On Resampling and Uncertainty Estimation in Linear System Identification ⋆

Simone Garatti [a], Robert R. Bitmead [b],

[a] *Dipartimento di Elettronica e Informazione - Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia.*

[b] *Deptartment of Mechanical & Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0411, USA.*

Abstract

Linear System Identification yields a nominal model parameter, which minimizes a specific criterion based on the single input-output data set. Here we investigate the utility of various methods for estimating the probability distribution of this nominal parameter using only the data from this single experiment. The results are compared to the actual parameter distribution generated by many Monte-Carlo runs of the data-collection experiment. The methods considered are collectively known as resampling schemes, which include Subsampling, the Jackknife, and the Bootstrap. The broad aim is to generate an empirical parameter distribution function via the construction of a large number of new data records from the original single set of data, based on an assumption that this data is representative of all possible data, and then to run the parameter estimator on each of these new records to develop the distribution function. The performance of these schemes is evaluated on a difficult, almost unidentifiable system, and compared to the standard results based on asymptotic normality. In addition to the exploration of this example as means to evaluate the strengths and weaknesses of these resampling schemes, some new theoretical results are proven and demonstrated for Subsampling schemes.

Key words: System Identification; Model Validation; Resampling; Monte-Carlo; Bootstrap; Subsampling

1 Introduction

Robust model-based control requires quantification of plant model uncertainty, [13,14,20,17]. System identification methods can be ill-equipped to provide a measure of parameter uncertainty other than that based on asymptotic-in-data variance formulæ derived from the Central Limit Theory, which in turn is based on a Taylor expansion of the empirical identification cost function about the correct parameter value [18,24]. Recent studies (in under-excited systems, [11,12]) have shown that cases can be found where the cost function is non-convex and these can have separated local minima. In such cases, the uncertainty characterization from asymptotic theory can be misleading.

Here we seek to develop an approach to the empirical calculation of the underlying distribution function of the parameter estimate, which is equally valid when the cost function is non-convex and which, asymptotically as the number of data points tends to infinity, fully characterizes the finite-data parameter distribution and, in the fixed-length case, yields a quantification of the error between the empirical distribution and the true underlying (and unknown) distribution. The approach is based on resampling ideas of the Bootstrap, the Jackknife, and Subsampling [21,29]. Our aim is to use the data to develop an approximation of the actual distribution function of the parameter estimate, based on the assumption that the data set is *representative* of the underlying stochastic processes.

We assume:

- We have N input-output pairs of data $\mathcal{X}^N = \{x_i = [u_i \quad y_i]^T, \quad i = 1, \ldots, N\}$, where u_i and y_i are scalars [1].

⋆ A preliminary version of this paper was presented at the 15th IFAC Symposium on System Identification. Corresponding author: S. Garatti. Tel. +39-02-2399-3650. Fax +39-02-2399-3412.

Email addresses: sgaratti@elet.polimi.it (Simone Garatti), rbitmead@ucsd.edu (Robert R. Bitmead).

[1] The multidimensional case (i.e. $u_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$.) presents no conceptual differences. We have preferred however to stick on the scalar input scalar output case to avoid

- These data are stationary and generated by a stable bivariate ARMA process, that is

$$A(z)\begin{bmatrix} u_t \\ y_t \end{bmatrix} = B(z)\eta_t, \tag{1}$$

where $A(z)$ and $B(z)$ are (2×2 and 2×1, respectively) polynomial matrixes of the forward shift operator z, and η_t is a bivariate i.i.d process. The (u_t, y_t) process above encompasses open-loop as well as closed-loop configurations.
- We seek to fit a fixed-order fixed-structure model parametrized by θ to the N-data set and to characterize the uncertainty in this parameter value. Specifically, we choose an empirical cost function $V(\theta, \mathcal{X}^N) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_{i|i-1}(\theta)\right)^2$, where $\hat{y}_{i|i-1}(\theta)$ is the optimal predictor based on the model corresponding to θ. If the data set to which the cost function refers is clear from the context, we shall write $V_N(\theta)$ in place of $V(\theta, \mathcal{X}^N)$. The minimizer of $V(\theta, \mathcal{X}^N)$ (assuming it is unique) is indicated by $\hat{\theta}_N$. Our goal is to reconstruct its probability distribution, hereafter indicated by $F_{\hat{\theta}_N}(\theta)$.

We present figures depicting distribution functions. To

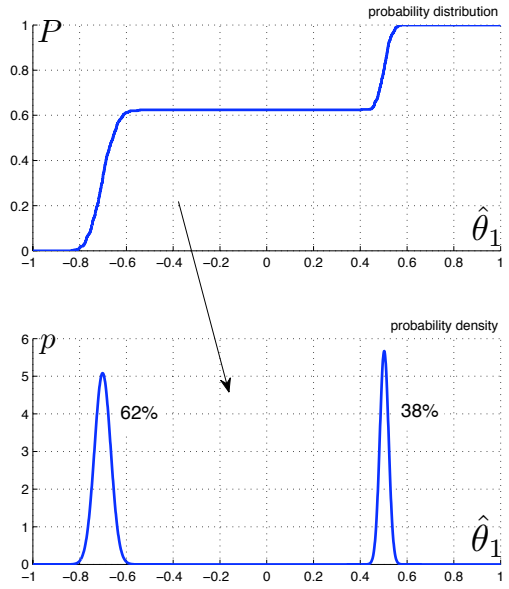

Figure 1. Distribution function (upper plot) and corresponding density function (lower plot) for the first parameter $\hat{a}_N = \hat{\theta}_N^1$. This illustrates that the steps in the distribution function correspond to peaks in the density with the step size related to the local density mass.

assist in the interpretation of these distribution functions, Figure 1 shows the density and distribution functions for the $\hat{\theta}_N^1 = \hat{a}_N$ component from an example to follow.

The paper is organized as follows. First, an example showing the limitations of the asymptotic theory of system identification is presented in Section 2. Then, some resampling strategies (namely; Monte-Carlo, Subsampling, Model-Based Jackknife, and Model-Based Bootstrapping) are briefly recalled in Section 3, with particular emphasis on their application in the system identification setting. The analysis of resampling techniques is given in Sections 4 and 5, while Section 6 provides a comparison based on the same example where asymptotic theory performed poorly.

2 Asymptotic theory and its limitations – the SMS example

The following example is taken from [11], with its eponym created as an acronym of the authors' first names. It shows a (somewhat contrived) situation where the blind use of the asymptotic theory of system identification as in [18,24], leads to an unreliable estimate of uncertainty unless the number of data is exceedingly large.

Consider the following data generating system:

$$y_t = \frac{b_0 z^{-1}}{1 + a_0 z^{-1}} u_t + (1 + h_0 z^{-1}) e_t, \tag{2}$$

where $\theta_0 = [a_0 \;\; b_0 \;\; h_0]^T = [-0.7 \;\; 0.3 \;\; 0.5]^T$ and $e_t \sim WGN(0,1)$, i.e. white gaussian noise with zero mean and unity variance. The system is operated in closed loop with the feedback law $u_t = r_t - y_t$, and with reference signal $r_t \sim WGN(0, 10^{-4})$ independent of e_t. The resulting closed-loop system is asymptotically stable. Note also that the variance of r_t is very small compared to the noise variance, so the system is poorly excited.

The identification experiment is as follows: $N = 2000$, (u, y) data points are collected, and a *full-order* model of the type

$$y_t = \frac{b z^{-1}}{1 + a z^{-1}} u_t + (1 + h z^{-1}) e_t$$

is identified by minimizing the empirical cost, $\hat{\theta}_N = \arg\min V_N(\theta)$, where $\theta = [a \; b \; h]^T$.

According to the asymptotic ($N \to \infty$) theory of system identification, the inflated estimation error, $\sqrt{N}(\hat{\theta}_N - \theta_0)$, is asymptotically distributed as a gaussian random variable with zero mean and covariance $P_\theta = \lambda_0 \cdot \left[\mathbf{E}\psi_t(\theta_0)\psi_t^T(\theta_0)\right]^{-1}$, with $\lambda_0 = \mathbf{E}e_t^2$ and

a cumbersome notation.

$\psi_t(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\hat{y}_{t|t-1}(\theta)$. Based on this theoretical result then, $\hat{\theta}_N$ is typically (and heuristically) presumed to be gaussian distributed too, with mean θ_0 and covariance $\frac{1}{N}P_\theta$. These values θ_0 and P_θ are replaced by their empirical counterparts (typically, $\hat{\theta}_N$ and $\sum_{i=1}^N (y_i - \hat{y}_{i|i-1}(\hat{\theta}_N))^2 \times \left[\sum_{i=1}^N \psi_i(\hat{\theta}_N)\psi_i^T(\hat{\theta}_N)\right]^{-1}$) so as to obtain an empirical estimate of the probability distribution of $\hat{\theta}_N$ based on available data only.

Though commonly used in practice, the above approach has only heuristic validity with N finite, and, in the present setting with $N = 2000$, it fails to return a sensible estimate of the distribution of $\hat{\theta}_N$. This is clearly depicted in Figure 2, where the empirical distribution estimate computed according to the rationale above is compared with the actual distribution of $\hat{\theta}_N$, reconstructed here through Monte-Carlo simulations. From the plot, a wide mismatch between the estimated distribution and the actual one is apparent, with the former being a gaussian centered on the estimate, $\hat{\theta}_N = [0.46 \ -0.84 \ -0.68]^T$, and the actual distribution being bi-modal and hence not gaussian. The empirical distribution estimate does not capture the mismatch between the obtained $\hat{\theta}_N$ and θ_0, so that the uncertainty evaluation via the asymptotic theory is unreliable in this case. Notably, the approximation by a gaussian is inadequate, even if this gaussian were to be computed based on the exact parameter values.

Let us briefly discuss the mechanism underlying the underperformance of the asymptotic theory. Asymptotic theory, as presented by Ljung for example [18], involves the Taylor expansion of the gradient of the identification cost function $V_N(\theta)$ about an isolated minimizing value θ_0 of the probabilistic cost function $\overline{V}(\theta) \triangleq \mathbb{E}[(y_k - \hat{y}_{k|k-1}(\theta))^2]$. That is (here V_N' denotes the gradient transpose and V_N'' is the Hessian matrix)

$$V_N'(\hat{\theta}_N) = 0 = V_N'(\theta_0) + V_N''(\xi_N)(\hat{\theta}_N - \theta_0), \qquad (3)$$

for some $\xi_N = \theta_0 + \mathrm{diag}([\epsilon_1 \ \epsilon_2 \ \cdots \ \epsilon_n])(\hat{\theta}_N - \theta_0)$, $\epsilon_i \in [0,1]$, $i = 1, \ldots, n$.

Equation (3) in turn can be rewritten as:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = -V_N''(\xi_N)^{-1} \cdot \sqrt{N}V_N'(\theta_0), \qquad (4)$$

and the explication of the terms $\sqrt{N}V_N'(\theta_0)$ and $V_N''(\xi_N)$ yields the standard results on asymptotic normality for the various system identification criteria.

In the asymptotic theory, one relies on two properties to recover the asymptotic variance formulæ used for parameter uncertainty estimation:

- the central limit theorem is used to describe the weak convergence of $\sqrt{N}V_N'(\theta_0)$ to a gaussian random variable;
- the quantity $V_N''(\xi_N)$ is replaced by $V_N''(\hat{\theta}_N)$, both of which are assumed to be close to $\overline{V}''(\theta_0)$.

It is this latter substitution which lies at the heart of the difficulty in applying these results with "small" N.

As shown in [11] in the SMS example setting, $\overline{V}(\theta)$ is a non-convex function with several minima. Precisely, if r_t had been a zero signal, there would have been two global minima, one corresponding to $\theta_0 = [-0.7 \ 0.3 \ 0.5]^T$ and the other to $\theta^* = [0.5 \ -0.9 \ -0.7]^T$, see [11]. When instead r_t is not zero but has only a small variance as in our example, θ_0 remains a global minimum while θ^* becomes just a local one, but $\overline{V}(\theta_0)$ and $\overline{V}(\theta^*)$ are very close. (Actually, their difference can be made as small as desired by reducing the variance of r_t.) This latter fact in turn implies that, because $V_N(\theta)$ is a perturbed version of $\overline{V}(\theta)$, the minimizer $\hat{\theta}_N$ of $V_N(\theta)$ will end up close to θ^* instead of θ_0 with non-vanishing probability. With $N = 2000$ this probability is about 38% as revealed by the actual distribution function of $\hat{\theta}_N$ plotted in Figure 2. When $\hat{\theta}_N \approx \theta^*$, it is no longer true that $V_N''(\xi_N) \approx \overline{V}''(\theta_0)$ nor that $V_N''(\xi_N) \approx V_N''(\hat{\theta}_N)$, and presuming the gaussianity of $\sqrt{N}(\hat{\theta}_N - \theta_0)$ from (4) also is no longer valid. Yet, in the neighborhood of each local minimum, the distribution is approximately gaussian.

It is worth noticing that, as N increases, with probability tending to one, $\hat{\theta}_N$ lies in the neighborhood of θ_0 and the asymptotic theory is valid. It should be clear that theoretical achievements of the asymptotic theory are not at issue in the SMS example. Our criticism regards only the heuristic use of asymptotic results with N finite. Clearly, the validity of the asymptotic theory for $N = 2000$ in this example is compromised by the paucity of excitation, leading to very weak identifiability of the correct parameter value. For sufficiently large N and even for this example, the asymptotic results are valid.

3 Resampling Strategies

As shown in Section 2, there are cases where one cannot rely on the asymptotic theory of system identification for a reliable description of the probability distribution of the identified parameter vector with seemingly large values of N. In particular, the SMS example reveals a circumstance where there are two closely competing but geometrically separated points θ^* and θ_0, and the asymptotic theory fails to reveal this dichotomy of solutions.

In order to provide a fair evaluation of uncertainty, some different tools have to be considered, and in this paper

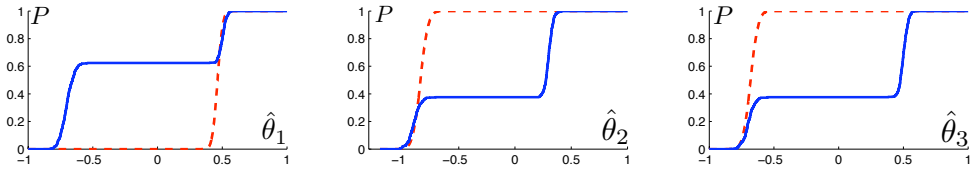

Figure 2. The actual distributions of each component of $\hat{\theta}_N$ (solid blue line) vs. the distributions returned by the asymptotic theory of system identification (dashed red line).

the focus is on resampling strategies, [8,9,10,21,23]. Resampling methods have recently attracted the attention of the systems and control community, [1,7,26,27,25,28]. Yet, they have not met with wide acceptance, at least not as in other fields such as statistics, econometrics, and signal processing. Our main objective here is to examine whether these methods overcome the difficulties of the asymptotic theory in contexts as challenging as the SMS example.

Four different resampling methodologies for the reconstruction of the underlying probability distribution of the identified parameter $\hat{\theta}_N$ are considered: namely, Monte-Carlo, Subsampling, Model-Based Jackknife, and Model-Based Bootstrapping. In the following, a brief description of each of them is provided for the sake of completeness. These approaches have been first developed in the context of independent data and then extended to the dependent case. Here, only the latter is treated for it is the framework of system identification problems. Should the reader desire further information on resampling with dependent data outside the system identification context, Lahiri's recent book [15] may be consulted.

3.1 Monte-Carlo

The Monte-Carlo procedure amounts to repeating the identification experiment m times so as to collect m independent N-long data sequences $(\mathcal{X}_1^N, \ldots, \mathcal{X}_m^N)$, which in turn, by minimizing $V(\theta, \mathcal{X}_i^N)$, $i = 1, \ldots, m$, yield m different parameter estimates $(\hat{\theta}_N^1, \ldots, \hat{\theta}_N^m)$. These estimates are then used to reconstruct the probability distribution of $\hat{\theta}_N$ as

$$F^{MC}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{[\hat{\theta}_N^i \leq \theta]},$$

where the vector inequality $\hat{\theta}_N^i \leq \theta$ is taken componentwise and $\mathbf{1}_{[.]}$ is the indicator function. The Monte-Carlo strategy scheme has been graphically depicted in Figure 3 to help the reader.

It is well known that $F^{MC}(\theta)$ is an unbiased and consistent (both mean square and almost sure) estimator of the actual probability distribution $F_{\hat{\theta}_N}(\theta)$ (see [23]). However, computing $F^{MC}(\theta)$ requires more data than those actually available, and thus is infeasible in general. The Monte-Carlo method has been introduced for comparison with other resampling methodologies.

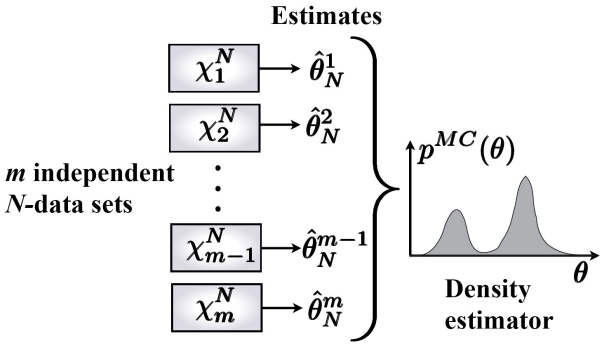

Figure 3. The Monte-Carlo procedure.

3.2 Subsampling

Subsampling was first introduced in [22], and despite its attractive properties, has not received much attention from the systems and control community yet except for [7]. Subsampling is quite intuitive and is reminiscent of the Monte-Carlo approach. A single data set, however, is used. Precisely, choose $m \leq N$ and $N_S \leq N - m + 1$. From the N-long available data set $\mathcal{X}^N$, the set of all N_S-long sub-sequences of consecutive data points is considered and, among these, m are extracted, that is:

$$(\mathcal{X}_1^{N_S}, \ldots, \mathcal{X}_m^{N_S}) \subseteq (X_1^{N_S}, X_2^{N_S}, \ldots, X_{N-N_S}^{N_S}),$$

where $X_j^{N_S} = \{x_{j+1} \quad x_{j+2} \cdots x_{j+N_S}\}$.

Starting from the chosen sub-sequences $(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S})$, m different parameter estimates $(\hat{\theta}_{N_S}^1, \ldots, \hat{\theta}_{N_S}^m)$ are derived by minimizing each time the identification cost criterion $V(\theta, \mathcal{X}_i^{N_S})$, $i = 1, \ldots, m$ based on N_S data points only. The distribution of $\hat{\theta}_N$ is then reconstructed as the empirical distribution of the $\hat{\theta}_{N_S}^i$s, i.e.

$$F^{SS}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{[\hat{\theta}_{N_S}^i \leq \theta]}.$$

The Subsampling scheme has been graphically depicted in Figure 4.

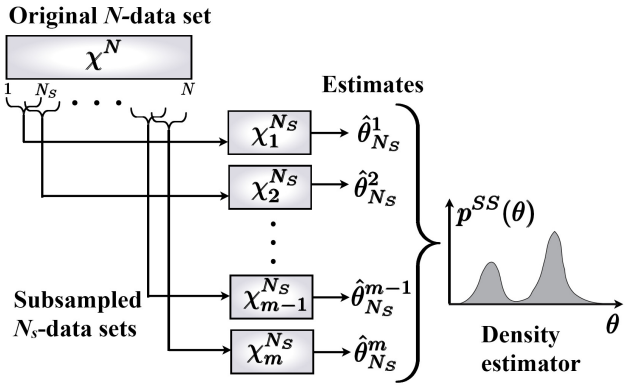

Figure 4. The Subsampling procedure.

The main point in Subsampling is that since the $\mathcal{X}_i^{N_S}$s are taken from the actual data set $\mathcal{X}^N$, they are distributed identically to the original data, although their length is reduced. The choice of N_S is a degree of freedom of Subsampling, and a sensible tuning of N_S is of paramount importance. Also, the choice of the $\mathcal{X}_i^{N_S}$s among $(X_1^{N_S}, X_2^{N_S}, \ldots, X_{N-N_S}^{N_S})$ is relevant to the final result, because of the inherent dependence between the data sets. These aspects will be treated in the next Section 4 where the analysis of Subsampling is provided. It is worth noting that Subsampling does not correspond to performing m Monte-Carlo N_S-long simulations, since the sub-sequences $\mathcal{X}_i^{N_S}$ are correlated in general (actually, they can even be overlapping). Showing that $F^{SS}(\theta)$ is unbiased and consistent is not straightforward.

3.3 Model-Based Jackknife & Bootstrap

Given the data sequence $\mathcal{X}^N$, estimates $\hat{A}(z)$ and $\hat{B}(z)$ of $A(z)$ and $B(z)$ in (1) are obtained according to some identification algorithm. This identification algorithm need not to be the same as that used for computing $\hat{\theta}_N$, and even the family of models from among which $\hat{A}(z)$ and $\hat{B}(z)$ are found can be different from that parametrized by θ, see [7,28]. These authors consider, for example, high-order modeling for the generation of the residual sequence, which then is applied via the Jackknife or Bootstrap for the development of further "data" sequences. To be precise, given the model estimate $\{\hat{A}(z), \hat{B}(z)\}$, the one-step prediction residuals $(\epsilon_1, \ldots, \epsilon_N)$ are computed according to the following equation:

$$\epsilon_t = \hat{B}(z)^{-1}\hat{A}(z)\begin{bmatrix} u_t \\ y_t \end{bmatrix}.$$

The residual sequence is asymptotically (as N increases) independent and equal to η_t provided $\hat{A}(z)$ and $\hat{B}(z)$ are consistent estimates of $A(z)$ and $B(z)$. Such a sequence of residuals is the basis of the Model-Based Jackknife and Bootstrap procedures.

In Model-Based Jackknife, from $(\epsilon_1, \ldots, \epsilon_N)$, a new artificial residual sequence $(\epsilon_1^R, \ldots, \epsilon_N^R) = (\epsilon_{i_1}, \ldots, \epsilon_{i_N})$ is generated by picking (with or without replacement) indices $i_1, \ldots, i_N$ at random from integers $1, \ldots, N$. The new residual sequence $(\epsilon_1^R, \ldots, \epsilon_N^R)$ can be thought of as new extracted samples from the noise process η_t and it can be used for computing a new (resampled) N-long input/output data sequence $(u_1^R, y_1^R, u_2^R, y_2^R, \ldots, u_N^R, y_N^R)$ according to the following mechanism:

$$\hat{A}(z)\begin{bmatrix} u_t^R \\ y_t^R \end{bmatrix} = \hat{B}(z)\epsilon_t^R.$$

This resampled data sequence in turn is used to produce a new parameter estimate $\hat{\theta}_N^1$ by minimizing the usual cost criterion. Repeating the residual resampling process m times yields a sequence of m parameter estimates $\hat{\theta}_N^1, \ldots, \hat{\theta}_N^m$ whose empirical distribution is used to reconstruct the probability distribution of $\hat{\theta}_N$:

$$F^{JK}(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_{[\hat{\theta}_N^i \le \theta]}.$$

See Figure 5 for a graphical representation of the Model-

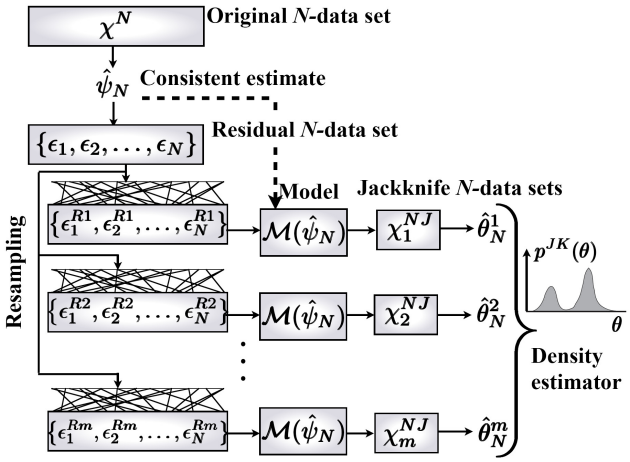

Figure 5. The Model-Based Jackknife procedure.

Based Jackknife algorithm.

If one uses selection without replacement for sequences of size N from a set of N residuals, then there are a maximum of $N!$ distinct sequences, which are clearly permutations of original sequence of residuals. With replacement, this maximal number of distinct residual sequences is N^N.

The main idea behind Model-Based Bootstrapping is similar to that for Model-Based Jackknife. That is, from $\mathcal{X}^N$ a consistent model is identified and its prediction

residuals $(\epsilon_1, \ldots, \epsilon_N)$ are used to generate artificial residual and data sequences; from these latter, the corresponding parameter estimates $\hat{\theta}_N^1, \ldots, \hat{\theta}_N^m$ are computed and their empirical distribution,

$$F^{BS}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[\hat{\theta}_N^i \leq \theta]},$$

reconstructs the actual distribution of $\hat{\theta}_N$. The difference between Bootstrapping and the Jackknife is that the artificial residual sequences are not obtained by resampling $(\epsilon_1, \ldots, \epsilon_N)$. Instead, this latter sequence is first used to reconstruct the complete distribution function, $\widehat{F}_{\epsilon_t}(\epsilon)$, of the residual error ϵ_t; then, artificial residual sequences are generated by extracting each time N samples according to $\widehat{F}_{\epsilon_t}(\epsilon)$. See Figure 6 for a graphical

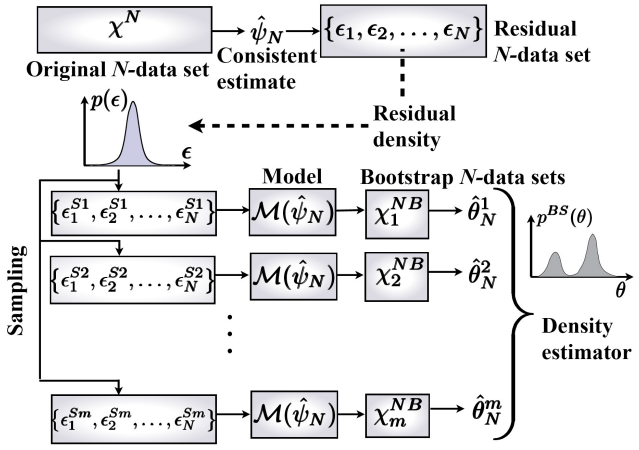

Figure 6. The Model-Based Bootstrap procedure.

representation of the Model-Based Bootstrap algorithm.

The reconstruction of the distribution of ϵ_t can be performed according to a number of techniques; empirical sum, L^1 approximation, kernel methods, etc. It is worth noticing that when the empirical sum is used (i.e. $\widehat{F}_{\epsilon_t}(\epsilon) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{[\epsilon^i \leq \epsilon]}$) Model-Based Bootstrap is the same as the Jackknife provided that $(\epsilon_1, \ldots, \epsilon_N)$ are resampled with replacement. When the empirical residual distribution is estimated in continuous functional form, then the number of distinct resampled residual sequences is unlimited, unlike the Jackknife. Because of the similarities between Model-Based Jackknife and Model-Based Bootstrap, the asymptotic analysis is considered only for the Bootstrap in Section 5.

Remark 1 (Terminology issues) *Normally, in the independent data case, the Jackknife is based on sequentially deleting one observation (or several observations in the block or grouped Jackknife) and is used for estimation of bias and variance. In the context considered, however, this standard procedure seems to be ill-suited due to the dependence between data, which requires that resampling must be applied to some independent reconstructed sequence like the residuals in Model-Based Jackknife and Bootstrap, and to the fact that we are interested in estimating the whole probability distribution of $\widehat{\theta}_N$ rather than its bias and variance. To the best of our knowledge, there is no universal acceptance of what the Jackknife is in the dependent case and for probability distribution reconstruction. Some authors prefer to call the procedure introduced above as "Resampling". We have, however, preferred to stick to "Jackknife" since "Resampling" refers also to all the introduced schemes (Monte-Carlo, Subsampling, Jackknife and Bootstrap).*

3.4 Alternative Bootstrap approaches

Perhaps, it is worth mentioning that some alternative schemes other than Model-Based Jackknife and Bootstrap have been introduced in the literature in order to cope with the data-dependent context. Among these, Moving-Block Bootstrap and Transformation-Based Bootstrap play a prominent role, see e.g. [16]. These two methods do not require the identification of a consistent model of the true system in order to compute residuals, and this makes them attractive in the presence of undermodeling. Nonetheless, Moving-Block Bootstrap and Transformation-Based Bootstrap have their own drawbacks, the most severe of which is that they have been developed/studied for times series analysis only, and the extension to the case of input/output system identification does not seem trivial. In particular, to the best of the authors' knowledge, no consistency results are available in this latter case. For the sake of completeness, however, Moving-Block Bootstrap and Transformation-Based Bootstrap will be reconsidered in the simulation results Section 6 by briefly introducing the corresponding algorithms and presenting an informal comparison based on the SMS example with the other approaches.

4 Analysis of the Subsampling method

In this section, we establish our main theoretical result concerning the consistency of the Subsampling procedure. To be precise, we will show that the probability distribution reconstructed via Subsampling, $F^{SS}(\theta)$, is a consistent estimate of the actual distribution of the parameter estimate identified with N_S data points, $F_{\hat{\theta}_{N_S}}(\theta)$. The proof relies on the fact that, in ARMAX processes, the dependence between data at two different time instants vanishes as the time lag between them increases, so that sufficiently distant Subsampled data subsequences behave as if they were obtained by means of independent experiments. In the following, we will give some preliminary results on α-mixing processes which characterize such dependence between data (Subsection 4.1). Based on these results, we will prove the consistency of Subsampling (Subsection 4.2). Finally,

some concluding considerations about Subsampling and its usage are provided (Subsection 4.3).

4.1 Preliminary definitions and results

We need the following preliminary definition, see e.g. [3].

Definition 1 (α-mixing) *Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary random process in $\mathbb{R}^l$ and let $\mathcal{A}_0$ be the σ-algebra generated by $\{Y_t\}_{t\leq 0}$ and $\mathcal{A}^\tau$ that generated by $\{Y_t\}_{t\geq\tau}$, $\tau\geq 0$. Then, the α- (or strong) mixing coefficient for $\{Y_t\}$ is defined as:*

$$\alpha_Y(\tau) \triangleq \sup_{A,B}\{|\mathbf{P}(A\cap B)-\mathbf{P}(A)\mathbf{P}(B)|,\ A\in\mathcal{A}_0, B\in\mathcal{A}^\tau\}.$$

If $\alpha_Y(\tau)\to 0$ as $\tau\to\infty$, then $\{Y_t\}$ is said to be α- (or strong) mixing. If in addition $\alpha_Y(\tau)\leq\rho^\tau$ for a certain $\rho\in(0,1)$ then $\{Y_t\}$ is said to be geometrically α- (or geometrically strong) mixing.

We have the following lemma, in line with [22], bounding the empirical distribution mean square error for α-mixing processes.

Lemma 1 *Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a stationary random process in $\mathbb{R}^l$ and suppose that Y_t is α-mixing. Let $\varphi:\mathbb{R}^l\to\mathbb{R}^k$ be any measurable function and let $\widehat{F}(x)=\frac{1}{m}\sum_{i=1}^m \mathbf{1}_{[\varphi(Y_i)\leq x]}$ be the empirical probability distribution of $\varphi(Y_t)$ and $F(x)=\mathbf{P}(\varphi(Y_t)\leq x)$ be the actual probability distribution. (Here, as usual, the vector inequalities are taken componentwise.) Then, for every x, we have that*

$$\mathbf{E}\big((\widehat{F}(x)-F(x))^2\big)\leq\frac{12}{m^2}\sum_{\tau=-m}^{m}(m-|\tau|)\cdot\alpha_Y(|\tau|).$$

Proof: Clearly, $\mathbf{P}(\varphi(Y_t)\leq x)=\mathbf{E}(\mathbf{1}_{[\varphi(Y_t)\leq x]})$. Let $Z_t=\mathbf{1}_{[\varphi(Y_t)\leq x]}-\mathbf{E}(\mathbf{1}_{[\varphi(Y_t)\leq x]})$. Z_t is zero mean and, moreover, it is stationary since Y_t is. Letting $\gamma_Z(\tau)=\mathbf{E}\big(Z_tZ_{t+\tau}\big)=\gamma_Z(-\tau)$ be the covariance function of Z_t, we have that

$$\begin{aligned}\mathbf{E}\big((\widehat{F}(x)-F(x))^2\big)&=\mathbf{E}\Big(\big(\frac{1}{m}\sum_{i=1}^m Z_i\big)^2\Big)\\&=\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m\mathbf{E}(Z_iZ_j)\\&=\frac{1}{m^2}\sum_{\tau=-m}^{m}(m-|\tau|)\gamma_Z(\tau)\\&\leq\frac{1}{m^2}\sum_{\tau=-m}^{m}(m-|\tau|)|\gamma_Z(\tau)|.\end{aligned}$$

Since $Z_t\in[-1,1]$ and is measurable with respect to the σ-algebra generated by Y_t, then we have that $|\gamma_Z(\tau)|\leq 12\alpha_Y(|\tau|)$ (See the Corollary of Lemma 2.1 in [6].) leading to the requisite bound. □

The following result is a straightforward consequence of Lemma 1

Corollary 1 *If $\frac{12}{m^2}\sum_{\tau=-m}^m(m-|\tau|)\cdot\alpha_Y(|\tau|)\to 0$ as $m\to\infty$, then $\hat{F}(x)$ is mean square convergent to $F(x)$*

4.2 Subsampling strategies and mixing conditions

We want now to apply Lemma 1 to the Subsampling reconstructed distribution function $F^{SS}(\theta)=\frac{1}{m}\sum_{i=1}^m\mathbf{1}_{[\hat{\theta}^i_{N_S}\leq\theta]}$. In this case, $\mathcal{X}_i^{N_S}$ plays the role of the process Y_t in Lemma 1 while $\hat{\theta}^i_N$, the parameter vector estimated from the subsequence $\mathcal{X}_i^{N_S}$, plays that of $\varphi(Y_t)$. Clearly, $\hat{\theta}^i_N$ is a measurable function of $\mathcal{X}_i^{N_S}$. As for the process $\mathcal{X}_i^{N_S}$ we need to check whether it is:

1. stationary;
2. α-mixing.

As for Point 1, recall that

$$(\mathcal{X}_1^{N_S},\ldots,\mathcal{X}_m^{N_S})\subseteq(X_1^{N_S},X_2^{N_S},\ldots,X_{N-N_S}^{N_S}),$$

where $X_j^{N_S}=\{x_{j+1}\quad x_{j+2}\ \cdots\ x_{j+N_S}\}$; $x_t=[u_t\ \ y_t]^T$, in turn, is generated as a stationary ARMA process:

$$A(z)x_t=B(z)\eta_t.$$

It easily follows that $X_j^{N_S}$ is stationary too, while $\mathcal{X}_i^{N_S}$ is stationary as long as the $\mathcal{X}_i^{N_S}$s are chosen from the $X_j^{N_S}$s in a equally time-spaced manner. That is, if $\mathcal{X}_i^{N_S}=X_{j_1}^{N_S}$ and $\mathcal{X}_{i+1}^{N_S}=X_{j_2}^{N_S}$, then the difference $j2-j1$ must be the same whatever i is. Some possible choices ensuring stationarity are the following ones where subsequences are overlapping in the first two cases and non-overlapping in the later two. In the following, $\lfloor\cdot\rfloor$ denotes the integer part and $k\leq N_S$.

$$\begin{aligned}&(\mathcal{X}_1^{N_S},\ldots\mathcal{X}_m^{N_S})=(X_1^{N_S},X_2^{N_S},\ldots,X_{N-N_S}^{N_S}),\\&(\mathcal{X}_1^{N_S},\ldots\mathcal{X}_m^{N_S})=(X_1^{N_S},X_{k+1}^{N_S},\ldots,X_{\lfloor\frac{N-N_S}{k}\rfloor\cdot k+1}^{N_S}),\\&(\mathcal{X}_1^{N_S},\ldots\mathcal{X}_m^{N_S})=(X_1^{N_S},X_{N_S+1}^{N_S},\ldots,X_{\lfloor\frac{N}{N_S}-1\rfloor\cdot N_S+1}^{N_S}),\\&(\mathcal{X}_1^{N_S},\ldots\mathcal{X}_m^{N_S})=(X_1^{N_S},X_{2N_S+1}^{N_S},\ldots,X_{\lfloor\frac{N}{2N_S}-1\rfloor\cdot 2N_S+1}^{N_S}).\end{aligned}$$

As for Point 2, note first that, since x_t is a stationary ARMA process, it is geometrically α-mixing as long as the mild assumption that the probability distribution of the noise η_t admits a probability density is satisfied, [3,19]. Thus, letting $\alpha_x(\tau)$ be the α-mixing coefficient of

x_t, we have that $\alpha_x(\tau) \leq \rho_x^\tau$ for a certain $\rho_x \in (0,1)$. (It is worth noticing that ρ_x is strictly related to the maximum modulus pole of the ARMA system in (1).) From the α-mixing property of x_t it easily follows that $X_t^{N_S}$ and, in turn, $\mathcal{X}_t^{N_S}$ are α-mixing too.

The α mixing coefficient of $\mathcal{X}_t^{N_S}$ (say $\alpha_{\mathcal{X}}(\tau)$), however, depends on how subsequences $(\mathcal{X}_1^{N_S}, \ldots, \mathcal{X}_m^{N_S})$ are chosen from $(X_1^{N_S}, X_2^{N_S}, \ldots, X_{N-N_S}^{N_S})$. With reference to the examples given above we have for $(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S})$:

$$
\begin{aligned}
&(X_1^{N_S}, X_2^{N_S}, \ldots) \Longrightarrow \alpha_{\mathcal{X}}(\tau) \leq \rho_x^{(\tau - N_S + 1)\cdot \mathbf{1}_{[\tau \geq N_S]}},\\
&(X_1^{N_S}, X_{k+1}^{N_S}, \ldots) \Longrightarrow \alpha_{\mathcal{X}}(\tau) \leq \rho_x^{[\tau k - N_S + 1]\cdot \mathbf{1}_{[\tau k \geq N_S]}},\\
&(X_1^{N_S}, X_{N_S+1}^{N_S}, \ldots) \Longrightarrow \alpha_{\mathcal{X}}(\tau) \leq \rho_x^{[(\tau-1) N_S + 1]\cdot \mathbf{1}_{[\tau > 0]}},\\
&(X_1^{N_S}, X_{2N_S+1}^{N_S}, \ldots) \Longrightarrow \alpha_{\mathcal{X}}(\tau) \leq \rho_x^{[(2\tau-1) N_S + 1]\cdot \mathbf{1}_{[\tau > 0]}}.
\end{aligned}
$$

Lemma 1 can now be invoked to prove that $F^{SS}(\theta)$ is a (mean-square) consistent estimate of $F_{\hat{\theta}_{N_S}}(\theta)$. Precisely, for the choice $(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S}) = (X_1^{N_S}, X_2^{N_S}, \ldots, X_{N-N_S}^{N_S})$, we have that

$$
\begin{aligned}
&\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big)\\
&\leq \frac{12}{m^2} \sum_{\tau=-m}^{m} (m - |\tau|) \rho_x^{(|\tau| - N_S + 1)\cdot \mathbf{1}_{[|\tau| \geq N_S]}}\\
&= \frac{12}{m^2} \sum_{\tau=-N_S+1}^{N_S-1} (m - |\tau|) + \frac{24}{m^2} \sum_{\tau=N_S}^{m} (m - \tau) \cdot \rho_x^{\tau - N_S + 1}\\
&= 12 \frac{2(N_S - 1)}{m} - 12 \frac{N_S(N_S-1)}{m^2} +\\
&\quad + \frac{24}{m^2} \sum_{i=1}^{m - N_S + 1} (m - N_S + 1 - i) \cdot \rho_x^i,\\
&\leq 12 \frac{N_S - 1}{m} \cdot \left(2 - \frac{N_S}{m}\right) + \frac{24}{m} \sum_{i=1}^{m - N_S + 1} \rho_x^i,\\
&\leq 12 \frac{N_S - 1}{m} \cdot \left(2 - \frac{N_S}{m}\right) + \frac{24}{m} \cdot \frac{\rho_x}{1 - \rho_x},\\
&= 12 \frac{N_S - 1}{N - N_S} \cdot \left(2 - \frac{N_S}{N - N_S}\right) + \frac{24}{N - N_S} \cdot \frac{\rho_x}{1 - \rho_x}, \quad (5)
\end{aligned}
$$

where the last term follows since $m = N - N_S$ in this case.

Equation (5) provides a *non-asymptotic* bound on the mean square mismatch between the Subsampling reconstructed distribution $F^{SS}(\theta)$ and $F_{\hat{\theta}_{N_S}}(\theta)$ for given N and N_S. The bound holds independently of the underlying data-generating mechanism, apart from the knowledge of ρ_x, a parameter which could be estimated. Besides, (5) implies that, as $N \to \infty$,

$$\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big) \to 0.$$

That is, $F^{SS}(\theta)$ is a mean-square consistent estimator of $F_{\hat{\theta}_{N_S}}(\theta)$ as long as N_S is chosen such that $\frac{N_S}{N} \to 0$ when $N \to \infty$. A typical choice for N_S guaranteeing this latter condition is $N_S = N^p$, where $p \in (0,1)$.

For reference, this result is stated as a theorem.

Theorem 1 *Suppose that sub-sequences are extracted from the available data according to the following scheme:* $(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S}) = (X_1^{N_S}, X_2^{N_S}, \ldots, X_{N-N_S}^{N_S})$. *Then, we have that*

$$
\begin{aligned}
&\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big)\\
&\leq 12 \frac{N_S - 1}{N - N_S} \cdot \left(2 - \frac{N_S}{N - N_S}\right) + \frac{24}{N - N_S} \cdot \frac{\rho_x}{1 - \rho_x}.
\end{aligned}
$$

If moreover N_S is such that $\frac{N_S}{N} \to 0$ as $N \to \infty$, then the reconstructed distribution $F^{SS}(\theta)$ is a mean-square consistent estimator of $F_{\hat{\theta}_{N_S}}(\theta)$.

Expressions like (5) can be similarly derived for all other choices of $(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S})$ given before, and correspondingly quantified theorems hold.

For instance, when

$$(\mathcal{X}_1^{N_S}, \ldots \mathcal{X}_m^{N_S}) = (X_1^{N_S}, X_{N_S+1}^{N_S}, \ldots, X_{\lfloor \frac{N}{N_S} - 1 \rfloor \cdot N_S + 1}^{N_S})$$

we have that:

$$
\begin{aligned}
&\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big)\\
&\leq \frac{12}{m^2} \sum_{\tau=-m}^{m} (m - |\tau|) \rho_x^{[(|\tau| - 1) N_S + 1]\cdot \mathbf{1}_{[|\tau| > 0]}}\\
&= \frac{12}{m} + \frac{24}{m^2} \sum_{\tau=1}^{m} (m - \tau) \cdot \rho_x^{(\tau-1) N_S + 1}\\
&\leq \frac{12}{m} + \frac{24}{m} \rho_x \sum_{i=0}^{m-1} (\rho_x^{N_S})^i,\\
&\leq \frac{12}{\lfloor \frac{N}{N_S} \rfloor} + \frac{24}{\lfloor \frac{N}{N_S} \rfloor} \cdot \frac{\rho_x}{1 - \rho_x^{N_S}}, \quad (6)
\end{aligned}
$$

where the last inequality follows since $m = \lfloor \frac{N}{N_S} \rfloor$ in this case. Similarly to the previous case, (6) gives a non-asymptotic bound for the reconstructed vs. actual distribution mean square error and from (6) it follows that if N_S is chosen such that $\frac{N_S}{N} \to 0$ then

$$\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big) \to_{N \to \infty} 0.$$

In other words, the following theorem holds.

Theorem 2 *Suppose that sub-sequences are extracted from available data according to the following scheme:* $(\mathcal{X}_1^{N_S}, \dots \mathcal{X}_m^{N_S}) = (X_1^{N_S}, X_{N_S+1}^{N_S}, \dots, X_{\lfloor \frac{N}{N_S}-1 \rfloor \cdot N_S+1}^{N_S})$. *Then, we have that*

$$\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big) \leq \frac{12}{\lfloor \frac{N}{N_S} \rfloor} + \frac{24}{\lfloor \frac{N}{N_S} \rfloor} \cdot \frac{\rho_x}{1-\rho_x^{N_S}}.$$

If moreover N_S *is such that* $\frac{N_S}{N} \to 0$ *as* $N \to \infty$, *then the reconstructed distribution* $F^{SS}(\theta)$ *is a mean-square consistent estimator of* $F_{\hat{\theta}_{N_S}}(\theta)$.

Remark 2 *The interpretation of this convergence of empirical distribution function to the actual ensemble value is in terms of the probability space describing the data-generating mechanism presented in (1). This captures the requirement that the data be representative of this underlying distribution.*

Remark 3 *Using either of these theorems and considering both* $N \to \infty$ *and* $N_S \to \infty$ *with* $N_S/N \to 0$ *exponentially, it should be possible using Tchebychev's inequality and the Kronecker Lemma to develop a large-N (i.e. asymptotic) theory in which the Subsampled distribution estimate converges almost surely to the ensemble distribution function. However to do so here, would take us too far afield from our finite-N, single-data-record emphasis.*

4.3 Critique of Subsampling

As previously seen, Subsampling has many appealing features.

- It is easily implementable at a low computational cost.
- The reconstructed distribution $F^{SS}(\theta)$ is a mean square consistent estimate of $F_{\hat{\theta}_{N_S}}(\theta)$.
- More importantly, the quantification of the mean square error $\mathbf{E}\big((F^{SS}(\theta) - F_{\hat{\theta}_{N_S}}(\theta))^2\big)$ is non-asymptotic, and depends only on a parameter, ρ_x, which might be retrieved from basic experiments on the data generating system.

Subsampling, however, has some drawbacks the most central of which is that it reconstructs the distribution of a different parameter. That is, $F_{\hat{\theta}_{N_S}}(\theta)$, the probability distribution of the parameter estimated with N_S data points only, in place of $F_{\hat{\theta}_N}(\theta)$, the distribution of $\hat{\theta}_N$. Clearly, there is a deep kinship between $\hat{\theta}_N$ and $\hat{\theta}_{N_S}$ as well as between $F_{\hat{\theta}_{N_S}}(\theta)$ and $F_{\hat{\theta}_N}(\theta)$, so that estimating the uncertainty of $\hat{\theta}_N$ with that of $\hat{\theta}_{N_S}$ is reasonable. However, the uncertainty of $\hat{\theta}_N$ is less than that of $\hat{\theta}_{N_S}$. In this respect, it is clear that N_S has to be chosen as a trade-off between two opposite effects:

1. too small an N_S means that $F^{SS}(\theta)$ is close to $F_{\hat{\theta}_{N_S}}(\theta)$ but $F_{\hat{\theta}_{N_S}}(\theta) \neq F_{\hat{\theta}_N}(\theta)$;
2. too large an N_S implies that $F_{\hat{\theta}_{N_S}}(\theta) \approx F_{\hat{\theta}_N}(\theta)$ but $F^{SS}(\theta) \neq F_{\hat{\theta}_{N_S}}(\theta)$.

One mechanism advocated to deal with this mismatch between distribution functions is to appeal to the underlying asymptotic theory based on the Central Limit Theorem and to rescale the variance, see [22]. According to asymptotic normality results, whose finite sample validity is in question here, the parameter estimate is asymptotically normally distributed with mean θ_0 and variance dependent on $\sqrt{N}$. For the Subsampled scheme, this variance is replaced by the same term multiplied by $\sqrt{N_S}/\sqrt{N}$, which is necessarily larger. If the estimated distribution, $F_{\hat{\theta}_{N_S}}(\theta)$, is close to gaussian, then such a scaling is possible. However, if it is wildly different from gaussian, then some other ad hoc approach is necessary, such as describing the distribution as a gaussian mixture, say. However, since the underlying problem is to assess the uncertainty in the parameter vector, the uncertainty associated with $F_{\hat{\theta}_{N_S}}(\theta)$ does provide an overbound on the actual uncertainty from $F_{\hat{\theta}_N}(\theta)$, as will be demonstrated in the subsequent Section 6.

5 Analysis of Model-Based Jackknife & Bootstrap

As remarked earlier, given the similarity between Model-Based Jackknife and Model-Based Bootstrapping, we shall concentrate solely on analytical results for the latter.

Differently from Subsampling, the consistency of the Bootstrap procedure has been intensively studied during the last two decades, and many results are available in the literature, [23]. In particular, we have the following result from [2], which mirrors Theorems 1-2 for Subsampling.

Theorem 3 ([2], Theorem 3.9) *Suppose that:*

- *the data are generated by an autoregressive (AR) process* $y_t = \theta_0^T \varphi_t + e_t$, $\varphi_t = [y_{t-1} \ \cdots \ y_{t-n}]^T$, *with roots inside the unit circle,*
- *the AR driving noise process,* e_t, *is independent and identically distributed with zero mean, unit variance, and has bounded* $(2s+1)$*th moment with* $s \geq 3$,
- *the variables* e_1 *and* e_1^2 *satisfy Cramèr's Condition, which is implied by their having probability distributions absolutely continuous with respect to Lebesgue measure.*

Denote the empirical Bootstrapped distribution function based on m resamplings of the N-long data sequence as

$F^{BS}(\hat{\theta}_N^{BS})$ and let the associated probability be $\mathbf{P}^{BS}$. Furthermore let Σ be the covariance of φ_t and let Σ_N^{BS} be the covariance of the Bootstrap version of φ_t (i.e. the regressor obtained from the model $y_t = \hat{\theta}_N^T \varphi_t + \epsilon_t$, with $\hat{\theta}_N$ the parameter estimate from actual data and ϵ_t the Bootstrapped residuals). Provided m is chosen sufficiently large for convergence of the estimate F^{BS}, then almost surely,

$$\begin{aligned} \sup_x \Big| & \mathbf{P}^{BS}\left(N^{1/2}\left[\Sigma_N^{BS}\right]^{1/2}\left(\hat{\theta}_N^{BS} - \hat{\theta}_N\right) \le x\right) \\ & - \mathbf{P}\left(N^{1/2}\Sigma^{1/2}\left(\hat{\theta}_N - \theta_0\right) \le x\right)\Big| \\ & = o\left(N^{-1/2}\right). \end{aligned} \tag{7}$$

The first comment about this result is to remark on its similarity to the earlier theorems on Subsampling. The underlying conditions on the stochastic processes are effectively the same. (The limitation to autoregressive processes is extensible to ARX, ARMA, and ARMAX with a small amount of work.) The quantification is marginally different and the result is almost sure rather than mean-square. Since Bootstrapping permits an extraordinarily large number of resampled data sequences, the limitation on m is not regarded as a problem. There are, however, some implied restrictions compared with Subsampling. The model structure of the true system and that of the Bootstrapping model must be identical, so under-modeling is not permitted in the theory. That being said, [28,7] both use high-order models to Bootstrap new data sequences and then explore the distribution of reduced-order models. Subsampling handles this distribution directly. Secondly, the Bootstrapping results provide no explicit, finite-N quantification of distribution error, and thus are contingent on asymptotic behavior.

In terms of quantifying the uncertainty in the parameter vector, we see that the result measures the Bootstrapped variable's deviation from the estimate $\hat{\theta}_N$ and compares this to the deviation about the true parameter value. Accordingly, there is an implicit requirement for near consistency of the initial parameter estimator before the Bootstrapped distribution estimator can be reliably applied. We shall see this feature demonstrated in the reconsideration of these estimators with the SMS Example next.

6 SMS Example Redux

Both Subsampling and the Jackknife/Bootstrapping have been applied to the SMS Example from Section 2 in order to reconstruct empirically the probability distribution of the identified model parameter $\hat{\theta}_N$, $N = 2000$. In this section, some results which permit better understanding of Subsampling and Bootstrapping estimators' performance are developed.

6.1 Subsampling Estimated Distributions

Figure 7 depicts the probability distribution $F^{SS}(\theta)$ reconstructed via Subsampling by setting $m = 250$, $N_S = 150$, and by choosing subsequences so as to achieve the smallest overlap compatibile with the number of collected data. The actual distribution of $\hat{\theta}_{N_S}$ (i.e. $F_{\hat{\theta}_{N_S}}(\theta)$) as well as that of $\hat{\theta}_N$ (i.e. $F_{\hat{\theta}_N}(\theta)$) are displayed too. The two reference distributions, $F_{\hat{\theta}_{N_S}}(\theta)$ and $F_{\hat{\theta}_N}(\theta)$, have been calculated by Monte-Carlo simulations with $m = 500$.

As is apparent and according to Theorem 1, $F^{SS}(\theta)$ and $F_{\hat{\theta}_{N_S}}(\theta)$ are quite close to each other, showing that Subsampling indeed provides a reliable estimate of the distribution function of $\hat{\theta}_{N_S}$ including capturing the local variances about the two modal points. On the other hand, $F_{\hat{\theta}_{N_S}}(\theta)$ and $F_{\hat{\theta}_N}(\theta)$ differ with the latter being more tightly centered on the modal points than the former. Consequently, the uncertainty reconstructed via Subsampling results as predicted in an oversized empirical local variance. As remarked earlier, one might contemplate rescaling these empirical distribution functions to accommodate this known feature. However, this would require firstly parametrizing the empirical distribution function as, say, a mixture of gaussians. This is a difficult problem to resolve. From our perspective of uncertainty estimation for control though, the central question about the quality of the plant parameter estimate is answered primarily by the detection of the two distinct modes.

If we increase the Subsample size, N_S, from 150 to 500, $F_{\hat{\theta}_{N_S}}(\theta)$ gets closer to $F_{\hat{\theta}_N}(\theta)$ for $N = 2000$; but, the reconstructed distribution $F^{SS}(\theta)$ does not match $F_{\hat{\theta}_{N_S}}(\theta)$ any more because N_S is now too big and the available set of representative Subsampled sequences is too small to achieve an accurate approximation. This is shown in Figure 8.

This behavior is more emphatic if we take $N_S = 1000$. In this case, $F_{\hat{\theta}_{N_S}}(\theta)$ and $F_{\hat{\theta}_N}(\theta)$ are almost identical (as is expected), but the reconstructed distribution $F^{SS}(\theta)$ is quite far away from both. See Figure 9.

6.2 Bootstrap Estimated Distributions

For the Jackknife/Bootstrap, we set $m = 500$ and generated this number of $N = 2000$-long data sets by resampling with replacement the empirical residual distribution generated as per the Jackknife procedure. We used the full order model corresponding to $\hat{\theta}_N$ with the original data as an estimate of the true data-generating system in computing the unresampled original residual sequence. The distribution of residuals was estimated by

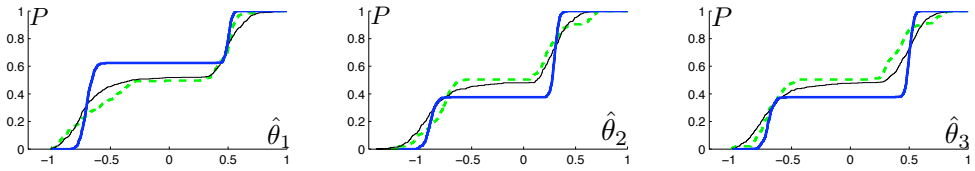

Figure 7. Distribution functions: Subsampling empirical distribution with $N_S = 150$ (dashed green line), the Monte-Carlo distribution of $\hat{\theta}_{N_S}$ (solid thin black line), and the Monte-Carlo distribution $\hat{\theta}_N$ with $N = 2000$ (solid thick blue line).

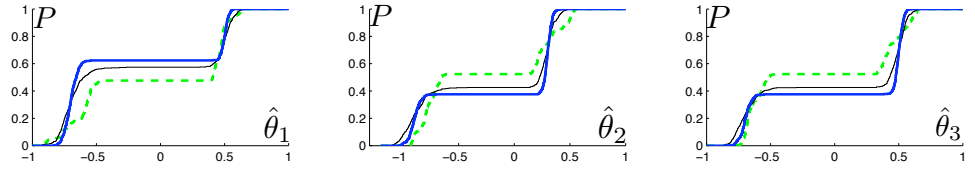

Figure 8. Distribution functions: Subsampling empirical distribution with $N_S = 500$ (dashed green line), the Monte-Carlo distribution of $\hat{\theta}_{N_S}$ (solid thin black line), and the Monte-Carlo distribution of $\hat{\theta}_N$ with $N = 2000$ (solid thick blue line).

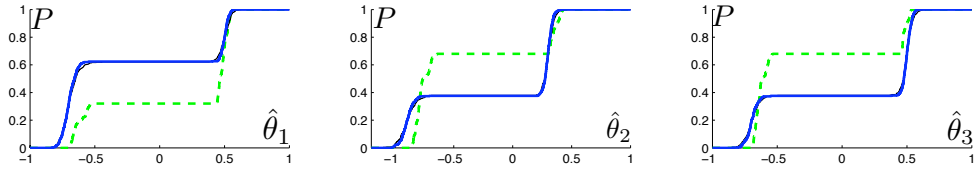

Figure 9. Distribution functions: Subsampling empirical distribution with $N_S = 1000$ (dashed green line), the Monte-Carlo distribution of $\hat{\theta}_{N_S}$ (solid thin black line), and the Monte-Carlo distribution of $\hat{\theta}_N$ with $N = 2000$ (solid thick blue line).

the empirical sum method. The reconstructed distribution $F^{BS}(\theta)$ along with the actual distribution of $\hat{\theta}_N$ is displayed in the next two figures.

We provide two separate plots. The first, in Figure 10 is based on the initial identified parameters being close to the correct value; $\hat{\theta}_N \approx \theta_0 = \begin{bmatrix} -0.7 & 0.3 & 0.5 \end{bmatrix}^T$. Here we see very close agreement between the Bootstrap empirical distribution function and the underlying actual parameter distribution, as determined by Monte-Carlo simulation. This includes the identification of the bi-modal distribution, the relative probabilities of the two modal points, and the local variances.

As remarked in Section 5, the accuracy of $F^{BS}(\theta)$ may be adversely affected by deviation of $\hat{\theta}_N$, the initial identified model parameter vector, from the true value, θ_0. In our particular experiment, poorer results are achieved with the actually identified parameter vector $\hat{\theta}_N = [0.46 \quad -0.84 \quad -0.68]^T$, which is significantly different from θ_0. This is depicted in Figure 11. The Monte-Carlo analysis shows that for this example set-up the likelihood of estimating such a distant parameter vector is 38%. The Bootstrap correctly picks up the bi-modality, but errs in estimating the probabilities and local variances.

This example shows that in situations like the SMS example, where $\hat{\theta}_N$ ends up far away from θ_0 with high probability, the Jackknife/Bootstrap can yield poor estimates for the distribution functions. Such behavior is comparable to that displayed by the asymptotic normal approximation discussed earlier and shown in Figure 2. Although, it must said that the Bootstrap proved quite capable of detecting the bi-modal behavior of the distribution, which is inherently not possible with a normal approximation. Of course, as the number of data, N, increases without bound, the probability of identifying the correct parameter value tends to one and then the asymptotic theory also takes hold in describing the local behavior around this point. However, for small N this need not be the case.

6.3 Comparison with other Bootstrap methods

For the sake of completeness, Moving-Block Bootstrap and Transformation-Based Bootstrap have been tested on the SMS example too. A description of the two approaches is briefly recalled prior to giving simulation results, see [16].

Moving-Block Bootstrap consists of splitting the original N-long 2-dimensional data sequence

$$\begin{Bmatrix} y(1) & y(2) & \cdots & y(N) \\ u(1) & u(2) & \cdots & u(N) \end{Bmatrix}$$

into N/l blocks whose length is equal to l. Then, resam-

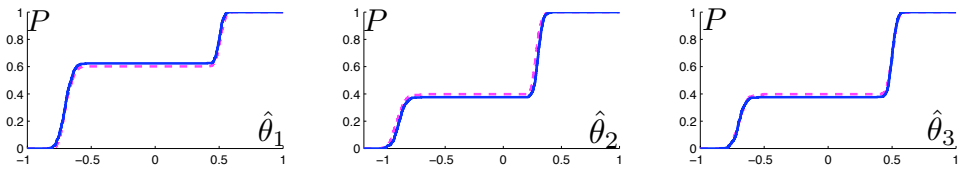

Figure 10. Bootstrap empirical distribution function (dashed magenta line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid blue line) for the case where $\hat{\theta}_N \approx \theta_0$.

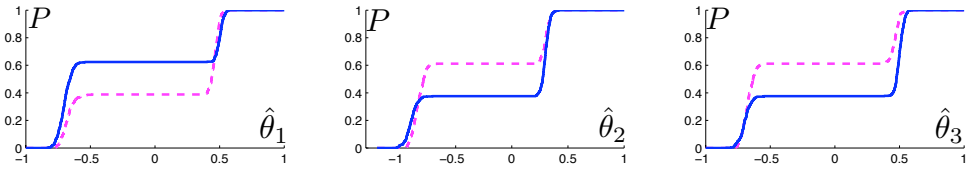

Figure 11. Bootstrap empirical distribution function (dashed magenta line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid blue line) for the case where $\hat{\theta}_N \approx \theta_\star \neq \theta_0$.

pling is performed over blocks, instead of single data points, so as to construct a number of different artificial data sequences, each made of shuffled blocks. This way, the correlation between data points is preserved within each block although discontinuities are created where two consecutive blocks are pasted together.

Figure 12 shows the result obtained for the SMS example with $l = 100$. As it appears, Moving-Block Bootstrap detects bi-modality but errs in estimating local probabilities. Yet, differently from Model-Based Bootstrap, we have noticed that the error in estimating local probabilities does not depend on whether $\widehat{\theta}_N$ is close to the true system parameter vector or not. Other choices of l do not seem to improve the result.

As for Transformation-Based Bootstrap, the idea is to compute the Discrete Fourier Transform of the available data sequence via the FFT algorithm and to divide it by the square root of the spectrum (obtained e.g. by averaging the periodogram of the data sequence over a moving window in the frequency domain). This way, we get a frequency-domain signal whose spectrum is approximately constant, and thus, in time-domain, it corresponds to an independent residual sequence. This latter can be recovered by applying the inverse FFT algorithm. Bootstrap is then performed over the obtained residuals, and bootstrapped data sequences are eventually retrieved by transforming the bootstrapped residual sequences via FFT, multiplying by the square root of the spectrum of the original data sequence, and taking the inverse FFT.

In the SMS example, Transformation-Based Bootstrap has been applied to the 2-dimensional input and output data sequence and the obtained results are depicted in Figure 13. Again, Transformation-Based Bootstrap detects bi-modality but errs in estimating local probabilities. Again, the error does not depend on the closeness of $\hat{\theta}_N$ to the true system parameter vector.

7 Conclusions

In this paper, we considered the problem of reconstructing the probability distribution of the identified model parameter $\hat{\theta}_N$ based on a single finite-length data record. After showing that the heuristic use (with N finite) of the classical asymptotic theory of system identification can be misleading, we introduced procedures based on resampling ideas and discussed their advantages and drawbacks. Theorems were developed on Subsampling and compared to the Bootstrap results. A somewhat pathological example was used as a vehicle for this evaluation.

In particular, in the Subsampling framework non-asymptotic guaranteed results can be given, although the estimated uncertainty tends to be oversized with respect to the actual one. Yet, Subsampling requires minimal assumptions to work properly, and the procedure presented in this paper can be applied verbatim in the presence of under-modeling.

The Jackknife and the Bootstrap may return sharper results than Subsampling, but they require that the true system be consistently estimated, i.e. no under-modeling, because of the need for the reconstruction of the residual sequence. Furthermore, the reconstructed distribution is only asymptotically guaranteed to converge to the actual one, and its quality with finite data depends conditionally on whether the system estimate is close enough to the true system itself.

Although this paper focused on probability distributions only, another important topic in uncertainty evaluation is the estimation of confidence regions for the true system parameter θ_0 (assuming that there is no under-modeling). A promising approach, would be to derive such regions from the reconstructed distribution of $\hat{\theta}_N$.

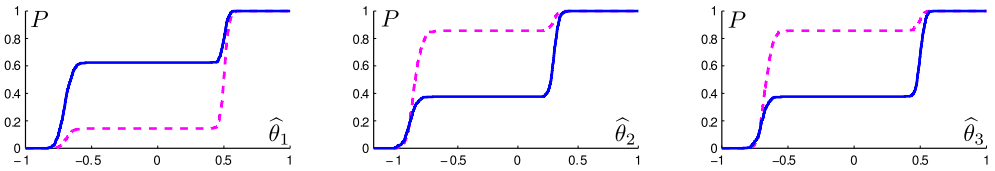

Figure 12. Moving-Block Bootstrap empirical distribution function (dashed magenta line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid blue line).

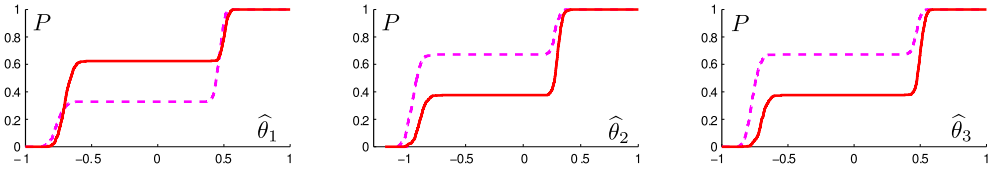

Figure 13. Transformation-Based Bootstrap empirical distribution function (dashed magenta line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid red line).

To this end, the guarantees on the probability distribution reconstructed via Subsampling could be useful to provide *non-asymptotic guaranteed* confidence region for θ_0. See also [4] and [5], where non-asymptotic guaranteed confidence regions for θ_0 are constructed following a different approach.

Acknowledgements

This work was supported by the National Research Council of Italy (CNR), the MIUR national project "Identification and adaptive control of industrial systems", and by the US Air Force Office of Scientific Research under Award No. FA9550-05-1-0401. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of AFOSR.

References

[1] S. Bittanti and M. Lovera. Bootstrap-based estimates of uncertainty in subspace identification methods. *Automatica*, 36:1605–1615, 2000.

[2] A. Bose. Egeworth correction by Bootstrap in autoregressions. *The Annals of Statistics*, 16:1709–1722, 1988.

[3] D. Bosq. *Non parametric statistics for stochastic processes*. Springer, New York, NY, USA, 1998.

[4] M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41:1751–1764, 2005.

[5] M.C. Campi and E. Weyer. Identification with finitely many data points: the lscr approach. In *Proceedings of the IFAC Symposium on System Identification, SYSID 2006, New Castle, Australia (semi-plenary presentation)*, 2006.

[6] Y.A. Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability and its Applications*, 13:691–696, 1968.

[7] W.J. Dunstan and R.R. Bitmead. Empirical estimation of parameter distributions in system identification. In *Proceedings of the 13th IFAC Symposium on System Identification, Rotterdam, The Netherlands*, 2003.

[8] B. Efron. *The Jackknife, the Bootstrap, and other Resampling plans*. SIAM NFS-CBMS, New York, NY, USA, 1982.

[9] B. Efron. Computer-intensive methods in statistical regression. *SIAM Review*, 30:421–449, 1988.

[10] B. Efron and R.J. Tibshirani. *An introduction to the Bootstrap*. Chapman and Hall/CRC, Montpelier, VT, USA, 1993.

[11] S. Garatti, M.C. Campi, and S. Bittanti. Assessing the quality of identified models trough the asymptotic theory - when is the result reliable? *Automatica*, 40:1319–1332, 2004.

[12] S. Garatti, M.C. Campi, and S. Bittanti. The asymptotic model quality assessment for instrumental variable identification revisited. *Systems & Control Letters*, 55:494–500, 2006.

[13] G.C. Goodwin, M. Gevers, and B. Ninness. Identification and robust control: bridging the gap. In *Proceedings of the 7th IEEE Mediterranian conference on control and automation, Haifa, Israel*, 1999.

[14] R.L. Kosut, G.C. Goodwin, and M.P. Polis, editors. *Special issue on system identification for robust control design*, volume 37(7) of *IEEE Transactions on Automatic Control*, 1992.

[15] S.N. Lahiri. *Resampling methods for dependent data*. Springer-Verlag, New York, NY, USA, 2003.

[16] S.N. Lahiri. Bootstrap methods: a review. In J. Fan and H.L. Koul, editors, *Frontiers in statistics*. Imperial College Press, 2006.

[17] L. Ljung. Model validation and model error modeling. In *report from the Åström Symposium on Control, Lund, Sweden*, 1999.

[18] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Upper Saddle River, NJ, USA, 1999.

[19] A. Mokkadem. Mixing properties of arma processes. *Stochastic processes and their applications*, 29:309–315, 1988.

[20] B. Ninness and G.C. Goodwin. Estimation of model quality. *Automatica*, 31:1771–1795, 1995.

[21] D.N. Politis. Computer-intensive methods in statistical analysis. *IEEE Signal Processing Magazine*, 15:39–55, 1998.

[22] D.N. Politis and J.P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22:2031–2050, 1994.

[23] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, NY, USA, 1995.

[24] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[25] F. Tjärnström. Computing uncertainty regions with simultaneous confidence degree using bootstrap. In *Proceedings of the 12th IFAC symposium on System Identification (SYSID), Santa Barbara, California, USA*, 2000.

[26] F. Tjärnström and U. Forssell. Comparison of methods for probabilistic uncertainty bounding. In *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, Arizona, USA*, 1999.

[27] F. Tjärnström and L. Ljung. Estimating the variance in case of undermodeling using bootstrap. In *Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, Arizona, USA*, 1999.

[28] F. Tjärnström and L. Ljung. Using the bootstrap to estimate the variance in the case of undermodeling. *IEEE Transaction on Automatic Control*, 47:395–398, 2002.

[29] A. Zoubir and B. Boashash. The Bootstrap and its application in Signal Processing. *IEEE Signal Processing Magazine*, 15:56–76, 1998.