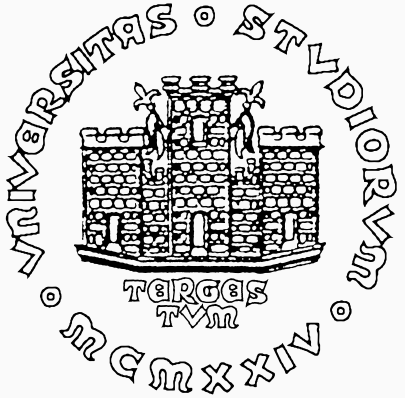
Systems Dynamics

Course ID: 267MI – Fall 2020

Thomas Parisini
Gianfranco Fenu

University of Trieste
Department of Engineering and Architecture



267MI –Fall 2020

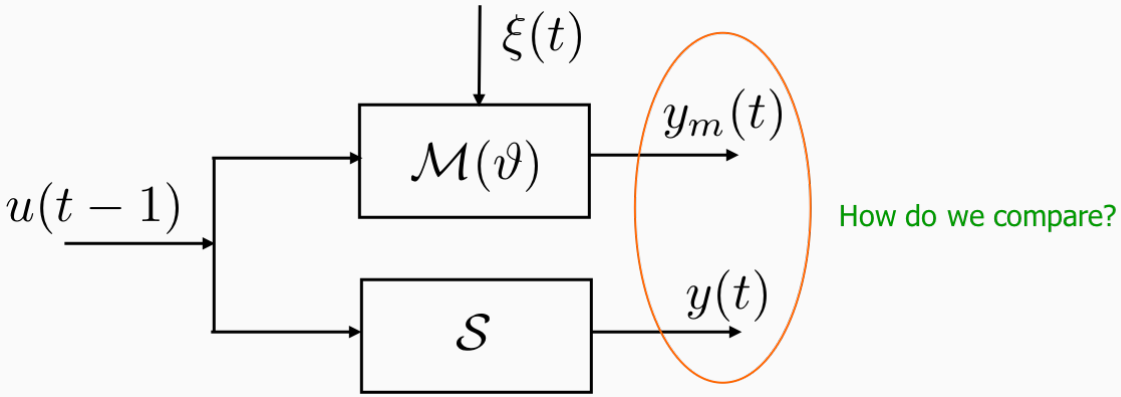
Lecture 11
Identification Based on Prediction Error Minimization (PEM)

Lecture 11: Table of Contents

Identification based on Prediction Error Minimization

Identification based on Prediction Error Minimization

- Consider the models class $\mathcal{M} = \{\mathcal{M}(\vartheta) : \vartheta \in \Theta\}$ of a **given complexity**.
- We want to determine the **best model** in the class $\mathcal{M}$, that is, the **best vector** $\bar{\vartheta} \in \Theta$ such that $\mathcal{M}(\bar{\vartheta})$ provides the best "interpretation" of the observed data.
- However, it is of customary importance to define in a precise way **how to compare** the true system (of which we observe the accessible data) with the model to be identified.
- One option could be to consider the scheme:



Identification based on Prediction Error Minimization (cont.)

- For given input variables $u(t)$ (if present) we could try to compare $y_m(t)$ with $y(t)$ trying to make $y_m(t)$ similar to $y(t)$ "in a suitable sense".
- However $\mathcal{M}(\vartheta)$ is a stochastic model and hence $y_m(t)$ is a random variable whereas $y(t)$ is a known numerical sequence.

A Trivial Approach

Let us compare $E[y_m(t)]$ with $y(t)$ (these quantities are both deterministic and hence comparable):

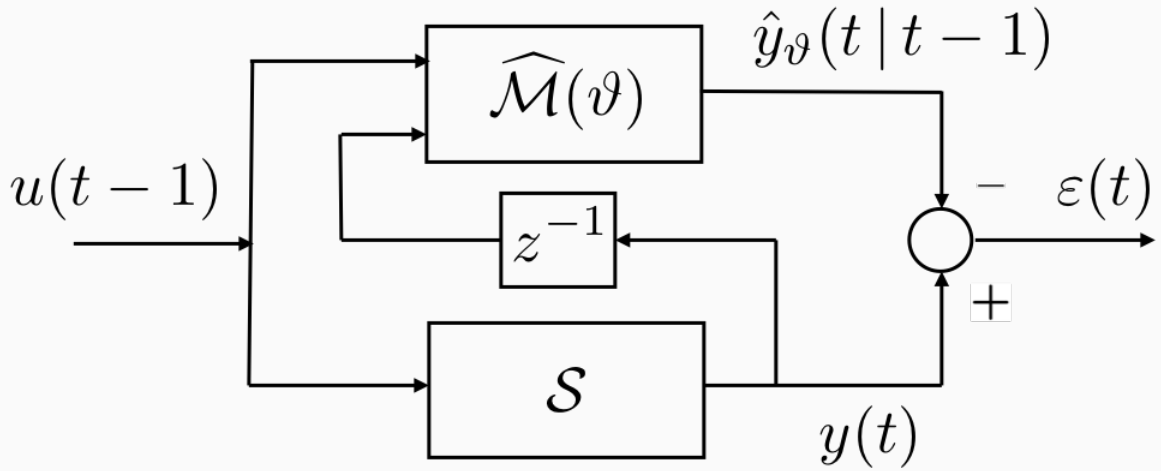
$$
\begin{aligned}
& A(z)\, y_m(t) = B(z)\, u(t-1) + C(z)\, \xi(t) \\
\Longrightarrow \quad & A(z)\, E[y_m(t)] = B(z)\, u(t-1) + C(z)\, E[\xi(t)] \\
\Longrightarrow \quad & E[y_m(t)] = \frac{B(z)}{A(z)}\, u(t-1)
\end{aligned}
$$

However, doing so, **the dependence on polynomial $C(z)$ would disappear** thus making it impossible to identity the stochastic part of the model.

Identification based on Prediction Error Minimization (cont.)

Predictive Approach to Systems Identification

- Given a class of models $\mathcal{M} = \{\mathcal{M}(\vartheta) : \vartheta \in \Theta\}$ we consider the corresponding class of models in prediction form (predictors for short) $\widehat{\mathcal{M}} = \left\{\widehat{\mathcal{M}}(\vartheta) : \vartheta \in \hat{\Theta}\right\}$
- Predictors are useful: $\hat{y}_\vartheta(t \mid t-1)$ is given by a **deterministic law** using past values of $y(\cdot)$ and of $u(\cdot)$ and hence the comparison is possible and well-posed.
- Then, the (very important) conceptual scheme is:



Identification based on Prediction Error Minimization (cont.)

Predictive Approach to Systems Identification

- The input to the predictor is made of the measurable variables $y(t-1)$ and $u(t-1)$; $\hat{y}_\vartheta(t\,|\,t-1)$ is generated using these **known** inputs (the subscript ϑ is enhanced to highlight the dependence on the vector of **unknown parameters**)
- From the **comparison** between $y(t)$ and $\hat{y}_\vartheta(t\,|\,t-1)$ we obtain the prediction error

$$\varepsilon_\vartheta(t) = y(t) - \hat{y}_\vartheta(t\,|\,t-1)$$

- The prediction error is exploited to determine the vector $\bar{\vartheta}$ for which the model $\mathcal{M}(\bar{\vartheta})$ associated with the predictor $\widehat{\mathcal{M}}(\bar{\vartheta})$ **"interprets" the observed data in the best way possible**.
- The vector $\bar{\vartheta}$ (hence the best model) is determined through the minimisation of a cost function taking on the form

$$J(\vartheta) = \frac{1}{N}\sum_{t=\tau}^{N}[\varepsilon_\vartheta(t)]^2 \quad \text{for a suitable} \quad \tau \geq 1$$

Identification based on Prediction Error Minimization

Remarks

Remarks

- **Conceptually** we identify the model $\mathcal{M}(\vartheta)$ but, from an **operational** viewpoint, we use the predictor $\widehat{\mathcal{M}}(\vartheta)$
- The minimization of the cost function on the pre-selected time-window is, of course, important, but it is very important as well that the prediction error is a stochastic process with characteristics that are as close as possible to the ones of a **white process**
- It is important to emphasize again that the identification procedure minimizing the prediction error (MPE) makes it possible to identify stochastic models by means of a **deterministic procedure**.

Asymptotic Theory for PEM Identification Methods

Asymptotic Theory for PEM Identification Methods

- Consider

$$\hat{\vartheta}_N = \arg \min_{\vartheta} J_N(\vartheta)$$

  where N is the size of the time-window and we suppose that the data $y(\cdot)$ and $u(\cdot)$ are stochastic processes; hence $\hat{\vartheta}_N$ is a random variable for any given value of N

- Assume that $y(\cdot)$ and $u(\cdot)$ are stationary ($\mathcal{S}$ stable) and assume also that $\widehat{\mathcal{M}}(\vartheta)$ is stable. Then:

$$\varepsilon_\vartheta(t) = y(t) - \hat{y}_\vartheta(t \,|\, t-1) \quad \text{is stationary}$$

  Hence:

$$J_N(\vartheta) = \frac{1}{N} \sum_{t=\tau}^{N} [\varepsilon_\vartheta(t)]^2 \quad \longrightarrow \quad E\left\{[\varepsilon_\vartheta(t)]^2\right\} \quad \text{for} \quad N \to \infty$$

Asymptotic Theory for PEM Identification Methods (cont.)

- Let $\bar{J}(\vartheta) = E\left\{[\varepsilon_\vartheta(t)]^2\right\}$. Clearly $\bar{J}(\vartheta)$ **does not depend on** t **because of the stationarity**
- $\bar{J}(\vartheta)$ (which coincides with **variance of the prediction error**) is a **deterministic function** of ϑ, that is, it does **not** depend on the result of the random experiment).

Fundamental Question

Does

$$\lim_{N\to\infty} J_N(\vartheta) = \bar{J}(\vartheta)$$

imply that

$$\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta^*$$

where $\vartheta^* \in \Delta$ with Δ being the **set of minima of** $\bar{J}(\vartheta)$, that is:

$$\Delta = \left\{\bar{\vartheta} :\ \bar{J}(\bar{\vartheta}) \leq \bar{J}(\vartheta),\ \forall\, \vartheta \in \Theta\right\}$$

Asymptotic Theory for PEM Identification Methods (cont.)

Asymptotic Theorem 1

Suppose that:

- $y(\cdot)$ and $u(\cdot)$ stationary stochastic processes
- $u(\cdot)$ independent from $\xi(\cdot)$
- $\xi(\cdot)$ white process
- $\Theta \subset \mathbb{R}^q$, Θ compact
- $\widehat{\mathcal{M}}(\vartheta)$ stable $\forall\, \vartheta \in \Theta$
- $\widehat{\mathcal{M}}(\vartheta) \in \mathcal{C}^3$ with respect to ϑ

Then:

$$\lim_{N \to \infty} \hat{\vartheta}_N \in \Delta \quad \text{a.s.}$$

Almost-sure asymptotic convergence (probability 1)
to the set of optimal parameters

Asymptotic Theory for PEM Identification Methods (cont.)

Asymptotic Theorem 2

Suppose that:

- Same assumptions of Asymptotic Theorem 1 hold
- Δ contains only one point
- $\exists\, \vartheta^\circ : \mathcal{S} = \mathcal{M}(\vartheta^\circ)$ (the true system belongs to the class in which we are looking for the best model)

Then:

- $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta^\circ \quad \text{a.s.}$
- The **innovation** $e(t) = y(t) - \hat{y}_{\vartheta^\circ}(t \,|\, t-1)$ is a **white process**

Almost-sure asymptotic convergence (probability 1)
to the true parametrization

Asymptotic Theory for PEM Identification Methods (cont.)

Sketch of the proof

- Consider $\varepsilon_\vartheta(t) = y(t) - \hat{y}_\vartheta(t \,|\, t-1)$ for a generic ϑ. Hence:

$$\begin{aligned}\varepsilon_\vartheta(t) &= y(t) - \hat{y}_{\vartheta^\circ}(t \,|\, t-1) + \hat{y}_{\vartheta^\circ}(t \,|\, t-1) - \hat{y}_\vartheta(t \,|\, t-1) \\ &= e(t) + [\hat{y}_{\vartheta^\circ}(t \,|\, t-1) - \hat{y}_\vartheta(t \,|\, t-1)]\end{aligned}$$

  where $e(t)$ is called **innovation** and represents the prediction error in case of use of the optimal predictor.
- From the optimality, it follows that $e(t)$ is uncorrelated from the past values of $y(\cdot)$ and $u(\cdot)$, while both $\hat{y}_{\vartheta^\circ}(t \,|\, t-1)$ and $\hat{y}_\vartheta(t \,|\, t-1)$ depend on such past values.
- Then, $e(t)$ and $[\hat{y}_{\vartheta^\circ}(t \,|\, t-1) - \hat{y}_\vartheta(t \,|\, t-1)]$ are uncorrelated and hence

$$\begin{aligned}\text{var}\,[\varepsilon_\vartheta(t)] &= \text{var}\,[e(t)] + \text{var}\,[\hat{y}_{\vartheta^\circ}(t \,|\, t-1) - \hat{y}_\vartheta(t \,|\, t-1)] \\ &\Longrightarrow \bar{J}(\vartheta) \geq \bar{J}(\vartheta^\circ)\end{aligned}$$

  Thus concluding that ϑ° is a minimum of $\bar{J}(\vartheta)$ and it is unique by assumption

Asymptotic Theory for PEM Identification Methods

Remarks

Remarks

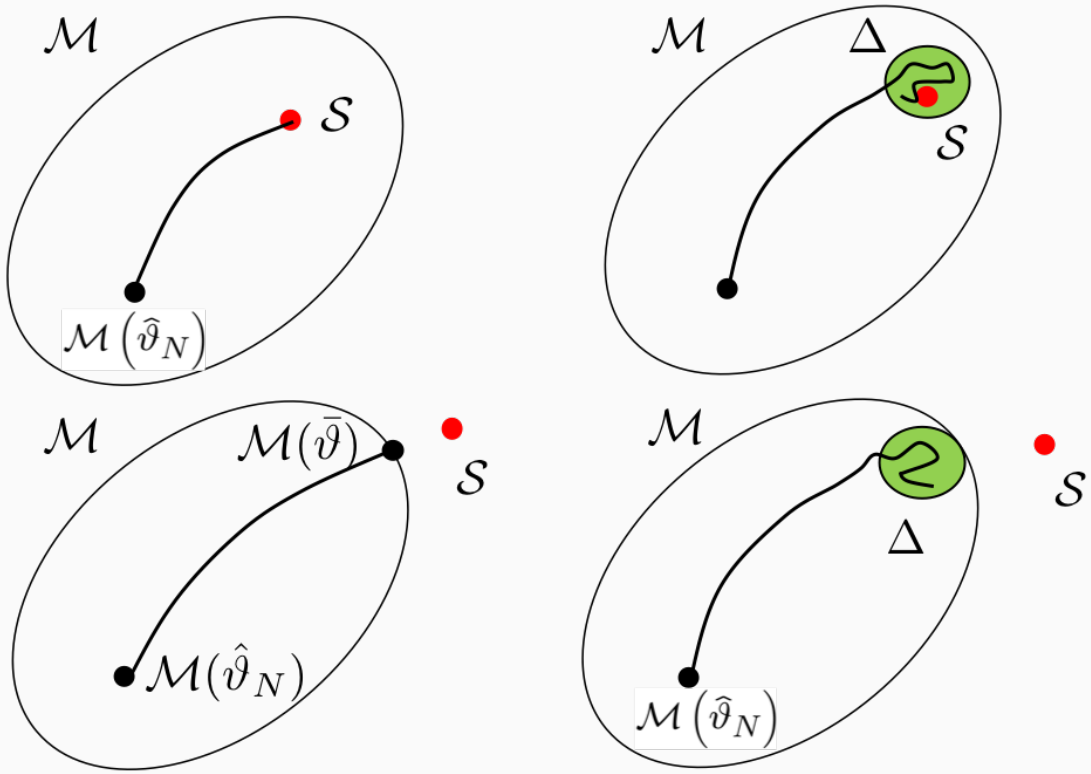
- The assumption $\mathcal{S} = \mathcal{M}(\vartheta^\circ)$ is an equality between transfer functions and ϑ° is called **true parametrization**.
- Let's keep the assumption $\exists\, \vartheta^\circ : \mathcal{S} = \mathcal{M}(\vartheta^\circ)$, but consider the case in which Δ is made of more than one point.
- In this case $\lim_{N\to\infty} \hat{\vartheta}_N \in \Delta \quad \text{a.s.}$ and it may happen that $\lim_{N\to\infty} \hat{\vartheta}_N = \vartheta^* \neq \vartheta^\circ \quad \text{a.s.}$, but it may also happen that $\hat{\vartheta}_N$ **does not converge, "cycling repeatedly" on points belonging to** Δ
- It is worth noting that, except in the case where ϑ° has a specific **physical meaning**, the convergence to $\vartheta^* \neq \vartheta^\circ$ is not necessarily a bad result. In fact, if $\bar{J}(\vartheta^*) = \bar{J}(\vartheta^\circ)$, it follows that $\mathcal{M}(\vartheta^\circ)$ and $\mathcal{M}(\vartheta^*)$ are **equivalent from the predictive point of view**.

Remarks (cont.)

- Let us now remove the assumption $\exists\, \vartheta^\circ : \mathcal{S} = \mathcal{M}(\vartheta^\circ)$, that is, consider the case $\nexists\, \vartheta^\circ : \mathcal{S} = \mathcal{M}(\vartheta^\circ)$; however, let's keep the assumption for which Δ is made of a single point: $\Delta = \{\bar{\vartheta}\}$
- The fact $\mathcal{S} \neq \mathcal{M}(\vartheta), \forall\, \vartheta \in \Theta$ means that $\mathcal{S}$ cannot be fully characterized in terms of models in the class $\mathcal{M}$:
  - Θ is not large enough
  - The order of model $\mathcal{M}(\vartheta)$ is not large enough
  - The class of models $\mathcal{M}$ is not rich enough
  - .....

Remarks (cont.)

- Thanks to asymptotic Theorem 1: $\lim_{N\to\infty} \hat{\vartheta}_N = \bar{\vartheta}$ a.s.
  Clearly $\bar{J}(\bar{\vartheta}) > \mathrm{var}\,[e(t)]$ but $\mathcal{M}(\bar{\vartheta})$ is anyway the model in the class $\mathcal{M}$ providing the **best approximation** of $\mathcal{S}$ in the sense of minimum prediction error
- Therefore, we have four possible cases:



Asymptotic Theory for PEM Identification Methods

Important Example

Important Example

Consider the process (true system):

$$\mathcal{S}: \quad y(t) = e(t) + \frac{1}{2}\, e(t-1)\,, \quad e(\cdot) \sim WN(0, \lambda^2)$$

and consider the class of models AR(1):

$$\mathcal{M}(\vartheta): \quad y(t) = a\, y(t-1) + \xi(t)$$

The corresponding class of models in prediction form is:

$$\widehat{\mathcal{M}}(\vartheta): \quad \hat{y}(t \,|\, t-1) = a\, y(t-1)$$

Hence:

$$\mathcal{S} \neq \mathcal{M}(\vartheta)$$

and we want to determine the set Δ of minima of $\bar{J}(\vartheta)$

Important Example (cont.)

We have:

$$\begin{aligned}
\bar{J}(\vartheta) &= E\left\{[\varepsilon_\vartheta(t)]^2\right\} = E\left\{[y(t) - \hat{y}(t\,|\,t-1)]^2\right\} \\
&= E\left\{\left[e(t) + \frac{1}{2}\,e(t-1) - a\,y(t-1)\right]^2\right\} \\
&= E\left\{\left[e(t) + \frac{1}{2}\,e(t-1) - a\,e(t-1) - \frac{1}{2}ae(t-2)\right]^2\right\} \\
&= E\left\{\left[e(t) + \left(\frac{1}{2} - a\right)\,e(t-1) - \frac{1}{2}ae(t-2)\right]^2\right\}
\end{aligned}$$

But $e(t), e(t-1), e(t-2)$ are uncorrelated. Hence:

$$\begin{aligned}
\bar{J}(\vartheta) &= \operatorname{var}[e(t)] + \left(\frac{1}{2} - a\right)^2 \operatorname{var}[e(t-1)] + \frac{1}{4}a^2 \operatorname{var}[e(t-2)] \\
&= \left(\frac{5}{4} + \frac{5}{4}a^2 - a\right) \operatorname{var}[e(t)]
\end{aligned}$$

Important Example (cont.)

Thus:

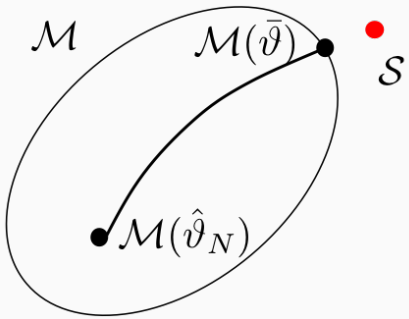
$$\frac{d\bar{J}}{d\vartheta} = \frac{d\bar{J}}{da} = \left(\frac{5}{2}a - 1\right) \operatorname{var}\left[e(t)\right] \quad \Longrightarrow \quad \bar{a} = \frac{2}{5}$$

Then:

$$\widehat{\mathcal{M}}(\bar{\vartheta}): \quad \hat{y}(t \,|\, t-1) = \frac{2}{5}\, y(t-1)$$

$$\Longrightarrow \quad \mathcal{M}(\bar{\vartheta}): \quad y(t) = \frac{2}{5}\, y(t-1) + \xi(t)$$

$\mathcal{M}(\bar{\vartheta})$ is the **best model in the class** $\mathcal{M} = AR(1)$ approximating the true system (recall that $\mathcal{S} \neq AR(1)$)



Important Example (cont.)

- The predictor is stable and this is consistent with the stationarity of $\mathcal{S}$
- The prediction error is given by:

$$\begin{aligned}\varepsilon_{\bar{\vartheta}}(t) &= y(t) - \hat{y}_{\bar{\vartheta}}(t \,|\, t-1) = y(t) - \hat{y}_{\bar{a}}(t \,|\, t-1) \\ &= e(t) + \frac{1}{2}\, e(t-1) - \frac{2}{5}\, y(t-1) \\ &= e(t) + \frac{1}{2}\, e(t-1) - \frac{2}{5}\left[e(t-1) + \frac{1}{2} e(t-2)\right] \\ &= e(t) + \frac{1}{10}\, e(t-1) - \frac{1}{5}\, e(t-2)\end{aligned}$$

Clearly, the process $\varepsilon_{\bar{\vartheta}}(t)$ **is not white** and this is not surprising because $\mathcal{S} \neq AR(1)$.

Identifiability

Identifiability

- To analyze the identifiability of a given system $\mathcal{S}$ through a given class of models $\mathcal{M}$ means to analyze the **cardinality of the set** Δ
- In general:

$$\left.\begin{array}{l}\textbf{Experimental conditions}\\[2ex]\textbf{Structure of the class of models}\end{array}\right\} \Longrightarrow \textbf{Cardinality of } \Delta$$

Identifiability: Experimental Conditions

Even if $\mathcal{S} \in \mathcal{M}$, this **does not imply** that $\Delta = \{\bar{\vartheta}\}$

Trivial Example

$$\mathcal{M}(\vartheta): \quad y(t) = G(z)\, u(t-1) + W(z)\, \xi(t)$$

- Suppose that the experimental conditions under which the identification procedure is conducted are such that $u(t) = 0\,,\ \forall t$
- Then, **any choice** of $G(z)$ would be admissible and hence the cardinality of the set Δ would be **infinite**

Identifiability

Remarks

Remarks

- If the experimental conditions could be constructed in such a way that $u(t)$ is **sufficiently rich**, then it is possible to guarantee that Δ contains a single element.
- On the other hand, if the experimental conditions cannot be constructed as above, it is then necessary to **reduce the models' complexity** (that it, the number of unknown parameters) thus limiting the identification procedure only to the actually identifiable parts.

Structure of the family of models to be identified

Assume that $\mathcal{S} \in \mathcal{M}$ but also that the chosen family has a **complexity larger than the one of the true system**

Example $\mathcal{S} = ARMAX(1,1,1), \quad \mathcal{M} = ARMAX(2,2,2)$

Clearly, irrespective of the experimental conditions, Δ will be necessarily made of an infinite number of elements because $\mathcal{S}$ can be described by an infinite number of models belonging to the family in which there are **common factors**.

It is important to guarantee that the family $\mathcal{M}$ **is not over-parametrised**

Concluding Remarks on Identifiability

- **Structural identifiability**:
  Uniqueness of the approximating model belonging to the pre-selected family of models (choice of model complexity)
- **Experimental identifiability**: Uniqueness of the vector of parameters with respect to the information conveyed by the observed data

To guarantee the uniqueness of the minimum it is necessary that both conditions above are satisfied.

Asymptotic Evaluation of Estimates' Uncertainty

Asymptotic Evaluation of Estimates' Uncertainty

- Beyond **point-wise convergence**, it is important to analyze the **uncertainty** of the estimates as well.
- Let $\psi(t,\vartheta) = -\left[\frac{\partial}{\partial\vartheta}\varepsilon_\vartheta(t)\right]^\top, \quad \bar{R}(\vartheta) = E\left[\psi(t,\vartheta)\,\psi(t,\vartheta)^\top\right]$

Theorem

- Same assumptions of Asymptotic Theorem 1 hold
- Δ contains only one point
- $\exists\,\vartheta^\circ : \mathcal{S} = \mathcal{M}(\vartheta^\circ)$

Then:

- $\lim_{N\to\infty} \sqrt{N}\left(\hat{\vartheta}_N - \vartheta^\circ\right) \sim G(0,\bar{P})$
- $\bar{P} = \text{var}\left[\varepsilon_{\vartheta^\circ}(t)\right]\,\bar{R}(\vartheta^\circ)^{-1}$

Hence, for N sufficiently large, the variance of the estimator is

$$\frac{1}{N}\text{var}\left[\varepsilon_{\vartheta^\circ}(t)\right]\,\bar{R}(\vartheta^\circ)^{-1}$$

Final Example

Important Example

Consider the process (true system):

$$\mathcal{S}: \quad y(t) = a^{\circ}\, y(t-1) + e(t)\,, \quad |a^{\circ}| < 1,\, e(\cdot) \sim WN(0, \lambda^2)$$

and consider the family of models AR(1):

$$\mathcal{M}(\vartheta): \quad y(t) = a\, y(t-1) + \xi(t)$$

The corresponding family of models in prediction form is:

$$\widehat{\mathcal{M}}(\vartheta): \quad \hat{y}(t\,|\,t-1) = a\, y(t-1)$$

Then, one has: $J_N(\vartheta) = \dfrac{1}{N} \displaystyle\sum_{t=1}^{N} \varepsilon(t)^2\,.$

But $\varepsilon(t) = y(t) - \hat{y}(t\,|\,t-1) = y(t) - ay(t-1)$ and hence:

$$J_N(\vartheta) = \frac{1}{N} \sum_{t=1}^{N} \left[y(t) - ay(t-1)\right]^2$$

Important Example (cont.)

Thus:

$$\frac{d}{da} J_N(\vartheta) = -\frac{2}{N} \sum_{t=1}^{N} \left[y(t) - a y(t-1) \right] y(t-1)$$

and hence

$$\frac{d}{da} J_N(\vartheta) = 0 \implies \hat{a}_N = \frac{\frac{1}{N} \sum_{t=1}^{N} \left[y(t)\, y(t-1) \right]}{\frac{1}{N} \sum_{t=1}^{N} \left[y(t-1) \right]^2} \implies \lim_{N \to \infty} \hat{a}_N = \frac{\gamma(1)}{\gamma(0)}$$

On the other hand:

$$\begin{aligned} y(t)\, y(t-1) &= a^\circ\, y(t-1)^2 + e(t)\, y(t-1) \\ \implies E\left[y(t)\, y(t-1) \right] &= a^\circ\, E\left[y(t-1)^2 \right] + E\left[e(t)\, y(t-1) \right] \\ \implies \gamma(1) &= a^\circ\, \gamma(0) \\ \implies \lim_{N \to \infty} \hat{a}_N &= a^\circ \end{aligned}$$

Important Example (cont.)

Concerning the **uncertainty of the estimate**:

$$\psi(t, a^\circ) = -\left.\frac{d}{da}\varepsilon_\vartheta(t)\right|_{\vartheta=a^\circ} = -\left.\frac{d}{da}\left[y(t) - ay(t-1)\right]\right|_{a=a^\circ} = y(t-1)$$

from which we have

$$\bar{R}(a^\circ) = E\left[\psi(t, a^\circ)\,\psi(t, a^\circ)^\top\right] = E\left[\psi(t, a^\circ)^2\right] = E\left[y(t-1)^2\right] = \gamma(0)$$

and then, for N sufficiently large, the **variance of the estimator** is

$$\mathrm{var}\left[\hat{a}_N\right] = \frac{1}{N}\mathrm{var}\left[\varepsilon_{a^\circ}(t)\right]\bar{R}(a^\circ)^{-1} = \frac{1}{N}\frac{\mathrm{var}\left[e(t)\right]}{\gamma(0)} = \frac{1}{N}\frac{\lambda^2}{\gamma(0)}$$

Therefore, the estimate's uncertainty is inversely proportional to the "signal-to-noise ratio" and asymptotically vanishes for $N \to \infty$

267MI –Fall 2020

Lecture 11
Identification Based on Prediction Error Minimization (PEM)

END