On Resampling and Uncertainty Estimation in Linear System Identification ⋆

Simone Garatti * **Robert R. Bitmead** **

* *Dipartimento di Elettronica ed Informazione, Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italy. E-mail:* sgaratti@elet.polimi.it.
** *Department of Mechanical & Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0411, USA. E-mail:* rbitmead@ucsd.edu.

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 sub-sampling 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 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.

Keywords: Uncertainty evaluation, Resampling techniques, System Identification.

1. INTRODUCTION

Robust model-based control requires quantification of plant model uncertainty, Ninness and Goodwin [1995]. 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 (Ljung [1999], Söderström and Stoica [1989]). Recent studies (in under-excited systems, Garatti et al. [2004, 2006]) have shown that cases can be found when the cost function is non-convex and can have separated local minima. In these 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 and Sub-sampling, Politis [1998], Zoubir and Boashash [1998]. 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 $\mathscr{Z}^N = \{x_i = [u_i \quad y_i]^T, \quad i = 1,\dots,N\}$, where $u_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$.
- These data are stationary and generated by a stable ARMA process
$$A(z)\begin{bmatrix} u_t \\ y_t \end{bmatrix} = B(z)\eta_t, \tag{1}$$
where η_t are i.i.d. The 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 $V(\theta, \mathscr{Z}^N) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{i|i-1}(\theta))^2$, where $\hat{y}_{i|i-1}(\theta)$ is the optimal predictor based on the model corresponding to θ. If the data set which the cost function refers to is clear from the context, we shall write $V_N(\theta)$ in place of $V(\theta, \mathscr{Z}^N)$.
The minimizer of $V(\theta, \mathscr{Z}^N)$ (assuming it is unique) is indicated by $\hat{\theta}_N$. Our goal is to reconstruct its probability distribution, hereafter indicated by $F_{\hat{\theta}_N}(\theta)$.

1.1 Structure of the paper

First, an example showing the limitations of the asymptotic theory of system identification is presented in Section 2. Then, some resampling strategies (namely; Mont-Carlo, sub-sampling, and Bootstrapping) are briefly recalled in Section 3, with particular emphasis to their application to 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 failed.

⋆ This work was supported by the National Research Council of Italy (CNR), by the MIUR national project "Identification and adaptive control of industrial systems", and by 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.

2. ASYMPTOTIC THEORY AND ITS LIMITATIONS – THE SMS EXAMPLE

The following example is taken from Garatti et al. [2004], 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 Ljung [1999], Söderström and Stoica [1989], 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, i.e. $\hat{\theta}_N = \arg\min V_N(\theta)$, where $\theta = [a\ b\ h]^T$.

According to the asymptotic ($N \to \infty$) theory of system identification, the empirical estimation of the probability distribution of $\hat{\theta}_N$, $\sqrt{N}(\hat{\theta}_N - \theta_0)$, 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 $\psi_t(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\hat{y}_{t|t-1}(\theta)$. 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 $\frac{1}{N}P_\theta$. Such 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 1, where

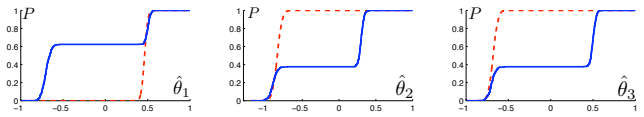

Figure 1. The actual probability distributions of each component of $\hat{\theta}_N$ (solid line) vs. the distributions returned by the asymptotic theory of system identification (dashed line).

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. 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 failure of the asymptotic theory. As shown in Garatti et al. [2004], in the SMS example, the cost $\overline{V}(\theta) \triangleq \mathbb{E}[(y_k - \hat{y}_{k|k-1}(\theta))^2]$ 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 Garatti et al. [2004]. 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 1.

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

In order to provide a fair evaluation of uncertainty, some different tools have to be considered, and in this paper the focus is on resampling strategies, Efron [1982, 1988], Efron and Tibshirani [1993], Politis [1998], Shao and Tu [1995]. Resampling methods have recently attracted the attention of the systems and control community, Bittanti and Lovera [2000], Dunstan and Bitmead [2003], Tjärnström and Forssell [1999], Tjärnström and Ljung [2002]. 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 difficult as the SMS example.

Three different resampling methodologies for the reconstruction of the underlying probability distribution of the identified parameter $\hat{\theta}_N$ are considered: namely, Monte-Carlo, Subsampling, and 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 this latter is treated for it is the framework of system identification problems.

3.1 The Monte-Carlo method

The Monte-Carlo procedure amounts to repeating the identification experiment m times so as to collect m independent N-long data sequences $(\mathscr{Z}_1^N, \ldots, \mathscr{Z}_m^N)$, which in turn, by minimizing $V(\theta, \mathscr{Z}_i^N)$, $i = 1, \ldots, m$, yields 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

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

where the vector inequality $\hat{\theta}_N^i \leq \theta$ is taken componentwise and $\mathbf{1}_{[\cdot]}$ is the indicator function.

It is well known that $F^{MC}(\theta)$ is an unbiased and consistent (both mean square and almost surely) estimator of the actual probability distribution $F_{\hat{\theta}_N}(\theta)$ (see Shao and Tu [1995]). 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 others resampling methodologies.

3.2 The Sub-sampling method

Sub-sampling has been first introduced in Politis and Romano [1994] and, despite its attractive properties, has not received much attention from the systems and control community yet except for Dunstan and Bitmead [2003]. Sub-sampling is quite intuitive and is reminiscent of the Monte-Carlo approach. A single data set, however, is used. Precisely, let $m \leq N$ and $N_S \leq N-m+1$. From the N-long available data set $\mathscr{X}^N$, the set of all N_S-long sub-sequences of consecutive data points is considered and, among these, m are extracted, that is:

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

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

Starting from the chosen sub-sequences $(\mathscr{X}_1^{N_S}, \ldots \mathscr{X}_m^{N_S})$, 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 $V(\theta, \mathscr{X}_i^{N_S})$, $i = 1, \ldots, m$ 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 main point in sub-sampling is that since the $\mathscr{X}_i^{N_S}$'s are taken from the actual data set $\mathscr{X}^N$, they are distributed identically to the original data, although their length is reduced. The choice of N_S is a degree of freedom of sub-sampling, and a sensible tuning of N_S is of paramount importance. Also, the choice of the $\mathscr{X}_i^{N_S}$'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 sub-sampling is provided. It is worth noticing that sub-sampling does not correspond to performing m Monte-Carlo N_S-long simulations, since sub-sequences $\mathscr{X}_i^{N_S}$ are correlated in general (actually, they can even be overlapping). Showing that $F^{SS}(\theta)$ is unbiased and consistent is not straightforward.

3.3 The Bootstrap method

Given the data sequence $\mathscr{X}^N$, 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 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 Dunstan and Bitmead [2003], Tjärnström and Ljung [2002]. Given the model estimate $\{\hat{A}(z), \hat{B}(z)\}$, the one-step prediction residuals $(\varepsilon_1, \ldots, \varepsilon_N)$ are computed according to the following equation:

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

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 $B(z)$.

From $(\varepsilon_1, \ldots, \varepsilon_N)$, the complete distribution function, $\widehat{F}_{\varepsilon_t}(\varepsilon)$, of the residual error ε_t is reconstructed (for this purpose one can use a number of techniques such as empirical sum, L^1 approximation, kernel methods, etc.); then, a new, bootstrapped, residual sequence is generated by extracting N samples $(\varepsilon_1^B, \ldots, \varepsilon_N^B)$ according to $\widehat{F}_{\varepsilon_t}(\varepsilon)$. This new bootstrapped residual sequence $(\varepsilon_1^B, \ldots, \varepsilon_N^B)$ can be thought of as new extracted samples from the noise process η_t and it can be used for computing a new bootstrapped N-long input/output data sequence $(u_1^B, y_1^B, u_2^B, y_2^B, \ldots, u_N^B, y_N^B)$ according to the following mechanism:

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

This bootstrapped 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 bootstrapping 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 $\hat{\theta}_N$:

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

4. ANALYSIS OF THE SUB-SAMPLING METHOD

In this section, we establish our main theoretical result, that $F^{SS}(\theta)$ is a consistent estimate of $F_{\hat{\theta}_{N_S}}(\theta)$, and provide some considerations about sub-sampling.

4.1 Preliminary definitions and results

We need the following preliminary definition, see e.g. Bosq [1998].

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

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

If $\alpha_Y(\tau) \to_{\tau \to \infty} 0$, 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 Politis and Romano [1994].

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, more-

over, it is stationary since Y_t is. Letting $\gamma_Z(\tau) = \mathbf{E}(Z_t Z_{t+\tau}) = \gamma_Z(-\tau)$ be the covariance function of Z_t, we have that

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

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 Davydov [1968].) leading to the sought 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 Sub-sampling strategies and mixing conditions

We want now to apply Lemma 1 to the sub-sampling 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, $\mathscr{X}_i^{N_S}$ plays the role of the process Y_t in Lemma 1 while $\hat{\theta}^i_N$, the parameter vector estimated from the subsequence $\mathscr{X}_i^{N_S}$, that of $\varphi(Y_t)$. Clearly, $\hat{\theta}^i_N$ is a measurable function of $\mathscr{X}_i^{N_S}$. As for the process $\mathscr{X}_i^{N_S}$ we need to check whether it is:

1. stationary;
2. α-mixing.

As for Point 1, recall that

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

where $X_j^{N_S} = [x_{j+1}^T \quad x_{j+2}^T \cdots x_{j+N_S}^T]^T$; $x_t = [u_t \;\; y_t]^T$, 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 $\mathscr{X}_i^{N_S}$ is stationary as long as the $\mathscr{X}_i^{N_S}$s are chosen from the $X_j^{N_S}$s in a equally time-spaced manner. That is, if $\mathscr{X}_i^{N_S} = X_{j_1}^{N_S}$ and $\mathscr{X}_{i+1}^{N_S} = X_{j_2}^{N_S}$, then the difference $j2-j1$ must be the same whatever i is. Some possible choices ensuring stationarity are the following (here $\lfloor\cdot\rfloor$ denotes the integer part and $k \leq N_S$):

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

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, Bosq [1998] and Mokkadem [1988]. 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 $\rho_x \in (0,1)$. Perhaps, 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, $\mathscr{X}_t^{N_S}$ are α-mixing too. The α mixing coefficient of $\mathscr{X}_t^{N_S}$ (say $\alpha_{\mathscr{X}}(\tau)$), however, depends on how subsequences $(\mathscr{X}_1^{N_S},\ldots,\mathscr{X}_m^{N_S})$ are chosen from $(X_1^{N_S},X_2^{N_S},\ldots,X_{N-N_S}^{N_S})$.
Lemma 1 can now be invoked in order to prove that $F^{SS}(\theta)$ is a (mean-square) consistent estimate of $F_{\hat{\theta}_{N_S}}(\theta)$. Precisely, for the choice $(\mathscr{X}_1^{N_S},\ldots\mathscr{X}_m^{N_S}) = (X_1^{N_S},X_2^{N_S},\ldots,X_{N-N_S}^{N_S})$, we have that

$$\begin{aligned}&\mathbf{E}\big((F^{SS}(\theta)-F_{\hat{\theta}_{N_S}}(\theta))^2\big)\\ &\leq \frac{12}{m^2}\sum_{\tau=-m}^{m}(m-|\tau|)\rho_x^{(|\tau|-N_S+1)\cdot\mathbf{1}_{[|\tau|\geq N_S]}}\\ &= \frac{12}{m^2}\sum_{\tau=-N_S+1}^{N_S-1}(m-|\tau|)+\frac{24}{m^2}\sum_{\tau=N_S}^{m}(m-\tau)\cdot\rho_x^{\tau-N_S+1}\\ &= 12\frac{N_S-1}{m}\cdot\left(2-\frac{N_S}{m}\right)+\frac{24}{m^2}\sum_{i=1}^{m-N_S+1}(m-N_S+1-i)\cdot\rho_x^i,\\ &= 12\frac{N_S-1}{N-N_S}\cdot\left(2-\frac{N_S}{N-N_S}\right)+\frac{24}{N-N_S}\cdot\frac{\rho_x}{1-\rho_x},\end{aligned} \tag{3}$$

Equation (3) provide a *non-asymptotic* bound of the mean square mismatch between the sub-sampling 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, (3) implies that

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

(i.e. $F^{SS}(\theta)$ 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 $p \in (0,1)$.
Expressions like (3) can be similarly derived for all other choices of $(\mathscr{X}_1^{N_S},\ldots\mathscr{X}_m^{N_S})$ given before, and correspondingly theorems hold. 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:

$$(\mathscr{X}_1^{N_S},\ldots\mathscr{X}_m^{N_S}) = (X_1^{N_S},X_2^{N_S},\ldots,X_{N-N_S}^{N_S}).$$

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 the subsequences are extracted according to:

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

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 $F_{\hat{\theta}_{N_S}}(\theta)$.

4.3 Critique of sub-sampling

As previously seen, sub-sampling 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.

Sub-sampling, however, has some drawbacks the most central of which is that it reconstructs the distribution of a different parameter: $F_{\hat{\theta}_{N_S}}(\theta)$, 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 $F^{SS}(\theta)\neq F_{\hat{\theta}_{N_S}}(\theta)$.

5. ANALYSIS OF THE BOOTSTRAP

We have the following result from Bose [1988], which mirrors Theorem 1 for sub-sampling.

Theorem 2. (Bose [1988], Theorem 3.9). Suppose that:

- the data is 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 $y_t=\hat{\theta}_N^T\varphi_t+\varepsilon_t$, being $\hat{\theta}_N$ 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,

$$\sup_x\left|\mathbf{P}^{BS}\left(N^{1/2}\left[\Sigma_N^{BS}\right]^{1/2}\left(\hat{\theta}_N^{BS}-\hat{\theta}_N\right)\leq x\right)-\mathbf{P}\left(N^{1/2}\Sigma^{1/2}\left(\hat{\theta}_N-\theta_0\right)\leq x\right)\right| = o\left(N^{-1/2}\right). \tag{4}$$

The first comment about this result is to remark on its similarity to the earlier theorems on sub-sampling. 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.

In terms of quantifying the uncertainty in the parameter vector, we see that the result measures the Bootstrapped variable's deviation from the sample mean 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 sub-sampling and the Bootstrap have been applied to the SMS Example from Section 2 in order to reconstruct the probability distribution of the identified model parameter $\hat{\theta}_N$, $N=2000$. In this section, some results which permit better understanding of sub-sampling and Bootstrapping estimators' performance are developed.

6.1 Sub-Sampling Estimated Distributions

Figure 2 depicts the probability distribution $F^{SS}(\theta)$ reconstructed via sub-sampling 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

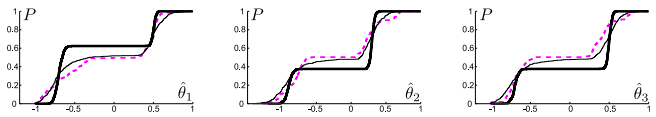

Figure 2. Distribution functions: Sub-sampling empirical distribution with $N_S=150$ (dashed line), the Monte-Carlo distribution of $\hat{\theta}_{N_S}$ (solid thin line), and the Monte-Carlo distribution $\hat{\theta}_N$ with $N=2000$ (solid thick line).

$F_{\hat{\theta}_N}(\theta)$, 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 sub-sampling 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 sub-sampling results as predicted in an oversized empirical variance. 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 sub-sample 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 sub-sampled sequences is too small to achieve an accurate approximation. This is shown in Figure 3.

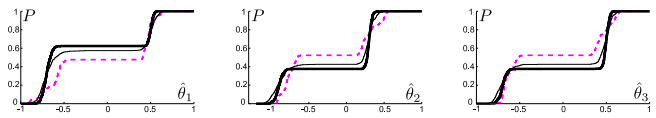

Figure 3. Distribution functions: Sub-sampling empirical distribution with $N_S = 500$ (dashed line), the Monte-Carlo distribution of $\hat{\theta}_{N_S}$ (solid thin line), and the Monte-Carlo distribution of $\hat{\theta}_N$ with $N = 2000$ (solid thick line).

6.2 Bootstrap Estimated Distributions

For the Bootstrap, we set $m = 500$ and generated this many $N = 2000$-long data sets by reconstructing the residual distribution with an empirical sum. 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 un-resampled original residual sequence. The distribution of residuals was estimated by 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 4 is based on the initial identified parameters being close to the correct value; $\hat{\theta}_N \approx \hat{\theta}_0 = [-0.7\ 0.3\ 0.5]^T$. 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.

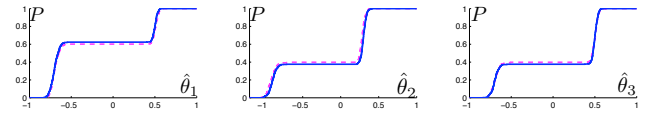

Figure 4. Bootstrap empirical distribution function (dashed line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid line) for the case where $\hat{\theta}_N \approx \theta_0$.

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 $\hat{\theta}_N = [0.46\ -0.84\ -0.68]^T$, which is significantly different from θ_0. This is depicted in Figure 5. 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.

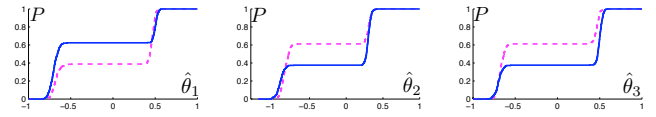

Figure 5. Bootstrap empirical distribution function (dashed line) and the Monte-Carlo distribution function of $\hat{\theta}_N$ (solid line) for the case where $\hat{\theta}_N \not\approx \theta_0$.

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 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 sub-sampling and compared to the Bootstrap results. A somewhat pathological example was used as a vehicle for this evaluation.

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

REFERENCES

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

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

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

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

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.

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

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

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

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.

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.

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

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

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

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

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.

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

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

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.

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.

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