MODEL STRUCTURE SELECTION AND MODEL VALIDATION

- **Reminder: Hypothesis Tests**

- **Is the Model Large Enough?** **Correlation Tests**

- **Is the Model Too Complex?** **- Cross-Validation**

  - FPE/AIC Criterion

  - BIC/MDL Criterion

PRELIMINARIES: HYPOTHESIS TESTS

Introduction

A *statistical hypothesis* is an assumption about the value(s) of one or more parameters of a statistical model

Let X be a random variable with p.d.f. $P\{x;\theta\}$ where $\theta \in \Theta$. From measurements $\mathbf{X} = [X_1 \quad \cdots \quad X_N]^T$, we want to compare the (*null*) *hypothesis* $H : \theta \in \Theta_0$ against the *alternative* (*hypothesis*) $K : \theta \in \Theta_1$. We need to decide whether to *reject* or *not reject* H

If $\Theta_0 = \{\theta_0\}$ $(\Theta_1 = \{\theta_1\})$, the hypothesis (or alternative) is *simple*; otherwise, it is *composite*

Problem: Does the experimental evidence support the rejection of the null hypothesis?

PRELIMINARIES: HYPOTHESIS TESTS (CONT.)

General Concepts

If H is true, $P\{\mathbf{X} \in D_c\} \approx 0$ (D_c: *critical region* of the test, and D_c^C: *region of acceptance*)

The test is: Reject H iff $\mathbf{X} \in D_c$

To check if $\mathbf{X} \in D_c$, we usually employ a *test statistic*, $T(\mathbf{X})$, such that

$$\mathbf{X} \in D_c \quad \Leftrightarrow \quad T(\mathbf{X}) \geq c \equiv \textit{critical value of the test}$$

PRELIMINARIES: HYPOTHESIS TESTS (CONT.)

Types of Errors:

	H **is true**	H **is false**
H **is not rejected**	✓	Type II Error
H **is rejected**	Type I Error	✓

Type I Errors:

$$\forall \theta_0 \in \Theta_0 : \; P\{X \in D_c \mid \theta = \theta_0\} \leq \alpha \equiv (significance)\, level\ of\ the\ test$$

$$\inf \alpha \equiv size\ of\ the\ test$$

Type II Errors:

$$P\{X \notin D_c \mid \theta = \theta'\} \equiv \beta(\theta') \equiv probability\ of\ Type\ II\ error \quad (for\ \theta' \in \Theta_1)$$

$$P(\theta') \equiv 1 - \beta(\theta') \equiv power\ of\ the\ test$$

In general a test is designed for a given size, and we want it to be as powerful as possible.

Typically: $\alpha = 0,05$ or $\alpha = 0,01$

PRELIMINARIES: HYPOTHESIS TESTS (CONT.)

Hypothesis Test v/s Confidence Intervals

If we know that

$$P\{\theta_1(\mathbf{X}) < \theta < \theta_2(\mathbf{X})\} = \gamma$$

then, assuming $H : \Theta = \{\theta_0\}$ we have that

$$P\{\theta_1(\mathbf{X}) < \theta_0 < \theta_2(\mathbf{X}) \mid H\} = \gamma$$

This gives the following test of size $\alpha = 1 - \gamma$:

$$\text{Reject } H \text{ iff } \theta_0 \notin (\theta_1(\mathbf{X}), \theta_2(\mathbf{X}))$$

IS THE MODEL SET LARGE ENOUGH?

To check if the model set is large enough to include (the main features of) the true system, we use statistical tests to verify some of the model assumptions, e.g.,

1. $\{\varepsilon_t\}$ is zero-mean white noise

2. $\{\varepsilon_t\}$ has a symmetric distribution

3. $\{\varepsilon_t\}$ is independent of past inputs

4. $\{\varepsilon_t\}$ is independent of all inputs

IS THE MODEL SET LARGE ENOUGH? (CONT.)

Autocorrelation Test: *Is* $\{\varepsilon_t\}$ *is zero-mean white noise?*

Let H_0: "$\{\varepsilon_t\}$ is zero mean white noise". Under H_0, it can be shown that

$$\hat{r}_0^{\varepsilon} = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^2, \qquad r := \begin{bmatrix} \hat{r}_1^{\varepsilon} \\ \vdots \\ \hat{r}_m^{\varepsilon} \end{bmatrix} := \frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix} \varepsilon_{t-1} \\ \vdots \\ \varepsilon_{t-m} \end{bmatrix}\varepsilon_t$$

satisfy

$$N\frac{r^T r}{(r_0^{\varepsilon})^2} \xrightarrow[N\to\infty]{d} \chi^2(m), \qquad \sqrt{N}\frac{\hat{r}_\tau^{\varepsilon}}{\hat{r}_0^{\varepsilon}} \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1)$$

Hence, one test of size α for H_0 is

$$\boxed{\text{reject } H_0 \text{ iff } \; N\frac{r^T r}{(r_0^{\varepsilon})^2} > \chi_\alpha^2(m)}$$

IS THE MODEL SET LARGE ENOUGH? (CONT.)

Cross-correlation Test: *Is* $\{\varepsilon_t\}$ *independent of* $\{u_t\}$*?*

Under H_0: "$\{\varepsilon_t\}$ is independent of $\{u_t\}$" (and assuming $\{\varepsilon_t\}$ is zero-mean white noise),

$$\hat{r}_0^{\varepsilon} = \frac{1}{N}\sum_{t=1}^{N}\varepsilon_t^2, \qquad r := \frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix} u_{t-\bar{\tau}-1} \\ \vdots \\ u_{t-\bar{\tau}-m} \end{bmatrix}\varepsilon_t, \qquad \hat{R}_u := \frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix} u_{t-1} \\ \vdots \\ u_{t-m} \end{bmatrix}\begin{bmatrix} u_{t-1} & \cdots & u_{t-m} \end{bmatrix}$$

(where $u_t = 0$ for $t \le 0$, and $\bar{\tau} \in \mathbb{Z}$ is given) satisfy

$$N r^T [\hat{r}_0^{\varepsilon} \hat{R}_u]^{-1} r \xrightarrow[N\to\infty]{d} \chi^2(m)$$

This statistic can be used as in the previous slide to obtain a test for H_0

IS THE MODEL SET LARGE ENOUGH? (CONT.)

Sign-change Test: *Has* $\{\varepsilon_t\}$ *a symmetric distribution?*

Let

$$\delta_t := \begin{cases} 1, & \text{if } \varepsilon_t \varepsilon_{t+1} < 0 \\ 0, & \text{if } \varepsilon_t \varepsilon_{t+1} > 0, \end{cases} \qquad \overline{x}_N := \sum_{t=1}^{N-1} \delta_t$$

Then, under H_0: "$\{\varepsilon_t\}$ has a symmetric distribution" (and assuming $\{\varepsilon_t\}$ is zero-mean white noise), it holds that

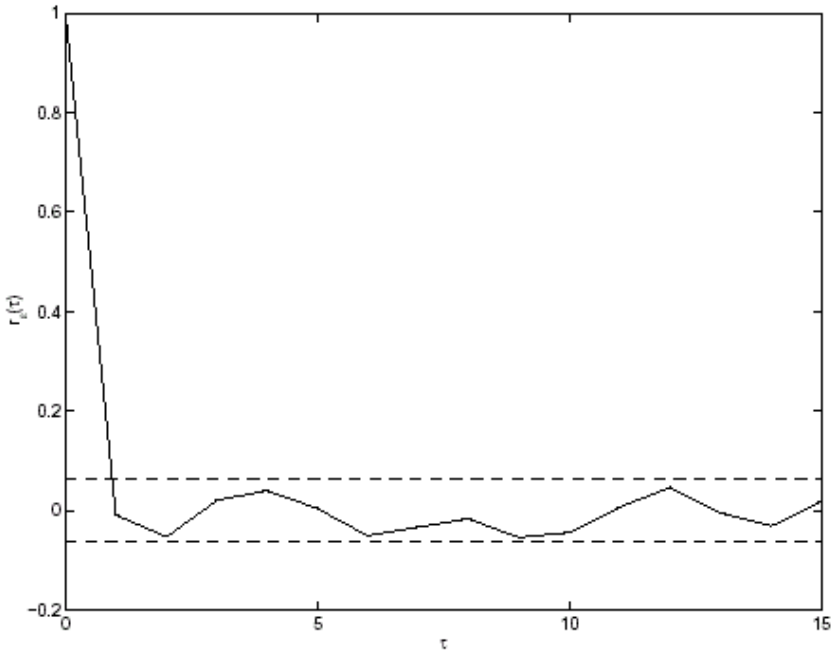
$$\frac{\overline{x}_N - N/2}{\sqrt{N}/2} \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1)$$

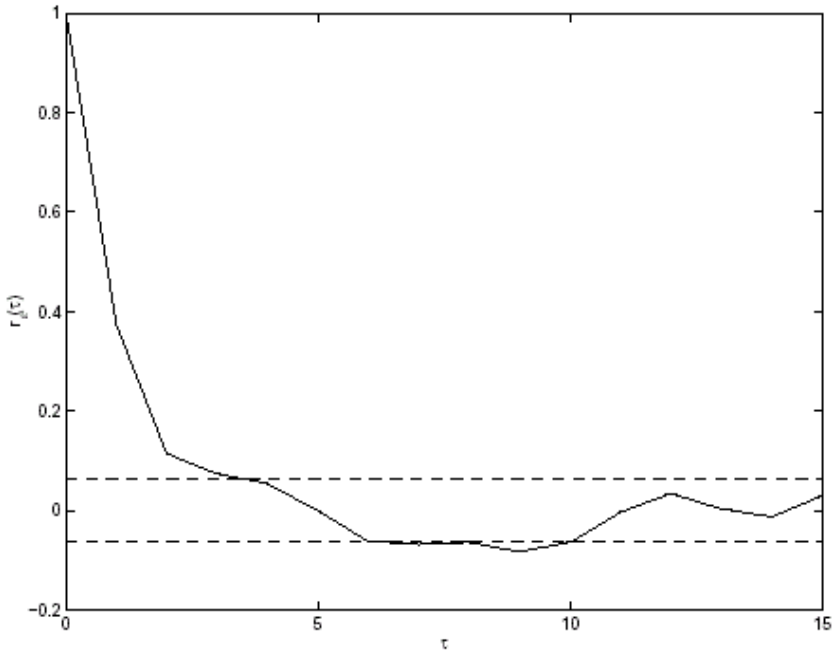
Hence, e.g., a test of size 0.05 for H_0 is

$$\boxed{\text{reject } H_0 \text{ iff } \left| \frac{\overline{x}_N - N/2}{\sqrt{N}/2} \right| > 1.96}$$

IS THE MODEL SET LARGE ENOUGH? (CONT.)

Autocorrelation plots





Left: White $\{\varepsilon_t\}$ Right: Correlated $\{\varepsilon_t\}$

IS THE MODEL TOO COMPLEX?

Comparing the PEM costs of different models does not work, since the cost function is a non-decreasing function of the number of parameters (for nested structures)

$\Rightarrow$ **Overfitting:** The model has been fitted to a particular realization of the data

Solution: *Cross-Validation*

The problem can be detected by splitting the data into:

- Estimation data (Z_e^N): Used for fitting the model
- Validation data (Z_v^N): Used for testing the estimated model

Mathematically, then, the idea is to compare:

$$W_N = V_N(\hat{\theta}_N(Z_e^N), Z_v^N)$$

FPE/AIC CRITERIA

Final Prediction Error (FPE) Criterion:

In order to avoid splitting the data, W_N can be estimated (from Z_e^N) as:

$$2V_N(\hat{\theta}_N(Z_e^N), Z_v^N) \approx 2V_N(\theta_0, Z_v^N) + 2V_N'(\theta_0, Z_v^N)(\hat{\theta}_N - \theta_0) + (\hat{\theta}_N - \theta_0)^T V_N''(\theta_0, Z_v^N)(\hat{\theta}_N - \theta_0)$$
$$\approx \lambda_0 + 2\bar{V}'(\theta_0)(\hat{\theta}_N - \theta_0) + \mathrm{tr}[(\hat{\theta}_N - \theta_0)(\hat{\theta}_N - \theta_0)^T \bar{V}''(\theta_0)]$$

Since $\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, \lambda_0 \bar{V}''(\theta_0))$, we have that

$$\boxed{\bar{E}\{W_N\} = \bar{E}\{V_N(\hat{\theta}_N(Z_e^N), Z_v^N)\} \approx \frac{\lambda_0}{2}\left(1 + \frac{n}{N}\right)}$$

FPE/AIC CRITERIA (CONT.)

To estimate λ_0, notice that

$$2V_N(\hat{\theta}_N(Z_e^N), Z_e^N) \approx \lambda_0 - 2V_N'(\hat{\theta}_N(Z_e^N), Z_e^N)(\hat{\theta}_N - \theta_0) - (\hat{\theta}_0 - \theta_N)^T V_N''(\hat{\theta}_N(Z_e^N), Z_e^N)(\hat{\theta}_0 - \theta_N)$$
$$\approx \lambda_0 - (\hat{\theta}_0 - \theta_N)^T \bar{V}''(\theta_0)(\hat{\theta}_0 - \theta_N)$$

hence,

$$\bar{E}\{V_N(\hat{\theta}_N(Z_e^N), Z_e^N)\} \approx \frac{\lambda_0}{2} - \frac{1}{2}\operatorname{tr}\bar{E}\{(\hat{\theta}_0 - \theta_N)(\hat{\theta}_0 - \theta_N)^T \bar{V}''(\theta_0)\} \approx \frac{\lambda_0}{2}\left(1 - \frac{n}{N}\right)$$

Therefore, we have that $\boxed{V_N(\hat{\theta}_N(Z_e^N), Z_e^N) \approx \frac{\lambda_0}{2}\left(1 - \frac{n}{N}\right)}$

In conclusion: $\boxed{FPE = W_N \approx V_N(\hat{\theta}_N(Z_e^N), Z_e^N)\frac{1 + n/N}{1 - n/N}}$ *FPE Criterion*

Rule: *Choose the model with smallest FPE*

FPE/AIC CRITERIA (CONT.)

Akaike Information Criterion (AIC):
Akaike proposed as cost function the log-likelihood

$$V_N(\theta, Z^N) = -\frac{1}{N}\sum_{t=1}^{N} \ln P\{y_t;\theta\}, \qquad Z^N = \{y_1, \ldots, y_N\}$$

which, as $N \to \infty$, becomes (almost) the Kullback-Liebler information

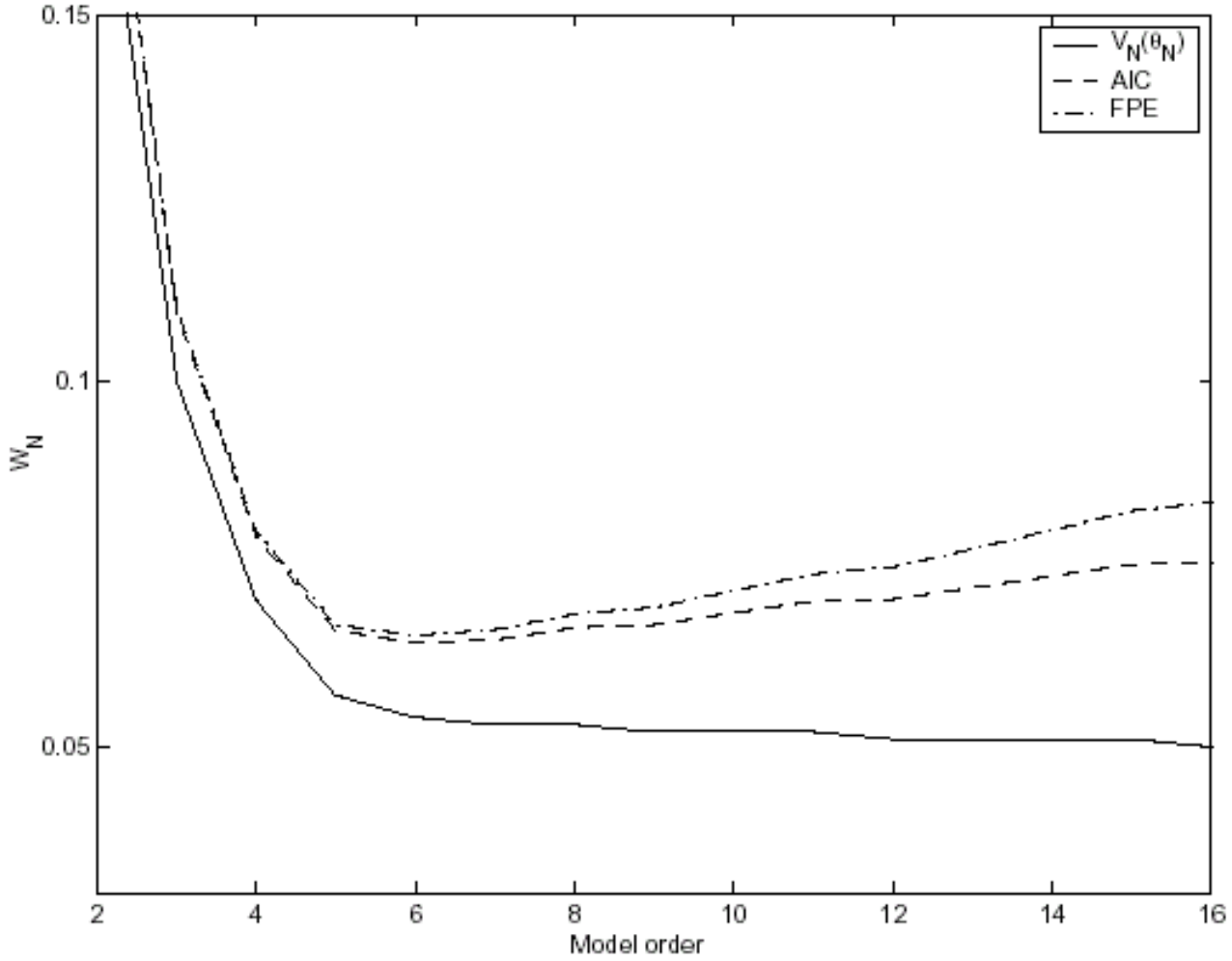
$$\bar{V}(\theta) = \bar{E}\{\ln P(y_t;\theta)\} = \int P_0(y) \ln P(y;\theta) dy$$

For finite N, this cost function gives the ML estimator. Thus, we obtain

$$\boxed{AIC = 2N\, E\{V_N(\hat{\theta}_N(Z_e^N), Z_v^N)\} \approx 2N V_N(\hat{\theta}_N(Z_e^N), Z_e^N) + 2n = -2l(\hat{\theta}_N) + 2n}$$

Rule: *Choose the model with smallest AIC*

FPE/AIC CRITERIA (CONT.)



BIC CRITERION

Consider different candidate model structures $\mathcal{M}_k$ for the data $Y^N := \{y_1, \ldots, y_N\}$. We want to select the one which most probably generated Y^N (in a Bayesian sense). Then

$$\begin{aligned}
\mathcal{M}_k^* &= \arg\max_k P\{\mathcal{M}_k \mid Y^N\} \\
&= \arg\max_k \frac{P\{Y^N \mid \mathcal{M}_k\} P\{\mathcal{M}_k\}}{\sum_i P\{Y^N \mid \mathcal{M}_i\} P\{\mathcal{M}_i\}} \\
&= \arg\max_k P\{Y^N \mid \mathcal{M}_k\} P\{\mathcal{M}_k\} \\
&= \arg\max_k P\{\mathcal{M}_k\} \int P\{Y^N \mid \theta^k, \mathcal{M}_k\} P\{\theta^k \mid \mathcal{M}_k\} d\theta^k
\end{aligned}$$

$P\{\mathcal{M}_k\}$ and $P\{\theta^k \mid \mathcal{M}_k\}$ are *prior* distributions, and $\theta^k \in \mathbb{R}^{n_k}$ is the parameter of $\mathcal{M}_k$. If we choose $P\{\mathcal{M}_k\} = \text{constant}$,

$$\mathcal{M}_k^* = \arg\max_k \int P\{Y^N \mid \theta^k, \mathcal{M}_k\} P\{\theta^k \mid \mathcal{M}_k\} d\theta^k$$

BIC CRITERION (CONT.)

To obtain a criterion independent of the choice of $P\{\theta^k \mid \mathcal{M}_k\}$, consider N large, so

$$\begin{aligned}
&\int P\{Y^N \mid \theta^k, \mathcal{M}_k\} P\{\theta^k \mid \mathcal{M}_k\} d\theta^k \\
&\quad = \int \exp[\ln P\{Y^N \mid \theta^k, \mathcal{M}_k\}] P\{\theta^k \mid \mathcal{M}_k\} d\theta^k \\
&\quad \approx \int \exp[\ln P\{Y^N \mid \hat{\theta}^k_{ML}, \mathcal{M}_k\} - (\theta^k - \theta^k_{ML})^T J^k (\theta^k - \theta^k_{ML})] P\{\theta^k \mid \mathcal{M}_k\} d\theta^k \\
&\quad = \int e^{-(\theta^k - \theta^k_{ML})^T J^k (\theta^k - \theta^k_{ML})} d\theta^k \, P\{Y^N \mid \hat{\theta}^k_{ML}, \mathcal{M}_k\} P\{\hat{\theta}^k_{ML} \mid \mathcal{M}_k\} \\
&\quad = (2\pi)^{n_k/2} \left|\det J^k\right|^{-1/2} P\{Y^N \mid \hat{\theta}^k_{ML}, \mathcal{M}_k\} P\{\hat{\theta}^k_{ML} \mid \mathcal{M}_k\}
\end{aligned}$$

where $\hat{\theta}^k_{ML}$ is the MLE of θ^k (for the model structure $\mathcal{M}_k$), and

$$J^k = -\left. \frac{\partial^2 \ln P\{Y^N \mid \theta^k, \mathcal{M}_k\}}{\partial \theta^k \partial (\theta^k)^T} \right|_{\theta^k = \hat{\theta}^k_{ML}}$$

BIC CRITERION (CONT.)

Hence,

$$\mathcal{M}_k^* \approx \arg\min_k \left[-2\ln P\{Y^N \mid \hat{\theta}_{ML}^k, \mathcal{M}_k\} + n_k \ln 2\pi + \ln\left|\det J^k\right| - 2\ln P\{\hat{\theta}_{ML}^k \mid \mathcal{M}_k\} \right]$$

$$\approx \arg\min_k \left[-2\ln P\{Y^N \mid \hat{\theta}_{ML}^k, \mathcal{M}_k\} + \ln\left|\det J^k\right| \right]$$

Furthermore,

$$\ln\left|\det J^k\right| = \ln\left| \det\left(-\left. \frac{\partial^2 \ln P\{Y^N \mid \theta^k, \mathcal{M}_k\}}{\partial\theta^k \partial(\theta^k)^T} \right|_{\theta^k = \hat{\theta}_{ML}^k} \right) \right|$$

$$= \ln\left| N^{n_k} \det\left(-\frac{1}{N} \left. \frac{\partial^2 \ln P\{Y^N \mid \theta^k, \mathcal{M}_k\}}{\partial\theta^k \partial(\theta^k)^T} \right|_{\theta^k = \hat{\theta}_{ML}^k} \right) \right|$$

$$= n_k \ln N + O(1)$$

BIC CRITERION (CONT.)

Summarizing:

$$\boxed{\mathcal{M}_k^* \approx \arg\min_k \left[-2 \ln P\{Y^N \mid \hat{\theta}_{ML}^k, \mathcal{M}_k\} + n_k \ln N \right]} \qquad \textit{BIC/MDL Criterion}$$

Discussion

AIC and BIC are the most commonly used criteria for model selection.
There is no consensus on which one to use. However:

- AIC tends to give the *best* model among the candidates (according to the cost function used), but it may give models of higher order than the true one

- BIC asymptotically finds the *true* model, if it is available among the candidate $\mathcal{M}_k$