STRUCTURE IDENTIFICATION USING SEPARATE VALIDATION DATA - ASYMPTOTIC PROPERTIES*

Tor A. Johansen †

Department of Engineering Cybernetics
Norwegian Institute of Technology
N-7034 Trondheim, Norway.
Email. Tor.Arne.Johansen@regtek.sintef.no

Erik Weyer

Department of Chemical Engineering
The University of Queensland
Brisbane QLD 4072, Australia.
Email: erikw@cheque.uq.oz.au

Keywords. System Identification, Non-linear Systems, Convergence, Stochastic Processes.

Abstract

Model structures are compared by estimating their expected prediction performance on a validation data sequence that is independent of the data sequence used for parameter estimation. Here we show that under reasonable assumptions, this common procedure asymptotically leads to a model with the best possible expected prediction performance within the given model set.

1 Introduction

The decomposition of a model set into a parametric and a structural level is common in system identification. The parameters are typically identified using a prediction error criterion, e.g. [1], while there are several possible criteria for structure identification. These include FPE [2], AIC [3], MDL [4], Bayesian criteria [5], cross-validation [6], the unbiasness criterion [7], as well as the simplest and perhaps most popular criterion, namely the the use of separate validation data to assess the prediction performance of the different identified models. The major drawback of this last procedure, compared to the rest, is that a longer data sequence is needed.

Suppose a bounded sequence of $2l$ input/output observations from a dynamical system is available:

$$Z_{2l} \quad = \quad (z_1, z_2, ..., z_{2l})$$

where $z_t = (u_t, y_t)$, $u_t \in R^r$ is the input vector, and $y_t \in R^m$ is the output vector, at time t. The problem we address is the one of identifying a model of the system on the basis of these observations. We will in particular study the identified model as l approaches infinity. We will assume that the given model set that can be decomposed into two levels, namely model structure and model parameters. The contribution of this paper is a novel convergence result for the hierarchical identification procedure based on splitting the data sequence Z_{2l} into separate independent data sequences for parameter and structure identification, under similar assumptions as in the asymptotic analysis of the parameter identification problem given in [8]. Indeed, such a result should be expected, since the use of separate validation data should give a good estimate of the expected prediction performance. It should be noted that some asymptotic results are given for related criteria based on separate validation data in [9, 10], in a significantly less general framework than in this paper. In particular, these results only hold for a finite number of model structures, and the observations are assumed to be independent. Related results for some of the other approaches mentioned initially are also known [11]. For instance, it is known that order selection using the MDL criterion [12] or Bayesian criterion [5] are consistent, in the sense of convergence to the simplest model structure that contains the true system. It is also known that AIC does not always converge to the simplest model structure that contains the true system [13].

The paper is organized as follows: First, in section 2, we present the main assumptions and discuss their immediate implications as well as the practical limitations imposed by these assumptions. Then the asymptotic properties of the hierarchical identification procedure is analysed, in section 3. The paper ends up in some conclusions.

2 Main Assumptions and Preliminary Results

Let $\mathcal{S}$ be a set of model structures. A model structure $S \in \mathcal{S}$ is a set of possibly non-linear equations, which may be algebraic, difference, differential, or a combination. The equations may contain some unknown parameters θ that belong to a parameter set Θ_S, which is a subset of a finite-dimensional Euclidean vector space, by assump-

*This work was partly supported by the Research Council of Norway under doctoral scholarship grant no. ST. 10.12.221718 given to the first author.

†Present address: SINTEF Automatic Control, N-7034 Trondheim, Norway.

tion. We will not explicitly require Θ_S to be compact, but as we shall see, this is typically required to satisfy the assumptions. This defines a model set

$$\mathcal{M} \quad = \quad \{M = (S, \theta) \mid S \in \mathcal{S}, \theta \in \Theta_S\}$$

Suppose that for each model M we are able to make a prediction $\hat{y}_t(M, Z_{t-1})$ of the output y_t on the basis of the data sub-sequence Z_{t-1}. Notice that the dependence of the model M on the predictor can be arbitrary. The one-step-ahead prediction error using the model M is defined by $\varepsilon_t(M, Z_t) = y_t - \hat{y}_t(M, Z_{t-1})$. In the following, we may let the dependence on Z_{t-1} be left implicit when it is more convenient, and use the notation $\hat{y}_t(M)$ and $\varepsilon_t(M)$.

Before we proceed, we need to introduce some notation. Let the γ-norm of a vector $x \in R^n$ be defined by $||x||_\gamma = (|x_1|^\gamma + ... + |x_n|^\gamma)^{1/\gamma}$ for all $\gamma \geq 1$, and the Euclidean norm be denoted by $||x|| = ||x||_2$.

The following assumption states the key property of the model set. It essentially limits the complexity of the model set, in the sense that there must exists a finite number of good approximations to the possibly infinite number of models in the model set.

Assumption M1. For any $\epsilon > 0$ the model set $\mathcal{M}$ can be covered by a finite ϵ-net with $N(\epsilon)$ elements, i.e. there exists a finite set of models $\mathcal{M}_\epsilon = \{M_1, M_2, ..., M_{N(\epsilon)}\}$ (not necessarily a subset of $\mathcal{M}$) such that for any $M \in \mathcal{M}$ there exists an $M_i(M) \in \mathcal{M}_\epsilon$ with the property

$$\sup_{t, Z_{t-1}} ||\hat{y}_t(M, Z_{t-1}) - \hat{y}_t(M_i(M), Z_{t-1})|| \quad \leq \quad \epsilon \qquad (1)$$

Please apologize for possible confusion by using ϵ and ε in two different meanings.

□

Before we proceed, let us discuss the limitations imposed by assumption M1 on the model set.

Lemma 1 *Suppose*

1. *$\mathcal{S}$ contains only a finite number of model structures.*
2. *For every $S \in \mathcal{S}$ the parameter set Θ_S is compact.*
3. *For every $S \in \mathcal{S}$ the predictor $\hat{y}_t(S, \theta, Z_{t-1})$ is uniformly continuous in θ on Θ_S.*

Then $\mathcal{M}$ can be covered by a finite ϵ-net.

Proof. Let $S \in \mathcal{S}$ be arbitrary. From the uniform continuity there exists for every $\epsilon > 0$ a $\delta > 0$ that depends only on ϵ, with the property that

$$||\theta_1 - \theta_2|| < \delta \quad \text{implies} \quad \sup_{t, Z_{t-1}} ||\hat{y}_t(S, \theta_1, Z_{t-1}) - \hat{y}_t(S, \theta_2, Z_{t-1})|| < \epsilon \qquad (2)$$

uniformly in Θ_S. Since Θ_S is compact, it can be covered by a finite number of δ-balls. The parameter vectors at the centre of these δ-balls define an ϵ-net for the model set $\mathcal{M}_S = \{M = (S, \theta) \mid \theta \in \Theta_S\}$. Since $\mathcal{S}$ contains only a finite number of model structures, the result follows.

□

Uniform continuity may exclude certain types of models and systems. Consider for example the simple model (predictor) $\hat{y}_t(S, \theta, Z_{t-1}) = \theta y_{t-1}$. We get

$$\sup_{t, Z_{t-1}} ||\hat{y}_t(S, \theta_1, Z_{t-1}) - \hat{y}_t(S, \theta_2, Z_{t-1})|| \quad \leq \quad \delta \sup_{y_{t-1}} ||y_{t-1}|| \; < \; \epsilon \qquad (3)$$

The system output must be bounded in order for uniform continuity to hold. However, we are free to choose the predictors ourselves, and may include for example saturation effects for very large outputs in the predictor to avoid such problems.

Lemma 2 *Suppose*

1. *There exists a compact set $\mathcal{Z}$ such that $z_t \in \mathcal{Z}$ for all t.*
2. *There exists a constant K such that*
$$\sup_{t, Z_{t-1}, M} ||\hat{y}_t(M, Z_{t-1})|| \quad \leq \quad K \qquad (4)$$
3. *For every $k \geq 1$ there exists a uniformly continuous function $f_k : \mathcal{M} \times \mathcal{Z}^k \to R^m$ with the property*
$$\sup_{t, Z_{t-1}, M} ||\hat{y}_t(M, z_{t-1}, ..., z_1) - f_k(M, z_{t-1}, ..., z_{t-k})|| \to 0$$
as $k \to \infty$.

Then $\mathcal{M}$ can be covered by a finite ϵ-net.

Proof. Let $\epsilon > 0$ be arbitrary. Choose k so large that

$$\sup_{t, Z_{t-1}, M} ||\hat{y}_t(M, z_{t-1}, ..., z_1) - f_k(M, z_{t-1}, ..., z_{t-k})|| \leq \epsilon/2$$

Provided the set of functions

$$\mathcal{F}_k \quad = \quad \{f_k(M, \cdot) : \mathcal{Z}^k \to R^m \mid M \in \mathcal{M}\} \qquad (5)$$

can be covered with a finite $\epsilon/2$-net, this $\epsilon/2$-net will also be an ϵ-net for $\mathcal{M}$. From the uniform continuity of f_k there exists a $\delta > 0$ such that

$$||Z^1 - Z^2|| < \delta \quad \text{implies} \quad \sup_{M \in \mathcal{M}} ||f_k(M, Z^1) - f_k(M, Z^2)|| < \epsilon/4 \qquad (6)$$

Cover $\mathcal{Z}^k$ with a finite number of boxes $\mathcal{Z}_1, \mathcal{Z}_2, ..., \mathcal{Z}_N$, where each box is contained within a δ-ball, and

$$\mathcal{Z} \subset \bigcup_i \mathcal{Z}_i \qquad \text{and} \qquad \mathcal{Z}_i \cap \mathcal{Z}_j = \emptyset \;\; \text{for } i \neq j \qquad (7)$$

Let μ_i be the characteristic function for $\mathcal{Z}_i$. From (4) and (6) it is clear that there exists indices $\{l(M, i) \mid i = 1, 2, ..., N\}$ such that

$$\sup_{M \in \mathcal{M}} \sup_{Z \in \mathcal{Z}} \left|\left| f_k(M, Z) - \sum_{i=1}^{N} \beta_{l(M,i)} \mu_i(Z) \right|\right| < \epsilon/4 \qquad (8)$$

where the constants $\beta_{l(M,i)} \in R^m$ are chosen from a finite set of values, determined by K and ϵ. Hence, the approximation $\sum_{i=1}^{N} \beta_{l(M,i)}\mu_i(Z)$ can be used to define a finite $\epsilon/2$-net for $\mathcal{F}_k$.

Remarks. It will later be required that the ϵ-net for $\mathcal{M}$ must consists of predictors that are smooth functions of past data, cf. M2. This may be achieved by replacing the μ_i-functions by smoothed versions. It should be mentioned that the use of a piecewise constant approximation in the proof is only one possibility, and several other approximators are expected to work as well.

□

The third assumption requires that the dependence of the predictor on past data must vanish asymptotically. Roughly speaking, this means that the predictor must be asymptotically stable. It also follows from this assumption that the predictor must be uniformly continuous. Moreover, it must be time-invariant, which is a quite restrictive assumption that need not hold when the predictor contains a state estimator, for example. On the other hand, boundedness of the data and prediction (cf. the first and second assumptions) are quite reasonable assumptions.

We would also like to stress that Lemma 1 and 2 are only examples of sufficient conditions for the existence of ϵ-nets that cover $\mathcal{M}$, and weaker conditions may very well exist. It must also be emphasized that it is sufficient to prove the existence of an ϵ-net. However, as we have seen, the proofs are typically constructive.

The asymptotic analysis is based on a law of large numbers that holds uniformly in the model set. We will in the next paragraphs state a set of assumptions that together with M1 ensure that such a law holds. We will introduce a stochastic framework that allows us to deal with non-stationary stochastic processes that may contain deterministic components and sequential correlations. In the following, let E denote expectation with respect to input/output observations, and following [1] we define the operator $\overline{E}$ by $\overline{E}x_t = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} Ex_t$ for any vector sequence $x = (x_t)$. Notice that it is implicitly assumed that the limit exists in all subsequent applications of this operator.

Definition [8]/[14]. A sequence (x_t) is said to be exponentially forgetting of order γ if there exists a doubly indexed sequence (x_t^s) where x_t^s is independent of the subsequences $(x_0, x_1, ..., x_s)$ and $(x_p^p, x_{p+1}^p, ..., x_s^p)$ for all $p \leq s$, $x_t^t = 0$, and for every $t \geq s$

$$E\,||x_t - x_t^s||_\gamma^\gamma \quad \leq \quad C\lambda^{t-s} \tag{9}$$

for some constants $C \geq 0$ and $\lambda \in (0,1)$.

□

This means that it must be possible to approximate the sequence (x_t) with a sequence (x_t^s) that is independent of the remote past. Moreover, the approximation must be so good that the "remote past is forgotten at an exponential rate". In [8], this is used as a definition of exponential stability of stochastic dynamical systems, and it is shown that in many cases it contains the classical definition of exponential stability.

Assumption S1. The sequences (u_t) and (y_t) are exponentially forgetting of order 4.

□

On the model set $\mathcal{M}$ and the ϵ-net $\mathcal{M}_\epsilon$ we need to make an additional assumption to ensure that the squared prediction error is an exponentially forgetting sequence when S1 holds.

Assumption M2. Define the sequence $h(M) = (h_t(M)) = (||\varepsilon_t(M)||^2)$, and suppose

1. The function $\hat{y}_t(M, Z_{t-1})$ is differentiable with respect to Z_{t-1} for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$.
2. For all $t \geq 0$ and $k \geq 1$, there exists a convex function $g_{t,k}$ such that for all models $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$
$$\left|\left|\frac{\partial}{\partial z_{t-k}} h_t(M, z_{t-1}, ..., z_1)\right|\right| \quad \leq \quad g_{t,k}(z_{t-1}, ..., z_1)$$
3. There exist constants C and $\lambda \in (0,1)$ such that for all $t \geq 0$ the set of functions $\{g_{t,k} \mid k \geq 1\}$ has the property
$$\begin{aligned} &E||z_t||^4 \ < \infty \quad \text{for all} \ \ t \\ &\qquad \text{implies} \quad E(g_{t,k}(z_{t-1}, ..., z_1))^4 \ \leq \ C\lambda^k \end{aligned}$$
4. Moreover, $(h_t(M, 0, 0, ..., 0))$ is a bounded sequence for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$.

□

First, assumption M2 means that the sensitivity of the squared prediction error with respect to the past inputs and outputs z_{t-k} must be bounded by a function that decays exponentially with k, which basically means that the predictor must be exponentially stable. In addition, M2 impose some growth conditions on the predictor. It is clearly sufficient that the derivative of the predictor is bounded. Furthermore, the prediction must depend smoothly on the past data. These are reasonable conditions that are likely to be fulfilled for many model sets, see also [14], p. 107.

Lemma 3 [14], pp. 46-48. *Suppose Assumptions S1 and M2 hold, then for all $M \in \mathcal{M} \cup \mathcal{M}_\epsilon$ the sequences $(||\varepsilon_t(M)||^2)$ and $(\sup_{M\in\mathcal{M}} ||\varepsilon_t(M)||^2)$ are exponentially forgetting of order 2.*

□

Exponential forgetting of order 2 is sufficient for the strong law of large numbers to hold:

Theorem 1 (Law of Large Numbers.) *Suppose a scalar sequence (h_t) is exponentially forgetting of order 2. Then*

$$\left|\overline{E}h_t - \frac{1}{l}\sum_{t=1}^{l} h_t\right| \to 0 \qquad \textit{w.p. 1} \quad \textit{as} \quad l \to \infty$$

Proof. Follows from a somewhat stronger result in [14], p. 66.
□

This law is a prerequisite for a uniform law of large numbers, which is the main tool in the asymptotic analysis.

Theorem 2 (Uniform Law of Large Numbers.) *Suppose $\mathcal{H}$ is a set of scalar sequences that are exponentially forgetting of order 2, and for each $\epsilon > 0$ there exists a finite set $\mathcal{H}_\epsilon$ of exponentially forgetting sequences, containing lower and upper approximations $h^{\epsilon,L} \in \mathcal{H}_\epsilon$ and $h^{\epsilon,U} \in \mathcal{H}_\epsilon$ to each $h \in \mathcal{H}$ such that*

$$h_t^{\epsilon,L} \leq h_t \leq h_t^{\epsilon,U} \quad \text{and} \quad \overline{E}\left(h_t^{\epsilon,U} - h_t^{\epsilon,L}\right) < \epsilon \quad (10)$$

Then

$$\sup_{h \in \mathcal{H}} \left| \overline{E} h_t - \frac{1}{l}\sum_{t=1}^{l} h_t \right| \rightarrow 0 \quad w.p.\ 1 \quad as \quad l \rightarrow \infty \quad (11)$$

Proof. Follows from a theorem in [15], p. 8, and Theorem 1.
□

The class of exponentially forgetting processes is closely related to mixing processes, and in particular L-mixing processes, [16]. However, as noted in [14], an equivalent to Lemma 3 appears to be more difficult to establish for such processes.

3 Structural and Parametric Identification

We define the expected squared one-step-ahead prediction error

$$\overline{V}(M) = \overline{E}\,||\varepsilon_t(M)||^2 \quad (12)$$

The ultimate goal is to find a model in $\mathcal{M}$ that makes $\overline{V}$ as small as possible. Notice that the results in this section can be easily reformulated for multi-step predictors as well.

Minimization of the expected squared prediction error $\overline{V}(M)$ will in many cases reflect the desired properties of the model, and is therefore suited as a criterion for model identification. Unfortunately, the underlying probability measure is assumed to be unknown, and consequently, the expectation cannot be computed. Consider instead the empirical average squared prediction error

$$V_l(S,\theta) = \frac{1}{l}\sum_{t=1}^{l} ||\varepsilon_t(S,\theta)||^2 \quad (13)$$

estimated on the basis of the first half of the data sequence Z_{2l}. We minimize for each model structure $S \in \mathcal{S}$ the criterion $V_l(S,\theta)$ with respect to θ in order to find a parameter estimate.

Assumption A1. For every l and $S \in \mathcal{S}$ the following global minimum exists with probability one

$$\hat{\theta}_{S,l} = \arg\min_{\theta \in \Theta_S} V_l(S,\theta)$$

□

For instance, A1 holds automatically if Θ_S is compact, and V_l is a continuous function of θ. Notice that we do not assume that the minimum is unique, but use the convention that if more than one value of θ minimizes $V_l(S,\theta)$, then is $\hat{\theta}_{S,l}$ an arbitrary minimum. Non-uniqueness should indeed be expected, because $\mathcal{M}$ may contain model structures of high complexity, possibly with an unbounded number of parameters. If this is the case, then for l smaller than the number of parameters in the model structure, the parameter identification problem is ill-posed and there may be several parameter vectors that give exact match to the data.

We define $\hat{M}_{S,l} = (S, \hat{\theta}_{S,l})$. This is the (not necessarily unique) model with structure S that on the basis of the l observations is believed to be the best. Next, we need to choose among the models in the set $\{\hat{M}_{S,l} \mid S \in \mathcal{S}\}$. We define a new criterion on the basis of the second half of the data sequence Z_{2l}

$$V_l'(S,\hat{\theta}_{S,l}) = \frac{1}{l-k}\sum_{t=l+k+1}^{2l} \left|\left|\varepsilon_t(S,\hat{\theta}_{S,l})\right|\right|^2 \quad (14)$$

throwing away the first k samples to ensure that the remaining samples are not too much correlated with $\hat{\theta}_{S,l}$, cf. the definition of exponentially forgetting sequences.

Assumption A2. For every l, the global minimum

$$\hat{S}_{2l} = \arg\min_{S \in \mathcal{S}} V_l'(S,\hat{\theta}_{S,l})$$

exists with probability one.
□

Again, this minimum need not be unique, and we use the convention that in the case of non-uniqueness, $\hat{S}_{2l}$ is any minimum. In summary, the model estimate $\hat{M}_{2l}$ is defined by the following hierarchical optimization procedure

$$\text{Parameter Identification:} \quad \hat{\theta}_{S,l} = \arg\min_{\theta \in \Theta_S} V_l(S,\theta) \quad (15)$$
$$\text{for all } S \in \mathcal{S}$$
$$\text{Structure Identification:} \quad \hat{S}_{2l} = \arg\min_{S \in \mathcal{S}} V_l'(S,\hat{\theta}_{S,l}) \quad (16)$$
$$\text{Identified Model:} \quad \hat{M}_{2l} = (\hat{S}_{2l}, \hat{\theta}_{\hat{S}_{2l},l}) \quad (17)$$

With a finite sample of empirical data available, it is evident that there will be a trade-off between the structural and parametric error components. A too simple model structure will typically give a dominating structural error, while the parametric error is typically small because of a low degrees of freedom to data ratio. On the other hand, a too complex model structure may lead to a small structural error, but since the degrees of freedom to data ratio is high, the parametric error will be large. Asymptotically (as $l \to \infty$), it is possible to simultaneously reduce both the parametric and structural errors. This requires that the identification procedure is able to find a delicate balance between structural and parametric error.

We are now in position to prove the main result:

Theorem 3 *Suppose Assumptions A1 and A2 hold, and*

$$\sup_{M \in \mathcal{M}} \left| V_l(M) - \overline{V}(M) \right| \to 0 \quad w.p.\ 1 \ as \ l \to \infty \quad (18)$$

$$\sup_{S \in \mathcal{S}} \left| V_l'(S, \hat{\theta}_{S,l}) - \overline{V}(S, \hat{\theta}_{S,l}) \right| \to 0 \quad w.p.\ 1 \ as \ l \to \infty \quad (19)$$

Then

$$\overline{V}(\hat{M}_{2l}) \quad \to \quad \inf_{M \in \mathcal{M}} \overline{V}(M) \quad w.p.\ 1 \ as \quad l \to \infty \qquad (20)$$

Proof. Successive applications of the triangle inequality gives

$$\begin{aligned}
\left| \overline{V}(\hat{M}_{2l}) - \inf_{M \in \mathcal{M}} \overline{V}(M) \right| &= \left| \overline{V}(\hat{M}_{2l}) - \inf_{S \in \mathcal{S}} \inf_{\theta \in \Theta_S} \overline{V}(S, \theta) \right| \\
&\leq \left| \overline{V}(\hat{M}_{2l}) - \inf_{S \in \mathcal{S}} \overline{V}(S, \hat{\theta}_{S,l}) \right| \\
&\quad + \left| \inf_{S \in \mathcal{S}} \overline{V}(S, \hat{\theta}_{S,l}) - \inf_{S \in \mathcal{S}} \inf_{\theta \in \Theta_S} \overline{V}(S, \theta) \right| \\
&\leq \left| \overline{V}(\hat{M}_{2l}) - V_l'(\hat{M}_{2l}) \right| \\
&\quad + \left| V_l'(\hat{M}_{2l}) - \inf_{S \in \mathcal{S}} V_l'(S, \hat{\theta}_{S,l}) \right| \\
&\quad + \left| \inf_{S \in \mathcal{S}} V_l'(S, \hat{\theta}_{S,l}) - \inf_{S \in \mathcal{S}} \overline{V}(S, \hat{\theta}_{S,l}) \right| \\
&\quad + \left| \inf_{S \in \mathcal{S}} \overline{V}(S, \hat{\theta}_{S,l}) - \inf_{S \in \mathcal{S}} V_l(S, \hat{\theta}_{S,l}) \right| \\
&\quad + \left| \inf_{S \in \mathcal{S}} V_l(S, \hat{\theta}_{S,l}) - \inf_{S \in \mathcal{S}} \inf_{\theta \in \Theta_S} V_l(S, \theta) \right| \\
&\quad + \left| \inf_{S \in \mathcal{S}} \inf_{\theta \in \Theta_S} V_l(S, \theta) - \inf_{S \in \mathcal{S}} \inf_{\theta \in \Theta_S} \overline{V}(S, \theta) \right|
\end{aligned}$$

The second and fifth terms are zero, with probability one, cf. (15) and (16). It is straightforward to prove that for any set $\mathcal{X}$, and functions $f, g : \mathcal{X} \to R$ that are bounded from below, that

$$\left| \inf_{x \in \mathcal{X}} f(x) - \inf_{x \in \mathcal{X}} g(x) \right| \quad \leq \quad \sup_{x \in \mathcal{X}} |f(x) - g(x)| \quad (21)$$

Hence, with probability one

$$\begin{aligned}
\left| \overline{V}(\hat{M}_{2l}) - \inf_{M \in \mathcal{M}} \overline{V}(M) \right| \quad &\leq \\
2 \sup_{S \in \mathcal{S}} &\left| V_l'(S, \hat{\theta}_{S,l}) - \overline{V}(S, \hat{\theta}_{S,l}) \right| \\
+2 \sup_{S \in \mathcal{S}} \sup_{\theta \in \Theta_S} &\left| V_l(S, \theta) - \overline{V}(S, \theta) \right| \qquad (22)
\end{aligned}$$

Both these terms go to zero with probability one as $l \to \infty$, cf. (18) and (19).

□

The result tells us that the expected squared prediction error when the model $\hat{M}_{2l}$ is used for prediction will tend towards the smallest possible with probability one as $l \to \infty$. In particular,

Corollary 1 *Suppose the assumptions in Theorem 3 hold, and in addition*

1. *There exists a unique $M^\star \in \mathcal{M}$ such that $\overline{V}(M^\star) = \inf_{M \in \mathcal{M}} \overline{V}(M)$.*
2. *The parameterizations of the model structures are such that a well-defined metric on $\mathcal{M}$ can be defined, and $\mathcal{M}$ is compact.*

Then $\hat{M}_{2l} \to M^\star$ with probability one as $l \to \infty$.

□

This result is commonly defined as consistency of the model estimate. The conditions are related to identifiability of the model set, see also [1, 8].

Next, we discuss assumptions under which (18) and (19) hold.

Lemma 4 *Suppose S1, M1, M2 and A1 hold, then (18) and (19) hold.*

Proof of (18): The proof is inspired by a result in [17]. Recall the definition of the sequence $h(M) = (||\varepsilon_t(M)||^2)$ and define the set of sequences $\mathcal{H} = \{h(M) \mid M \in \mathcal{M}\}$. Of course, the main tool is Theorem 2, and the proof is basically to show that $h(M)$ satisfies the assumptions of this theorem.

Pick an arbitrary $\epsilon > 0$ and $M \in \mathcal{M}$, and let $M_i(M) \in \mathcal{M}_{\epsilon'}$ denote the model in the ϵ'-net $\mathcal{M}_{\epsilon'}$ that is closest to M (in the sense of (1)), where ϵ' is yet unspecified. Now

$$\begin{aligned}
h_t(M_i(M)) - 2\epsilon' \sqrt{h_t(M_i(M))} \quad &\leq \\
h_t(M) \ &\leq \ \left(\sqrt{h_t(M_i(M))} + \epsilon' \right)^2 \qquad (23)
\end{aligned}$$

and since $\sqrt{x} \leq 1 + x$, lower and upper approximations to the sequence $h(M)$ are given by

$$\begin{aligned}
h_t^{\epsilon,L} \ = \ h_t(M_i(M)) - 2\epsilon'(1 + h_t(M_i(M))) \ &\leq \ h_t(M) \\
\leq \ h_t(M_i(M)) + 2\epsilon'(1 + h_t(M_i(M))) + (\epsilon')^2 \ &= \ h_t^{\epsilon,U}
\end{aligned}$$

Taking the expectation, we get

$$\overline{E} \left(h_t^{\epsilon,U} - h_t^{\epsilon,L} \right) \quad = \quad (\epsilon')^2 + 4\epsilon' + 4\epsilon' \overline{V}(M_i(M)) \quad (24)$$

Now

$$\overline{V}(M_i(M)) \quad \leq \quad \sup_{M \in \mathcal{M}} \overline{V}(M_i(M)) \ \leq \ (K + \epsilon')^2 \quad (25)$$

since it follows directly that $(\sup_{M \in \mathcal{M}} h_t(M))$ is exponentially forgetting of order 2 from Lemma 3, and it therefore exists a constant K such that

$$\sup_{M \in \mathcal{M}} \overline{V}(M) \quad \leq \quad K \qquad (26)$$

Choosing $\epsilon' = \epsilon'(K, \epsilon)$ sufficiently small ensures

$$\overline{E}\left(h_t^{\epsilon,U} - h_t^{\epsilon,L}\right) \quad < \quad \epsilon \tag{27}$$

Hence, the conditions of Theorem 2 are fulfilled, and (18) follows.

Proof of (19): The terms in V_l' will make up a sequence $(||\varepsilon_t(S, \hat{\theta}_{S,l})||^2)$ that is exponentially forgetting of order 2, and since the model subset $\{M = (S, \hat{\theta}_{S,l}) \mid S \in \mathcal{S}\} \subset \mathcal{M}$ can be covered by a finite ϵ-net, the rest of the proof is similar to that of (18).

□

Eq. (18) says that the empirical average V_l converges with probability one to its expectation $\overline{V}$ simultaneously for all possible models in the model set. This is often referred to as uniform convergence, or as a strong uniform law of large numbers. Its main implication is that V_l will be a close approximation to $\overline{V}$ for all models in $\mathcal{M}$, for large l. Hence, minimization of V_l and $\overline{V}$ are closely related, asymptotically. The key result of [8], which applies to a fixed model structure and some slightly different other assumptions, is a special case:

Corollary 2 *Suppose S1 and M2 hold. Consider a fixed model structure $S \in \mathcal{S}$, suppose Θ_S is compact and $\hat{y}_t(S, \theta)$ is differentiable with respect to θ for all t and $\theta \in \Theta_S$. Then*

$$\sup_{\theta \in \Theta_S} \left| V_l(S, \theta) - \overline{V}(S, \theta) \right| \to 0 \quad \textit{w.p.} \ 1 \quad \textit{as} \quad l \to \infty$$

□

The data sequence is split into two sequences of approximately equal length. However, the analysis holds as long as the length of both sequences go to infinity as l goes to infinity. In order to find a good split, one should consider the size of the model structure set relative to the parameter sets, and take into account the convergence rate, cf. (22). Unfortunately, accurate estimates of the convergence rates are not easily attainable. There are some convergence results available in [18, 15], but they suffer from two major drawbacks: They are potentially very conservative, and applicable for independent data points z_t only. It is not clear how to extend these results to exponentially forgetting processes.

4 Conclusions

We have analyzed the hierarchical identification of model structure and parameters on the basis of separate data sequences. We have shown that this procedure essentially has the same asymptotic convergence properties as the parameter identification procedure of [8]. We have introduced one additional assumption, namely the existence of a finite ϵ-net that covers the model set, but also illustrated that this assumption holds under some reasonable conditions on the model set, cf. Lemmas 1 and 2.

References

[1] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Inc., Englewood Cliffs, NJ., 1987.

[2] H. Akaike, "Fitting autoregressive models for prediction", *Ann. Inst. Stat. Math.*, vol. 21, pp. 243–247, 1969.

[3] H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automatic Control*, vol. 19, pp. 716–723, 1974.

[4] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, pp. 465–471, 1978.

[5] R. L. Kashyap, "A Bayesian comparison of different classes of dynamic models using empirical data", *IEEE Trans. Automatic Control*, vol. 5, pp. 715–727, 1977.

[6] P. Stoica, P. Eykhoff, P. Janssen, and T. Söderström, "Model-structure selection by cross-validation", *Int. J. Control*, vol. 43, pp. 1841–1878, 1986.

[7] A. G. Ivakhnenko and Y. P. Yurachkovsky, "System structure identification by sets of observation data on the base of unbiasness principles", in *Preprints IFAC Symposium on Identification and System Parameter Estimation, Beijing*, 1988, pp. 953–965.

[8] L. Ljung, "Convergence analysis of parametric identification methods", *IEEE Trans. Automatic Control*, vol. 23, pp. 770–783, 1978.

[9] V. S. Stepashko, "Asymptotic properties of external criteria for model selection", *Soviet J. Automation and Information Sciences*, vol. 21, pp. 84–92, 1988.

[10] T. I. Aksenova, "Sufficient covergence conditions for external criteria for model selection", *Soviet J. Automation and Information Sciences*, vol. 22, pp. 49–53, 1989.

[11] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[12] J. Rissanen, "Consistent order-estimates of autoregressive processes by shortest decscription of data", in *Analysis and Optimization of Stochastic Systems*, O. L. R. Jacobs, M. H. A. Davis, M. A. H. Dempster, C. J. Harris, and P. C. Parks, Eds., pp. 451–461. Academic Press, London, 1980.

[13] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion", *Biometrica*, vol. 63, pp. 117–126, 1976.

[14] H. Hjalmarsson, *Aspects on Incomplete Modeling in System Identification*, PhD thesis, University of Linköping, 1993.

[15] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

[16] L. Gerencsér, "On a class of mizing processes", *Stochastics*, vol. 26, pp. 165–191, 1989.

[17] E. Weyer, R. C. Williamson, and I. M. Y. Mareels, "System identification in the behavioural framework via risk minimization", In Preparation, 1995.

[18] V. Vapnik, *Estimation of Dependences based on Empirical Data*, Springer-Verlag, New York, 1982.