should provide a mathematical description of how the dynamic uncertainty ball is distributed in the frequency domain, rather than, as more traditionally done, in the parameter space, about the nominal identified model.

Essential ingredients #3: The Model Selection Criterion

In these lectures we shall take the probabilistic point of view and model uncertainty with the apparatus of probability theory. In this framework **identification is essentially a problem of mathematical statistics**. General idea: *minimize criteria based on a notion of distance between the data and the model class*.

Trivial example: Least squares fitting. Note that in Gauss' work least squares come out as a solution method for optimally fitting a certain class of *density functions* to the observed data (maximum likelihood).

Nota Bene: the basic problem of identification is, much more than designing algorithms which fit models to observed data (the easy part), the quantification of the *uncertainty bounds* or the description of the *dynamic errors* which will be incurred when using the model with generic data. Any sensible identification method should provide some mathematical description of how uncertainty is distributed in time or frequency about the nominal identified model. In this respect statistics and probability offer an ideal framework. Describing a probability distribution is the same as modeling uncertainty.

A common critique

It has been argued that the abstract "urn model" of probability theory looks inadequate to deal with situations like the one we have envisaged, where there is just one experiment and there is really no sample space around from which the results of the experiment could possibly have been drawn.
The critique has the merit of criticizing large sectors of the literature where the statistical framework is often imposed dogmatically.
In our opinion however, the critique originates from a tendency to confuse physical reality with mathematical modeling. In fact the urn model (i.e. the underlying probability space) is just a mathematical device which is *not required to have any physical interpretation* and could in principle be used to model things which, to be described deterministically, would require extremely complicated mathematical models with myriads of variables.
On the same grounds it could be questioned if there are in nature objects like differential or difference equations.

INGREDIENT #1: DATA

$\{u(1), y(1), ..., u(T), y(T)\}$ where $u(t)$ will be called "input" and $y(t)$ will be called "output" are sample paths from (wide-sense) **stationary** and **ergodic** stochastic processes $\mathbf{u}$ and $\mathbf{y}$ respectively.

Wide-sense stationary random processes

$\mathbf{y} = \{\mathbf{y}(t,\omega)\}$ discrete-time m-dimensional random process $t \in [t_0, +\infty)$.
Expected value: $\mathbb{E}\mathbf{y}(t) = \int_\Omega \mathbf{y}(t,\omega)\, dP = \mu(t)$
can be subtracted off. All random quantities will be **zero mean**. Assume a finite Covariance function:

$$\mathbb{E}\mathbf{y}(t)\mathbf{y}(s)^\top = \Lambda(t,s)\,, \qquad m \times m \quad \text{matrix function}\,.$$

This is the basic mathematical description of the process. A *second order process* is the equivalence class of all stochastic process having (zero mean and) the same covariance function. Contains a Gaussian representative. Second order processes can be described by Linear models.

$\mathbf{y}$ is a (wide sense) stationary process if its covariance function depends on the difference $t-s$: $\Lambda(t,s) \equiv \Lambda(t-s)$.
We shall study stationary processes on the time line $\mathbb{Z}\,(t_0 = -\infty)$.

Hilbert space setting for second order processes

The closure in $L^2(\Omega, P)$ of all finite linear combinations of the random variables $\mathbf{y}_k(t), k = 1, 2, \ldots, m, t \in \mathbb{Z}$, is a Hilbert space

$$\mathbf{H}(\mathbf{y}) := \operatorname{span}\{\mathbf{y}_k(t)\,;\, k = 1, 2, \ldots, m\,;\, t \in \mathbb{Z}\} \equiv \operatorname{span}\{\mathbf{y}(t)\,;\, t \in \mathbb{Z}\}$$

with inner product $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle = \mathbb{E}\{\boldsymbol{\xi}\bar{\boldsymbol{\eta}}\}$.
The shift operator $\mathbf{U} : \mathbf{H}(\mathbf{y}) \to \mathbf{H}(\mathbf{y})$ is the linear extension of

$$\mathbf{U}\mathbf{y}_k(t) := \mathbf{y}_k(t+1)\,, \qquad k = 1, 2, \ldots, m, t \in \mathbb{Z}.$$

$\mathbf{U}$ is Unitary (preserves inner product).

Notation: $\mathbf{H}_t^-(\mathbf{y}) := \operatorname{span}\{\mathbf{y}_k(s)\,;\, k = 1, 2, \ldots, m\,;\, s < t\}$

Hilbert space for second order processes: projections and orthogonality

Let A, B, C be a closed subspaces of $\mathbf{H}(\mathbf{y})$. Define the following symbols:

$(\circ)$: For $\boldsymbol{\eta}, \boldsymbol{\xi} \in \mathbf{H}(\mathbf{y})$, we write $\boldsymbol{\eta} \perp \boldsymbol{\xi} \Leftrightarrow \mathbb{E}\,\boldsymbol{\eta}\boldsymbol{\xi} = 0$

$(\circ)$: $A \perp B$ if any element of A is orthogonal to any element of B.

$(\circ)$: For $\boldsymbol{\eta} \in \mathbf{H}(\mathbf{y})$, $\hat{\boldsymbol{\eta}} := \hat{\mathbb{E}}[\boldsymbol{\eta}|A]$ is the *orthogonal projection* of $\boldsymbol{\eta}$ onto A, i.e. the unique element of A such that $\boldsymbol{\eta} - \hat{\boldsymbol{\eta}} \perp A$

$(\circ)$: $\boldsymbol{\eta} \perp \boldsymbol{\xi}|A$ (conditional orthogonality) if $\boldsymbol{\eta} - \hat{\mathbb{E}}[\boldsymbol{\eta}|A] \perp (\boldsymbol{\xi} - \hat{\mathbb{E}}[\boldsymbol{\xi}|A])$

$(\circ)$ $B \perp C|A$ if, for any $\boldsymbol{\eta} \in B$, $\boldsymbol{\xi} \in C$, $\boldsymbol{\eta} \perp \boldsymbol{\xi}|A$ holds.

NOTA BENE: $\hat{\mathbb{E}}[\boldsymbol{\eta}|A]$ is the conditional expectation in the Gaussian case.

Hilbert space for second order processes: linear prediction

Projection Theorem: the best (minimum variance) linear estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta} \in \mathbf{H}(\mathbf{y})$ based in $\mathbf{y}(s)$, $s < t$, which has the form

$$\hat{\boldsymbol{\eta}} := \sum_{k=1}^{\infty} h_k^{\top} \mathbf{y}(t-k)$$

is given by the orthogonal projection

$$\hat{\boldsymbol{\eta}} = \hat{\mathbb{E}}\left[\boldsymbol{\eta} | \mathbf{H}_t^{-}(\mathbf{y})\right]$$

Purely non deterministic stationary random processes

A stationary random process is purely non deterministic (p.n.d) if it can be represented as the output of a causal ℓ^2-stable linear system driven by a white noise

$$\mathbf{y}(t) = \sum_{k=-\infty}^{t} W(t-k)\mathbf{w}(k)$$

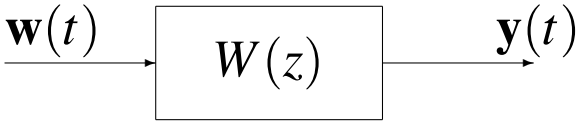
$\{\mathbf{w}(t)\}$ p-dimensional **white noise** process of variance $\mathbb{E}\,\mathbf{w}(t)\mathbf{w}(s)^\top = I_p\delta(t-s)$.
The $m \times p$ impulse response $W(t)$ is a causal function in ℓ^2: $W(t) = 0$ for $t < 0$.
The Fourier transform has an analytic extension $W(z)$ to $\{|z| > 1\}$ in H^2.
The representation is highly non unique. Input white noise is a *latent variable*; special white input is the innovation process = one step prediction error given the infinite past.

FIRST ENCOUNTER WITH INGREDIENT #2: MODELS

This is just one particular instance, we shall see more later on.

Spectrum from shaping filters

Shaping filter representation



Wiener-Kintchine formula gives the **spectral density matrix** $\Phi(e^{j\theta})$ of $\{\mathbf{y}(t)\}$

$$\Phi(e^{j\theta}) = \sum_{k=-\infty}^{+\infty} e^{-j\theta\tau}\Lambda(\tau) = W(e^{j\theta})W(e^{-j\theta})^{\top} \qquad \text{spectral factorization.}$$

FACT: every shaping filter $W(z)$ is a **spectral factor** of $\Phi(z)$.

The covariance function $\Lambda(\tau)$ of a p.n.d process admits Fourier transform.

Positivity: $\Phi(e^{j\theta}) = W(e^{j\theta})W(e^{-j\theta})^{\top} \geq 0$.

Shaping filters and ARMA models

Assume $W(z)$ is a rational matrix function. Since $W(z)$ is **stable** i.e. analytic in $\{|z|>1\}$, can be written as a ratio of polynomial matrices $W(z) = D(z)^{-1} N(z)$ with $\det D(z) \neq 0$ in $\{|z|>1\}$;

$$D(z) = Iz^{\nu} + \sum_{1}^{\nu} A_k z^{\nu-k} \qquad N(z) = N_0 z^{\nu} + \sum_{1}^{\nu} N_k z^{\nu-k}$$

$\{\mathbf{y}(t)\}$ may be described by a (multivariabile) **ARMA model**

$$\mathbf{y}(t) + \sum_{1}^{\nu} A_k \mathbf{y}(t-k) = N_0 \mathbf{w}(t) + \sum_{1}^{\nu} N_k \mathbf{w}(t-k) \quad .$$

There are *many* ARMA model representations !

Shaping filters and ARMA models

In symbolic form we write

$$\mathbf{y}(t) = W(z)\mathbf{w}(t)$$

For the purpose of exposition let us consider the scalar case $m = 1$ (in this course we shall only deal with scalar processes)

We shall be interested in one *special* ARMA model $W_-(z)$, together with its driving noise $\mathbf{e}(t)$ such that both

$$\mathbf{y}(t) = W_-(z)\mathbf{e}(t) \quad \mathbf{e}(t) = W_-^{-1}(z)\mathbf{y}(t)$$

can be interpreted as **causal** systems (of course for this to hold both W_- and W_-^{-1} has to be BIBO stable).

Shaping filters and ARMA models

What do we need?

$$W_-(z) = \frac{\sum_{k=0}^{\nu} b_k z^{-k}}{1 + \sum_{k=1}^{\nu} a_k z^{-k}} = \frac{\sum_{k=0}^{\nu} b_k z^{\nu-k}}{z^{\nu} + \sum_{k=1}^{\nu} a_k z^{\nu-k}}$$

It is necessary and sufficient that

1. $W_-(z)$ is analytic *outside* the closed unit disc $|z| \geq 1$, i.e. $W_-(z)$ is the transfer function of a causal and BIBO stable linear system.

2. $W_-^{-1}(z)$ is analytic *outside* the closed unit disc $|z| \geq 1$, i.e. $W_-^{-1}(z)$ is the transfer function of a causal and BIBO stable linear system (caution with zeros on the unit circle...)

i.e. $W_-(z)$ is causal and with causal inverse. This implies that $\mathbf{y}(t)$ is a (stable) function of the past of $\mathbf{e}(t)$ and viceversa, so that

$$\mathbf{H}_t^-(\mathbf{y}) = \mathbf{H}_t^-(\mathbf{e}) \tag{2}$$

Prediction for ARMA models

PROBLEM: want to compute the linear **one-step-ahead predictor** of $\mathbf{y}(t)$ given the past values $\mathbf{y}(s)$, $s < t$. By the **Projection Theorem** this is given by:

$$\hat{\mathbf{y}}(t|t-1) := \hat{\mathbb{E}}\left[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{y})\right]$$

Theorem: $\hat{\mathbf{y}}(t|t-1)$ can be written as

$$\hat{\mathbf{y}}(t|t-1) = (W_-(z)-1)W_-^{-1}(z)\mathbf{y}(t)$$

Proof: from $\mathbf{y}(t) = W_-(z)\mathbf{e}(t)$, defining $\{w_-(k)\}_{k\in\mathbb{Z}^+} := \mathscr{Z}^{-1}[W_-(z)]$, and observing that $w_-(0) = 1$, we have that

$$\mathbf{y}(t) = \mathbf{e}(t) + \sum_{k=1}^{\infty} w_-(k)\mathbf{e}(t-k)$$

Therefore, using (2) we have:

$$\begin{aligned}\hat{\mathbf{y}}(t|t-1) &:= \hat{\mathbb{E}}[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{y})] = \hat{\mathbb{E}}[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{e})] = \sum_{k=1}^{\infty} w_-(k)\mathbf{e}(t-k) \\ &= [W_-(z)-1]\mathbf{e}(t) = [W_-(z)-1]W_-^{-1}(z)\mathbf{y}(t)\end{aligned}$$

HOMEWORK: use the argument above to compute $\mathbf{y}(t+k|t-1)$, $k = 1, 2, \ldots$.

Purely deterministic stationary random processes

$\mathbf{y}$ is a purely deterministic (p.d) process if it has zero innovation. Can be predicted exactly based on the infinite past.

Example (elementary)

$$\mathbf{y}(t) = \sum_{k=1}^{\nu} \mathbf{x}_k \cos \omega_k t + \mathbf{z}_k \sin \omega_k t \,, \qquad \mathbb{E}\,\mathbf{x}_k^2 = \mathbb{E}\,\mathbf{z}_k^2 = \sigma_k^2$$

all random variables $\{\mathbf{x}_k \,, \mathbf{z}_k \,; k = 1, 2, \ldots, \nu\}$ mutually uncorrelated.

$$\mathbf{H}(\mathbf{y}) := \operatorname{span}\{\mathbf{x}_k \,, \mathbf{z}_k \,; k = 1, 2, \ldots, \nu\} = \mathbf{H}_t^-(\mathbf{y}) = \mathbf{H}_t^+(\mathbf{y})$$

The spectral density does not exists. Formally is a sum of delta functions (spectral lines).

Wold decomposition

Theorem 1 (Wold decomposition) *Every stationary process can be decomposed uniquely as*

$$\mathbf{y}(t) = \mathbf{y}_{pnd}(t) + \mathbf{y}_{pd}(t) \quad ,$$

where $\{\mathbf{y}_{pnd}(t)\}$ *is p.n.d.,* $\{\mathbf{y}_{pd}\}$ *is p.d. and* $\{\mathbf{y}_{pnd}(t)\}$ *and* $\{\mathbf{y}_{pd}\}$ *are uncorrelated, i.e.* $\mathbb{E}\,\mathbf{y}_{pnd}(t)\mathbf{y}_{pd}^{\top}(s) = 0$, $\forall t, s \in \mathbb{Z}$, *and* subordinate *to* $\mathbf{y}$, *i.e.*

$$\mathbf{H}_t^-(\mathbf{y}_{pnd}) \subseteq \mathbf{H}_t^-(\mathbf{y}) \quad \mathbf{H}_t^-(\mathbf{y}_{pd}) = \mathbf{H}_{-\infty}(\mathbf{y}_{pd}) \subseteq \mathbf{H}_t^-(\mathbf{y})$$

The spectrum of $\mathbf{y}$ is the sum of an absolutely continuous part (spectral density) plus a singular part (spectral lines + ..). If the logarithm of the absolutely continuous part of the spectrum $\Phi_{\mathbf{y}}(e^{j\theta})$ satisfies

$$\int_0^{\pi} \log\Phi_{\mathbf{y}}(e^{j\theta})\, d\theta > -\infty$$

then $\Phi_{\mathbf{y}}(e^{j\theta})$ is also the spectrum of the p.n.d. component, i.e. $\Phi_{\mathbf{y}}(e^{j\theta}) = \Phi_{\mathbf{y}_{pnd}}(e^{j\theta})$.

(WIDE-SENSE) ERGODICITY (I)

Consider the (time invariant) functions

$$f_t(\mathbf{y}) := \mathbf{y}(t)\mathbf{y}^\top(t-\tau)$$

The (wide-sense) stationary process $\mathbf{y}$ is (wide-sense) ergodic if

$$\lim_{T\to\infty} \frac{1}{T}\sum_{1}^{T} f_t(\mathbf{y}) = \lim_{T\to\infty} \frac{1}{T}\sum_{1}^{T} \mathbf{y}(t)\mathbf{y}^\top(t-\tau) = \mathbb{E}\,\mathbf{y}(t)\mathbf{y}^\top(t-\tau) = \mathbb{E}\,f_t(\mathbf{y}) \quad w.p.1 \tag{3}$$

NOTA BENE 1: Sufficient conditions for this to hold are not very simple.

(WIDE-SENSE) ERGODICITY (II)

Consider the (time invariant) functions

$$f_t(\mathbf{y}) := \mathbf{y}(t)\mathbf{y}^\top(t-\tau)$$

The stationary process $\mathbf{y}$ is (wide-sense) ergodic if

$$\lim_{T\to\infty} \frac{1}{T} \sum_{1}^{T} f_t(\mathbf{y}) = \mathbb{E} f_t(\mathbf{y}) \quad w.p.1$$

NOTA BENE 2: There is a more general definition of Ergodicity (strict sense) which requires strict stationarity as an assumption and considers general measurable functions f, *which do not depend explicitly on time*, of the variables $\{\mathbf{y}(\tau)\,, \tau \in \mathbb{I}\}, \quad \mathbb{I} \subseteq \mathbb{Z}$.

ERGODICITY (III)

$$\lim_{T\to\infty} \frac{1}{T} \sum_{1}^{T} f_t(\mathbf{y}) = \mathbb{E}\, f_t(\mathbf{y}) \quad w.p.1$$

NOTA BENE 3: This is really a theorem (Birkhoff's ergodic theorem). The almost sure convergence is a consequence of the more general definition of ergodicity which requires that the only invariant variables of the process are deterministic constants.

DO STOCHASTIC SIGNALS EXIST?

Let $y(t)$ be a "deterministic" signal which admits the limits

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} y(t) &:= 0 \\ \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} y(t+\tau)y^{\top}(t) &:= r(\tau) \quad \tau \in \mathbb{Z} \end{aligned} \tag{4}$$

If (4) holds we say the the deterministic signal $y(t)$ is **second order stationary**.

It is a theorem (which goes back to Wiener 1930) that if such a limit exists, $r(\tau)$ is a positive definite function, i.e.

$$\mathbf{R}_n := \begin{bmatrix} r(0) & r(1) & \dots & r(n-1) \\ r(1) & r(0) & r(1) & \dots & r(n-2) \\ \vdots & & & \vdots & \\ r(n-1) & r(n-2) & \dots & & r(0) \end{bmatrix} \geq 0 \quad \forall n$$

DO STOCHASTIC SIGNALS EXIST? (II)

A positive (semi-)definite function $r(\tau)$ is called a *bona-fide* covariance function.

$$\Updownarrow$$

One can think that $y(t)$ is one sample trajectory of a stationary ergodic process $\mathbf{y}$ which admits $r(\tau)$ as covariance function and by Herglotz theorem there is a monotonically non-decreasing (on $[-\pi, \pi]$) function $F_y(e^{j\theta})$, the spectral distribution of $\mathbf{y}$, such that:

$$r(\tau) = \int_{-\pi}^{\pi} e^{j\theta\tau} dF_u(e^{j\theta}) .$$

If this spectral distribution is absolutely continuous then one recovers the more classical representation

$$r(\tau) = \int_{-\pi}^{\pi} e^{j\theta\tau} \Phi_y(e^{j\theta}) d\theta .$$

where $\Phi_y(e^{j\omega}) = \frac{dF_y(e^{j\theta})}{d\theta}$ is the (power) spectral density.

DYNAMICAL MODELS FOR SYSTEM IDENTIFICATION

From now on:

1. - Output variables (symbol $\mathbf{y}$): variables which are to be modeled
   - eXogenous variables (symbol $\mathbf{u}$) : variables we are not interested in but which influence $\mathbf{y}$

2. Data are sample trajectories of* *second order stationary stochastic processes*, i.e. we consider the equivalence class of processes having the same second order moments.

3. Additional assumptions may be needed to guarantee (wide sense) ergodicity which become relevant to study the asymptotic properties.

*This is not a limitation for the reasons discussed a few slides earlier.

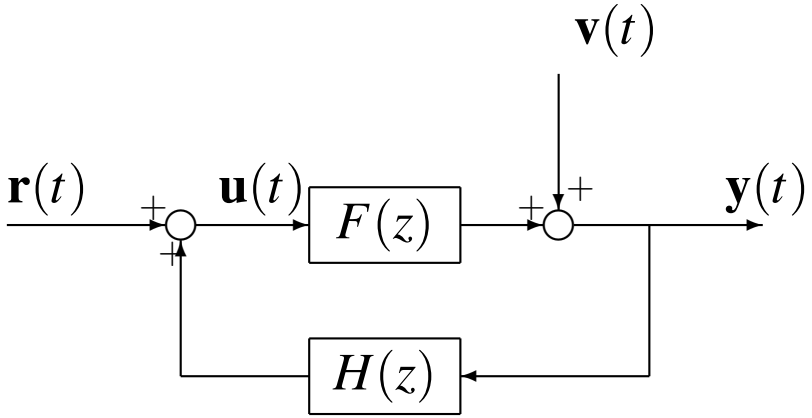
LINEAR DYNAMICAL MODELS FOR SECOND ORDER PROCESSES

Second order statistics (mean, in general assumed to be zero, and covariance function) can be described ONLY using linear models.

EXAMPLE: the Wold representation for p.n.d. processes is a linear model; it is independent of the probability distribution.

FEEDBACK MODELS, PREDICTION AND CAUSALITY

FEEDBACK MODELS AND PREDICTION



How do we find a feedback model for the p.n.d. "processes" $\mathbf{z} := [\mathbf{y}^\top, \mathbf{u}^\top]^\top$?

Can do prediction of $\mathbf{z}$ using the same procedure we used earlier for $\mathbf{y}$:

$$\begin{array}{lcl} \hat{\mathbf{y}}(t|t-1) & := & \hat{\mathbb{E}}\left[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{y}),\mathbf{H}_t^-(\mathbf{u})\right] = P_{11}(z)\mathbf{y}(t) + P_{12}(z)\mathbf{u}(t) \\ \hat{\mathbf{u}}(t|t-1) & := & \hat{\mathbb{E}}\left[\mathbf{u}(t)|\mathbf{H}_{t+1}^-(\mathbf{y}),\mathbf{H}_t^-(\mathbf{u})\right] = P_{21}(z)\mathbf{y}(t) + P_{22}(z)\mathbf{u}(t) \end{array} \tag{5}$$

where $P_{ij}(z)$ is analytic outside the (closed) unit disc and

$$P_{11}(\infty) = P_{12}(\infty) = P_{22}(\infty) = 0 \tag{6}$$

Define now the *innovations* $\mathbf{e}(t) := \mathbf{y}(t) - \hat{\mathbf{y}}(t|t-1)$ and $\mathbf{n}(t) := \mathbf{u}(t) - \hat{\mathbf{u}}(t|t-1)$ it is easy to show that $\mathbf{e}$ and $\mathbf{n}$ are completely uncorrelated.

Therefore:

$$\begin{aligned} \mathbf{y}(t) &= P_{11}(z)\mathbf{y}(t) + P_{12}(z)\mathbf{u}(t) + \mathbf{e}(t) \\ \mathbf{u}(t) &= P_{21}(z)\mathbf{y}(t) + P_{22}(z)\mathbf{u}(t) + \mathbf{n}(t) \end{aligned} \tag{7}$$

so that

$$\begin{aligned} \mathbf{y}(t) &= \frac{P_{12}(z)}{1-P_{11}(z)}\mathbf{u}(t) + \frac{1}{1-P_{11}(z)}\mathbf{e}(t) \\ \mathbf{u}(t) &= \frac{P_{21}(z)}{1-P_{22}(z)}\mathbf{y}(t) + \frac{1}{1-P_{22}(z)}\mathbf{n}(t) \end{aligned}$$

which yields the feedback scheme in the previous slide with

$$\begin{aligned} F(z) &:= \frac{P_{12}(z)}{1-P_{11}(z)} \quad \mathbf{v}(t) := \frac{1}{1-P_{11}(z)}\mathbf{e}(t) = G(z)\mathbf{e}(t) \\ H(z) &:= \frac{P_{21}(z)}{1-P_{22}(z)} \quad \mathbf{r}(t) := \frac{1}{1-P_{22}(z)}\mathbf{n}(t) = K(z)\mathbf{n}(t) \end{aligned} \tag{8}$$

where the rightmost equations define $G(z)$ and $K(z)$.

Nota Bene: From (6) and (7) we have $F(\infty) = 0$, $G(\infty) = I$, $K(\infty) = I$, $G^{-1}(z)$ and $G^{-1}(z)F(z)$ stable

INTERNALLY STABLE FEEDBACK MODELS

From Wold's representation of the joint process* $\mathbf{z} := (\mathbf{y}^\top, \mathbf{u}^\top)^\top$ we have that

$$\mathbf{H}_t^-(\mathbf{y}, \mathbf{u}) = \mathbf{H}_t^-(\mathbf{e}, \mathbf{n}) \tag{9}$$

where $\mathbf{e}, \mathbf{n}$ is the joint innovation process. Note that equation (7) can be rewritten as:

$$\begin{bmatrix} 1 - P_{11}(z) & -P_{12}(z) \\ -P_{21}(z) & 1 - P_{22}(z) \end{bmatrix} \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = [I - P(z)] \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{u}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{e}(t) \\ \mathbf{n}(t) \end{bmatrix} \tag{10}$$

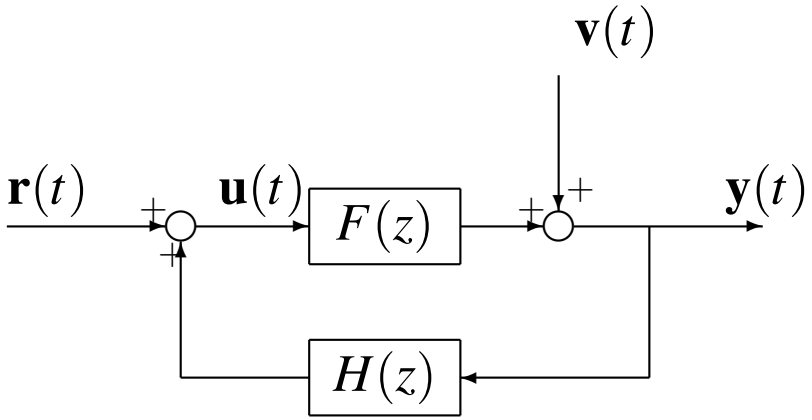
Theorem: Both $(I - P(z))$ and $(I - P(z))^{-1}$ are proper and analytic in $|z| \geq 1$

Proof: From (10) $(I - P(z))$ is the transfer function from $\mathbf{z}$ to its innovation while $(I - P(z))^{-1}$ constructs $\mathbf{z}$ from its innovation. From (9), there is a causal relation between the past spaces $\mathbf{H}_t^-(\mathbf{y}, \mathbf{u})$ and $\mathbf{H}_t^-(\mathbf{e}, \mathbf{n})$; Therefore, both $(I - P(z))$ and $(I - P(z))^{-1}$ have to be causal and stable.

*We assume $\mathbf{z}$ to be of *full rank*, see *Rozanov, 1967* and, for simplicity, we assume also $\Phi_{\mathbf{z}} > cI$ (coercive).

INTERNALLY STABLE FEEDBACK MODELS

Hence, the feedback interconnection



with

- $F(z)$, $H(z)$ uniquely defined from $P(z)$ as in (8)
- $\mathbf{r}$, $\mathbf{v}$ completely uncorrelated

is **internally stable**.

Conversely:

Theorem: If the joint process $\mathbf{z}$ is stationary and the model

$$\begin{aligned}\mathbf{y}(t) &= F(z)\mathbf{u}(t)+\mathbf{v}(t) = F(z)\mathbf{u}(t)+G(z)\mathbf{e}(t)\\ \mathbf{u}(t) &= H(z)\mathbf{y}(t)+\mathbf{r}(t)\end{aligned}$$

is given such that:

1. $F(\infty) = 0$

2. $\mathbf{e} \perp \mathbf{r}$ and $G(z)\Lambda_{\mathbf{e}}G^{\top}(1/z) = \Phi_{\mathbf{v}}(z)$, $G(\infty) = I$

3. $G^{-1}(z)F(z)$ and $G^{-1}(z)$ are analytic in $|z| \geq 1$

Then

$$\begin{aligned}\mathbf{y}(t) &= G^{-1}(z)F(z)\mathbf{y}(t)+(G(z)-1)G^{-1}(z)\mathbf{u}(t)+\mathbf{e}(t)\\ &= \hat{\mathbf{y}}(t|t-1)+\mathbf{e}(t)\end{aligned}$$

Proof: Note that from joint stationarity of $\mathbf{z}(t)$ is follows that the transfer function from $(\mathbf{e},\mathbf{r})$ to $\mathbf{z}$ is analytic in $|z| \geq 1$ so that $\mathbf{H}_t^-(\mathbf{y},\mathbf{u}) \subseteq \mathbf{H}_t^-(\mathbf{e},\mathbf{r})$. Therefore $\mathbf{e}(t) \perp \mathbf{H}_t^-(\mathbf{e},\mathbf{r}) \rightarrow \mathbf{e}(t) \perp \mathbf{H}_t^-(\mathbf{y},\mathbf{u})$. In addition, using

$$\mathbf{e}(t) = G^{-1}(z)\mathbf{y}(t) - G^{-1}(z)F(z)\mathbf{u}(t)$$

$\mathbf{y}(t)$ can be written in the form

$$\begin{aligned}
\mathbf{y}(t) &= F(z)\mathbf{u}(t) + (G(z)-1)\mathbf{e}(t) + \mathbf{e}(t) \\
&= F(z)\mathbf{u}(t) + (G(z)-1)G^{-1}(z)\left(\mathbf{y}(t) - F(z)\mathbf{u}(t)\right) + \mathbf{e}(t) \\
&= G^{-1}(z)F(z)\mathbf{y}(t) + (G(z)-1)G^{-1}(z)\mathbf{u}(t) + \mathbf{e}(t)
\end{aligned}$$

From the assumptions $G^{-1}(z)F(z)\mathbf{y}(t) + (G(z)-1)G^{-1}(z)\mathbf{u}(t) \in \mathbf{H}_t^-(\mathbf{y},\mathbf{u})$ and the fact that $\mathbf{e}(t) \perp \mathbf{H}_t^-(\mathbf{y},\mathbf{u})$ we obtain

$$\hat{y}(t|t-1) := \hat{\mathbb{E}}\left[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{y},\mathbf{u})\right] = G^{-1}(z)F(z)\mathbf{y}(t) + (G(z)-1)G^{-1}(z)\mathbf{u}(t)$$

FEEDBACK MODELS AND CAUSALITY

Problem: Understand when variables "influences each other". In which sense? is "plain correlation" the correct concept? NO!

DEFINITION: We say that $\mathbf{y}$ *does not Granger-cause* $\mathbf{u}$ (or also that there is *absence of feedback* from $\mathbf{y}$ to $\mathbf{u}$) if

$$\hat{\mathbf{u}}(t|t-1) = P_{12}(z)\mathbf{y}(t) + P_{22}(z)\mathbf{u}(t) = P_{22}(z)\mathbf{u}(t)$$

i.e. $P_{12}(z) = 0$.

FEEDBACK MODELS AND CAUSALITY

The following are equivalent conditions:

1. $\mathbf{y}$ *does not Granger-cause* $\mathbf{u}$

2. $H(z) = 0$ in the internally stable feedback model

3. $\mathbf{H}_t^+(\mathbf{u}) \perp \mathbf{H}_t^-(\mathbf{y}) | \mathbf{H}_t^-(\mathbf{u})$

4. $\hat{\mathbb{E}}[\mathbf{y}(t)|\mathbf{H}(\mathbf{u})] = \hat{\mathbb{E}}[\mathbf{y}(t)|\mathbf{H}_t^-(\mathbf{u})] = F(z)\mathbf{u}(t)$, $F(z)$ strictly causal and stable

5. $\mathbf{e} \perp \mathbf{u}$

FEEDBACK MODELS AND CAUSALITY

Can be extended to more general interconnections. Define

$$\mathbf{z}^\top(t) := [\mathbf{z}_1(t)^\top, \mathbf{z}_2(t)^\top, .., \mathbf{z}_k(t)^\top]$$

and consider the predictor

$$\hat{\mathbf{z}}_i(t|t-1) := \sum_j \mathbf{H}_{ij}(z)\mathbf{z}_j(t) + \mathbf{e}_i(t)$$

We say that $\mathbf{z}_\ell$ *does not Granger-cause* $\mathbf{z}_i$ if

$$\hat{\mathbf{z}}_i(t|t-1) = \sum_{j\neq\ell} \mathbf{H}_{ij}(z)\mathbf{z}_j(t) \quad \mathbf{H}_{i\ell}(z) = 0 \tag{11}$$

This is a "dynamic" version of static conditional orthogonality conditions. Note that, in the static Gaussian case, these conditional orthogonality corresponds to zeros in the inverse covariance matrix, see the old paper by A.P. Dempster, *Biometrics, 1972*, on the so-called *covariance selection* problem.

Note that *causality* conditions of the form (11) can be encoded in a graphical way where nodes are time series $\mathbf{z}_j$ and there is an arc from $\mathbf{z}_j$ to $\mathbf{z}_i$ if $\mathbf{H}_{ij}(z) \neq 0$.

SYSTEM IDENTIFICATION

INGREDIENTS OF THE PROBLEM

- DATA: $\{u(1), y(1), ..., u(T), y(T)\}$

- MODELS: linear/nonlinear, parametric/non parametric etc.

- CRITERIA: how do we choose a good model? how do we choose the model class and its complexity (order for linear parametric models)?

SYSTEM IDENTIFICATION

INGREDIENTS OF THE PROBLEM

A CLASSICAL PERSPECTIVE

- DATA: $\{u(1), y(1), ..., u(T), y(T)\} \Longrightarrow$ *Trajectories from a stationary and ergodic stochastic process*

- MODELS: $\Longrightarrow$ *Linear Parametric Models; We shall discuss linear non-parametric models tomorrow (see Pillonetto/De Nicolao)*

- CRITERIA: $\Longrightarrow$ *Prediction Error Methods and order selection criteria (AIC, BIC, MDL, GCV...)*

LINEAR PARAMETRIC MODELS

We shall consider linear parametric models (in innovation form) with the structure

$$\mathbf{y}(t) = F_\theta(z)\mathbf{u}(t) + G_\theta(z)\mathbf{e}(t)\,, \qquad \theta \in \Theta \subset \mathbb{R}^p \qquad (\dagger) \tag{12}$$

$F_\theta(z)$ e $G_\theta(z)$ are rational functions of fixed order. We assume that the parametrization is *regular*, i.e. continuous and differentiable as many times as needed.

We shall be interested in the "direct chain" of the feedback interconnection which represents the joint process $\mathbf{z} := (\mathbf{y}^\top, \mathbf{u}^\top)^\top$. In general one might have a collection of variables $\mathbf{z}^\top(t) := [\mathbf{z}_1(t)^\top, \mathbf{z}_2(t)^\top, .., \mathbf{z}_k(t)^\top]$ and be interested in modeling *one* (say $\mathbf{z}_i$) as a function of the others, in the form

$$\mathbf{z}_i(t) = \sum_{j \neq i} F_{ij}(z)\mathbf{z}_j(t) + G_i(z)\mathbf{e}_i(t)$$

REMARKS:

1. Several well-known model classes are contained in (12) ARMAX, ARX, OE, Box-Jenkins, ARIMA, Orthonormal Basis (Laguerre, Kautz) etc., see *Ljung, Söderström, Box-Jenkins....*

2. The assumption that the model order is fixed and known is an *unrealistic* assumption. In practice it has to be estimated from data using well known order estimation criteria AIC, BIC, MDL, AICC, GCV. Properties of estimators which follows model selection (PMSE: Post Model Selection Estimator) are not entirely trivial (see Leeb-Pötcher)

IDENTIFIABILITY ISSUES

Two important issues arise (NOT TOUCHED UPON IN THESE LECTURES)

1. *A Priori Identifiability* $\Longleftrightarrow$ the parametrization $\theta \to [F_\theta, G_\theta]$ is *injective* (globally/locally), i.e.

$$[F_{\theta_0}(z), G_{\theta_0}(z)] = [F_{\theta_2}(z), G_{\theta_2}(z)] \ \forall z \in \mathbb{C} \Leftrightarrow \theta_1 = \theta_2 \quad \text{globally/locally}$$

2. The "inputs" (or external excitations $\mathbf{r}$ and $\mathbf{v}$ in the "feedback" interconnections) are *rich enough* so that, under condition 1 the predictor can be uniquely determined, i.e.

$$\mathbb{E}\,\|\hat{y}_{\theta_1}(t|t-1) - \hat{y}_{\theta_2}(t|t-1)\|^2 = 0 \Leftrightarrow \theta_1 = \theta_2 \quad \text{globally/locally} \tag{13}$$

DEFINITION: Condition (13) is called *Identifiability*

Nota Bene: Condition 2 can be "strengthened" to an optimal experiment design problem: *how do I design the experiment so that certain properties of the system are estimated with the least uncertainty?*

IDENTIFIABILITY ISSUES: HOMEWORK

Assume there is no feedback: prove that any p.n.d. input signal $\mathbf{u}$ is "sufficiently exciting" so that a priori identifiability is necessary and sufficient for identifiability

PREDICTION ERROR METHODS (PEM)

Principle: given a (parametric??) model $M(\theta)$, $\theta \in \Theta$ (e.g. specified through F_θ and G_θ), where the domain Θ may account for constraints of various form (e.g. stability, positivity etc.) and given a sequence of input-output data

$$y^N := \{y(t)\,;\, t = 1,2,\ldots,N\}\,, \qquad u^N := \{u(t)\,;\, t = 1,2,\ldots,N\}$$

do:

1. Compute the *best* (e.g. linear minimum variance) predictor $\hat{y}_\theta(t \mid t-1)$ based on the given model $M(\theta)$ (note that $\hat{y}_\theta(t \mid t-1)$ is a *deterministic function* of the data and of θ. We shall use the symbol $\hat{\mathbf{y}}_\theta(t|t-1)$ for the same function of the random variables $\{\mathbf{y}(s), \mathbf{u}(s), s < t\}$ rather than of the sample values $\{y(s), u(s), s < t\}$.

2. Compute the *prediction errors*:

$$\varepsilon_\theta(t) := y(t) - \hat{y}_\theta(t)\,; \qquad t = 1,2,\ldots,N$$

which, similarly to the predictor, can be regarded as random quantities and denoted with bold symbols: i.e. $\boldsymbol{\varepsilon}_\theta(t)$.

For the *parametric* model class (12) the prediction error takes the form:

$$\begin{aligned} \boldsymbol{\varepsilon}_\theta(t) &= \mathbf{y}(t) - \hat{\mathbf{y}}(t \mid t-1) = \mathbf{y}(t) - G_\theta(z)^{-1}\left[F_\theta(z)\mathbf{u}(t) + (G_\theta(z)-1)\mathbf{y}(t)\right] \\ &= G_\theta(z)^{-1}\left[\mathbf{y}(t) - F_\theta(z)\mathbf{u}(t)\right] \end{aligned} \tag{14}$$

3. Minimize w.r.t. θ the **mean squared prediction error** which quantifies how well the model is able to predict the next data point:

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \varepsilon_\theta(t)^2$$

REMARK: sometimes weighted versions are considered which include forgetting factors $\beta(N,t)$,

$$V_N(\theta) := \frac{1}{N}\sum_{t=1}^{N} \beta(N,t)\varepsilon_\theta(t)^2 \qquad \beta(t,N) > 0 \tag{15}$$

to account for effect of transients (e.g. mishandling of initial conditions and/or to time slowly time-varying parameters); also clipped and/or filtered residuals are sometimes used to (i) reduce the effect of outliers and (ii) focus on specific frequency bands of interests.

Bottom line, the *minimum prediction error* estimator $\hat{\theta}_N$ is obtained solving

$$\hat{\theta}_N := \text{Arg} \min_{\theta} V_N(\theta) \tag{16}$$

WARNING: this turns out to be a non-linear, non-convex optimization problem with local minima etc.

4. An estimator of the innovation variance $\lambda^2 = \text{var}\{\mathbf{e}(t)\}$, is taken to be the *mean of the squared residuals* , i.e.

$$\hat{\lambda}_N^2 := V_N(\hat{\theta}_N) \tag{17}$$

WHY PREDICTION ERROR METHODS?

Based on “nice” statistical properties:

1. PEM estimator coincide (asymptotically in N) with the Maximum Likelihood (ML) estimator for Gaussian innovations, as such it inherits nice properties of ML:

2. Consistency: under some reasonable assumptions (if “true” model belong to model class)

$$\hat{\theta}_N \xrightarrow{P} \theta_o$$

   otherwise converges to a point of “minimum distance” from the “true” model in terms of Kullback-Leibler divergence (for Gaussian innovations)

3. Asymptotic normality (under some reasonable assumptions)

$$\sqrt{N}(\hat{\theta}_N - \theta_o) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \Sigma)$$

i.e. the normalized error $\sqrt{N}(\hat{\theta}_N - \theta_o)$ converges in law to a normal random variables with zero mean ($\sqrt{N}$-consistency) and variance Σ

4. Asymptotic efficiency (for Gaussian innovations): the asymptotic variance Σ is equal, for Gaussian innovations, to the Cramèr-Rao lower bound (theoretical lower bound on the minimum variance achievable by *any* unbiased estimator of θ_o).

CAVEATS:

The properties above are

1. Asymptotic in the number of data

2. Hold when the *model order/complexity* (say n) is known

When the "true" model order n_o is replaced with and estimator $\hat{n}$ (consistent), $\hat{\theta}_N(\hat{n})$ is called a *PMSE (post model selection estimator)*: these properties do not hold *uniformly* in the parameter space (see Leeb-Pötcher) and have unbounded (normalize) maximal risk:

$$\lim_{N\to\infty} \sup_{\theta} N\,\mathbb{E}\,\|\hat{\mathbf{y}}_{\theta}(t|t-1) - \hat{\mathbf{y}}_{\hat{\theta}_N(\hat{n})}(t|t-1)\|^2 = \infty$$

NB: similar results hold for *any* unbounded loss function $\ell(\hat{\mathbf{y}}_{\theta}(t|t-1) - \hat{\mathbf{y}}_{\hat{\theta}_N(\hat{n})}(t|t-1))$.

MODEL ORDER SELECTION CRITERIA

Typically of the form ($\theta \in \mathbb{R}^n$)

$$C_N(n) = -\frac{2}{N}\log p_{\hat{\theta}_N(n)}(y) + \frac{\alpha(N)}{N}n$$

which, in the Gaussian case reduces to

$$C_N(n) = \log\left(\frac{1}{N}\sum_{t=1}^{N}\left(y(t) - \hat{y}_{\hat{\theta}_N(n)}(t|t-1)\right)^2\right) + \frac{\alpha(N)}{N}n$$

Possible choices:

1. AIC: $\alpha(N) = 2$. It is obtained minimizing (and estimate of) the KL-divergence between the model parametrized by θ and the "true" model. It is, w.r.t. estimation of a "regression function", min-max rate optimal = *"the (worst-case) risk of this estimator when predicting the output converges at the same rate as that of an "optimal" estimator which minimizes the worst case (min-max) risk" (see next slides)*. AIC *over-estimates* the order with positive probability.

2. BIC/MDL: $\alpha(N) = \log(N)$. It derives from a "Bayesian" viewpoint where models are assigned a certain prior and the model class is estimated by maximizing the marginal posterior of the model (i.e. once the "parameters" are averaged out), *Schwartz, 1978*. It is consistent, i.e.:

$$\lim_{N \to \infty} \mathbb{P}[\hat{n} = n_o] = 1$$

FACT: no order estimator can be at the same time min-max rate optimal and consistent (Yang, 2005)

RISK and OPTIMALITY

Let $\mathbf{y}_{\theta_o}(t) := \hat{\mathbf{y}}_{\theta_o}(t|t-1) + e_o(t)$ denotes the output process when the "true" parameter is θ_o. Consider the risk (final prediction error), where expectation is taken w.r.t. the "innovation process"

$$FPE(\theta_o, N, \hat{n}) := \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left(\mathbf{y}_{\theta_o}(t) - \hat{\mathbf{y}}_{\hat{\theta}_N(\hat{n})}(t|t-1)\right)^2$$

The rule $\hat{n}$ is called *min-max rate optimal* w.r.t the parameter set Θ if $\sup_{\theta_o \in \Theta} FPE(\theta_o, N, \hat{n})$ converges at the same rate as

$$\inf_{\hat{\theta}} \sup_{\theta_o \in \Theta} \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left(\mathbf{y}_{\theta_o}(t) - \hat{\mathbf{y}}_{\hat{\theta}}(t|t-1)\right)^2$$

where $\hat{\theta}$ ranges through all measurable functions of the available data.

MINIMIZING RISK (FPE)?

Definition: **minimum Final Prediction Error**

$$\hat{n}_{FPE} := \arg\min_{n} \widehat{FPE}(\theta_o, N, n)$$

is the estimator of the order which minimizes an unbiased estimator $\widehat{FPE}$ of FPE. This criterion is known as *Final Prediction Error* because it is (an estimate of) the expected prediction error when the estimated model is used to predict new data.

Theorem:

$$\hat{n}_{FPE} = \arg\min_{n} \frac{N+n}{N-n} \hat{\sigma}^2_{ML}(n)$$

where

$$\hat{\sigma}^2_{ML}(n) := \frac{1}{N} \sum_{t=1}^{N} \left(y(t) - \hat{y}_{\hat{\theta}_N(n)}(t|t-1) \right)^2$$

is the Maximum Likelihood (under Gaussian innovations) estimator of the innovation variance σ^2.

Proof: We provide the proof under the simplifying assumption that $y_{\theta_o}(t) := \phi^\top(t)\theta_o + e_o(t)$. Let

$$\hat{\theta}_N(n) := \left(\sum_{t=1}^{N} \phi(t)\phi^\top(t) \right)^{-1} \sum_{t=1}^{N} \phi(t) y_{\theta_o}(t)$$

be the least squares estimator and let $\tilde{\theta}_N(n) := \theta_o - \hat{\theta}_N(n)$ be the error with variance

$$\mathbb{E}\, \tilde{\theta}_N(n)\tilde{\theta}_N^\top(n) = \sigma^2 \left(\sum_{t=1}^{N} \phi(t)\phi^\top(t) \right)^{-1}$$

Then

$$\begin{aligned} FPE(\theta_o,N,n) &= \tfrac{1}{N}\sum_{t=1}^{N}\mathbb{E}\left(\phi^\top(t)\tilde{\theta}_N(n)+e_o(t)\right)^2 \\ &= \tfrac{1}{N}\mathrm{Tr}\left[\sum_{t=1}^{N}\phi(t)\phi^\top(t)\,\mathbb{E}\,\tilde{\theta}_N(n)\tilde{\theta}_N^\top(n)\right]+\sigma^2 \\ &= \tfrac{1}{N}\mathrm{Tr}\left[\sum_{t=1}^{N}\phi(t)\phi^\top(t)\sigma^2\left(\sum_{t=1}^{N}\phi(t)\phi^\top(t)\right)^{-1}\right]+\sigma^2 \\ &= \sigma^2\left(1+\tfrac{n}{N}\right) \end{aligned}$$

An unbiased estimator of $FPE(\theta_o,N,n)$ is

$$\widehat{FPE}(\theta_o,N,n)=\frac{N}{N-n}\left(1+\frac{n}{N}\right)\hat{\sigma}^2_{ML}=\frac{N+n}{N-n}\hat{\sigma}^2_{ML}(n)$$

This concludes the proof.

Proposition: AIC and FPE are asymptotically equivalent

Proof: Homework!

REFERENCES

1. Ljung L. Ljung, *System Identification - Theory For the User*. Prentice Hall, 1999.

2. T. Soderstrom and P. Stoica, *System Identification*. Prentice Hall, 1989.

3. G. Box and G. Jenkins, *Time series analysis: Forecasting and control*, San Francisco: Holden-Day, 1970.

4. T. Soderstrom , *Discrete Time Stochastic Systems*. Springer , 2002.

5. Y. A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.

6. G. Picci, *Metodi Statistici per l'Identificazione dei Sistemi Lineari*. Disponibili su www.dei.unipd.it/∼picci

7. G.D. Birkhoff. *Proof of the ergodic theorem*. Proc. Nat. Acad. Sciences (USA), 17:565600, 1931.

8. C.W.J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*. Econometrica 37, 424-438, 1969.

9. J. Geweke, *Measurement of linear dependence and feedback between multiple time series*. Journal of the American Statistical Association 77, 304-313, 1980.

10. M. Gevers and B.D.O. Anderson, *On jointly stationary feedback-free stochastic processes*. IEEE Trans. Aut. Contr., 27:431–436, 1982.

11. U. Forsell and L. Ljung, *Closed loop identification revisited*. Automatica, 35:1215–1242, 1999.

12. A. Chiuso and G. Picci, *Consistency Analysis of some Closed-loop Subspace Identification Methods*, Automatica: special issue on System Identification, **41** pp. 377-391, 2005.

13. A. Chiuso and G. Pillonetto, *A Bayesian approach to sparse dynamic network identification*, Automatica, in press, 2012.

14. H. Akaike, *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 19:716–723, 1974.

15. G. E. Schwarz, *Estimating the dimension of a model*. Annals of Statistics 6 (2): 461464, 1978.

16. H. Leeb and B. M. Pötscher, *Model Selection and Inference : Facts and Fiction.* Econometric Theory, 21, 2159, 2005.

17. H. Leeb and B. M. Pötscher, *Sparse estimators and the oracle property, or the return of Hodges' estimator*. Journal of Econometrics, 142(1):201 – 211, 2008.

18. Y. Yang, *Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation*. Biometrika,92(4): 937-950, 2005.

Acknowledgments

A special thank goes to Giorgio Picci for providing material for these slides and for stimulating discussions on the subject