Model Order Selection in System Identification: New and Old Techniques

Student : Giulia Prando

Supervisor : Prof. Alessandro Chiuso

Co-Supervisors : Prof. Tianshi Chen,
Prof. Lennart Ljung
Prof. Niels Kjølstad Poulsen

Academic Year 2012-2013

Università degli Studi di Padova
Dipartimento di Ingegneria dell'Informazione
Via Gradenigo, 6/B, 35131 Padova Italy
Phone +39 0498277600, Fax +39 0498277699
www.dei.unipd.it

To Bruno and Luciana

Abstract

Model order selection has always represented an important and difficult problem, both in system identification and statistics; for these reasons, it has been widely studied in literature. This thesis faces the problem in a system identification perspective, with the aim of providing a quite extensive study of classical and innovative techniques, which are adopted for model order selection. Among the classical methods, cross-validation, information criteria, the F-test and the statistical tests on the residuals are considered. Newly introduced techniques are also evaluated, such as the so-called PUMS criterion (Parsimonious Unfalsified Model Structure Selection), the kernel-based estimation and its connection with the prediction error method approach (PEM). A theoretical description of these methods is provided and accompanied by an experimental analysis, which exploits a versatile data bank, containing both systems and data sets. The order selection methods are not evaluated according to their ability to determine the true order of a system, but to select a complexity which leads to a good reproduction of the input-output properties (impulse response) of the true system. Two combinations of the considered order selection techniques are also introduced and the results based on the data bank prove that the simultaneous adoption of two methods reduces the risk of wrong order choices. Particular attention is also reserved to the tuning of the significance level to be adopted in the order selection criteria based on statistical tests.

Preface

This thesis was prepared at the department of Control Systems at Linköping University in fulfilment of the requirements for acquiring an M.Sc. in Mathematical Modelling and Computation at the Technical University of Denmark.

The thesis deals with the issue of model order selection in system identification. Both classical and new techniques adopted for this purpose are considered. Two combinations of these methods are also introduced.

The thesis consists of a theoretical description of various model order selection techniques, equipped with an experimental analysis on a particular data bank.

Linköping, 20-August-2013

Giulia Prando

Student: Giulia Prando
Supervisor: Prof. Alessandro Chiuso
Co-Supervisors: Prof. Tianshi Chen,
Prof. Lennart Ljung,
Prof. Niels Kjølstad Poulsen

Acknowledgements

This thesis represents the completion of a path which started two years ago. The last two years have constituted a fundamental step in my life, for the experiences I have done, for the people I have met and for my academic education.

Among the people I have met during this period or who have shared with me this journey, some deserve a special thanks for having contributed to make these two years so important for me.

First of all, a particular thanks go to my two supervisors at Linköping University, Prof. Lennart Ljung and Prof. Tianshi Chen, who have received me at their department and who have followed with passion my thesis work (reading and commenting my thesis even during their holidays!). I have really appreciated the willingness showed by both of them: by Lennart, who has put his experience to use in my thesis work and by Tianshi, with whom I have had many interesting discussions and who has given me various incentives to improve my work. I would like to thank both of them also for the help they gave me in solving some practical issues that I encountered because of my strange "status" at Linköping University.
For this reason, I want also to thank all the Professors and the PhD students of the Automatic Control Division at Linköping University, who have kindly greeted me within their group. A particular mention goes to the Division Chief, Prof. Svante Gunnarsson, who took upon himself in order to remedy a difficult situation that I faced.

I would like also to thank Prof. Alessandro Chiuso, who helped me to plan my thesis work and who allowed me to get in contact with Prof. Lennart Ljung and Prof. Tianshi Chen.

A thanks goes also to Prof. Niels Kjølstad Poulsen, who helped me to fulfill the administrative requirements at DTU and who kindly organized my defence at Linköping University.

Many other people have scarred my experience abroad in the last two years: the list of people that I would like to thank would be too long, but I can't avoid to cite Basilio, who has been a great companion in this adventure, both in everyday life matters and in the study experience. Our discussions at DTU library have represented for me a great resource for sharing knowledge, impressions and learnings. Moreover, our long study nights spent at DTU and our incredible dinners will be hard to forget!

My experience abroad has been made possible thanks also to the contribution of my parents. A big thanks goes to them and my family who, beyond financially supporting me, have always made me feel their affection and support, even if far away from me.

Last but not least, a special thanks goes to my boyfriend Nicola, who accepted without reserves my decision to live abroad for two years. During this period I could always count on his fundamental moral support, without forgetting also the important help that he gave me in solving many practical matters.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Model order selection has always been a challenging issue in statistical studies based on data mining. This discipline aims at finding a model able to describe a specific set of data: this model can then be used to extrapolate new information from the given data but the most interesting use is its application on new sets of data for prediction of new information.
The complexity of the model used to describe the analyzed data is of crucial importance for these techniques, since both too simple and too complex models have their disadvantages. On the one hand, a simple model is easier to estimate and to handle with, but it could not be able to completely extrapolate the features of the data. On the other hand, a complex model requires a large computational effort and a large amount of data for its estimation (this issue is known as *curse of dimensionality*) but it would probably be able to explicate the data very well. However, this ability is not always beneficial, especially when the model is applied on new data, different from the ones used for estimation. In this case complex models could be affected by overfitting, i.e. they could find difficulties in the explanation of new data, since they are too adherent to their estimation data (namely to the specific noise realization present in the data): they don't have the so-called generalization ability.

In the control system field the issue of model order selection is present in connection to system identification practices, which rely on the estimation of a mathematical model for a dynamical system, starting from experimental input-output data. This issue is relevant for parametric system identification methods, which employ a finite-dimensional parameter vector in the search for the best description of the system. Such techniques require the choice of the model type (linear or non linear, polynomial or state-space model, etc.), of the model order (i.e. the number of parameters describing the system) and of the model parametrization (i.e. the formulation of the model as a differentiable function of the parameter vector, with a stable gradient). These choices can be done according to:

- a priori considerations, which are independent from the particular set of data used;
- a preliminary data analysis, which can help in the determination of the model order and also in the choice between the use of a linear or a non linear model;
- a comparison among various model structures, which relies on the estimation of different types of models and on the comparison based on a pre-defined fit function;
- the validation of the estimated model, which uses the original estimation data to evaluate how well the model is able to describe their features, i.e. how much the data obtained from the estimated system agree with the estimation data.

Model structure determination within the system identification field has been widely treated in literature; more detailed discussions can be found in [8, Ch. 16], [13, Ch.11], [1], [6].

1.1 Problem statement

System identification

Consider a linear single input-single output system $\mathcal{S}$ described as

$$y(t) = G_0(q)u(t) + H_0(q)e(t) \tag{1.1}$$

where

$$G_0(q) = \sum_{k=1}^{\infty} g(k)q^{-k}, \qquad H_0(q) = 1 + \sum_{k=1}^{\infty} h(k)q^{-k}$$

are the transfer functions of the system and $e(t)$ is white noise with variance σ^2. q is the shift operator, such that $u(t-1) = q^{-1}u(t)$.

Given a set of input-output data $Z^N = \{u(1), y(1), u(2), y(2), ..., u(N), y(N)\}$, system identification aims at estimating $G_0(q)$ and $H_0(q)$ as well as possible. In other words, indicating the two estimates with $\widehat{G}(q)$ and $\widehat{H}(q)$, the identification procedure should maximize the functions $\mathcal{F}(G_0, \widehat{G})$ and $\mathcal{F}(H_0, \widehat{H})$, defined in (3.1).

When parametric methods are used for model estimation, the transfer functions to be estimated are defined as functions of a parameter vector $\theta \in D_{\mathcal{M}} \subset \mathbb{R}^{d_{\mathcal{M}}}$, i.e. $G(q, \theta)$ and $H(q, \theta)$. Thus, the identification procedure reduces to the determination of the value $\widehat{\theta}_N$, for which $\widehat{G}(q) = G(q, \widehat{\theta}_N)$ and $\widehat{H}(q) = H(q, \widehat{\theta}_N)$ are closest to $G_0(q)$ and $H_0(q)$.

In the identification field a system description in terms of prediction is generally preferred to the one given in (1.1); keeping the parametric approach, the k-step ahead predictor is such defined:

$$\widehat{y}(t|t-k) = W_u(q, \theta)u(t) + W_y(q, \theta)y(t) \tag{1.2}$$

where $W_u(q, \theta)$ and $W_y(q, \theta)$ represent the predictor transfer functions. In particular, the one-step ahead predictor is mainly exploited:

$$\widehat{y}(t|t-1) = H^{-1}(q, \theta)G(q, \theta)u(t) + [1 - H^{-1}(q, \theta)]y(t) \tag{1.3}$$

Indeed, the most common methods adopted to determine $\widehat{\theta}_N$ are based on the minimization of the prediction errors and for this reason are known as *prediction-error methods* (PEM). According to PEM, $\widehat{\theta}_N$ is determined as the minimizer of the function

$$V_N(\theta, Z^N) = \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon_F(t, \theta)) \tag{1.4}$$

i.e.,

$$\widehat{\theta}_N(Z^N) = \arg \min_{\theta \in D_{\mathcal{M}}} V_N(\theta, Z^N) \tag{1.5}$$

where $l(\cdot)$ is a norm function and

$$\varepsilon_F(t, \theta) = L(q)\varepsilon(t, \theta), \quad 1 \leqslant t \leqslant N \tag{1.6}$$

is a filtered version of the prediction error at sample t, when the one-step ahead prediction is adopted:

$$\begin{aligned} \varepsilon(t, \theta) &= y(t) - \widehat{y}(t|t-1) && (1.7) \\ &= H^{-1}(q, \theta)\left[y(t) - G(q, \theta)u(t)\right] && (1.8) \end{aligned}$$

Model structure definition

When no physical information is given about the system to be identified, a set of so-called "black-box" model structures is defined, i.e. flexible descriptions that can be suitable for a large variety of systems. Formally, a model structure $\mathcal{M}$ is defined as a differentiable mapping from the open subset $D_{\mathcal{M}}$ of $\mathbb{R}^{d_{\mathcal{M}}}$ to a model set $\Xi_{\mathcal{M}}$,

$$\begin{aligned} \mathcal{M} : D_{\mathcal{M}} &\rightarrow \Xi_{\mathcal{M}} \\ \theta &\mapsto \mathcal{M}(\theta) = \left[\begin{array}{cc} W_u(q,\theta) & W_y(q,\theta) \end{array} \right] \end{aligned} \tag{1.9}$$

with the constraint that the filter

$$\Psi(q,\theta) = \left[\begin{array}{cc} \frac{d}{d\theta}W_u(q,\theta) & \frac{d}{d\theta}W_y(q,\theta) \end{array} \right] \tag{1.10}$$

exists and is stable for $\theta \in D_{\mathcal{M}}$. The estimation of θ based on N measurement data, $\hat{\theta}_N$, gives rise to a specific model $m = \mathcal{M}(\hat{\theta}_N)$.

Typical examples of linear model structures are transfer-function and state-space models. Transfer-function models fall into the general definition given by

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \tag{1.11}$$

where

$$\begin{aligned} A(q) &= 1 + a_1 q^{-1} + ... + a_{n_a} q^{-n_a} && (1.12) \\ B(q) &= b_{n_k} q^{-n_k} + ... + b_{n_k+n_b-1} q^{-n_k-n_b+1} && (1.13) \\ C(q) &= 1 + c_1 q^{-1} + ... + c_{n_c} q^{-n_c} && (1.14) \\ D(q) &= 1 + d_1 q^{-1} + ... + d_{n_d} q^{-n_d} && (1.15) \\ F(q) &= 1 + f_1 q^{-1} + ... + f_{n_f} q^{-n_f} && (1.16) \end{aligned}$$

with n_k being the delay contained in the dynamics from u to y. The most common specifications of the general formulation (1.11) are:

ARMAX models - where $D(q) = F(q) = 1$, such that

$$G(q,\theta) = \frac{B(q)}{A(q)}, \quad H(q,\theta) = \frac{C(q)}{A(q)} \tag{1.17}$$

ARX models - where $C(q) = D(q) = F(q) = 1$, such that

$$G(q,\theta) = \frac{B(q)}{A(q)}, \quad H(q,\theta) = \frac{1}{A(q)} \tag{1.18}$$

OE models - where $A(q) = C(q) = D(q) = 1$, such that

$$G(q,\theta) = \frac{B(q)}{F(q)}, \quad H(q,\theta) = 1 \tag{1.19}$$

FIR models - where $A(q) = C(q) = D(q) = F(q) = 1$, such that

$$G(q,\theta) = B(q), \quad H(q,\theta) = 1 \tag{1.20}$$

For both transfer-function and state-space models a direct parametrization exists, i.e. a formulation in terms of a parameter vector θ can be defined. More precisely, for transfer function model structures θ contains the polynomial coefficients, while for state-space models θ includes the elements of the involved matrices. Furthermore, some transfer-function model structures, such as ARX and FIR, allow the formulation of the one-step ahead predictor $\hat{y}(t|\theta) = \hat{y}(t|t-1)$ as a linear regression [8, Ch. 4], i.e. as a scalar product between a known data vector $\varphi(t)$ and the parameter vector θ:

$$\hat{y}(t|\theta) = \varphi(t)^T\theta \tag{1.21}$$

where x^T denotes the transpose of the vector x.
State-space models also allow a definition in terms of linear regression [8, p. 208]. When $L(q) = 1$ in (1.6) and $l = \frac{1}{2}\varepsilon^2$ in (1.4) and in view of (1.21), the loss function (1.4) can be rewritten as

$$V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{2}\left[y(t) - \varphi(t)^T\theta\right]^2 \tag{1.22}$$

Thus, the minimization problem (1.5) becomes a least-squares problem, which admits an analytic solution, given by (assuming that the inverse exists)

$$\hat{\theta}_N^{LS} = \arg\min_{\theta\in D_{\mathcal{M}}} V_N(\theta, Z^N) = \left[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi(t)^T\right]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) \tag{1.23}$$

However, the one-step ahead predictor for a general transfer-function model can only be expressed as a pseudo-linear regression, i.e.

$$\hat{y}(t|\theta) = \varphi(t,\theta)^T\theta \tag{1.24}$$

with the regressors depending on the parameter vector itself. In this case (1.5) is a non-convex problem, for which solutions that are actually local minima can be found.

Model structure selection

As previously illustrated, in a system identification problem a set of model structures is first defined; among them, the optimal one should then be selected. This choice includes three steps, which can be done at different stages of the identification procedure:

1. The choice of the type of model set, i.e. whether a non-linear or a linear model has to be adopted; in the latter case, a further choice between input-output, state-space models, etc. should be done.
2. The determination of the model order, i.e. of the length $d_{\mathcal{M}}$ of the parameter vector θ ($\dim \theta = d_{\mathcal{M}}$), from which the order of the estimated model depends.
3. The choice of the model parametrization, i.e. the selection of a model structure, whose range equals the chosen model set.

The present project is dedicated to the second point, i.e. to model order selection. The focus is on methods based on the comparison of different model structures and on the validation of the obtained models. In particular, the classical methods used for this purpose, such as cross-validation, information criteria and various statistical tests will be tested and compared on data coming from four data sets with specific characteristics. New techniques will be also evaluated on those datasets: they range from kernel-based estimation, which circumvents the order selection problem thanks to regularization, to statistical tests performed on noiseless simulated data coming from an estimated high-order model.

The aim of the project is to provide an analysis of the classical order selection techniques used in system identification and to illustrate also newly introduced methods. A practical perspective is mainly adopted, since a detailed investigation of experimental results is drawn.

1.2 Report structure

The next chapters of the report are organized as follows:

- Chapter 2 explains how the classical model order selection procedures work and the concepts on which they are based.

- Chapter 3 illustrates the data bank used for the simulations described in the successive chapters; three fit functions are also introduced in order to assess the quality of the estimated models;

- Chapter 4 is dedicated to the experimental comparison of the methods introduced in Chapter 2, when they have to discriminate among OE models with different complexities.

- Chapter 5 describes the so-called kernel-based estimation, which is directly related to the regularized estimation. A combination with the classical PEM procedures is also illustrated. A theoretical description is first given, followed by the analysis of the experimental results achieved on the data sets introduced in Chapter 3.

- Chapter 6 illustrates a new order selection method, called PUMS, which exploits the "parsimony" principle and a statistical test appropriately defined. After a theoretical description of the method, it is applied on the data sets introduced in Chapter 3.

- Chapter 7 summarizes the main results observed in Chapters 4, 5 and 6.

Chapter 2

Classical model order selection techniques: Theoretical description

This chapter is dedicated to the description of the classical model order selection methods which will be tested and compared in the following chapters.

The procedures here described can be divided into two categories, namely:

Model validation methods - They evaluate the ability of an estimated model in the description of the estimation data, usually called estimation data.

Comparison methods - They compare models with different complexities by means of specific criteria, selecting the model giving the best value of the criterion used.

Among the first type of procedures, the project will consider residual analysis testing the whiteness of residuals and their independence from past input data. The comparison methods here evaluated are cross-validation, FPE, AIC, BIC and other information criteria, and the F-test performed on two models with different order.

A detailed explanation of the techniques listed above will be provided in the following sections with reference to the identification of a single input-single output system.

2.1 Model validation methods

A first approach for model order selection involves the examination of the goodness of the estimated model: this analysis should be performed exploiting the available information, that could be the estimation data, some a priori knowledge about the true system and its behaviour, or the purpose of modelling itself. If a model proves to be suitable with respect to this analysis, it can be considered a valid candidate for the representation of the true system.

The procedures that assess the quality of a model are generally called *model validation* methods. Some of them exploit estimation data in order to evaluate the agreement between the data and the estimated model, by means of statistical tests or simple simulations, which compare the measured output and the one obtained from the model. Previous knowledge about the true system can also be used: for instance, if this knowledge regards the values of some parameters involved in the model, a comparison between the expected and the estimated value can help in the validation of the model.

Among the various model validation procedures, the most powerful one, especially when estimation is performed using PEM, is the analysis of the *residuals*, i.e. of the prediction errors evaluated for the parameter estimate $\widehat{\theta}_N$:

$$\varepsilon(t) = \varepsilon(t, \widehat{\theta}_N) = y(t) - \widehat{y}(t|\widehat{\theta}_N), \qquad t = 1, ..., N \tag{2.1}$$

Notice that the last expression is equivalent to (1.7) when $\widehat{y}(t|t-1)$ is calculated for $\widehat{\theta}_N$. The name "residuals" underlines the fact that these quantities represent what remains to be explained from the data. Therefore, a first confirm of the goodness of a certain model comes from a "small" value of its residuals, computed for a certain data set. In this sense, the maximal value assumed by them or their average are useful quantities to assess the entity of the residuals on the chosen set of data. However, one would like to generalize this property to all the possible data for which they can be computed; in other words, one would like to prove that the residuals are small, independently from the data for which they are evaluated. According to this consideration, it seems reasonable to test their independence from past inputs in order to both validate their values for all the possible inputs and to prove that all the information coming from past inputs have been included in the model; if indeed this is not the case, the residuals would include traces of the past inputs. Furthermore, if no more information can be gained from the data, $\{\varepsilon(t)\}$, $t = 1, ..., N$, will be a sequence of independent random variables with zero mean, i.e. a white noise sequence with zero mean. This means that no correlation should be found between $\varepsilon(t)$ and $\varepsilon(t-\tau)$, $\tau \neq 0$, otherwise $y(t)$ could be better predicted from the data.

The residual analysis is particularly useful in practical applications, because it allows to evaluate the agreement of the model with the estimation data, but it also gives an insight on the generalization ability of the model, thanks to the cited independence tests. Therefore, by means of it, it is possible to draw conclusions on the behaviour of the model on new data, by only exploiting the estimation ones.

Next sections will describe how these tests on the residuals should be performed.

2.1.1 Residual analysis testing whiteness

The test for the whiteness of the residuals is based on their auto-correlation, defined as

$$\hat{R}_\varepsilon^N(\tau) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)\varepsilon(t-\tau) \tag{2.2}$$

If $\{\varepsilon(t)\}$, $t = 1, ..., N$, is a white noise sequence, then the auto-correlation values (2.2) are "small" for all $\tau \neq 0$. However, it is necessary to define what "small" means in a numerical context. For this purpose, the typical statistical framework of hypothesis testing should be adopted. Namely, a null hypothesis H_0 should be tested against an alternative hypothesis H_1, which is supposed to be less probable than H_0. In this context, the null hypothesis H_0 will be

H_0: $\{\varepsilon(t)\}$, $t = 1, ..., N$, are white with zero mean and variance σ^2

to be tested again the alternative hypothesis H_1 of correlation among the residuals.

Defining

$$r_\varepsilon^{N,M} = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} \begin{bmatrix} \varepsilon(t-1) \\ \vdots \\ \varepsilon(t-M) \end{bmatrix} \varepsilon(t) = \sqrt{N} \begin{bmatrix} \hat{R}_\varepsilon^N(1) \\ \vdots \\ \hat{R}_\varepsilon^N(M) \end{bmatrix} \tag{2.3}$$

it can be proved that, if H_0 holds, then [8, p. 512]

$$r_\varepsilon^{N,M} \xrightarrow{dist} \mathcal{N}(0_{M\times 1}, \sigma^4 I_M) \quad \text{as } N \to \infty \tag{2.4}$$

i.e. the rows of $r_\varepsilon^{N,M}$ are asymptotically independent Gaussian random variables. $0_{M\times 1}$ represents the null vector of size M, while I_M denotes the M-dimensional identity matrix. Therefore, under H_0,

$$\frac{N}{\sigma^4}\sum_{\tau=1}^{M}\left(\widehat{R}_\varepsilon^N(\tau)\right)^2 = \frac{N}{\sigma^4}\left(r_\varepsilon^{N,M}\right)^T r_\varepsilon^{N,M} \xrightarrow{dist} \chi^2(M) \quad \text{as } N\to\infty \tag{2.5}$$

Since the true variance is not known, it can be replaced by its estimate $\widehat{R}_\varepsilon^N(0)$ without affecting the asymptotic validity of the expression; to be precise, when N is small, the χ^2-distribution should be replaced by the F-distribution.
The result (2.5) can be directly exploited for the whiteness test. First, let us define the significance level α as

$$\alpha = P(x > \chi_\alpha^2(M)) \tag{2.6}$$

with x being a χ^2-distributed random variable with M degrees of freedom. Figure 2.1 gives a graphical representation of the definition of α for a generic χ^2-distribution with M degrees of freedom: α is given by the yellow area in the plot. In this context, α represents the risk of rejecting the null hypothesis H_0 when it holds; its value has a great influence on the efficacy of the test and it is usually chosen very small, between 0.01 and 0.1, thus limiting the described risk and also the probability of accepting H_1.
Then, the null hypothesis H_0 of the whiteness test is accepted at a significance level α if

$$x_\varepsilon^{N,M} = \frac{N}{\left(\widehat{R}_\varepsilon^N(0)\right)^2}\sum_{\tau=1}^{M}\left(\widehat{R}_\varepsilon^N(\tau)\right)^2 \leqslant \chi_\alpha^2(M) \tag{2.7}$$

Since the estimate $\widehat{R}_\varepsilon^N(0)$ of σ^2 is larger than the true value σ^2 that would be obtained as $N\to\infty$, the risk of rejecting H_0 when it holds is smaller than the expected one, but at the same time is larger the risk of accepting H_0 when it is not true [13, p. 427], [12].

While the test (2.7) holds for the whiteness of the residuals for lags τ that go from 1 to M, a test for a single value of τ can also be derived, observing that

$$\sqrt{N}\widehat{R}_\varepsilon^N(\tau) \xrightarrow{dist} \mathcal{N}(0,\sigma^2) \quad \text{as } N\to\infty \tag{2.8}$$

Therefore, the null hypothesis H_0 for the independence between $\varepsilon(t)$ and $\varepsilon(t-\tau)$ can be accepted at a significance level α if

$$\sqrt{N}\frac{\left|\widehat{R}_\varepsilon^N(\tau)\right|}{\sqrt{\widehat{R}_\varepsilon^N(0)}} \leqslant \mathcal{N}_\alpha(0,1) \tag{2.9}$$

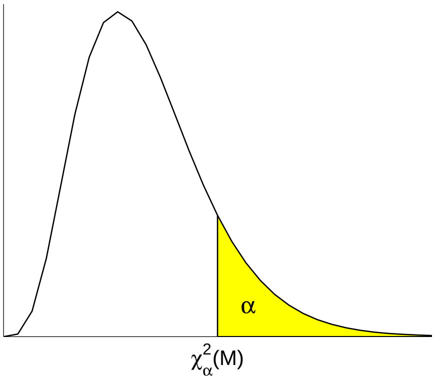


Figure 2.1: Graphical illustration of the definition of the significance level α for a generic χ^2-distribution with M degrees of freedom.

where $\mathcal{N}_\alpha(0,1)$ is a constant defined by

$$\alpha = P(|y| > \mathcal{N}_\alpha(0,1)) \tag{2.10}$$

with y being a Gaussian random variable with zero mean and unit variance.

2.1.2 Residual analysis testing independence from past inputs

The test for the independence of the residuals from the past inputs can be derived in a similar way to what was done for the whiteness test. First, it should be noticed that when the independence holds, the covariance function

$$\hat{R}^N_{\varepsilon u}(\tau) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t)u(t-\tau) \tag{2.11}$$

assumes small values. Again, by means of a statistical test it is possible to numerically assess the entity of the correlation between inputs and residuals. In this case, the null hypothesis is:

H_0 : the residuals $\{\varepsilon(t)\}$, $t = 1, ..., N$ are independent from past inputs, i.e. $E\varepsilon(t)u(s) = 0,\ t > s$

In a similar way to what was done for the whiteness test, let us define the vector

$$r_{\varepsilon u}^{N,M} = \frac{1}{\sqrt{N}} \sum_{t=1}^{N} \begin{bmatrix} u(t-M_1) \\ \vdots \\ u(t-M_2) \end{bmatrix} \varepsilon(t) = \sqrt{N} \begin{bmatrix} \hat{R}_{\varepsilon u}^{N}(M_1) \\ \vdots \\ \hat{R}_{\varepsilon u}^{N}(M_2) \end{bmatrix} \tag{2.12}$$

When H_0 holds, it can be proved that [8, p. 513]

$$r_{\varepsilon u}^{N,M} \xrightarrow{dist} \mathcal{N}(0_{M\times 1}, P_{\varepsilon u}) \quad \text{as } N \to \infty \tag{2.13}$$

where $M = M_2 - M_1 + 1$, while the covariance matrix $P_{\varepsilon u}$

$$P_{\varepsilon u} = \lim_{N\to\infty} E\left[r_{\varepsilon u}^{N,M} \left(r_{\varepsilon u}^{N,M} \right)^T \right] \tag{2.14}$$

depends on the properties of the residuals. If they constitute a white noise sequence with zero mean and variance σ^2, then [13, p. 427]

$$P_{\varepsilon u} = \sigma^2 \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} E \begin{bmatrix} u(t-M_1) \\ \vdots \\ u(t-M_2) \end{bmatrix} \begin{bmatrix} u(t-M_1) & \cdots & u(t-M_2) \end{bmatrix} \tag{2.15}$$

If instead, the residuals are not white, but they can be expressed as

$$\varepsilon(t) = \sum_{k=0}^{\infty} f_k e(t-k) \tag{2.16}$$

with $f_0 = 1$ and $e(t)$ being white noise with variance σ^2, then

$$P_{\varepsilon u} = \sigma^2 \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} E\phi(t)\phi(t)^T, \qquad \phi(t) = \sum_{k=0}^{\infty} f_k \begin{bmatrix} u(t+k-M_1) \\ \vdots \\ u(t+k-M_2) \end{bmatrix} \tag{2.17}$$

Therefore, if the null hypothesis H_0 holds

$$x_{\varepsilon u}^{N,M} = \left(r_{\varepsilon u}^{N,M} \right)^T P_{\varepsilon u}^{-1} r_{\varepsilon u}^{N,M} \xrightarrow{dist} \chi^2(M) \quad \text{as } N \to \infty \tag{2.18}$$

and H_0 is accepted at a significance level α if

$$x_{\varepsilon u}^{N,M} \leqslant \chi_{\alpha}^2(M) \tag{2.19}$$

where α is defined by (2.6).
Again, the test is still valid when the asymptotic covariance matrix $P_{\varepsilon u}$ is replaced by its estimate computed for a finite, but large, value of N. Furthermore, the value of α, with its significant impact on the efficacy of the test, should be

carefully chosen: in light of this, a specific analysis on the selection of this value will be conducted in Chapter 4.

The test can be performed also for a given value of τ, observing that, if H_0 holds, then

$$\sqrt{N}\hat{R}^N_{\varepsilon u}(\tau) \xrightarrow{dist} \mathcal{N}(0, P_\tau) \quad \text{as } N \to \infty \tag{2.20}$$

where P_τ is the τ-th diagonal element of the matrix $P_{\varepsilon u}$ [8, p. 513]. Therefore, the null hypothesis H_0 of independence between $\varepsilon(t)$ and $u(t-\tau)$, $t = 1, ..., N$ will be accepted at a significance level α if

$$\left|\hat{R}^N_{\varepsilon u}(\tau)\right| \leqslant \sqrt{\frac{P_\tau}{N}}\mathcal{N}_\alpha(0,1) \tag{2.21}$$

where $\mathcal{N}_\alpha(0,1)$ is defined in (2.10).

When evaluating the cross-correlation between inputs and residuals using estimation data, a specific mention should be given to the choice of τ. In particular, when $\tau < 0$ and $u(t)$ is white, then $\hat{R}^N_{\varepsilon u} = 0$ even if the model is inaccurate. Moreover, if $\tau < 0$ and the system operates in closed loop during the measurements, then $\hat{R}^N_{\varepsilon u} \neq 0$ even for a precise model. On the other hand, when $\tau > 0$ and the model is estimated by least-squares, then $\hat{R}^N_{\varepsilon u} = 0$ for $\tau = 1, ..., n_b$, because of the uncorrelation between the residuals and the regressors that arises from the least-squares procedure. Indeed, the regressors used in PEM contain the n_b past input values, where n_b is the order of the polynomial convolved with the inputs in transfer function models [8, p. 514], [13, p. 426].
These considerations should be kept in mind also for the choice of the numbers M_1 and M_2.

2.1.2.1 Use of validation methods for model order selection

In the previous sections the tests for model validation have been presented mainly as methods for assessing the goodness of an isolate model. However, a specific procedure for model order selection should evaluate many model structures with different complexities in order to identify the most suitable one for the system to be identified. For this purpose, it is possible to extend the model validation procedures to an order selection criterion, by iteratively performing one of the described tests on model structures of increasing complexity ($\mathcal{M}_0 \subset \mathcal{M}_1 \subset ...\mathcal{M}_j \subset ...$): such a procedure will stop when a certain model structure passes the considered test and the corresponding order will be returned.

Another possible application of these validation tests is the combination between them and one of the comparison methods described in the next section: namely, by performing the residual analysis on the model structure selected by a comparison method, the quality of that model structure can be further confirmed or called into question. Section 2.3.2 will specifically describe this application of the residuals tests.

2.2 Comparison methods

Comparison methods require the definition of a quality measure, which evaluates the goodness of an estimated model in terms of description of the estimation data and generalization to new data. In other words, the function should measure how well the model is able to reproduce both the data used for its estimation (called *estimation* or *training* data) and also new data, denoted as *test* or *validation* data. Since the common identification procedures are based on prediction models, a suitable quality measure should evaluate the prediction ability of the estimated model.

Assuming that the true system can be completely described by the model structure $\mathcal{M}$, i.e. that exists $\theta_0 \in D_{\mathcal{M}} \subset \mathbb{R}^{d_{\mathcal{M}}}$ such that $\mathcal{M}(\theta_0)$ coincides with the true system $\mathcal{S}$ ($\mathcal{M}(\theta_0) \equiv \mathcal{S}$), a proper quality measure $J(m) = J(\hat{\theta}_N)$ should be a smooth function of θ and it should be minimized by θ_0:

$$J(\theta) \geqslant J(\theta_0), \quad \forall\theta \tag{2.22}$$

Indicating with $\hat{y}_k(t|m) = \hat{y}(t|t-k)$ the k-step ahead prediction for the model m, a first quality measure based on the prediction ability of the model is defined as

$$J_k(m) = \frac{1}{N}\sum_{t=1}^{N}\{y(t) - \hat{y}_k(t|m)\}^2 \tag{2.23}$$

i.e. as the sum of squared prediction errors on a certain set of data [8, p. 500]. When computed for the estimation data, $J_k(m)$ represents the *estimation error* (or *training error*), since it provides information about the ability of the model to reproduce the data used for its estimation.
The quality measure (2.23) is here defined for a generic k-step ahead predictor, but if $k = 1$ it coincides with the loss function (1.4) (when a quadratic norm is used and no filtering is performed on the residuals), i.e. with the criterion adopted for the model estimation. Thus, when $J_1(m)$ is computed on the estimation data, its value will decrease when a more complex model is adopted, since the minimization (1.5) is performed on a larger set of values; in other words, a more complex model has more degrees of freedom by means of which it can better adjust to the estimation data. If on the one hand this property

could seem positive, on the other hand the risk to reproduce also non-relevant features, such as the particular noise realization present in the estimation data is higher with a flexible model. This phenomenon is known as *overfitting*. It follows that a small value of $J_1(m)$ when evaluated on estimation data is not always a right indicator of the goodness of a model: in order to exploit the information coming from $J_1(m)$, one should be able to distinguish when the decrease of $J_1(m)$ in correspondence to a more complex model is due to the capture of relevant features and when instead it is due to the adaptation to the noise realization.

The last observation suggests that $J_k(m)$ can be a reliable indicator of the quality of a model m when it is evaluated on a new set of data, independent from the estimation ones and usually denoted as *validation data* or *test data*. However, since the definition of $J_k(m)$ makes it dependent on the data for which it is computed, a more general measure of the quality of the model m is given by its expectation with respect to the data, i.e.

$$\bar{J}_k(m) = EJ_k(m) \tag{2.24}$$

This quantity is referred to as *generalization error* or as *test error*. Here the estimation data set is fixed and therefore the model $m = \mathcal{M}(\widehat{\theta}_N)$ is considered as a deterministic quantity. However, it depends from $\widehat{\theta}_N$, which actually is a random variable, since it is estimated from a certain set of data records, in which a noise component is present. Taking this observation into account, a quality measure for the model structure $\mathcal{M}$, depending on $d_{\mathcal{M}} = \dim \theta$ parameters, can be defined as

$$\bar{J}_k(\mathcal{M}) = E_m\left[\bar{J}_k(m)\right] = E_m\left[\bar{J}_k(\mathcal{M}(\widehat{\theta}_N))\right] \tag{2.25}$$

where E_m indicates the expectation with respect to the model m described by $\widehat{\theta}_N$. This quantity is also known as *expected prediction error* (EPE) or *expected test error*, since it averages the quality measures $\bar{J}_k(m_i)$ of the models m_i, $i = 1, ..., n_i$, estimated on different estimation data sets [4, p. 220].

Traditionally, the expected prediction error admits a decomposition into a *bias* part and a *variance* one, whose values strictly depend on the model complexity. We assume here that the observed data are described by

$$y(t) = G(q, \theta_0)u(t) + e(t) \tag{2.26}$$

where $E[e(t)] = 0$ and $E[e(t)e(s)] = \sigma^2\delta_{t,s}$, with $\delta_{t,s}$ representing the Kronecker delta. θ_0 is the true parameter vector that has to be estimated, while the input $u(t)$ is considered a known deterministic quantity.

Let $(u(t_0), y(t_0))$ be a data point coming from the validation data set, which is assumed to be independent from the estimation data set. The EPE computed on the data point $(u(t_0), y(t_0))$, assuming $N = 1$, is given by:

$$\begin{aligned} \bar{J}_k(t_0;\mathcal{M}) &= E_m\left[E J_k(t_0;m)\right] \\ &= E_m\left[E\left[\{y(t_0) - \hat{y}_k(t_0;m)\}^2\right]\right] && (2.27) \\ &= E_m\left[\sigma^2 + E\left[\{E[y(t_0)] - \hat{y}_k(t_0;m)\}^2\right]\right] && (2.28) \\ &= \sigma^2 + E\left[E_m\left[\{E[y(t_0)] - \hat{y}_k(t_0;m)\}^2\right]\right] && (2.29) \end{aligned}$$

Notice that expression (2.28) has been derived exploiting the independence between the estimation and the validation data sets. Furthermore, in (2.29) the order of the two expectations has been exchanged; this is possible because they are calculated for a non-negative term. Thus, the inner expectation becomes:

$$\begin{aligned} E_m\left[\{E[y(t_0)] - \hat{y}_k(t_0;m)\}^2\right] &= \{E[y(t_0)] - E_m\left[\hat{y}_k(t_0;m)\right]\}^2 + \\ &+ E_m\left[\{E_m\left[\hat{y}_k(t_0;m)\right] - \hat{y}_k(t_0;m)\}^2\right] && (2.30) \end{aligned}$$

where the fact that the quantity $E[y(t_0)] - E_m\left[\hat{y}_k(t_0;m)\right]$ is constant w.r.t. m has been exploited. The two addends appearing in (2.30) are respectively the squared *bias* and the *variance*. Namely, the bias

$$E[y(t_0)] - E_m\left[\hat{y}_k(t_0;m)\right] = G(q,\theta_0)u(t_0) - E_m\left[\hat{y}_k(t_0;m)\right] \qquad (2.31)$$

gives the amount by which the average of the estimates done over many data sets differs from the true system. The variance

$$E_m\left[\{E_m\left[\hat{y}_k(t_0;m)\right] - \hat{y}_k(t_0;m)\}^2\right] \qquad (2.32)$$

represents the extent to which the different estimations done over many data sets vary around their mean.

Therefore, the expected prediction error computed at t_0 can be expressed as the sum of three quantities

$$\begin{aligned} \bar{J}_k(t_0;\mathcal{M}) = \sigma^2 &+ E\left[\{G(q,\theta_0)u(t_0) - E_m\left[\hat{y}_k(t_0;m)\right]\}^2\right] + \\ &+ E\left[E_m\left[\{E_m\left[\hat{y}_k(t_0;m)\right] - \hat{y}_k(t_0;m)\}^2\right]\right] && (2.33) \end{aligned}$$

where σ^2 is a data-dependent quantity and can not be minimized by any analytical procedure, while the other two addends are respectively the expected squared bias and the expected variance. Their values are strongly influenced by the complexity of the model: low order models usually have high bias and low variance, while complex models lead to low bias and high variance. On the one

hand, the few degrees of freedom that are available for simple models can not be sufficient to properly reproduce the true system but, on the other hand, they also limit the variability in the obtained estimations. On the contrary, complex models can exploit more degrees of freedom, which allow to properly catch the dynamics of the true system, but which also lead to very different estimates according to the used data.
In light of these considerations, the minimization of the EPE translates into a trade-off between bias and variance, which in turn can be controlled by the model order selection.

The previous observations suggest an ideal procedure for the choice of the best model complexity, i.e. of the best model structure in the set $\{\mathcal{M}_j\},\ j = 1, ..., n_j$. This procedure requires the subdivision of the whole set of available data into three parts: an estimation set, a validation set and a test set. The estimation set should be further split into equally-sized subsets, each of which should be used for the estimation of a model $m_i = \mathcal{M}_j(\widehat{\theta}_i),\ i = 1, ..., n_i$. For each of them, the corresponding generalization error should be computed on the validation set. This step should be repeated for all the model structures $\mathcal{M}_j$ that has to be compared. For each of them the expected prediction error is then computed by averaging the generalization errors achieved for the models $m_i = \mathcal{M}_j(\widehat{\theta}_i),\ i = 1, ..., n_i$, and the model structure $\mathcal{M}_j = \mathcal{M}^*$ giving the lowest EPE is selected. The test set is finally used for assessing the generalization ability of the chosen model structure $\mathcal{M}^*$ by evaluating the generalization error on it. Thus, the test set should be used only during the ultimate step of the procedure in order to have a reliable measure of the generalization capability of the chosen model structure [4, Sec. 7.2,7.3].
If this procedure appears theoretically very useful, in practice it is applicable only when a large amount of data is available. This could not always be the case and even if the amount of data that can be exploited is quite large, it is usually preferable to use as many data as possible for model estimation. In order to overcome this issue, many methods have been developed in order to approximate the validation step of the procedure: these range from cross-validation to the various information criteria.

2.2.1 Cross-validation

Cross-validation is probably the most common method for model order selection and is based on the prediction ability of the tested models. It does not need any probabilistic setting and it represents a simplification of the procedure previously described, which leads to a rough estimation of the expected prediction error (EPE). Cross-validation simply requires the split of the data into two sets (usually equally-sized), here referred to as the estimation set and the validation one. For each of the possible orders, i.e. model structures $\mathcal{M}_j$, a corresponding model $m_j = \mathcal{M}_j(\hat{\theta}_j)$, $j = 1, ..., n_j$ is estimated using the estimation set. For each of them, the EPE is estimated by computing $J_k(m_j) = J_k(\mathcal{M}_j(\hat{\theta}_j))$ on the validation data and the model structure $\mathcal{M}_j = \mathcal{M}^*$ which gives the lowest EPE is chosen.

With respect to the procedure previously illustrated, which needed the split of the data into three sets and a further subdivision of the estimation set into smaller subsets, this method simply requires the split of the data into two parts and the whole estimation set is used for the estimation of a certain model.

The focus of this method is not directly on the model order but mainly on its generalization ability: the selected model could actually not capture all the information present in the estimation data, but this could lead to better performances when applied on new data.

Even if a proper application of this method requires less data than the ideal procedure previously described, its major drawback is still the fact that not all the available data are used for estimation; thus, to avoid a serious loss of information a quite large dataset is still required. Some variants of cross-validation, such as the *leave-one-out* and the *K-fold* cross-validation, allow an alleviation of this issue, by means of an intelligent exploitation of the data.
With *K-fold* cross-validation the estimation set is split into K equally sized parts and for each model structure $\mathcal{M}_j$, $j = 1, ..., n_j$, to be evaluated, K models m_i, $i = 1, ..., K$, are estimated using $K - 1$ subsets; for each model the prediction error $J_k(m_i)$ is computed on the remaining part. The EPE for the model structure $\mathcal{M}_j$ is then estimated as the average of the prediction errors $J_k(m_i)$ and the model structure giving the lowest estimated EPE is finally selected. In *leave-one-out* cross-validation, for each model structure $\mathcal{M}_j$ a number of models m_i equal to the size N of the estimation data set is estimated, by using $N - 1$ samples for each of them and exploiting the remaining sample to compute the prediction error. The EPE for the evaluated model structure $\mathcal{M}_j$ is estimated by averaging the prediction errors of the models m_i, $i = 1, ...N$; finally, the model structure $\mathcal{M}_j = \mathcal{M}^*$ giving the lowest EPE is selected.

2.2.2 Information criteria

Information criteria constitute another way of overcoming the drawback described for cross-validation: indeed, they are based on an analytic approximation of the expected prediction error, obtained using only estimation data. In this way, they allow to use all the available data for estimation, without the

need to keep aside a portion of them for validation purposes.
The application of an information criterion for model order selection within the set $\{\mathcal{M}_j\}$, $j = 1, ..., n_j$ requires first to compute the value of the specific criterion for each model structure $\mathcal{M}_j$; then, the structure $\mathcal{M}_j = \mathcal{M}^*$ leading to the lowest value of the criterion is selected.

The derivation of the information criteria is based on the assumption that the comparison function $J_k(m)$ coincides with the loss function $V_N(\theta, Z^N)$, which is minimized in order to obtain the estimated model. In most cases this assumption holds; for instance, referring to the loss function defined in (1.4), $V_N(\theta, Z^N)$ equals $J_k(m)$ when:

- $k = 1$ in $J_k(m)$;
- no filtering is performed on the residuals in $V_N(\theta, Z^N)$, i.e. $L(q) = 1$ in (1.6);
- a quadratic norm is used in (1.4), i.e.
$$l(\varepsilon_F(t,\theta), t, \theta) = \varepsilon_F^2(t,\theta) = \varepsilon^2(t,\theta) \tag{2.34}$$

As previously observed, in this case the evaluation of the quality measure $J_k(m)$ on the estimation data provides an under-estimation of the generalization error, since the same data are used both for estimation and prediction. In particular, the gap between the true generalization error and the estimation error can be represented by an expression that holds asymptotically for $N \to \infty$, leading to the following approximation of the expected prediction error for the model structure $\mathcal{M}$ [8, p. 501]:

$$\begin{aligned}\bar{J}_k(\mathcal{M}) = E_m\left[\bar{V}(\hat{\theta}_N)\right] &= E_m\left[\lim_{N\to\infty} E\left[V_N(\hat{\theta}_N, Z^N)\right]\right] \\ &\approx E_m\left[V_N(\hat{\theta}_N, Z^N)\right] + \frac{1}{N}\,\mathrm{Tr}\left[\bar{V}''(\theta^*)P_\theta\right] \end{aligned} \tag{2.35}$$

In the equation above $V_N(\theta, Z^N)$ is the loss function defined in (1.4) with the condition set in (2.34), while θ^* is the minimizer of $\bar{V}(\theta)$,

$$\bar{V}(\theta) = \lim_{N\to\infty} E\left[V_N(\theta, Z^N)\right]$$

and P_θ is the limiting value of the covariance of the parameter estimate $\hat{\theta}_N$,

$$P_\theta = \lim_{N\to\infty} E\left[N(\hat{\theta}_N - \theta^*)(\hat{\theta}_N - \theta^*)^T\right]$$

Here $E_m\left[\bar{V}(\hat{\theta}_N)\right]$ represents the expected prediction error, that is the average of the loss functions when computed on many validation data sets. Moreover, since $E_m\left[V_N(\widehat{\theta}_N, Z^N)\right]$ is the average estimation error over many data sets, the term $\frac{1}{N}\operatorname{Tr}\left[\bar{V}''(\theta^*)P_\theta\right]$ actually represents the quantity that has to be added to the estimation error to make it a reliable estimate of the expected prediction error.

FPE

The general expression (2.35), which holds when the comparison criterion $J_k(m)$ coincides with the loss function $V_N(\theta, Z^N)$, can be specialized to different cases, according to the different formulation of V_N and $J_k(m)$ that can be used.

Akaike's final prediction error criterion (FPE) holds when the conditions previously listed for the equality between $J_k(m)$ and the loss function $V_N(\theta, Z^N)$ are satisfied. Moreover, FPE is applicable when it is assumed that the validation data have the same second order properties of the estimation data.
Supposing that the model structure $\mathcal{M}$ can fully describe the true system, i.e. that the true parameter value coincides with the minimizer θ^* of the loss function computed for $N \to \infty$ ($\theta^* = \theta_0$), and that the parameters are identifiable (i.e. that $\bar{V}''(\theta_0)$ is invertible), expression (2.35) takes the specific form

$$\bar{J}_1(\mathcal{M}) = E_m\left[\bar{V}(\widehat{\theta}_N)\right] \approx E_m\left[V_N(\widehat{\theta}_N, Z^N)\right] + \frac{2d_{\mathcal{M}}}{N}\sigma^2 \tag{2.36}$$

since [8, p. 284]

$$\begin{aligned} P_\theta &= \sigma^2\left\{\lim_{N\to\infty} E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right]\right\}^{-1} \\ &= \sigma^2\left\{\lim_{N\to\infty} E\left[\frac{V''(\theta_0)}{2}\right]\right\}^{-1} \\ &= 2\sigma^2\left[\bar{V}''(\theta_0)\right]^{-1} \end{aligned} \tag{2.37}$$

and

$$\operatorname{Tr}\left[\bar{V}''(\theta_0)P_\theta\right] = 2\sigma^2\operatorname{Tr}\left[\bar{V}''(\theta_0)\left[\bar{V}''(\theta_0)\right]^{-1}\right] = 2\sigma^2\dim\theta = 2\sigma^2 d_{\mathcal{M}}$$

where $\sigma^2 = E\left[e^2(t)\right] = \bar{V}(\theta_0)$ and $\psi(t,\theta_0) = -\frac{d}{d\theta}\varepsilon(t,\theta)|_{\theta=\theta_0}$.
A further simplification of (2.36) can be obtained disregarding the average of the training errors computed on many estimation sets and considering the error computed on the unique available data set:

$$\bar{J}_1(\mathcal{M}) \approx V_N(\widehat{\theta}_N, Z^N) + \frac{2d_{\mathcal{M}}}{N}\sigma^2 \tag{2.38}$$

The last expression shows how the gap between the estimation error and the expected prediction error depends on the model complexity $d_{\mathcal{M}}$: the reduction of $V_N(\widehat{\theta}_N, Z^N)$ that follows from an increase of the model complexity $d_{\mathcal{M}}$ is counterbalanced by an increase of the second term, that acts as a penalty on model complexity. Moreover, referring to the bias-variance decomposition described in (2.30), $V_N(\widehat{\theta}_N, Z^N)$ represents the bias contribution to the expected prediction error, while $\frac{2d_{\mathcal{M}}}{N}\sigma^2$ accounts for the variance term, whose value increases with $d_{\mathcal{M}}$.

In practice, (2.38) can not be computed because σ^2 is not known but it can be estimated by [8, p. 504]

$$\widehat{\sigma}_N^2 = \frac{V_N(\widehat{\theta}_N, Z^N)}{1 - \frac{d_{\mathcal{M}}}{N}}$$

leading to the definitive version of the final prediction-error criterion (FPE) introduced by Akaike

$$\bar{J}_1(\mathcal{M}) \approx \frac{1 + \frac{d_{\mathcal{M}}}{N}}{1 - \frac{d_{\mathcal{M}}}{N}} V_N(\widehat{\theta}_N, Z^N) = \frac{1 + \frac{d_{\mathcal{M}}}{N}}{1 - \frac{d_{\mathcal{M}}}{N}} \left[\frac{1}{N} \sum_{t=1}^{N} \varepsilon^2(t, \widehat{\theta}_N) \right] \tag{2.39}$$

It is worth to notice that when $d_{\mathcal{M}} << N$ the expression above can be approximated by

$$\bar{J}_1(\mathcal{M}) = V_N(\widehat{\theta}_N, Z^N) \left[1 + \frac{\frac{2d_{\mathcal{M}}}{N}}{1 - \frac{d_{\mathcal{M}}}{N}} \right] \approx V_N(\widehat{\theta}_N, Z^N) \left[1 + \frac{2d_{\mathcal{M}}}{N} \right] \tag{2.40}$$

The model order selection reduces to find the structure $\mathcal{M}^*$ for which (2.39) is minimized:

$$\mathcal{M}^* = \arg\min_{\mathcal{M} \in \Xi} \frac{1 + \frac{d_{\mathcal{M}}}{N}}{1 - \frac{d_{\mathcal{M}}}{N}} V_N(\widehat{\theta}_N^{\mathcal{M}}, Z^N) \tag{2.41}$$

or, equivalently, when $d_{\mathcal{M}} << N$

$$\mathcal{M}^* = \arg\min_{\mathcal{M} \in \Xi} V_N(\widehat{\theta}_N^{\mathcal{M}}, Z^N) \left[1 + \frac{2d_{\mathcal{M}}}{N} \right] \tag{2.42}$$

where $\Xi = \{\mathcal{M}_j\},\ j = 1, ..., n_j$.

AIC

The Akaike Information Criterion is applicable when the estimation criterion is based on the log-likelihood function of the prediction errors $\varepsilon(t,\theta)$, i.e.

$$\begin{aligned} V_N(\theta, Z^N) &= -\frac{1}{N} L_N(\theta, Z^N) \\ &= -\frac{1}{N} \sum_{t=1}^{N} \log f_e(\varepsilon(t,\theta), t; \theta) \\ &= -\frac{1}{N} \log \left[\prod_{t=1}^{N} f_e(\varepsilon(t,\theta), t; \theta) \right] \end{aligned} \tag{2.43}$$

where $f_e(x,t;\theta)$ is the PDF of the prediction errors $\varepsilon(t,\theta) = y(t) - \widehat{y}(t|\theta)$.

Again, let us assume that the model structure $\mathcal{M}$ can fully describe the true system (i.e. $\theta^* = \theta_0$), that the parameters are identifiable (so that $\bar{V}''(\theta_0)$ is invertible) and that the validation data have the same 2nd order properties as the estimation data. In this context the Cramer-Rao inequality can be exploited [8, p. 214]:

$$\begin{aligned} E\left[(\widehat{\theta}_N - \theta_0)(\widehat{\theta}_N - \theta_0)^T\right] &\geqslant -E\left[\frac{d^2}{d\theta} \log f_y(\theta; y^N)\right]\Bigg|_{\theta=\theta_0} \\ &\geqslant -E\left[\frac{d^2}{d\theta} \log \left(\prod_{t=1}^{N} f_e(\varepsilon(t,\theta), t; \theta) \right)\right]\Bigg|_{\theta=\theta_0} \\ &\geqslant -\left\{E\left[L_N''(\theta_0)\right]\right\}^{-1} \end{aligned} \tag{2.44}$$

where it is assumed that the true joint PDF for the observations $y^N = \begin{bmatrix} y(1) & \cdots & y(N) \end{bmatrix}^T$ is $f_y(\theta_0; y^N) = \prod_{t=1}^{N} f_e(\varepsilon(t,\theta_0), t; \theta_0)$. When $N \to \infty$ the inequality holds as equality and

$$\begin{aligned} P_\theta &= \lim_{N\to\infty} NE\left[(\widehat{\theta}_N - \theta_0)(\widehat{\theta}_N - \theta_0)^T\right] \\ &= -N\left\{\lim_{N\to\infty} E\left[L_N''(\theta_0)\right]\right\}^{-1} \\ &= N\left\{N\bar{V}''(\theta_0)\right\}^{-1} = \left\{\bar{V}''(\theta_0)\right\}^{-1} \end{aligned} \tag{2.45}$$

Thus,

$$\operatorname{Tr}\left[\bar{V}''(\theta_0) P_\theta\right] = \operatorname{Tr}\left[\bar{V}''(\theta_0)\left\{\bar{V}''(\theta_0)\right\}^{-1}\right] = \dim\theta = d_{\mathcal{M}}$$

and (2.35) assumes the following expression

$$E_m\left[\bar{V}(\widehat{\theta}_N)\right] \approx V_N(\widehat{\theta}_N, Z^N) + \frac{d_{\mathcal{M}}}{N} = -\frac{1}{N} L_N(\widehat{\theta}_N, Z^N) + \frac{d_{\mathcal{M}}}{N} \tag{2.46}$$

Therefore, the model order selection problem translates into the following minimization problem

$$\mathcal{M}^* = \arg\min_{\mathcal{M}\in\Xi} -\frac{1}{N} L_N(\hat{\theta}_N^{\mathcal{M}}, Z^N) + \frac{d_{\mathcal{M}}}{N} \tag{2.47}$$

with $\Xi = \{\mathcal{M}_j\},\ j = 1, ..., n_j$.

A specific formulation of (2.46) can be given when the prediction errors $\varepsilon(t,\theta)$ are assumed to be Gaussian with zero mean and unknown variance σ^2; in this case,

$$L_N(\theta, Z^N) = -\frac{1}{2}\sum_{t=1}^{N} \frac{\varepsilon(t,\theta')}{\sigma^2} - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi \tag{2.48}$$

with $\theta = [\theta' \ \sigma^2]$ being the unknown parameters. Replacing them with their estimations,

$$\begin{aligned} \hat{\theta}'_N &= \arg\min_{\theta\in D_{\mathcal{M}}} \sum_{t=1}^{N} \varepsilon^2(t,\theta) \\ \hat{\sigma}_N^2 &= \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\hat{\theta}'_N) \end{aligned}$$

(2.46) becomes

$$E_m\left[\bar{V}(\hat{\theta}_N)\right] \approx \frac{1}{2} + \frac{1}{2}\log 2\pi + \frac{1}{2}\log\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\hat{\theta}_N)\right] + \frac{d_{\mathcal{M}}}{N} \tag{2.49}$$

Thus, the optimal model structure $\mathcal{M}^*$ according to this criterion is chosen as

$$\mathcal{M}^* = \arg\min_{\mathcal{M}\in\Xi} \log\left[\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\hat{\theta}_N^{\mathcal{M}})\right] + \frac{2d_{\mathcal{M}}}{N} \tag{2.50}$$

or, equivalently when $d_{\mathcal{M}} << N$, as

$$\mathcal{M}^* = \arg\min_{\mathcal{M}\in\Xi} \log\left[\left(1 + \frac{2d_{\mathcal{M}}}{N}\right)\frac{1}{N}\sum_{t=1}^{N}\varepsilon^2(t,\hat{\theta}_N^{\mathcal{M}})\right] \tag{2.51}$$

The concept behind the use of this criterion is analogous to the one described for FPE: the reduction of the normalized sum of squared prediction errors evaluated on the estimation data, which arises from an increase of $d_{\mathcal{M}}$, is penalized by the term $2d_{\mathcal{M}}/N$. Therefore, a larger order is selected only if it gives rise to a considerable reduction of the first term, with respect to the increase detected in $2d_{\mathcal{M}}/N$: this should allow to recognize when a further increase of $d_{\mathcal{M}}$ will only capture unnecessary features, such as the noise component.

If the final formulations of the FPE and AIC criteria are derived by different considerations, the two criteria actually are almost analogous, as can be seen from (2.42) and (2.51), where the AIC criterion can be derived from the FPE one, by simply taking its logarithm.

BIC and other information criteria

Although FPE and AIC are the most common information criteria, many other criteria have been defined according to the type of penalty inflicted to the model order increase.

Starting from the formulation (2.42) of the FPE criterion, a general expression for the information criteria can be derived as:

$$W_N(\widehat{\theta}_N, \mathcal{M}, Z^N) = V_N(\widehat{\theta}_N, Z^N)\left(1 + U_N(\mathcal{M}, Z^N)\right) \tag{2.52}$$

where $V_N(\widehat{\theta}_N, Z^N)$ represents the loss function (1.4) which is minimized to obtain the estimate $\widehat{\theta}_N$, while $U_N(\mathcal{M}, Z^N)$ is a function that is properly defined in order to penalize the model complexity. An alternative formulation can be also derived from the AIC criterion, namely

$$W_N(\widehat{\theta}_N, \mathcal{M}, Z^N) = \log\left[V_N(\widehat{\theta}_N, Z^N)\right] + T_N(\mathcal{M}, Z^N) \tag{2.53}$$

with $T_N(\mathcal{M}, Z^N)$ having the same role as $U_N(\mathcal{M}, Z^N)$. According to how $U_N(\mathcal{M}, Z^N)$ and $T_N(\mathcal{M}, Z^N)$ are chosen, the increase of model complexity can be penalized in a heavier or lighter way. In particular, the choice of $U_N(\mathcal{M}, Z^N) = d_{\mathcal{M}} \frac{\log N}{N}$ leads to the so-called MDL (*Minimum description length*) criterion:

$$W_N(\widehat{\theta}_N, \mathcal{M}, Z^N) = V_N(\widehat{\theta}_N, Z^N)\left(1 + d_{\mathcal{M}} \frac{\log N}{N}\right) \tag{2.54}$$

Thanks to the larger penalty on $d_{\mathcal{M}}$, this criterion tends to select very simple models, namely the smallest model able to sufficiently reproduce the available data: its name is actually due to this property. In literature, this criterion is also known as BIC (*Bayesian Information Criterion*), since it can also be derived using a Bayesian approach for model order selection [4, Sec. 7.7].

Alternative definitions of the function $U_N(\mathcal{M}, Z^N)$ can also be found in literature; for instance [7]:

$$\begin{aligned} U1_N(\mathcal{M}, Z^N) &= d_{\mathcal{M}} \frac{d_{\mathcal{M}}^{\frac{1}{3}}}{N} \\ U2_N(\mathcal{M}, Z^N) &= d_{\mathcal{M}} \frac{2 \log d_{\mathcal{M}}}{N} \\ U3_N(\mathcal{M}, Z^N) &= d_{\mathcal{M}} \frac{\log d_{\mathcal{M}} \log N}{N} \end{aligned} \tag{2.55}$$

2.2.3 F-test

While the information criteria allow the direct comparison of many model structures $\{\mathcal{M}_j\}$, $j = 1, ..., n_j$, statistical tests can be used to compare two model structures $\mathcal{M}_0$ and $\mathcal{M}_1$, again exploiting the information coming from the loss function (1.4) and from the model complexity $d_{\mathcal{M}} = \dim \theta$.

Let us assume that $\mathcal{M}_0 \subset \mathcal{M}_1$, i.e. $d_{\mathcal{M}_0} < d_{\mathcal{M}_1}$. As previously mentioned, statistical tests are based on the definition of a null hypothesis H_0, on which a higher degree of confidence is posed and which has to be tested against an alternative hypothesis H_1. Since $\mathcal{M}_0 \subset \mathcal{M}_1$, it is preferable to choose $\mathcal{M}_0$, if it can be proved that it can properly reproduce the available data. Thus, for this application the two hypothesis will be:

H_0 : the data are generated by $\mathcal{M}_0(\widehat{\theta}_N^{(0)})$

H_1 : the data are generated by $\mathcal{M}_1(\widehat{\theta}_N^{(1)})$

where $\widehat{\theta}_N^{(0)}$ and $\widehat{\theta}_N^{(1)}$ are the parameter estimates for the two evaluated model structures, i.e. the minimizers of the loss functions $V_N(\theta^{(0)}, Z^N)$ and $V_N(\theta^{(1)}, Z^N)$, respectively. For the comparison the following test quantity is usually considered:

$$v = N \cdot \frac{V_N(\widehat{\theta}_N^{(0)}, Z^N) - V_N(\widehat{\theta}_N^{(1)}, Z^N)}{V_N(\widehat{\theta}_N^{(1)}, Z^N)} \tag{2.56}$$

A large value of v is an indication of a large gap between the two loss functions, meaning that the adoption of a more complex model is beneficial; on the other hand, when v is small, there is no significant advantage in the choice of a more complex model and the simpler model is preferable.

The numerical analysis on the value of v is based on the fact that v is asymptotically χ^2-distributed if the true system $\mathcal{S}$ can be correctly described by $\mathcal{M}_0 \subset \mathcal{M}_1$, i.e. if $\mathcal{S} \in \mathcal{M}_0$; more precisely,

$$v = N \frac{V_N(\widehat{\theta}_N^{(0)}, Z^N) - V_N(\widehat{\theta}_N^{(1)}, Z^N)}{V_N(\widehat{\theta}_N^{(1)}, Z^N)} \xrightarrow{dist} \chi^2(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}), \text{ as } N \to \infty \quad (2.57)$$

where $d_{\mathcal{M}_0} = \dim \theta^{(0)}$ and $d_{\mathcal{M}_1} = \dim \theta^{(1)}$, [8, p. 560, Lemma II.4]. Thus, the previous distinction between a large and a small value of v translates into the choice of the significance level α at which the null hypothesis is tested: namely, the smallest model structure $\mathcal{M}_0$ is chosen if

$$v \leqslant \chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \quad (2.58)$$

where $\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})$ is defined by

$$\alpha = P\left(x > \chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})\right) \quad (2.59)$$

with x being a $\chi^2(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})$-distributed random variable.
The significance level α represents the maximal risk of rejecting the null hypothesis H_0 when it is true, i.e. of choosing the more complex model structure $\mathcal{M}_1$ when $\mathcal{M}_0$ is sufficiently able to reproduce the given data. This means that a small value of α will reduce the probability of selecting the model structure $\mathcal{M}_1$.

For small values of N, the statistical test is more appropriate if based on the F-distribution: in this case a new test quantity has to be used, namely

$$f = \frac{V_N(\widehat{\theta}_N^{(0)}, Z^N) - V_N(\widehat{\theta}_N^{(1)}, Z^N)}{V_N(\widehat{\theta}_N^{(1)}, Z^N)} \cdot \frac{N - d_{\mathcal{M}_1}}{d_{\mathcal{M}_1} - d_{\mathcal{M}_0}} \quad (2.60)$$

which is asymptotically F-distributed under the hypothesis H_0 [8, p. 560, Lemma II.4]

$$f \xrightarrow{dist} F(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}, N - d_{\mathcal{M}_1}) \quad \text{as } N \to \infty \quad (2.61)$$

Moreover, when $\widehat{\theta}_N$ is estimated by linear regression, f is exactly F-distributed. Assuming that x is an F-distributed random variable with degrees of freedom $d_{\mathcal{M}_1} - d_{\mathcal{M}_0}$ and $N - d_{\mathcal{M}_1}$, i.e. $x \in F(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}, N - d_{\mathcal{M}_1})$, let us define

$$\alpha = P\left(x > F_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}, N - d_{\mathcal{M}_1},)\right) \quad (2.62)$$

Then, similarly to the χ^2 test, $\mathcal{M}_0$ is selected at a significance level α if

$$f \leqslant F_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}, N - d_{\mathcal{M}_1}) \quad (2.63)$$

In order to exploit these tests for model order selection among many structures of increasing complexity, $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \ldots \subset \mathcal{M}_{n_j}$, the tests should be performed starting from the simplest structure: the procedure is stopped when (2.58) (or (2.63)) is verified for the first time and the smallest model structure involved in that test is selected.

2.2.4 Relationship between F-test and information criteria

It can be shown that each information criterion (illustrated in Section 2.2.2) is asymptotically equivalent to the test (2.58) when a specific significance level α is chosen.

Let us concentrate on the choice between two model structures $\mathcal{M}_0$ and $\mathcal{M}_1$ with $\dim \theta^{(0)} = d_{\mathcal{M}_0} < d_{\mathcal{M}_1} = \dim \theta^{(1)}$. According to the test (2.58), $\mathcal{M}_0$ is chosen if

$$V_N(\widehat{\theta}_N^{(0)}, Z^N) \leqslant V_N(\widehat{\theta}_N^{(1)}, Z^N) \left[1 + \frac{1}{N}\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})\right] \tag{2.64}$$

On the other hand, using the criterion (2.52), $\mathcal{M}_0$ is selected if

$$\begin{aligned} V_N(\widehat{\theta}_N^{(0)}, Z^N)\left[1 + U_N(\mathcal{M}_0, Z^N)\right] &\leqslant V_N(\widehat{\theta}_N^{(1)}, Z^N)\left[1 + U_N(\mathcal{M}_1, Z^N)\right] \\ V_N(\widehat{\theta}_N^{(0)}, Z^N) &\leqslant V_N(\widehat{\theta}_N^{(1)}, Z^N)\frac{1 + U_N(\mathcal{M}_1, Z^N)}{1 + U_N(\mathcal{M}_0, Z^N)} \end{aligned} \tag{2.65}$$

The two criteria are therefore equivalent if

$$1 + \frac{1}{N}\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) = \frac{1 + U_N(\mathcal{M}_1, Z^N)}{1 + U_N(\mathcal{M}_0, Z^N)} \tag{2.66}$$

i.e. if

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) = N \cdot \frac{U_N(\mathcal{M}_1, Z^N) - U_N(\mathcal{M}_0, Z^N)}{1 + U_N(\mathcal{M}_0, Z^N)} \tag{2.67}$$

The last expression defines the value of the significance level α of the F-test for which it becomes equivalent to the condition stated by a generic information criterion.

Equation (2.67) can be specialized for the criteria previously illustrated. Namely, in the FPE criterion $U_N(\mathcal{M}, Z^N) = 2d_{\mathcal{M}}/N$, and therefore

$$\begin{aligned} \chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) &= N \cdot \frac{2d_{\mathcal{M}_1}/N - 2d_{\mathcal{M}_0}/N}{1 + 2d_{\mathcal{M}_0}/N} \\ &= 2N \cdot \frac{d_{\mathcal{M}_1} - d_{\mathcal{M}_0}}{N + 2d_{\mathcal{M}_0}} \\ &\approx 2(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \end{aligned} \tag{2.68}$$

where the last approximation holds for $N \to \infty$. Therefore, FPE is asymptotically equivalent to the F-test with a significance level α defined by (2.68) and (2.59).
For what regards BIC criterion instead, since $U_N(\mathcal{M}, Z^N) = d_{\mathcal{M}} \log N/N$ we have

$$\begin{aligned}
\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) &= N \cdot \frac{(d_{\mathcal{M}_1} \log N)/N - (d_{\mathcal{M}_0} \log N)/N}{1 + d_{\mathcal{M}_0} \log N/N} \\
&= N \cdot \frac{\log N(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})}{N + d_{\mathcal{M}_0} \log N} \\
&\approx \log N(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \qquad (2.69)
\end{aligned}$$

The analogy can be derived also for the criterion (2.53), which can be characterized into the AIC criterion. According to (2.53), the simplest structure $\mathcal{M}_0$ is chosen if

$$\begin{aligned}
\log\left[V_N(\widehat{\theta}_N^{(0)}, Z^N)\right] + T_N(\mathcal{M}_0, Z^N) &\leqslant \log\left[V_N(\widehat{\theta}_N^{(1)}, Z^N)\right] + T_N(\mathcal{M}_1, Z^N) \\
V_N(\widehat{\theta}_N^{(0)}, Z^N) &\leqslant V_N(\widehat{\theta}_N^{(1)}, Z^N)\frac{\exp\{T_N(\mathcal{M}_1, Z^N)\}}{\exp\{T_N(\mathcal{M}_0, Z^N)\}}
\end{aligned}$$

Thus, the criteria (2.58) and (2.53) are equivalent if

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) = N\left[\frac{\exp\{T_N(\mathcal{M}_1, Z^N)\}}{\exp\{T_N(\mathcal{M}_0, Z^N)\}} - 1\right] \qquad (2.70)$$

In the case of AIC criterion, for which $T_N(\mathcal{M}, Z^N) = 2d_{\mathcal{M}}/N$, (2.70) becomes

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) = N\left[\exp\left\{\frac{2}{N}(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})\right\} - 1\right] \approx 2(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \qquad (2.71)$$

where the last approximation is valid for large values of N [13, p. 444,445].

From these considerations, it follows that it is possible to estimate the asymptotic probability that a certain information criteria will select a too complex model structure.

This analogy between the F-test and the information criteria highlights the advantages of the latter ones w.r.t. the first one. In particular, the information criteria allow to simultaneously compare all the tested model structures and does not suffer from early stopping, due to the presence of local minima, as is the case of the F-test. In addition, the tuning of the significance level α is implicitly done in the formulation of the information criteria: if on the one hand, this could be an advantage, since no tuning is required to the user, on the other hand, this could also be a limitation, because better performances could maybe be possible with a specific tuning.

2.2.5 Consistency analysis

The considerations done in Sections 2.2.3 and 2.2.4 can be exploited to drive a consistency analysis on the information criteria illustrated in 2.2.2.

For an order selection method to be *consistent*, the probability that it selects the correct order should tend to 1 when the amount of available data tends to infinity.

Let us consider again the choice between two model structures $\mathcal{M}_0$ and $\mathcal{M}_1$, with $\mathcal{M}_0 \subset \mathcal{M}_1$. Referring to (2.58) and (2.6), α represents the risk of overfitting, i.e. of choosing the largest model when the smallest one $\mathcal{M}_0$ can properly describe the true system $\mathcal{S}$ ($\mathcal{S} \in \mathcal{M}_0$). Thus, when $N \to \infty$, α should tend to 0 in order to avoid the overfitting or, equivalently,

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \longrightarrow \infty \quad \text{as } N \to \infty \tag{2.72}$$

On the other hand, the probability of underfitting, i.e. of choosing $\mathcal{M}_0$ when $\mathcal{S} \notin \mathcal{M}_0$, $\mathcal{S} \in \mathcal{M}_1$, is equal to

$$1 - \alpha = P\left(\frac{V_N(\hat{\theta}_N^{(0)}, Z^N) - V_N(\hat{\theta}_N^{(1)}, Z^N)}{V_N(\hat{\theta}_N^{(1)}, Z^N)} \leqslant \frac{1}{N}\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \right) \tag{2.73}$$

If $\mathcal{S} \notin \mathcal{M}_0$, the difference $V_N(\hat{\theta}_N^{(0)}, Z^N) - V_N(\hat{\theta}_N^{(1)}, Z^N)$ is significant, i.e.

$$V_N(\hat{\theta}_N^{(0)}, Z^N) - V_N(\hat{\theta}_N^{(1)}, Z^N) = O(1) \tag{2.74}$$

since it does not tend to zero when N tends to infinity. Therefore, to eliminate the risk of underfitting,

$$\frac{1}{N}\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \longrightarrow 0 \quad \text{as } N \to \infty \tag{2.75}$$

For FPE and AIC criteria, it was found in Section 2.2.4 that

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \approx 2(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})$$

Since of the two conditions (2.72) and (2.75), only (2.75) is verified, FPE and AIC are not consistent order selection rules but they avoid the risk of underfitting.
BIC criterion is instead consistent, because for it,

$$\chi^2_\alpha(d_{\mathcal{M}_1} - d_{\mathcal{M}_0}) \approx \log N(d_{\mathcal{M}_1} - d_{\mathcal{M}_0})$$

and both the conditions (2.72) and (2.75) are satisfied [13, p. 449].

2.3 Combinations of the order selection methods

This section introduces two combinations of the order selection methods previously illustrated. These techniques will then be applied in the following chapters to perform order selection of OE models.

2.3.1 Combination of the comparison and the validation methods

As was clarified in Section 2.1, the primary purpose of the validation methods is the assessment of the quality of a model w.r.t. the description of the available data. Then, this property can be exploited for the determination of the most suitable model complexity, as described in Section 2.1.2.1. On the other hand, comparison methods have been introduced in order to directly estimate an appropriate model complexity. These considerations suggest that a combination of these two families of model order selection methods could give more reliable results, decreasing the risk of inappropriate choices. The combination here considered equips one of the comparison methods with a validation stage in order to further assess the quality of the model structures that it originally returns. In particular, early stopping due to local minima is avoided; actually, the models deriving from local minima can be rejected if the tests on the residuals are not passed.

All the types of validation procedures will be tested in order to evaluate the efficacy of each one when adopted in this context. Therefore, the quantities $x_{\varepsilon}^{N,M}$ in (2.7) or $x_{\varepsilon u}^{N,M}$ in (2.18) (or both) are computed on the estimation data for the model structure $\mathcal{M}^*$ selected by a certain comparison method. Then, when the residuals whiteness is evaluated, $\mathcal{M}^*$ is unfalsified if the condition in (2.7) holds, i.e. if

$$x_{\varepsilon}^{N,M} \leqslant \chi_{\alpha}^{2}(M) \tag{2.76}$$

If instead the independence between residuals and past inputs is analyzed, the condition (2.19) has to be verified, i.e.

$$x_{\varepsilon u}^{N,M} \leqslant \chi_{\alpha}^{2}(M) \tag{2.77}$$

If both whiteness and independence tests are considered, the two conditions have to hold simultaneously for the model structure $\mathcal{M}^*$ to be unfalsified.
When a model structure selected by a certain criterion does not pass the validation stage, a new order is selected: in the case of cross-validation or information criteria, the new order will be the one giving the second lowest value of the prediction error or of the criterion, respectively; if instead the F-test is applied,

the new order will correspond to the one of the first larger model structure which passes the test. These newly chosen model structures will be definitively selected if they pass a new validation test.

The method here illustrated can be also re-formulated in the following way: among the model structures which pass the whiteness test (2.76) or the independence test (2.77), the one giving the lowest value of the considered information criterion (or of the prediction error in case of cross-validation) is chosen. For what regards the combination of the F-test with the validation methods, the first model structure which passes both the F-test and the considered test on the residuals is selected.

A different approach has also been tested when this technique is applied on OE and ARMAX models. Since they are estimated solving non-convex optimization problems, local minima can be returned. According to this approach, whenever a model structure does not pass a validation test, a re-estimated model is first evaluated, before passing directly to another complexity. The re-estimation is obtained using the MATLAB routine `init`, which randomly perturbs the initial parameter estimate of the old model, in order to obtain the initial estimate for the new estimation. However, it has been observed that the re-estimation tends to decrease the impulse response fit w.r.t. to the true system (defined in (3.1)). Thus, the immediate choice of a new complexity appears more beneficial.

2.3.2 Combination with the F-test

This combination exploits the F-test to perform a sort of "local search" around the order of the model structure returned by a specific order selection technique, be it a comparison or a validation method. Namely, when more complex model structures are considered, starting from the selected order, an F-test is applied to compare the current model structure with the next larger one: if the F-test is passed, i.e. if the smallest model structure is chosen, the procedure is stopped and the original order is confirmed; if instead the F-test is not passed by the simplest structure, the procedure is iterated evaluating model structures of increasing orders, until the smallest structure between the two compared in a certain test is selected. On the other hand, when simpler model structures are compared, the procedure is iterated if the smallest model is chosen, otherwise it is stopped and the most complex structure in the current comparison is returned.

When the F-test is applied in this context, an initial suggestion for the model complexity is already available, thanks to the previous application of another order selection method. One should then decide whether to test simpler or more complex model structures; this choice can be done according to the behaviour of the previously applied method: namely, if this tends to undermodel, the F-test should compare model structures of increasing complexity, while lower complexities should be tested if the first method presents a tendency to overmodel. Furthermore, the significance level α adopted in the F-test can be properly tuned in order to favour or limit the selection of higher or lower orders. Indeed, a large value of α increases the probability of accepting the most complex model structure between the two compared by an F-test. Therefore, to favour the selection of higher orders a large value of α should be chosen, while it should be small to facilitate the choice of simpler model structures.

Since both the described combinations require the setting of the significance level α for the statistical tests that they involve, a detailed experimental analysis of its choice will be carried out in the next chapters.

Chapter 3

Simulation settings

3.1 Data sets used for the simulations

The order selection methods here evaluated are tested on the data bank used in [2], which includes measurement data and the corresponding systems, from which the data were generated. The knowledge of the true system allows a direct comparison between it and the estimated model by means of the functions described in Section 3.2.

The data bank consists of 5000 SISO systems of 30th order and corresponding input-output data. The systems are sampled versions of original continuous-time systems generated with the MATLAB routine `rss`; sampling time was set to three times the bandwidth of the system. These systems were split into "fast" ones (identified as S1), having all their poles within a circle of radius 0.95 and "slow" ones (called S2), which have at least one pole outside the circle of radius 0.95.

The measurement data include output data generated by the systems when simulated with Gaussian white noise with unit variance as input. Output data are corrupted by additive white Gaussian noise with different variances, in order to distinguish between data with low SNR, where the additive output noise has the same variance of the noise-free output and data with high SNR, for which the additive noise variance is a tenth of the one of the noise-free output.

The different systems properties and output noise characteristics above described allow the subdivision of the whole data bank into four data sets:

S1D1 : It includes fast systems with high SNR, i.e. SNR=10; each set of records consists of 500 input-output measurements;

S1D2 : It contains fast systems with low SNR, i.e. SNR=1; each set of records consists of 375 input-output measurements;

S2D1 : It includes slow systems with high SNR (SNR=10); each set of records consists of 500 input-output measurements;

S2D2 : It contains slow systems with low SNR (SNR=1); each set of records consists of 375 input-output measurements;

While S1D1 presents the most favourable conditions for identification, S2D2 represents the most difficult data set to exploit for identification, because of the slow nature of the systems, which require a large amount of data to be properly identified, but also because of the low SNR and the relative small records set.

The bank of data considered should provide a good sample of the different systems and conditions that can be found in real-life scenarios: in particular, it contains complex systems, that can be quite well approximated by low-order models. This property can highlight the tendency to overfit of the order selection methods that will be tested.

3.2 Fit functions used to compare the tested criteria

The comparison of the different methods used for model order selection will be carried out by means of certain measures of goodness of the estimated models, that are described in this section. For each order selected by a specific criterion, the corresponding model will be estimated and the fit measures here introduced will be computed: their values will be then compared with the ones achieved by models having orders selected by other criteria.
Three types of fit measures will be exploited: one describing the closeness of the estimated model to the true system and other two evaluating the prediction ability of the estimated models.

Therefore, the model order selection criteria will not be evaluated according to their ability to identify the true order of the systems, but according to their capacity to find orders that allow a good reproduction of the input-output properties of the true systems.

For each kind of fit and for each model type that will be tested on a certain data set, an "*oracle*" criterion will indicate the model structure giving the maximal fit, thanks to the knowledge of the true system. This maximal fit will be considered as the upper bound achievable with the specific set conditions.

3.2.1 Impulse response fit

The quality of an estimated model is evaluated using a measure which compares the impulse response of the true model and of the estimated one. Indicating with g_k^0 and $\hat{g}_k$ the impulse response coefficients of the true system and of the estimated one, respectively, such measure function is defined as

$$\mathcal{F}(G_0, \widehat{G}) = 100 \left[1 - \left(\frac{\sum_{k=1}^{n} |g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^{n} |g_k^0 - \bar{g}^0|^2} \right)^{\frac{1}{2}} \right] \tag{3.1}$$

with

$$\bar{g}^0 = \frac{1}{n} \sum_{k=1}^{n} g_k^0$$

It should be remarked that G_0 and $\widehat{G}$ represents the set of impulse response coefficients of the true system and of the estimated one, respectively.
A perfect fit between the first n impulse response coefficients of the true model ad of the estimated one will lead to $\mathcal{F}(G_0, \widehat{G}) = 100$.
The number n of compared coefficients should be chosen sufficiently large in order to take into account all the coefficients which are significantly different from zero. In the tests here conducted n is set to 125.

3.2.2 Type 1 prediction fit

The prediction ability of an estimated model is evaluated using two types of measures. For the definition of the first one, let z_k^0 indicate the output produced by the true system when it is simulated with the input data used for the estimation and without the addition of output noise. Furthermore, let $\hat{z}_k$ represent the output generated by the estimated model in the same simulation setting.

The first type of prediction fit function is defined as

$$\mathcal{F}_1(Z_0, \widehat{Z}) = 100 \left[1 - \left(\frac{\sum_{k=1}^{N} |z_k^0 - \hat{z}_k|^2}{\sum_{k=1}^{N} |z_k^0 - \bar{z}^0|^2} \right)^{\frac{1}{2}} \right] \tag{3.2}$$

with

$$\bar{z}^0 = \frac{1}{N} \sum_{k=1}^{N} z_k^0$$

$N = 500$ when the fit is computed for systems of data sets S1D1 and S2D1, while $N = 375$ for the data sets S1D2 ad S2D2.

3.2.3 Type 2 prediction fit

The second type of prediction fit is analogous to the first one, but instead of using the input data coming from the estimation set, new input data are exploited, but of the same type as the estimation ones.
Indicating with y_k^0 ad $\hat{y}_k$ the outputs produced respectively by the true system and by the estimated one when they are fed with new input data, the fit function is defined as

$$\mathcal{F}_2(Y_0, \widehat{Y}) = 100 \left[1 - \left(\frac{\sum_{k=1}^{N} |y_k^0 - \hat{y}_k|^2}{\sum_{k=1}^{N} |y_k^0 - \bar{y}^0|^2} \right)^{\frac{1}{2}} \right] \tag{3.3}$$

with

$$\bar{y}^0 = \frac{1}{N} \sum_{k=1}^{N} y_k^0$$

Again, $N = 500$ for data sets S1D1 and S2D1, while $N = 375$ for S1D2 and S2D2.

3.3 Settings used for the statistical tests

This section illustrates some specifications which will be adopted in the order selection criteria involving statistical tests, such as the residual analysis testing whiteness and independence from past inputs and the F-test comparing two model structures.

First of all, it could happen that none of the evaluated model structures is unfalsified by one of these tests: in this case, the most complex model structure is chosen, since this should be theoretically unfalsified with higher probability.

These tests also require the definition of the significance level α at which the null hypothesis H_0 is accepted or rejected, which has a significant impact on the effectiveness of the test; for this reason, the analysis described in the following chapters focuses also on the selection of α.

When model structure selection is performed by means of the statistical tests on the residuals, these are applied using the estimation data. Furthermore, the use of these tests also requires the setting of the interval of lags for which the correlation is computed. Namely, for the whiteness test, the value of the maximal lag M introduced in (2.3) has to be set, while for the test for the independence between residuals and past inputs, the minimal and maximal lags M_1 and M_2 (introduced in (2.12)) have to be specified. In the tests performed for this project M_1 is always set to 1, such that $M = M_2 - M_1 + 1 = M_2$. Therefore, in the following the maximal lag M_2 will be denoted by M also for this type of test.
The impact of the value of M on the performances of these tests when they are used for model order selection will be investigated in Section 4.1.2.

Moreover, two different routines are used to perform the test for the independence of residuals from past inputs: the MATLAB routine `resid`, in which the whiteness of the residuals is assumed and no model for the residuals (such as (2.16)) is estimated, and a routine which estimates a noise model (2.16), computing $P_{\varepsilon u}$ as described in (2.17). It should be noticed that the correct use of `resid` for this kind of test requires first the assessment of the residuals whiteness; however, the following analysis will evaluate also the performances of the independence test when implemented through `resid` but without its combination with the whiteness test, even if this use is not theoretically correct.

Chapter 4

Classical model order selection techniques: Application on OE models

The methods described in Chapter 2 will be here tested and compared on OE models (see (1.19)). The tests are performed using 200 sets of data coming from each of the four data sets described in Section 3.1. For all the identified systems, a one-sample delay is assumed to be present in the dynamics from u to y, i.e. $n_k = 1$ in $B(q)$ (1.13).

The order selection criteria have to discriminate among orders that go from 1 to 40. For simplicity, the choice is limited to model structures, whose polynomials have all the same orders, i.e. $n_b = n_f$ in (1.13) and (1.16).

An analogous analysis has been performed also on FIR, ARX and ARMAX models. The results are reported in Appendix A.

The following acronyms are introduced to indicate the different model order selection methods:

RAW - The whiteness test on the residuals (2.7) is used for model order selection, as described in Section 2.1.2.1. The test is applied on the estimation data.

$RAI1$ - Model order is selected by means of the test which verifies the independence between residuals and past inputs, as described in Section 2.1.2.1. The test is applied on the estimation data and it assumes (without a proper check) that the residuals constitute a white noise sequence, with the covariance matrix $P_{\varepsilon u}$ computed as in (2.15).

$RAI2$ - This method is analogous to the previous one, but the independence between residuals and past inputs is evaluated without assuming residuals whiteness, with the covariance matrix $P_{\varepsilon u}$ computed as in (2.17).

RA - The tests for residuals whiteness (RAW) and for the independence between them and past inputs (RAI1) are combined, using the same value of the significance level α: model structures of increasing complexities are tested and the first one that passes both the tests is selected.

F - The F-test is applied to compare two model structures of consecutive complexities, starting from the simplest one in the range of the considered ones; as soon as the condition (2.63) is verified, the procedure is stopped and the simplest model structure between the two compared is chosen.

CV - Cross-validation is applied according to the way illustrated at the beginning of Section 2.2.1. It should be specified that the available set of data is split into two equally-sized subsets in order to obtain the estimation and the validation set.

FPE - The model structure is selected according to (2.42).

AIC - Order selection is performed by means of (2.50).

BIC - The model structure which minimizes (2.54) is chosen.

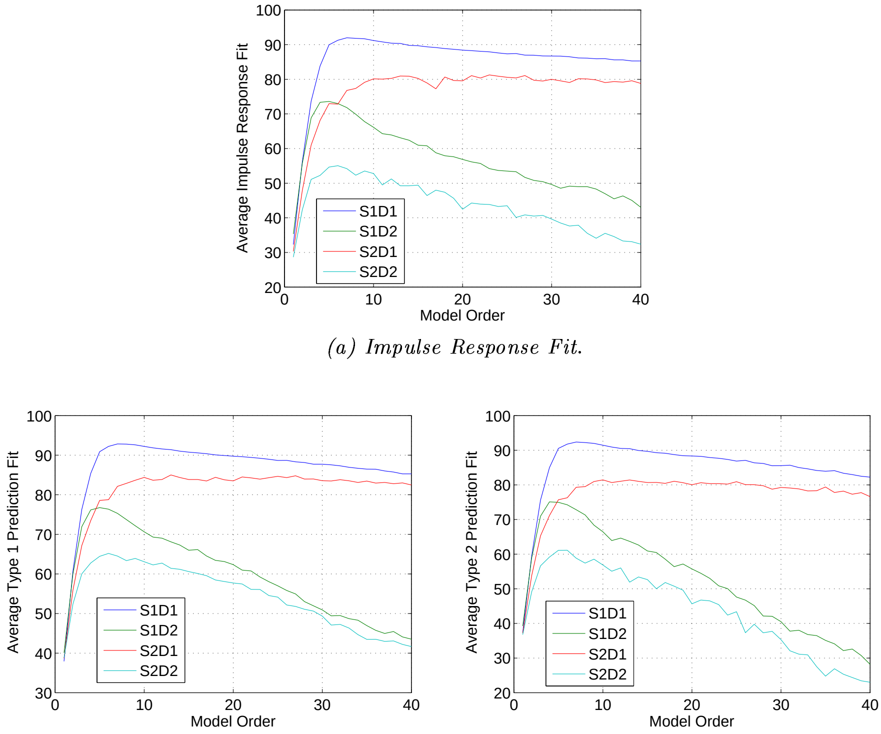
4.1 Order estimation for OE models

The analysis that follows is based on OE models estimated using the routine `pem`, with the option that avoids the estimation of a noise model.

4.1.1 Influence of the model order in the estimation

In order to understand the importance of a correct choice of the model complexity, Figure 4.1 shows the average fits obtained in the identification of 200 systems in each data set as function of the model orders.

It can be seen that all the three kinds of fit assume comparable values.



(a) Impulse Response Fit.

(b) Type 1 Prediction Fit.

(c) Type 2 Prediction Fit.

Figure 4.1: OE models - Average fits achieved in the estimation of 200 systems in each data set for model orders ranging from 1 to 40.

The plots highlight a significant difference between noisy and non-noisy data sets: while for the latter ones, high orders do not significantly deteriorate the performances, for the noisy sets, high-order OE models suffer from the so-called overfitting issue, leading to a significant worsening of the fits. As expected, this issue mostly affects type 2 prediction fit, which encounters a more relevant decrease for high complexities.

Moreover, when less noisy data are available, two different trends are observed, depending on the slow or fast dynamics of the true system. Namely, when slow systems have to be identified, quite high orders are still adequate for the reproduction of the true nature of the systems, while in case of fast systems, too high orders give rise to overfitting.

4.1.2 Influence of the maximal correlation lag M in the tests on the residuals

The validation methods described in Section 2.1 are based on statistical tests on the residuals, which in turn depend on two parameters that have to be set by the user, once the measurement data are given. These parameters are the significance level α of the test and the maximal lag M for which the auto- or the cross-correlation are computed. This section focuses on the impact of M on the performances of the tests, taking into account also the influence of the significance level.

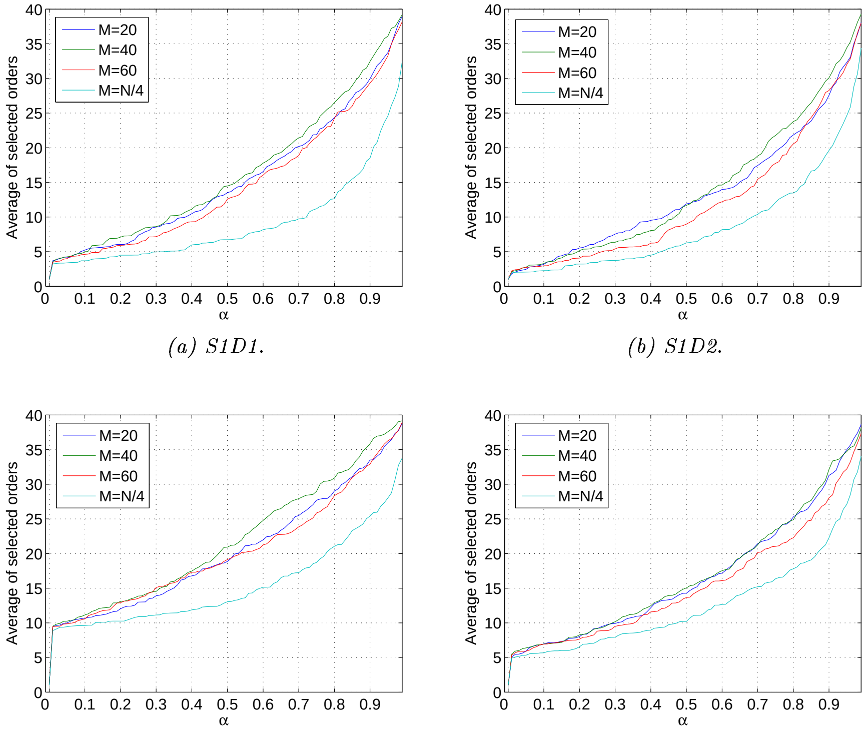
In literature no specific indications can be found for the choice of M; in [13] values between 5 and $N/4$ are suggested, being N the number of available data. The only constraint for M regards its use in the test for the independence between residuals and past inputs, when the model to be validated is estimated through least-squares (such as for FIR and ARX models). In this case M has to be chosen larger than n_b, the order of the polynomial $B(q)$ (defined in (1.13)), in order to make the test reliable. Indeed, n_b past inputs are used as regressors $\varphi(t)$ for the estimation of the predicted output $\widehat{y}(t|\theta)$ in (1.21). Therefore, the independence between residuals and past inputs until a lag equal to n_b is already guaranteed by the independence between residuals and regressors achieved by the least-squares procedure.

The analysis that follows is performed only for OE models, but analogous conclusions can be drawn for the other model types considered in Appendix A.

Whiteness test

Figure 4.2 shows the average orders selected by the whiteness test on the residuals (RAW) as function of the significance level adopted in the test and for four different values of M, while Figure 4.4 illustrates the corresponding average impulse response fits. Figure 4.2 clearly highlights how too large values of M generally lead to the selection of smaller orders, as is the case for $M = 60$ and $M = N/4$. It should be observed that the value of $\chi^2_\alpha(M)$ (defined in (2.6)) increases with the degrees of freedom M of the χ^2 distribution, when α is fixed. On the other hand, also the quantity $x_\varepsilon^{N,M}$ (defined in (2.7)) increases with M, since it is the normalized sum of the auto-correlations computed for lags from 1 to M. However, since the major correlation components among the residuals are detected for small lags, increasing M after a certain value $\widetilde{M}$ will not give rise to a significant growth of $x_\varepsilon^{N,M}$; therefore, the condition (2.7),

$$x_\varepsilon^{N,M} \leqslant \chi^2_\alpha(M)$$



(a) S1D1.

(b) S1D2.

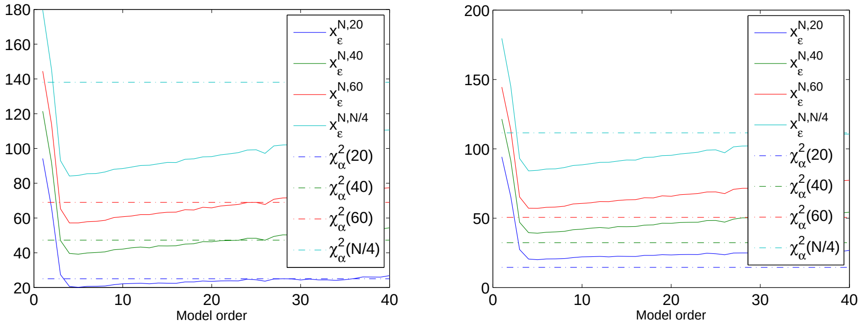
(c) S2D1.

(d) S2D2.

Figure 4.2: OE models - Average of the orders selected by the whiteness test on the residuals (RAW) as function of the significance level α and for different values of the maximal lag M for which the auto-correlation is computed. The average is calculated from the identification of 200 systems in each data set.

is more easily verified for large M.
The only exception to the behaviour just described can be observed in data set $S2D1$, where for small significance levels, setting $M = 60$ leads to the selection of larger orders w.r.t. the case with $M = 20$. This is probably the consequence of the slow nature of the systems contained in that data set, which gives rise to relevant correlation components also for large lags, when the estimated model is not appropriate. When M is set to 20, these components are not considered, causing the unfalsification of simpler model structures.

Figure 4.2 also illustrates how the increase of the significance level leads to a considerable growth of the average of the selected orders. This is a consequence of the default selection of the largest order (i.e. 40), which is done whenever none of the evaluated model structures is unfalsified by the whiteness test. Indeed,



(a) $\alpha = 0.2$.

(b) $\alpha = 0.8$.

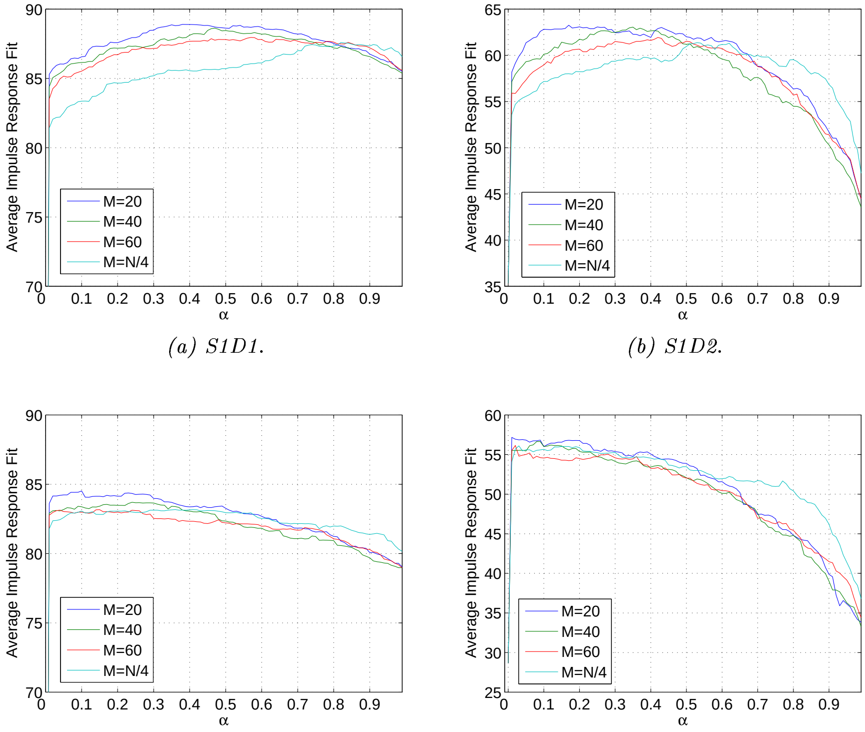
Figure 4.3: OE models, Data set S1D2 - Average values of $x_\varepsilon^{N,M}$ (solid lines) and values of $\chi^2_\alpha(M)$ (dashed-dotted lines) for different values of the maximal lag M. The significance level α is fixed. The average is calculated from the identification of 200 systems.

when a large value of α is adopted, the probability of unfalsification significantly decrease. The situation is illustrated by Figure 4.3, which refers to data set S1D2. The two plots show the average trends of $x_\varepsilon^{N,M}$ and the values of $\chi^2_\alpha(M)$ for different values of M. In the left plot, α is set to 0.2 and for all the considered values of M the condition (2.7) is verified in correspondence to a certain model order. In the right plot, α is equal to 0.8 and only the curve $x_\varepsilon^{N,N/4}$ crosses the line $\chi^2_\alpha(N/4)$, meaning that the condition (2.7) is verified only for $M = N/4$. According to these considerations, high values of M favour the unfalsification of at least one of the evaluated model structures, but they are not always beneficial in terms of the achieved impulse response fits, as Figure 4.4 proves.

Figure 4.4 highlights how the choice of M has a quite considerable impact on the performances achieved by the whiteness test when it is used as an order selection method. In all the four data sets here considered $M = 20$ leads to the best performances, when the significance level is properly tuned (see Section 4.1.3).

Test for independence between residuals and past inputs

The analysis of the influence of the maximal lag M on the test for the independence between residuals and past inputs is here based on the second implementation of the test considered in this project, i.e. the one denoted by RAI2.



(a) S1D1.

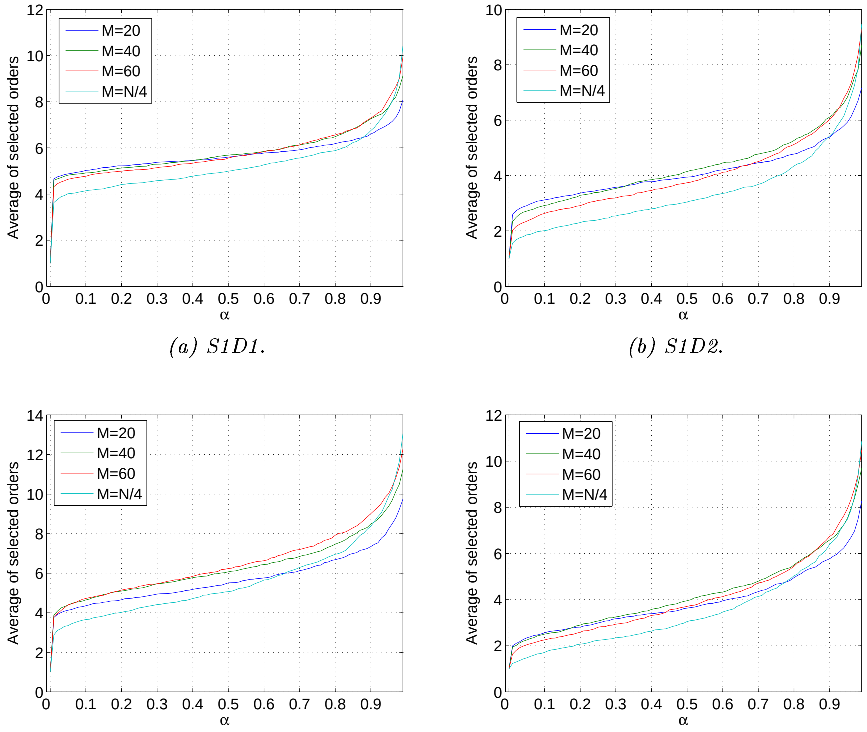
(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 4.4: OE models - Average impulse response fits achieved by the whiteness test on the residuals (RAW) as function of the significance level α and for different values of the maximal lag M for which the auto-correlation is computed. The average is calculated from the identification of 200 systems in each data set.

Figure 4.5 illustrates the average of the orders selected by this criterion as function of the significance level α and for different values of M. A specific trend can be observed in the plots relative to S1D1, S1D2 and S2D2: for small significance levels, small values of M lead to the selection of larger orders, while the situation inverts for large significance levels, since large M favour the selection of more complex models. Indeed, as previously observed for the whiteness test, it is more probable that the main correlation components are present for small lags. Therefore, for small significance levels the situation is analogous to the one described for the whiteness test. However, differently from what observed in that case, for large lags the cross-correlation between residuals and past inputs generally exhibits larger components than the auto-correlation of the residuals. Therefore, when the probability to unfalsify a certain model structure decreases, i.e. when large significance levels are adopted, more complex model structures



(a) S1D1.

(b) S1D2.

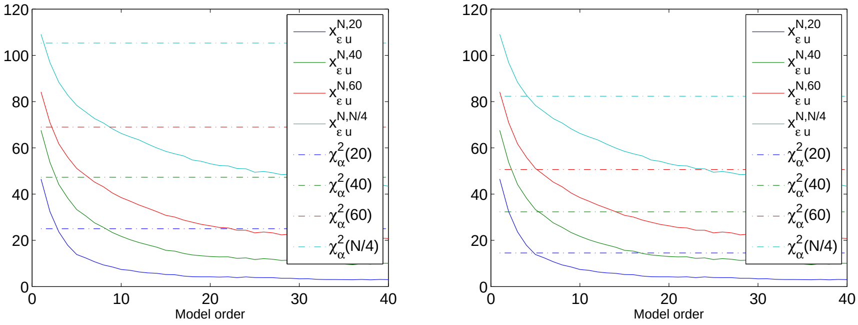
(c) S2D1.

(d) S2D2.

Figure 4.5: OE models - Average of the orders selected by the test for independence of the residuals from past inputs (RAI2) as function of the significance level α and for different values of the maximal lag M for which the cross-correlation is computed. The average is calculated from the identification of 200 systems in each data set.

are required in order to eliminate those correlation components.

The described situation is illustrated by Figure 4.6, which refers to data set S2D2. It shows the average trends of $x_{\varepsilon u}^{N,M}$ and the values of $\chi^2_\alpha(M)$ for different values of M. It can be noticed how $x_{\varepsilon u}^{N,20}$ encounters a faster decrease than $x_{\varepsilon u}^{N,60}$ for small model orders. Indeed, simple models can give rise to a considerable correlation between a residual $\varepsilon(t)$ and a past input $u(t-\tau)$ also for small τ. These correlation components can be considerably reduced by increasing a bit the model complexity. On the other hand, when correlation is present also for large values of τ, the model order should be further increased in order to reduce it. This explains the slow decrease of $x_{\varepsilon u}^{N,60}$ noticeable in Figure 4.6. When α is set to a high value, as in Figure 4.6.(b), the corresponding $\chi^2_\alpha(M)$ decreases and makes $x_{\varepsilon u}^{N,60}$ crossing $\chi^2_\alpha(60)$ in correspondence to larger orders than what is observed for $x_{\varepsilon u}^{N,20}$.



(a) $\alpha = 0.2$.

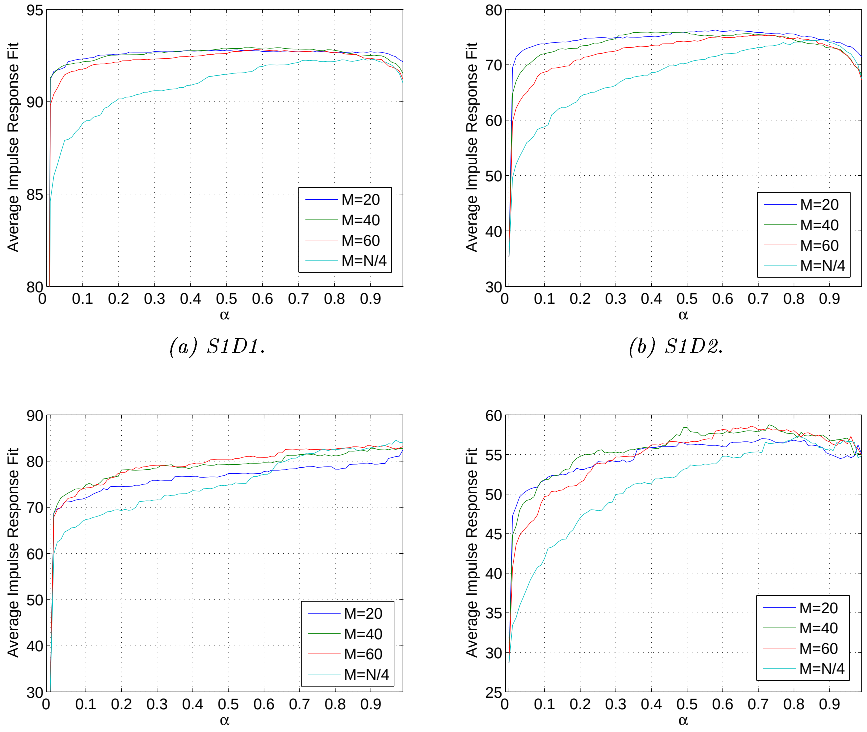
(b) $\alpha = 0.8$.

Figure 4.6: OE models, Data set S2D2 - Average values of $x_{\varepsilon u}^{N,M}$ (solid lines) and values of $\chi^2_\alpha(M)$ (dashed-dotted lines) for different values of the maximal lag M. The significance level α is fixed. The average is calculated from the identification of 200 systems.

The trends observed in Figure 4.5 for dataset S2D1 are different from the ones just described, since the average of the orders selected setting $M = 40$ or $M = 60$ is always greater than the one observed for $M = 20$. Because of the slow dynamics of the systems present in that data set, particularly large correlation components are exhibited also for large lags. When M is set equal to 20, these components are not taken into account in the value of $x_{\varepsilon u}^{N,M}$, making more probable the unfalsification of simple models. This makes the test less reliable, as it is confirmed by Figure 4.7, which shows how large values of M (but smaller than $N/4$) lead to better impulse response fits in the "slow" data sets. However, Figure 4.7 also illustrates how small values of M are preferable when fast systems have to identified, since they give rise to the highest fits and they also appear more robust to the tuning of the significance level.

Value of M used in the experimental tests

In the following, all the tests on the residuals adopt a maximal lag M equal to $n_b + 20$, being n_b the order of the polynomial $B(q)$ in (1.13). This choice allows to correctly use the independence test when it is applied on a model estimated by least-squares, such as a FIR or an ARX model. Since the model orders that are evaluated in the following tests range from 1 to 40, this means that M will range from 21 to 60. As shown by the previous analysis, this choice could penalize the performances of RAI1 or RAI2 in data sets S2D1 and S2D2, because it could lead to the selection of too simple models.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

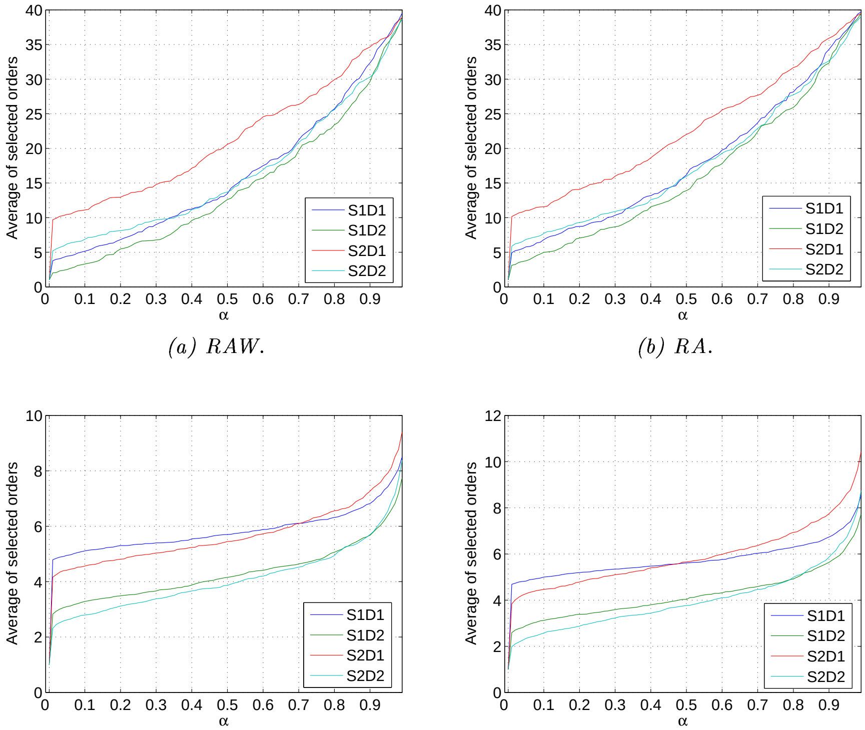
Figure 4.7: OE models - Average impulse response fits achieved by the test for independence of the residuals from past inputs (RAI2) as function of the significance level α and for different values of the maximal lag M for which the cross-correlation is computed. The average is calculated from the identification of 200 systems in each data set.

Moreover, Figures 4.4 and 4.7 have shown how the dependence of the performances of the tests from the significance level α is very similar for M ranging from 20 to 60. Therefore, the analysis that will be conducted in the next sections for the selection of α holds independently of the value of M.

4.1.3 Selection of the significance level α used in the statistical tests

A new analysis is carried out in order to define the optimal values of the significance level α in the statistical tests used for model order selection.

The investigations done in Section 4.1.2 have already shown how α significantly influences the order selection done by the tests on the residuals (namely RAW and RAI2).



(a) RAW.

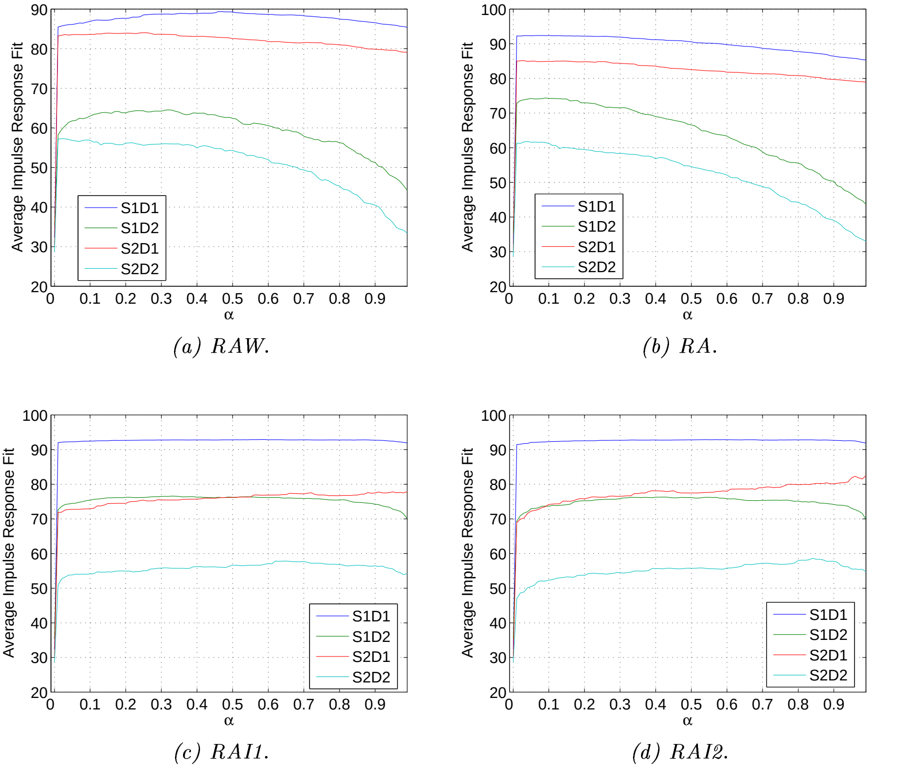
(b) RA.

(c) RAI1.

(d) RAI2.

Figure 4.8: OE models - Average of the orders selected by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

Figures 4.8 and 4.9 contain respectively the average of the selected orders and the average impulse response fits achieved for different values of α used in the statistical tests on the residuals ($0 \leqslant \alpha \leqslant 0.99$). Ideally, a small value of α leads to a minor probability of rejecting the null hypothesis H_0, meaning that simple model structures are more easily selected. The extreme situation in this sense is obtained when α is equal to 0, meaning that all the model structures will pass the test and order 1 will be always chosen, as it is the first one to be considered; all the plots in Figure 4.8 clearly show this phenomenon. Figure 4.9 illustrates that also the fit encounters a drastic drop when $\alpha = 0$, since model structures of order 1 in most cases are not suitable for an appropriate description of the



(a) RAW.

(b) RA.

(c) RAI1.

(d) RAI2.

Figure 4.9: OE models - Average impulse response fits achieved by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

underlying true system.

Figure 4.8 shows how the rate of growth of the average orders for increasing α is larger for the whiteness test (RAW) or for its combination with the independence test (RA). The trends here observed are actually analogous to the ones noticed in Figure 4.2. Indeed, as was previously explained, the increase of α makes more probable the default selection of model structures of order 40, since the whiteness test is not able to unfalsify any of the evaluated model structures. S2D1 appears as the data set which is most affected by this issue, since the systems it contains can be properly described by high-order OE models. In particular, for the identification of some systems, the problem is still present even if a very small value of α (such as 0.01) is chosen. Though data set S2D2 also contains slow systems, the described issue is less frequent because of the more relevant presence of noise in the measurement data, which can favour the whitening of the residuals. Indeed, the noise component in the residuals (the so-called *data error*) can be larger than the approximation component, leading to the whitening of the residuals.

Further inspecting the plots in Figure 4.8, it can be noticed how the average of the orders selected by the two independence tests RAI1 and RAI2 is considerably lower for the noisy datasets S1D2 and S2D2 w.r.t. the one reported for S1D1 and S2D1, thus showing that overfitting is avoided. An explanation of this behaviour can probably be found by looking at the explicit expression of the residuals; namely, let $B_0(q)$ and $F_0(q)$ be the polynomials which give the best description of the true system and let $\widehat{B}(q)$ and $\widehat{F}(q)$ be their estimates, then

$$\begin{aligned} \varepsilon(t)u(t-\tau) &= \left[\frac{B_0(q)}{F_0(q)}u(t) + e(t) - \frac{\widehat{B}(q)}{\widehat{F}(q)}u(t)\right]u(t-\tau) \\ &= \left[\frac{B_0(q)}{F_0(q)} - \frac{\widehat{B}(q)}{\widehat{F}(q)}\right]u(t)u(t-\tau) + e(t)u(t-\tau) \qquad (4.1) \end{aligned}$$

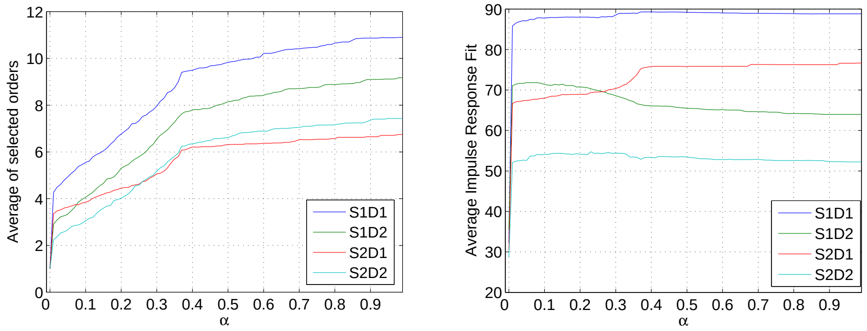
If the estimated model is very close to the true system, the first term in the summation (4.1) is small; in case of overfitting instead, $\frac{\widehat{B}(q)}{\widehat{F}(q)}$ would depart from $\frac{B_0(q)}{F_0(q)}$, thus increasing the first term in the summation and in turn the correlation component $\varepsilon(t)u(t-\tau)$.

It should also be noticed that for small values of α the average of the orders chosen for the slow systems in data sets S2D1 and S2D2 is lower than the corresponding one for the "fast" data sets S1D1 and S1D2, respectively. This is due to the choice of the maximal lag M considered for the cross-correlation between residuals and past inputs. As was previously noticed in Section 4.1.2, when slow systems have to be identified, there could be some relevant correlation between a residual $\varepsilon(t)$ and a past input $u(t-\tau)$, with τ particularly large. If the maximal lag M for which the cross-correlation is computed is too small, these correlation components could not be detected, leading to the premature unfalsification of too small model structures. Since the maximal lag considered for these tests depends on the complexity of the estimated model, simple model structures are more affected by this issue. When large values of α are adopted, the probability to unfalsify a certain model structure decreases, thus reducing the described risk.

Figure 4.9 provides further insight about the impact of α on the goodness of the estimated models. Comparing the plots relative to the whiteness test (RAW) and to its combination with the independence test (RA), different trends can be observed for the "fast" data sets. While for RA very small values of α lead to the best average impulse response fit, when the whiteness test is used alone (RAW)

larger values of α lead to better fits. Indeed, lacking of the equipment with the independence tests on the residuals, RAW tends to select a bit smaller orders than RA, thus needing a larger significance level to compensate this lack. This difference between RAW and RA is more evident for S1D1 and S1D2, because in the case of slow systems it is alleviated by the default selection of order 40, done whenever none of the evaluated model structures can be unfalsified. Indeed, the optimal significance values for the "slow" data sets are very low, in order to increase the probability to unfalsify at least a model structure.
It should also be observed that a unique significance level is used for both the whiteness and the independence test in RA. Figure 4.9 shows that this is not the optimal choice and that two different values of the significance level could lead to better performances. Namely, small values of α are preferable for RAW, while larger ones work better for RAI1.

For what regards the independence tests, similar fits can be achieved by the two implementations considered. The significance level does not seem so influential for the identification of fast systems, even if too small and too large values should be avoided, since they respectively lead to the selection of too small or too complex model complexities. Different considerations hold for data sets S2D1 and S2D2, for which large significance levels are more indicated in order to reduce the undermodelling tendency due to a non proper choice of the maximal lag M (as observed in Section 4.1.2).



(a) Average of the selected orders.

(b) Average impulse response fits.

Figure 4.10: OE models - Average of the selected orders and average impulse response fits achieved by the F-test, as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

Figure 4.10 shows the average of the orders selected by the F-test and the average impulse response fits that are obtained for different values of the significance level α. For all the data sets $\tilde{\alpha} \approx 0.38$ seems to be a critical value, beyond which only a minor increase of the model complexities can be reached. The average fits

achieved are in turn influenced by this value, showing a step in correspondence of $\tilde{\alpha}$. Figure 4.10 shows that too large values of α lead also to overfitting in case of noisy measurements. Therefore, the adoption of significance levels lower than 0.5 is suggested for noisy data sets, while larger α are more suitable for non-noisy data sets.
It should be observed that relative small orders are chosen by the F-test for the data set S2D1, when also compared to the choices done for the other data sets. Since the F-test is based on the loss function, a better understanding of this result can be obtained by looking at the average values assumed by the loss. After a rapid decrease observed for very small orders, the decrease in data set S2D1 is much less significant than the one observed for the other data sets, especially w.r.t. the noisy ones, where the reduction is mainly due to the adaptation to the noise realization. Since for S2D1 no relevant difference is detected between the loss of two model structures with similar orders, the F-test will choose the simplest one, thus explaining the trend observed in Figure 4.10.(a).

4.1.4 Analysis of the order selection methods

The analysis of the different order selection methods is presented by means of the box plots of the impulse response and type 1 prediction fits (Figures 4.11 and 4.12) and of the histograms of the orders chosen by each criterion (Figures 4.13, 4.15, 4.17 and 4.19). The histograms of the differences between the orders chosen by the oracle for impulse response fit and the ones selected by the criteria are also reported (Figures 4.14, 4.16, 4.18 and 4.20).

The results relative to the criteria which involve statistical tests are obtained with the significance level α which guarantees the best average impulse response fits. Table 4.1 summarizes the values of α that are used in the tests giving the results presented in the following. Except for RA, which benefits when low values of α are used, the optimal significance level for the other statistical tests significantly varies from a data set to the other one. However, in the case of the independence tests, the value adopted for the significance level is not so influential, as was observed analyzing Figures 4.9.(c) and (d); moreover, the large values of α appearing in the table for data set S2D1 are chosen in order to remedy the non-optimal choice of the maximal lag M.

Tables 4.2, 4.3 and 4.4 summarize the average fits reached by each of the model order selection criteria in the considered setting. As was previously observed during the analysis of Figure 4.1, all the three types of fit assume very similar values, showing consistency between the system description ability and the prediction one. The largest discrepancy among the three types of fits is observed

	α				
Dataset	*RAW*	*RA*	*RAI1*	*RAI2*	*F*
S1D1	0.46	0.07	0.58	0.54	0.39
S1D2	0.32	0.09	0.33	0.43	0.05
S2D1	0.25	0.03	0.99	0.99	0.99
S2D2	0.02	0.04	0.62	0.84	0.23

Table 4.1: OE models - Values of the significance level α which guarantee the best average impulse response fits when adopted in the statistical tests used for model order selection.

Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	93.9	89.4	92.9	92.9	92.4	89.3	92.4	84.6	84.7	93.1
S1D2	79.9	64.6	76.6	76.3	74.3	71.8	73.6	38.4	38.4	71.9
S2D1	89.9	84	77.8	82.5	85.1	76.7	79.5	80.7	80.8	87.6
S2D2	69.9	57.3	57.8	58.6	61.8	54.7	55.5	30.5	30.5	62.9

Table 4.2: OE models - Average impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	94.5	90.1	93.7	93.7	93.2	90.4	93.2	84.2	84.6	93.8
S1D2	81.8	67.4	78.9	78.7	76.8	74.6	76.5	39.4	39.7	74.8
S2D1	91.6	87.7	82.5	85.6	88.8	81.5	84.3	83.6	83.9	90.8
S2D2	75.7	66	66.6	67.1	70.4	64.6	64.3	39.5	40.1	70.5

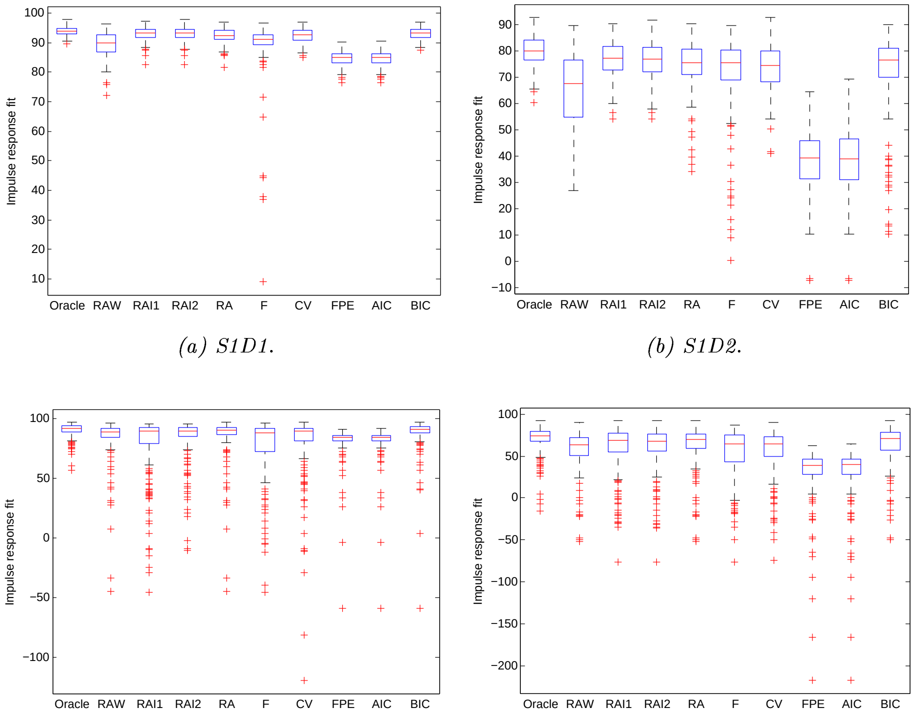
Table 4.3: OE models - Average type 1 prediction fits achieved by the evaluated criteria when 200 systems are identified in each data set.

Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	94.3	89	93.4	93.4	92.8	89.5	92.8	80.4	80.8	93.4
S1D2	81.2	64.6	78.2	78	75.4	73.5	75.3	20.2	20.2	71.4
S2D1	90	84.4	80	83	86	78.5	81.2	76.6	77.2	87.7
S2D2	72.8	60.2	62.8	63.2	64.4	60.6	59.6	18.4	18.8	64.9

Table 4.4: OE models - Average type 2 prediction fits achieved by the evaluated criteria when 200 systems are identified in each data set.

in data set S2D2, where the choice of quite small orders to avoid overfitting limits the proper description of the true system but favours the generalization capability.

The performances achieved by the oracle give an indication of the adequateness of OE models in the description of the true systems present in the data sets considered. As expected, the fits decrease in the noisy data sets S1D2 and S2D2 with respect to the corresponding less noisy data sets S1D1 and S2D1. On the other hand, Table 4.2 also shows a certain difference between the average fits



(a) S1D1.

(b) S1D2.

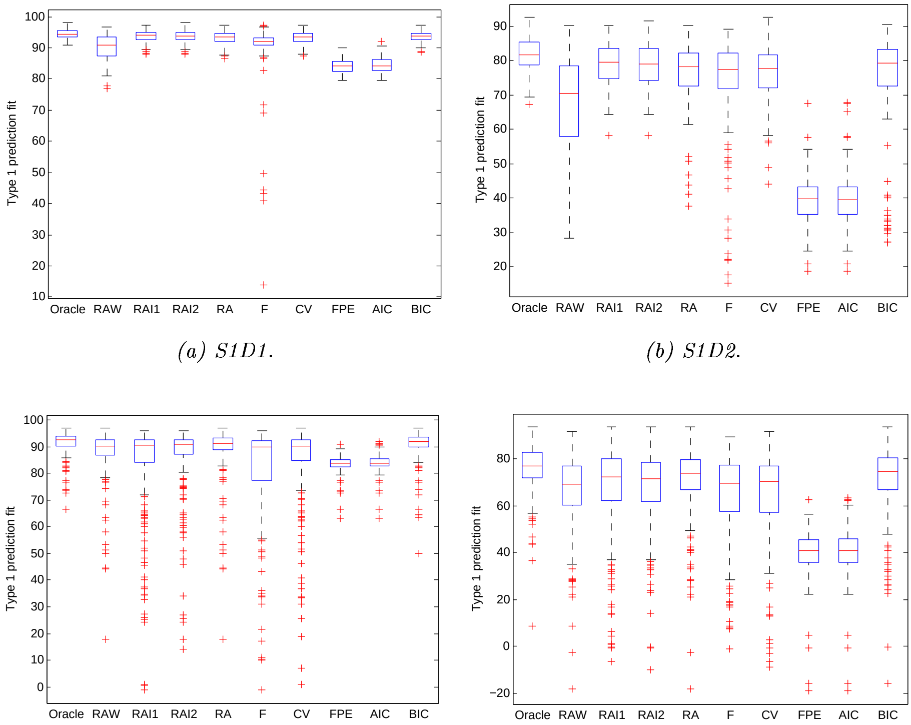
(c) S2D1.

(d) S2D2.

Figure 4.11: OE models - Box plots of the impulse response fits achieved by the analyzed criteria when 200 systems are identified in each data set.

achieved by oracle in the "fast" and "slow" data sets. Moreover, the plots in Figure 4.11 and 4.12 relative to the "slow" data sets present a large number of outliers, which correspond to systems with all zeros and poles placed very close to the unit circle and to the positive real axis or, generally with overlapping zeros and poles. The identification of these systems appears particularly difficult: for them the oracle tends to choose quite complex model structures in data set S2D1, while it usually selects low orders in S2D2. However, in both cases, the other criteria choose even lower complexities, which result in lower fits and in the longer tails that are visible in the box plots.

Analyzing the performances of the various order selection criteria, the BIC criterion appears the most effective one, except in data set S1D2, where the independence test on the residuals lead to the highest average fits. Among the other criteria also the combination of the whiteness and the independence test on the



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 4.12: OE models - Box plots of the type 1 prediction fits achieved by the analyzed criteria when 200 systems are identified in each data set.

residuals (RA) is effective in all the data sets.

Starting from the whiteness test on the residuals a more detailed analysis of the results obtained by each criterion is now conducted.

In all the data sets RAW gives rise to lower fits than the ones reached by RA, with the minor gap observed in S2D1. The reason of this discrepancy lies in the tendency to undermodel that RAW shows when a too small value of the significance level α is adopted for the test. However, increasing its value, the probability to unfalsify at least a model structure decreases, leading to the default selection of order 40, which is not appropriate in case of noisy data. On the other hand RA benefits from the combination with the independence test, since it limits the undermodelling tendency, even if a small value for α is adopted. In particular, also the histograms in Figures 4.14, 4.16, 4.18 and 4.20

show a significant agreement between the oracle choices and the ones done by RA.

The two implementations of the independence test on the residuals, RAI1 and RAI2, lead to very similar performances, especially in the "fast" data sets, while in the "slow" ones better performances are achieved by RAI2. This behaviour can be justified by the high number of model structures that are unfalsified by RAI1, even if the whiteness test is not passed. Therefore, in these cases the reliability of RAI1 is limited.
Thanks to the general tendency to select quite low orders, the independence test on the residuals provides very good performances on the "fast" data sets (being the most effective criterion in S1D2), but encounters some difficulties in the complexity estimation when slow systems have to be identified. This behaviour could be in part explained by the adoption of a too small value for the maximal lag M until which the cross-correlation between residuals and past inputs is computed. Indeed, the tests in Section 4.1.2 have shown that a larger M improves the performances in data sets S2D1 and S2D2.

The F-test suffers of undermodelling when applied for the identification of slow systems, leading to the unsatisfying fits detected in S2D2 and especially in S2D1. However, also the box-plots relative to data sets S1D1 and S1D2 in Figures 4.11 and 4.12 present a significant number of outliers. These arise because of the selection of too low orders as a consequence of local minima solutions. Indeed, whenever a local minimum is returned by the estimation procedure, an increase in the loss function is detected w.r.t. the value assumed with simpler complexities. When the F-test encounters these situations, the procedure for order selection is early stopped because the test is passed by the smallest model structure in the comparison.

Cross-validation seems to be more effective when fast systems have to be identified; indeed the fits observed for S2D1 and S2D2 are quite low compared to the best criteria for those data sets. Figures 4.11 and 4.12 illustrate the presence of many outliers for cross-validation in data sets S2D1 and S2D2. They probably correspond to very slow systems, which can be appropriately identified only when a large amount of measurement data is available. Since cross-validation estimates a temporary model only on half of the available data, this could be quite different from the one estimated using all the data, leading to an imprecise estimation of the prediction errors and in turn of the optimal complexity.

FPE and AIC criteria, giving almost equivalent fits, appear as the worst methods in all the data sets, except in S2D1. The reason of the unsatisfying performances is the marked overmodelling tendency that they present (evident in Figures 4.14, 4.16, 4.18 and 4.20) which is beneficial only in data set S2D1. This result is in line with the consistency analysis done in Section 2.2.5, which has shown how

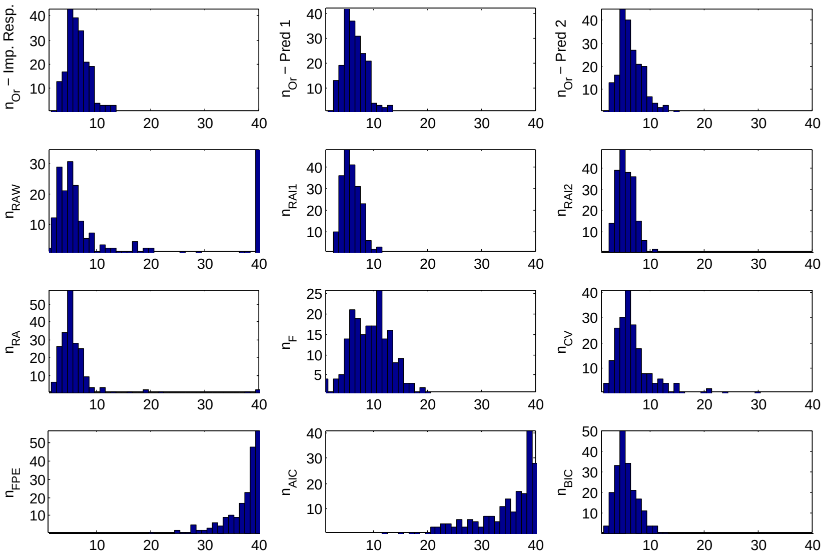


Figure 4.13: OE models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S1D1.

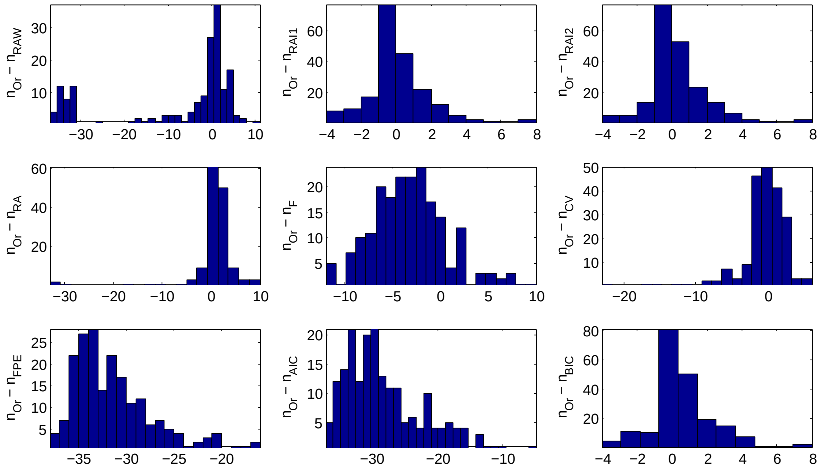


Figure 4.14: OE models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S1D1.

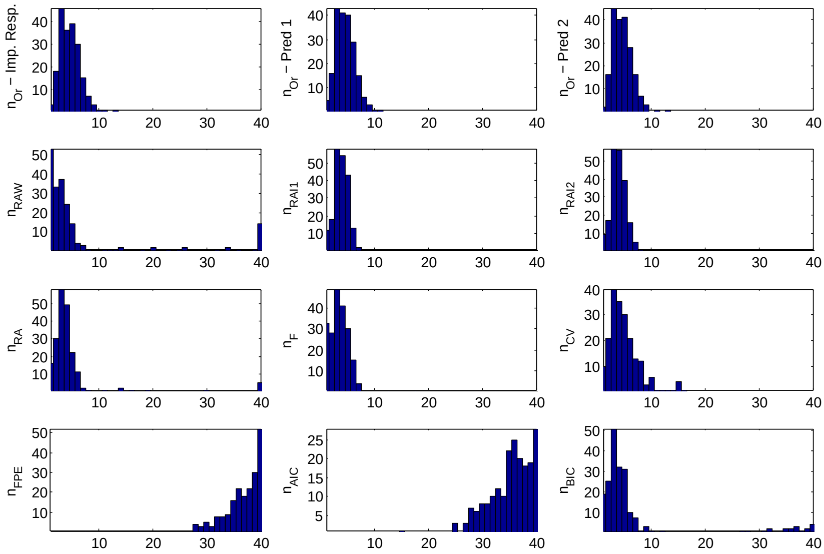


Figure 4.15: OE models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S1D2.

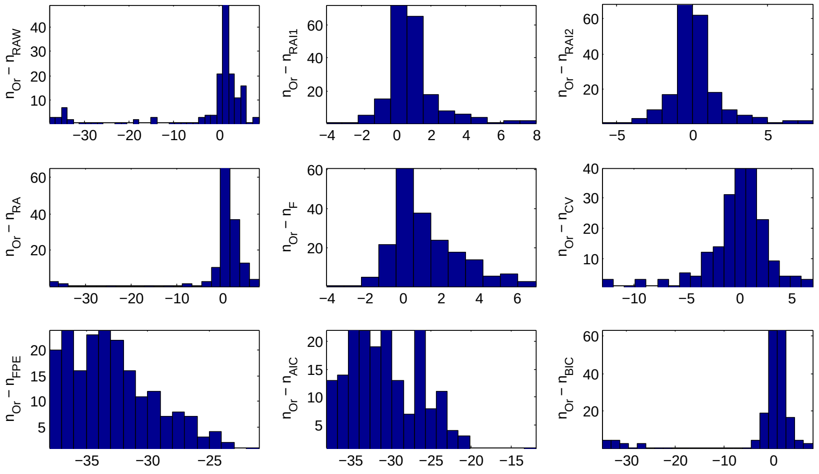


Figure 4.16: OE models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S1D2.

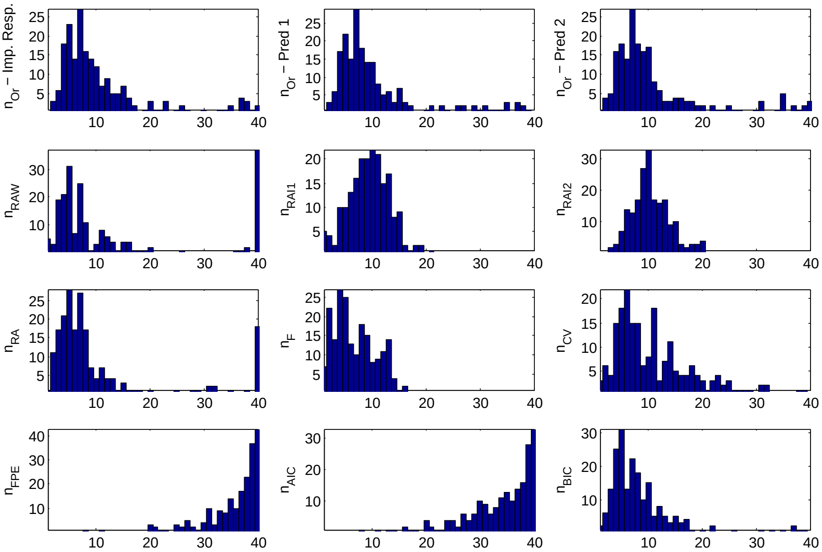


Figure 4.17: OE models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S2D1.

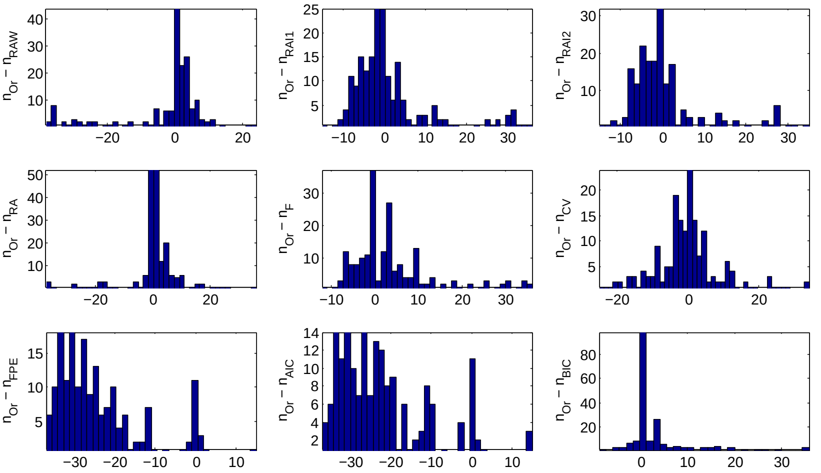


Figure 4.18: OE models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S2D1.

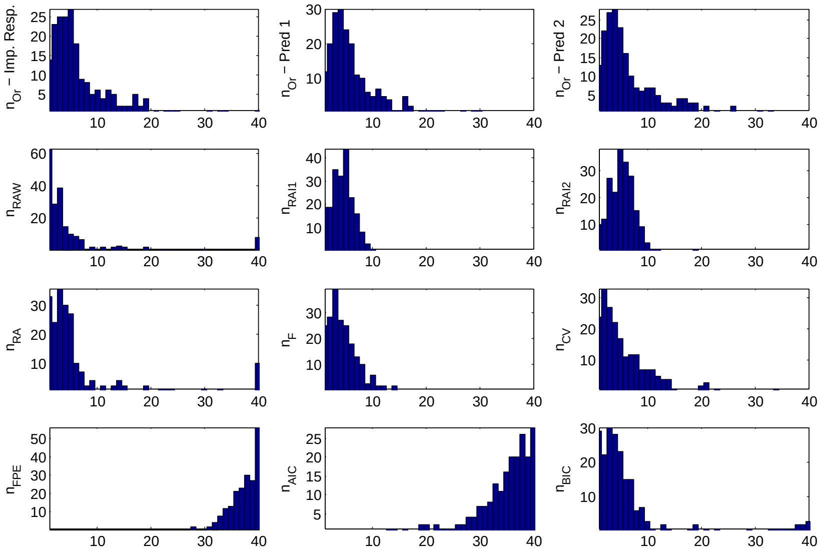


Figure 4.19: OE models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S2D2.

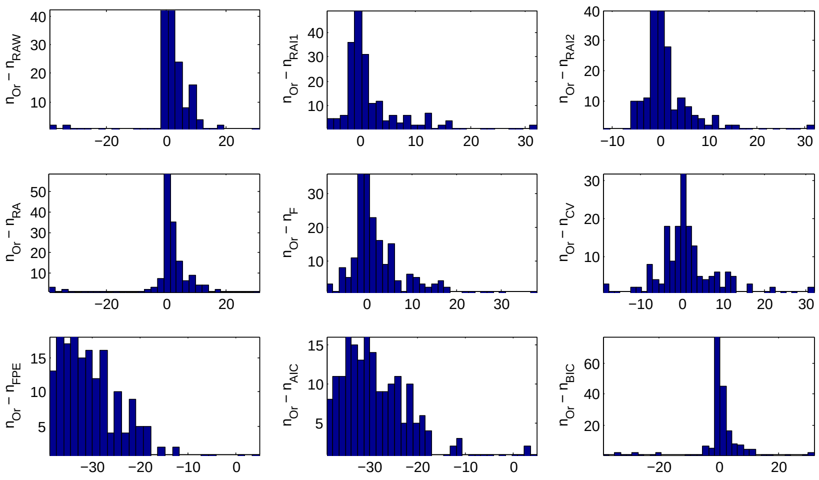


Figure 4.20: OE models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S2D2.

FPE and AIC criteria are not able to avoid overmodelling even when an infinite amount of data is available.
These findings suggest that the penalty inflicted by the two criteria on model complexity is not sufficiently large to avoid the selection of too complex models.

As previously observed, the BIC criterion allows to achieve the highest fits in all the data sets, except in S1D2, where it shows an overfitting tendency. While the penalization on high complexities is effective on the less noisy data sets S1D1 and S2D1, it is not sufficiently large when noisy data are used, resulting in the more frequent selection of too high orders. If this phenomenon is not so detrimental for S2D2, because of the slow nature of the systems that have to be identified, it appears problematic in S1D2, as the average fit in Table 4.2 and the box plot in Figure 4.11 prove. This overfitting tendency is present only when noisy data are exploited, because in these cases the loss function encounters a significant decrease for large orders, thanks to the adaptation to the specific noise realization present in the data. Therefore, a larger penalty on high complexities is needed in order to counterbalance the loss reduction.

4.1.5 Combination of the comparison and validation methods

The order estimation of OE models is now performed by combining the comparison methods (F-test, cross-validation and the information criteria) with the validation procedures, according to the way described in Section 2.3.1. The results are still based on the identification of 200 systems from each of the four available data sets. The statistical tests on the residuals evaluate the auto- and cross-correlation until a maximal lag M again equal to $n_b + 20$, being n_b the order of the $B(q)$ polynomial.

F-test

The analysis done in the previous section highlighted how local minima lead to a premature stopping of the procedure exploited by the F-test for model order selection. This issue significantly penalizes the performances of the F-test. Therefore, its equipment with the validation methods could help to avoid this issue.
This combination leads to the final selection of the first model structure which passes both the F-test and the considered test on the residuals.

Table 4.5 summarizes the average impulse response fits that are achieved by the F-test alone and by its combination with one or both the tests on the residuals.

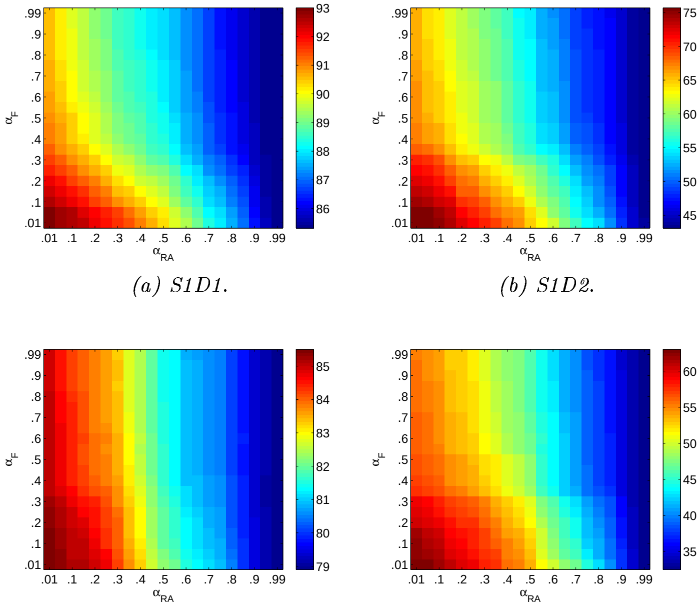
The results are obtained choosing for each data set and for each statistical test the significance level leading to the highest average impulse response fit.
In all the data sets the equipment of the F-test with a validation procedure leads to an improvement of the average fits: while the combination with the independence test (RAI1 or RAI2) appears more effective on "fast" data sets, the adoption of both the tests on the residuals (RA) leads to better results in S2D1 and S2D2. These performances are in line with the ones previously observed for the application of the residuals tests alone.
It should also be observed that, except for data set S2D1, the average fits appearing in Table 4.5 are comparable or even better than the best ones present in Table 4.2.

A clarification should be given with regards to the average of the selected orders for data set S1D1 that are shown in Table 4.5. Theoretically, the combination of the F-test with a validation method should lead to the choice of higher orders; however, the values in the table show a decrease, which is justified by the adoption of different significance levels α_F for the F-test.

Set	Average impulse response fit					Average of the selected orders				
	F	*F + RAW*	*F + RA*	*F + RAI1*	*F + RAI2*	*F*	*F + RAW*	*F + RA*	*F + RAI1*	*F + RAI2*
S1D1	89.3	92.3	93	93	93.1	9.5	5.6	5.4	5.7	5.9
S1D2	71.8	74.2	75.7	76.6	76.5	3.4	4.5	4.4	3.7	3.8
S2D1	76.7	85.2	85.5	81.6	84.5	6.4	10.7	10.5	12.4	10.4
S2D2	54.7	61.6	63	61.1	60.4	4.9	6.4	6.2	4.5	4.4

Table 4.5: OE models - Average impulse response fits and average of the selected orders when validation methods are combined with the F-test for model order selection in the identification of 200 systems in each data set.

Since statistical tests are here involved, a specific analysis should be devoted to the choice of the significance levels to adopt. For this purpose Figure 4.21 shows the average impulse response fits that are achievable for different values of the significance levels α_F and α_{RA} respectively adopted in the F-test and in both the tests on the residuals (RA), when they are combined as described in Section 2.3.1. A significant match can be observed among the four data sets for what regards the dependence between the fits and the value of α_F and α_{RA}. Namely, particularly small values should be adopted for both the significance levels. The optimal choices here suggested for α_{RA} agree with the ones indicated by Figure 4.9.(*b*), when RA is applied alone as an order selection method. On the other hand, comparing Figures 4.10.(*b*) and 4.21, a slight disagreement is detected in the optimal values of α_F suggested in data sets S1D1 and S2D1; indeed, when the F-test is combined with a validation method, a large value of α_F should be avoided, since it could lead to overmodelling.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 4.21: OE models - Average impulse response fits achieved for different values of the significance levels when both the tests on residuals (RA) are combined with the F-test for model order selection in the identification of 200 systems in each data set.

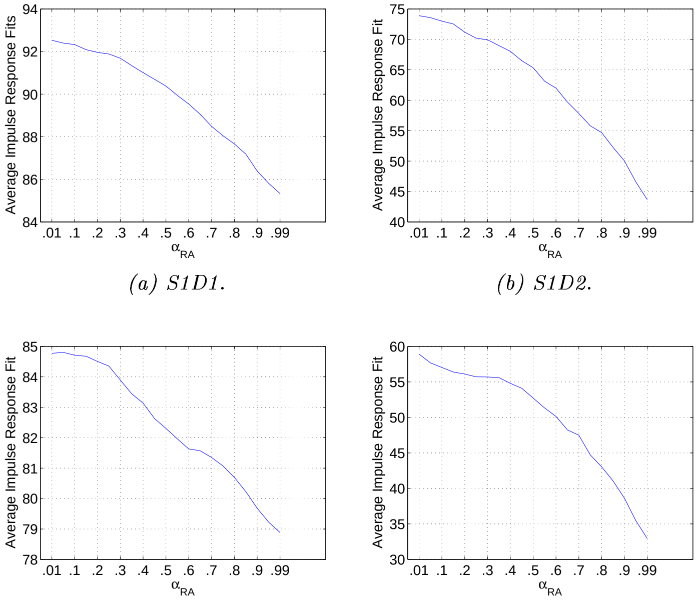
Cross-validation

The equipment of cross-validation with the validation methods according to the way described in Section 2.3.1 is now analyzed. The average impulse response fits present in Table 4.6 again prove the efficacy of the combination and confirm what was detected with the F-test: namely, the independence test on the residuals, RAI1 and RAI2, are more effective on S1D1 and S1D2, while performing both the tests on the residuals (RA) is more indicated when slow systems have to be identified (S2D1 and S2D2). The values in Table 4.6 are still obtained setting the significance levels of the statistical tests to the values leading to the highest impulse response fits.
In this application the performances improvement is guaranteed by an increase of the complexities chosen for the estimated models. Indeed, whenever the model structure returned by cross-validation is not unfalsified by the adopted validation method, the newly chosen model structure is the one leading to the second lowest value of the normed prediction errors. When the average of this

function over the 200 measurement sets considered in each data set is plotted vs. model complexity, one observes a rapid initial decrease, followed by a significant growing trend for high orders. Therefore, the newly chosen model structures are generally more complex than the original ones returned by cross-validation.

Set	Average impulse response fit					Average of the selected orders				
	CV	$CV+RAW$	$CV+RA$	$CV+RAI1$	$CV+RAI2$	CV	$CV+RAW$	$CV+RA$	$CV+RAI1$	$CV+RAI2$
S1D1	92.4	92.4	92.5	92.6	92.6	7	7.1	7.2	7.3	7.5
S1D2	73.6	73.4	73.9	74.5	74.5	4.9	5.3	5.3	5.3	5.3
S2D1	79.5	84.8	84.8	82.3	82.7	11.1	14.5	14.9	13.1	12.4
S2D2	55.5	58.6	59.9	56.9	56.8	5.7	9	9.4	8.2	7.3

Table 4.6: OE models - Average impulse response fits and average of the selected orders when validation methods are combined with cross-validation for model order selection in the identification of 200 systems in each data set.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

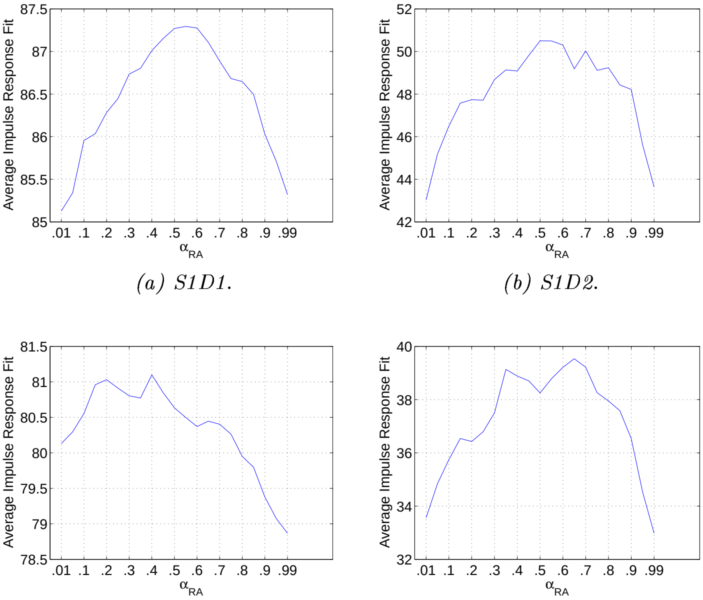
Figure 4.22: OE models - Average impulse response fits achieved for different values of the significance level α_{RA} used in both the tests on residuals (RA) when they are combined with cross-validation for model order selection in the identification of 200 systems in each data set.

Figure 4.22 illustrates the average impulse response fits achieved in the four data sets by the combination of cross-validation with both the tests on the residuals

(RA). The fits are plotted as functions of the significance level α_{RA} adopted in both the statistical tests. The trends visible in Figure 4.22 are in line with the ones reported in Figure 4.9.(b), since low values of the significance level α_{RA} lead to the highest fits. It should be observed that a wrong choice of α_{RA} can lead to a decrease of the fit originally obtained by cross-validation alone.

FPE

The performances achieved by the FPE criterion on OE models can be improved only by combining it with the whiteness test (RAW) or by both the tests on the residuals (RA). The independence tests alone are not effective in this context, since the model structures returned by FPE are all unfalsified by that test, thanks to the high complexity that they have.



(a) S1D1. *(b) S1D2.*

(c) S2D1. *(d) S2D2.*

Figure 4.23: OE models - Average impulse response fits achieved for different values of the significance level α_{RA} used in both the tests on the residuals (RA) when they are combined with FPE for model order selection in the identification of 200 systems in each data set.

Figure 4.23 shows the average impulse response fits that are achievable by combining the FPE criterion with RA as function of the significance level α_{RA}

adopted in the two tests. Compared to the values relative to FPE and reported in Table 4.2, an increase in the average impulse response fit is detected, independently from the value chosen for α_{RA}. The only exception in this sense is observed for data set S2D1, where the combination is not as beneficial as in the other data sets, because the performances obtained by the FPE criterion in these data set were already quite good.
Differently from what observed for the F-test and the cross-validation, here the most suitable values for the significance level seem to range around 0.5. Indeed, these values allow to select new model structures, whose complexities are lower than the ones originally returned by the FPE criterion. If too large values of α_{RA} are adopted, the probability to unfalsify at least one of the evaluated model structures significantly decrease, leading to the default selection of too complex models, which are not beneficial in terms of fit.

It should be remarked that even if the combination here described gives rise to a considerable improvement of the performances achieved by FPE alone, the final average fits are still lower than the ones obtained with the other criteria, especially in the noisy data sets.

AIC

When the AIC criterion is combined with the validation methods, analogous results to the ones described for FPE are obtained. This conformity is a direct consequence of the analogies detected between the two criteria in the analysis in Section 4.1.4.

BIC

The combination of the BIC criterion with a validation method does not lead to significant improvements but can also give rise to a performance worsening if a wrong significance level is adopted in the statistical tests on the residuals. The reasons for this lack of efficacy are different for the two statistical tests applied on the residuals. Indeed, the independence tests (RAI1 and RAI2) with a low significance level generally unfalsify the model structures returned by BIC, thus confirming its choice. On the other hand, the whiteness test does not tend to unfalsify the model structures chosen by BIC. In those cases, the consequent choice of more complex model structures is not beneficial. However, when a very small significance level is adopted, the previous results obtained by BIC are confirmed, since the unfalsification probability is quite high. The only exception to this behaviour is detected in data set S1D2, where BIC tends to overmodel, as was observed in Section 4.1.4. In this case, its combination with RAW or RA with low significance levels leads to an average impulse response fit equal to 75, thanks to the choice of simpler complexities. Indeed, in the noisy data set S1D2, it happens that the loss function (on which BIC is based) assumes very low values in correspondence to complex model structures, because of the adaptation to the noise realization. However, these models don't give rise to good reproductions of the true system. In some cases, this issue can be detected by the whiteness test, because a correlation between the noise and the estimated model is more easily detectable.

Conclusions

The combination between a comparison method and a validation one has proved to be useful especially with the F-test and with cross-validation. FPE and AIC also benefit from it but the performances achieved with them are still unsatisfying.

Among the tests on the residuals to be exploited in this context, the independence test has proved more effective when fast systems have to be identified, while the use of both whiteness and independence test (RA) has led to better results for the identification of slow systems. These findings are in line with the performances obtained by the tests on the residuals when applied alone for model order selection.

For what regards the significance level to be adopted in the statistical tests on the residuals, the analysis performed has shown a quite clear analogy between the values to use in this context and the ones used when the tests are exploited alone for model order selection.

4.1.6 Combination with the F-test

The combination illustrated in Section 2.3.2 between a model order selection method and the F-test is now evaluated for the order estimation of OE models.

Whiteness test on the residuals (RAW)

In the analysis conducted in Section 4.1.4 the whiteness test has shown an undermodelling tendency, whenever it was able to unfalsify at least one of the evaluated model structures. This trend could be alleviated by increasing the significance level α_{RAW} of the test, but this in turn would increase the rate of model structures that can not be unfalsified. Therefore, the application of the F-test after the whiteness test could be effective to alleviate this undermodelling property.

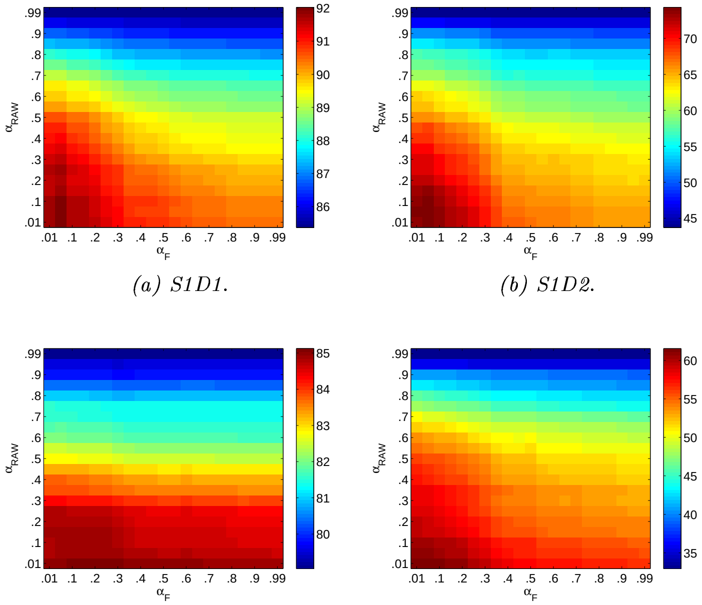
	Average impulse response fit		Average of the selected orders	
Set	*RAW*	*RAW + F (Up)*	*RAW*	*RAW + F (Up)*
S1D1	89.4	92	12.5	5.5
S1D2	64.6	74.3	7	4.1
S2D1	84	85.1	13.8	11.2
S2D2	52.3	61.5	5.5	6.1

Table 4.7: OE models - Average impulse response fits and average of the selected orders when the F-test is applied after the whiteness test (RAW) to perform model order selection in the identification of 200 systems in each data set.

Table 4.7 shows the improvement that can be achieved when the F-test is used to evaluate model structures with increasing complexities, starting from the one returned by RAW. The results are obtained choosing the significance levels α_{RAW} and α_F which guarantee the best average impulse response fits.
The table illustrates that the application of the F-test leads to a decrease of the average of the selected orders: this is due to the use of different significance levels α_{RAW} when the whiteness test is used alone and when it is combined with the F-test. More precisely, as suggested by Figure 4.24, small significance levels are adopted for RAW, thus increasing the probability to unfalsify at least a model structure. In this way, the default selection of models of order 40 is significantly reduced, explaining the decrease previously noticed in the average of the chosen orders (Table 4.7). The consecutive application of the F-test allows to refine the "rough" order selection done by the whiteness test. Figure 4.24 also illustrates that values of α_F lower than 0.5 are more indicated for the F-test, otherwise too complex models are chosen. The only exception in this sense is detected in data set S2D1, where large values of α_F are also indicated in order to favour the selection of more complex models.

Whiteness and independence test on the residuals (RA)

The application of both the tests on the residuals for model order selection has proved to be particularly effective on OE models. The addition of the independence test allows a reduction of the undermodelling trend characterizing RAW. However, performing the F-test after RA to compare model structures of increasing complexities can further improve the performances achieved by RA alone, as Table 4.8 proves. Again, the reduction noticed in the average of the achieved fits after the use of the F-test is due to the adoption of different significance levels for RA.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 4.24: OE models - Average impulse response fits achieved for different values of the significance levels when the F-test is applied after the whiteness test on the residuals (RAW) in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

	Average impulse response fit		Average of the selected orders	
Set	*RA*	*RA + F (Up)*	*RA*	*RA + F (Up)*
S1D1	92.4	93	6	5.4
S1D2	74.3	75.8	4.7	4.2
S2D1	85.1	85.5	10.7	10.7
S2D2	61.8	62.8	6.5	6.2

Table 4.8: OE models - Average impulse response fits and average of the selected orders when the F-test is applied after both whiteness and independence tests on the residuals (RA) to perform model order selection in the identification of 200 systems in each data set.

The dependence between the average impulse response fits and the significance levels adopted in the statistical tests is analogous to the one reported in Figure 4.24; therefore, the suggestions coming from Figure 4.9.(*b*) regarding the significance level to be used for RA are confirmed by the trends seen in Figure 4.24.

Independence test on the residuals (RAI1 and RAI2)

The analysis done in Section 4.1.4 has shown that the independence test on the residuals leads to very good fits when fast systems have to be identified; however, some difficulties have been detected when slow systems are estimated. Therefore, it could be expected that RAI1 or RAI2 benefit of their combination with the F-test especially in S2D1 and S2D2. This is actually confirmed by Table 4.9, which shows how the application of the F-test after the independence one for the comparison of more complex model structures leads to a more relevant increase in the performances obtained for the "slow" data sets. Again, the values in Table 4.9 are achieved setting the significance levels to the values leading to the highest average impulse response fits.

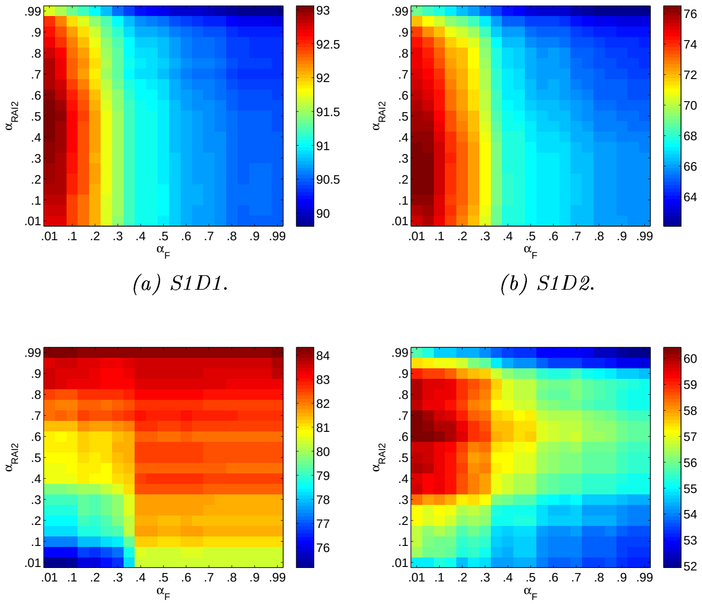
Set	Average impulse response fit				Average of the selected orders			
	RAI1	*RAI2*	*RAI1 + F (Up)*	*RAI2 + F (Up)*	*RAI1*	*RAI2*	*RAI1 + F (Up)*	*RAI2 + F (Up)*
S1D1	92.9	92.9	93	93.1	5.8	5.6	6.1	5.9
S1D2	76.6	76.3	76.6	76.5	3.7	3.9	3.8	3.8
S2D1	77.8	82.5	81.5	84.3	9.4	10.4	11	10.7
S2D2	57.8	58.6	60.8	60.4	4.3	5.3	4.4	4.5

Table 4.9: OE models - Average impulse response fits and average of the selected orders when the F-test is applied after the independence test on the residuals (RAI1 or RAI2) to perform model order selection in the identification of 200 systems in each data set.

The dependence between the average impulse response fits and the significance levels adopted in the tests is shown in Figure 4.25 for the second implementation of the independence test here used, RAI2, but analogous plots can be observed for RAI1. The fact that different trends are noticed in the four data sets is a consequence of the non-uniform trends noticed also in Figures 4.9.(c) and (d) for the four data sets. It should be observed that the highest fits in Figures 4.25 are obtained setting the significance level α of RAI2, α_{RAI2}, to the values suggested also by Figure 4.9.(d). For what regards the F-test instead, small values are preferable, since there is no need to test too complex model structures. Again, the exception is given by data set S2D1, for which complex model structures are suitable.

Cross-validation

Figures 4.14, 4.16, 4.18 and 4.20 do not highlight a clear overmodelling or undermodelling tendency for cross-validation; however, in the "slow" data sets underfitting is more evident. This suggests the application of the F-test after



(a) S1D1.

(b) S1D2.

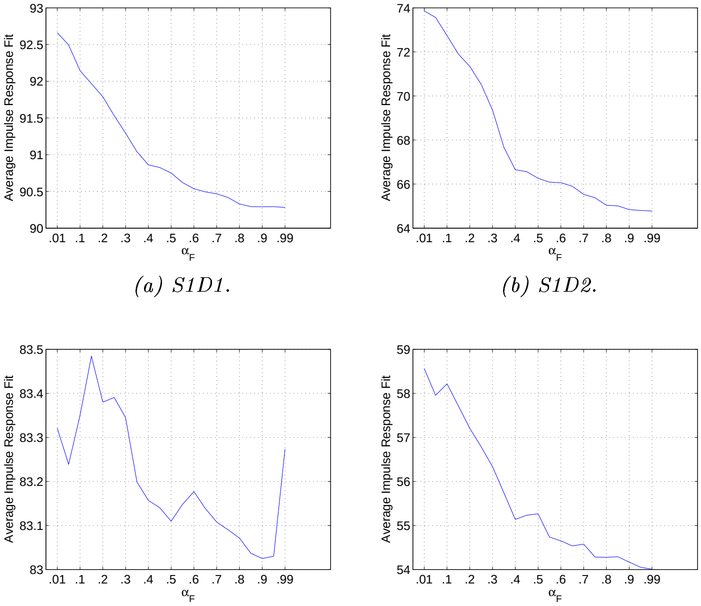
(c) S2D1.

(d) S2D2.

Figure 4.25: OE models - Average impulse response fits achieved for different values of the significance levels when the F-test is applied after the independence test on the residuals (RAI2) in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

cross-validation to compare model structures with increasing complexities. Figure 4.26 shows the dependency between the average impulse response fits achievable thanks to this combination and the significance level α_F adopted in the F-test. Small values of the significance level are preferable, since when α_F is too high too complex model structures are selected. However, while for S1D1 and S1D2 too large values of α_F lead to a significant decrease of the performances previously achieved by cross-validation, S2D1 and S2D2 are less affected by a wrong choice of α_F. This is a consequence of the major undermodelling tendency noticed in the "slow" data sets.

Table 4.10 contains the highest average impulse response fits that are achievable appropriately setting α_F; the values confirm that the combination is more effective on data sets S2D1 and S2D2.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 4.26: OE models - Average impulse response fits achieved for different values of the significance level α_F used in the F-test when it is applied after cross-validation in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

	Average impulse response fit		Average of the selected orders	
Set	*CV*	*CV + F (Up)*	*CV*	*CV + F (Up)*
S1D1	92.4	92.7	7	7.4
S1D2	73.6	73.9	4.9	5.3
S2D1	79.5	83.5	11.1	12
S2D2	55.5	58.6	5.7	6.2

Table 4.10: OE models - Average impulse response fits and average of the selected orders when the F-test is applied after cross-validation to perform model order selection in the identification of 200 systems in each data set.

FPE and AIC

Both FPE and AIC present a marked overfitting tendency, which is beneficial only in data set S2D1. Therefore, the F-test could be exploited to evaluate model structures of lower complexities with respect to the ones returned by the two information criteria. However, when applied in this context, the F-test is not able to guarantee a significant decrease of the selected orders, thus giving no improvements in terms of fit.

BIC

When BIC criterion is combined with the F-test according to the way described in Section 2.3.2, no significant improvement can be observed in the average of the achieved fits. The only exception is given by data set S1D2, where BIC encounters difficulties derived from overfitting; in this case, testing for simpler models leads to a slight performances improvement.
The reason for which BIC does not benefit from the successive application of the F-test is probably the good matching that already exists between the orders chosen by the oracle and by BIC. Therefore, the probability to accept different model structures is very low when the significance level for the F-test is small. On the other hand, when a large α_F is adopted, the newly chosen models tend to worsen the achieved fits, because they deviate more from the oracle choices.

Conclusions

When the combination illustrated in Section 2.3.2 between a model order selection method and the F-test is applied to estimate the complexities of OE models, it still gives rise to improvements w.r.t. the fits originally achieved applying only a model order selection procedure. The only exception has been found with the information criteria, which do not benefit from its combination with the F-test.
When model structures of increasing complexities are evaluated by the F-test, significance levels lower than 0.5 should be used, since larger values lead to the selection of too complex models.

Chapter 5

New model order selection techniques: Kernel-based model order selection

This chapter presents and tests a model order selection method which combines the kernel-based approach for system identification and the classical PEM method.

A theoretical description of the kernel-based estimation is first given, followed by the illustration of its combination with PEM. In section 5.2 these two techniques are evaluated by applying them on the four data sets introduced in Section 3.1. In addition, Sections 5.2.1 and 5.2.2 report the results achieved when the order selection method introduced in this chapter is combined with the validation methods and with the F-test.

5.1 Theoretical description

Preliminaries

Let us consider again the identification of the system described in (1.1), given the set of input-output data $Z^N = \{u(1), y(1), u(2), y(2), ..., u(N), y(N)\}$. In this case we consider the estimation through a FIR model, i.e. no noise modelling is performed ($H(q,\theta) = 1$), while $G_0(q)$ is estimated through the finite-order polynomial $B(q)$ defined in (1.13), $G(q,\theta) = B(q)$. For simplicity, let us assume that a one-sample delay exists in the dynamics from u to y, i.e. $n_k = 1$ in (1.13). Therefore, the model here adopted is given by

$$y(t) = G(q,\theta)u(t) + e(t) = B(q)u(t) + e(t) \tag{5.1}$$

with $E[e(t)] = 0$ and $E[e(t)e(s)] = \sigma^2 \delta_{t,s}$.
As observed in Section 1.1, the one-step ahead predictor for a FIR model can be represented by a linear regression model,

$$\hat{y}(t|\theta) = G(q,\theta)u(t) = \varphi(t)^T \theta$$

with the vector of regressors given by

$$\varphi(t) = \begin{bmatrix} u(t-1) & u(t-2) & \cdots & u(t-n_b) \end{bmatrix}^T$$

and the parameter vector θ containing the coefficients of the polynomial $B(q)$, or equivalently, the first n_b impulse response coefficients:

$$\theta = \begin{bmatrix} b_1 & b_2 & \cdots & b_{n_b} \end{bmatrix}^T \tag{5.2}$$

Therefore, assuming that the loss function $V_N(\theta, Z^N)$ takes the form (1.22), the minimization problem (1.5) admits the least-squares solution (1.23).
The estimation (1.23) is applicable when the data Z^N are "pre-windowed" adding n_b zeros such that the regressors vectors $\{\varphi(t),\ t = 1, ..., n_b\}$ can be formed. As an alternative, the summation in the loss function (1.22) can be defined starting from $t = n_b + 1$, leading to the following least-squares solution, that will be here adopted:

$$\hat{\theta}_N^{LS} = \left[\frac{1}{N} \sum_{t=n_b+1}^{N} \varphi(t)\varphi(t)^T \right]^{-1} \frac{1}{N} \sum_{t=n_b+1}^{N} \varphi(t)y(t) \tag{5.3}$$

For ease of notation, let us pass to the matrix formulation, by defining

$$\begin{aligned} Y_N &= \begin{bmatrix} y(n_b+1) & y(n_b+2) & \cdots & y(N) \end{bmatrix}^T \\ \Phi_N &= \begin{bmatrix} \varphi(n_b+1) & \varphi(n_b+2) & \cdots & \varphi(N) \end{bmatrix} \end{aligned} \tag{5.4}$$

Thus, we can write

$$\hat{\theta}_N^{LS} = (\Phi_N \Phi_N^T)^{-1} \Phi_N Y_N = R_N^{-1} F_N \tag{5.5}$$

with

$$R_N = \Phi_N \Phi_N^T = \sum_{t=n_b+1}^{N} \varphi(t)\varphi(t)^T \tag{5.6}$$

$$F_N = \Phi_N Y_N = \sum_{t=n_b+1}^{N} \varphi(t)y(t) \tag{5.7}$$

Bayesian linear regression

When estimation is done in a Bayesian setting, the parameter to be estimated is considered as a random variable, for which a prior distribution is known; given the available data, its posterior distribution is then estimated.

According to this procedure, let us assume that the vector θ defined in (5.2) is a Gaussian distributed random variable,

$$\theta \in \mathcal{N}(\theta^{ap}, P_{n_b}) \tag{5.8}$$

where the mean θ^{ap} and the covariance matrix P_{n_b} should contain the a-priori knowledge of the parameter vector. θ^{ap} is usually set to 0 ($\theta^{ap} = 0$), while different choices can be done for P_{n_b}, as will be illustrated in the following.
Moreover, assume that the noise $e(t)$ is independently Gaussian distributed, i.e. $e(t) \in \mathcal{N}(0, \sigma^2)$. Defining

$$\Lambda_N = \begin{bmatrix} e(n_b+1) & e(n_b+2) & \cdots & e(N) \end{bmatrix}^T \tag{5.9}$$

we can express the model (5.1) through a matrix formulation:

$$Y_N = \Phi_N^T \theta + \Lambda_N, \qquad \Lambda_N \in \mathcal{N}(0_{N-n_b\times 1}, \sigma^2 I_{N-n_b}) \tag{5.10}$$

with

$$Y_N \in \mathcal{N}\left(0_{N-n_b\times 1}, \Phi_N^T P_{n_b} \Phi_N + \sigma^2 I_{N-n_b}\right) \tag{5.11}$$

in view of (5.8). Therefore, Y_N and θ are jointly Gaussian random variables,

$$\begin{bmatrix} \theta \\ Y_N \end{bmatrix} \in \mathcal{N}\left(\begin{bmatrix} 0_{n_b\times 1} \\ 0_{N-n_b\times 1} \end{bmatrix}, \begin{bmatrix} P_{n_b} & P_{n_b}\Phi_N \\ \Phi_N^T P_{n_b} & \Phi_N^T P_{n_b} \Phi_N + \sigma^2 I_{N-n_b} \end{bmatrix}\right) \tag{5.12}$$

The posterior distribution of θ can then be determined by conditioning θ on Y_N,

$$\theta | Y_N \in \mathcal{N}\left(\hat{\theta}_N^{apost}, P_N^{apost}\right) \tag{5.13}$$

with

$$\begin{aligned}
\hat{\theta}_N^{apost} &= 0_{n_b\times 1} + P_{n_b}\Phi_N \left(\Phi_N^T P_{n_b} \Phi_N + \sigma^2 I_{N-n_b}\right)^{-1} (Y_N - 0_{N-n_b\times 1}) && (5.14) \\
&= \left(P_{n_b}\Phi_N\Phi_N^T + \sigma^2 I_{n_b}\right)^{-1} P_{n_b}\Phi_n Y_N \\
&= \left(R_N + \sigma^2 P_{n_b}^{-1}\right)^{-1} F_N && (5.15) \\
P_N^{apost} &= P_{n_b} - P_{n_b}\Phi_N \left(\Phi_N^T P_{n_b}\Phi_N + \sigma^2 I_{N-n_b}\right)^{-1} \Phi_N^T P_{n_b} && (5.16)
\end{aligned}$$

where R_N and F_N are respectively defined in (5.6) and (5.7). Expressions (5.14) and (5.16) are directly derived from the classical result relative to conditioned jointly Gaussian variables.

The computation of $\hat{\theta}_N^{apost}$ and P_N^{apost} requires the definition of P_{n_b}. When the a-priori knowledge does not allow to precisely define P_{n_b}, an estimation is made possible by the Bayesian setting, exploiting the so-called empirical Bayes methods. Namely, assume that P_{n_b} depend on an unknown hyper-parameter vector β; from (5.11) we have

$$Y_N \in \mathcal{N}\left(0_{N-n_b\times 1}, \Phi_N^T P_{n_b}(\beta)\Phi_N + \sigma^2 I_{N-n_b}\right) \tag{5.17}$$

Therefore, β can be estimated by maximum likelihood, i.e. maximizing the likelihood function of the observations Y_N given β:

$$\hat{\beta} = \arg\min_{\beta} Y_N^T \Sigma(\beta)^{-1} Y_N + \log\det\Sigma(\beta) \tag{5.18}$$

with $\Sigma(\beta) = \Phi_N^T P_{n_b}(\beta)\Phi_N + \sigma^2 I_{N-n_b}$. The noise variance σ^2 can be estimated by including it in the vector β or as the sample variance of an estimated high-order FIR model.

Choice of the kernel

As previously illustrated, P_{n_b} can be formulated as function of some hyper-parameters which are then estimated by maximum-likelihood. The problem that still remains open is the way in which P_{n_b} is defined. In the following, two connections between the Bayesian regression and other estimation approaches will be presented, thus providing further insight on the formulation of P_{n_b}.

A first result is that the a-posteriori estimate $\hat{\theta}_N^{apost}$ coincides with the solution of a regularized least-squares problem, $\hat{\theta}_N^{reg}$, with a specific choice of the regularization matrix. Indeed, when regularization is applied, the parameters are estimated by minimizing the function

$$V_N^{reg}(\theta, L) = \sum_{t=n_b+1}^{N} \left(y(t) - \varphi(t)^T\theta\right)^2 + \theta^T L\theta \tag{5.19}$$

where L is the so-called regularization matrix, i.e. a positive semi-definite matrix which acts as a penalty term on the parameter complexity. L determines the amount of shrinkage (towards zero) that is imposed on the elements of the parameter vector. In this case, the estimate becomes

$$\begin{aligned} \hat{\theta}_N^{reg} &= \left(R_N + L\right)^{-1} F_N && (5.20) \\ &= \left(R_N + L\right)^{-1} R_N \hat{\theta}_N^{LS} && (5.21) \end{aligned}$$

where again R_N and F_N are defined in (5.6) and (5.7), while $\hat{\theta}_N^{LS}$ is given in (5.5). Comparing equations (5.15) and (5.20) it is clear that they coincide if

$$L = \sigma^2 P_{n_b}^{-1} \tag{5.22}$$

This connection between the regularized estimation and the Bayesian regression gives an insight on how to choose both the regularization matrix L and the a-priori covariance matrix P_{n_b}. Indeed, on the one hand the elements of L should be chosen according to the degree of shrinkage which is desired for the elements of the parameter vector θ; for simplicity, L is usually a diagonal matrix. On the other hand, P_{n_b} should incorporate the a-priori knowledge about the correlation among the elements of θ, i.e. among the impulse response coefficients in the application here considered.

Exploiting the relation between the two cited estimation approaches, three types of formulations for P_{n_b} are introduced in [2].
In particular, starting from the definition of the regularization matrix L, a possible choice for it can be done in order to minimize the mean square error

$$MSE(\hat{\theta}_N^{reg}) = E\left[\left(\hat{\theta}_N^{reg} - \theta_0\right)\left(\hat{\theta}_N^{reg} - \theta_0\right)^T\right] \tag{5.23}$$

assuming that the true system is described by a FIR model of order n_b. Let us keep this assumption and also suppose that for large N there exists a certain $\mu > 0$, such that

$$\frac{1}{N - n_b} R_N \approx \mu I_{n_b} \tag{5.24}$$

Adopting a diagonal matrix L, $L = diag(l_1, l_2, ..., l_{n_b})$, the (k, k)-th element of $MSE(\hat{\theta}_N^{reg})$ becomes

$$MSE(\hat{b}_k^{reg}) \approx \frac{\sigma^2 \mu(N - n_b) + l_k^2 (b_k^0)^2}{[\mu(N - n_b) + l_k]^2} \tag{5.25}$$

where b_k^0 indicates the true k-th impulse response coefficient. The minimizer of $MSE(\hat{b}_k^{reg})$ w.r.t l_k is given by $l_k = \frac{\sigma^2}{(b_k^0)^2}$. Furthermore, assuming that the system is stable and indicating with λ the absolute value of the dominant pole of the true system, there exists a $c > 0$ such that $|b_k^0| < c\lambda^k$. Therefore, the diagonal of L should face an exponential increase:

$$l_k = \left(\frac{\sigma}{c\lambda^k}\right)^2, \qquad k = 1, ..., n_b \tag{5.26}$$

In [2], these considerations are directly exploited for the formulation of P_{n_b}, leading to the following diagonal covariance matrix:

$$P_{DI} = diag(p_1, p_2, ..., p_{n_b}), \qquad p_k = c\lambda^k \tag{5.27}$$

with $c \geqslant 0$ and $0 \leqslant \lambda \leqslant 1$ being the hyper-parameters estimated by means of (5.18).

Moreover, assuming a smooth impulse response, a formulation of the non-diagonal elements of P_{n_b} is also given. Namely, the highest correlations should be present in the diagonals close to the main one. Thus, a possible choice for P_{n_b} can be the following:

$$P_{DC}(k,j) = c\rho^{|k-j|}\lambda^{\frac{k+j}{2}} \tag{5.28}$$

with $c \geqslant 0$, $0 \leqslant \lambda \leqslant 1$ and $|\rho| \leqslant 1$ being hyper-parameters to be estimated. In (5.28) the term $\rho^{|k-j|}$ gives rise to larger values for the elements close to the main diagonal, while the term $c\lambda^{\frac{k+j}{2}}$ allows to include the previous considerations done for the diagonal of L.
Imposing the relation $\rho = \lambda^{\frac{1}{2}}$, a third formulation of P_{n_b} is derived:

$$P_{TC}(k,j) = c\min(\lambda^k, \lambda^j) \tag{5.29}$$

A second connection can be established between the Bayesian regression previously illustrated and the Gaussian process regression (GPR). Namely, the estimates (5.14) and (5.16) coincide with the Gaussian process estimate of any collection of impulse response coefficients.
GPR is applied to infer a certain function $f(x)$ from a set of measurements $\{y_k,\ k = 1, ..., N\}$. The name of the method comes from the fact that the function $f(x)$ is a-priori modelled as a Gaussian process with a certain mean and covariance function, which is called *kernel* in this context. The posterior distribution of $f(x)$ given the measurements $\{y_k,\ k = 1, ..., N\}$ can be then computed using the rules for conditioning jointly Gaussian random variables.
In [10], GPR is exploited to infer the impulse response of a stable linear system: the first n_b coefficients of the impulse response are modeled as a Gaussian process, as was done in (5.8) for the Bayesian linear regression. Given the observations Y_N, the posterior distribution of the impulse response is then computed as was done in (5.14) and (5.16).
The application of GPR requires also the choice of the kernel P_{n_b}, from which the name of the method here described comes. Thanks to the analogy between the kernel in GPR and the a-priori covariance matrix in the Bayesian regression, the formulations previously given for P_{n_b} can be adopted also in the GPR setting. However, in GPR other choices are possible. In [10] three possible kernels are considered, starting from the common "cubic spline" kernel

$$P_{CS}(k,j) = \begin{cases} c\frac{k^2}{2}\left(j - \frac{k}{3}\right), & k \geqslant j \\ \\ c\frac{j^2}{2}\left(k - \frac{j}{3}\right), & k < j \end{cases} \tag{5.30}$$

which can be modified into the so-called "stable spline" kernel

$$P_{SS}(k,j) = \begin{cases} c\frac{e^{-2\nu k}}{2}\left(e^{-\nu j} - \frac{e^{-\nu k}}{3}\right), & k \geqslant j \\ \\ c\frac{e^{-2\nu j}}{2}\left(e^{-\nu k} - \frac{e^{-\nu j}}{3}\right), & k < j \end{cases} \tag{5.31}$$

Finally, a Gaussian kernel is also discussed [11]:

$$P_{SE}(k,j) = ce^{-\frac{(k-j)^2}{2\lambda^2}} \tag{5.32}$$

In the previous equations $c \geqslant 0$, $0 \leqslant \lambda \leqslant 1$ and $\nu > 0$ act as hyper-parameters that are estimated using (5.18).

Connection with PEM methods

When the kernel-based method previously described is applied for system identification, the value of n_b is set quite large, since a complexity reduction is then guaranteed by the regularization imposed by the matrix $\sigma^2 P_{n_b}^{-1}$. Therefore, this method does not require to the user a specific model order selection, because it is implicitly done by regularization. The order selection technique which is introduced in [3] exploits this property in order to determine the optimal complexity of models estimated by PEM.

Assume that a set of possible models $\mathcal{m}_d = \mathcal{M}(\widehat{\theta}_N(d))$ estimated by (1.5) is given, where d represents the complexity, i.e. $d = \dim \widehat{\theta}_N(d)$. Furthermore, let $\widehat{G}_d$ and $\widehat{G}_{KB}$ indicate the set of the first n impulse response coefficients of the model $\mathcal{m}_d$ and of the model $\mathcal{m}_{KB}$ estimated by kernel-based regularization, respectively. The optimal complexity d^* can then be selected as the one which maximizes the impulse response fit between $\mathcal{m}_d$ and $\mathcal{m}_{KB}$:

$$d^* = \arg\max_d \mathcal{F}(\widehat{G}_{KB}, \widehat{G}_d) \tag{5.33}$$

$\mathcal{F}$ represents the function defined in (3.1), with G_0 and $\widehat{G}$ replaced by $\widehat{G}_{KB}$ and $\widehat{G}_d$, respectively.

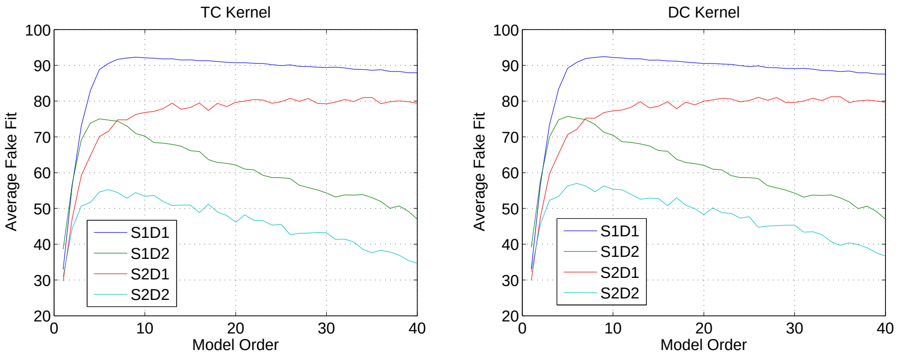
For easiness of notation, the model order selection criterion defined by (5.33) will be denoted by KB+PEM in the following. This notation refers to the use of a generic kernel in the Bayesian estimation.

5.2 Experimental results

The order selection method described in the previous section is now applied on the data sets introduced in Section 3.1. As usual, 200 systems are identified in each of the four data sets. The covariance matrices P_{TC} and P_{DC} are considered in the tests, since they give rise to the best results, according to the tests performed in [2].

In the following the abbreviations TC and DC will denote the system identification performed using the kernel-based estimation with $P_{n_b} = P_{TC}$ and $P_{n_b} = P_{DC}$, respectively. FIR models are estimated by this procedure. Furthermore, TC+PEM and DC+PEM will indicate the combination of the two kernel-based estimation methods with PEM, which leads to the complexity selection defined in (5.33) and to the estimation of OE models. As previously done, these are estimated by means of the routine `pem` without noise modelling.

Figure 5.1 shows the average values of $\mathcal{F}(\widehat{G}_{KB}, \widehat{G}_d)$ computed using OE models with orders ranging from 1 to 40, i.e. for d ranging from 2 to 80. The similarity between the plots in Figures 5.1 and 4.1.(a) proves the consistency of the order selection method arising from the combination of kernel-based estimation and PEM. Indeed, this proves that the maximization problem formulated in (5.33) leads to the selection of model complexities which maximize also the impulse response fit computed w.r.t. the true system (defined in (3.1)).



(a) TC kernel.

(b) DC kernel.

Figure 5.1: Average values of $\mathcal{F}(\widehat{G}_{KB}, \widehat{G}_d)$ for d going from 2 to 80, when 200 systems in each data set are identified.

Tables 5.1, 5.2 and 5.3 respectively illustrate the average impulse response fits, type 1 and type 2 prediction fits achieved by the evaluated identification methods. In addition, Figures 5.2 and 5.3 contain the box plots of the impulse response fits and of the type 2 prediction ones, respectively.
Table 5.1 shows how the combination with PEM gives rise to an improvement of the impulse response fit achieved by kernel-based estimation methods in the less-noisy data sets S1D1 and S2D1. However, in the noisy data set S2D2 the

Set	*TC*	*TC + PEM*	*DC*	*DC + PEM*
S1D1	91.4	91.9	91.7	92.3
S1D2	76	76.3	76.5	75.8
S2D1	80	83.5	80.2	83.7
S2D2	64.9	62.1	65.3	62

Table 5.1: Average impulse response fits achieved by the kernel-based estimation methods using P_{TC} and P_{DC} and by their combination with PEM procedures when 200 systems are identified in each data set.

Set	*TC*	*TC + PEM*	*DC*	*DC + PEM*
S1D1	93.1	92.6	93.4	93
S1D2	80.9	78.2	81.6	77.7
S2D1	84.1	86.9	84.4	87.2
S2D2	76	71.1	76.6	70.8

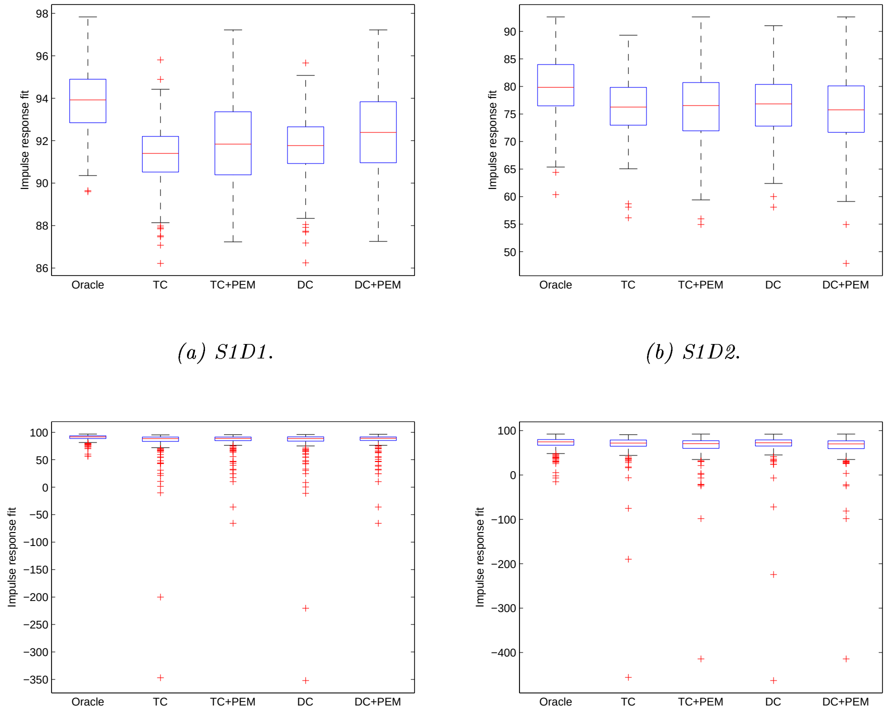
Table 5.2: Average type 1 prediction fits achieved by the kernel-based estimation methods using P_{TC} and P_{DC} and by their combination with PEM procedures when 200 systems are identified in each data set.

Set	*TC*	*TC + PEM*	*DC*	*DC + PEM*
S1D1	92.8	92.3	93.1	92.8
S1D2	80.2	77.6	80.8	77
S2D1	81.6	83.8	81.8	84.3
S2D2	71.4	67	72	66.4

Table 5.3: Average type 2 prediction fits achieved by the kernel-based estimation methods using P_{TC} and P_{DC} and by their combination with PEM procedures when 200 systems are identified in each data set.

combination leads to a decrease of the average fit w.r.t. the one achieved using only kernel-based methods adopting TC and DC kernels. The same result is detected also in data set S1D2 when a DC kernel is used.

Comparing the performances achieved by kernel-based methods alone with the two types of kernel, the DC one generally leads to better fits than the TC one. The higher number of degrees of freedom that the DC formulation of P_{n_b} allows is probably helpful in this sense. On the other hand, comparing the average fits reported in Table 5.1 in columns TC+PEM and DC+PEM and the fits achieved by the classical order selection methods with OE models (Table 4.2), the combination described in this section beats the classical procedures in noisy data sets S1D2 and S2D2, while the opposite situation is detected in S1D1 and S2D1. The regularized estimation performed by kernel-based methods is probably beneficial when a relevant measurement noise is present in the data. This behaviour also partially explains the reduced impulse response fits achievable when the kernel-based methods are combined with PEM procedures in data set



(a) S1D1.

(b) S1D2.

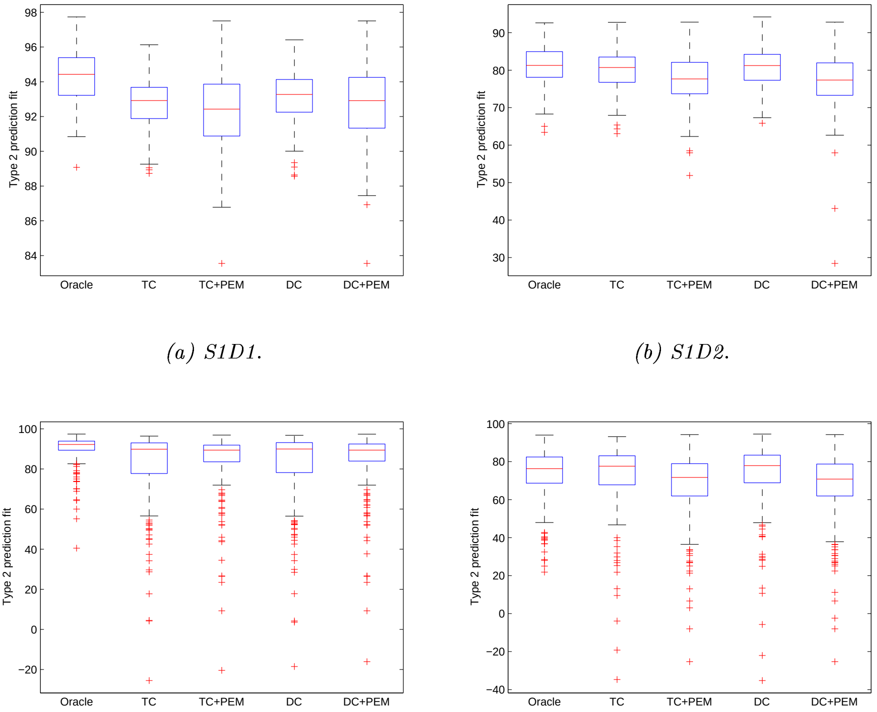
(c) S2D1.

(d) S2D2.

Figure 5.2: Box plots of the impulse response fits achieved by the kernel-based estimation methods using P_{TC} and P_{DC} and by their combination with PEM procedures when 200 systems are identified in each data set.

S2D2. Indeed, the fits achieved by the oracle for OE models in that set are lower than the ones obtained by the kernel methods alone for almost half of the identified systems. The percentage of systems for which this result is found significantly decreases in the less-noisy data sets, while it is around 25% in S1D2. In these cases, even if the optimal complexity is chosen for OE models, they do not allow to reach the performances achievable with kernel-based methods.

The good behaviour observed in the noisy data sets for the technique here tested is also confirmed by the histograms in Figure 5.4: the differences between the orders chosen by the oracle for OE models and the corresponding orders selected by TC+PEM and DC+PEM are much more centered around 0 in data sets S1D2 and S2D2 than in S1D1 and S2D1. Actually, in the latter data sets an overmodelling tendency is much more noticeable than in the noisy sets. This overmodelling trend, which is present also in the noisy sets, is the other cause of



(a) S1D1.

(b) S1D2.

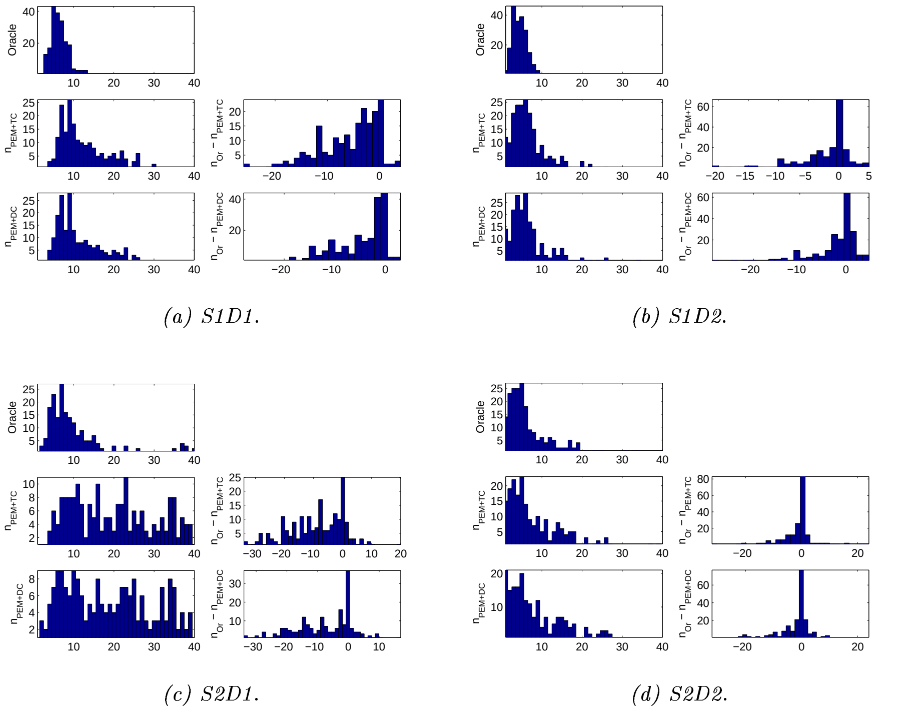
(c) S2D1.

(d) S2D2.

Figure 5.3: Box plots of the type 2 prediction fits achieved by the kernel-based estimation methods using P_{TC} and P_{DC} and by their combination with PEM procedures when 200 systems are identified in each data set.

the performances worsening detected in S2D2 after the combination with PEM.

Inspecting the prediction fits in Tables 5.2 and 5.3, it is clear how the models estimated by kernel-based methods are more suitable for prediction than the OE models estimated using PEM. Again, the regularization performed by means of a specific kernel is probably beneficial for prediction. An opposite result is detected only in data set S2D1, since the overmodelling tendency characterizing TC+PEM and DC+PEM is beneficial. Indeed, complex OE models allow to better reproduce the slow dymanics of the systems present in that data set, as it is proved by the substantial improvement reached by the impulse response fit after the combination of kernel-based methods with PEM. The better reproduction of the system dynamics favours also the good performances of the estimated model in terms of prediction.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 5.4: Histograms of the orders selected by the combination of kernel-based estimation methods (with TC and DC kernels) with PEM (TC+PEM, DC+PEM). The histograms of the differences w.r.t. to the oracle choices are also shown.

In addition, it should be pointed out that the outliers appearing in the box plots of the impulse response fits in Figures 5.2.(c) and (d) refer to particular systems, which have all zero and poles concentrated in a very small region close to the unit circle. Thus, a proper identification of these systems appear particularly difficult; however, the box plots in Figures 5.3.(c) and (d) prove that the corresponding estimated models don't give so bad performances in terms of prediction, when inputs with the same properties of the estimation ones are adopted.

The following sections will present the results achieved when the methods TC+PEM and DC+PEM are combined with the validation methods or with the F-test, according to the ways described in Section 2.3.

5.2.1 Combination with the validation methods

In this context the validation methods are applied, using only estimation data, on the model structure returned by TC+PEM or DC+PEM: if it does not pass the considered test on the residuals, new complexities are evaluated, starting from the one leading to the second best value of $\mathcal{F}(\widehat{G}_{KB}, \widehat{G}_d)$ and then following the complexities order defined by the maximization problem (5.33). The first of these newly chosen model structures which passes the statistical test on the residuals is then definitively selected.
Notice that the combination here analyzed is analogous to the one introduced in Section 2.3.1, with comparison methods replaced by KB+PEM.
Let us also recall that the value of M used in the statistical tests on the residuals (see (2.7) and (2.19)) is again set to $n_b + 20$, with n_b being the order of the OE model estimated by the routine `pem` without noise modelling.

Set	Average impulse response fit					Average of the selected orders				
	TC + PEM	*+ RAW*	*+ RA*	*+ RAI1*	*+ RAI2*	*TC + PEM*	*+ RAW*	*+ RA*	*+ RAI1*	*+ RAI2*
S1D1	91.9	91.9	91.9	91.9	91.9	12.5	12.4	12.4	12.5	12.6
S1D2	76.3	76.1	76.2	76.4	76.4	6.9	6.9	6.9	6.9	7
S2D1	83.5	84.5	84.4	83.5	83.5	20	19.2	19.4	20	20
S2D2	62.1	62.5	62.7	62.6	62.5	8.3	9.2	9.3	8.7	8.7

Table 5.4: Average impulse response fits and average of the selected orders when validation methods are combined with TC+PEM for model order selection in the identification of 200 systems in each data set.

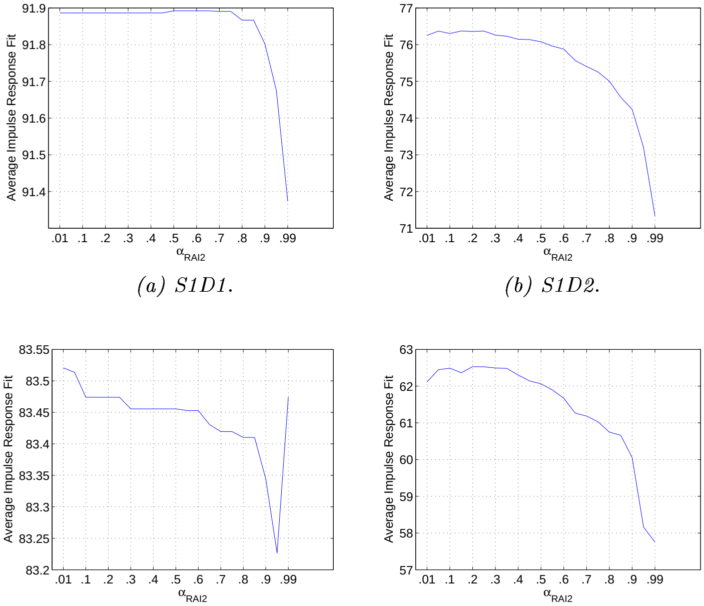
Set	Average impulse response fit					Average of the selected orders				
	DC + PEM	*+ RAW*	*+ RA*	*+ RAI1*	*+ RAI2*	*DC + PEM*	*+ RAW*	*+ RA*	*+ RAI1*	*+ RAI2*
S1D1	92.3	92.3	92.3	92.3	92.3	11.4	11.3	11.3	11.4	11.4
S1D2	75.8	75.6	75.6	75.8	75.8	7	7.1	7.1	7	7.2
S2D1	83.7	84.7	84.7	83.7	83.7	19.2	19.3	19.3	19.2	19.2
S2D2	62	62	62.1	62.2	62.2	9.1	9.8	9.9	9.2	9.2

Table 5.5: Average impulse response fits and average of the selected orders when validation methods are combined with DC+PEM for model order selection in the identification of 200 systems in each data set.

Tables 5.4 and 5.5 respectively show the average impulse response fits achieved by TC+PEM and DC+PEM with the combination just described. It is clear that the combination is not helpful in the "fast" data sets S1D1 and S1D2, while it leads to improvements in the impulse response fit in the "slow" sets. The combination is particularly helpful in data set S2D2, in order to alleviate the performance worsening which is detected with TC+PEM and DC+PEM w.r.t. the fits achieved by kernel-based methods alone.

Among the tests on the residuals, the use of both the whiteness and the independence tests (RA) gives rise to the best performances in S2D1 and S2D2, but it also leads to a slight decrease of the average impulse response fit in S1D2 w.r.t. the ones originally achieved by TC+PEM and DC+PEM. The application of the independence test (RAI1 or RAI2) in this context appears more robust, since it does not worsen the performances of TC+PEM and DC+PEM, even if it does not improve their fits in S2D1. The reason of this behaviour is the minor impact brought by this test on the previous choices of DC+PEM and TC+PEM: indeed, the test for the independence of the residuals from past inputs unfalsifies a very large percentage of the estimated model structures, leading to very rare selections of different complexities. This property is beneficial in the noisy data sets, where the similarity between the plots in Figure 5.1 and in Figure 4.1.(a) proves that substantial modifications of the complexity choices done by TC+PEM and DC+PEM would also deteriorate the impulse response fit. This observation is in particular true for data set S1D2, since the range of complexities which give rise to the best impulse response fits is very narrow, even w.r.t. the one observed in S2D2.
Different considerations hold when both the tests on the residuals (RA) are applied after TC+PEM and DC+PEM: the probability that RA unfalsifies the model structures that they return is much lower and this in turn increases the percentage of changes in the complexities originally selected. This explains both the performances worsening observed in S1D2 and the improvement achieved in S2D1, where the overmodelling tendency of TC+PEM and DC+PEM is very significant (see Figure 5.4).

As was previously done in Section 4.1.5, the impact of the significance level adopted in the tests on the residuals is again investigated. The plots in Figure 5.5 show the average impulse response fits achieved by the combination of TC+PEM with the independence test on the residuals (RAI2), as function of the significance level α_{RAI2} adopted in the test. The plots indicate that small values of α_{RAI2} are more beneficial, since they lead to very few changes in the complexity choices already done by TC+PEM. Moreover, when α_{RAI2} is lower than 0.5, this technique appears quite robust w.r.t. the choice of the significance level. The same result is not detected when TC+PEM is combined with RA: indeed, in this case, values of α_{RA} larger than 0.05 lead to a more considerable worsening of the average impulse response fit. This is a consequence of the large percentage of changes in the complexities returned by TC+PEM. Therefore, when RA is applied in this context, very small values of the significance level are suggested.
The considerations here done hold also for DC+PEM.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 5.5: OE models - Average impulse response fits achieved for different values of the significance level α_{RAI2} when the independence test on residuals (RAI2) is combined with TC+PEM for model order selection in the identification of 200 systems in each data set.

5.2.2 Combination with the F-test

The application of the F-test according to the way illustrated in Section 2.3.2 is tested also with the order selection technique here introduced. Figure 5.4 clearly shows how the two methods denoted by TC+PEM and DC+PEM tend to overmodel, as it is proved by the histograms relative to the differences between the orders chosen by the oracle and by these procedures. Therefore, the F-test could be helpful for the evaluation of model structures with decreasing complexity. Tables 5.6 and 5.7 show that the application of the F-test for testing model structures of lower complexities w.r.t. the ones returned by TC+PEM and DC+PEM leads to a decrease of the average impulse response fit in the "fast" data sets S1D1 and S1D2 but to an increase in S2D1 and S2D2.

The substantial decrease detected in S1D2 is again explicable by the narrow range of orders which are suitable to properly reproduce the true systems in this data set. In addition, Figure 5.1 proves that the choices done by the meth-

	Average impulse response fit		Average of the selected orders	
Set	*TC + PEM*	*+ F (Down)*	*TC + PEM*	*+ F (Down)*
S1D1	91.9	91.5	12.5	10.3
S1D2	76.3	75.1	6.9	6.7
S2D1	83.5	83.8	20	18.8
S2D2	62.1	63.5	8.3	7.2

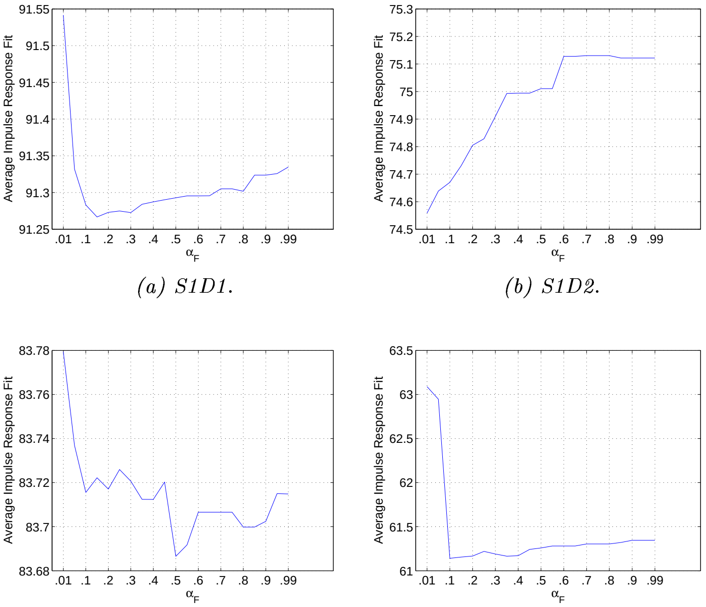
Table 5.6: Average impulse response fits and average of the selected orders when the F-test is applied after TC+PEM in order to perform model order selection in the identification of 200 systems in each data set.

	Average impulse response fit		Average of the selected orders	
Set	*DC + PEM*	*+ F (Down)*	*DC + PEM*	*+ F (Down)*
S1D1	92.3	92	11.4	9.3
S1D2	75.8	74.6	7	6.8
S2D1	83.7	84.1	19.2	18.1
S2D2	62	62.6	9.1	8

Table 5.7: Average impulse response fits and average of the selected orders when the F-test is applied after DC+PEM in order to perform model order selection in the identification of 200 systems in each data set.

ods TC+PEM and DC+PEM in S1D2 are already consistent. Therefore, only small modifications of the complexities returned by these two methods could be beneficial. This purpose can be achieved by adopting a large significance level α_F for the F-test, which makes less probable the selection of lower orders w.r.t. the ones originally returned by TC+PEM and DC+PEM. Figure 5.6 confirms this consideration, showing that very large values of α_F lead to the highest average impulse response fit in S1D2, when the F-test is applied after TC+PEM. However, even setting $\alpha_F = 0.99$, the combination with the F-test will choose simpler model structures w.r.t. the ones returned by TC+PEM, whenever the latter ones coincide with a local minimum. This explains the performance worsening detected in S1D2 after the combination of the F-test with TC+PEM. Analogous considerations hold for DC+PEM.

Further investigating the plots in Figure 5.6, it is clear that small values of α_F give rise to the highest average impulse response fits in data sets S1D1, S2D1 and S2D2. Indeed, with this choice of α_F the modifications of the complexities returned by TC+PEM are favoured in order to alleviate the overmodelling tendency observed in the less noisy data sets S1D1 and S2D1 and to refine the choices done in S2D2 by TC+PEM.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 5.6: OE models - Average impulse response fits achieved for different values of the significance level α_F when the F-test is applied after TC+PEM in order to evaluate model structures with decreasing complexity. The average is calculated from the identification of 200 systems in each data set.

5.2.3 Conclusions

The results illustrated in this section have shown how the kernel-based estimation is very effective when noisy measurement data are available. In some cases, neither the optimal complexity choice for an OE model can beat a regularized FIR model in the reproduction of the true system. The combination of the kernel-based estimation with the classical PEM (KB+PEM) appears beneficial when the measurement data are almost noiseless, even if it is affected by overmodelling, when compared with the optimal choices. Moreover, when the order selection technique introduced in this chapter and denoted by KB+PEM is equipped with a test on the residuals or with the F-test, a slight performances improvement is detected in the "slow" data sets, w.r.t. the results achieved by KB+PEM alone. It is worth to observe that the main enhancement is obtained in data set S2D2, where PEM has proved to be less effective than the kernel-based estimation.

Chapter 6

New model order selection techniques: PUMS - Parsimonious Unfalsified Model Structure Selection

This chapter illustrates and tests a new model order selection procedure, that has been introduced in [5].
A theoretical description of the method is given in Section 6.1. The technique is then applied on the four data sets presented in Section 3.1 and its performances are analyzed in Section 6.2.3. PUMS is also combined with the validation methods and the F-test, following procedures analogous to the ones illustrated in Section 2.3. The performances of these combinations are respectively presented in Sections 6.2.4 and 6.2.5. An equipment of PUMS with another test is also considered in Section 6.2.6. Finally, in Section 6.2.7 a different initialization of the MATLAB routine `pem` is illustrated and the corresponding results are presented.

6.1 Theoretical description

Given the measurement data $Z^N = \{u(t), y(t),\ t = 1, ..., N\}$, let us assume that the output data $\{y(t),\ t = 1, ..., N\}$ can be expressed through a linear regression model:

$$y(t) = \varphi(t)^T g_0 + e(t) \tag{6.1}$$

where $\varphi(t) \in \mathbb{R}^d$ is a known regressors vector, $g_0 \in \mathbb{R}^d$ is the unknown parameter vector and $e(t) \in \mathcal{N}(0, \sigma^2)$ is white additive noise, independent from $\varphi(t)$. To simplify the notation, let us pass to a matrix formulation. Namely, defining

$$\begin{aligned} Y_N &= [y(1)\ \cdots\ y(N)]^T && (6.2)\\ \Phi_N &= [\varphi(1)\ \cdots\ \varphi(N)] && (6.3)\\ \Lambda_N &= [e(1)\ \cdots\ e(N)]^T, \quad \Lambda_N \in \mathcal{N}(0_{N\times 1}, \sigma^2 I_N) && (6.4) \end{aligned}$$

where I_N denotes the N-dimensional identity matrix, we can write

$$Y_N = \Phi_N^T g_0 + \Lambda_N \tag{6.5}$$

Using the maximum likelihood approach (ML), g_0 can be estimated as

$$\hat{g}_N^{ML} = \arg\min_g\ J_{ML}(g) \tag{6.6}$$

where

$$J_{ML}(g) = \left(Y_N - \Phi_N^T g\right)^T \left(Y_N - \Phi_N^T g\right) \tag{6.7}$$

It is well known that $\hat{g}_N^{ML}$ coincides with the least-squares estimate $\hat{g}_N^{LS}$:

$$\begin{aligned} \hat{g}_N^{ML} = \hat{g}_N^{LS} &= \left(\Phi_N \Phi_N^T\right)^{-1} \Phi_N Y_N && (6.8)\\ &= g_0 + \left(\Phi_N \Phi_N^T\right)^{-1} \Phi_N \Lambda_N && (6.9) \end{aligned}$$

$\hat{g}_N^{ML}$ represents the unstructured estimate of g_0, since no model structure has been specified yet. In particular, assuming that the data are "pre-windowed" or that the input data $\{u(-n_k - d + 2), ..., u(1 - n_k)\}$ are known, then $\varphi(t)$ can be defined as

$$\varphi(t) = \left[\ u(t - n_k) \quad \dots \quad u(t - n_k - d + 1)\ \right]^T, \qquad t = 1, ..., N \tag{6.10}$$

with n_k being the input-output delay of the true system. When the regressors $\varphi(t),\ t = 1, ..., N$, are formulated as in (6.10), $\hat{g}_N^{ML}$ represents a FIR model estimate. However, $\hat{g}_N^{ML}$ is here referred to as an unstructured estimate of g_0, since d is assumed to be large enough to describe the non-trivial impulse response of the true system.

Let us now consider the structured estimation, in which a model structure $\widetilde{\mathcal{M}}$ is defined as a twice continuously differentiable mapping of the parameter vector θ into a model set $\Xi_{\widetilde{\mathcal{M}}}$:

$$\Xi_{\widetilde{\mathcal{M}}} = \left\{ g : g = \widetilde{\mathcal{M}}(\theta),\ \theta \in D_{\widetilde{\mathcal{M}}} \subset \mathbb{R}^{d_{\widetilde{\mathcal{M}}}} \right\} \tag{6.11}$$

Differently from the definition given in (1.9), let us consider $\widetilde{\mathcal{M}}$ simply a transformation of the parameter vector θ: thus, $\widetilde{\mathcal{M}}(\theta)$ is a vector containing the coefficients of the predictor polynomials $W_u(q, \theta)$ and $W_y(q, \theta)$. Further assuming that $\widetilde{\mathcal{M}}$ is an invertible function, this observation allows us to directly include g in the ML criterion:

$$J_{ML}(g) = \left(Y_N - \Phi_N^T g\right)^T \left(Y_N - \Phi_N^T g\right) \tag{6.12}$$

The structured ML estimate of g_0 is thus given by

$$\begin{aligned} \hat{g}_{N,\widetilde{\mathcal{M}}}^{ML} = \quad & \arg \quad \min_g J_{ML}(g) \\ & s.t. \quad g \in \Xi_{\widetilde{\mathcal{M}}} \end{aligned} \tag{6.13}$$

Notice that $\hat{g}_N^{ML} = \hat{g}_{N,\widetilde{\mathcal{M}}}^{ML}$ if $\hat{g}_N^{ML} \in \Xi_{\widetilde{\mathcal{M}}}$.

To be clearer, some examples are discussed. For instance, when FIR models of order $n_b = d_{\widetilde{\mathcal{M}}} < d$ have to be specified, the mapping $\widetilde{\mathcal{M}}$ becomes

$$\begin{aligned} \widetilde{\mathcal{M}}^{FIR} : \mathbb{R}^{d_{\widetilde{\mathcal{M}}}} & \rightarrow \mathbb{R}^d \\ \theta & \mapsto g \end{aligned} \tag{6.14}$$

with

$$\theta = \begin{bmatrix} b_1 & \cdots & b_{n_b} \end{bmatrix}^T \tag{6.15}$$

$$g = \begin{bmatrix} b_1 & \cdots & b_{n_b} & 0 & \cdots & 0 \end{bmatrix}^T \tag{6.16}$$

An analogous formulation of $\widetilde{\mathcal{M}}$ can be given also for OE models of order $n_b = n_f = \frac{d_{\widetilde{\mathcal{M}}}}{2}$. Namely, let $\theta \in \mathbb{R}^{d_{\widetilde{\mathcal{M}}}}$, $d_{\widetilde{\mathcal{M}}} < d$, contain the coefficients of the frequency response,

$$\theta = \begin{bmatrix} f_1 & \cdots & f_{n_f} & b_1 & \cdots & b_{n_b} \end{bmatrix}^T \tag{6.17}$$

then $\widetilde{\mathcal{M}}$ can be defined as

$$\begin{aligned} \widetilde{\mathcal{M}}^{OE} : \mathbb{R}^{d_{\widetilde{\mathcal{M}}}} & \rightarrow \mathbb{R}^d \\ \theta & \mapsto g = \begin{bmatrix} g_1 & \cdots & g_d \end{bmatrix}^T \end{aligned} \tag{6.18}$$

with

$$g_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{b_1 e^{-j\omega} + \cdots + b_{n_b} e^{-jn_b\omega}}{1 + f_1 e^{-j\omega} + \cdots + f_{n_f} e^{-jn_f\omega}} e^{i\omega k} d\omega \qquad k = 1, ..., d \tag{6.19}$$

Let us consider again the unstructured ML estimate $\widehat{g}_N^{ML}$. Since it coincides with the least-squares estimate $\widehat{g}_N^{LS}$, it is unbiased, i.e.:

$$E\left[\widehat{g}_N^{ML}\right] = E\left[\widehat{g}_N^{LS}\right] = E\left[g_0 + (\Phi_N\Phi_N^T)^{-1}\Phi_N\Lambda_N\right] = g_0 \tag{6.20}$$

Therefore, $\widehat{g}_N^{ML}$, together with the noise variance estimate

$$\widehat{\sigma}^2 = \frac{1}{N-d}\left(Y_N - \Phi_N^T\widehat{g}_N^{ML}\right)^T\left(Y_N - \Phi_N^T\widehat{g}_N^{ML}\right) \tag{6.21}$$

is a sufficient statistic for $\{g_0, \sigma^2\}$. Hence, the structured ML estimate of g_0 based on $\{\widehat{g}_N^{ML}, \widehat{\sigma}^2\}$ will have the same statistical properties as $\widehat{g}_{N,\widetilde{\mathcal{M}}}^{ML}$. Such ML estimate is given by

$$\begin{aligned} \widehat{g}_{N,\widetilde{\mathcal{M}}}^{MR} = \quad & \arg \quad \min_g J_{MR}(g) \\ & s.t. \quad g \in \Xi_{\widetilde{\mathcal{M}}} \end{aligned} \tag{6.22}$$

where

$$J_{MR}(g) = \left(\widehat{g}_N^{ML} - g\right)^T \Phi_N\Phi_N^T\left(\widehat{g}_N^{ML} - g\right) \tag{6.23}$$

Notice that the problem (6.22) can be viewed as a model reduction problem in which the unstructured estimate $\widehat{g}_N^{ML}$ is projected onto the set $\Xi_{\widetilde{\mathcal{M}}}$. Moreover, defining $\widehat{Y}_N = \Phi_N^T\widehat{g}_N^{ML}$, we can rewrite $J_{MR}(g)$ as

$$J_{MR}(g) = (\widehat{Y}_N - \Phi_N^T g)^T(\widehat{Y}_N - \Phi_N^T g) \tag{6.24}$$

i.e. as a ML criterion in which the original measurement vector Y_N is replaced by the estimated one $\widehat{Y}_N$.
Using (6.9), $J_{MR}(g)$ can be also rewritten in the following way:

$$\begin{aligned} J_{MR}(g) &= (g_0 - g)^T\Phi_N\Phi_N^T(g_0 - g) + 2\Lambda_N^T\Phi_N^T(g_0 - g) + \\ &+ \Lambda_N^T\Phi_N^T\left(\Phi_N\Phi_N^T\right)^{-1}\Phi_N\Lambda_N \qquad (6.25) \\ &= J_{ML}(g) - W_N^T W_N \qquad (6.26) \end{aligned}$$

with

$$W_N = \left(I_N - \Phi_N^T\left(\Phi_N\Phi_N^T\right)^{-1}\Phi_N\right)\Lambda_N \tag{6.27}$$

Therefore, $J_{MR}(g)$ and $J_{ML}(g)$ have the same optima, meaning that $\widehat{g}_{N,\widetilde{\mathcal{M}}}^{ML} = \widehat{g}_{N,\widetilde{\mathcal{M}}}^{MR}$. For ease of notation, both the estimates will be denoted as $\widehat{g}_N^{\widetilde{\mathcal{M}}}$ in the following.

The previous considerations about the structured estimation refer to the case in which a model structure $\widetilde{\mathcal{M}}$ is fixed. However, in a model structure selection setting, model structures of different complexities are defined and a criterion is adopted to discriminate among them. In this context, the use of the ML criterion

$$\begin{aligned} \left\{\hat{d}_{\widetilde{\mathcal{M}}}, \hat{g}_N^{\widetilde{\mathcal{M}}}\right\} = \quad & \arg \quad \min_{d_{\widetilde{\mathcal{M}}}, g} J_{ML}(g) \\ & s.t. \quad g \in \Xi_{\widetilde{\mathcal{M}}} \end{aligned} \tag{6.28}$$

is generally not beneficial. Indeed, if for example the least-squares estimate $\hat{g}_N^{LS} = \hat{g}_N^{ML} \in \Xi_{\widetilde{\mathcal{M}}}$ for some $d_{\widetilde{\mathcal{M}}}$, then $\hat{g}_N^{LS}$ will be always selected by the ML criterion. However, this is usually not a good choice, because of the high variance which characterizes the least-squares estimate.

In view of the previous consideration, different approaches should be considered for model order selection. For this reason, notice that the following expression, obtained inserting (6.8) in (6.23),

$$J_{MR}(g) = \left(Y_N - \Phi_N^T g\right)^T \Phi_N^T \left(\Phi_N \Phi_N^T\right)^{-1} \Phi_N \left(Y_N - \Phi_N^T g\right) \tag{6.29}$$

$$= \mathcal{E}(g)^T \Phi_N^T \left(\Phi_N \Phi_N^T\right)^{-1} \Phi_N \mathcal{E}(g) \tag{6.30}$$

suggests to use $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ as a test statistic. Indeed the vector $\Phi_N \mathcal{E}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ contains the sample correlations between the regressors in Φ_N and the residuals in $\mathcal{E}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ computed for different lags.
In particular, when the regressors $\varphi(t)$ are defined as in (6.10), then

$$\Phi_N \mathcal{E}(\hat{g}_N^{\widetilde{\mathcal{M}}}) = \sqrt{N} r_{\varepsilon u}^{N,M}, \qquad M_1 = 1,\ M_2 = d,\ M = d \tag{6.31}$$

Moreover, if the residuals are white with unit variance, then $\Phi_N \Phi_N^T \sim N P_{\varepsilon u}$ with the formulation of $P_{\varepsilon u}$ given in (2.15) and $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}}) \sim x_{\varepsilon u}^{N,M}$ (with $x_{\varepsilon u}^{N,M}$ defined in (2.18)). Therefore, $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ can be used as a test statistic for assessing the independence between the residuals and the input signal, i.e. to evaluate if traces of the input signal are still present in the residuals. For this purpose, the statistical properties of $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ have to be established under the assumption that the true system can be described by the specified model structure. To check whether $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ is consistent with these properties, a statistical test is performed. If a model $\hat{g}_N^{\widetilde{\mathcal{M}}}$ passes the test, the corresponding model structure $\widetilde{\mathcal{M}}$ is unfalsified.

Let us derive the statistical properties of $J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}})$ for the model structure $\widetilde{\mathcal{M}}^{FIR}$ defined in (6.14). The same properties can be derived also for rational models, using parametrization in (6.17) and (6.18).

Assume that the true system $\mathcal{S}$ is given by a FIR model of order $d_{\widetilde{\mathcal{M}}}$, i.e. $\mathcal{S} \equiv g_0 \in \Xi_{\widetilde{\mathcal{M}}}$; then, there exists a $\theta_0 \in \mathbb{R}^{d_{\widetilde{\mathcal{M}}}}$ such that $\widetilde{\mathcal{M}}^{FIR}(\theta_0) = g_0$. In the following we omit the superscript FIR, $\widetilde{\mathcal{M}}^{FIR} = \widetilde{\mathcal{M}}$, to simplify the notation. For the same reason, let A_N be the square-root of $R_N = \Phi_N \Phi_N^T$ and introduce

$$\Delta = A_N R_N^{-1} \Phi_N \Lambda_N \tag{6.32}$$

Notice that $\Delta \in \mathcal{N}(0_{d\times 1}, \sigma^2 I_d)$.
We can rewrite equation (6.25) as

$$J_{MR}(g) = (g_0 - g)^T R_N (g_0 - g) + 2\Delta^T A_N (g_0 - g) + \Delta^T \Delta \tag{6.33}$$

or, equivalently, as

$$J_{MR}(\theta) = \left(g_0 - \widetilde{\mathcal{M}}(\theta)\right)^T R_N \left(g_0 - \widetilde{\mathcal{M}}(\theta)\right) + 2\Delta^T A_N \left(g_0 - \widetilde{\mathcal{M}}(\theta)\right) + \Delta^T \Delta \tag{6.34}$$

We have

$$\begin{aligned} J'_{MR}(\theta) &= -2\left(g_0 - \widetilde{\mathcal{M}}(\theta)\right)^T R_N \widetilde{\mathcal{M}}'(\theta) - 2\Delta^T A_N \widetilde{\mathcal{M}}'(\theta) && (6.35)\\ J''_{MR}(\theta) &= -2g_0^T R_N \widetilde{\mathcal{M}}''(\theta) + 2\left[\widetilde{\mathcal{M}}'(\theta)\right]^T R_N \widetilde{\mathcal{M}}'(\theta) \\ &+ 2\left[\widetilde{\mathcal{M}}(\theta)\right]^T R_N \widetilde{\mathcal{M}}''(\theta) - 2\Delta^T A_N \widetilde{\mathcal{M}}''(\theta) && (6.36)\end{aligned}$$

Hence,

$$\begin{aligned} J_{MR}(\theta_0) &= \Delta^T \Delta && (6.37)\\ J'_{MR}(\theta_0) &= -2\Delta^T A_N \widetilde{\mathcal{M}}'(\theta_0) && (6.38)\\ J''_{MR}(\theta_0) &= 2\left[\widetilde{\mathcal{M}}'(\theta_0)\right]^T R_N \widetilde{\mathcal{M}}'(\theta_0) && (6.39)\end{aligned}$$

Let indicate with $\hat{\theta}_N$ the estimate obtained solving (6.22), that is $\widetilde{\mathcal{M}}(\hat{\theta}_N) = \hat{g}^{MR}_{N,\widetilde{\mathcal{M}}} = \hat{g}_N^{\widetilde{\mathcal{M}}}$. Exploiting the Taylor expansion, we have

$$\begin{aligned} J_{MR}(\theta_0) &= J_{MR}(\hat{\theta}_N) + \frac{1}{2}(\theta_0 - \hat{\theta}_N)^T J''_{MR}(\eta)(\theta_0 - \hat{\theta}_N) && (6.40)\\ J'_{MR}(\theta_0) &= J'_{MR}(\hat{\theta}_N) + (\theta_0 - \hat{\theta}_N)^T J''_{MR}(\xi) = (\theta_0 - \hat{\theta}_N)^T J''_{MR}(\xi) && (6.41)\end{aligned}$$

for some ξ and η in-between θ_0 and $\hat{\theta}_N$. Notice that from (6.41), we have

$$\theta_0 - \hat{\theta}_N = \left[J''_{MR}(\xi)\right]^{-1} J'_{MR}(\theta_0)^T \tag{6.42}$$

Hence,

$$\begin{aligned} J_{MR}(\widehat{\theta}_N) &= J_{MR}(\theta_0) - \frac{1}{2} J'_{MR}(\theta_0) \left[J''_{MR}(\xi)\right]^{-1} J''_{MR}(\eta) \left[J'_{MR}(\xi)\right]^{-1} J'_{MR}(\theta_0)^T \\ &\approx J_{MR}(\theta_0) - \frac{1}{2} J'_{MR}(\theta_0) \left[J''_{MR}(\theta_0)\right]^{-1} J'_{MR}(\theta_0)^T \qquad (6.43) \\ &= \Delta^T \left\{ I_d - A_N \widetilde{\mathcal{M}}'(\theta_0) \left[\left[\widetilde{\mathcal{M}}'(\theta)\right]^T R_N \widetilde{\mathcal{M}}'(\theta) \right]^{-1} \left[\widetilde{\mathcal{M}}'(\theta_0)\right]^T A_N \right\} \Delta \end{aligned}$$

Notice that the term between brackets is idempotent and its trace is equal to $d - d_{\widetilde{\mathcal{M}}}$; in addition, recalling that $\Delta \in \mathcal{N}(0_{d\times 1}, \sigma^2 I_d)$, it can be shown that

$$\frac{J_{MR}(\widehat{\theta}_N)}{\sigma^2} \sim \chi^2(d - d_{\widetilde{\mathcal{M}}}), \qquad d = \dim \widehat{g}_N^{ML}, \quad d_{\widetilde{\mathcal{M}}} = \dim \widehat{\theta}_N \tag{6.44}$$

when $\mathcal{S} \equiv g_0 \in \Xi_{\widetilde{\mathcal{M}}}$. The proof is analogous to the one given for Lemma II.3 in [8, p.556].
Observe that because of the approximation $\widehat{\theta}_N \approx \eta \approx \xi \approx \theta_0$ considered in (6.43), expression (6.44) may hold only for $N \to \infty$, when (6.5) represents a non-linear regression problem.

From the properties of the least-squares estimate, we also know that [8, p.556]

$$\frac{J_{ML}(\widehat{g}_N^{ML})}{\sigma^2} \in \chi^2(N - d) \tag{6.45}$$

Now notice from (6.25) that the terms of $J_{MR}(g)$ in which g appears depend on the noise Λ_N only through $\Phi_N \Lambda_N$; this holds also for $\widehat{g}_N^{\widetilde{\mathcal{M}}}$. Moreover, since Λ_N is Gaussian and $E[W_N(\Phi_N\Lambda_N)^T] = 0$, we conclude hat $J_{MR}(\widehat{g}_N^{\widetilde{\mathcal{M}}}) = J_{MR}(\widehat{\theta}_N)$ is independent from W_N. From (6.23) and (6.23) we also have that $J_{ML}(\widehat{g}_N^{ML}) = W_N^T W_N$, hence $J_{MR}(\widehat{g}_N^{\widetilde{\mathcal{M}}}) = J_{MR}(\widehat{\theta}_N)$ is also independent from $J_{ML}(\widehat{g}_N^{ML})$.
Therefore, using the rule for the division of two independent χ^2-distributed variables, we have

$$\frac{\frac{J_{MR}(\widehat{\theta}_N)}{\sigma^2(d-d_{\widetilde{\mathcal{M}}})}}{\frac{J_{ML}(\widehat{g}_N^{ML})}{\sigma^2(N-d)}} = \frac{J_{MR}(\widehat{\theta}_N)}{J_{ML}(\widehat{g}_N^{ML})} \cdot \frac{N-d}{d-d_{\widetilde{\mathcal{M}}}} \in F(d - d_{\widetilde{\mathcal{M}}}, N - d) \tag{6.46}$$

Moreover, exploiting equation (6.26) and recalling that $J_{ML}(\widehat{g}_N^{ML}) = W_N^T W_N$, we can rewrite

$$J_{MR}(\widehat{\theta}_N) = J_{MR}(\widehat{g}_N^{\widetilde{\mathcal{M}}}) = J_{ML}(\widehat{g}_N^{\widetilde{\mathcal{M}}}) - J_{ML}(\widehat{g}_N^{ML}) \tag{6.47}$$

from which

$$\frac{J_{ML}(\widehat{g}_N^{\widetilde{\mathcal{M}}}) - J_{ML}(\widehat{g}_N^{ML})}{J_{ML}(\widehat{g}_N^{ML})} \cdot \frac{N-d}{d-d_{\widetilde{\mathcal{M}}}} \in F(d - d_{\widetilde{\mathcal{M}}}, N - d) \tag{6.48}$$

This is the classical F-test used to discriminate between two model structures with complexities d and $d_{\widetilde{\mathcal{M}}}$ $(d > d_{\widetilde{\mathcal{M}}})$.

6.1.1 Model order selection procedure

The previous considerations can be directly exploited in a model selection setting. Namely, suppose to have a set of model structures $\left\{\widetilde{\mathcal{M}}_k,\ k = 1, 2, ...\right\}$ and let $\chi^2_\alpha(d - d_{\widetilde{\mathcal{M}}_k})$ be defined by

$$\alpha = P\left(x > \chi^2_\alpha(d - d_{\widetilde{\mathcal{M}}_k})\right) \tag{6.49}$$

with x being a χ^2-distributed random variable. Then, the hypothesis test derived from (6.44) allows to reject the model structures $\widetilde{\mathcal{M}}_k$ for which the inequality

$$\frac{J_{MR}(\hat{g}_N^{\widetilde{\mathcal{M}}_k})}{\hat{\sigma}^2} \leqslant \chi^2_\alpha(d - d_{\widetilde{\mathcal{M}}_k}) \tag{6.50}$$

does not hold, i.e. the ones that are not able to properly "explain" the data. However, what still has to be defined is the way in which to discriminate among the unfalsified model structures, that is the ones for which the inequality (6.50) holds. Assume that $\left\{\widetilde{\mathcal{M}}_k,\ k = 1, 2, ...\right\}$ is a set of nested model structures, which can be ordered according to their complexities: $\widetilde{\mathcal{M}}_1 \subset \widetilde{\mathcal{M}}_2 \subset ...$, since $D_{\widetilde{\mathcal{M}}_1} \subset D_{\widetilde{\mathcal{M}}_2} < ...$. In this case, the so-called "parsimony" principle is exploited: hence, among the unfalsified model structures, the simplest one is chosen. This makes also clear the meaning of the name "PUMS" (Parsimony Unfalsified Model Structure Selection), given to this order selection criterion.

The hypothesis test (6.50) based on the χ^2-distribution can be performed also exploiting the F-distribution, but using the value in (6.46).

6.2 Experimental results

6.2.1 Implementation of the method

The method previously described is here applied to discriminate among OE models of different complexities. As previously done, 200 systems in each of the four data sets illustrated in Section 3.1 are identified; furthermore, OE models

with orders ranging from 1 to 40 are evaluated.
Furthermore, as was done for the other criteria involving statistical tests, when none of the evaluated model structures satisfies the condition (6.50) for a fixed significance level α, the most complex model structure is chosen by default.

The order selection technique denoted by PUMS has been implemented through the following steps:

1. The unstructured least-squares estimate $\hat{g}_N^{ML} = \hat{g}_N^{LS}$ is determined by estimating a high-order FIR model (the order is set to $N/3$);

2. The model described by $\hat{g}_N^{ML}$ is simulated, obtaining the output data $\hat{Y}_N = \Phi_N^T \hat{g}_N^{ML}$.

3. Starting from the smallest complexity among the considered ones, an OE model $\hat{g}_N^{\widetilde{\mathcal{M}}_k}$ is estimated by solving

$$\begin{aligned} \hat{g}_N^{\widetilde{\mathcal{M}}_k} = \quad & \arg \quad \min_g (\hat{Y}_N - \Phi_N^T g)^T (\hat{Y}_N - \Phi_N^T g) \\ & s.t. \quad g \in \Xi_{\widetilde{\mathcal{M}}_k} \end{aligned} \tag{6.51}$$

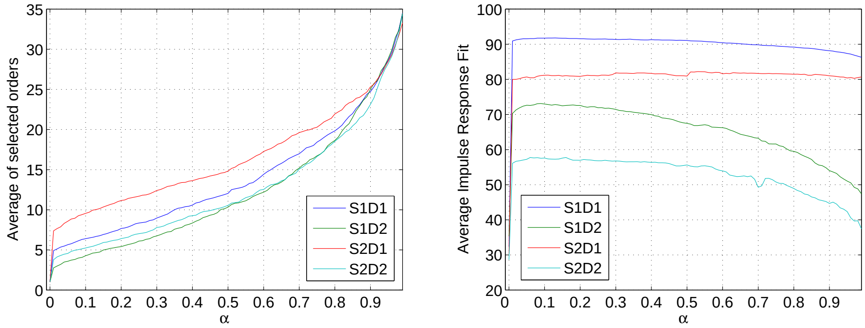
   In practice the OE model $\hat{g}_N^{\widetilde{\mathcal{M}}_k}$ is estimated from $\hat{Y}_N$ using the routine `pem` with no noise modelling.

4. If the inequality (6.50) holds, the procedure is stopped and the model structure $\widetilde{\mathcal{M}}_k$ is chosen, otherwise the procedure is iterated on more complex model structures.

5. A state-space model with no noise description and of the chosen complexity is finally estimated from the original data using the routine `pem`.

6.2.2 Selection of the significance level

In Chapter 4 a specific analysis has been conducted for the choice of the significance level, when an order selection criterion involved a statistical test. Since PUMS is based on the test (6.50), an analogous investigation is here performed. The previous investigations have shown that a large significance level could decrease too significantly the rate of unfalsified model structures among the tested ones, thus making the test ineffective. On the other hand, a too small significance level, combined with the "parsimony" property of PUMS could give rise to an undermodelling tendency.



(a) Average of the selected orders.

(b) Average impulse response fits.

Figure 6.1: OE models - Average of the selected orders and average impulse response fits achieved by PUMS, as function of the significance level α adopted in the statistical test (6.50). The average is calculated from the identification of 200 systems in each data set.

Figure 6.1.(b) shows the average impulse response fits achieved in the four data sets, as functions of the significance level α adopted in the test (6.50). The trends are similar to the ones reported in Figure 4.9.(b) for the combination of the whiteness and the independence tests on the residuals: while the selection of the significance level is not so influential in the less-noisy data sets S1D1 and S2D1, the adoption of too large values for α leads to a significant decrease of the impulse response fit. Indeed, as was previously observed for the whiteness test and for the combination denoted by RA, when α is too large, the risk that no model structure can be unfalsified by the test increases, causing the default selection of the most complex model structure. This is further confirmed by the plot showing the averages of the selected orders as functions of α (Figure 6.1.(a)): the fast increase which is observed for growing values of α is the direct consequence of the default selection of the largest order. While this choice is not so detrimental for the "slow" and non-noisy data set S2D1, it gives rise to overfit in S1D2 and S2D2.

The previous considerations, combined with the results observed in Figure 6.1.(b) suggest that the test in (6.50) is more effective when small significance levels α are adopted. A further confirm of this result comes also from Table 6.1 which contains the values of α leading to the highest average impulse response fit in each data set. The unique exception is detected in S2D1, where the selection of large complexities is beneficial and it is made possible by the use of a large significance level.

Set	α
S1D1	0.13
S1D2	0.09
S2D1	0.53
S2D2	0.06

Table 6.1: OE models - Values of the significance level α which guarantee the best average impulse response fits when adopted in the statistical test (6.50).

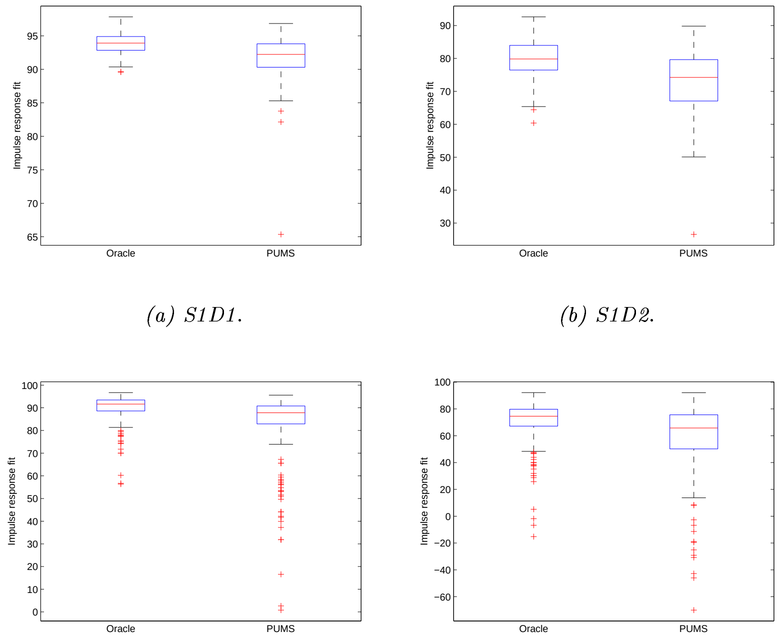
6.2.3 Analysis of PUMS performances

	Impulse Response Fit		Type 1 Pred. Fit		Type 2 Pred. Fit	
	Oracle	*PUMS*	*Oracle*	*PUMS*	*Oracle*	*PUMS*
S1D1	93.9	91.8	94.5	92.5	94.3	92.3
S1D2	79.9	73.1	81.8	75.9	81.2	74.5
S2D1	89.9	82.2	91.6	85.6	90	82.5
S2D2	69.9	57.8	75.7	66.1	72.8	61.5

Table 6.2: OE models - Average fits achieved by PUMS when 200 systems are identified in each data set.

Table 6.2 contains the average fits achieved by PUMS in the identification of 200 systems in each of the four data sets; the values of α reported in Table 6.1 are adopted in the test (6.50). Compared to the average fits achieved by the classical order selection criteria and reported in Table 4.2, PUMS behaves worse than BIC and RA, while it leads to performances almost equivalent to the ones reached by cross-validation in the "fast" data sets, outperforming it in the "slow" ones.
A further comparison also shows that PUMS and the independence tests on the residuals (RAI1 and RAI2) achieve very similar results, even if RAI2 (but also RAI1, apart in data set S2D1) outperforms PUMS. Indeed, an analogy between these two tests was previously pointed out: namely, below equation (6.30) it has been highlighted how the quantities $J_{MR}(\widehat{g}_N^{\widetilde{\mathcal{M}}})$ (exploited by PUMS) and $x_{\varepsilon u}^{N,M}$ (used by the independence test in (2.19)) are comparable, under certain assumptions. However, thanks to the interpretation of the estimation problem (6.22) as a model reduction problem, a different test is exploited in PUMS.

The box plots of the impulse response fits in Figure 6.2 prove a certain robustness of PUMS in the "fast" data sets S1D1 and S1D2, since very few outliers are visible. This property is further confirmed by Figure 6.3, where the histograms of the differences between the orders chosen by the oracle and the ones selected by PUMS are well centered around 0, even if a slight overmodelling is observed. The situation is different for the "slow" sets S2D1 and S2D2, whose box plots present more outliers. However, this kind of performances deterioration in S2D1



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 6.2: Box plots of the impulse response fits achieved by PUMS when 200 systems are identified in each data set.

and S2D2 was detected also in Chapter 4 for the classical order selection methods.

The increased number of outliers detected in Figures 6.2.(c) and (d) is mainly due to the adoption of the "parsimony" principle: the choice of the simplest model structure which satisfies the condition in (6.50) gives rise to undermodelling, when slow systems have to be reproduced. The adoption of a large significance level α in S2D1 reduces this problem, thanks also to the default selection of the largest order, when no model structure among the tested ones is unfalsified by the test. However, the outliers visible in the box plot of the fits achieved by PUMS in S2D1 (Figure 6.2.(c)) are due to the selection of lower complexities w.r.t. the oracle choices. This justification is valid also for the outliers present for PUMS in Figure 6.2.(d). Again, Figure 6.3 confirms these considerations: the histograms of the differences between the orders choices done by the oracle and by PUMS show an undermodelling tendency in data sets S2D1 and S2D2, which is not present in S1D1 and S1D2.

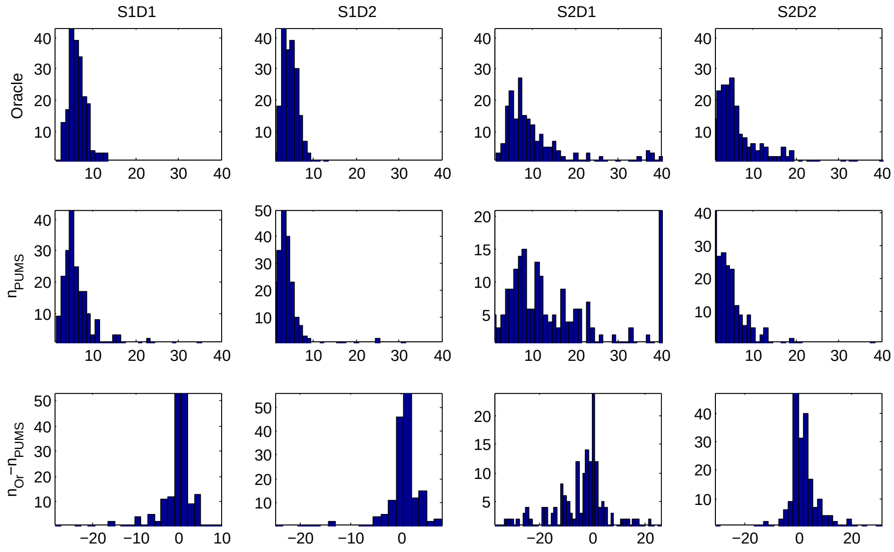


Figure 6.3: Histograms of the orders selected by PUMS in the identification of 200 systems in each data set. The histograms of the differences w.r.t. to the oracle choices are also shown.

In spite of the undermodelling observed in the "slow" data sets, PUMS typically does not suffer of early stopping due to local minima, which was instead detected for the F-test. Therefore, the PUMS criterion appears robust also w.r.t. to the so-called local minima issue.

6.2.4 Combination with the validation methods

The analysis of PUMS performances done in Section 6.2.3 has highlighted how the parsimony property adopted by PUMS can limit its performances, especially when slow systems have to be reproduced. Therefore, the equipment of PUMS with some other tests or its combination with other order selection criteria could reveal itself more effective. The combination of PUMS with the validation methods (i.e. with the tests on the residuals) could be helpful in this sense; hence it is analyzed in the following.

In this setting the tests on the residuals exploit the estimation data in order to evaluate the model estimated at point 5 of the procedure in Section 6.2.1. The final model structure selected by the joint use of PUMS and a validation method is the simplest one for which both the PUMS test in (6.50) and the considered

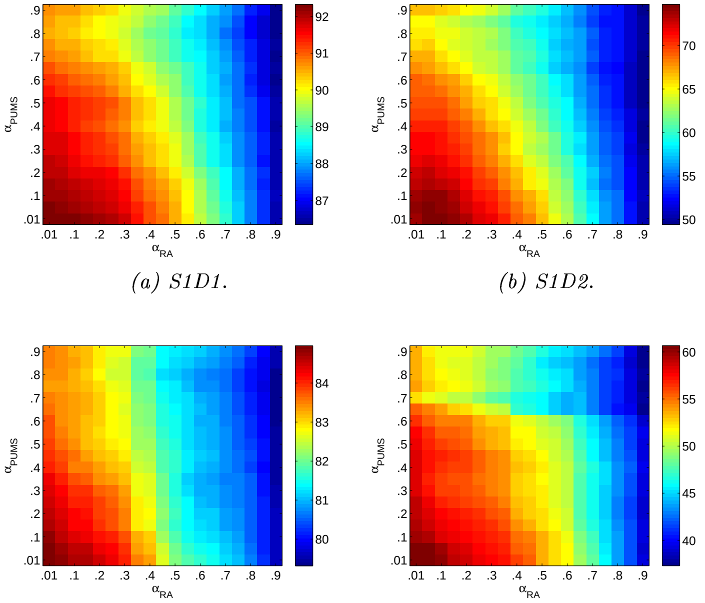
test on the residuals are passed.
Again, the latter one, that could be the whiteness test in (2.7) or the independence test in (2.19) (or both) is applied setting $M = n_b + 20$; recall that n_b is the order of the polynomial $B(q)$ in (1.13) and M is the maximal number of lags for which the auto- or the cross-correlation of the residuals is computed.

Table 6.3 illustrates the average impulse response fits achieved by the equipment of PUMS with the different tests on the residuals. The results refer to the use of the optimal significance levels in both the statistical tests involved, that is in the test (6.50) and in the residuals test. It is clear that the combination is beneficial in all the four data sets and with all the types of tests on the residuals, since the average fits are always higher than the ones which refer to PUMS criterion alone. The best fits are reached using the independence test on the residuals in the "fast" data sets S1D1 and S1D2, while RA leads to the best performances in the "slow" sets. This kind of result was already observed in Chapter 4, when the validation tests were combined with the classical order selection techniques. Indeed, the tendency of RA to select quite complex model structures is more beneficial when slow systems have to be identified. On the other hand, the independence test on the residuals generally unfalsifies quite high-order models, thus leading to few changes when this type of models are already selected by the previously applied criterion. In the "slow" data sets, the adoption of a larger value for M could probably lead to better performances when RAI1 or RAI2 are applied after PUMS.

	Average impulse response fit					Average of the selected orders				
Set	*PUMS*	*+RAW*	*+RA*	*+RAI1*	*+RAI2*	*PUMS*	*+RAW*	*+RA*	*+RAI1*	*+RAI2*
S1D1	91.8	91.9	92.3	92.9	92.8	6.7	7	5.3	6	5.9
S1D2	73.1	73.1	74.7	76.5	76.4	4.1	4.9	3.9	3.9	4.1
S2D1	82.2	85	85.3	82.3	83.5	15.6	12.3	11.2	9.3	8.3
S2D2	57.8	59.8	60.7	60.4	59.8	4.8	7.2	6.4	5.9	5.8

Table 6.3: OE models - Average impulse response fits and average of the selected orders when validation methods are combined with PUMS for model order selection in the identification of 200 systems in each data set.

Table 6.3 also shows the average of the orders selected in the four data sets by PUMS and by its combination with one of the four validation methods here considered. In some data sets an unexpected decrease of the average is observed after the equipment with the tests on the residuals: this is due to the adoption of a different significance level in the statistical test exploited by PUMS. As the achieved fits proved, this decrease is beneficial in data sets S1D1 and S1D2, where an overmodelling tendency of PUMS was highlighted by the histograms in Figure 6.3. In addition, the selection of simpler structures w.r.t. the ones originally chosen by PUMS is beneficial also in data set S2D1: indeed, the joint use of two statistical tests allows to adopt a smaller significance level in the



(a) S1D1.

(b) S1D2.

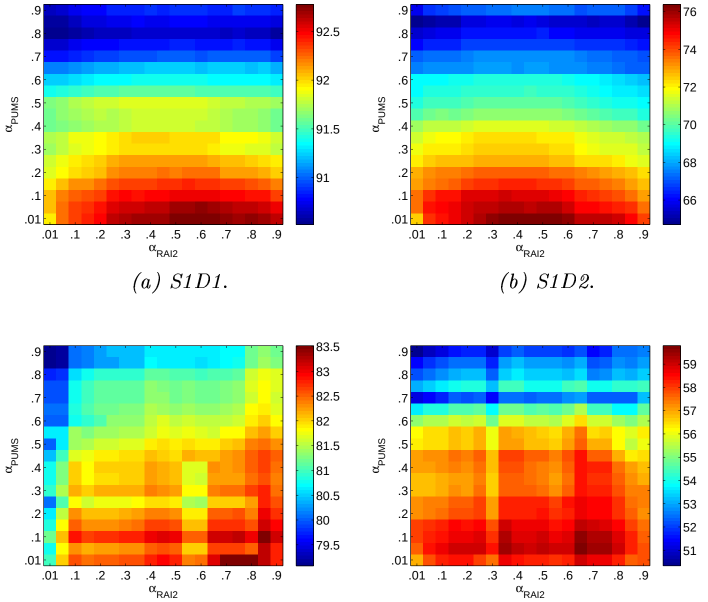
(c) S2D1.

(d) S2D2.

Figure 6.4: OE models - Average impulse response fits achieved for different values of the significance levels when both the tests on residuals (RA) are combined with PUMS for model order selection in the identification of 200 systems in each data set.

test for PUMS, thus increasing the unfalsification rate of the test. Previously, when no model structure was unfalsified by the test used by PUMS, the most complex model structure was by default selected. In this case, instead, thanks to the exploitation of another test, these "default" choices are replaced by more precise ones, leading to improvements in the impulse response fit.

The previous considerations find a further confirm in Figure 6.4. It shows the average impulse response fits achieved by the combination of PUMS with RA for different values of the significance levels used in the two tests. In all the data sets very small values are preferable for both the tests, otherwise too complex model structures are selected.
Different trends are observed in Figure 6.5, which refers to the equipment of PUMS with the second implementation of the test for the independence of the residuals from past inputs (RAI2). While small significance levels are still preferable for PUMS, values of α_{RAI2} around 0.5 or even larger are more beneficial for the effectiveness of the test on the residuals. Indeed, in this way the rate of model structures that are unfalsified by that test decreases, thus leading to



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 6.5: OE models - Average impulse response fits achieved for different values of the significance levels when the independence test on the residuals (RAI2) is combined with PUMS for model order selection in the identification of 200 systems in each data set.

some modifications in the previous order choices done by PUMS.

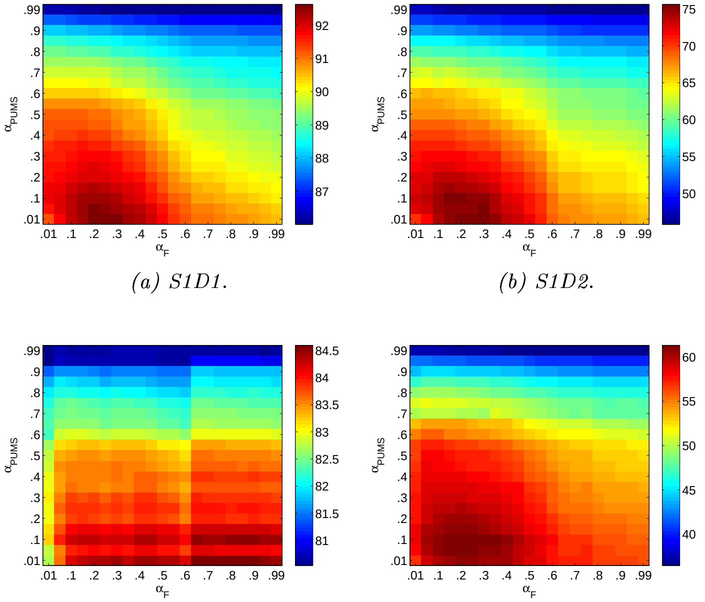
6.2.5 Combination with the F-test

The considerations done at the beginning of Section 6.2.4 suggested the equipment of PUMS with some other tests or criteria, in order to alleviate the limitations derived from the use of the "parsimony" principle. Therefore, this section evaluates the combination of PUMS with the F-test, according to the modality described in Section 2.3.2. By means of the F-test, model structures of increasing complexities are evaluated, starting from the ones returned by PUMS, i.e. the ones estimated at point 5 of the procedure in Section 6.2.1. As soon as the most complex model structures between the two compared ones does not pass the F-test, the procedure is stopped and the simplest one is finally chosen.

Table 6.4 contains the average impulse response fits achieved by PUMS and by

	Average impulse response fit		Average of the selected orders	
Set	*PUMS*	*PUMS + F (Up)*	*PUMS*	*PUMS + F (Up)*
S1D1	91.8	92.6	6.7	6.9
S1D2	73.1	75.7	4.1	4.1
S2D1	82.2	84.6	15.6	9.9
S2D2	57.8	61.3	4.8	5.7

Table 6.4: OE models - Average impulse response fits and average of the selected orders when the F-test is applied after PUMS to perform model order selection in the identification of 200 systems in each data set.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 6.6: OE models - Average impulse response fits achieved for different values of the significance levels when the F-test is applied after PUMS in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

its equipment with the F-test, when the optimal values of the significance levels are adopted in both the tests. In all the data sets the combination leads to improvements in terms of impulse response fits.

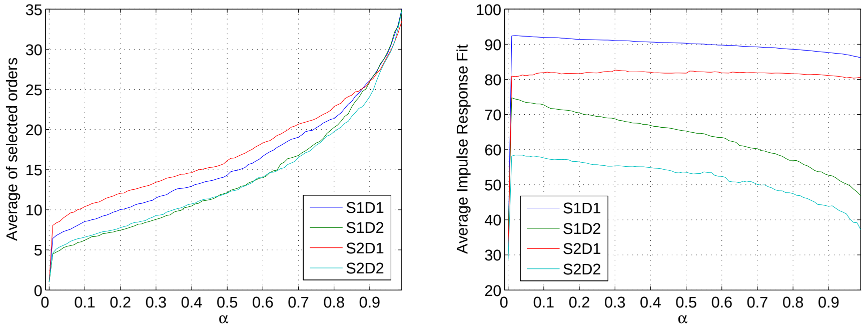
Again, the decrease observed in the average of the selected orders in data set S2D1, after the application of the F-test, is due to the use of a different significance level in the test for PUMS. As observed for the combination of PUMS with the validation methods, the exploitation of two tests allows to adopt a smaller significance level in the test used by PUMS, since undermodelling is avoided by the successive application of the F-test.

The analysis of the combination here considered is completed by Figure 6.6, which shows the impact of the significance levels used in the test for PUMS and in the F-test on the achievable performances. Except in data set S2D1, the analysis done in Section 6.2.2 is here confirmed, since PUMS is still more effective when small values of the significance level are adopted. Also the F-test should be applied with a small value of α_F, in order to avoid the selection of too complex model structures. Different considerations hold for data set S2D1, where α_F can be chosen quite large, in order to favour the selection of higher model orders. This result is in line with the findings observed in Chapter 4, when the F-test was applied after the classical order selection methods.

6.2.6 Modification of PUMS criterion

The previous analysis of PUMS performances has highlighted the need to combine it with some other methods, in order to enforce the check of the model quality and to alleviate the undermodelling tendency detected when slow systems have to be identified. Therefore, in addition to the two combinations described in Sections 6.2.4 and 6.2.5, the joint use of PUMS with other methods has been also tested. Particular attention has been devoted to the application of tests on the model $\widehat{g}_N^{\widetilde{\mathcal{M}}_k}$ estimated at point 3 of the procedure in Section 6.2.1. The idea is to evaluate whether those tests are more effective when applied on models estimated from simulated data, which theoretically are free of the measurement noise.
Among the evaluated combinations, the simultaneous use of the test in (6.50) and of the F-test applied on the model $\widehat{g}_N^{\widetilde{\mathcal{M}}_k}$ has proved the most effective one. According to this order selection method, the complexity of the first model structure which passes both the tests is selected; then, a state-space model of that complexity (with no noise modelling) is estimated from the original input-output data.
In the following this order selection method will be denoted as PUMS+F(Sim).

As usual, a first analysis is conducted on the choice of the significance level to adopt in the two statistical tests involved in this criterion. Specific experiments were done, in which the two tests were applied with different significance levels. However, the achieved performances were equivalent to the ones obtained setting the same significance level for both the tests. Therefore, the analysis is here conducted with this simplification.



(a) Average of the selected orders.

(b) Average impulse response fits.

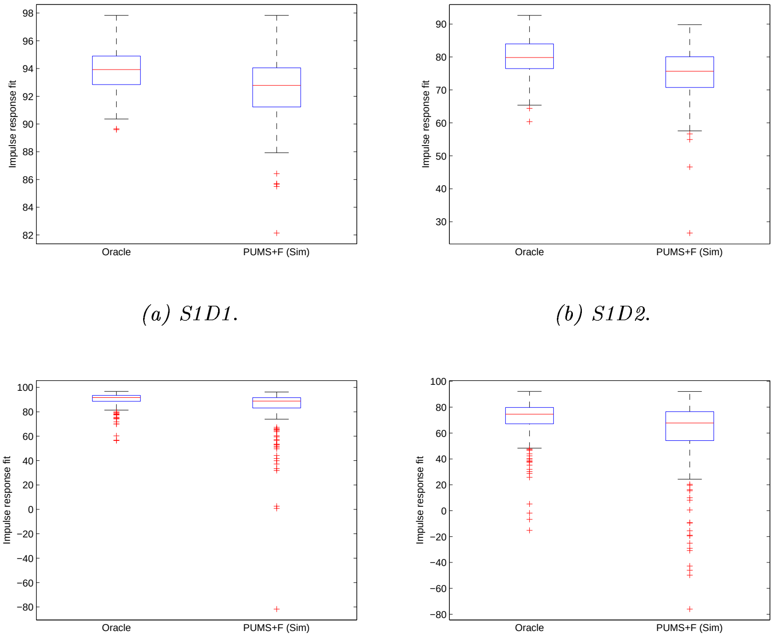
Figure 6.7: OE models - Average of the selected orders and average impulse response fits achieved by PUMS+F(Sim) for different values of the significance level α adopted in both tests (6.50) and (2.63). The average is calculated from the identification of 200 systems in each data set.

Figure 6.7 shows results which agree with the ones in Figure 6.1: small values of the significance level used in both the tests are preferable in data sets S1D1, S1D2 and S2D2, while the performances reached in data set S2D1 appear less affected by the choice of this value.

	Impulse Response Fit			Type 1 Pred. Fit			Type 2 Pred. Fit		
Set	*Or.*	*PUMS*	*PUMS + F (Sim)*	*Or.*	*PUMS*	*PUMS + F (Sim)*	*Or.*	*PUMS*	*PUMS + F (Sim)*
S1D1	93.9	91.8	92.5	94.5	92.5	93.2	94.3	92.3	92.9
S1D2	79.9	73.1	74.8	81.8	75.9	77.5	81.2	74.5	76.5
S2D1	89.9	82.2	82.6	91.6	85.6	86.1	90	82.5	83
S2D2	69.9	57.8	58.5	75.7	66.1	67.5	72.8	61.5	62.6

Table 6.5: OE models - Average fits achieved by PUMS and PUMS+F(Sim) when 200 systems are identified in each data set.

The average fits achieved by the procedure PUMS+F(Sim) are reported in Table 6.5, together with the ones reached by the oracle and by PUMS. A performances improvement is observed in all the data sets w.r.t. the results obtained by PUMS. Indeed, the box plots of the impulse response fit reported in Figure 6.8 appear more compact, when compared to the ones related to PUMS (Figure 6.2). This property is mainly observed for the "fast" data sets S1D1 and S1D2, while some outliers are still present in the box plots referring to data sets S2D1 and S2D2. Figure 6.9 helps to understand these findings. The histograms of the differences between the orders chosen by the oracle and the ones selected by PUMS+F(Sim) prove a significant agreement between the two criteria in data sets S1D1 and S1D2, while PUMS+F(Sim) suffers of undermodelling in



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure 6.8: Box plots of the impulse response fits achieved by PUMS+F(Sim) when 200 systems are identified in each data set.

S2D2. Different results are observed in data set S2D1, where both over- and undermodelling is detected; indeed, further investigations have proved that the outliers present in Figure 6.8.(c) are due to both these tendencies.

6.2.7 Different initialization for PEM

The so-called PUMS criterion has been tested also using a different initialization of the MATLAB routine `pem`. The previous results have been obtained exploiting the default initialization of `pem` (refer to [9] for details).
The results presented in this section are obtained by initializing the routine `pem` with a state-space model returned by a model reduction problem.

From (6.9) notice that

$$\widehat{g}_N^{ML} - g_0 = \widehat{g}_N^{LS} - g_0 = (\Phi_N \Phi_N^T)^{-1} \Phi_N \Lambda_N = R_N^{-1} \Phi_N \Lambda_N \tag{6.52}$$

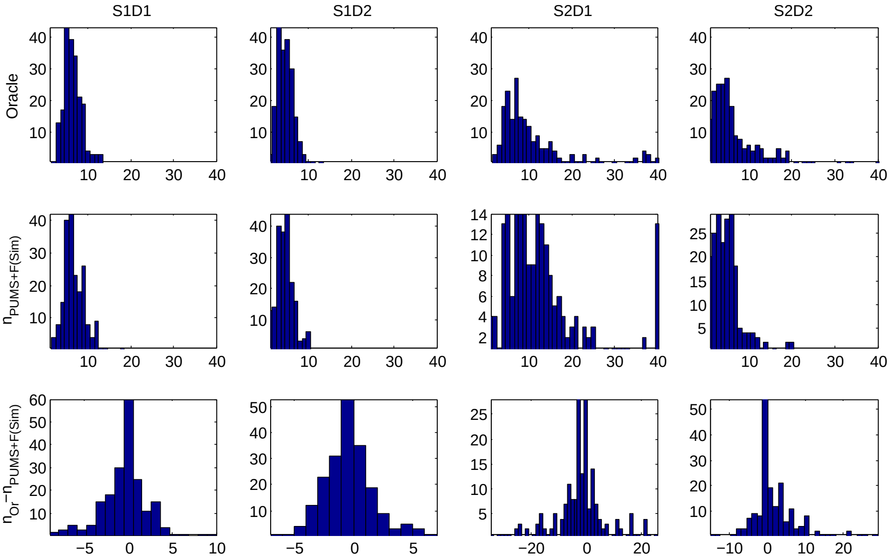


Figure 6.9: Histograms of the orders selected by PUMS+F(Sim) in the identification of 200 systems in each data set. The histograms of the differences w.r.t. to the oracle choices are also shown.

with $R_N = \Phi_N \Phi_N^T$. Hence, recalling that $\Lambda_N \in \mathcal{N}(0_{N\times 1}, \sigma^2 I_N)$, we have

$$\begin{aligned} E\left[(\hat{g}_N^{ML} - g_0)(\hat{g}_N^{ML} - g_0)^T\right] &= E\left[(R_N^{-1}\Phi_N\Lambda_N)(R_N^{-1}\Phi_N\Lambda_N)^T\right] \\ &= \sigma^2 R_N^{-1}\Phi_N\Phi_N^T R_N^{-1} \\ &= \sigma^2(\Phi_N\Phi_N^T)^{-1} \end{aligned} \tag{6.53}$$

From (6.52) and (6.53), we derive that

$$\hat{g}_N^{ML} - g_0 \in \mathcal{N}(0_{d\times 1}, \sigma^2 R_N^{-1}) \tag{6.54}$$

from which

$$\begin{aligned} \frac{J_{MR}(g_0)}{\sigma^2} &= \frac{1}{\sigma^2}(\hat{g}_N^{ML} - g_0)^T R_N (\hat{g}_N^{ML} - g_0) \\ &= \frac{1}{\sigma^2}(\hat{g}_N^{ML} - g_0)^T \Phi_N\Phi_N^T(\hat{g}_N^{ML} - g_0) \in \chi^2(d) \end{aligned} \tag{6.55}$$

with $d = \dim \hat{g}_N^{ML} = \dim g_0$.
An initial state estimate for the routine `pem` is thus derived by solving the following minimization problem

$$\begin{aligned} \arg \quad & \min_g \|H(g)\|_* \\ s.t. \quad & \frac{1}{\hat{\sigma}^2}(\hat{g}_N^{ML} - g)^T \Phi_N\Phi_N^T(\hat{g}_N^{ML} - g) \leqslant \chi^2_\alpha(d) \end{aligned} \tag{6.56}$$

where $H(g)$ denotes the Hankel matrix of the model represented by g, $\|\cdot\|_*$ denotes the nuclear norm operator and $\chi^2_\alpha(d)$ is the α-level of the $\chi^2(d)$-distribution.
The optimization problem (6.56) tries to find the simplest model g which is considered able to "explain" the data, i.e. which falls into the ellipsoids in $\mathbb{R}^d$ defined by

$$\frac{1}{\widehat{\sigma}^2}(\widehat{g}_N^{ML} - g)^T \Phi_N \Phi_N^T (\widehat{g}_N^{ML} - g) \leqslant \chi^2_\alpha(d) \tag{6.57}$$

In order to exploit this new initialization, points 3 and 5 of the implementation procedure illustrated in Section 6.2.1 are modified. Namely, before applying the routine `pem`, the optimization problem (6.56) is solved and the returned FIR model is realized as a state-space model. By means of the routine `balred`, this is then truncated to the desired complexity $d_{\widetilde{\mathcal{M}}_k}$ and the obtained model is passed as initial estimate to the routine `pem` exploited at points 3 and 5.

The results that will be reported in the following refer to the use of both the initializations, i.e. the `pem` default one and the one here illustrated. Namely, at points 3 and 5 of the procedure in Section 6.2.1 both the initializations are tried and the estimated model giving the lowest loss is then kept. It has to be clarified that, at point 4 of the procedure, the inequality (6.50) is evaluated only on the model which gives the lowest loss.

	Impulse Response Fit	
Set	*PUMS - Default Init*	*PUMS - Both Init*
S1D1	91.6	91.8
S1D2	72.6	73.1
S2D1	80.7	87.4
S2D2	57.2	65.3

Table 6.6: OE models - Average impulse response fits achieved by PUMS in the identification of 200 systems in each data set, when different initializations are used for the MATLAB routine *pem*.

	Impulse Response Fit	
Set	*PUMS+RAI2 - Default Init*	*PUMS+RAI2 - Both Init*
S1D1	92.4	92.7
S1D2	75.5	76
S2D1	82.2	88.4
S2D2	58.9	67.1

Table 6.7: OE models - Average impulse response fits achieved by the combination of PUMS with RAI2 in the identification of 200 systems in each data set, when different initializations are used for the MATLAB routine *pem*.

Table 6.6 contains the average impulse response fits achieved by PUMS when only the internal initialization of pem is used and when both the initializations are exploited. The test (6.50) is applied with $\alpha = 0.05$. A significant improvement is detected in the "slow" data sets when the two initializations are combined. When slow systems have to be reproduced, the initialization introduced in this section is probably more indicated, since it exploits a high-order FIR model which better catches the slow dynamics of the true system.

The performances reported in Table 6.6 can be further improved when PUMS is combined with the validation methods, according to the way described in Section 6.2.4. Table 6.7 illustrates the average fits achieved when PUMS is combined with the second implementation of the independence test on the residuals (RAI2): the two columns respectively refer to the cases in which only the own initialization of the routine pem is used and when also the one here introduced is exploited. The average fits reported in the table are obtained with the significance levels $\alpha_{PUMS} = 0.05$ and $\alpha_{RAI2} = 0.2$. Again, the use of both the initializations is much more beneficial in the "slow" data sets S2D1 and S2D2, where the average fits achieved are the best ones so far observed.

6.2.8 Conclusions

The experimental analysis here conducted on PUMS has shown how it is more effective when a small significance level is adopted in the statistical test (6.50). Moreover, PUMS appears robust when applied for the identification of fast systems, while it is a bit penalized by its parsimony when slow systems have to be reproduced. It overall does not suffer from early stopping due to local minima.

When PUMS is combined with other methods, such as the validation ones or the F-test, as illustrated in Sections 6.2.4 and 6.2.5, a significant improvement is detected in all the four data sets, especially in the "slow" ones. Again, in this setting, small significance levels are preferable for the test used by PUMS and also for the tests with which it is combined (with the only exception given by RAI2).

The exploitation of a new type of initialization for the routine pem has proved to be beneficial, since better impulse response fits have been reached, especially in the "slow" data sets. In addition, when PUMS is combined with a validation method and when also this new initialization is exploited, a further performances improvement can be obtained.

Chapter 7

Conclusions

The thesis has focused on the techniques which are adopted for model order selection in system identification. Both classical and innovative methods have been presented: among the first ones cross-validation, the information criteria (FPE, AIC and BIC), the F-test and the residual analysis have been considered (Chapters 2 and 4), while new techniques, such as kernel-based estimation and PUMS have been illustrated and evaluated (Chapters 5 and 6, respectively).

A theoretical description of these techniques has been provided and accompanied by an experimental analysis. For this purpose, four different data sets have been exploited (see Section 3.1): they contain a wide range of systems, which are commonly faced and identified in practical situations; however, the available data are generally less than the ones used in real-life scenarios.
Those systems have been mainly identified through OE models, thus testing the order selection methods on this specific model type. However, Appendix A illustrates also the performances of the classical order selection techniques, when they are applied to choose an appropriate order for FIR, ARX and ARMAX models.

Three fit measures have been introduced in order to quantify the adherence between the true system and the estimated one: one of them compares their impulse responses, while the other two measures evaluate the prediction ability of the estimated models. Major importance has been given to the impulse response fit, since it directly compares the intrinsic properties of the true system and of the estimated one: indeed, a high degree of correspondence between the impulse responses leads also to a high agreement in terms of prediction.

It should be specified that the conclusions that follow are drawn from the experiments on the specific data sets illustrated in Section 3.1. They reproduce a special situation, since they contain relatively few data for the identification of quite complex systems. The model order selection criteria have not been evaluated according to their ability to identify the true order of the systems, but to find orders that allow a good reproduction of the input-output properties (impulse responses) of the true systems. Therefore, the criteria may behave differently with other kinds of data sets, such as longer ones.

Classical order selection techniques

The analysis done in Chapter 4 in relation to OE models and in Appendix A for FIR, ARX and ARMAX models has shown how none of the classical order selection techniques generally outperforms the others, independently from the specific model type. However, the method here denoted as RA, which exploits both the whiteness and the independence tests on the residuals appears among the best methods, when applied on OE, ARX and ARMAX models, but it shows some difficulties when it has to discriminate among FIR models. Analogous performances have been detected for RAI2, i.e. the method which tests the independence of the residuals from past inputs, when the residuals are not assumed white: it gives good performances when it is applied on OE, ARX and ARMAX models but it is outperformed by RAI1 on FIR models, i.e. by the other implementation of the independence test which assumes the residuals whiteness.

If the order selection criteria based on the residual analysis have proved to be among the most effective ones, their performances are strongly influenced by the tuning of two parameters: the maximal lag M for which the residuals auto- or cross-correlation is computed and the significance level α adopted in the statistical tests. Specific analysis have been conducted in order to understand their impact on the performances of the criteria.
Section 4.1.2 has shown how small values of M are preferable when residuals whiteness has to be assessed: namely, the highest average impulse response fits were reached when $M = 20$. On the other hand, when the independence between residuals and past inputs is evaluated, small values of M are more suitable ($20 \leqslant M \leqslant 40$) when fast systems have to be identified, while larger values of M ($M \geqslant 40$) are preferable for slow systems. These considerations have been drawn with reference to OE models, but they hold also for the other model types
.

For what regards the significance level α instead, its tuning should take into account the estimated model type. Indeed, small values of α increase the unfalsification probability of the whiteness test on FIR and OE models, resulting in more reliable order estimations. On the other hand, the residuals obtained for ARX and ARMAX models benefit of the whitening effect performed by the polynomials $A(q)$ and $C(q)$: thus, to avoid the so-called undermodelling tendency, large significance levels are preferable in this case. Passing to the independence test (RAI1 or RAI2), its performances appear quite robust w.r.t. the value of the significance level when it is applied on OE and ARMAX models, even if too large and too small values should be avoided. The only exception has been detected when slow systems have to be identified and the measurement noise in the data is limited: in this case, large values of α are preferable, in order to alleviate the undermodelling tendency. For the same reason large significance levels are always suggested when the order has to be selected for FIR and ARX models. When both the whiteness and the independence tests are exploited, the tuning of the significance level has to be done according to the test which leads to the selection of larger orders, when applied alone: namely, according to the whiteness test for FIR and OE models and to the independence test for ARX and ARMAX models.

A summary of the performances reached by the classical order selection methods is now provided. The acronyms here used refer to the ones introduced at the beginning of Chapter 4.

RAW - When applied alone on all the considered model types, it does not lead to high impulse response fits. Its equipment with RAI1 allows a significant performances improvement. Moreover, when the criterion is applied on FIR and OE models, the test sometimes is not able to unfalsify at least one of the evaluated model structures, even when small significant levels are adopted. This issue is more frequent when slow systems have to be identified.

$RAI1$, $RAI2$ - The two implementations of the test for the independence of the residuals from past inputs lead to very similar performances with all the model types here tested. The major discrepancy has been observed with FIR models, for which RAI1 outperforms RAI2. This order selection method appears among the best criteria with ARX and ARMAX models, but it gives satisfying performances also with FIR and OE models. It is a bit penalized when slow systems are identified, since the tests here performed exploit a too small value of M.

RA - It is among the best criteria for OE, ARX and ARMAX models, while it is penalized by the unfalsification problem brought by the whiteness test, when it is applied on FIR models. Indeed, this issue is more frequent with this model type than with the other ones. Furthermore, it generally outperforms RAI1 and RAI2 when slow systems have to be identified.

F - When applied alone, it does not lead to good performances. With OE and ARMAX models it also suffers from early stopping due to the presence of local minima. For what regards the tuning of the significance level, it is not so straightforward for OE models, while large values are preferable when the F-test is applied on ARX, ARMAX and FIR models.

CV - It behaves quite well with all the model types, when fast systems are identified, while its performances significantly deteriorate when slow systems are reproduced. When OE or ARMAX models are estimated, it can be penalized by the presence of local minima.

FPE - It is one of the worst criteria, because of its strong overmodelling tendency. It gives satisfying performances only when ARX and ARMAX models are exploited to identify slow systems, using almost noise-free data.

AIC - It behaves quite well with FIR and ARX models, while it is heavily penalized by its overmodelling tendency when it is applied on OE and ARMAX models.

BIC - It is the best criterion when applied on OE models, even if it has shown some difficulties when fast systems have to be identified exploiting noisy data. It gives good performances also with ARMAX models, while it suffers from undermodelling when applied on FIR and ARX models.

New order selection techniques

In addition to the classical order selection methods, the thesis has also presented and tested two new techniques. Both of them have been applied only for the choice of OE model orders.

The kernel-based estimation methods, illustrated in Chapter 5, have proved to be the most effective techniques when noisy data are exploited, even when compared to the classical methods. Their combination with PEM procedures (described at the end of Section 5.1) is not beneficial in these conditions. On the other hand, when the noise present in the data is not so relevant, BIC, RAI1, RAI2 and RA outperform the kernel-based methods, which in this setting benefit from their combination with PEM.

The tests performed adopting PUMS method in Chapter 6 have shown how this criterion leads to quite good performances in all the considered data sets, even if it is generally outperformed by RAI2, RA and BIC. The combination of the test

exploited by PUMS with the F-test according to the way illustrated in Section 6.2.6 has led to an improvement of the performances achieved by PUMS alone. Since PUMS is based on a statistical test, a brief analysis of its impact on the effectiveness of the criterion has shown how small values are in this case preferable.
In Chapter 6 PUMS has been tested also exploiting a new initialization for the MATLAB routine `pem` (see Section 6.2.7): the simultaneous use of this new initialization and of the default one has proved to be very effective, leading to the best impulse response fits observed in the data sets containing slow systems.

Techniques combinations

In Section 2.3 two combinations of the classical model order selection techniques have been presented: the first one equips them with the tests on the residuals (Section 2.3.1), while the second-one adopts the F-test after the application of a classical order selection technique, in order to evaluate simpler or more complex model structures (2.3.2). These combinations have been considered also in Chapters 5 and 6, where they have been applied replacing the classical model order selection methods with the techniques denoted as KB+PEM and PUMS, respectively (see Sections 5.2.1 and 6.2.4).

These techniques combinations have been introduced with the purpose of reducing the risk of wrong order choices w.r.t. the application of a single method.

Combination with the validation methods

The combination involving the tests on the residuals has proved to be effective when the performances achieved by a certain order selection method are not satisfying. For instance, this is the case of the F-test, whose order choices are generally far away from the optimal ones with all the tested model types. Therefore, it significantly benefits from its equipment with the validation methods. Another example in this sense is given by cross-validation which encounters difficulties in the selection of OE model orders for the identification of slow systems: its performances encounter a certain improvement, when it is combined with the tests on the residuals. An analogous situation is detected with ARX and ARMAX models, while the improvement achieved after the combination is less significant with FIR models, since cross-validation already gives rise to good impulse response fits. This situation is encountered also in relation to BIC, whose performances on FIR, ARX and ARMAX models can be improved by its combination with the validation methods. On the other hand, no improvement

is observed on OE models, since the performances reached by BIC with this model type are already good; the only exception in this sense has been detected when fast systems have to be identified exploiting noisy data: since BIC encounters some difficulties in the order estimation in this setting, its equipment with the validation methods appears beneficial.
Passing to the combination of the tests on the residuals with the new order selection techniques, it gives rise to an improvement of the performances reached by PUMS in all the four data sets here considered. On the other hand, KB+PEM slightly benefits from this combination only in presence of noisy data, even if the application of kernel-based methods alone is still preferable in these conditions. Therefore, the exploitation of this technique combination is suggested in order to achieve satisfying performances, whenever the model order selection methods initially adopted is not the most suitable one for the specific identification setting.

Among the two tests on the residuals that can be applied in this combination, the adoption of the test for independence of the residuals from past inputs has generally proved more effective when fast systems are identified; on the other hand, for slow systems, the use of both the tests on the residuals is preferable. However, the independence test alone could be effective also with slow systems, if a larger value of M is adopted.

Since the combination here cited is based on statistical tests, an important role for its efficacy is played also by the adopted significance levels. The performed tests have shown how their optimal values in this setting agree with the ones observed when the tests have been applied alone.

Combination with the F-test

As already cited in its illustration in Section 2.3.2, the application of this technique requires to choose whether to evaluate simpler or more complex model structures, starting from the one returned by the firstly applied order selection method. The tests performed in Chapters 4, 5 and 6 have proved how this choice has to be taken according to the properties of that method: namely, if it tends to undermodel, more complex structures should be evaluated, while simpler ones should be considered, if an overmodelling tendency is observed.
The application of the F-test in this setting has proved to be particularly effective with the whiteness test (RAW) or with its combination with the independence test (RA) on FIR and OE models. In particular, the exploitation of another test allows to increase the unfalsification rate of the whiteness test, by adopting a smaller significance level.
Furthermore, also PUMS benefits from the successive application of the F-test

for evaluating more complex model structures, since the eventual limitations derived by the use of the parsimony principle can be alleviated in this way.

The efficacy of this combination still depends on the significance level adopted for the F-test: namely, its tuning should be done according to the degree of change which is desired w.r.t. to the order choices done by the firstly applied method. Generally, when more complex model structures have to be evaluated, small values of the significance level avoid the risk of overfitting, especially when noisy data are exploited. The only exceptions to this trend have been observed with FIR models and when slow systems have to be identified starting from almost noise-free data: in these cases, large significance levels for the F-test reduce the undermodelling risk.

Appendix A

Classical model order selection techniques: Application on FIR, ARX and ARMAX models

A.1 Order estimation for FIR models

The model order selection criteria are here evaluated on FIR models. The orders among which each criterion has to discriminate range from 1 to 120, with an exception when cross-validation is applied. More precisely, when it is used on the short data sets S1D2 and S2D2, the range of orders goes from 1 to 90 because the data length is too small to obtain a precise estimation of larger parameter vectors.

As for OE models, the average values that will be reported refer to the identification of 200 systems in each of the four data sets introduced in Section 3.1. Furthermore, in the statistical tests performed on the residuals (RAW, RAI1, RAI2 and RA) the maximal lag M for which auto- or cross-correlation is computed is again set to $n_b + 20$, with n_b being the order of the polynomial $B(q)$ in (1.13).

A.1.1 Influence of the model order in the estimation

As was previously done during the analysis based on OE models, an investigation about the influence of the model order on the achievable fits is first conducted. Figure A.1 shows the average of the impulse response fit introduced in Section 3.2.1 as function of the FIR model orders.

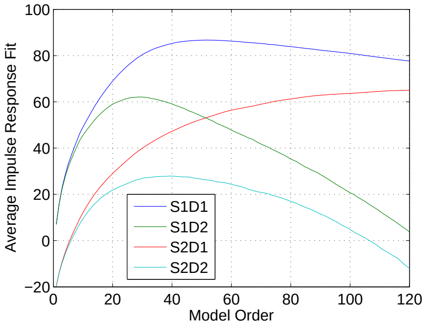


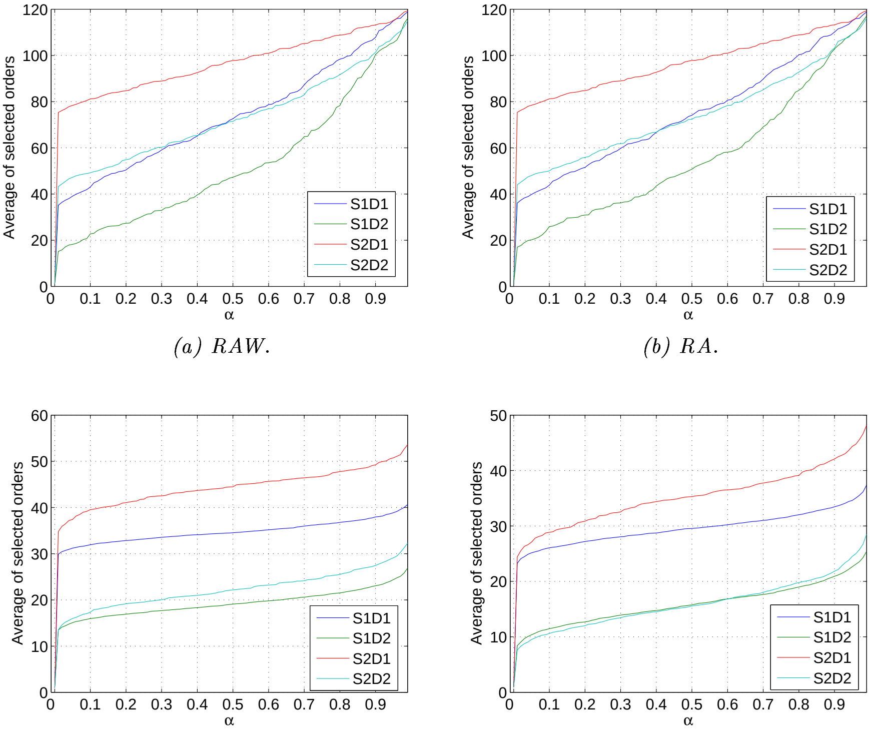
Figure A.1: FIR models - Average impulse response fit achieved in the estimation of 200 systems in each data set for model orders ranging from 1 to 120.

With respect to the plots in Figure 4.1, referring to OE models, the graphs in Figure A.1 are smooth, because the minimization problems that are solved for the estimation of FIR models are convex and therefore no local minima can be found. OE models are instead estimated solving pseudo-linear regression problems, that are non-convex and can lead to local minima.

For what regards the dependence of the fit on the model complexities, analogous considerations to the ones done w.r.t. OE models can be done: namely, the order selection is much more critical when noisy measurements are exploited, for which low complexities are suitable to avoid overfitting. In relation to the less noisy data sets, the trend for data set S2D1 in Figure A.1.(a) actually suggests that larger orders than the ones here evaluated could lead to even better impulse response fits.

A.1.2 Selection of the significance level α used in the statistical tests

Again, the analysis for the choice of the significance level α to be used in the statistical tests is carried out by observing the mean values of the selected orders and of the impulse response fits achieved for α ranging from 0 to 0.99 with steps



(a) RAW.

(b) RA.

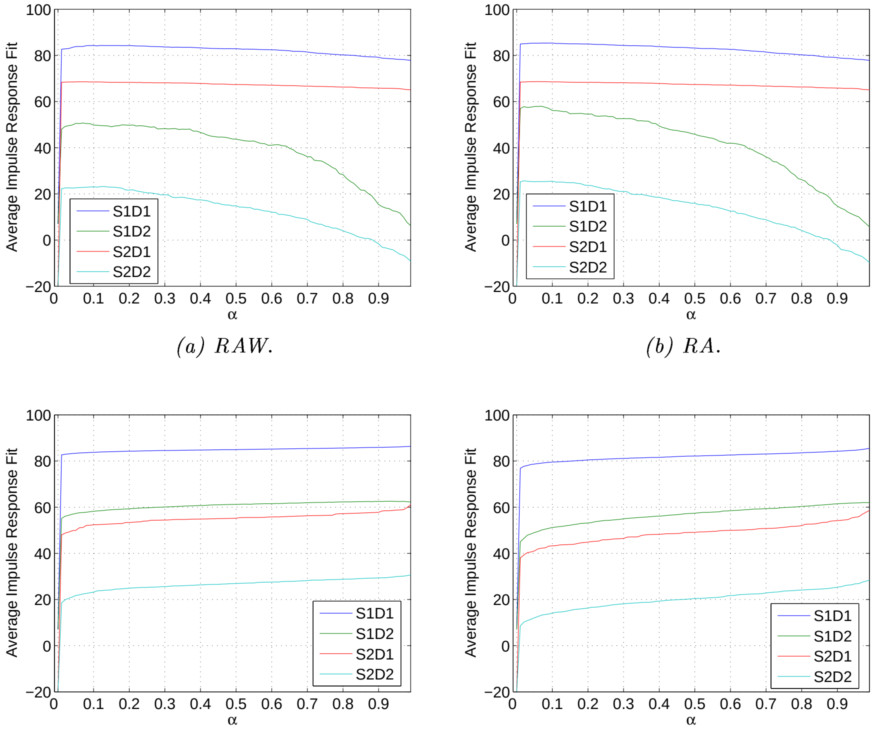
(c) RAI1.

(d) RAI2.

Figure A.2: FIR models - Average of the orders selected by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

of 0.01.

Figure A.2, which refers to the tests on the residuals, presents trends similar to the ones observed in Figure 4.8 for OE models. For what regards the whiteness test (RAW) or its combination with the independence test (RA), the significant difference present between the average of the orders chosen for the data set S2D1 and the one observed for the other data sets is due to the issue previously described with OE models that still affects the whiteness test: in many cases it is not able to unfalsify at least one of the evaluated model structures, causing the default selection of the highest order. However, with FIR models, this behaviour is partially justified by the trend in Figure A.1.(a), which suggests the adoption of very high-order FIR models for the reproduction of the true systems in data set S2D1. Again, data set S2D2 is less affected by these issues, because of the



(a) RAW.

(b) RA.

(c) RAI1.

(d) RAI2.

Figure A.3: FIR models - Average impulse response fits achieved by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

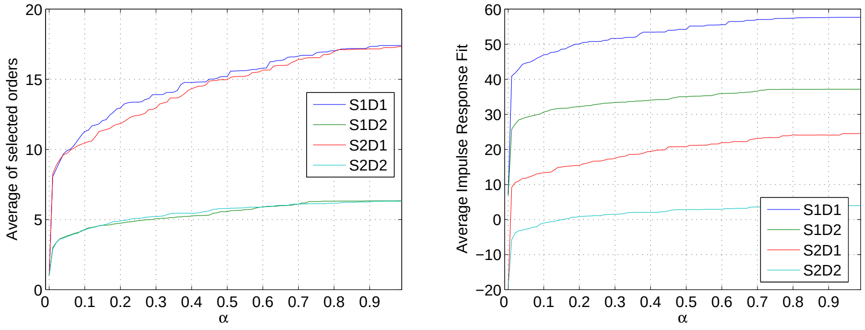
larger noise component present in the measurement data.

For what regards the orders selection done by the independence test on the residuals, a clear distinction between noisy and non-noisy data sets can be observed and explained by the same arguments presented for OE models (refer to (4.1)). Furthermore, as expected, when slow systems are estimated, model structures of higher orders are chosen w.r.t. the ones selected for fast systems. A similar behaviour in the orders selection is seen for the whiteness test and for RA even if the averages of the orders chosen for S1D1 and S2D2 are very similar, because S2D2 is more affected by the problem that none of the evaluated model structures can be unfalsified by the whiteness test, even for small values of α.

Passing to the analysis of the average impulse response fits, which will provide

the most useful indication for the choice of α, clear trends can be identified. While for the whiteness test a small value of α seems to give the best average fits in all the data sets, for the independence test the value of α does not significantly impact the average fit, but large values are preferable. When the whiteness test or its combination with the independence test is applied, low values of α allow to reduce the rate of models that are not unfalsified, thus limiting the default selection of models with order 120. When α is very high, order 120 is assigned by default to most of the estimated models: if this choice is beneficial for the non-noisy data sets, the same does not hold for the noisy ones, since it gives rise to overfitting.
Further tests, analogous to the ones illustrated in Section 4.1.2 for OE models, have shown that the use of a larger value for the maximal lag M makes possible the unfalsification of at least one of the evaluated model structures. With this choice of M fits comparable to the best ones observed in Figure A.3 can be achieved, but adopting a large significance level.

In relation to the independence tests (RAI1 and RAI2), the fact that the best average fits can be obtained when $\alpha = 0.99$ suggests a tendency to undermodel of this selection method; in this case, values between 0.99 and 1 could probably give even better fits. This tendency also explains the fact that the achieved fits are larger for the "fast" data sets than for the "slow" ones.



(a) Average of the selected orders.

(b) Average impulse response fits.

Figure A.4: FIR models - Average of the selected orders and average impulse response fits achieved by the F-test, as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

Figure A.4 contains the average of the orders selected and the average fits achieved by means of the F-test for different values of the significance level α. The orders selection is very well distinguished in two trends, one for the non-noisy and long data sets and one for the noisy and short data sets. For the latter ones, lower orders are chosen w.r.t. the ones selected for S1D1 and

S2D1, proving that no overfitting issue is encountered. If, on the one hand, the F-test seems able to correctly discern between noisy and non-noisy data, on the other hand, no distinction between fast and slow systems is detected in the orders choices. Since the F-test directly compares the loss functions of two model structures, an explanation for the described behaviour can be found by investigating the average trend of the loss function in each data set: after the fast decrease that it encounters for very small complexities in all the data sets, in the noisy ones the loss settles down in correspondence to smaller orders than what observed for non-noisy sets. Therefore, the procedure for order selection is stopped earlier in noisy data sets, since no significant difference in the loss of the compared model structures is detected.
The F-test also shows a general tendency to undermodel, as it is further proved by the fact the the highest average impulse response fits are achieved for very large values of α.

A.1.3 Analysis of the order selection methods

The analysis is carried out by means of Table A.2, which summarizes the average impulse response fits achieved by the different criteria in the identification of 200 systems in each of the four data sets considered. Further information about the achieved fits are provided by the box-plots of the impulse response fits in Figure A.5. The orders selected by the evaluated methods are investigated by the histograms in Figures A.6, A.8, A.10 and A.12; in the end, Figures A.7, A.9, A.11 and A.13 report the histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the other order selection criteria.

Table A.1 contains the values of the significance level α adopted in the statistical tests presented in this section: the chosen values correspond to the ones leading to the highest average impulse response fit in each data set. The significance level α adopted for the two independence tests on the residuals (RAI1 or RAI2) and for the F-test ($\alpha = 0.99$) is very high when compared with the values commonly used in statistical tests. However, in this case a high value of α guarantees the selection of model structures of higher complexity, thus limiting the undermodelling issue that characterizes these two order selection methods; this issue is particularly problematic for FIR models, which need to be enough complex in order to properly describe the systems in the data sets. The adoption of a larger value of the maximal lag M for which the cross-correlation is computed alleviates this undermodelling tendency in the slow data sets: in this way correlation components corresponding to large lags are detected. However, specific tests have proved that large significance levels are still more appropriate.

On the other hand, the whiteness test (RAW) and its combination with the independence test (RA) lead to better performances when low significance levels are adopted, thus limiting the problem previously cited, i.e. favouring the unfalsification of at least one of the evaluated model structures.

	α				
Dataset	*RAW*	*RA*	*RAI1*	*RAI2*	*F*
S1D1	0.12	0.10	0.99	0.99	0.99
S1D2	0.07	0.07	0.92	0.99	0.99
S2D1	0.06	0.06	0.99	0.99	0.99
S2D2	0.13	0.02	0.99	0.99	0.99

Table A.1: FIR models - Values of the significance level α which guarantee the best average impulse response fits when adopted in the statistical tests used for model order selection.

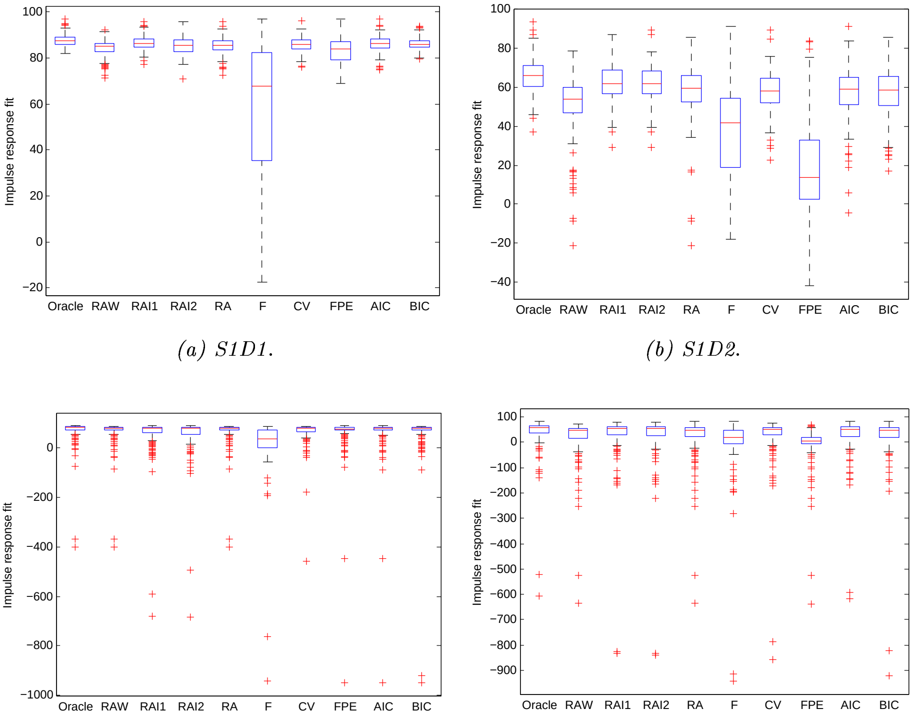
Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	87.7	84.3	86.3	85.5	85.3	57.7	85.9	83.1	86.1	86
S1D2	66	50.7	62.5	62	57.9	37.4	57.8	19.4	57.4	56.8
S2D1	71.2	68.6	61	58.6	68.6	24.5	65.4	64	65.6	62.3
S2D2	39.7	23.2	30.6	28.4	25.7	3.9	27.8	-2.12	30	23.6

Table A.2: FIR models - Average impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

The inspection of the performances reported for the oracle in Table A.2 allows to detect the difficulties encountered by FIR models in the description of slow systems, especially the ones contained in data set S2D1. Figure A.10 shows that the highest order in the interval of the evaluated ones, i.e. 120, is most often selected by the oracle; this suggests that probably higher orders would have been more appropriate for the description of the true systems in that data set, as also Figure A.1.(a) indicates.

Figure A.5 highlights a considerable difference between the fits achieved on the "fast" data sets and the ones reached on the "slow" ones. Namely, in the case of S2D1 and S2D2, the box plots of the impulse response fits contain many outliers for all the order selection criteria. These outliers correspond to models for which the oracle chooses very high orders and for which FIR models of orders larger than 120 could be even more suitable for the reproduction of the true systems. Therefore, the choice of lower orders, which is done by the criteria here analyzed, leads to very small values of the impulse response fit.

Passing to the analysis of the order selection methods, the criteria which lead to the best impulse response fits seem to be cross-validation and AIC for S1D1 and S2D1, while the two independence tests on the residuals are more effective on the noisy data sets S1D2 and S2D2. Indeed, while the tendency of cross-



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.5: FIR models - Box plots of the impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

validation and AIC to choose quite high orders is beneficial for the data sets characterized by output measurement data with high SNR (especially for the slow systems in S2D1), the low-complexity models selected by the independence tests on the residuals avoid the adaptation to the specific noise realization, thus leading to a better description of the system properties.

On the other hand, the worst criteria are FPE for the noisy sets S1D2 and S2D2 and the F-test for S1D1 and S2D1. The reasons of the unsatisfying fits achieved are opposite for the two criteria: on the one hand, FPE has a marked tendency to select complex model structures, which give rise to overfitting in the noisy datasets, but it is beneficial especially for the slow systems of S2D1; on the other hand, the F-test usually selects very simple model structures, which avoid overfitting when noisy data are used but at the same time are less suitable for a proper description of the true systems. It should be noticed how the overfitting

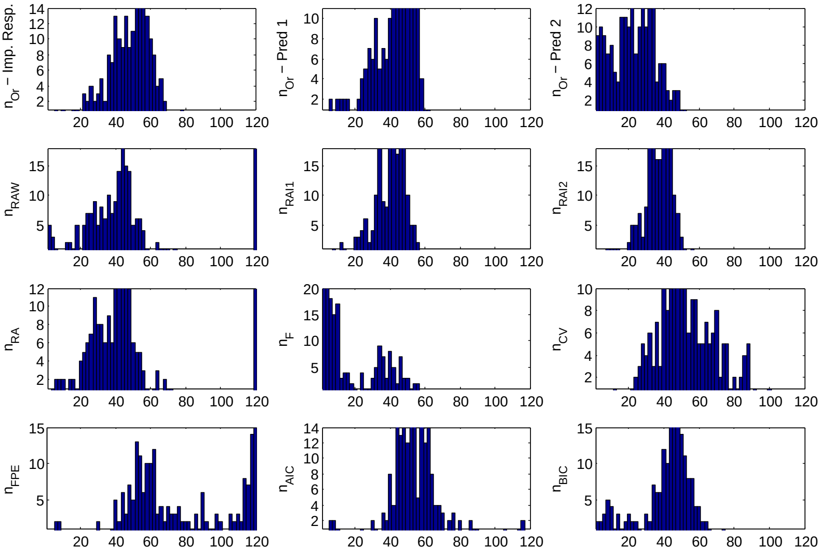


Figure A.6: FIR models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S1D1.

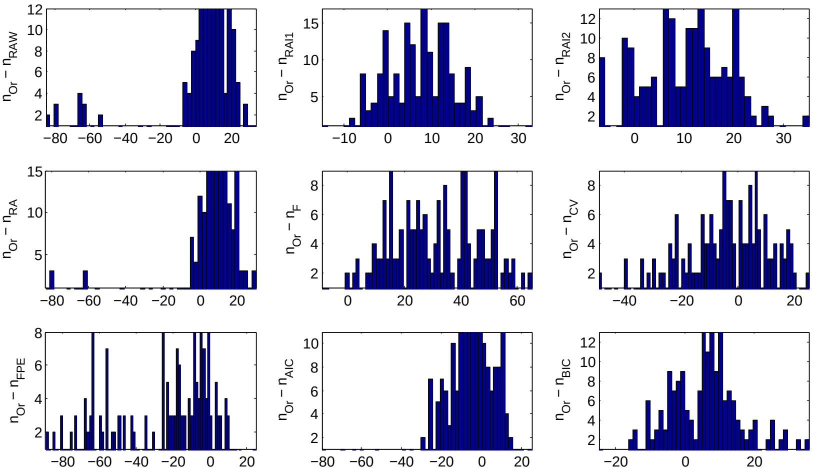


Figure A.7: FIR models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S1D1.

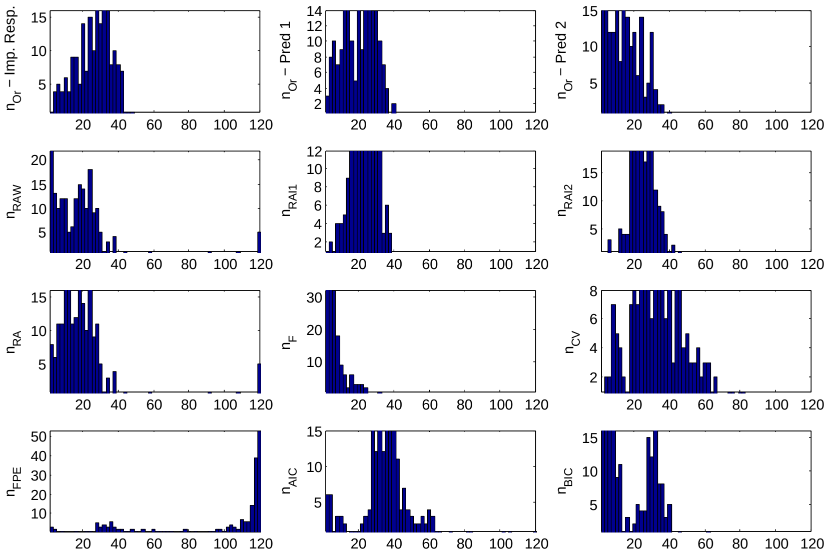


Figure A.8: FIR models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S1D2.

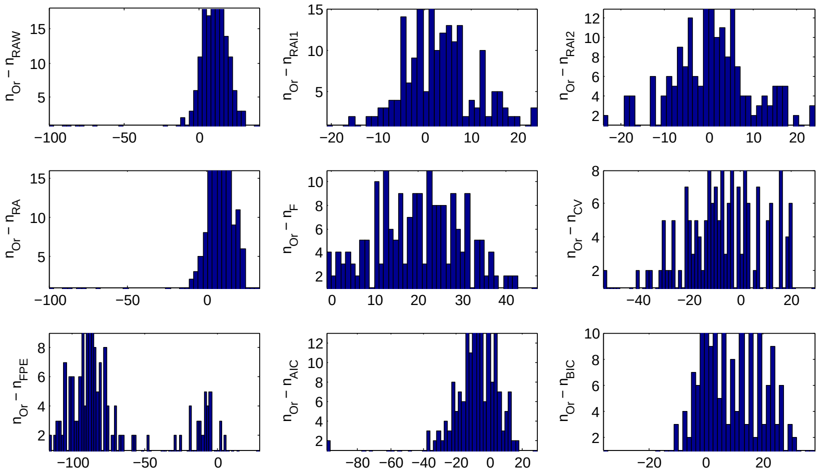


Figure A.9: FIR models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S1D2.

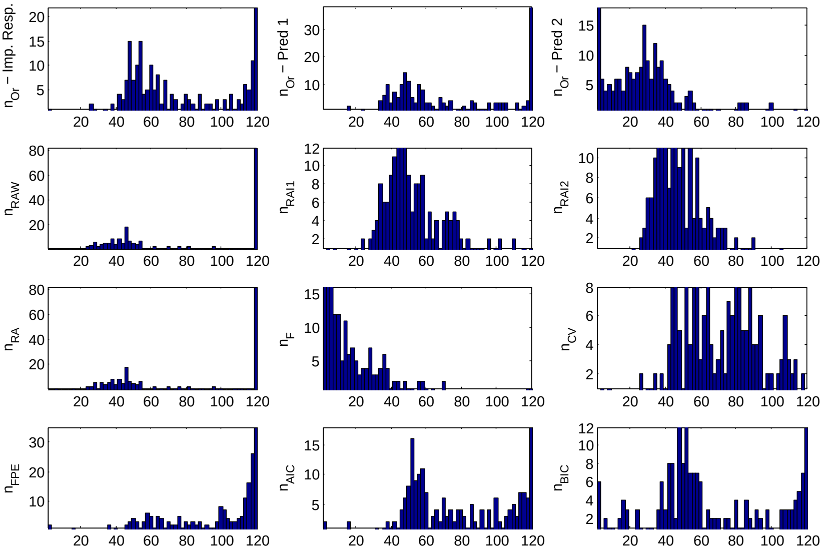


Figure A.10: FIR models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S2D1.

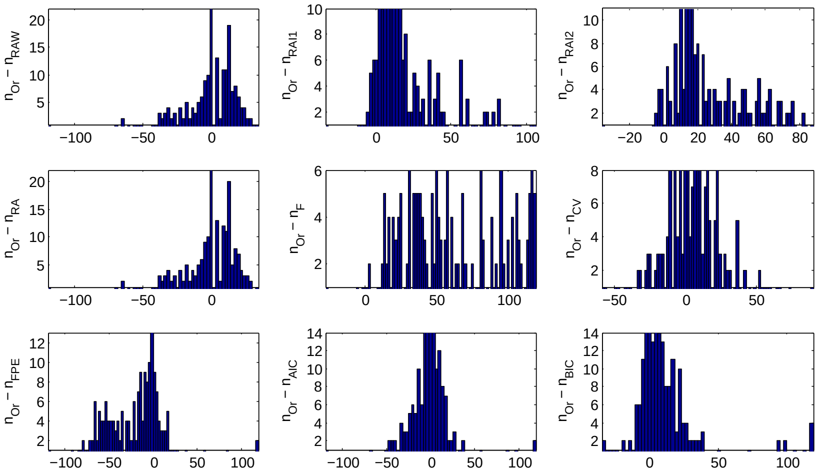


Figure A.11: FIR models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S2D1.

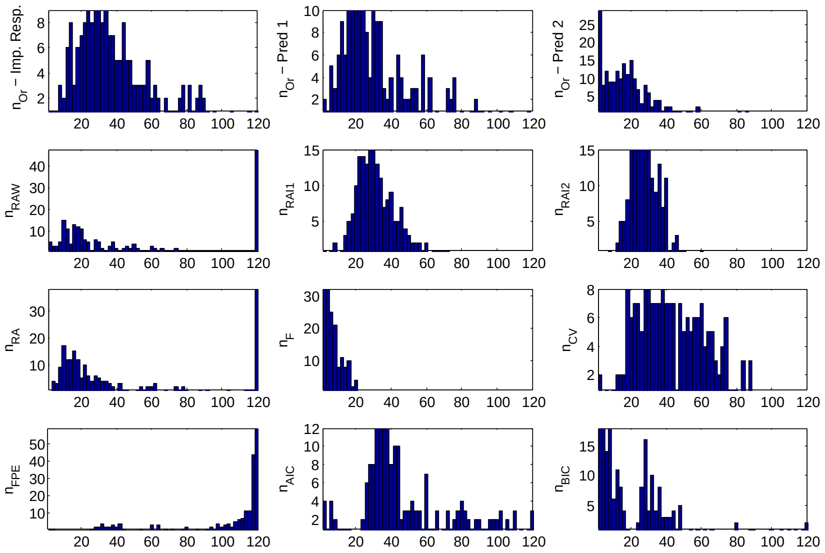


Figure A.12: FIR models - Histograms of the orders selected by the analyzed criteria when 200 systems are identified in data set S2D2.

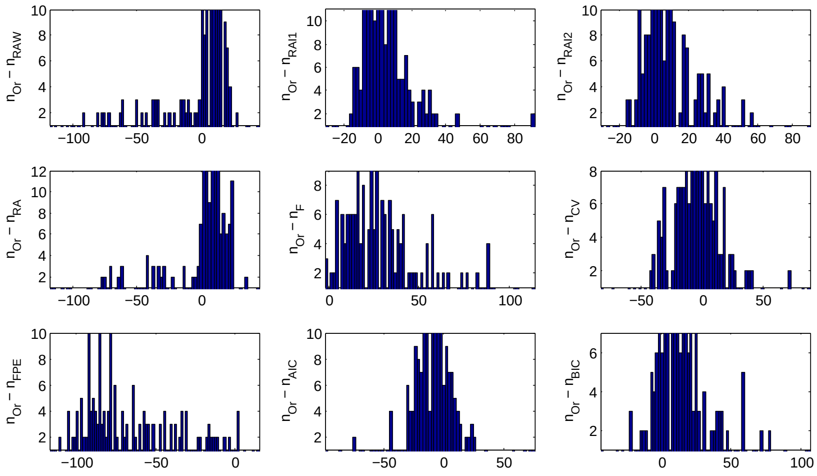


Figure A.13: FIR models - Histograms of the differences between the orders selected by the oracle for impulse response fit and the ones chosen by the analyzed criteria when 200 systems are identified in data set S2D2.

trend detected for FPE is even more relevant in S1D2 and S2D2. Indeed, for the noisy data sets the loss function $V_N(\widehat{\theta}_N, Z^N)$ shows a more marked decrease for models of increasing complexity because also the specific noise realization is modelled; since the penalty inflicted by the FPE criterion on model complexity is not sufficiently large to counterbalance this overfitting effect, model structures of high order still give rise to the lowest values of the FPE criterion.
The overmodelling tendency that AIC criterion has shown when applied on OE models is less marked with FIR models. In the latter case, the number $d_{\mathcal{M}}$ of parameters involved in the largest FIR model here tested is usually larger than the ones appearing in the most complex OE model tested in Section 4.1. Therefore, AIC seems able to appropriately penalize only very large complexities, as the ones considered with FIR models.

RAW and RA give almost comparable performances in terms of impulse response fits, with RA leading to better results thanks to the inclusion of the independence test of the residuals from past inputs, which allows to further validate the choice done by the whiteness test. In data set S2D1 they give rise to the best performances, because they are favoured by the default selection of order 120 which is done whenever they are not able to select a model structure among the tested ones. This behaviour is justified for data set S2D1, since the oracle chooses in many cases the highest order, 120, but the same argument does not hold for the systems in S2D2, which can be properly described by simpler model structures, as proved by the oracle choices (Figure A.12). Moreover, for data set S2D2 the default choice of order 120 is not so beneficial as the average fits in Table A.2 and in Figure A.1 prove.

Another observation that should be done regards the results achieved by the two independence tests on the residuals (RAI1 and RAI2): RAI1 leads to slightly better fits than the ones obtained by RAI2, even if the application of RAI1 without its combination with the whiteness test appears wrong from a theoretical point of view. A first explanation of the practical results lies in the fact that the model orders chosen by RAI1 are usually a bit higher than the ones selected by RAI2 (except in data set S1D2), thus being more in line with the selections done by the oracle.

A.1.4 Combination of comparison and validation methods

The combination presented in Section 2.3.1 is now evaluated on FIR models. The average values reported still refer to the identification of 200 systems in each of the four data sets. Furthermore, for the tests on the residuals the correlation is computed until the lag $M = n_b + 20$.

F-test

When the F-test is combined with the validation methods, the best average impulse response fits in all the four data sets are achieved by the combination with the two independence tests, RAI1 and RAI2. The only exception is detected in data set S2D1, where RAW and RA guarantee the best fits thanks to the default selection of order 120, done in most cases.

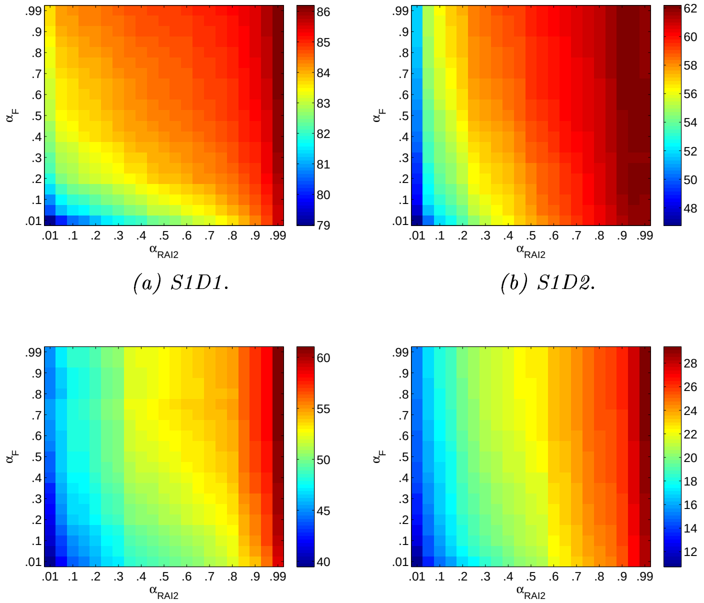
Figure A.14 illustrates the average fits achieved for different values of α_F and α_{RAI2}, the significance levels used respectively in the F-test and the independence test on the residuals which is applied in a second stage. The trend is similar for all the four data sets: namely, α_F is not so influential when a large α_{RAI2} is adopted for the independence test later applied. It also should be noticed that, independently from the set significance level for both the two statistical tests, the average fits achieved are always larger than the ones achieved by the F-test alone, reported in Table A.2. Furthermore, the trend observed w.r.t. the value of α_{RAI2} is analogous to the one previously observed in Figure A.3, when RAI2 was applied alone as an order selection method.

Table A.3 summarizes the average fits achieved when the F-test is used alone for model order selection and when it is equipped with a validation stage based on residuals analysis. The average of the selected orders is also reported. These results are obtained choosing for the two statistical tests the significance levels which lead to the best average impulse response fit for each data set.

	Average impulse response fit					Average of the selected orders				
Set	*F*	*F +* *RAW*	*F +* *RA*	*F +* *RAI1*	*F +* *RAI2*	*F*	*F +* *RAW*	*F +* *RA*	*F +* *RAI1*	*F +* *RAI2*
S1D1	57.7	85.3	85.8	86.7	86.2	17.4	38.5	39	43.1	40.3
S1D2	37.2	56.6	59.3	62.8	62.2	6.3	20.9	18.8	24.3	23.9
S2D1	24.5	68.9	68.9	62.46	60.9	17.3	80.4	80.4	57.9	52.7
S2D2	3.9	26.4	27.1	31.6	29.4	6.3	45.5	45.9	34.7	30.6

Table A.3: FIR models - Average impulse response fits and average of the selected orders when validation methods is combined with the F-test for model order selection in the identification of 200 systems in each data set.

The combination of the two model order selection methods is very beneficial in terms of the average impulse response fits achieved in all the four data sets; these fits are in line with the ones obtained by the tests on residuals alone and reported in Table A.2. However, thanks to the slight improvement w.r.t. the fits in Table A.2, this combination gives rise to the best performances detected with the until now analyzed criteria, when applied on FIR models.
The improvement is reached thanks to the selection of more complex model structures with respect to the ones chosen by the F-test alone. Between the two



(a) S1D1. *(b) S1D2.*

(c) S2D1. *(d) S2D2.*

Figure A.14: FIR models - Average impulse response fits achieved for different values of the significance levels when RAI2 is combined with the F-test for model order selection in the identification of 200 systems in each data set.

implementations of the independence test, RAI1 generally selects higher model orders than RAI2, thus being more effective in terms of impulse response fit. However, the selection of too high orders is not everywhere beneficial, as the results in data set S2D2 prove: the combination of the F-test with the whiteness test on the residuals leads to the selection of the highest orders in that data set but the achieved fits are lower because of the overfitting which arises.

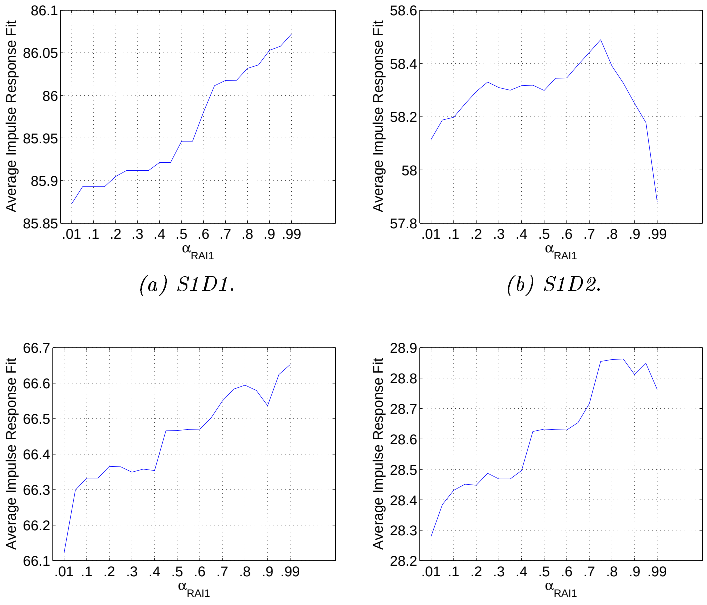
Cross-validation

The impulse response fits achieved when cross-validation is used for model order selection can be improved when it is combined with a validation stage, as Table A.4 proves. However, the improvements are less significant than the ones detected in the same setting with the F-test. In this case, all the tests on the residuals are effective when applied in a second stage for model validation. The main differences in terms of performances regard the whiteness test and its combination with the independence test (RA), which lead to the best fits in S2D1

but to the worst ones in S2D2; this behaviour is analogous to the one previously described for the combination with the F-test and can be motivated by the same arguments.
Table A.4 shows that the improvement in the fit obtained after the validation stage is again accompanied by the increase of the average of the selected orders.

Set	Average impulse response fit					Average of the selected orders				
	CV	$CV +$ RAW	CV $+$ RA	$CV +$ $RAI1$	$CV +$ $RAI2$	CV	$CV +$ RAW	CV $+$ RA	$CV +$ $RAI1$	$CV +$ $RAI2$
S1D1	85.9	86.1	86.1	86.1	86	53.8	53.8	53.8	55.1	54.6
S1D2	57.8	58	58	58.5	58.3	34.5	35.9	35.9	36.2	35.8
S2D1	65.4	69.1	69.1	66.7	66.2	70.7	85.6	85.6	73.2	72
S2D2	27.8	26.2	26.3	28.9	28.9	43.1	58.8	58.8	44.7	44.4

Table A.4: FIR models - Average impulse response fits and average of the selected orders when validation methods are combined with cross-validation for model order selection in the identification of 200 systems in each data set.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.15: FIR models - Average impulse response fits achieved for different values of the significance level α_{RAI1} used in the independence test RAI1 when it is combined with cross-validation for model order selection in the identification of 200 systems in each data set.

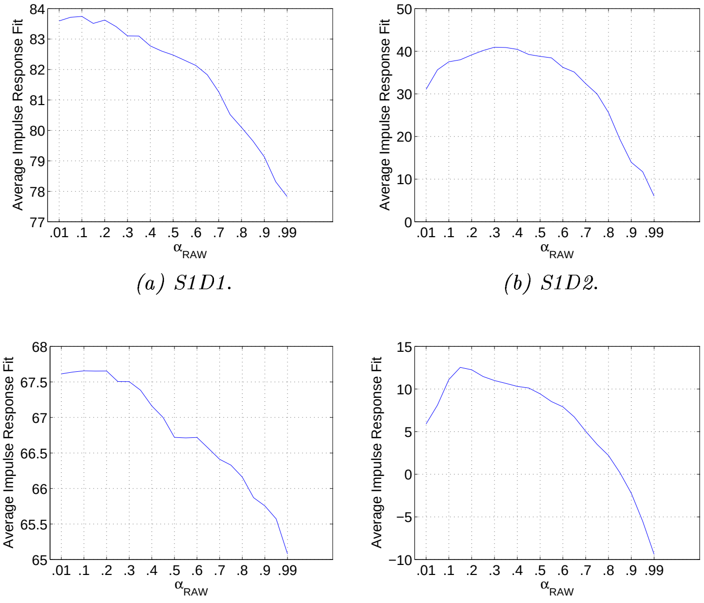
Since the different tests on the residuals, when combined with cross-validation,

lead to almost equivalent impulse response fits (as Table A.4 shows), the analysis of the impact of the significance level on these tests is carried out considering the first implementation of the independence test on the residuals (RAI1). Figure A.15 illustrates the average fits achieved for different values of α when RAI1 is applied after cross-validation. While small values of α_{RAI1} favour the validation of the model structures selected by cross-validation, large values of α_{RAI1} increase the rate of new model structures that have to be chosen, since the ones originally returned by cross-validation are validated with less probability.
Large values of α_{RAI1} seem to be more appropriate, except in data set S1D2, where a too large value of the significance level leads to the selection of too high orders for the noisy measurements and fast systems of that data set. However, independently from the value of α_{RAI1}, the average fits achieved by the combination of cross-validation with the validation stage are larger than the ones obtained by cross-validation alone in all the data sets. Similar trends can be observed when the validation stage is performed by means of RAI2.

It should be pointed out that the results reported in Table A.4 are obtained setting the significance level to the value giving the highest average impulse response fit in each data set.

FPE

The FPE criterion only benefits from its combination with the whiteness test (RAW) or with both the tests on the residuals (RA), while the independence test alone does not give rise to significant improvements in the achieved impulse response fits. Figure A.16 shows that low values of the significance level α_{RAW} adopted in the whiteness test lead to higher values of the average impulse response fit. Indeed, when α_{RAW} is very high, the issue characterizing the whiteness test becomes more frequent and order 120 is in most cases chosen by default. This causes the decrease in the average fit, which is illustrated by the plots. This drop is of course more evident for the noisy data sets, whose plots also present a different trend w.r.t. the one observed for S1D1 and S2D1. Indeed, the FPE criterion tends to choose very high orders for the noisy data sets and its combination with the whiteness test with low values of α_{RAW} allows to invert this tendency, favouring the selection of lower orders.
Differently from what was previously observed with the F-test and cross-validation, in case of the FPE criterion a wrong choice of the significance level could lead to a decrease of the impulse response fit originally achieved by FPE alone.



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.16: FIR models - Average impulse response fits achieved for different values of the significance level α_{RAW} used in the whiteness test (RAW) when it is combined with FPE for model order selection in the identification of 200 systems in each data set.

AIC

When AIC is combined with a validation stage no improvements are obtained in data sets S1D1, S1D2 and S2D2, while an average impulse response fit equal to 68.2 is achieved in S2D1 thanks to the combination of AIC with RAI2, with $\alpha_{RAI2} = 0.99$. It should be underlined that in this case a wrong choice of the significance level could lead to a slight worsening of the impulse response fits previously achieved by AIC; however, the optimal value of α_{RAI2} here cited agrees with what observed in Figure A.3.

For what regards the whiteness test RAW or the application of both the tests on the residuals (RA), their combination with the AIC criterion is beneficial only in data set S2D1, where an average impulse response fit equal to 69.1 is reached with $\alpha_{RAW} = 0.01$. Again, this good performance is favoured by the default selection of order 120, done in many cases.

BIC

When also BIC is equipped with validation methods, the impulse response fits encounter a general improvement. Among the validation tests, the two independence tests RAI1 and RAI2 are again the most effective, as shown in Table A.5, whose results are the best ones achievable in each data set by appropriately setting the significance level. Again, the improvement is reached thanks to the selection of more complex model structures.

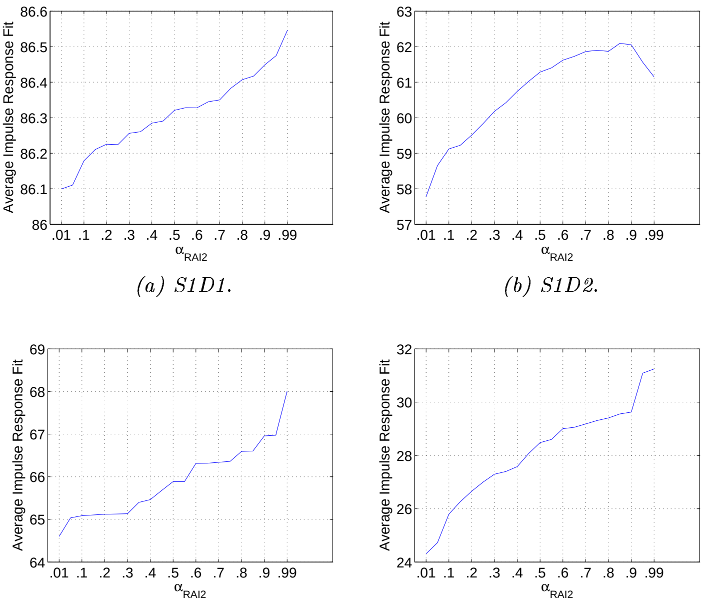
Set	Average impulse response fit					Average of the selected orders				
	BIC	*BIC* + *RAW*	*BIC* + *RA*	*BIC* + *RAI1*	*BIC* + *RAI2*	*BIC*	*BIC* + *RAW*	*BIC* + *RA*	*BIC* + *RAI1*	*BIC* + *RAI2*
S1D1	86	86.3	86.3	86.7	86.5	41.9	44.3	44.4	46.3	45
S1D2	56.8	59.2	60	62.3	62.1	17.6	22.8	23.5	26	26.1
S2D1	62.3	69.3	69.3	66.7	68	64.3	80.9	80.9	69.7	70.1
S2D2	23.6	26.2	26.6	31.8	31.2	23.1	48.7	49	39.9	40.5

Table A.5: FIR models - Average impulse response fits and average of the selected orders when validation methods are combined with BIC for model order selection in the identification of 200 systems in each data set.

Table A.5 shows that the average fits achieved by the combination of BIC and a validation method are analogous or even better (as in data sets S2D1 and S2D2) than the best ones observed in Table A.2.
According to the values in Table A.5, the performances of the two implementations of the independence test on the residuals are almost equivalent, except in data set S2D1, where RAI2 gives better results. However, the combination of BIC criterion with RAI2 is more indicated since the influence of the significance level α_{RAI2} on the effectiveness of the technique is clearer, as Figure A.17 illustrates. Indeed, it can be immediately noticed how large values of α_{RAI2} are more beneficial, even if the average impulse response fits achieved are always better than the ones reached by BIC alone, independently from the value of α_{RAI2}.

Conclusions

The analysis about the combination described in Section 2.3.1 has shown that it gives rise to improvements of the impulse response fit with respect to the single application of a certain comparison technique. In this context, the test for the independence of the residuals from past inputs has proved to be the most effective validation technique, except for the FPE criterion, whose performances can



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.17: FIR models - Average impulse response fits achieved for different values of the significance level α_{RAI2} used in the independence test RAI2 when it is combined with BIC for model order selection in the identification of 200 systems in each data set.

be improved by equipping it with the whiteness test on the residuals. A singularity has been detected in data set S2D1, where the whiteness test has always led to the best performances, because it is favoured by the default selection of order 120.

Furthermore, the way in which the significance level α used in the statistical tests on the residuals influences the effectiveness of the technique appears clear and is analogous to the one observed when the tests are applied alone for model order selection.

A.1.5 Combination with the F-test

The combination presented in Section 2.3.2 is now analyzed on FIR models.

Whiteness test on the residuals (RAW)

Figures A.7, A.9, A.11 and A.13, showing the differences between the orders selected by the oracle and the whiteness test on the residuals, underline how this method tends to underfit, whenever a certain model structure can be unfalsified. This tendency could be alleviated by increasing the significance level α used in the statistical test, but this would also worsen the issue present in the slow data sets and lead to a decrease in the average fit, as Figure A.3.(a) has shown. Therefore, in this context, the application of the F-test can be exploited to increase in a second stage the complexities of the model structures returned by the whiteness test.

Figure A.18 illustrates the average values of the impulse response fits achieved for different values of the significance levels α_{RAW} and α_F respectively adopted in the whiteness and in the F-test. A similar trend can be observed for all the data sets: namely, small values of α_{RAW} are suggested to increase the probability of unfalsification of at least one of the tested model structures, while values of α_F larger than 0.5 favour the final selection of larger model structures.

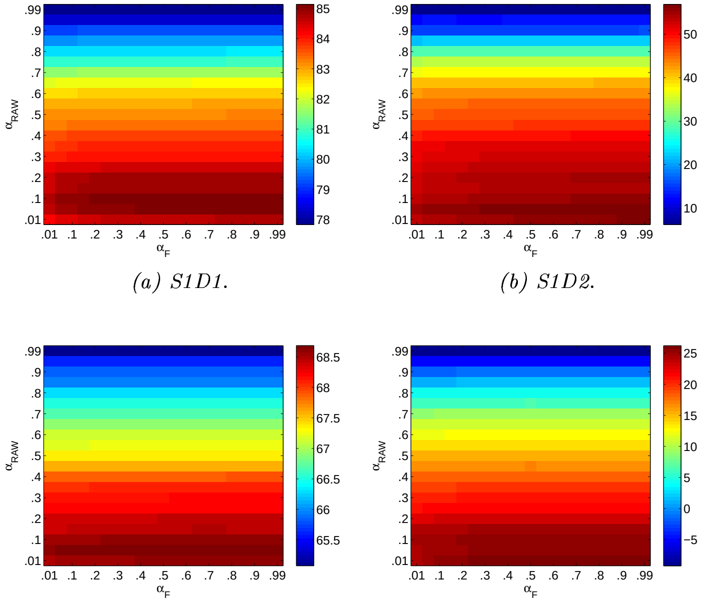
	Average impulse response fit		Average of the selected orders	
Set	*RAW*	*RAW + F (Up)*	*RAW*	*RAW + F (Up)*
S1D1	84.3	85.1	43	43.8
S1D2	50.7	56.8	19.3	19.9
S2D1	68.6	68.7	78.8	79.8
S2D2	23.2	26.2	49.1	50.6

Table A.6: FIR models - Average impulse response fits and average of the selected orders when the F-test is applied after the whiteness test (RAW) to perform model order selection in the identification of 200 systems in each data set.

Table A.6 summarizes the average impulse response fits and the average of the selected orders when the whiteness test is applied alone and when it is combined with the F-test. The results are achieved with the significance levels of both tests set to the value which leads to the best mean of the impulse response fit. The growth detected in the fits confirms the efficacy of this combination, even if the newly obtained fits are still lower than the best ones reported in Table A.2.

Whiteness and independence test on the residuals (RA)

When the F-test is applied on the model structures selected by both the statistical tests on the residuals, analogous results to the ones described for the whiteness test are observed. Indeed, as previously observed, the two methods behave in a very similar way and the dependence between the average impulse



(a) S1D1. *(b) S1D2.*

(c) S2D1. *(d) S2D2.*

Figure A.18: FIR models - Average impulse response fits achieved for different values of the significance levels when the F-test is applied after the whiteness test on the residuals (RAW) in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

response fits and the significance levels adopted in the two statistical tests is analogous to the one shown in Figure A.18. Also the values in Table A.7 are in line with the ones previously analyzed in Table A.6.

	Average impulse response fit		Average of the selected orders	
Set	*RA*	*RA + F (Up)*	*RA*	*RA + F (Up)*
S1D1	85.3	85.6	43.8	44.3
S1D2	57.9	59.3	20.2	21.3
S2D1	68.6	68.7	78.9	78.9
S2D2	25.7	27.1	45.1	45.1

Table A.7: FIR models - Average impulse response fits and average of the selected orders when the F-test is applied after both tests on the residuals (RA) to perform model order selection in the identification of 200 systems in each data set.

Independence tests on the residuals (RAI1 and RAI2)

Previously, it was observed that the two implementations of the independence test tend to undermodel even if a large significance level is adopted. However, this property has proved to be beneficial for the noisy data sets, where these tests gave rise to the best impulse response fits when applied for model order selection.
These observations suggest that the application of the F-test on model structures with higher complexities should lead to an improvement of the previously obtained fits. Table A.8 confirms this expectation, even if the observed improvements are not significant. When the best values of the significance levels are adopted, the increase in the complexity of the newly chosen model structures is less relevant for the noisy data sets.

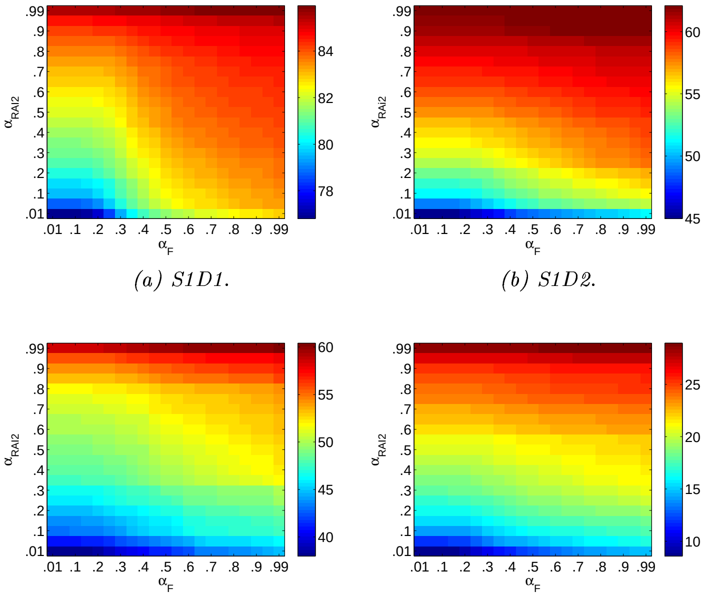
Set	Average impulse response fit				Average of the selected orders			
	RAI1	*RAI2*	*RAI1 + F (Up)*	*RAI2 + F (Up)*	*RAI1*	*RAI2*	*RAI1 + F (Up)*	*RAI2 + F (Up)*
S1D1	86.3	85.5	86.4	85.7	40.6	37.4	40.9	37.9
S1D2	62.5	62	62.8	62.2	23.4	25.4	24.1	23.6
S2D1	61	58.6	61.1	58.7	53.6	48.2	53.8	48.4
S2D2	30.6	28.4	30.7	28.7	32.3	28.5	32.8	29.4

Table A.8: FIR models - Average impulse response fits and average of the selected orders when the F-test is applied after the independence test on residuals (RAI1 or RAI2) to perform model order selection in the identification of 200 systems in each data set.

From Figure A.19 the tuning of the significance levels appears quite clear: namely, the one for the independence test, α_{RAI2}, should be high (as was previously observed also from Figure A.3.(d)), as well as the one for the F-test, α_F, thus favouring the choice of more complex model structures.

Cross-validation

Cross-validation does not generally benefit from its combination with the F-test. In particular, it is not clear whether an improvement of the performances achieved by cross-validation alone can be obtained by testing simpler or more complex model structures. In this sense, Figures A.7, A.9, A.11 and A.13 are explanatory, since they highlight how cross-validation does not present a clear tendency to under- or overmodel. In particular, in data set S2D1, the evaluation of more complex model structures is clearly helpful.



(a) S1D1.

(b) S1D2.

(c) S2D1.

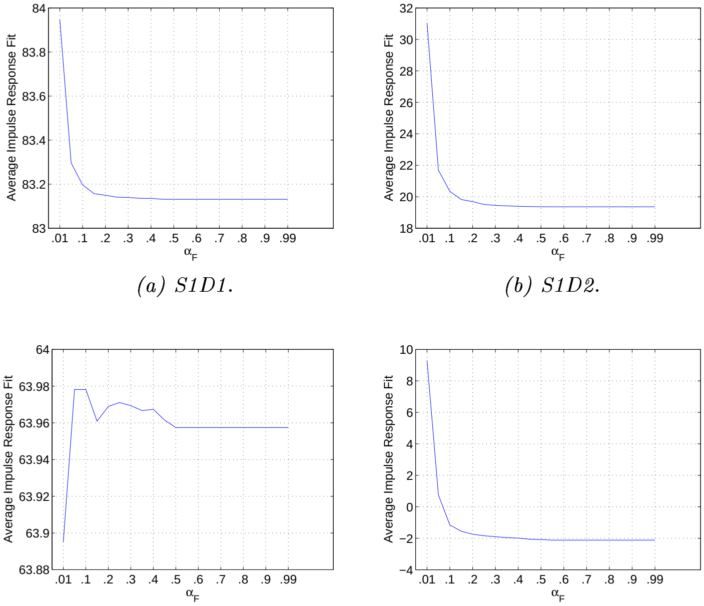
(d) S2D2.

Figure A.19: FIR models - Average impulse response fits achieved for different values of the significance levels when the F-test is applied after the independence test on the residuals (RAI2) in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

FPE

The previous analysis of the model order selection methods has clearly shown how the FPE criterion tends to overfit. Therefore, its combination with the F-test could be exploited to refine its order selection by testing model structures of decreasing complexity. As could be expected, Figure A.20 suggests that very small values of the significance level for the F-test, α_F, are preferable in order to favour the selection of simpler model structures. The only exception detected in the figure regards the data set S2D1, where the overfitting tendency of the FPE criterion has proved to be beneficial; therefore, the choice of smaller orders for the estimated models leads to a very slight decrease of the impulse response fits achieved, as Figure A.20.(c) illustrates.

The values in Table A.9 prove the improvement that can be reached thanks to the combination here described. However, the newly achieved fits are still sig-



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.20: FIR models - Average impulse response fits achieved for different values of the significance level α_F used in the F-test when it is applied after FPE in order to evaluate model structures with decreasing complexity. The average is calculated from the identification of 200 systems in each data set.

nificantly lower than the best ones obtained with other order selection methods.

	Average impulse response fit		Average of the selected orders	
Set	*FPE*	*FPE + F (Down)*	*FPE*	*FPE + F (Down)*
S1D1	83.1	83.9	76.1	69.4
S1D2	19.4	31.1	94.7	78.4
S2D1	64	64	95.8	94.3
S2D2	-2.1	9.3	102.4	86.1

Table A.9: FIR models - Average impulse response fits and average of the selected orders when the F-test is applied after FPE to perform model order selection in the identification of 200 systems in each data set.

AIC

AIC was one of the criteria giving the best results when applied alone for order selection analysis, especially in the non-noisy data sets. For them, the selection of model structures with lower complexity by means of the application of the F-test in a second stage is not beneficial, while it is helpful in data sets S1D2 and S2D2, where AIC shows a more marked overfitting tendency.

BIC

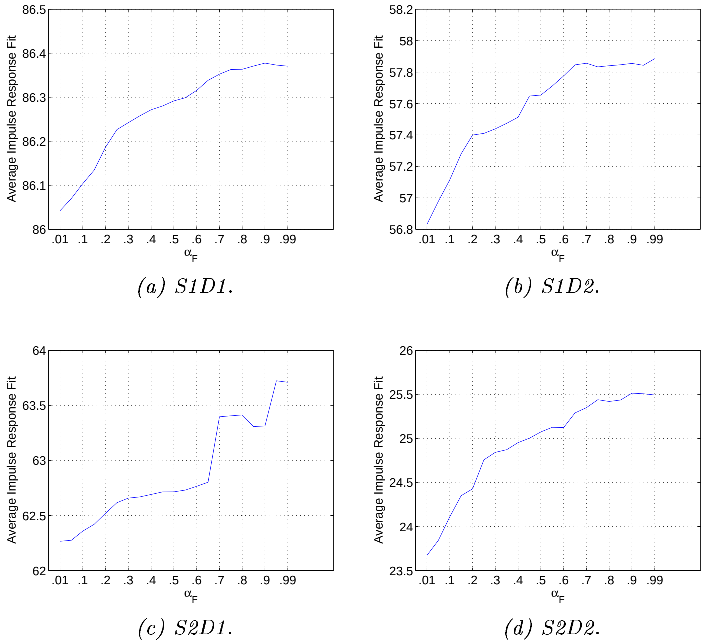
BIC criterion has been designed in order to identify the lowest complexity for which the corresponding model can properly reproduce the features of the original system. However, the large penalty which it inflicts to high complexities can also lead to the choice of too simple models, especially when slow systems has to be reproduced (as Figures A.11 and A.13 illustrate). These considerations suggest that the BIC criterion should benefit from its combination with the F-test for the evaluation of model structures of increasing complexity. Indeed, Figure A.21 confirms this analysis, showing also how a large value of the significance level α_F is more indicated for this application. This trend is particularly relevant in data set S2D1, whose systems are better described by complex model structures. Table A.10 also shows that the largest relative improvement w.r.t. the original fit achieved by BIC is detected in data set S2D1. Again, the results in Table A.10 are obtained with α_F set to the value leading to the largest average of the impulse response fit in each data set.

	Average impulse response fit		Average of the selected orders	
Set	*BIC*	*BIC + F (Up)*	*BIC*	*BIC + F (Up)*
S1D1	86	86.4	41.9	44
S1D2	56.8	57.9	17.6	19
S2D1	62.3	63.7	64.3	67.3
S2D2	23.6	25.5	23.1	24.7

Table A.10: FIR models - Average impulse response fits and average of the selected orders when the F-test is applied after BIC in order to perform model order selection in the identification of 200 systems in each data set.

Conclusions

The tests performed on FIR models for the combination between a model order selection method and the F-test (illustrated in Section 2.3.2) have proved its efficacy. Improvements have been detected in the impulse response fits obtained



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.21: FIR models - Average impulse response fits achieved for different values of the significance level α_F used in the F-test when it is applied after BIC in order to evaluate model structures with increasing complexity. The average is calculated from the identification of 200 systems in each data set.

thanks to this combination w.r.t. the ones achieved using only an order selection method.
Particular attention should be given to the choice of the significance level α_F for the F-test, since a wrong value could lead to a worsening of the fits previously achieved using only an order selection method. However, the analysis has suggested to adopt quite high significance levels when more complex model structures have to be tested, while low values of α_F are more suitable to test simpler models.

A.2 Order estimation for ARX models

A brief analysis of the order estimation for ARX models is done in this section. The analysis still exploits 200 sets of data in each of the four data sets introduced in Section 3.1. An order ranging from 1 to 40 has to be chosen for the estimated models. Again, the tests on the residuals evaluate the auto- and the cross-correlation until a lag $M = n_b + 20$, with n_b being the order of the polynomial $B(q)$ in (1.13).

A.2.1 Influence of the model order in the estimation

Figure A.22 shows the average fits achieved by ARX models with orders ranging from 1 to 40. In this case, larger complexities than OE models are more suitable. Indeed, with an ARX model the system dynamics and the noise are partially described by the same polynomial $A(q)$, thus limiting the degrees of freedom devoted to the description of the system dynamics. Because of this, larger complexities are required in order to achieve a good adherence with the true system. In particular, when slow systems have to be modelled and the measurement noise is not so significant (as for S2D1), higher complexities could have led to even better results.

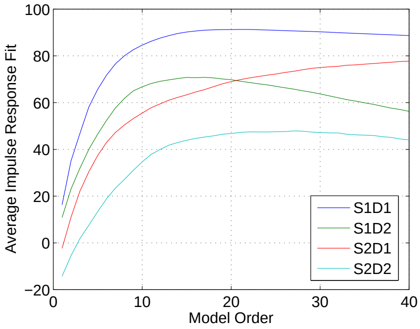
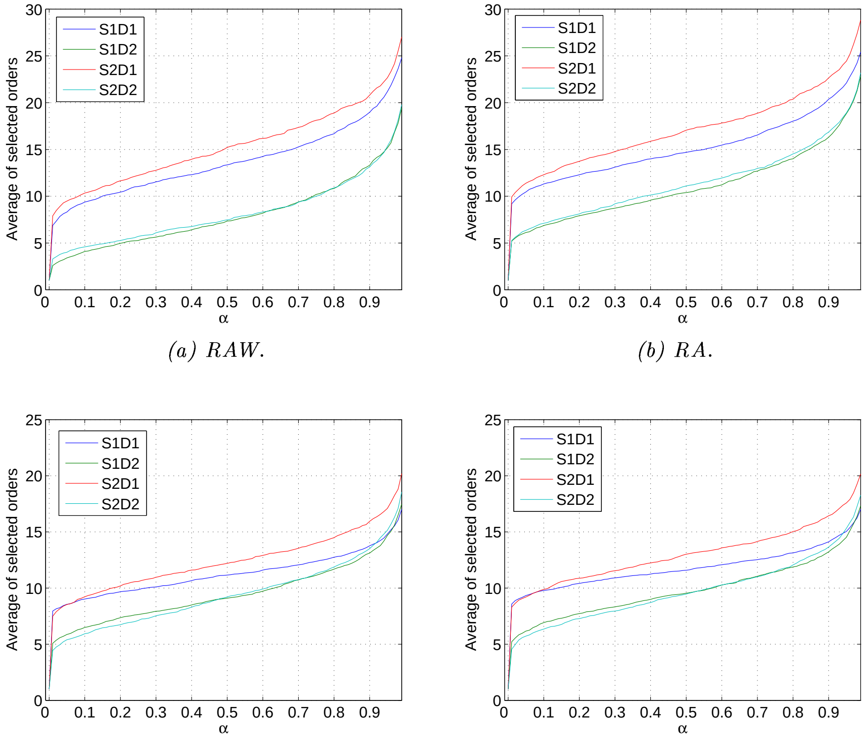


Figure A.22: ARX models - Average impulse response fits achieved in the estimation of 200 systems in each data set for model orders ranging from 1 to 40.

A.2.2 Selection of the significance level α used in the statistical tests

Figures A.23 and A.24 respectively show the average of the orders selected for the ARX models and the corresponding average fits achieved when the tests on the residuals with different significance levels are exploited for model order selection.

The main difference detected with ARX models with respect to the model structures analyzed in the previous sections regards the whiteness test. While with



(a) RAW.

(b) RA.

(c) RAI1.

(d) RAI2.

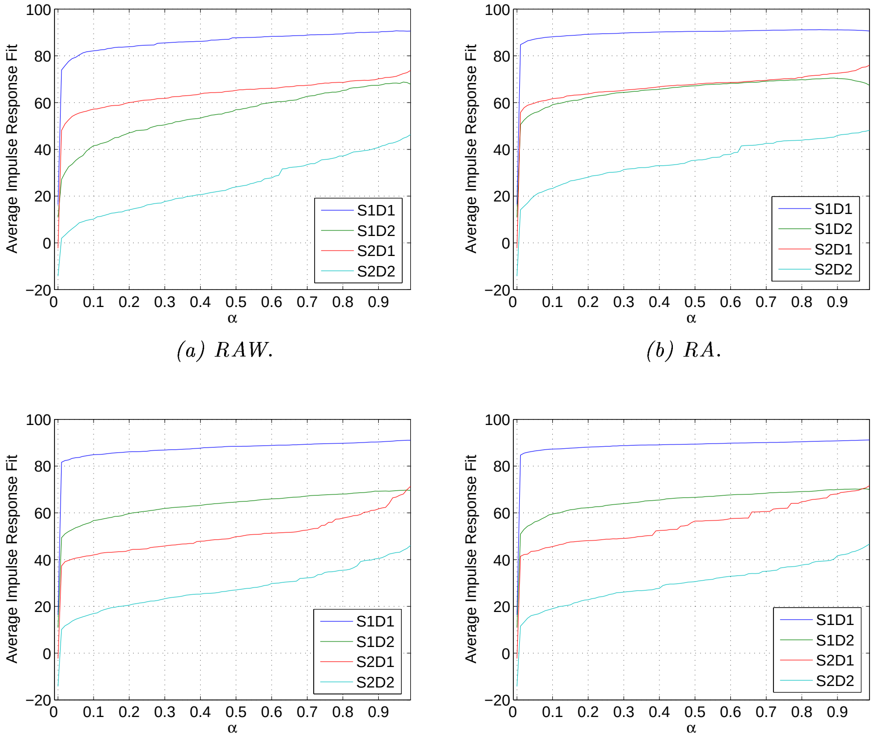
Figure A.23: ARX models - Average of the orders selected by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

FIR and OE models, that test sometimes was not able to unfalsify at least one of the evaluated model structures, with ARX models this issue is not present, thanks to the whitening effect provided by the polynomial $A(q)$ in the residuals,

$$\varepsilon(t) = A(q)\left(y(t) - \frac{B(q)}{A(q)}u(t)\right) \tag{A.1}$$

Because of this effect, RAW generally presents an undermodelling tendency that can be reduced using a quite high significance level α for the test.

The two implementations of the independence test on the residuals still choose lower complexities than the ones selected by the whiteness test. The choice of the maximal lag M for which the correlation is computed has proved to be quite influential: small values (such as 20) are more indicated when fast systems



(a) RAW.

(b) RA.

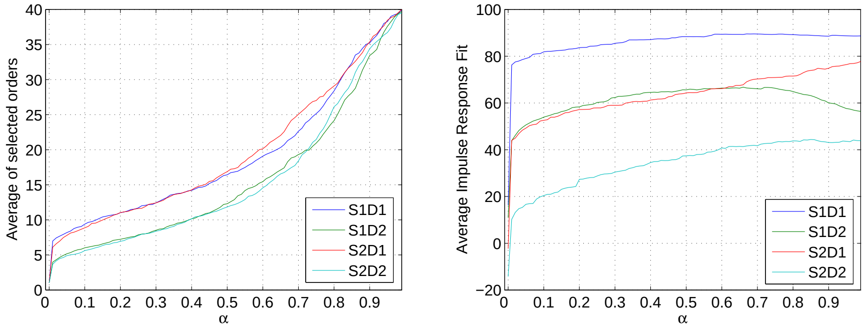
(c) RAI1.

(d) RAI2.

Figure A.24: ARX models - Average impulse response fits achieved by the statistical tests on the residuals as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

have to be identified, while larger ones are more suitable for slow systems. Figures A.24.(c) and (d) illustrate how a large significance level α leads to a more relevant improvement in the "slow" data sets than in the "fast" ones. Indeed, the adoption of a large α for data sets S2D1 and S2D2 can alleviate the performance reduction due to a too small value of M. However, when fast systems have to be identified, small values of M are more appropriate because the possible correlation between a residual and a past input acquires more weight in the total correlation. In this case, large significance levels are still preferable in order to reduce the undermodelling tendency that generally characterizes this order selection criterion.

Large significance levels have proved to be more effective also for the F-test, as Figure A.25.(b) illustrates: in particular, when slow systems have to be identified, very large significance levels are preferable. Figure A.25.(a) also shows how the order selection done for the noisy data sets S1D2 and S2D2 is again well distinguished from the one done for S1D1 and S2D1 when small significance levels are adopted. The reason of this behaviour is analogous to the one given for FIR models, since the loss function for data sets S1D1 and S2D1 undergoes a more significant decrease when low complexities models are evaluated.



(a) Average of the selected orders. *(b) Average impulse response fits.*

Figure A.25: ARX models - Average of the selected orders and average impulse response fits achieved by the F-test, as function of the adopted significance level α. The average is calculated from the identification of 200 systems in each data set.

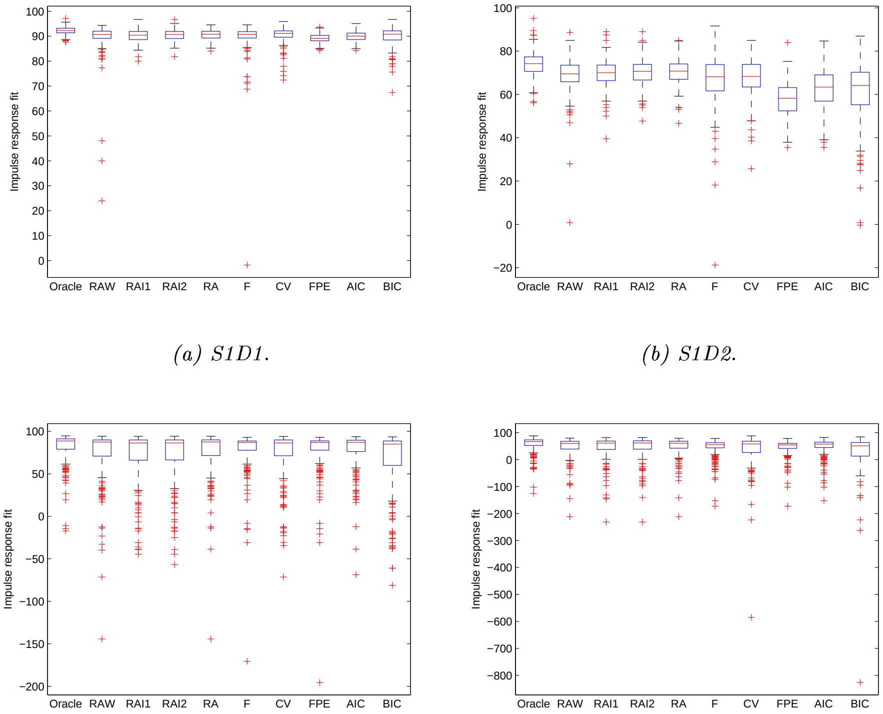
A.2.3 Analysis of the order selection criteria used in the statistical tests

Figure A.26 and Table A.11 respectively illustrate the box plots and the averages of the impulse response fits achieved by the different order selection techniques. Criteria which are based on statistical tests adopt the significance levels which lead to the highest average impulse response fit for each data set.

Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	92.3	90.7	91.1	91.2	91.2	89.6	90.5	89.2	89.9	90
S1D2	74.2	68.8	69.7	70.3	70.5	66.8	67.7	57.9	63	61.3
S2D1	81.2	73.8	71.3	71.6	76	77.7	73	77.3	77.7	66.6
S2D2	56.7	46.3	46	46.7	48.1	44.4	38.7	44.8	47.5	28.5

Table A.11: ARX models - Average impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

The criterion which gives the most satisfying performances is the combination of the whiteness and independence test on the residuals (RA). The independence



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.26: ARX models - Box plots of the impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

test alone (RAI1 or RAI2) leads to fits analogous to the ones reached by RA in the "fast" data sets, while its performances in the "slow" ones are less satisfying. This behaviour is due to the use of a too small value of the maximal lag M for which the cross-correlation between residuals and past inputs is computed. Further tests have proved that increasing the value of M gives rise to average fits that are comparable with RA even in S2D1 and S2D2.

FPE and AIC criteria still present the tendency to overfit, which penalizes the performances when fast systems have to be identified. However, this behaviour is particularly beneficial in data set S2D1, where the two criteria reach the best average impulse response fit. Though the overfitting tendency is again more evident for FPE than for AIC, AIC criterion does not appear able to appropriately penalize too high complexities.

On the other hand, BIC criterion is heavily penalized by its underfitting tendency, which makes BIC the worst criterion in the "slow" data sets S2D1 and S2D2.

Cross-validation behaves quite well when applied for the identification of fast systems, while it leads to undermodelling when slow systems have to be identified. This undermodelling could be justified by the fact that cross-validation is based on the prediction ability. Indeed, when models which contain a noise description are used (such as ARX models), low complexities are more beneficial in terms of prediction.

A.2.4 Combination of comparison and validation methods

The combination illustrated in Section 2.3.1 turns out to be effective also for the order estimation of ARX models. Indeed, the F-test, cross-validation and BIC criterion benefit from their combination with both the statistical tests on the residuals (RA) or with the independence test alone (RAI2). As partially observed also with the other model types, no positive effect is detected when a validation method is applied after FPE and AIC criterion: the overfitting tendency of these criteria lead to the choice of model structures that are unfalsified by the statistical tests on the residuals and therefore no order re-estimation is performed.

The use of both the statistical tests on the residuals (RA) leads to the best performances, especially in the "slow" data sets S2D1 and S2D2; also the independence test alone (RAI2) is effective in the "fast" data sets, while its performances deteriorate on the "slow" data sets, as was observed in the previous section for the use of this test alone.

For what regards the choice of the significance levels to be adopted in the tests, quite high values are suggested in order to favour the falsification of the originally chosen model structures and the consequent order re-estimation. However, too high values are not suitable when the measurement noise in the data is relevant.

A.2.5 Combination with the F-test

When the F-test is applied in order to "refine" the order selection done by the other criteria evaluated in this chapter, some improvements are detected, but a careful choice of the significance level used in the test is required. This choice is particularly awkward when the F-test is combined with the statistical tests on the residuals (RAW, RA, RAI1 and RAI2), because both the significance levels used in the two tests impact the performances. In this context, it is preferable to apply the F-test for the evaluation of more complex model structures than the ones originally selected by the tests on the residuals, since they have shown a light undermodelling tendency. While the combination of the tests on the residuals with the F-test does not lead to significant improvements in the achieved impulse response fits in data sets S1D1 and S1D2, the combination appears effective in data sets S2D1 and S2D2. In particular, the application of the F-test after RAI1 or RAI2 can remedy a non-optimal choice of the maximal lag M used for the correlation. For what regards the significance levels to be adopted here in the F-test, α_F should not be too large for data sets S1D1 and S1D2, in order to limit the selection of too complex models. On the other hand, significance levels larger than 0.5 are indicated for the identification of the systems in S2D1 and S2D2.

Also cross-validation and BIC criterion benefit when the F-test is applied after them to test more complex model structures. For what regards the choice of the significance level for the F-test, analogous considerations to the ones done for the tests on the residuals hold.

The F-test can be also exploited to refine the selection of the FPE and AIC criteria in S1D1 and S1D2, testing model structures of smaller complexities w.r.t. the ones returned by the two information criteria, which are known for their overmodelling tendency. In this case small significance levels are suggested for the F-test, in order to favour the selection of simpler models.

A.3 Order estimation for ARMAX models

The model order selection is now analyzed when ARMAX models are used to estimate the systems of the four data sets described in Section 3.1. In each of them 200 sets of measurement data are considered and an order between 1 and 40 has to be chosen by the order selection criteria here tested. The maximal lag M considered in the tests on the residuals is again $M = n_b + 20$.

A.3.1 Influence of the model order in the estimation

Figure A.27 contains the average of the impulse response fits obtained for different model orders. The clear non-smoothness of the plots proves how the estimation of ARMAX models is very sensible to the local minima issues, arising in the resolution of the non-convex optimization problems exploited for the estimation. In particular, the problem appears more relevant for complex models.

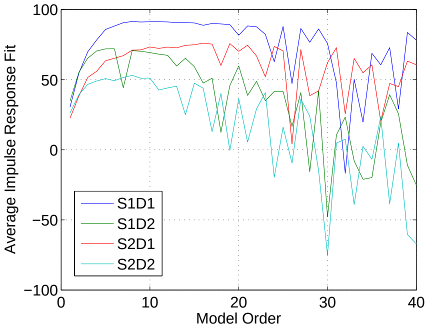


Figure A.27: ARMAX models - Average impulse response fits achieved in the estimation of 200 systems in each data set for model orders ranging from 1 to 40.

A.3.2 Selection of the significance level α used in the statistical tests

An analogy between ARX and ARMAX models is detected in relation to the whiteness test. Namely, at least a model structure among the evaluated ones is unfalsified. Actually, quite small orders are chosen by this criterion because the two polynomials used to model the noise component in the data, $C(q)$ and $A(q)$, act as a whitening filter on the residuals. Therefore, the whiteness test RAW shows an undermodelling tendency when it is used for model order selection; this tendency can be alleviated by the use of a large significance level, since it reduces the probability to unfalsify too simple model structures.

Differently from what observed with ARX models, the significance level adopted for the independence test on the residuals does not significantly affect the performances, but too small and too large values (lower than 0.2 and larger than 0.8) are not suggested. The only exception is detected in data set S2D1, where large values of α favour the selection of more complex model structures, that are more suitable for the description of the slow systems in S2D1.
The difference observed w.r.t. ARX models is due to the general tendency of this criterion to select quite low orders. Since with ARX models complex model structures were more suitable for the description of the true systems in the data sets, the slight undermodelling tendency penalized the performances of this criterion; therefore, a large significance level was suggested to alleviate the problem. On the other hand, low order ARMAX models can properly reproduce

the true systems of the data sets here considered, thus explaining the reduced influence of the significance level.

For what regards the use of both the statistical tests on the residuals, RA, analogous considerations to the ones done for RAI1 and RAI2 can be done.

The F-test still presents an undermodelling tendency, even if less marked than the one observed with ARX models; however, large significance levels are preferable also when applied on ARMAX models.

A.3.3 Analysis of the order selection criteria

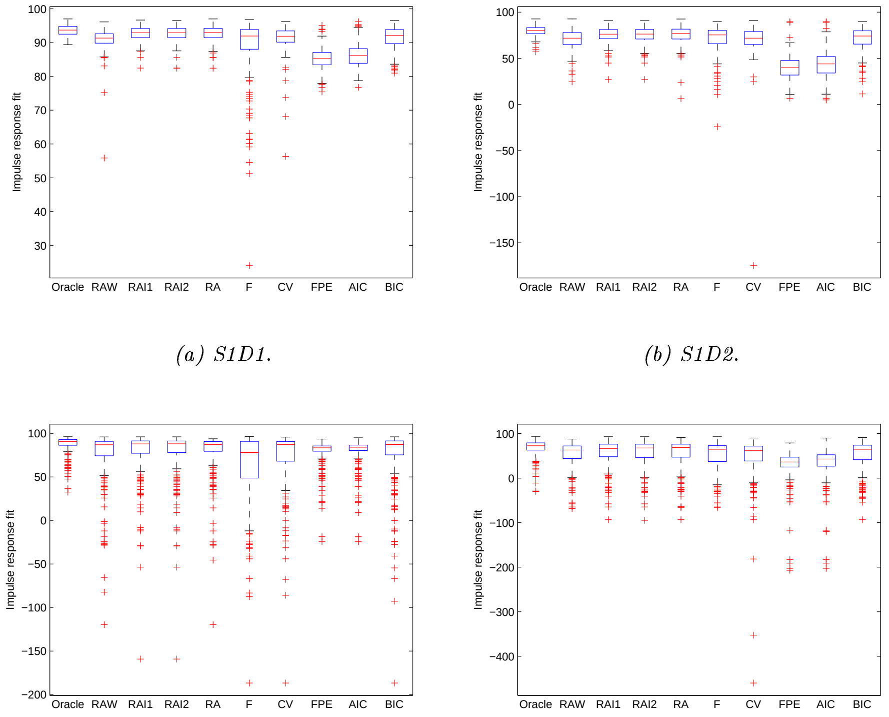
Table A.12 contains the average impulse response fits achieved by the evaluated order selection criteria, while Figure A.28 shows the corresponding box plots of the obtained fits. Again, the methods which are based on statistical tests are applied setting the significance level to its optimal value.

Set	*Or*	*RAW*	*RAI1*	*RAI2*	*RA*	*F*	*CV*	*FPE*	*AIC*	*BIC*
S1D1	93.7	90.9	92.7	92.7	92.7	88.7	91.3	85.2	86.4	91.6
S1D2	79.6	70.6	75.6	75.3	75.2	71	69.9	39.4	43.4	71.1
S2D1	87.3	73.9	76.1	76.1	77.4	62.4	70.3	78.3	79	72
S2D2	67.6	52.7	54.7	54.9	55.4	53	42.3	27.6	33	52.4

Table A.12: ARMAX models - Average impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

As observed for ARX models, the best performances in terms of impulse response fit are achieved by the independence test (RAI1 and RAI2) and by both the tests on the residuals (RA). In particular, the latter criterion performs better than the first one on the "slow" data sets, thanks to the selection of more complex model structures. On the other hand, in these data sets, the independence test tends to undermodel.
A further improvement of the performances given by RA, RAI1 and RAI2 on the "slow" data sets can be achieved by choosing a larger value of the maximal lag M for which the correlation is computed. However, larger values of M are less appropriate for the identification of the fast systems in S1D1 and S1D2, since they lead to the selection of simpler model structures. These considerations are in line with the findings described in Section 4.1.2, where the influence of the value of M was analyzed for OE models.

For what regards the information criteria, FPE and AIC show the classical overfitting tendency, which is beneficial only in data set S2D1. BIC criterion confirms its undermodelling property, which is here less detrimental than for



(a) S1D1.

(b) S1D2.

(c) S2D1.

(d) S2D2.

Figure A.28: ARMAX models - Box plots of the impulse response fits achieved by the evaluated criteria when 200 systems are identified in each data set.

ARX models, because low order ARMAX models allow a good reproduction of the true systems in the data sets. As expected, the data set in which BIC gives rise to the worst performances is S2D1.

As observed with ARX models, cross-validation leads to good impulse response fits on the "fast" data sets, while it is ineffective on the "slow" data sets. Partially, this behaviour can be justified by the same consideration done for ARX models. However, ARMAX models can be negatively affected by the local minima issue previously cited. Since cross-validation exploits a preliminary model estimate based on half of the available data, this model can be significantly different from the one estimated using all the data, especially when a local minimum is found in one of the two estimations. This consideration explains the presence of some outliers also in the box plots of the impulse response fits for data sets S1D1 and S1D2.

A.3.4 Combination of comparison and validation methods

As observed for ARX models the combination illustrated in Section 2.3.1 is effective for the F-test, cross-validation and BIC criterion, while FPE and AIC criteria do not benefit from the successive application of a validation method.
In this setting, the use of both the tests on the residuals (RA) leads to the best impulse response fits, but satisfying performances are achieved also by the test for the independence of the residuals from past inputs.
Since too complex ARMAX models are not indicated for a good reproduction of the systems in data sets S1D1, S1D2 and S2D2, significance levels around 0.5 or lower should be used in the statistical tests performed on the residuals. A different consideration holds for data set S2D1, for which larger significance levels are more suitable in order to favour the choice of more complex models.

A.3.5 Combination with the F-test

When the F-test is applied on the model structure returned by a certain model order selection criterion according to the way described in Section 2.3.2, some improvements are observed, w.r.t. the fits illustrated in Table A.12 and in Figure A.28.

The analysis of this procedure on the previously considered model types has underlined the critical importance of the significance level adopted in the F-test. The tests performed with ARMAX models confirm this property, especially when the F-test is applied after another criterion which involves a statistical test and whose performances in turn depend on the value chosen for the significance level, such as RAW, RAI1, RAI2 and RA.

The independence test on the residuals (RAI1 or RAI2) generally shows an undermodelling tendency, that is emphasized when the value of the maximal lag M for which the correlation is computed is not appropriately set. Therefore, the evaluation of model structures with increasing complexity by means of the F-test turns out to be effective after the application of RAI1 or RAI2. In this case significance levels larger than 0.5 are suggested for the F-test when it is applied on data set S2D1, while values smaller than 0.5 are more appropriate in the other data sets, since there is not a relevant necessity to choose more complex model structures. For what regards the significance level to be used for RAI1 and RAI2 instead, the analysis done in Section A.3.2 remains valid.

The observations done in Section A.3.3 in relation to cross-validation have highlighted the negative impact brought by the local minima issue. Thus, the F-test could be effective to refine the order selection originally done by cross-validation, since it could allow to depart from models that correspond to local minima. When the F-test is applied for this purpose, it is not clear whether simpler or more complex model structures should be evaluated. Indeed, both the choices lead to improvements in the impulse response fits achieved thanks to the successive application of the F-test. Also the choice of the significance level to be adopted in the F-test does not appear clear, but the performed tests have shown that improvements w.r.t. the fits reported in Table A.12 are achieved independently from the specific value of α.

Since FPE and AIC criteria show an overmodelling tendency also with ARMAX models, the F-test could be applied in a second stage to reduce the complexity of the selected model structures. However, this combination has not proved to be effective in the performed tests. On the contrary, when the F-test is exploited to alleviate the undermodelling property which characterizes BIC criterion, it is effective, especially when a large significance level is adopted for the F-test.

Bibliography

[1] T. Bohlin. *Interactive System Identification: Prospects and Pitfalls.* Springer-Verlag, Berlin, 1991.

[2] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and gaussian processes-revisited. *Automatica*, 48(8):1525–1535, 2012.

[3] T. Chen, G. Pillonetto, and L. Ljung. Kernel-based model order selection for linear system identification. *11th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing, ALCOSP'13, Caen, France*, July 2013.

[4] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[5] H. Hjalmarsson. Simplest unfalsified model selection. *Unpublished notes*, KTH, Stockholm, 2013.

[6] R. L. Kashyap and A. R. Rao. *Dynamic stochastic models from empirical data.* Academic Press, New York, 1976.

[7] Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res.*, 98888:1037–1057, June 2012.

[8] L. Ljung. *System Identification (2nd ed.): theory for the user.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 1999.

[9] L. Ljung. *System Identification Toolbox User's Guide, Version 8.2 (R2013a).* The MathWorks Inc., 3 Apple Hill Drive Natick, MA, USA, 2013.

[10] G. Pillonetto and G. D. Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46(1):81–93, 2010.

[11] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[12] T. Söderström and P. Stoica. On covariance function tests used in system identification. *Automatica*, 26(1):125–133, 1990.

[13] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.