Proceedings of the 38th
Conference on Decision & Control
Phoenix, Arizona USA • December 1999

Finite Sample Properties of System Identification Methods

Erik Weyer
Department of Electrical and Electronic Engineering, University of Melbourne
Parkville VIC 3052, Australia, e.weyer@ee.mu.oz.au

M.C. Campi
Department of Electrical Engineering for Automation, University of Brescia
Via Branze 38, 25123 Brescia, Italy, campi@bsing.ing.unibs.it

Abstract

In this paper we study the performance of system identification methods on a *finite* data sample. Our results are of the following form: with a probability not less than $1-\delta$, minimising the empirical identification cost leads to an estimate which is within an accuracy ϵ from the theoretical optimal estimate. Explicit expressions for the accuracy ϵ are derived, revealing its dependence on the data generation characteristics and the choices made in the system identification procedure. This is believed to be the first contribution delivering a finite sample identification theory applicable to a general linear time-invariant setting.

Keywords: System identification, finite sample properties

1 Introduction

In this paper we study the properties of quadratic identification methods applied to a linear model class for a *finite* data sample. Asymptotic properties of these methods have been extensively studied over the last three decades (see e.g. Ljung (1987)) and are now well understood. However, almost nothing is known regarding the finite sample properties.

Our main result (Section 2) quantitatively assesses the discrepancy between minimising a theoretical identification cost and minimising its empirical counterpart. It is shown that a number of factors affect the result, including the size of the adopted model class and the upper bound on the absolute value of the singularities of the model and the data generation mechanism. All these dependencies have meaningful interpretations.

The approach we take is to cover the parameter space by a ρ-net and to use generalised exponential inequalities together with continuity results. This is a general approach, and we believe it can be extended to other identification criteria and nonlinear model structures. This is an area of current research.

We refer the reader to Campi and Weyer (1999) for the full proofs.

1.1 The data generation mechanism

We assume that the observed data are generated by a linear system

$$y(t) = G_0 u(t) + H_0 e(t)$$

where the input signal $u(t)$ is deterministic and bounded by $|u(t)| \leq U$ and $e(t)$ is a sequence of independent Gaussian random variables with zero mean and variance σ_e^2. G_0 and H_0 are transfer functions in the backward shift operator q^{-1}, i.e. $q^{-1}y(t) = y(t-1)$; however, for the sake of readability, we omit throughout to explicitly indicate the dependence on q^{-1}. Moreover, G_0 and H_0 can be written as

$$G_0 = \frac{B_0}{A_0}, \quad H_0 = \frac{C_0}{D_0}$$

where

$$\begin{aligned} A_0 &= 1 + a_{01}q^{-1} + \cdots + a_{0n_0}q^{-n_0} \\ B_0 &= b_{01}q^{-1} + \cdots + b_{0n_0}q^{-n_0} \\ C_0 &= 1 + c_{01}q^{-1} + \cdots + c_{0n_0}q^{-n_0} \\ D_0 &= 1 + d_{01}q^{-1} + \cdots + d_{0n_0}q^{-n_0} \end{aligned}$$

0-7803-5250-5/99/$10.00 © 1999 IEEE

and n_0 is an upper bound on the degrees. Moreover, we assume that the zeros of A_0, C_0 and D_0 are inside a circle of a known radius $\eta < 1$. The zeros of B_0 is assumed to be inside a circle of known radius μ, where μ might be larger than 1. Finally we assume that $|b_{01}|$ is bounded by a known constant B. For simplicity we assume that $B \leq \mu$.

1.2 Model class

The model class considered is

$$y(t) = G(\theta)u(t) + H(\theta)w(t) \qquad (1)$$

where $w(t)$ is a sequence of independent Gaussian random variables with zero mean and

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

For convenience we parameterise the polynomials in terms of their zeros (and the leading coefficient of $B(\theta)$), i.e.

$$A(\theta) = \prod_{i=1}^{n_1}(1 - a_i q^{-1}) \qquad (2)$$

$$B(\theta) = b_1 q^{-1} \prod_{i=2}^{n_1}(1 - b_i q^{-1}) \qquad (3)$$

$$C(\theta) = \prod_{i=1}^{n_1}(1 - c_i q^{-1}) \qquad (4)$$

$$D(\theta) = \prod_{i=1}^{n_1}(1 - d_i q^{-1}) \qquad (5)$$

We let $\theta \in \mathbb{C}^{4n_1}$ contain the unknown parameters of the polynomials.

We impose the requirement that θ belongs to a compact set Θ such that the polynomials have real coefficients, and further that $|a_i| \leq \eta, i = 1, \cdots, n_1$, $|b_1| \leq B$, $|b_i| \leq \mu, i = 2, \ldots, n_1$, $|c_i| \leq \eta, i = 1, \cdots, n_1$, $|d_i| \leq \eta, i = 1, \cdots, n_1$. These requirements are all in agreement with our prior information about the data generating system.

1.3 The identification criterion

From a system identification perspective, the most important feature of the above models is their associated predictors which are given by

$$\hat{y}(t, \theta) = (1 - H^{-1}(\theta))y(t) + H^{-1}(\theta)G(\theta)u(t)$$

and the corresponding prediction errors are

$$\begin{aligned} \varepsilon(t, \theta) &= y(t) - \hat{y}(t, \theta) \\ &= H^{-1}(\theta)y(t) - H^{-1}(\theta)G(\theta)u(t) \end{aligned} \qquad (6)$$

Ideally, one would like to choose θ such that the following theoretical identification cost

$$\bar{V}_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} E\varepsilon^2(t, \theta) \qquad (7)$$

is minimised, where N is the number of data points.

As the data generation mechanism is unknown, one cannot compute the expected value in (7) and, in its place, the sample version

$$V_N(\theta) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t, \theta) \qquad (8)$$

is used. Clearly, the minimisation of $V_N(\theta)$ may be expected to be equivalent to the minimisation of $\bar{V}_N(\theta)$ only when $N \to \infty$ (and, this is indeed the case under mild assumptions, see e.g. Ljung (1987)). On the other hand, in real applications the number of data points is finite ($N < \infty$) and, therefore, one question that arises naturally is to quantify the deterioration in the model quality caused by a finite data sample. In order to answer this question, quantitative bounds on

$$\bar{V}_N(\hat{\theta}_N) - \bar{V}_N(\bar{\theta}_N) \qquad (9)$$

is required where

$$\hat{\theta}_N := \arg\min_{\theta \in \Theta} V_N(\theta)$$

and

$$\bar{\theta}_N := \arg\min_{\theta \in \Theta} \bar{V}_N(\theta)$$

2 The main result

The bound for (9) is given in Theorem 2.2 below. The proof of the bound is immediate from the more general - and perhaps of independent interest - result presented next

Theorem 2.1 *With probability at least* $1 - \delta$, *we have*

$$\sup_{\theta \in \Theta} |V_N(\theta) - \bar{V}_N(\theta)| < \epsilon(\delta, \eta, n_0, n_1, N)$$

where $\epsilon(\delta, \eta, n_0, n_1, N)$ *is such that*

$$\lim_{\delta \to 0} \epsilon(\delta, \eta, n_0, n_1, N) = \infty$$

$$\lim_{\eta \to 1} \epsilon(\delta, \eta, n_0, n_1, N) = \infty$$

$$\lim_{n_0 \to \infty} \epsilon(\delta, \eta, n_0, n_1, N) = \infty$$

$$\lim_{n_1 \to \infty} \epsilon(\delta, \eta, n_0, n_1, N) = \infty$$

$$\lim_{N \to \infty} \epsilon(\delta, \eta, n_0, n_1, N) = 0$$

The actual expression for $\epsilon(\delta, \eta, n_0, n_1, N)$ can be obtained by the derivation of the result provided in Section 3.

From Lemma 3.3 to 3.6 below it follows that the actual expressions are $\epsilon = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$ and $\delta = \kappa_1 + \kappa_2 + \kappa_4$ where γ_i and κ_i are given in the lemmas. Although these expressions are complicated, they can be explicitly evaluated for any finite number of data points, and they have a natural functional dependence on the important variables as shown in the above theorem.

Theorem 2.2 *With probability at least $1 - \delta$, we have*

$$\bar{V}_N(\hat{\theta}_N) - \bar{V}_N(\bar{\theta}_N) < 2\epsilon(\delta, \eta, n_0, n_1, N)$$

Proof: Apply Theorem 2.1 twice. ∎

It is important to stress that the result holds true for any *finite* data sample of size N, that is, it is not asymptotic in N.

The form of the result deserves to be further commented upon. With a certain confidence $1 - \delta$, minimising the empirical cost (8) corresponds to minimising the theoretical cost (7) to within an accuracy 2ϵ. The presence of a confidence coefficient δ is natural and stems from the stochastic nature of the problem. When the data sample is finite there is always a nonzero (even though, possibly small) probability that the noise plays against the identification objective, resulting in a deterioration of the accuracy of the estimate. Not surprisingly, ϵ increases as δ decreases and tends to infinity when $\delta \to 0$. This is in agreement with the fact that no level of accuracy can be guaranteed with confidence 1 for a finite data sample. This is in contrast to the asymptotic theory, where the assumption $N \to \infty$ leads to results valid with probability 1.

Besides δ, ϵ depends on η, n_0, n_1, and N. We have that $\epsilon \to \infty$ both when the system and/or model complexity (as measured by the parameters n_0 and n_1) tends to infinity and when $\eta \to 1$. This behavior can be easily understood. Increasing n_0 and/or n_1, or sending η to 1, leads to a prediction error process (6) with a long correlation tail. When this happens, the averaging effect on the noise is decreased and a larger number of data points is necessary to guarantee a certain accuracy level. Finally, $\epsilon \to 0$ when $N \to \infty$, as is expected from averaging effects.

Identification and estimation methods have been analysed using techniques similar to those employed in this paper in e.g. Campi and Kumar (1998) and Modha and Masry (1998). However, the finite sample properties were not explicitly considered in those papers. Finite sample properties has been considered in the present setting in Weyer et al (1999) under more restrictive conditions.

3 Derivation of the main result

In this section, we sketch the proof of the fundamental Theorem 2.1.

Lemma 3.1 *Under the assumptions in section 1.2, the set $\Theta \subset \mathbb{C}^{4n_1}$ can be covered by*

$$M(\rho) = \left(\left[\left(\frac{2\eta}{\sqrt{2}\rho} + 1\right)\left(\frac{\eta}{\sqrt{2}\rho} + 1\right)\right]\right)^{\frac{3n_1}{2}} \cdot \left(\left[\left(\frac{2\mu}{\sqrt{2}\rho} + 1\right)\left(\frac{\mu}{\sqrt{2}\rho} + 1\right)\right]\right)^{\frac{n_1 - 1}{2}} \cdot ([B/\rho] + 1)$$

($[x]$ stands for the integer part of x) balls of radius ρ (in the infinity norm).

Proof: Follows from straightforward calculations. ∎

Denote the balls in Lemma 3.1 by $\bar{B}_i$, $i = 1, \ldots, M(\rho)$ and their centres by θ_i, and let $B_i = \bar{B}_i \cap \Theta$. Without loss of generality we can assume that $\theta_i \in B_i$. Let $\bar{H}_\theta = H^{-1}(\theta)H_0$. We have the following intermediate result (throughout, $\max_i$ stands for $\max_{i \in [1, M(\rho)]}$)

Lemma 3.2

$$\sup_{\theta \in \Theta} |V_N(\theta) - \bar{V}_N(\theta)| \leq$$

$$\max_i \sup_{\theta \in B_i} \left| \frac{1}{N} \sum_{t=1}^{N} (\tilde{H}_\theta e(t))^2 - (\tilde{H}_{\theta_i} e(t))^2 \right| + \tag{10}$$

$$\max_i \left| \frac{1}{N} \sum_{t=1}^{N} (\tilde{H}_{\theta_i} e(t))^2 - E(\tilde{H}_{\theta_i} e(t))^2 \right| + \tag{11}$$

$$\max_i \sup_{\theta \in B_i} \left| \frac{1}{N} \sum_{t=1}^{N} E(\tilde{H}_{\theta_i} e(t))^2 - E(\tilde{H}_\theta e(t))^2 \right| + \tag{12}$$

$$\max_i \sup_{\theta \in B_i} \left| \frac{2}{N} \sum_{t=1}^{N} H^{-1}(\theta)(G_0 - G(\theta))u(t) \cdot H^{-1}(\theta) H_0 e(t) \right| \tag{13}$$

Proof: Follows from straightforward manipulations. ∎

Roughly, the main idea is that we can bound the discrepancy between the theoretical identification criterion and its sample version for θ_i, $i = 1, \ldots, M(\rho)$, and then extend the bound to Θ, since for every $\theta \in \Theta$ there is a θ_i such that $||\theta - \theta_i|| \leq \rho$, and there should only be small variations in the theoretical criterion and its sample version when the parameters are close.

Intuitively, (10) and (12) should be small when $||\theta - \theta_i|| \leq \rho$ is small and the difference (11) between the expected and empirical value should be small when the number of data points is large. Notice that the $\max_i$ in (11) is over a finite number of parameters. Moreover, (13) is a sum of zero mean Gaussian random variables, and is therefore also expected to be small.

As Lemma 3.3 to 3.6 below show this is indeed the case, and Theorem 2.1 follows by combining these lemmas, which separately bound the four terms (10) to (13).

Lemma 3.3 *For any $\epsilon_1 > 0$ and $\epsilon_2 > 0$,*

$$\max_i \sup_{\theta \in B_i} \left| \frac{1}{N} \sum_{t=1}^{N} (\tilde{H}_\theta e(t))^2 - (\tilde{H}_{\theta_i} e(t))^2 \right| \leq \gamma_1$$

with probability at least $1 - \kappa_1$ where

$$\gamma_1 = 2 \frac{2^{2n_0+3n_1+1}\rho}{(1-\eta)^{2n_0+2n_1}} \left(\frac{4\epsilon_2 \sqrt{2\sigma_e^2/(N\pi)}}{(1-\eta)^{n_0+n_1+1}} + \frac{8\sigma_e^2}{N(1-\eta)^{2(n_0+n_1+1)}\pi} + \epsilon_1 + \sigma_e^2 + \epsilon_2^2 \right) \tag{14}$$

$$\kappa_1 = 4e^{\frac{-\epsilon_2^2 N}{\frac{(2n_0+2n_1)!8(1-2/\pi)\sigma_e^2}{((n_0+n_1)!)^2(1-\eta^2)^{2n_0+2n_1+1}} + c_{F_1}\sqrt{2N}\sigma_e\epsilon_2}} + 4e^{\frac{-\epsilon_1^2 N^2}{8\sigma_e^4 N + 4\sigma_e^4 \epsilon_1}} \tag{15}$$

$$c_{F_1} = \max_{\tau=1,2,\ldots} (\tau+1)\cdots(\tau+n_1)\eta^\tau \tag{16}$$

Sketch of Proof. Let $F_1(\theta, \theta_i) := (H^{-1}(\theta) - H^{-1}(\theta_i))H_0$, $F_2(\theta, \theta_i) := (H^{-1}(\theta) + H^{-1}(\theta_i))H_0$, $v_1(t) := F_1(\theta, \theta_i)e(t)$ and $v_2(t) := F_2(\theta, \theta_i)e(t)$. We have that

$$\left| \frac{1}{N} \sum_{t=1}^{N} (\tilde{H}_\theta e(t))^2 - (\tilde{H}_{\theta_i} e(t))^2 \right| \leq \sqrt{\frac{1}{N} \sum_{t=1}^{N} v_1^2(t)} \sqrt{\frac{1}{N} \sum_{t=1}^{N} v_2^2(t)} \tag{17}$$

Furthermore let

$$V_{1,N}(\omega) := \frac{1}{\sqrt{N}} \sum_{t=1}^{N} v_1(t) e^{-i\omega t}$$

From Parseval's equality and Theorem 2.1 in Ljung (1987) we have

$$\frac{1}{N}\sum_{t=1}^{N} v_1^2(t) = \frac{1}{N}\sum_{k=1}^{N} \left| V_{1,N}\left(\frac{2\pi k}{N}\right) \right|^2 \leq \frac{2}{N}\sum_{k=1}^{N} \left| F_1(\theta, \theta_i, e^{\frac{i2\pi k}{N}}) E_N\left(\frac{2\pi k}{N}\right) \right|^2 + \frac{2}{N}\sum_{k=1}^{N} \left| R_N\left(\frac{2\pi k}{N}\right) \right|^2$$

where $R_N(\omega) = V_{1,N}(\omega) - F_1(\theta, \theta_i, e^{i\omega}) E_N(\omega)$. After some more algebraic computations the last expression can be bounded by using Bernstein's inequality (Bosq (1998), Theorem 1.2) and bounding $||F_1(\theta, \theta_i, e^{-i\omega})||_\infty$. The bound on $\sqrt{\frac{1}{N}\sum_{t=1}^{N} v_2^2(t)}$ follows using the same approach. ∎

Comment on Lemma 3.3. In Lemma 3.3, ϵ_1 and ϵ_2 are arbitrary positive real numbers. Let $\epsilon_1 = \epsilon_2 = \epsilon$ and make equation (15) explicit with respect to ϵ as a function of κ_1, η, n_0, n_1, and N: $\epsilon = \epsilon(\kappa_1, \eta, n_0, n_1, N)$. The following results readily come from the expression for

κ_1: $\epsilon \to \infty$ as $\kappa_1 \to 0$, $\epsilon \to \infty$ as $\eta \to 1$, $\epsilon \to \infty$ as n_0 and/or $n_1 \to \infty$, and $\epsilon \to 0$ as $N \to \infty$

Substitute now $\epsilon(\kappa_1, \eta, n_0, n_1, N)$ in the expression for γ_1 and further let ρ depend on N such that $\rho(N) \to 0$ as $N \to \infty$ (decreasing the radius of the balls - and thereby increasing their number - when $N \to \infty$ corresponds to the intuitive idea that a larger number of expected values can be estimated from data when the sample size becomes larger). Then, it is clear that: $\gamma_1 \to \infty$ as $\kappa_1 \to 0$, $\gamma_1 \to \infty$ as $\eta \to 1$, $\gamma_1 \to \infty$ as n_0 and/or $n_1 \to \infty$, and $\gamma_1 \to 0$ as $N \to \infty$.

Lemma 3.4 *For each integer* $q \in [1, N/2]$ *and* $k \geq 3$, *and any* $\gamma_2 > 0$

$$\max_i \left| \frac{1}{N} \sum_{t=1}^{N} \left((\tilde{H}_{\theta_i} e(t))^2 - E(\tilde{H}_{\theta_i} e(t))^2 \right) \right| \leq \gamma_2$$

with probability at least $1 - \kappa_2$ *where*

$$\begin{aligned}
\kappa_2 &= M(\rho) \left(a_1 e^{\frac{-q\gamma_2^2}{50K_H^2\sigma_e^4 + 10K_H\sigma_e^2\gamma_2}} + \right. \\
&\quad \left. a_2(k)\alpha([N/(q+1)])^{\frac{2k}{2k+1}} \right) \\
a_1 &= 2N/q + \\
&\quad 2\left(1 + \frac{\gamma_2^2}{50K_H^2\sigma_e^4 + 10K_H\sigma_e^2\gamma_2}\right) \\
a_2(k) &= 11N\left(1 + \frac{5kK_H\sigma_e^2}{\gamma_2}\right)^{\frac{k}{2k+1}} \\
K_H &= 2^{2n_{01}} \frac{(2n_{01}-2)!}{((n_{01}-1)!)^2} \frac{1}{(1-\eta^2)^{2n_{01}-1}} \\
n_{01} &= n_0 + n_1
\end{aligned}$$

$\alpha(l)$ *is given in Lemma A.1.*

Sketch of Proof. The bound follows by using Theorem 1.4 in Bosq (1998) applied to $X_t = (\tilde{H}_{\theta_i} e(t))^2 - E(\tilde{H}_{\theta_i} e(t))^2$ for $i = 1, \ldots, M(\rho)$. The theorem in Bosq (1998) is a generalisation of Bernstein's inequality to α-mixing processes, and it follows from Lemma A.1 that $\tilde{H}_{\theta_i} e(t)$ and hence $(\tilde{H}_{\theta_i} e(t))^2$ is α-mixing. ■

Comment on Lemma 3.4. Let q depend on N in such a way that $q(N) \to \infty$ and that $q(N)/N \to 0$ as $N \to \infty$. Also, let ρ depend on N and make the expression for κ_2 explicit with respect to γ_2. Then, it is easy to show that if $\rho(N)$ tends to zero at a sufficiently slow rate, γ_2 exhibits a limiting behavior similar to the one for γ_1.

Lemma 3.5

$$\max_i \sup_{\theta \in B_i} \left| \frac{1}{N} \sum_{t=1}^{N} E(\tilde{H}_{\theta_i} e(t))^2 - E(\tilde{H}_{\theta} e(t))^2 \right| \leq \gamma_3$$

where

$$\gamma_3 = K_1 \frac{(2^{n_1}\rho + 2\eta)\rho\sigma_e^2}{(1-\eta^2)^{2n_0+2n_1}} \tag{18}$$

$$K_1 = 2^{2n_0+3n_1+1} \frac{(2n_0+2n_1-1)!}{(n_1+n_0)!(n_0+n_1-1)!} \tag{19}$$

Sketch of Proof. Follows by bounding $\left| \, ||\tilde{H}_{\theta_i}||_2^2 - ||\tilde{H}_{\theta}||_2^2 \, \right|$ ■

Comment on Lemma 3.5. Let as usual ρ depend on N, $\rho = \rho(N)$, where $\rho(N)$ is any function tending to zero as $N \to \infty$. Then, it is clear that: $\gamma_3 \to \infty$ as $\eta \to 1$, $\gamma_3 \to \infty$ as n_0 and/or $n_1 \to \infty$, and $\gamma_3 \to 0$ as $N \to \infty$

Due to space limitations we omit the exact expressions in the next lemma.

Lemma 3.6

$$\max_i \sup_{\theta \in B_i} \left| \frac{2}{N} \sum_{t=1}^{N} H^{-1}(\theta)(G_0 - G(\theta))u(t-1) \cdot \tilde{H}_{\theta} e(t) \right| \leq \gamma_4$$

with probability at least $1 - \kappa_4$.

Sketch of proof. The lemma can be proved using the same techniques as in the proof of Lemma 3.3 and 3.4. ■

Under the same assumptions as in the comment to Lemma 3.4 it can be shown that γ_4 exhibits a limiting behavior similar to γ_2.

Theorem 2.1 now follows by noting that the probability that one or more of the inequalities in Lemma 3.3 - 3.6 is violated is at most $\sum_{i=1}^{4} \kappa_i$, and hence they must simultaneously hold with probability at least $1 - \sum_{i=1}^{4} \kappa_i$.

4 Concluding remarks

In this paper, we have studied the quality of system identification models obtained using a quadratic prediction error criterion. The main feature of our results is that they hold true for a finite data sample and are not asymptotic.

In this concluding remark, we would like to remark on a technical aspect of our approach which may be of general interest. The main result is that the empirical and theoretical identification criteria are close to one another uniformly in θ with high probability provided that a certain (finite) number of data points is available. The key aspect of this result is its *uniformity* in θ. Mathematically, this can be rephrased by saying that we have studied a problem of *uniform* convergence of empirical expectations to their true values. This type of problems have received increasing attention in the control community, see e.g. Vidyasagar (1997), Weyer et al (1999). However, differently from the cited literature, we have not used the notion of Vapnik-Chervonenkis (or Pollard) dimension in order to measure the complexity of the function class under consideration (i.e. $\bar{V}_N(\theta)$). Instead, a ρ-net in the Θ space has been used in conjunction with generalized Hoeffding or Bernstein inequalities (Bosq (1998)) and continuity results. In doing so, we have exploited the fundamental fact that in these inequalities the probability of a bad multisample decreases exponentially with the number of data points or, equivalently, the number of data points necessary to guarantee that the result holds true with a certain confidence $1-\delta$ increases only as $\log(1/\delta)$ (popularly phrased as "confidence is cheap"). It is the opinion of the authors that the same approach can be exploited in other identification/control settings to work out results tighter than those obtained via the Vapnik-Chervonenkis/Pollard dimension.

Acknowledgement: This work was supported by the Australian Research Council.

References

[1] Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes - Estimation and Prediction,*. Lecture Notes in Statistics 110. Springer Verlag.

[2] Campi M. C., and P. R. Kumar (1998). "Learning Dynamical Systems in a Stationary Environment" *Systems and Control Letters*, Vol. 34, pp. 125-132.

[3] Campi M. C., and E. Weyer (1999). "Finite sample properties in linear system identification." In preparation.

[4] Ljung, L. (1987). *System Identification - Theory for the User*. Prentice Hall.

[5] Modha D. S. and E. Masry (1998). "Memory-Universal Prediction of Stationary Random Processes," *IEEE Trans. Information Theory*, vol. 44, pp. 117-133.

[6] Vidyasagar M. (1997). *A theory of Learning and Generalization*. Springer Verlag.

[7] Weyer, E, R. C. Williamson, and I. M. Y. Mareels (1999). "Finite sample properties of linear model identification." To appear in *IEEE Trans. on Automatic Control.*

A α-mixing coefficients

Given a stochastic process $X_t, t \in \mathbb{Z}$, the associated $\alpha(k)$-mixing coefficients are defined as:

$$\alpha(k) := \sup_{t\in\mathbb{Z}} \sup_{B\in\sigma X_s, s\le t \;\; C\in\sigma X_s, s\ge t+k} |P(B\cap C) - P(B)P(C)|.$$

The process is said to be α-mixing if $\alpha(k) \to 0$, when $k \to 0$.

Lemma A.1 *Let $z(t) := H^{-1}(\theta)H_0 e(t)$. The sequence $z(t)$ is α-mixing and for $k \ge 0$ the mixing coefficients are bounded by*

$$\alpha(k+n_0+n_1+1) \le C(k+2)\cdots(k+n_0+n_1)\eta^{k+1} \tag{20}$$

where

$$C = \frac{2^{2n_0+2n_1+1}(2n_0+2n_1-1)!}{((n_0+n_1-1)!)^3} \cdot \frac{1}{(1-\eta^2)^{2n_0+2n_1}(1-\eta)^{n_0+n_1}}$$

Proof: See Campi and Weyer (1999). ∎