Minimum Description Complexity

by

Soosan Beheshti

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2002

Author .....
Department of Electrical Engineering and Computer Science
September 3, 2002

Certified by .....
Munther A. Dahleh
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Minimum Description Complexity

by

Soosan Beheshti

Submitted to the Department of Electrical Engineering and Computer Science
on September 3, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The classical problem of model selection among parametric model sets is considered. The goal is to choose a model set which best represents observed data. The critical task is the choice of a criterion for model set comparison. Pioneer information theoretic based approaches to this problem are Akaike information criterion (AIC) and different forms of minimum description length (MDL). The prior assumption in these methods is that the unknown true model is a member of all the competing sets.

We introduce a new method of model selection: minimum description complexity (MDC). The approach is motivated by the Kullback-Leibler information distance. The method suggests choosing the model set for which the model set relative entropy is minimum. We provide a probabilistic method of MDC estimation for a class of parametric model sets. In this calculation the key factor is our prior assumption: unlike the existing methods, no assumption of the true model being a member of the competing model sets is needed. The main strength of the MDC calculation is in its method of extracting information from the observed data.

Interesting results exhibit the advantages of MDC over MDL and AIC both theoretically and practically. It is illustrated that, under particular conditions, AIC is a special case of MDC. Application of MDC in system identification and signal denoising is investigated. The proposed method answers the challenging question of quality evaluation in identification of stable LTI systems under a fair prior assumption on the unmodeled dynamics. MDC also provides a new solution to a class of denoising problems. We elaborate the theoretical superiority of MDC over the existing thresholding denoising methods.

Thesis Supervisor: Munther A. Dahleh
Title: Professor

Acknowledgments

I thank my supervisor Professor Munther Dahleh sincerely for all his guidance and advice throughout my studies. Munther is a great researcher, mentor and friend, all at the same time, and I've learned a great deal from him over these past five years. I would also like to express my gratitude to Professors Sanjoy Mitter and George Verghese for their helpful comments and guidance as members of my PhD committee. I am grateful to my academic advisor Professor Alvin Drake for supporting me and believing in me. He always made the registration day a nice start of the semester.

It was my privilege to be a member of LIDS and I would like to thank the members of the control systems group and in particular my officemates, Sridevi Sarma, Nuno Martins, Georgios Kotsalis and Neal Jameson for all the academic and non-academic discussions. Special thanks go to the administrative staff of LIDS and the graduate office of the EECS department for all their great help.

Over the past several years, my life has been enriched by many wonderful friends and I am very grateful to them. In particular, I thank Nicole Lazo, Payman Kassaei, Farnaz Haghseta, Nazbeh Taghizadeh and Navid Laskarian. Living at Greenhall during my student life was a great experience. In Greenhall I met many interesting people of different backgrounds. I wish to thank all of them who became like a family to me, especially Ana-Maria Castravet who lived there the longest.

Last but not least, I thank my family. I am very grateful to my parents Mahin Chitsaz and Reza Beheshti for their love and support and for being my first teachers. Thanks to my sister Sima for being my close friend and my roommate for so many years. Also, thanks to my sister Nushin and my brother Ali for always believing in their oldest sister. I have to thank my niece, little Sadaf, for the joy she has brought to me and my family. Most importantly, I thank my husband Farinam for being him.

Contents

List of Figures

Chapter 1

Introduction

From biological to huge man-made systems, complex systems are all around us. Understanding the behavior of these systems is necessary for the purposes such as simulation, prediction and control. Describing these systems with a model of lower complexity becomes essential in practical applications. When no prior knowledge, or a partial knowledge, of the system's physics is available, the only source of understanding the system is through observation of its inputs and outputs. In this scenario the crucial question is how to *extract the most information* about the complex system, under a realistic prior assumption about the system. The motivation of this thesis primary was the search for a proper answer to this question.

System identification approaches to this problem are divided in two fields based on the prior assumption on the additive noise properties. In deterministic identification the additive noise belongs to a set with a bounded norm. In stochastic identification the additive noise is a sample of a random variable. In both approaches estimation of possibly a complex system in a parametric model set is investigated. The complexity of the true system is taken into account by prior assumptions on the unmodeled dynamics in the competing parametric sets. It is an important quality of an identification method to offer an efficient estimation method in each low-complexity model set and provide a proper information and comparison method on the estimation errors of the competing model sets [34]. In next chapter we thoroughly discuss several deterministic and stochastic approaches which attempt to satisfy this quality.

In practical problems more can be said about the correlation of the noise with itself and with the input compare to only a bounded norm definition. The conservative definition of additive noise in deterministic approaches prevents the method to provide any results on the convergence of the estimates in each competing set robustly, as the length of the data grows. On the other hand, in quality evalution of the estimates the stochastic approaches fail to address the unmodeled dynamics effects properly. It seems that all the stochastic and deterministic methods lack a proper, fair assumption on either the additive noise or on the model structure.

Continuing the search for a suitable quality evaluation method leads us to study information theoretic approaches to order estimation problem. Here the problem of parametric model selection is considered. The competing sets are parametric sets of different order. The question is that by using one observation of a random variable,

which is generated by a parametric model, which model set best represents the data. It is clear that our quality evaluation problem in system identification is a special case of this order estimation problem.

The well-known existing information theoretic model selection methods are Akaike information criterion(AIC), Bayesian information criterion(BIC) and different forms of minimum description length(MDL). AIC provides a method to estimate the Kullback-Liebler distance of the true model and the estimate of model asymptotically and suggests to use that as a comparison criterion. BIC is based on calculation of the probability that the model set includes the true model. In this method a prior probabilistic assumption on the parameters is also needed. MDL is introduced based on an idea to define a description length for the observed data given each model set. It suggests to choose the model set for which the description length is minimum. In this thesis we thoroughly study the theory and motivation behind each of these methods. The main common drawback of the theory of these methods is that calculation of these criteria is under the prior assumption that the unknown true model is a member of all the competing sets.

We invest on defining a new model selection method which can overcome the observed drawbacks. The first step is to define a proper distance measure between the true model and any given parametric model set. We define this distance, the description complexity, based on the Kullback- Liebler information distance. The next important step is to provide a method of estimation of this distance using only the observed data. Comparison of this distance for the competing sets leads to the choice of the set for which the minimum description complexity(MDC) is obtained.

In the last part of the thesis we illustrate the application of MDC in signal denoising. The problem of estimating an unknown signal embedded in Gaussian noise has received a great deal of attention in numerous studies. The denoising process is to separate an observed data sequence into a "meaningful" signal and a remaining noise. The choice of the denoising criterion depends on the properties of the additive noise, smoothness of the class of the underlying signal and the selected signal estimator.

The pioneer method of wavelet denoising was first formalized by Donoho and Johnstone [11]. The wavelet thresholding method removes the additive noise by eliminating the basis coefficients with small absolute value which tend to be attributed to the noise. The method assumes a prior knowledge of the variance of the additive white Gaussian noise. Hard or soft thresholds are obtained by solving a min-max problem in estimation of the expected value of the reconstruction error [12]. The suggested optimal hard threshold for the basis coefficient is of order $\sqrt{2\log N/(N)}$. The method is well adapted to approximate piecewise-smooth signals. The argument however fails for the family of signals which are not smooth, i.e., the family of signals for which the noiseless coefficients might be nonzero, very small, and comparable with the noise effects, for a large number of basis functions.

The approach to the denoising problem in [29] proposes a thresholding method for any family of basis functions. Here the attempt is to calculate the mean-square reconstruction error of the signal as a function of any given threshold. It provides heuristic estimates of such error for different families of basis functions such as wavelet and local cosine bases. The choice of the optimum threshold is given experimentally.

For the best basis search the suggestion is to compare the error estimates for different families of bases and choose the one which minimizes such criterion.

A different denoising approach is recommended by Rissanen in [43]. In each subspace of the basis functions the normalized maximum likelihood (NML) of the noisy data is considered as the description length of the data in that subspace. The Minimum description length (MDL) denoising method suggests to choose the subspace which minimizes this description length. Here noise is defined to be a part of the data that can not be compressed with the considered basis functions, while the meaningful information-bearing signal need not to be smooth. The method provides a threshold which is almost half of the suggested wavelet threshold in [11].

The new method of denoising in this thesis is based on subspace comparison rather than thresholding. We suggest to use the proposed information theoretical approach, MDC, for denoising. Our focus is not on setting a threshold for the coefficients beforehand, but to find the estimation error in each subspace separately and choose the subspace for which the error is minimized. Similar to MDL denoising no prior assumption on the smoothness of the noiseless part of the data is needed.

The thesis is organized as follows. In chapter 2 the identification and quality evaluation problem is defined. In chapter 3, the methods of order estimation, AIC, BIC and MDL are discussed. Chapter 4 proposes a new method of quality evaluation for the identification problem. In chapter 5 we introduce the new method of parametric model selection MDC. We also introduce a new minimum description length which is consistent with the notion of Kolmogorov complexity. Chapter 6 addresses the denoising problem. We provide the new method of denoising based on the information theoretic approaches introduced in chapter 5. Finally, chapter 7 is the conclusion and future work.

Chapter 2

Background: System Identification and Model Quality Evaluation

2.1 System Identification

In this section we briefly review the basic approaches to the system identification in both stochastic and deterministic settings. The following methods provide parametric estimates for the impulse response of a stable linear time invariant(LTI) system. The input of the system is assumed to be persistently existing of order N, the length of the data [32].

2.1.1 Notations and Definitions

For vectors and matrices $(\cdot)^T$ denotes transpose. For vector y and a linear subspace Z_m of order m, $\hat{y}_{Z_m} = \arg\min_{z \in Z_m} ||y - z||^2$ is the orthogonal projection of y into Z_m, where $||\cdot||$ denotes the l_2-norm. The Q function is $Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{-\frac{u^2}{2}} du$ which is the probability that an element of a normal distribution is within $|x|$ distance of zero. Finally, $O(f(N))$ is a function of N such that $\lim_{N\to\infty} \frac{O(f(N))}{f(N)} = 0$.

2.1.2 Stochastic Methods

A finite set of observation, input $u^N = [u_1, \cdots, u_N]$ and output $y^N = [y_1, \cdots, y_N]$ of a system is given. It is assumed that θ_0, an element of a parametric model set Θ, with a probability density function(PDF) $f_y(y^N|\theta_0, u^N)$, generated the data. The goal is to find an estimate of θ_0 in the set Θ by using the observed data.

Maximum a posteriori (MAP) and maximum likelihood (ML) estimators are two basic, important estimators. If by the prior assumption there is a probability distribution for θ in Θ, $f_\Theta(\theta)$, then the MAP estimate is

$$\hat{\theta}(\text{MAP}) = \arg\max_{\theta} f_\theta(\theta|y^N, u^N) \tag{2.1}$$

where $f_\theta(\theta|y^N, u^N)$ is the conditional PDF of the parameter $\theta \in \Theta$ given the observed

data $z^N = (y^N, u^N)$. If the distribution of θ is unknown or θ is not a random variable, the ML estimator is used. The ML estimate of θ_0 is

$$\hat{\theta}(\text{ML}) = \arg\max_{\theta} f_y(y^N|\theta, u^N) \tag{2.2}$$

where $f_y(y^N|\theta, u^N)$ is the conditional PDF of the output using the observed data. Note that if θ is a deterministic parameter this PDF is written in form of $f_y(y^N; \theta, u^N)$. Also note that when $f_\Theta(\theta)$ is a uniform distribution on Θ, ML and MAP estimators are the same.

Consider the noisy output of a stable LTI system

$$\begin{aligned} y_n &= \sum_{k=0}^{\infty} h(k)u_{n-k} + \sum_{k=0}^{\infty} g(k)e_{n-k} && (2.3) \\ &= H(q,\theta_0)u_n + G(q,\theta_0)e_n = H(q,\theta_0)u_n + w_n && (2.4) \end{aligned}$$

where h is the impulse response of the system, e is a sequence of independent random variable with zero mean and unit variance. The sequence $w_n = G(q,\theta_0)e_n$ represents the additive colored noise of the system. The transfer function of h is

$$H(q,\theta_0) = \sum_{k=1}^{\infty} h(k)q^{-k} \tag{2.5}$$

where θ_0 in $H(q,\theta_0)$ is a parameter which represents the impulse response h. For example, it can be the taps of the impulse response $\theta_0 = h$, or can be the zeros (and/or poles) of the system transfer function. What is the best estimator of θ_0, using the finite length data u^N and y^N?

In most identification methods for such systems the impulse response is deterministic and no prior PDF for θ is assumed. Therefore, MAP is not used for the estimation. The ML estimator is given by (2.2) for PDF

$$f_y(y^N|\theta, u^N) = \frac{1}{\sqrt{\det \Sigma_w}} e^{(y^N - H(q,\theta)u^N)^T \Sigma_w^{-1} (y^N - H(q,\theta)u^N)} \tag{2.6}$$

where Σ_w is the covariance matrix of noise w and itself is a function of $G(q,\theta)$. When the covariance matrix is a function of θ the ML estimator is obtained by solving a nonlinear optimization problem which is not trivial.

The conventional method of calculation of $\hat{\theta}$ in this setting is based on minimum prediction error(MPE) estimation. For any θ, the output predictor at point n is a function of y^{n-1}, u^n and θ, $\hat{y}_n(y^{n-1}, y^n, \theta)$. MPE method is to choose the $\hat{\theta}$ which minimizes the output prediction error,

$$\hat{\theta}(\text{MPE}) = \arg\min_{\theta} \sum_{n=1}^{N} |y_n - \hat{y}_n(y^{n-1}, u^n, \theta)|^2 \tag{2.7}$$

In [32] the output predictor $\hat{y}_n(y^{n-1}, u^n, \theta)$ is obtained based on Bayes' Least-square

method, also known as minimum mean square error(MMSE). The MMSE is computed as follows. For each θ find the output predictor of $\hat{y}_n$ such that the expected value of the least square error is minimized,

$$\hat{y}_n(y^{n-1}, u^n, \theta) = \arg\min_v \int_{x=-\infty}^{\infty} (v-x)^2 f_{y_n}(x|y^{n-1}, u^n, \theta) dx \tag{2.8}$$

where $f_{y_n}(x|y^{n-1}, u^n, \theta)$ is the conditional PDF of y_n given y^{n-1}, u^n and θ. The solution to this minimization is

$$\hat{y}_n(\text{BLS}) = E(Y_n|y^{n-1}, u^n, \theta), \tag{2.9}$$

where $E(Y_n|y^{n-1}, u^n, \theta)$ is the expected value of y_n given y^{n-1}, u^n, θ. For this setting, using (2.4),

$$\hat{y}_n(\text{BLS}) = G^{-1}(q,\theta)H(q,\theta)u_n + (1 - G^{-1}(q,\theta))y_n. \tag{2.10}$$

Note that the transfer function $G(q,\theta)$ is assumed to be inversely stable and monic[1]

The MPE estimator, in (2.7), using BLS output predictor is

$$\hat{\theta}(\text{MPE}) = \arg\min_\theta \sum_{n=1}^{N} (y_n - \hat{y}_n(BLS))^2. \tag{2.11}$$

As it is mentioned before, when $G(q,\theta) \neq 1$ the ML estimator is the solution of a nonlinear optimization. However, when $G(q,\theta) = 1$ and θ_0 is the taps of the impulse response h,

$$\hat{y}(\text{BLS}) = \sum_{k=0}^{\infty} \theta(k) u_{n-k} \tag{2.12}$$

and the ML and MPE estimators are the same.

In [32] and [49] the asymptotic behavior of MPE is investigated. It is shown that as N grows the estimate of the impulse response, $\hat{h}^N$, approaches h with probability one and the random variable $h - \hat{h}^N$ can be approximated by a "zero-mean" random variable with a normal distribution,

$$\sqrt{N}(h - \hat{h}^N) \approx N(0, P_h). \tag{2.13}$$

where P_h is the covariance matrix.

[1] A transfer function of the impulse response g is monic if $g(i) = 0$, for $i < 0$ and $g(0) = 1$. Therefore, the first element of impulse response of $G^{-1}(q,\theta)$ is one. As a result for $L(q,\theta) = 1 - G^{-1}(q,\theta)$, $l(0)$ is zero, and in calculation of $(1 - G^{-1}(q,\theta))y_n$ only y^{n-1}, and not y_n, is needed.

2.1.3 Deterministic Methods

Deterministic system identification methods avoid stochastic assumption on the additive noise in favor of non-stochastic magnitude bounded disturbance. This assumption led to various methods of set-membership identification, worst-case identification, H_∞ identification, l_1 identification, [22, 53, 27], *etc.*

The basic idea in these identification methods is to use the observed data, a parameterized model set and with a prior assumption on the bounded magnitude error, find a region in the parameter space that is compatible with the prior assumption.

One example of set-membership identification in H_∞ is as follows [18]. Assume that the input-output relation is given by

$$y_n = H(q, \theta_0)u_n + w_n \tag{2.14}$$

where $H(q, \theta_0)$ is the transfer function of the impulse response of the system h, $\theta_0 \in \Theta$, and $[w_1, \cdots, w_N]$ is an element of set W

$$W = \{v| \ \frac{1}{\sqrt{N}}||v||_2 \leq \epsilon\}. \tag{2.15}$$

Given the finite length input and output of the system, u^N, y^N, the acceptable region in the parameter space, feasible parametric set(FPS), is

$$FPS = \{\theta \in \Theta| \ \ ||y^N - H(q, \theta)u^N||_2 \leq \sqrt{N}\epsilon\}. \tag{2.16}$$

Also an estimator Φ provides an estimate of h using the finite data, $\Phi(u^N, y^N, \epsilon) = \hat{h}^N$. If for such estimator the following property holds

$$\lim_{N\to\infty, \epsilon\to 0} \sup ||H(q, \theta_0) - \hat{H}^N(q)||_\infty \to 0, \quad \text{for all } \theta_0 \in \Theta \tag{2.17}$$

the algorithm is called *robustly convergent* [22]. The optimality of the least-square estimator for H_∞ identification is investigated in [51].

An example of worst-case identification is given in [53]. It investigates the asymptotic results in l_1 identification. The prior assumption is that h is the finite impulse response of an LTI system and $w \in W$, $W = \{v| \ ||v||_\infty \leq \epsilon\}$. In this scenario [53] proves that there exists an optimal algorithm Φ, $\Phi(u^N, y^N, \epsilon) = \hat{h}^N$, such that

$$\epsilon \leq \lim_{N\to\infty} \sup ||h - \hat{h}^N||_1 \leq 2\epsilon, \quad \text{for FIR filters.} \tag{2.18}$$

2.2 Model Quality Evaluation in System Identification

The major motivation for deterministic identification methods is to provide an estimate of the system and also provide bounds on the accuracy of the estimate. Conventional system identification approach claims that the answer to this question is in

the estimate of the variance of the impulse response estimate in (2.13). The argument is that the region of confidence, provided by this variance, plays the same role as the feasible parametric set in deterministic identification (2.16). However, this argument in general is not valid. Why? The answer lies on the source of disturbance. Here the prior assumption on the model set plays an important role. If the true model is an element of the model set, the argument given above is correct. If the true model is not an element of the model set, the unmodeled dynamics effects always can be considered as a bounded disturbance, however, it can not be considered as additive stochastic noise which is independent of input. While deterministic identification captures such scenario, the stochastic identification fails to address this case. Therefore, to compare the estimate of the true model in competing model sets, it is important to separate the effects of additive noise and the effects of unmodeled dynamics. In [26] the results of model quality evaluation for some deterministic and stochastic settings is compared.

Here we review some important existing quality evaluation methods. The first method, which is proposed by the conventional system identification approach, is the bias-variance tradeoff method. The second identification method is proposed by Venkatesh [55]. The attempt is to estimate a system with unknown dimension in a parametric model set. The prior probabilistic assumption on the unmodeled dynamics is suggested in [20] and we briefly review this method. There are set-membership identification methods which consider the unmodeled dynamics effects of the system in the parametric identification. Examples of such methods of quality evaluation are in [18, 18, 61, 5, 16]. Here we discuss the most recent of these methods.

2.2.1 Bias-Variance Tradeoff

Following the conventional system identification method, reviewed in section 2.1.2, [33] suggests a method of calculating the quality of the estimated model.

Problem Statement

The collected data $y^N = (y_1, \cdots, y_N)$ and $u^N = (u_1, \cdots, u_N)$ of system in (2.4) is available. The posterior information in terms of the frequency function $H(e^{j\omega})$ is given either with a fixed known hard bound

$$|\hat{H}_m^N(e^{j\omega}) - H(e^{j\omega})|^2 \leq W(e^{j\omega}), \tag{2.19}$$

or with a probabilistic bound

$$\hat{H}_m^N(e^{j\omega}) - H(e^{j\omega}) \in N(0, P(e^{j\omega})), \tag{2.20}$$

where $\hat{H}_m^N(e^{j\omega})$ is the estimate of the system impulse response in the parametric model set of order m, S_m. Here $N(\mu, P)$ is a normal distribution with mean μ and variance P.

Quality Evaluation

To search for the most powerful unfalsified model in a model set of order m, [33] suggests to find a nominal model $\hat{H}_m^N(e^{j\omega})$ which minimizes $\max_w W_m(e^{j\omega})$ in (2.19) for when the hard bound assumption is considered, or to minimize the average variance in (2.20) for the probabilistic bound. For the probabilistic prior assumption the method is equivalent to using the mean square error(MSE) criterion, i.e., minimizing $J(S_m)$,

$$\begin{aligned} J(S_m) &= \int_{-\pi}^{\pi} P_m(e^{j\omega}) d\omega \\ P_m(e^{j\omega}) &= E|\hat{H}_m^N(e^{j\omega}) - H(e^{j\omega})|^2. \end{aligned} \tag{2.21}$$

In this case the MSE can be split into two terms, bias and variance contributions

$$J(S_m) = J_B(S_m) + J_V(S_m) \tag{2.22}$$

where the bias contribution is $J_B(S_m)$ and the variance is $J_V(S_m)$. From [33]

$$J_V(S_m) \approx \frac{m}{N}. \tag{2.23}$$

(It seems that here the variance of the noise is assumed to be one and no information about the input power is given). In [33] it is claimed that in most cases the best trade-off is obtained when the conflicting contributions are of the same magnitude

$$J_B(S_m) \approx J_V(S_m). \tag{2.24}$$

As a result, three methods for choosing the order of the system are suggested

- *Cross Validation*: Evaluate the criterion (2.21) (without the expectation E) on a fresh set of data for each model $\hat{H}_m^N(e^{j\omega})$. The role of H is then played by the new, observed data (typically $J(S_m)$ is the MSE between the model's output and the observed one). The role of "E" is played by the fact that a fresh data is used. Then pick order m that minimizes this estimate of $J(S_m)$.

- *Penalty Criteria*: Suppose that the criterion $J(S_m)$ in fact is the expected value of the squared prediction errors. Then the decrease in $J(S_m)$ due to $J_B(S_m)$ can be measured by $J(S_m)$ itself, when evaluated on the data set that produced the model. The contribution from $J_v(S_m)$ must then be computed and added artificially, using (2.23), as a model complexity penalty. This leads to criteria of the Akaike-Rissanen[2, 38] type

$$\min_m \min_\theta \sum_1^N \epsilon^2(n,\theta) + \frac{f(m)}{N} \tag{2.25}$$

  where

$$\epsilon(n,\theta) = y_n - \hat{y}_n(\theta) \tag{2.26}$$

and $\hat{y}_n(\theta)$ is the estimate of output using input and parametric model with parameter θ.

- *Hypothesis Tests, Model Validation*: The most traditional statistical approach is to start with small m, let the obtained model go through *model validation*, i.e., try to falsify the hypothesis that the data have been generated by an mth order model. The model order is then increased until we find a model that cannot be falsified. The most common tests are constructed so that they are testing whether a possible contribution is significantly larger than the variance contribution.

2.2.2 Minimize-Unmodeled Dynamics Principle(MUDP)

Minimize-unmodeled dynamics(MUD) principle is proposed for parametric identification of a system with possibly infinite dimension [54].

Problem Statement

Finite, N, points of a noisy output of a linear system, T, with input u is given

$$y_n = Tu + w_n = \sum_{1}^{m} h(k)u_{n-k} + \sum_{m+1}^{n} h(k)u_{n-k} + w_n, \tag{2.27}$$

where noise w belongs to W_N

$$W_N = \{w \in R^N \mid \sup_{q \in Q^l} |\frac{1}{\sqrt{N}\log(N)} \sum_{k=1}^{N} w_k e^{jq_k}| \leq 1\}, \tag{2.28}$$

and Q^l is the class of polynomials in N of order l over field of reals[2]. For each N the estimate of the system, $\hat{h}_m^N$, in a subspace of order m is obtained using the least square method

$$\hat{h}_m^N = \arg\min_{g \in S_m} ||y - y(g)||^2 \tag{2.29}$$

where $y_n(g) = \sum_{k=1}^{m} g(k)u_{n-k}$ and S_m is the set of all FIR filters of length m. Define h_m as follows

$$h_m = \arg\min_{g \in S_m} ||h - g||_1 \tag{2.30}$$

where $||X||_1 = \sum |x_i|$ is the l_1 norm of X. The prior knowledge in this setting is

$$||h - h_m||_1 \leq \gamma, \tag{2.31}$$

[2]in the definition of W, j is $\sqrt{-1}$

for a given γ. The input of the system is

$$u_n = \exp(j\alpha n^2), \quad \alpha \in R, \tag{2.32}$$

which satisfies the following property

$$\max_{0\leq\tau\leq N} |r_u^N(\tau)| \leq L(\alpha)\frac{\log(N)}{N}, L(\alpha) > 0 \tag{2.33}$$

for almost all $\alpha > 0$ except for a set of Lebesgue zero measure where $r_u^N(\tau) = \frac{1}{N}\sum_1^N u(\tau+i)u(i)$. An upper bound on the rate of convergence of $\hat{h}_m^N$ to h_m is provided as follows.

Rate of Convergence of the Error In subspace of order m,

$$\hat{h}_m^N - h_m = ((A_m(N))^T A_m(N))^{-1}(A_m(N))^T(B_m(N)\Delta_m^N + w^N), \tag{2.34}$$

where $h_m^N = [h(1), \cdots, h(m)]^T$, $\Delta_m^N = [h(m+1), \cdots, h(N)]^T$, $A_m(N)$ is a $N \times m$ matrix, $B_m(N)$ is a $N \times N-m$ matrix, $A_m(N)$ is the first m columns of the Toeplits matrix U_N and $B_m(N)$ is the last $N-m$ columns of U_N

$$U_N = \begin{bmatrix} u_1 & 0 & \cdots & \cdots \\ u_2 & u_1 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ u_N & \cdots & \cdots & u_1 \end{bmatrix}. \tag{2.35}$$

The prior information in (2.31) is

$$||\Delta_m^N||_1 \leq \gamma. \tag{2.36}$$

Therefore, we have

$$\begin{aligned} ||\hat{h}_m^N - h_m||_1 &= ||((A_m(N))^T A_m(N))^{-1}(A_m(N))^T(B_m(N)\Delta_m^N + w^N)||_1 & (2.37) \\ &\leq C(||\frac{1}{N}A_m(N))^T(B_m(N)\Delta_m^N||_1 + ||\frac{1}{N}(A_m(N))^T w||_1), & (2.38) \end{aligned}$$

where for a given N_0,

$$C = \sup_{N\geq N_0} ||(\frac{1}{N}A_m(N))^T A_m(N))^{-1}||_\infty. \tag{2.39}$$

Using the prior information (2.36) and the input property in (2.33) for the $m \times 1$ vector $\frac{1}{N}A_m(N))^T(B_m(N)\Delta_m^N = [v_1, \cdots, v_m]^T$, we have

$$||\frac{1}{N}(A_m(N))^T B_m(N)\Delta_m^N||_1 = ||[v_1, \cdots, v_m]^T||_1 \leq m \max_{1\leq i\leq m} |v_i| \tag{2.40}$$

$$\leq \quad m\gamma \max_{m\leq\tau\leq N} |r_u^N(\tau)| \leq m\gamma L(\alpha)\frac{log(N)}{N}. \quad (2.41)$$

Also using the noise property in (2.28) with $q[k] = k^2$, for the $m \times 1$ vector $||\frac{1}{N}(A_m(N))^T w||_1 = ||[z_1, \cdots, z_m]^T||$,

$$||\frac{1}{N}(A_m(N))^T w||_1 \quad = \quad ||[z_1, \cdots, z_m]^T||_1 \leq m \max_{1\leq i\leq m} |z_i| \quad (2.42)$$

$$\leq \quad m\frac{log(N)}{N}. \quad (2.43)$$

Finally, from (2.43) and (2.41) the upper bound for the subspace impulse error in (2.38) is

$$||\hat{h}_m^N - h_m||_1 \leq Cm(\gamma L(\alpha) + 1)\frac{log(N)}{N} \quad (2.44)$$

2.2.3 Set-membership Identification

H_∞ Identification

We follow an example of deterministic approach to the quality evaluation in system identification from [18]. Assume that the impulse response of the LIT system can be written in form of

$$h = \sum_{i=1}^{m} p_i F_i + \delta_m, \quad (2.45)$$

where F_is are orthonormal basis functions and p_i's are the parametric representation of h in a model set of order m, S_m. For simplicity of presentation we follow the method with F_i being a vector for which all its elements are zero except the ith element which is 1, therefore, $p_i = h(i)$. The input-output relationship is as follows

$$y_n = \sum_{1}^{m} h(k)u_{n-k} + \sum_{m+1}^{n} h(k)u_{n-k} + w_n, \quad (2.46)$$

and finite length data, input and output of the system, is available. The assumptions are that input has power one, and noise, w, has a bounded power less than ϵ. In each subspace S_m, an identification algorithm is an operator Φ_m which maps the available data to an estimate $\hat{h}_m = \Phi(y^N, u^N)$.

Other important assumption is on the H_∞ norm of Δ_m

$$||\Delta_m||_{H_\infty} \leq \epsilon_m \quad (2.47)$$

where Δ_m is the transfer function of δ_m in (2.45). In each subspace S_m define the

feasible parameter set(FPS) as

$$FPS(S_m) = \{g \in S_m, \frac{1}{N}||y^N - y^N(g)||_2 \leq \epsilon + \epsilon_m\} \tag{2.48}$$

where $y^N(g) = [y_1(g), \cdots, y_N(g)]$, $y_n(g) = \sum_1^m g(k)u_{n-k}$. The proposed criterion for the quality evaluation of estimation in each model set is

$$J(S_m) = \sup_{g \in FPS} \sup_{||\Delta_m||_\infty \leq \epsilon_m} ||G + \Delta_m - \hat{H}_m||_\infty \tag{2.49}$$

where H is the transfer function of h, and G is the transfer function of g.

The method suggests to choose the model set which minimizes $J(S_m)$.

Bounds on the Unmodeled Dynamics Induced Norm

Comparison of the FPS in (2.16) and (2.48) shows that while the norm of the additive noise is upperbounded by ϵ for exact modeling, here the effects of both noise and unmodeled dynamics is upperbounded by $\epsilon + \epsilon_m$, for each model class S_m. The proposed algorithm is sensitive to the choice of upper bound ϵ_m in (2.47).

In [18] it is claimed that an operator Φ for which the minimum of $J(S_m)$ is attained is the least square method (optimality of least-square method). With this estimator the following method of estimation of ϵ_m is suggested. In each subspace S_m find

$$|\bar{\Delta}_m(\omega)| = \sup_{\delta_m \in FES} |\Delta_m(\omega)|, |\underline{\Delta}_m(\omega)| = \inf_{\delta_m \in FES} |\Delta_m(\omega)|, \tag{2.50}$$

where the feasible error set(FES) is

$$FES = \{\delta_m, ||y^N - y^N(\hat{h}_m) - u * \delta_m||_2 \leq \sqrt{N}\epsilon\}. \tag{2.51}$$

Then

$$\sup_{0 \leq \omega \pi} |\underline{\Delta}_m(\omega)| \leq \epsilon_m \leq \sup_{0 \leq \omega \pi} |\bar{\Delta}_m(\omega)| \tag{2.52}$$

It is then suggested to calculate the estimate of ϵ_m as follows

$$\hat{\epsilon}_m = \sup_{0 \leq \omega \leq 2\pi} |\hat{\Delta}_m(\omega)| \tag{2.53}$$

where $\hat{\Delta}_m = \frac{1}{2}[|\bar{\Delta}_m| + |\underline{\Delta}_m|]$.

H_2 Identification

In [16] the prior assumption is that the disturbance sequence $w^N = [w_1, \cdots, w_N]^T$ in

$$y_n = \sum_1^\infty h(k)u_{n-k} + w_n, \tag{2.54}$$

belongs to W_ϵ

$$W_\epsilon = \{e^N : \frac{1}{N}||e^N||_2 \leq \epsilon\}, \tag{2.55}$$

and a priori information on the system is expressed as follows

$$K = \{h : ||h||_{2,\rho} \leq L, L > 0, 0 < \rho < 1\}, \tag{2.56}$$

where

$$||h||_{2,\rho} = \sqrt{\sum h(i)^2 \rho^{-2i}}. \tag{2.57}$$

An identification mapping $\psi(y^N)$ provides an approximation of h based on the observed data. The *global worse case identification error* $E(\psi, \epsilon)$ is defined as

$$E(\psi, \epsilon) = \sup_{h \in K} \sup_{e \in W_\epsilon} ||h - \psi(y^N)||^2. \tag{2.58}$$

Let Ψ_{S_m} be the set of all admissible identification algorithms ψ which maps the data to an element of set S_m where S_m is a m-dimensional subspace of the impulse response space. The conditional radius of information is defined as

$$r(S_m, \epsilon) = \inf_{\psi \in \Psi_{S_m}} E(\psi_m, \epsilon) \tag{2.59}$$

The goal is to find an estimate of this criterion for comparison of different model sets. [16] provides upper and lower bounds on $r(S_m, \epsilon)$ for any class of models linear in parameters for identification of exponentially stable systems. However, no results on convergence of the bounds can be provided and in the simulation results the bounds for model sets of different order are monotonically decreasing or increasing as the order of the model set grows.

2.2.4 Probabilistic Prior Assumption on the Unmodeled Dynamics

In [20] a probabilistic assumption on the unmodeled dynamics is chosen in attempt to provide a new method of quality evaluation in identification.

Problem Statement The output of the system is in form

$$y_n = Tu + w = \Phi_n^T \theta_0 + \Psi_n^T \eta + w \tag{2.60}$$

where the additive noise is a Gaussian random variable, an element of $N(0, \sigma_w^2)$ and

$$T(e^{-j\omega}) = H(e^{-j\omega}, \theta_0) + \Delta(e^{-j\omega}) \tag{2.61}$$

with assumption that $\Delta(e^{-j\omega})$ is a zero mean stochastic process

$$E(\Delta(e^{-j\omega})) = 0 \tag{2.62}$$

$$\Delta(q) = \sum_{k=1}^{L} \eta_k q^{-k} \tag{2.63}$$

where η_k is zero mean with variance

$$E(\eta_k^2) = \alpha\lambda^k \tag{2.64}$$

and $L \leq N$. The prediction model $H(q,\theta)$ is a member of the model set

$$S_m = \{H(q,\theta) : \theta \in R^m\} \tag{2.65}$$

The estimate of θ_0, $\hat{\theta}_m^N$ is obtained by using the least-square method and the goal is to estimate

$$V_{S_m}(\omega) = E\left(|H_T(e^{-j\omega}) - H(e^{-j\omega}, \hat{\theta}_m^N)|^2\right) \tag{2.66}$$

for model sets of different order S_m.

Solution It is shown that to estimate (2.66), the estimate of $\xi = (\alpha, \lambda, \sigma_w)$ is needed. Consider

$$W_{S_m} = R^T\epsilon = R_{S_m}^T(y^N - \Phi_{S_m}\hat{\theta}) \tag{2.67}$$

where R is any matrix whose columns span the subspace orthogonal to columns of Φ and ϵ is the output error. The estimate of $\xi = (\alpha, \lambda, \sigma_w)$ is obtained using the output error as following

$$\hat{\xi} = \arg\max_{\xi} L(W_{S_m}|U,\xi) \tag{2.68}$$

where $L(W_{S_m}|U,\xi)$ is the likelihood function

$$L(W_{S_m}|U,\xi) = -\frac{1}{2}\ln\det\Sigma_\eta - \frac{1}{2}W_{S_m}^T\Sigma_\eta^{-1}W_{S_m} + \text{const} \tag{2.69}$$

and

$$\Sigma_\eta = R^T\Psi C_\eta(\alpha,\lambda)\Psi^T R + \sigma_w^2 R^T R \tag{2.70}$$

$$C_\eta(\alpha,\lambda) = \text{diag}(\alpha\lambda, \alpha\lambda^2, \cdots, \alpha\lambda^L). \tag{2.71}$$

The suggested method of comparing estimates in different model classes is to compare

$$J(S_m) = \int_{-\pi}^{\pi} V_{S_m}(\omega)d\omega \tag{2.72}$$

for the model classes.

2.3 Discussion

In this section we reviewed some basic system identification methods for stable LTI systems. Here we discuss the model quality evaluation methods presented in this section. Some drawbacks of these identification methods, from a practical point of view and for robust control purposes, are discussed in [34]. Although some of the methods can be used in practical problems for a subclass of stable LTI systems, here we discuss some of the shortcomings in generalizing their application.

Bias-Variance Tradeoff

In the stochastic approaches with the finite length data $\hat{h}_m^N$, an estimate in each model set S_m, is obtained. The estimator variance is such that as N grows it becomes smaller and eventually zero. Therefore with probability one, and for some $\bar{h}_m$, we have

$$\lim_{N\to\infty} \hat{h}_m^N \to \bar{h}_m \tag{2.73}$$

The error in estimation of impulse response h is

$$h - \hat{h}_m^N = (h - \bar{h}_m) + (\bar{h}_m - \hat{h}_m^N) \tag{2.74}$$

In bias-variance method the hard band on the norm of frequency response of this error $W(e^{j\omega})$ in (2.19) is given as a prior and no method of estimation of such bound is given. The probabilistic assumption on such error is given in (2.20). The main assumption is that such error is zero mean. Considering the structure of error in (2.74), the expected value of this error can be zero only if as N grows

$$E(\bar{h}_m - \hat{h}_m^N) = 0 \tag{2.75}$$

$$h - \bar{h}_m = 0 \tag{2.76}$$

The second assumption implies that the true system is an element of the model set S_m. It seems that the correct prior assumption is to consider that only the subspace error $\bar{h}_m - \hat{h}_m^N$ has a zero mean asymptotically. In this case the bias-variance method observes the behavior of the variance of $\bar{h}_m - \hat{h}_m^N$, not of the total error $h - \hat{h}_m^N$. The variance of the error $\bar{h}_m - \hat{h}_m^N$, has two elements, a noise effect, called variance, and effects of the unmodeled dynamics, $h - \bar{h}_m$, called bias term.

Note that in off-line identification, the error $\bar{h}_m - \hat{h}_m^N$ can be estimated using several samples of output obtained from several input sets. However, in comparison of model sets, the important error is $h - \hat{h}_m^N$ which this bias-variance method can not estimate.

Minimize-Unmodeled Dynamics Principle

Similar to the bias-variance method this method focuses on the estimate of the subspace error $\bar{h}_m - \hat{h}_m^N$. The main difference is that here an upper bound for the bias term is provided based on a prior assumption on the norm of the unmodeled dynamics. However, since the upper bound , γ, is the same for all the model sets, the upper bound for error increases as the order of the model set increases. Therefore the provided upper bound can not be used as a criterion for comparison of the estimates in different model sets.

Another issue is on the choice of $\log(N)$ in defining the input and noise in (2.33) and (2.28). Instead of $\log(N)$ in these definitions it can be shown that any function $\beta(N)$, for which $\lim_{\to\infty} \beta(N) = \infty$ and $\lim_{\to\infty} \frac{\beta(N)}{N} = 0$, can be used. Then the noise definition is

$$W_N(\beta_N) = \{w \in R^n \mid \sup_{q \in Q^m} |\frac{1}{\sqrt{N}\beta(N)} \sum_{k=1}^{N} w[k] e^{jq[k]}| \leq 1\}. \tag{2.77}$$

Therefore the rate of convergence of the error in (2.44) can be generalized to

$$||\hat{h}_m^N - h_m||_1 \leq Cm(\gamma L(\alpha) + 1)\frac{\beta(N)}{N} \tag{2.78}$$

Set membership Identification

Similar to the definition of "robustly convergence" given in (2.17), we define the robustly convergent property as

$$\lim_{N\to\infty, \epsilon\to 0} \sup ||H(q, \theta_0) - \hat{H}^N(q)||_\infty \to ||\Delta_m||_\infty \tag{2.79}$$

None of the available H_∞ or H_2 identification methods can prove whether the method is robustly convergent. It is not possible to provide any rate of convergence for the estimate of $||\Delta_m^N||$ as N grows. One of the main factors causing this problem is the definition of noise. Boundedness of the norm of the noise is a very conservative prior assumption and it ignores other properties of the additive noise such as independence from the input. Even the correlation of noise with itself provides some advantages in the stochastic approach that enables us to prove the variance of the estimate approaching zero with rate $\frac{1}{N}$. Therefore, even if an estimate of unmodeled dynamics norm is available, the noise effects limits the performance of the error estimators similar to what is discussed for the worst-case l_1 identification in (2.18).

Probabilistic Prior Assumption on the Unmodeled Dynamics

In this setting not only the additive noise but also the unmodeled dynamics is probabilistic. Unlike the bias-variance method the goal is to estimate the error between the true model and the estimates in each model class, $h - \hat{h}_m^N$. However, the main draw-

back is in the structural and probabilistic assumption on the unmodeled dynamics

$$\Delta(q) = \sum_{1}^{L} \eta_k q^{-k} \tag{2.80}$$

With this structure, the unmodeled dynamics and the estimate of the model are not orthogonal. Therefore, even in noiseless case, distinguishing the overlap between $\hat{h}_m^N(q)u$ and $\Delta(q)u$ is not possible.

On the other hand the prior assumption that $E(\eta_k)$ is zero discards the bias term in $E(h - \hat{h}_m^N)$ and the unmodeled dynamics is more of a kind of additive noise with structure different than the additive white Gaussian noise. Because of the structure of the unmodeled dynamics, the algorithm is very sensitive to the choice of L the length of the unmodeled dynamics impulse response in (2.80. The algorithm estimates the variance of noise and the unmodeled dynamics, $E(\eta_k^2) = \alpha\lambda^k$. Using the likelihood function, the estimates of $\xi = (\alpha, \lambda, \sigma_w)$ changes as L changes.

Conclusion

We summarize the quality evaluation problem for the competing model sets with the following final words. The first N taps of the impulse response is an element of R^N and in all the discussed methods, $\hat{h}_m^N$, the estimate of this parameter in S_m, a subset of R^N, is calculated. We believe that the important criterion for comparison of these model sets is the impulse response error

$$h - \hat{h}_m^N \tag{2.81}$$

in different model sets. The prior assumption on the structure of S_m plays an important role in calculation of this error. The only discussed method in this chapter which focuses on calculation of this error is the last method with probabilistic assumption on the unmodeled dynamics and all the other methods fail to calculate this error. However, the calculation of the error becomes ad-hoc due to the prior assumption on the structure of S_m and the unmodeled dynamics. If S_m is a subspace of R^N, then as figure 2-1 shows the unmodeled dynamics of h, Δ_m^N, is a member of R^N which is not in subspace S_m. $\bar{h}_m$ is defined as

$$\bar{h}_m = \lim_{N\to\infty} \hat{h}_m^N. \tag{2.82}$$

In MUDP method an upper bound on the subspace impulse response error(SIRE)

$$\bar{h}_m - \hat{h}_m^N \tag{2.83}$$

is provided based on a prior assumption on the l_2 norm of the unmodeled dynamics, $||\Delta_m^N|| \leq \gamma$. As a result the provided upperbound on the error is monotonically increasing as the dimension of subspace S_m grows. Note that while this error(SIRE) asymptotically is zero, if h is not an element of S_m the IRE does not approach zero. Our goal in the following chapters is to find an estimate of the l_2 norm of the

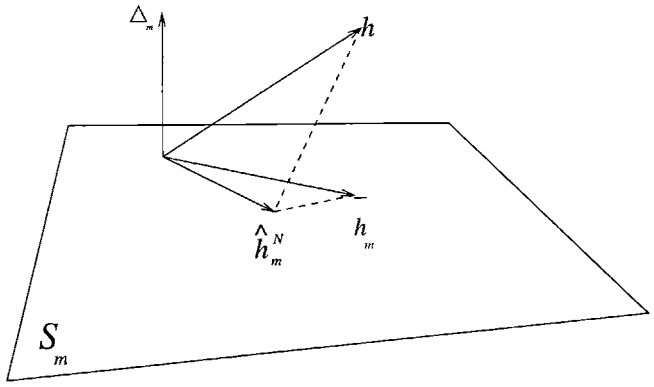

Figure 2-1: *Impulse response estimate in subspace* S_m.

unmodeled dynamics using the observed data. We aim to provide an estimate not only for subspace impulse response error in 2.83 but also for the impulse response error in 2.81 as a criterion for the quality evaluation. Our prior assumption on the structure of the model set and structure of the unmodeled dynamics is similar to assumptions of the MUDP approach. We also plan to search for a rich deterministic noise definition to provide a deterministic identification method which is robustly convergent.

Chapter 3

Order Estimation Methods

An observed data is produced by an unknown parametric model set. The problem of selecting an appropriate model set among parametric model sets with different dimension to fit this data is of interest. This classical order estimation problem has been studied extensively, and a variety of methods have been developed in this area. If model sets are nested, the maximum likelihood(ML) principle always leads to the choice of the model set with the highest possible dimension.

In this chapter we review three important information theoretic approaches to the order estimation problem: Akaike Information Criteria (AIC), Bayesian Information Criteria(BIC), and minimum description length (MDL). All these methods use the ML estimate of the parametric model in each model set. However, for the purpose of model set comparison, they implement criteria different from the ML principle.

3.1 Model Validity Criterion

Akaike extends the Maximum likelihood (ML) principle to a pioneering method of model order selection, Akaike Information Criterion (AIC) [2].

Let $z^N = (u^N, y^N)$ denote the observed data, where $u^N = (u_1, \cdots, u_N)$ and $y^N = (y_1, \cdots, y_N)$ are the input and output of a parametric model. The observed data is an element of a random variable $Z^N = (u^N, Y^N)$. Note that u^N in the following arguments is considered to be a deterministic signal. However, in general u^N itself can be an element of a random variable U^N. It is assumed that the data is generated by a parametric model set with parameter θ. The goal is to pick an element of competing parametric model sets of different order, as the best model which represents the given data. For any parameter γ in the competing model sets the distribution of the output is defined with $f(Y^N; u^N, \gamma)$ and the output is produced by the unknown parameter $\gamma = \theta$. A competing model set with order m is called S_m.

A cost function $V_m^N(\theta_{S_m}, z^N)$ is considered, where $\theta_{S_m} \in S_m$. The estimate of the true parameter θ in subspace S_m, $\hat{\theta}_{S_m}^N$, is obtained by minimizing the cost function

$$\hat{\theta}_m^N(z^N) = \arg\min_{\theta_{S_m}} V_m^N(\theta_{S_m}, z^N). \tag{3.1}$$

When the cost function is $V_m^N(\theta_{S_m}, z^N) = \frac{1}{N}$(log likelihood function), where likelihood function is the inverse of probability distribution $f(y^N; u^N, \theta_{S_m})$, the estimate is the maximum likelihood (ML) estimate.

For some families of cost functions, Akaike proposes to compare the model sets of different order by comparing the expected value of the estimate of the cost function of $\hat{\theta}_m^N$ in each subspace asymptotically, $\lim_{N\to\infty} V_m^N(\hat{\theta}_m^N, z^N)$. Define $\bar{V}_m^N(\theta_{S_m})$ as

$$\bar{V}_m^N(\theta_{S_m}) = \lim_{N\to\infty} EV_m^N(\theta_{S_m}, Z^N) \tag{3.2}$$

where $EV_m^N(\theta_{S_m}, z^N)$ is the expected value of $V_m^N(\theta_{S_m}, z^N)$ using the true probability distribution of the random variable $f(Y^N; u^N, \theta)$.

The criterion for comparison of the model sets is $E\bar{V}_m(\hat{\theta}_m^N)$ over all possible y^N, which evaluates the cost function "on the average" for each model set S_m. The estimate is provided as

$$J(S_m) = E\bar{V}_m^N(\hat{\theta}_m^N) \approx EV_m^N(\hat{\theta}_m^N, Z^N) + tr((\bar{V}_m^N)''(\theta_m^*)\frac{P_{\theta_m^*}}{N}) \tag{3.3}$$

where $\theta_m^* = \arg\min_{\theta_{S_m}} \bar{V}_m$, $(\bar{V}_m^N)''$ is the second derivative of $\bar{V}_m$, $P_{\theta_m^*}$ is the covariance matrix of for the following random variable $\sqrt{N}(\hat{\theta}_m^N - \theta_m^*) \to N(0, P_{\theta_m^*})$ and $tr(A)$ is the trace of matrix A. Details of the estimation is in appendix A.1.

The criterion $J(S_m)$ is called the model validity criterion [32]. Akaike calculates this criterion for two different cost functions, maximum likelihood and least square error. These criteria are called AIC and final prediction error (FPE) respectively. We briefly review calculation of these criteria.

3.1.1 Akaike's Information Theoretic Criterion(AIC)

AIC is the validity criterion, in (3.3), when the cost function is $V_m^N(\theta_m, Z^N) = \frac{1}{N}$(log likelihood function). For this cost function provided that

- The true system is in the model class of order m, i.e., $\theta_m^* = \theta$ (strong assumption),
- $(\bar{V}_m^N)''(\theta)$ is invertible,

it can be shown that $J(S_m)$, in (3.3), is

$$J(S_m) \approx -\frac{1}{N} L_N(\hat{\theta}_m^N, z^N) + \frac{m}{N}. \tag{3.4}$$

where $L_N(\hat{\theta}_m^N, Z^N) = \frac{1}{N} \log f(Y^N; u^N, \hat{\theta}_m^N)$ is the log likelihood function.

Since θ is an element of S_m, $\theta_m^* = \theta$ and by using the Cramer-Rao inequality [1] for calculation of $J(S_m)$ in (3.3) we have

$$tr(\bar{V}_m''(\theta)\frac{P_\theta}{N}) \approx tr(\bar{V}_m''(\theta)\frac{(P^*)^{-1}}{N}) = \frac{1}{N}tr(\bar{V}_m''(\theta)(\bar{V}_m''(\theta))^{-1}) = \frac{m}{N} \tag{3.6}$$

where $P^* = -E\frac{d^2}{d\theta^2}\log f_y(\theta; Y^N)|_\theta = (\bar{V}_m''(\theta))^{-1}$.

The AIC criterion can be viewed as the expected value of the Kullback-Leibler distance of the true probability distribution and the estimated probability distribution. Kullback-Leibler distance of two pdfs $f_1(z)$ and $f_2(z)$, of random variable z, when f_1 is the true pdf is

$$E\log\frac{f_1}{f_2} = \int f_1(z)\log\frac{f_1(z)}{f_2(z)}dz. \tag{3.7}$$

It can be shown that the information distance between the true pdf $f(y^N; u^N, \theta)$ and $f(y^N; u^N, \hat{\theta}_m^N)$ is $H(\theta) + E\bar{V}_m(\hat{\theta}_m^N)$ asymptotically, where $H(\theta)$ is the entropy of the true system output, $H(\theta) = E\log f(y^N; u^N, \theta)$. Since $H(\theta)$ is a fixed number, comparison of the expected value of the Kullback-Leibler distance for different model sets is equivalent to comparison of $E\bar{V}_m(\hat{\theta}_m^N)$ which is the AIC.

AIC for Gaussian Innovations Consider the ARMAX model

$$y_n + a_1y_{n-1} + \cdots + a_py_{n-p} = b_0u_n + \cdots + b_qu_{n-q} + e_n + c_1e_{n-1} + \cdots + c_le_{n-l} \tag{3.8}$$

where e_n is the additive white Gaussian noise(AWGN) with unknown variance σ^2. Therefore, the dimension of the parameter is $k = p + q + l + 1$ and

$$\theta = (a_1, \cdots, a_p, b_0, \cdots, b_q, c_1, \cdots, c_l). \tag{3.9}$$

The ML function is

$$V_m^N(\theta_m, z^N) = -\frac{1}{N}\log\frac{1}{(\sqrt{2\pi\sigma^2})^N}e^{-\frac{\sum_n \epsilon^2(n,\theta_m)}{2\sigma}} \tag{3.10}$$

$$= -\frac{1}{2}\sum_n\frac{\epsilon^2(n,\theta_m)}{\sigma} - \frac{N}{2}\log\sigma - \frac{N}{2}\log 2\pi \tag{3.11}$$

[1]

Cramer-Rao Inequality Let $\hat{\theta}(z^N)$ be the estimate of θ. It is important to assume that the estimator is unbiased, $E(\hat{\theta}_m(Z^N)) = \theta$. Under this condition Cramer-Rao Inequality states that

$$P_{\hat{\theta}_m^N} \geq (P^*)^{-1} \tag{3.5}$$

where $P^* = -E\frac{d^2}{d\beta^2}\log f_y(\beta; Y^N)|_{\beta=\theta}$ is the fisher information.

where $\frac{1}{N}\sum_n \epsilon^2(n,\theta_m)$ is the output error

$$\frac{1}{N}\sum_n \epsilon^2(n,\theta_m) = \frac{1}{N}\sum_n (y_n - \hat{y}_n(\theta_m))^2 \tag{3.12}$$

and $\hat{y}_n(\theta_m)$ is obtained by using θ_m and the input in ARMAX model (3.8) when noise is zero.

The ML estimate of the parameter

$$\hat{\theta}_m^N(z^N) = \arg\min_{\theta_m} \frac{1}{N}\sum_n \epsilon^2(n,\theta_m), \quad \hat{\sigma}_N^2(z^N) = \frac{1}{N}\sum_n \epsilon^2(n,\hat{\theta}_m^N) \tag{3.13}$$

Therefore the AIC, in (3.4), is

$$J(S_m) = \frac{1}{2}(1 + \log 2\pi + \log\left[\frac{1}{N}\sum \epsilon^2(n,\hat{\theta}_m^N)\right]) + \frac{m}{N}. \tag{3.14}$$

If the variance of the additive noise, σ^2, is known, then $\hat{\theta}_m^N = \arg\min_{\theta_m}\sum \epsilon^2(n,\theta_m)$ and the model validity criterion is

$$J(S_m) = \frac{1}{2}(\log\sigma + \log 2\pi + \frac{1}{N\sigma}\sum \epsilon^2(n,\hat{\theta}_m^N)) + \frac{m}{N} \tag{3.15}$$

3.1.2 Final Prediction Error(FPE)

If the cost function is $V_m^N(\theta_m, Z^N) = \frac{1}{N}\sum \frac{1}{2}\epsilon^2(n,\theta_m)$, where $\epsilon^2(n,\theta_m)$ is defined in (3.12), the validity criterion, in (3.3), is called FPE. For the ARMAX model in (3.8), if θ is an element of the parametric model set, [32]

$$J(S_m) = \frac{1+m/N}{1-m/N}\frac{1}{N}\sum \frac{1}{2}\epsilon^2(n,\hat{\theta}_m^N). \tag{3.16}$$

In [1] Akaike considers an AR model and shows that by implementing the FPE criterion, the probability of adapting a model set with order smaller than the correct model order goes to zero as N grows. But the probability of FPE(S_M) $\leq$ FPE(true model set) for model sets which include the true model and have higher dimension, goes to a non-zero constant number and therefore the estimation is not consistent. Note that when $m/N << 1$, the above model validity criterion (FPE) is the same as AIC criterion. This proves that AIC is also not consistent.

3.2 Bayesian Information Criteria (BIC)

Here the problem of selecting the dimension of a model is treated by calculation and comparison of the Bayesian Information Criterion (BIC) [44]. A prior probability distribution for the model parameter $\theta \in \Theta$ is assumed. In his classical paper [44] Schwarz argues that there is no intrinsic linear structure in a general parameter space

and assumes that the observed output $y^N = (y_1, \cdots, y_N)$ comes from a Koopman-Darmois family, i.e., relative to some fixed measure on the sample space they posses a density function of the form

$$f(y^N, \theta) = e^{(\theta' \cdot l(y^N) - b(\theta))} \tag{3.17}$$

where θ is an element of Θ a convex subset of M-dimensional Euclidean space, b is a scalar function of θ, $l(y_n)$ is a sufficient statistic [2] of y^N and has the same order as θ and $\theta' \cdot l(y^N)$ is the inner product of the two vectors, θ and $l(y^N)$.

Competing model sets are $S_m \subset \Theta$, with order m and $P(S_m) = \alpha_m$ is the prior probability of S_m being the correct model set. The prior distribution of θ in S_m is $f(\theta|S_m)\mathrm{d}\theta = \mathrm{d}\mu_{\mathrm{m}}(\theta_{\mathrm{m}})$ for a prior known function μ_m. Hence, the a priori density of θ is $f(\theta) = \sum \alpha_i \mathrm{d}\mu_{\mathrm{i}}(\theta)$.

With this prior assumption, the probability of θ being an element of S_m, given the observed data y^N, is

$$P(S_m|y^N) = \frac{f(y^N, S_m)}{f(y^N)}. \tag{3.18}$$

Bayesian information criterion for each model set S_m is the probability of the model set S_m being the correct model set given the observe data. Since the denominator of such criterion in (3.18) is the same for all S_m's, the criterion in comparison of BIC equivalently is

$$S(y^N, S_m) = \log f(y^N, S_m) \quad = \quad \log \int_{\theta_m \in S_m} f(y^N, \theta) d\theta \tag{3.19}$$

Using (3.17), Schwarz replaces $S(y^N, S_m)$ with

$$S(y^N, S_m) \quad = \quad \log \alpha_m + \log \int e^{(\theta' \cdot l(y^N) - b(\theta))N} \mathrm{d}\mu_{\mathrm{m}}(\theta) \tag{3.20}$$

However, (3.20) is obtained from (3.19) only if the prior assumption in (3.17) is replaced with the following prior assumption

$$f(y^N|\theta) = e^{(\theta' \cdot l(y^N) - b(\theta))} \tag{3.21}$$

which is a conditional probability distribution assumption rather than a joint probability distribution assumption.

Note that the conditional prior assumption in (3.17) is not consistent with separate prior assumptions for α_m and μ_m. If the joint distribution of y^N and θ, is known then

[2] sufficient statistic l relative to the parametric probability distribution family $f(y^N, \theta)$ is a function of the random variable y^N for which the conditional density of the random variable given l is independent of θ, i.e., mutual information of θ and y^N is the same as mutual information of θ and l.

α_m and μ_m can be calculated as follows

$$\alpha_m = P(S_m) = \int_{y^N, \theta_m \in S_m} f(y^N, \theta_m) d\theta_m dy^N \tag{3.22}$$

$$f(\theta|S_m) = \frac{f(\theta, S_m)}{P(S_m)} = \frac{1}{\alpha_m} \int_{y^N} f(y^N, \theta) dy^N \tag{3.23}$$

Therefore, in the following, we continue with the prior assumption in (3.21) which is consistent with the prior assumption on α_m and μ_m.

BIC Criterion Assume that for each model set S_m, μ_m is the Lebesgue measure on S_m. With the observed data y^N for each model set, as N goes to infinity, we have

$$S(y^N, S_m) \to N \sup_{\theta \in S_m} (\theta'_m \cdot l(y^N) - b(\theta_m)) - \frac{1}{2} m \log N + R \tag{3.24}$$

where $R = R(y^N, N, m)$ is bounded in N for fixed y^N, m. The method suggests to pick the model set which minimizes this asymptotic estimate of $S(y^N, S_m)$. The proof is first given for a family of probability distribution, $f(y^N|\theta)$, for which the distribution defined in (3.21) is in form

$$\theta'_m \cdot l(y^N) - b(\theta_m) = C_m - \gamma ||\theta - \hat{\theta}_m||^2 \tag{3.25}$$

where $C_m = \hat{\theta} \cdot l(y^N)_m - b(\hat{\theta}_m) = \max_{\theta_m \in S_m} (\theta \cdot l(y^N) - b(\theta_m))$ for some $\hat{\theta}_m = (\hat{\theta}_{m1}, \cdots, \hat{\theta}_{mm}) \in S_m$. For such family of distributions Schwarz claims that the criterion in (3.20) is

$$S(y^N, S_m) = \log \alpha_m + NC_j + \log \left(\sqrt{\frac{\pi}{N\gamma}} \right)^m. \tag{3.26}$$

BIC for Linear Gaussian Models Consider the following model structure

$$y_n = b_0 u_n + \cdots + b_q u_{n-q} + w_n, \tag{3.27}$$

where w_n is the additive white Gaussian noise(AWGN) with variance σ^2, u_n is the input which is independent identically distributed(i.i.d) with zero mean and unit variance and

$$\theta = (b_0, \cdots, b_q) \tag{3.28}$$

with dimension $k = q + 1$. In a model set of order m, $\theta_m = [\theta_m(1), \cdots, \theta_m(m)]$, the conditional density of y_n is $f_\theta(y_n|u^N, \theta_m)$, where for a given input, $E_\theta(y_n|\theta_m; u^N) = [u_n, \cdots, u_{n-m}]^T \cdot \theta_m$, With the prior assumption that θ is an element of the model set we have

$$f(y^N|\theta_m) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\frac{1}{2\sigma^2} ||y^N - A_m \theta_m||^2}. \tag{3.29}$$

where A_m is the Toeplits matrix

$$A_m = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ u_2 & u_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ u_N & u_N & \cdots & u_{N-m} \end{bmatrix} \tag{3.30}$$

In each model set of order m there exists a $\hat{\theta}_m$ which maximizes $f(y^N|\theta_m)$,

$$\hat{\theta}_m = (A_m^T A_m)^{-1} A_m^T y^N, \quad f(y^N|\hat{\theta}_m) = e^{NC_m}. \tag{3.31}$$

As N goes to infinity, $C_m \to \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2N\sigma^2}(y^N)^T A_m (A_m^T A_m)^{-1} A^T y^N$, since the process is stationary the limit exists. Also

$$||y^N - A_m\theta_m||^2 = ||y^N - A_m\hat{\theta}_m||^2 + \frac{N}{2\sigma^2}(\theta_m - \hat{\theta}_m)^T \frac{A_m^T A_m}{N}(\theta_m - \hat{\theta}_m). \tag{3.32}$$

As N goes to infinity, since the input is IID, $\frac{A_m^T A_m}{N} \to I_{m\times m}$. Therefore

$$f(y^N|\theta_m) = e^{N\left(C_m - \frac{1}{2\sigma^2}||\theta_m - \hat{\theta}_m||^2\right)}. \tag{3.33}$$

This conditional density function is from the family of probability distributions in (3.25) with $\gamma = \frac{1}{2\sigma^2}$. Therefore the BIC is

$$\begin{aligned} S(y^N, S_m) = -\frac{1}{2\sigma^2}||y^N - A_m\hat{\theta}_m||^2 \quad &- \quad \frac{m}{2}\log\left(\frac{N}{2\pi\sigma^2}\right) + N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \\ &+ \quad \log\alpha_m + \log\left(\sqrt{\frac{\pi}{n\gamma}}\right)^m \end{aligned} \tag{3.34}$$

3.3 Minimum Description Length

In this section we describe the principle of minimum description length (MDL) and study the two-stage MDL. Some examples are accompanied with several MDL criteria to show the fundamentals behind each description. We start with some background on information theory [8].

3.3.1 Background

Consider stationary random processes $Y^N(\theta_m) = Y_1, \cdots, Y_N$ where $\theta_m = (\theta(1), \cdots, \theta(m))$ is a real-valued vector parameter. Range of Y_i is a finite or countable set and each outcome $y^N = [y_1, \cdots, y_n]$ has a time invariant probability of $P_\theta(y^N)$. If Y_i is not finite or countable we use some number α of fractional digits in their binary representation for truncation and work with $p_\theta(y^N)$ obtained from $f_\theta(Y^N)$. Issues

related to the truncation of Y_i are addressed in [39].

The important question is how to code each string to a binary string such that the expected value of the code length is minimized. When the probability distribution of the string is known, i.e., θ is known, Shannon suggests a method of coding to minimize the expected value of length of codes as N tends to infinity.

Shannon coding theory

Consider a random variable y with finite (or countable) elements and string of $Y^N = [Y_1, \cdots, Y_N]$ an IID stochastic process. Entropy of Y, with probability distribution p_θ, is defined as $H(\theta) = -\sum_y p_\theta(y) \log p_\theta(y)$.

Theorem

$$\lim_{N\to\infty} (-\frac{1}{N} \log p(y_1, y_2, \cdots, y_N)) = H(\theta). \tag{3.35}$$

In other words as N goes to infinity only the elements of the typical set have non-zero probability. *Typical set* $A_\epsilon^N(\theta)$ is the set of all $(y_1, \cdots, y_N)$s such that

$$2^{-N(H(\theta)+\epsilon)} \leq p_\theta(y_1, \cdots, y_N) \leq 2^{-N(H(\theta)-\epsilon)} \tag{3.36}$$

and $Pr(A_\epsilon^N(\theta)) > 1 - \epsilon$ for sufficiently large enough N. This important theorem is a result of *the weak law of large numbers.*

Prefix code: A code maps all elements of Y^N to a set of binary codewords with length $L(y^N)$. If no codeword is the prefix of any other, then it is uniquely decidable. Any code satisfying this codeword condition is referred to as a *prefix code.* The code is prefix if and only if the code length satisfies Kraft's inequality. If we use binary alphabet for coding each y^N with code length $L(y^N)$, then *Kraft's inequality* is

$$\sum_{y^N} 2^{-L(y^N)} \leq 1. \tag{3.37}$$

Shannon shows that when y^N is generated by a source with probability distribution $p_\theta(y^N)$, and a prefix code $L(y^N)$ is used for coding, then as N goes to infinity the expected value of the length code is lower bounded by the entropy of the source

$$E_\theta(L(Y^N)) \geq H(\theta), \tag{3.38}$$

and the lower bound is achieved with a code satisfying

$$L(y^N) = -\log P_\theta(y^N). \tag{3.39}$$

A random prefix code can be associated with any probability distribution $Q(y^N)$. If $L(y^N)$ is the codeword length for y^N, define the probability for y^N as

$$Q(y^N) = \frac{2^{-L(y^N)}}{\sum 2^{-L(y^N)}}. \tag{3.40}$$

By using the kraft's inequality, which holds for codeword $L(y^N)$, it can be shown that

$$E_\theta(L(Y^N)) \geq H(\theta) + D(p_\theta||Q) \tag{3.41}$$

where $D(p_\theta||Q) = \sum p_\theta(Y^N) \log \frac{p_\theta(Y^N)}{Q(Y^N)}$ is the Kullback-Leibler distance of p_θ and Q. This inequality provides a lower band for the expected value of the codelength when the true parameter is not used for the coding.

Huffman Coding Note that for a finite length observation Huffman codes are the optimum codes minimizing the expected value of the code length [8]. In finite observation, Huffman code assigns a shorter length code to the non-typical set elements than Shannon coding. However, as N grows, the non-typical set becomes smaller and smaller and the effects of this set in the expected value of the code length can be ignored. Therefore, the expected value of the code length for Shannon and Huffman code becomes the same in the limit.

3.3.2 The MDL Principle

Here we consider the problem of coding the string $y^N = (y_1, \cdots, y_N)$ when the distribution generated the data is not known but y^N is a string generated by a parametric probability distribution p_θ, and θ belongs to a compact subset of R^m. Following the Shannon coding, Rissanen obtains a lower bound for $E_\theta(L(Y^N))$ when the model parameter θ is not known. The estimate of θ, $\hat{\theta}$, defines a probability distribution Q for y^N. In the following theorem the principle of two-stage MDL is given and a lower bound is derived for $D(p_\theta||Q)$ in (3.41). The bound is provided under particular assumption on the rate of convergence of $\hat{\theta}$ to θ. The MDL principle is to estimate the obtained lower bound as a criterion to compare different model sets of different order.

Rissanen's First Theorem

Consider a family of parametric models with parameter $\theta \in \Omega^m$, where Ω^m is a compact subset of R^m. Let central limit theorem hold for the maximum likelihood estimator $\hat{\theta}(y^N)$ for each θ in the interior of Ω^m so that the distribution $\sqrt{N}(\hat{\theta}(y^N) - \theta)$ converges to a zero-mean normal distribution with covariance matrix $\Sigma(\theta)$ as N grows. Then for any prefix code with codelength $L(y^N)$ and for all positive ϵ

$$E_\theta \frac{1}{N} L(Y^N) \geq \frac{1}{N} H(\theta) + (1-\epsilon)\frac{m}{2N} \log N, \tag{3.42}$$

where $H(\theta)$ is the entropy of the data generated by θ. The inequality holds for all but a set of θs with a Lebesgue measure which approaches zero as the length of y^N, N, grows. The proof of this theorem first presented in [38] and some corrections was made by Telatar in [50]. Rissanen gives a more comprehensive proof in his book [42].

In [38] Rissanen suggests an optimal code which achieves the lower bound given in theorem, as N grows. If $\hat{\theta}_d(y^N)$ is the ML estimate $\hat{\theta}(y_N)$, truncated to $\lfloor \log \sqrt{N} \rfloor$ and $C_1(N)$ is a normalized factor to make the code length satisfy the Kraft inequality, define

$$L(y^N) = -\log p_{\hat{\theta}_d}(y^N) + \frac{1}{2} m \log N + C_1(N). \tag{3.43}$$

where $C_1(N)/N$ goes to zero as N grows.

To compare different model sets with different order, the integer number m is coded by using *Elias's universal representation of integers*. The number of bits for such coding is $\log^* m = \log m + \log\log m + \cdots + c$. The summation is over all nonnegative $\log\log\cdots\log m$ and constant c is added so that the coding satisfies the Kraft inequality[3]. It is then suggested to choose the set with dimension m for which the code length is minimized

$$L(y^n) = \min_m \{-\log p_{\hat{\theta}_d}(y^N) + \frac{1}{2} m \log N + \log^* m + C_2(N)\} \tag{3.44}$$

Example of Model Selection

In finding the two-stage MDL criterion it is assumed that there is an equal probability that any of the models being the correct model. Therefore for each string generated by any of the models we use a fixed number of bits. In comparison of the code length of the data, we can ignore this code length.

In stage one, given each model set S_m, θ is encoded by first discretizing the compact parameter space of dimension m, with precision $\delta = 1/\sqrt{N}$ per dimension. So in the first stage we estimate θ using a method such as ML or Bayes procedure. The estimate $\hat{\theta}$ is then truncated to precision θ $\lfloor \hat{\theta} \rfloor$. Assuming that probability distribution of θ in each model set is uniform, i.e., $f(\theta|S_m)$ is constant, we have

$$L(\hat{\theta}) \approx \frac{m}{2} \log N. \tag{3.45}$$

In the second stage, the estimated distribution is used to encode Y^N using Shannon's coding method. A continuous data is discretized to some precision δ_d. Therefore

$$\log P(y^N) \approx -\log f(y_1, \cdots, y_N | \lfloor \hat{\theta} \rfloor) - N \log \delta_d. \tag{3.46}$$

[3]Example of Elias coding : code 1010010000 is used for N=16 (log N=4, log log N=2), we start with the first two digit, the first element is 10= binary representation of $j = \log\log n$, this tells us that the second $j+1 = 3$ elements to be read which is 100. So we should read the next $j+1 = 4+1$ digits 10000 which is 16 and after that zero means that the code is ended.

Since $N \log \delta_d$ is constant for all the model classes, we can ignore it in comparing the model sets. Then

$$L(y^N|\hat{\theta}) \approx -\log f(y_1, \cdots, y_N|\hat{\theta}). \tag{3.47}$$

Combining the two steps, the two-stage MDL criterion is

$$L(y^N) = -\log f(y_1, \cdots, y_n|\hat{\theta}) + \frac{m}{2} \log N. \tag{3.48}$$

Note that this criterion is the same as BIC.

Let's assume that $y^N = (y_1, \cdots, y_N)$ is an IID observation from a normal distribution with variance one and mean θ, $N(\theta, 1)$, for some $\theta \in R$. We want to choose between the models $M_0 = \{N(0,1)\}$ and $M_1 = \{N(\theta,1) : \theta \neq 0\}$. Note that if we maximize the likelihoods of both models and choose the larger maximized likelihood, M_1 is always chosen unless $\bar{y}_N = \frac{y_1 + \cdots + y_N}{N}$ is 0, which is an event with zero probability even when M_0 is true!.

By using the two-stage description length criterion (3.48) for model M_0 and M_1 we have

$$L_0(y^N) = \frac{1}{2}\sum_1^n y_i^2 + \frac{N}{2}\log(2\pi). \tag{3.49}$$

$$L_1(y^N) = \frac{1}{2}\sum_1^n (y_i - \bar{y}_N)^2 + \frac{N}{2}\log(2\pi) + \frac{1}{2}\log N. \tag{3.50}$$

where $\bar{y}_N = \frac{y_1 + \cdots + y_N}{N}$ and $\hat{\theta} = \bar{y}^N$ is the ML estimate of θ. Following the MDL principle, we choose M_0 over M_1 if

$$|\bar{y}_N| < \sqrt{\log(N)/N}. \tag{3.51}$$

In this example the MDL criterion takes the form of a likelihood ratio test and it can be shown that the method is consistent.

3.3.3 Rissanen's Second Theorem

The second theorem is the result of the first theorem when it is applied for ARMA models. Here the upper bound of expected value of the code length, provided in theorem one, gives an approximation for the rate of convergence of the output error to zero [38]. Consider an ARMA process

$$y_n + a_1 y_{n-1} + \cdots = b_0 e_n + b_1 e_{n-1} + \cdots + b_q e_{n-q} \tag{3.52}$$

where e_ns are independent zero-mean Gaussian random variables with variance $\sigma^2(\theta)$ and $\theta = (a_1, \cdots, a_p, b_0, \cdots, b_q)$ ranges over a compact set Ω^m. The two polynomials have no common factors and all the roots are outside the unit circle, i.e., the system

is assumed to be minimal phase. The prediction error is

$$V^N(\theta) = \sum_{n=1}^{N-1} E_\theta(y_{n+1} - \hat{y}_{n+1})^2. \tag{3.53}$$

The second theorem in [42] provides a lower bound for $V^N(\hat{\theta})/N$.

Theorem Consider the ARMA model in (3.52). The predictor function $\hat{y}_n$ is a function of past observations and $\hat{\theta}$ the estimate of θ converges θ as N grows. Then the following inequality holds.

$$\frac{1}{N}\sum_{n=1}^{N-1} E_\theta(y_{n+1} - \hat{y}_{n+1})^2 \geq \sigma^2(\theta)[1 + \frac{p+q-\epsilon}{N}\ln N] \tag{3.54}$$

for all θs except elements of the typical set $A_\epsilon(N)$ defined in (3.36) which its Lebesgue measure goes to zero as N grows.

Example Related to the Second Theorem The second theorem provides a lower bound for rate of convergence of output error to zero. The prior assumption is that the true system is an element of the model set. This prior assumption plays an important role in calculation of the lower bound. When the true system is not a member of the model sets the error in limit is a nonzero number. To elaborate this point we give the following example of calculation of the limit of the output error. Consider the following ARMA model

$$y_n + a_0 y_{n-1} = e_n + c_0 e_{n-1} \tag{3.55}$$

where e_n is an AWGN with variance σ. Then

$$R_{ye}(0) = \sigma, \quad R_{ye}(1) = (c_0 - a_0)\sigma, \quad R_{ye}(n) = 0, n \neq 0, 1 \tag{3.56}$$

$$R_{yy}(0) = \sigma\frac{1 + c_0^2 - 2a_0c_0}{1 - a_0^2}. \tag{3.57}$$

where for the stationary processes x, z, $R_{xz}(\tau) = E(x_n z_{n+\tau})$. Here we consider three different model sets and use the least square (LS) estimation in each of the sets and calculate the prediction error(PE) as N goes to infinity.

- The first model set has order two, $\theta = (a, c)$ and the structure of the model set is in form of

  $$y_n + a y_{n-1} = e_n + c e_{n-1} \tag{3.58}$$

  for each a,c. Therefore $\theta_0 = (a_0, c_0)$ is an element of this model set. By using the minimum variance estimator of the output [32]

  $$\hat{y}_n = E(y_n | y_{n-1}, \cdots, y_0) = c e_{n-1} - a y_{n-1} \tag{3.59}$$

where e_{n-1} can be found from $e_{n-1} = \frac{z^{-1}+az^{-2}}{1+cz^{-1}}Y(z)$ i.e.

$$e_{n-1} = (y_{n-1} + ay_{n-2}) * ([1, c, c^2, ...]) \tag{3.60}$$

($A * B$ denotes the convolution of signals A and B). Therefore for $\hat{y}$ we have

$$\hat{y}_n - \hat{y}_{n-1} = (c - a)y_{n-1} \tag{3.61}$$

and

$$\hat{y}_n - y_n = -e_n + (c - c_0)e_{n-1} + (a_0 - a)y_{n-1}. \tag{3.62}$$

We minimize the output error

$$V_m^N(\theta, z^N) = \frac{1}{N}\sum(\hat{y}_n - y_n)^2 \tag{3.63}$$

to find the estimator. As N goes to infinity the output error converges to its mean

$$\begin{aligned} \frac{1}{N}\sum(\hat{y}_n - y_n)^2 &= E(\hat{y}_n - y_n)^2 \qquad (3.64)\\ &= \sigma(1 + (c - c_0)^2 + (a_0 - a)^2 R_y(0) + 2(c - c_0)(a_0 - a)) \end{aligned}$$

and the LS estimator is $c = c_0$ and $a = a_0$ with

$$\min \frac{1}{N}\sum(\hat{y}_n - y_n)^2 = \sigma \tag{3.65}$$

So with this model set $\hat{\theta} \to \theta_0$ as N grows and the result is consistent with MDL approach.

- Consider a model of form $\theta = c$, i.e., M_2 is one dimensional model set with $a = 0$ in (3.55). So the model set is in form of

$$y_n = e_n + ce_{n-1}. \tag{3.66}$$

In this case

$$\begin{aligned} \hat{y}_n = ce_{n-1} &= c(y_{n-1} + ay_{n-2}) * ([1, c, c^2, \cdots]) \qquad (3.67)\\ \hat{y}_n - y_n &= -e_n + (c - c_0)e_{n-1} + a_0 y_{n-1} \qquad (3.68)\\ E(\hat{y}_n - y_n)^2 &= \sigma(1 + (c - c_0)^2) + a_0^2 R_{yy}(0) + \sigma 2a_0(c - c_0) \qquad (3.69) \end{aligned}$$

Here the LS estimator, as N goes to infinity, is $\hat{a} = 0, \hat{c} = c_0 - a_0$, which can cause a large bias.

$$\min \frac{1}{N}\sum(\hat{y}_n - y_n)^2 = \sigma(1 + a_0^2) + R_{yy}(0) \tag{3.70}$$

- Consider $\theta = a$, i.e., M_3 is a one dimensional model set with $c = 0$ a model class of form

$$y_n + ay_{n-1} = e_n. \tag{3.71}$$

Then

$$\begin{aligned} \hat{y}_n &= ay_{n-1} && (3.72)\\ \hat{y}_n - y_n &= -e_n - c_0 e_{n-1} + (a_0 - a)y_{n-1} && (3.73)\\ E(\hat{y}_n - y_n)^2 &= \sigma(1 + c_0^2) + (a_0 - a)^2 R_{yy}(0) - 2\sigma a_0(c - c_0) && (3.74) \end{aligned}$$

The LS estimator as N goes to infinity is $\hat{c} = 0, \hat{a} = a_0 - \frac{c_0\sigma}{R_{yy}(0)} = a_0 - \frac{c_0(1-a_0^2)}{1+c_0^2-2a_0c_0}$ and

$$\min \frac{1}{N}\sum(\hat{y}_n - y_n)^2 = \sigma(1 + c_0^2) + \frac{(\sigma c_0)^2}{R_{yy}(0)} \tag{3.75}$$

This example shows the effects of under modeling in the estimation for an ARMA process. For the last two model sets, M_2 and M_3 the lower bounds on the output error in (3.70) and (3.75) are nonzero and functions of the unmodeled elements. Therefore it proves that the lower bound suggested in Rissanen's theorems is not applicable in these cases.

3.3.4 Other Forms of Description Length

The two-stage MDL was the first MDL method suggested when the MDL principle was introduced. Here we briefly review several coding schemes that was introduced after two-stage description length. The methods provide description lengths for a data string based on classes of probability models. All these forms of MDL share many aspects of both frequentist and Bayesian approaches to the model selection. See [24] for more details and simulation results on these methods.

Mixture MDL and Stochastic Information Complexity

In mixture MDL a prior probability distribution for θ in each model set is considered. The description of a data string y^N in a model set M is based on a distribution that is obtained by using a mixture of the model set members with respect to a distribution $w(\theta)$ on the parameters

$$m(y^N) = \int_M f_\theta(y^N|\theta)w(\theta)\mathrm{d}\theta. \tag{3.76}$$

It should be emphasized that $w(\theta)$ is not as a prior in the Bayesian approach but rather as a device for creating a distribution for the data based on the model class. The description length is

$$-\log[m(y^N)] = -\log\int_M f(y^N|\theta)w(\theta)\mathrm{d}\theta. \tag{3.77}$$

An analytical approximation to the mixture $m(\cdot)$ in (3.76) is obtained by Laplace's expansion when $w(\theta)$ is smooth [41]. With this approximation, for the defined description length, we arrive at a two-stage description length which we will call the *Stochastic Information Complexity*:

$$SIC(y^N) = -logf(y^N|\hat{\theta}) + \frac{1}{2}\log\det(\hat{\Sigma}_N) \tag{3.78}$$

where $\hat{\theta}_N$ is the MLE and $\hat{\Sigma}_N$ is the Hessian matrix of $-\log f(y^N|\theta)$ evaluated at $\hat{\theta}_N$. For IID observations from a regular parametric family, as N tends to infinity

$$\frac{1}{2}\log\det(\hat{\Sigma}_N) = \frac{1}{2}\log\det(NI(\hat{\theta}))(1+o(1)) = \frac{k}{2}\log N1 + o(1)) \tag{3.79}$$

where $I(\cdot)$ is the Fisher information matrix of a single observation. In this case SIC is approximately the same as BIC or two-stage description length.

Predictive Description Length

In [40] Rissanen introduces another description length for strings generated by parametric models. Here instead of minimizing the log likelihood over the complete data i.e. the joint distribution $f(y^n, \theta)$, at each time t we estimate $\hat{\theta}(t)$ from the first $t-1$ elements of y^n. Therefore we have

$$L(y^N) = -\sum_{t=1}^{N} \log f_{\hat{\theta}_t}(y_{t+1}|y^t) \tag{3.80}$$

as the cost of encoding the data string y^N. The MDL model selection criterion based on this form of description is called predictive MDL.

Consider the ARMA model, with β the parameter to be estimated. Then the PMDL is

$$PMDL(y^N) = -\sum_{1}^{N} \log f(y_t|\hat{\beta}). \tag{3.81}$$

PMDL is closely related to the so-called accumulated prediction error(APE) of the form

$$APE(y^N) = \sum_{1}^{N}(y_t - \hat{y}_t)^2, \tag{3.82}$$

which was introduced by Akaike as a final prediction error (see section 3.1.2). The computational cost of PMDL for general ARMA models is enormous since the parameter estimate must be updated for each new observation.

Normalized MDL

Recently Rissanen developed an MDL criterion based on normalized maximum likelihood coding scheme. The NML description of a data string is provided by restricting the second stage of coding to a data region identified by the parameter estimate, i.e., by the typical set of strings generated by the estimated parameter.

The description length of normalized MDL, nMDL, for a linear regression model is derived in [24] and [3]. Assume that the output of the system is generated by

$$y_i = \sum_{n=1}^{k} \beta_n u_{i-n} + w \tag{3.83}$$

where variance of the zero mean additive white Gaussian noise w is σ_w^2. Consider $\hat{\beta}$ and $\hat{\sigma}^2$ the maximum likelihood estimates of β and σ_w^2,

$$\hat{\beta}(y^N) = (U'U)^{-1}U'y^N, \quad \hat{\sigma}^2 = ||y^N - U\hat{\beta}||^2/N. \tag{3.84}$$

where U is the Toeplits matrix of input u. The Gaussian density of y^N corresponding to the model set S_m is $f_{S_m}(y^N; u^N, \beta, \sigma^2)$. Then the normalized maximum likelihood function is

$$\hat{f}(y^N; u^N, S_m) = \frac{f_{S_m}(y^N; u, \hat{\beta}(y), \hat{\sigma}^2(y))}{\int_{S(r,\sigma_0^2)} f_{S_m}(z; u^N, \hat{\beta}(z), \hat{\sigma}^2(z))dz} \tag{3.85}$$

where

$$S(r, \sigma_0^2) = \{z | \hat{\beta}'(z)U'U\hat{\beta}(z)/N \leq r, \hat{\sigma}^2 \geq \sigma_0^2\}, \tag{3.86}$$

and r and σ_0^2 are chosen such that the Ml estimates fall within $S(r, \sigma_0^2)$. Calculation of nMDL and comparison of this description length with other forms of MDL, for the linear regression problem, is given in [24].

3.3.5 Consistency

One important issue in model selection procedure is consistency of the method when a finite-dimensional model has generated the data. As the sample size gets larger, a consistent procedure chooses the correct model class with probability approaching one. Akaike in [1] shows that with FPE, the probability of choosing a model with higher dimension is nonzero, and hence this method is not consistent. AIC and FPE are the same for the ARMA models, when the number of observed data is much larger than the dimension of the true model set. Therefore AIC also is not consistent for ARMA models.

In [48] it is proven that the two-stage MDL(and therefore the BIC), predictive and mixture form of MDL are consistent methods for linear regression problems.

3.4 Discussion

We reviewed three methods of model order selection AIC, BIC and MDL, with emphasis on the application for linear regression. All these methods assume that the true model has a finite dimension, and there exists a model set among the chosen model sets that is "close enough" to the true system. Not much related research is done for the case that the true system is infinite dimensional and the bias decays gradually but never approaches zero. For such cases Shibata [45],[46] shows that in terms of prediction error AIC is optimal.

3.4.1 Comments on BIC

BIC in section 3.2 is in form of

$$S(y^N, S_m) = \log \alpha_m + NC_j + \log \left(\sqrt{\frac{\pi}{N\gamma}}\right)^m. \tag{3.87}$$

However, in appendix A.2 we prove that the calculation of $S(y^N, S_m)$ leads to

$$S(y^N, S_m) = \log \alpha_m + NC_j - \log \mu_m + \log \left(\sqrt{\frac{\pi}{N\gamma}}\right)^m. \tag{3.88}$$

The extra element $-\log \mu_m$ in the criterion, which was ignored in (3.26), plays a critical role in comparison of model sets of different order.

The extra term especially becomes important when the estimation is over all model sets with dimension less than the true model set. Assume that the true model set is $\theta = (.5, .5)$ and our prior information is that $\theta \in [0, 1] \times [0, 100]$. Consider two model sets $m_1 = 0 \times [0, 100]$, $m_2 = [0.1] \times 0$. Also assume that we assigned the same probability for each of these models to be the correct one, i.e., $\alpha_1 = \alpha_2$. If the BIC defined in (3.87) is the same for the two model classes, $A = A_1 = A_2$, and it occurs at two points (0,a) in m_1 and (b,0) in m_2, then the new BIC is $S(y^N, m_1) = A_1 - \log \mu_1 = A - \log 100$ and $S(y^N, m_2) = A_2 - \log \mu_1 = A$, which implies that we choose point (a,0). It means that we pick the subset which has the smaller size(smaller Lebesgue measure). However the criterion is the same if we use the BIC in (3.87) for both models.

3.4.2 MDL Principle and the Role of "$\frac{\log(N)}{N}$"

The inequality which is proved in theorem one is valid for all but the non-typical set out of $D_\epsilon(N)$ whose Lebesgue measure is small and as N grows approaches zero. This prior assumption might be wrong in cases that the some of the elements of $D_\epsilon(N)$ has a very high probability density (for example Dirac delta at some θ). This problem is addressed in [10]. The theorem can be further strengthened to the form that not only the volume of the non-typical set tends to zero, but even the union of these sets over $n \geq N$ goes to zero as N also tends to infinity. The proof is tied to the rate at which

the distribution of the estimator approaches its limit. More on this issue is discussed by Rissanen in [41].

The lower bound provided in the first theorem which introduces the MDL principle, is of form

$$E_\theta \frac{1}{N} L(Y^N) - \frac{1}{N} H(\theta) \geq (1-\epsilon)\frac{m}{2N}\log N, \tag{3.89}$$

with probability one for $0 \leq \epsilon < 1$. One important fact about the probability of this event is that for a fixed N the probability is an increasing function of ϵ. If ϵ is close to 1 the rate of convergence of the probability as a function of N is much faster than the rate of convergence of that of ϵs close to zero.

Note that as we discussed in 3.1.1, AIC estimates $E_\theta \frac{1}{N} L(Y^N) - \frac{1}{N} H(\theta)$ to be

$$E_\theta \frac{1}{N} L_{\hat{\theta}}(Y^N) - \frac{1}{N} H(\theta) \approx \frac{m}{N}, \tag{3.90}$$

when the estimator of θ is unbiased. MDL is obtained by using an unbiased estimator for θ for which the lower bound in (3.89) is achieved for $\epsilon \approx 0$ and as N grows

$$E_\theta \frac{1}{N} L(Y^N) - \frac{1}{N} H(\theta) \geq \frac{m}{2N}\log N, \tag{3.91}$$

with probability which approaches one as N grows. In the following theorem we prove that $\log(N)$ in Rissanen's first theorem can be replaced with a family of functions:

Theorem 3.4.2 Consider a family of parametric models with parameter $\theta \in \Omega^m$, where Ω^m is a compact subset of R^m. Let central limit theorem hold for the maximum likelihood estimator $\hat{\theta}(y^N)$ for each θ in the interior of Ω^m so that the distribution $\sqrt{N}(\hat{\theta}(y^N) - \theta)$ converges to a zero-mean normal distribution with covariance matrix $\Sigma(\theta)$ as N grows. Then for any prefix code with codelength $L(y^N)$ and for all positive ϵ the following inequality holds

$$E_\theta \frac{1}{N} L(Y^N) \geq \frac{1}{N} H(\theta) + (1-\epsilon)\frac{\log(\beta(N))}{N}, \tag{3.92}$$

where $H(\theta)$ is the entropy of the data generated by θ. The inequality holds for all but a set of θs with a Lebesgue measure which approaches zero as the length of y^N, N, grows and $\beta(N)$ is a function of N satisfying the following conditions

$$\lim_{N\to\infty} \beta(N) = \infty \tag{3.93}$$

$$\lim_{N\to\infty} \frac{1}{(1-\epsilon/2)} \frac{(\beta(N))^{\frac{1-\epsilon}{1-\epsilon/2}}}{N^{m/2}} = 0 \tag{3.94}$$

Proof In appendix B. $\diamondsuit$

An example of $\beta(N) = N^{m/2}$ results the lower bound which is given in first theorem. However, this $\beta(N)$ does not provide the tightest lower bound in the limit. For example $\beta(N) = (\log(N))^{\frac{m}{2}}$ which is smaller than $N^{m/2}$ as N grows also satisfies

the conditions.

An information theoretical approach to this theorem is given in [9]. Here Rissanen first theorem is stated as follows.

Theorem Let $\{P_\theta\}_{\theta\in\Theta}$ be any family of random processes, not necessarily IID, possibly not even stationary, where $\Theta \in R^m$. Suppose that for each $N \geq N_0$, there exists an estimator $\hat{\theta}_N(y^N)$, function of observed data y^N, with

$$E_\theta||\hat{\theta}_N(y^N) - \theta||^2 \leq \frac{c(\theta)}{N}. \tag{3.95}$$

Then, for every $\epsilon > 0$, there is a constant $K > 0$ such that for $N \geq N_0$ and for every probability density or mass function g we have

$$E_\theta \log \frac{P_\theta(Y^N)}{g(Y^N)} \geq \frac{m}{2}\log N - K \tag{3.96}$$

except possibly for a set of parameters θ of Lebesgue measure less than ϵ. The fixed number K is a function of m and $c(\theta)$. It is shown in the proof that

$$K = B + \frac{m}{2}\log c(\theta) \tag{3.97}$$

where B itself is a function of $c(\theta)$ and m.

Note that in practical cases $c(\theta)$ in (3.95) is a function of the variance of the estimate and can be a chosen as a function of N. One valid example is $c(\theta) = Nf(N)$. For the variance of estimate to be a finite number $\lim_{N\to\infty}\frac{Nf(N)}{N}$ has to be finite. For this example one $c(\theta)$ dependent element of K, $\log c$, is

$$\frac{m}{2}\log(c(\theta)) = \frac{m}{2}\log f(N) + \frac{m}{2}\log(N) \tag{3.98}$$

Using (3.97) and (3.98), the element $\frac{m}{2}\log(N)$ is eliminated from the lower bound in the theorem in (3.96).

We conclude that the mysterious number $\frac{\log(N)}{N}$ in the first theorem can be replaced by a family of functions $\beta(N)$ provided in theorem 3.4.2.

3.4.3 Conclusion

All the information theoretic approaches discussed in this chapter heavily rely on one prior assumption: the true model belongs to all the comparing model sets. With such assumption as N grows an estimate of the rate of convergence of the estimate to the true parameter is estimated. The results provided for all the methods is asymptotic for large enough N.

Can the criterion provided for such model sets be used for comparison of model sets which do not include the true model? In practical problems we do not know which of the competing model sets has the minimum order and at the same time

includes the true model. That is why we use the order estimation methods. In practice the criterion which is obtained for model classes which include the true model is used for all the competing model sets. Here we elaborate the drawbacks of such implementation through the following example. The output of an LTI system is generated as follows

$$y_n = \sum_{i=1}^{M} u_{n-i} h(i) + w_n \tag{3.99}$$

where $h(i)$ is the taps of the finite impulse response of the system and w is the additive white Gaussian noise which belongs to $N(0, \sigma_w^2)$. The goal is to estimate the impulse response and its length M using the input and output of length N. The competing model sets S_m are the systems with impulse response of length m. No delay is considered.

All the information theoretic methods at some point of calculation of the criterion calculate the pdf of y^N given the estimate of the impulse response $\hat{h}_m^N$ in each subspace. The distribution of y^N for subspaces, when $M \leq m$, is

$$\frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\hat{y}_{S_m}^N - \bar{y}^N||^2}{2\sigma_w^2}} \tag{3.100}$$

where $\bar{y} = \sum_{i=1}^{M} u_{n-i} h(i)$ is the mean and $\hat{y}_{S_m}$ is the estimate of output using the estimate of impulse response $\hat{h}_m^N$. The output provided by any subspace S_m is of form

$$y_{S_m} = \sum_{i=1}^{m} u_{n-i} h_{S_m}(i) \tag{3.101}$$

Note that the mean of the random variable $\hat{Y}_{S_m}$, $\bar{y}^N$, can not be generated by any element of subspace S_m when $M > m$. Therefore, by using the pdf in (3.100) for all the model sets, the unmodeled dynamics effects in the estimation of output is considered as a part of the noise. Such assumption provides a larger variance for the additive noise than σ_w^2 for model sets S_m, $m < M$, which is an inconsistent result.

In the following chapters our goal is to provide a new information theoretic order estimation method which avoids this problem and does not rely on a prior assumption that the true model is an element of any given model set. Also, from the review of the existing methods it seem that a proper definition of distance measure for the parametric family of probability distributions is needed. For all these purposes we will introduce a new method of order estimation based on an information theoretic distance measure in chapter 5.

Chapter 4

New Quality Evaluation Method

4.1 Problem Statement

We consider a stable causal, single-input/single-output, linear time-invariant, discrete-time system. Input and output of the system are related as follows

$$y_n = \sum_{i=1}^{n} h_i u_{n-i+1} + w_n, \tag{4.1}$$

where $h = [h_1, \cdots]^T$ is the impulse response of the system and $w = [w_1, \cdots]^T$ is the zero-mean additive white Gaussian noise(AWGN). Each w_i has variance σ_w^2, and is independent of the input. The input is assumed to be a quasi-stationary signal [32].

Finite length data, input $[u_1, \cdots, u_N]$, and output $[y_1, \cdots, y_N]$, is available and u_i is zero for $i \leq 0$. There is no assumption on the length of the impulse response. However, note that only the first N elements of h, h^N, relate the finite N points of the input and output.

Consider subspace S_m of order m in space R^N. An estimate of h^N in this subspace is $\hat{h}_{S_m}$. Our goal is to find an estimate for the estimation error

$$||h^N - \hat{h}_{S_m}||^2. \tag{4.2}$$

In the following sections we provide probabilistic bounds on this error as a function of length of data N, the structure of S_m, and the input u. We are able to provide probabilistic bounds on this criterion. The provided bounds converge to the criterion as N grows. We are also able to provide the rate of convergence. The performance of estimates in different subspaces can be compared by comparison of the error estimate. With this criterion the best subspace is the one which minimizes the estimation error.

4.2 Impulse Response Error

Consider the space R^N, which is the space of the first N taps of the impulse response, and S_m a subspace of order m of this space. The following argument can be used

for any subspace of R^N, S_m. However, for simplicity of presentation, let S_m be a subspace which includes the first m taps of the impulse response. Form (4.1) the input-output relationship for the finite available data is

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ u_2 & u_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_N & u_{N-1} & \cdots & u_1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} \tag{4.3}$$

$$y^N = \begin{bmatrix} A_m(N) & B_m(N) \end{bmatrix} \begin{bmatrix} h_m^N \\ \\ \Delta_m^N \end{bmatrix} + w^N \tag{4.4}$$

where $y^N = [y_1, \cdots, y_N]^T$, $h^N = [h_m^N; \Delta_m^N]$, $h_m^N = [h_1, \cdots, h_m]^T$, $\Delta_m^N = [h_{m+1}, \cdots, h_N]^T$, $A_m(N)$ is an $N \times m$ matrix, $B_m(N)$ is an $N \times N - m$ matrix and w^N is the additive white noise. The least-square method is used to find the estimate of first m taps of the impulse response, h_m^N,

$$\begin{aligned} \hat{h}_m^N &= ((A_m(N))^T A_m(N))^{-1}((A_m(N))^T y^N), \\ &= h_m^N + ((A_m(N))^T A_m(N))^{-1}(A_m(N))^T (B_m(N)\Delta_m^N + w^N) \end{aligned} \tag{4.5}$$

(From here we drop N from $w^N, A_m(N), B_m(N)$). We define two errors: subspace impulse response error (SIRE), $||\hat{h}_m^N - h_m^N||_2^2$, and impulse response error(IRE), $||\hat{h}_m^N - h^N||_2^2$:

$$||\hat{h}_m^N - h_m^N||_2^2 = (B_m\Delta_m^N + w)^T C_m (B_m\Delta_m^N + w) \tag{4.6}$$

$$||\hat{h}_m^N - h^N||_2^2 = ||\hat{h}_m^N - h_m^N||_2^2 + ||\Delta_m^N||^2 \tag{4.7}$$

where

$$C_m = A_m(A_m^T A_m)^{-1}(A_m^T A_m)^{-1} A_m^T. \tag{4.8}$$

The goal is to estimate these errors given the observed data.

Asymptotic Behavior Before we estimate bounds on IRE and SIRE, it is informative to investigate on the asymptotic behavior of the two errors as the length of data grows. As N approaches infinity the terms which are noise dependent approach zero asymptotically

$$\lim_{N\to\infty} 2w^T C_m B_m \Delta_m^N = 0, \quad \lim_{N\to\infty} w^T C_m w = 0 \tag{4.9}$$

Therefore

$$\lim_{N\to\infty} ||\hat{h}_m^N - h_m^N||_2^2 = (\Delta_m^\infty)^T (\lim_{N\to\infty} B_m^T C_m B_m)\Delta_m^\infty, \tag{4.10}$$

$$\lim_{N\to\infty} ||\hat{h}_m^N - h||_2^2 = \lim_{N\to\infty} ||\hat{h}_m^N - h_m^N||_2^2 + ||\Delta_m^\infty||^2. \tag{4.11}$$

The second component of the impulse response error in (4.11) is the norm of the unmodeled dynamics of the system. The first component, however, is a function of both the input and unmodeled dynamics. Since the input is quasi-stationary $\lim \frac{1}{N} A_m^T B_m$ and $\lim \frac{1}{N} A_m^T A_m$ exist, therefore, $B_m^T C_m B_m$ has a limit. If the input is such that

$$\lim_{N\to\infty} \frac{1}{N} A_m^T B_m = 0 \tag{4.12}$$

then $B_m^T C_m B_m$ vanishes as N approaches infinity and SIRE in (4.10) approaches zero asymptotically. Such scenario happens for a subspace of quasi-stationary inputs such as independent identically distributed (IID) inputs. We will elaborate on the properties of such input in the following sections. If $\lim \frac{1}{N} A_m^T B_m$ does not approach zero asymptotically, there is a fixed bias in SIRE as N goes to infinity.

4.2.1 Output Error

The only available data to estimate the impulse response error is the observed input and output of the system. Here we study the behavior of the output error and its relationship with the IRE in different subspaces. The output error is

$$\frac{1}{N}||y - \hat{y}_m^N||_2^2 = \frac{1}{N}(B_m\Delta_m^N + w)^T G_m B_m(\Delta_m^N + w) \tag{4.13}$$

where $\hat{y}_m^N = A_m \hat{h_m^N}$ is the estimate of the output in subspace S_m and

$$G_m = (I - A_m(A_m^T A_m)^{-1} A_m^T). \tag{4.14}$$

In absence of the additive noise, SIRE, IRE (4.6),(4.7) and the output error are decreasing functions of m. Assume that there exists M such that $h_M \neq 0$, $h_i = 0$, $i > M$. If $M < N$, then all errors are none zero for $m < M$ and zero for $m \geq M$. If $M \geq N$, then all errors are decreasing functions of m. In this noiseless scenario comparing the output error, which is available, is equivalent to comparing the IRE of different subspaces and to find the model set with minimum order m^*, which minimizes the IRE, we can use the output error. If the output error is non-zero for all m, then $m^* = N$, otherwise, the smallest m for which output error is zero is $m^* = M$. Figure(4-1) shows the behavior of both the output and impulse response error. Figure(4-2) shows the output and IRE behavior in presence of the additive noise. In this scenario the output error is a decreasing function of m. However, regardless of M, which can be less than or greater than N, the IRE is minimized at some point m^*. The optimum order m^* is less than or equal to M and might be less than or equal to N. In the next section we elaborate the relationship between the output error and the IRE in presence of the additive noise.

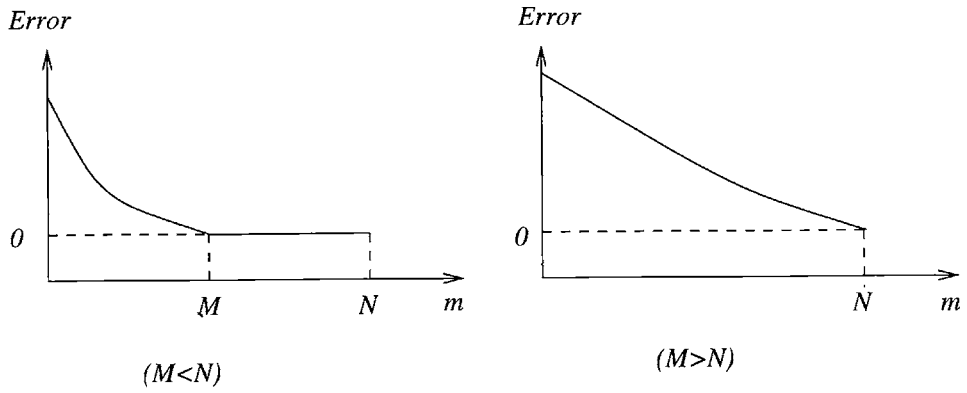

Figure 4-1: *Output and impulse response errors in absence of an additive noise: Left figure shows the "behavior" of both output error and IRE when the length of h, M, is less than N. Right figure shows the behavior of the output error when $M \geq N$. In this case the IRE is also a decreasing function of m. Although here the output error is zero for $m = N$, the impulse response error might still be none zero for all m.*

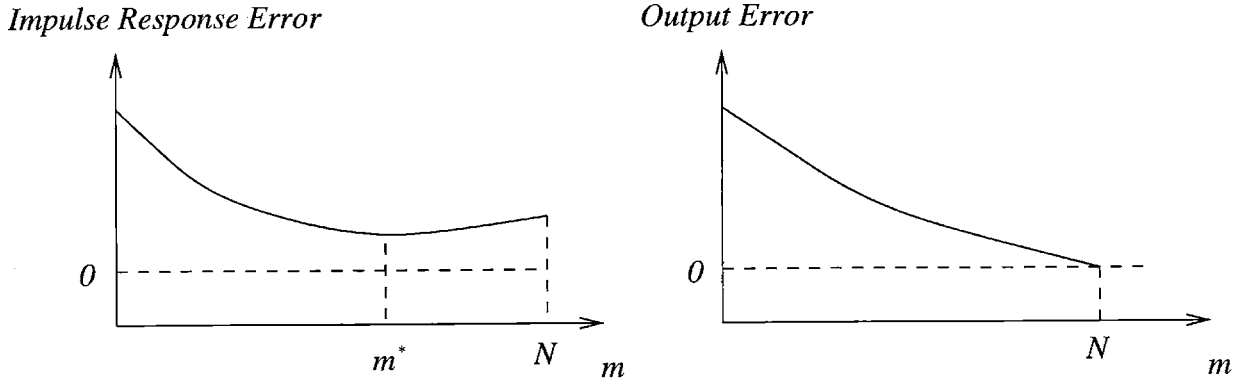

Figure 4-2: *Impulse response error and output error for a subspace of order m*

4.2.2 Impulse Response and Output Spaces

Figure(4-3) shows the output space and the impulse response space for a finite data of length N (both are R^N). In the output space, $\bar{y}$ is the noiseless output, $\bar{y} = h^N * u$. As the figure shows

$$h^N = \begin{bmatrix} h_m^N \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \Delta_m \end{bmatrix}. \tag{4.15}$$

Transformation of the impulse response by the two matrices A_m and B_m,in (4.4), to an element in the output space results

$$\bar{y} = [A_m \;\; B_m] \left(\begin{bmatrix} h_m^N \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \Delta_m \end{bmatrix} \right) \tag{4.16}$$

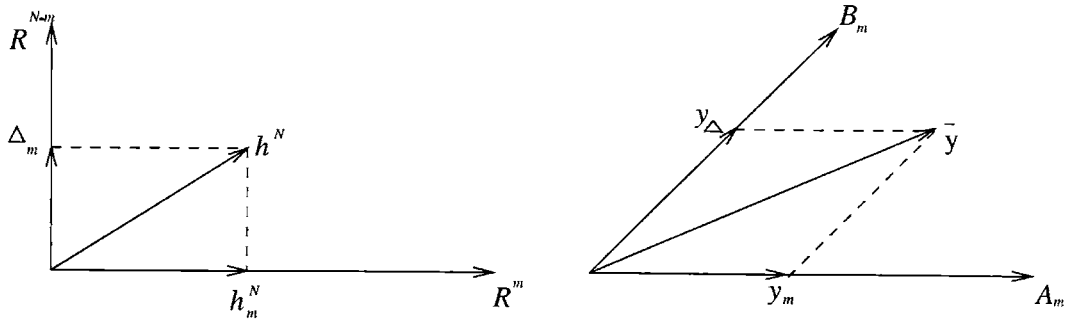

Figure 4-3: *Left figure: h^N, the first N elements of the impulse response. Right figure: $\bar{y}$, the noiseless output.*

where A_m forms a basis for the outputs resulted from elements of subspace S_m. Therefore, matrix $[A_m \quad B_m]$, which is a $N \times N$ full rank matrix, transforms h^N to two elements

$$y_m = A_m h_m^N, \quad y_\Delta = B_m \Delta_m. \tag{4.17}$$

While the elements of the impulse response, h_m^N and Δ_m, are orthogonal, the orthogonality of the elements of the output, y_m and y_Δ, depends on orthogonality of the matrices A_m and B_m.

The observed output is the output of the system corrupted by the additive white Gaussian noise. The least-square estimate of the impulse response, $\hat{h}_m^N$, is obtained such that the distance between the output estimate, $\hat{y}_m = A_m \hat{h}_m^N$, and the output, y, is minimized

$$\hat{h}_m^N = \arg\min_{\bar{h}} ||y - A_m \bar{h}||^2. \tag{4.18}$$

The solution is found by projecting the output, y, on the space spanned by A_m. The solution provides both the projection, $\hat{y}_m$, and the impulse response estimate, $\hat{h}_m^N$. Figure(4-4) shows the estimate of the output using the least-square method. The left figure shows one sample of the noisy output, y, and the projection, $\hat{y}_m$. As the figure shows, the estimate of the impulse response in each subspace is biased, i.e., $|y_m - \hat{\bar{y}}_m| \neq 0$, if and only if columns of A_m and B_m are dependent. If subspaces spanned by A_m and B_m are orthogonal then the estimate is unbiased. Later we show that these two subspaces are asymptotically orthogonal for independent identically distributed (IID) inputs. The thick segment in the right figure shows the impulse response estimates for additive noise such that $|w| \leq \sigma_w$, where σ_w is the noise variance. Figure(4-5) shows the behavior of the IRE in presence of noise. The figure shows the error results for the same system when the output is corrupted by two noises with different variances, σ_1 and σ_2, where $\sigma_1 < \sigma_2$. We call the two cases setting 1

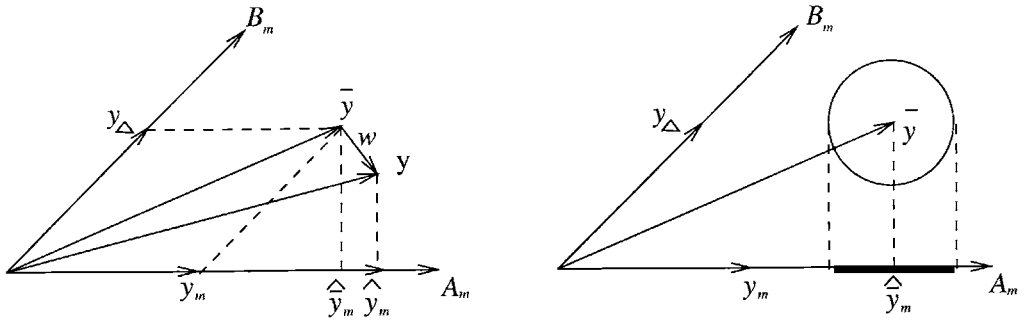

Figure 4-4: *Left figure: an example of the observed noisy data, $y = \bar{y} + w$, and the estimation result for subspace S_m. Right figure: the thick segment in direction of y_m is the estimated output, $\hat{y}_m$, for when the noise belongs to the set $W_\sigma = \{w| \ ||w|| \leq \sigma_w\}$.*

and setting 2 respectively. The left figure is the IRE results of the first setting and the right figure is the IRE results of the second setting. The thick segments on the subspace S_m in both cases are the impulse response estimates when the noise belongs to $W_1 = \{w| \ ||w|| \leq \sigma_1\}$ for the first setting and $W_2 = \{w| \ ||w|| \leq \sigma_2\}$ for the second setting. It is worth mentioning that the probability of these two sets happening is the same. The circles in both figures represent all possible impulse response estimates in space R^N with the same assumption on the additive noise. We can compare the

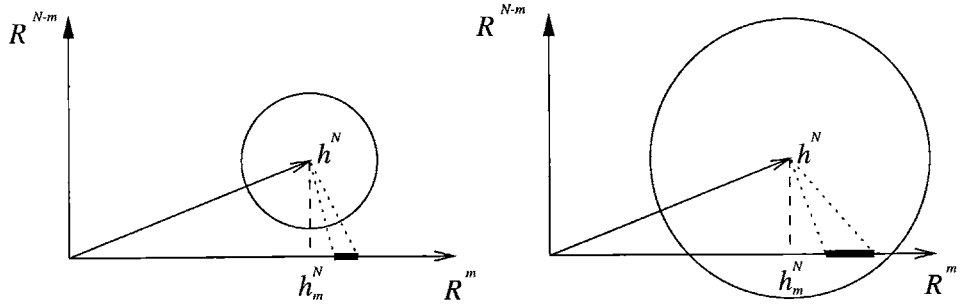

Figure 4-5: *The impulse response estimates of a system in two settings. Setting 1: Additive noise has variance σ_1 and noise belongs to $W_1 = \{w| \ ||w|| \leq \sigma_1\}$. Setting 2: Noise has variance σ_2 where $\sigma_1 < \sigma_2$ and belongs to $W_2 = \{w| \ ||w|| \leq \sigma_2\}$. Left figure shows the impulse response for the first setting and right figure shows the impulse response for the second setting. The thick segments are the impulse response estimates in S_m, $\hat{h}^N_m$. The circles are the impulse response estimates in S_N.*

worst case error of IRE in each setting for the estimate in R^N and S_m. As the figure shows, in the first setting the worst case error of the estimate in S_m is larger than that of R^N. However for the second setting the worst case error in R^N is larger than that in S_m.

4.3 Bounds on the Impulse Response Error

Both SIRE $v_m = ||\hat{h}_m^N - h_m^N||^2$ in (4.6) and IRE $z_m = ||\hat{h}_m^N - h^N||^2$ in (4.7) are random variables which are quadratic forms of the additive Gaussian random variable w. If σ_{mi} and u_{mi} are the ith singular value and corresponding vector of the singular decomposition of C_m in (4.8), then we have

$$v_m = \sum_{i=1}^{m} (\mu_i + m_i)^2 \tag{4.19}$$

$$z_m = v_m + ||\Delta_m||^2 \tag{4.20}$$

where $\mu_i = \sqrt{\sigma_{mi}} u_{mi}^T w$.

If the probability distribution of these random variables are known, and the estimate of h_m in each subspace is calculated $\hat{h}_m$, which subspace is representing h the best?, *i.e.*, which subspace estimate is eventually picked? The important random variable to be checked is IRE, z_m. One answer to this question is to pick the subspace for which expected value of this error is minimum,

$$S_{m^*} = \arg\min_m E(Z_m). \tag{4.21}$$

On the other hand variance of this random variable also plays an important role in this comparison. We know that the estimate $\hat{h}_m$ produces one sample of the random variable Z_m. How close this sample is to the expected value? We can assign the probabilistic region of confidence around the mean, i.e., with probability $P1$, the following holds

$$|z_m - E(Z_m)| \leq J_1(m). \tag{4.22}$$

Therefore, to compare estimates of different subspaces we suggest comparing

$$E(Z_m) + J_1(m) \tag{4.23}$$

of the different subspaces and pick the subspace for which this criterion is minimized

$$S_{m^*} = \arg\min_m E(Z_m) + J_1(m) \tag{4.24}$$

Then it can be said that with probability of confidence of $P1$ the worst case error is minimized for subspace S_{m^*}.

What if only the second order statistics of these random variables are available?

One option is to use the Chebychev inequality [1]:

$$S_{m^*} = \arg\min_m E(Z_m) + \beta\sqrt{\text{var}Z_m} \tag{4.25}$$

for which the confidence region is given with probability $P1 \geq 1 - \frac{1}{\beta^2}$.

Also, if m is large enough the Central Limit Theorem can be used to estimate the distribution of these random variables with Gaussian distributions (See appendix C.2). In this case the same criterion in (4.25) valid with probability $P_1 = Q(\beta)$. In the following we calculate the expected value and variance of SIRE and IRE.

Expected value and variance of SIRE and IRE The expected value and the variance of these random variables are (see appendix C.1).

$$\begin{aligned} E(||\hat{H}_m^N - h_m^N||^2) &= tr(C_m)\sigma_w^2 + (B_m\Delta_m^N)^T C_m B_m \Delta_m^N, & (4.26)\\ E(||\hat{H}_m^N - h^N||^2) &= tr(C_m)\sigma_w^2 + (B_m\Delta_m^N)^T C_m B_m \Delta_m^N + ||\Delta_m^N||^2, & (4.27) \end{aligned}$$

where $tr(F)$ is the trace of matrix F, $\sum_i F_{ii}$, and the variances are

$$\begin{aligned} \text{var}(||\hat{H}_m^N - h^N||^2) &= \text{var}(||\hat{H}_m^N - h_m^N||^2)\\ &= \text{var}(w^T C_m w) + 4(B_m\Delta_m^N)^T C_m^2 B_m \Delta_m^N \sigma_w^2, & (4.28) \end{aligned}$$

The noise related components of the expected values and variances, $tr(C_m)\sigma_w^2$ and $\text{var}(w^T C_m w)$, can be computed using the input and the AWGN second order statistics (Appendix C.4). The goal is to estimate the unmodeled dynamics effects, in the expected values and variance, which are quadratic forms

$$\begin{aligned} mc &= (B_m\Delta_m^N)^T C_m B_m \Delta_m^N, & (4.29)\\ v_c &= (B_m\Delta_m^N)^T C_m^2 B_m \Delta_m^N, & (4.30) \end{aligned}$$

and $||\Delta_m^N||^2$. These elements are in form of

$$(\Delta_m^N)^T D_i \Delta_m^N, \tag{4.31}$$

where

$$\begin{aligned} m_c &: D_1 = B_m^T C_m B_m, & (4.32)\\ v_c &: D_2 = B_m^T C_m^2 B_m, & (4.33) \end{aligned}$$

[1] For a random variable x with expected value E_x and variance of σ_x the Chebychev inequality is

$$\text{Prob}(|x - E_x| \geq t) \leq (\frac{\sigma_x}{t})^2.$$

Note that for $t = \beta\sigma_x$

$$\text{Prob}(|x - E_x| \leq \beta\sigma_x) \geq 1 - \frac{1}{\beta^2}.$$

$$||\Delta_m^N||^2 \quad : \quad D_3 = I_{(N-m)\times(N-m)}. \tag{4.34}$$

In our problem setting the prior assumption is that $||\Delta_m^N||^2$ is bounded but no prior upper bound is available.

4.3.1 The Output Error

The output error (4.13) is a Chi-square random variable of order $N-m$ for which

$$E(\frac{1}{N}||Y-\hat{Y}_m^N||_2^2) = (1-\frac{m}{N})\sigma_w^2 + g_m, \tag{4.35}$$

$$\text{var}(\frac{1}{N}||Y-\hat{Y}_m^N||_2^2) = (1-\frac{m}{N})\frac{2\sigma_w^4}{N} + \frac{4\sigma_w^2}{N}g_m., \tag{4.36}$$

where g_m is the effect of the unmodeled dynamics

$$g_m = \frac{1}{N}(B_m\Delta_m^N)^T G_m B_m \Delta_m^N. \tag{4.37}$$

See appendix C.2 for the details on calculation of the expected value and the variance.

How can we use one sample of this random variable to estimate unmodeled dynamics effects for SIRE and IRE? We suggest to use the observed output error to validate g_m probabilistically. Bounds on this quadratic form of the unmodeled dynamics can then be used to provide bounds on the quadratic forms m_c,v_c and $||\Delta_m^N||^2$ in (4.32), (4.33) and (4.34).

The first step is to choose probability of validation $P2$. Next is to find the bounds for which $X_m(g_m) = \frac{1}{N}||Y-\hat{Y}_m^N||_2^2$ is around its mean with probability $P2$. For such probability we can use the table of Chi-square random variables of order $N-m$ and find $L1$ such that

$$\Pr(|X_m(g_m) - E(X_m(g_m))| \leq L1) = P2 \tag{4.38}$$

Therefore $L1$ is a function of $P2, \sigma_w, N, m, g_m$. Next step is validation of g_ms for which the observed $x_m = \frac{1}{N}||y-\hat{y}_m^N||_2^2$ lies in the region $E(X_m) \pm L1$. Such validation provides an upper and an lower bounds on g_m.

By using the Central Limit Theorem, the cumulative distribution function(cdf) of the output error can be estimated with the cdf of a Gaussian random variable asymptotically (see appendix C.2). Therefore instead of table of Chi-square distribution we can use the $Q(\cdot)$ function.

If Gaussian random variable X_m has mean m_X and variance σ_X^2, then

$$\Pr(m_X - \alpha\sigma_X < X_m < m_X + \alpha\sigma_X) = Q(\alpha). \tag{4.39}$$

Given the observed output error x_m, we find the feasible set of g_ms for which x_m is

within $\alpha\sqrt{\text{var}X_m}$ distance of its mean, i.e., we calculate g_ms for which

$$|x_m - (g_m + m_w)| \leq \alpha\sqrt{4\sigma_w^2\frac{g_m}{N} + v_m}, \tag{4.40}$$

where

$$m_w = (1 - \frac{m}{N})\sigma_w^2, \quad v_m = (1 - \frac{m}{N})\frac{2\sigma_w^4}{N}. \tag{4.41}$$

The expected value g_m+m_w and variance $4\sigma_w^2\frac{g_m}{N}+v_m$are resulted from (4.35) and(4.36). If the output error is estimated with a Gaussian distribution, this feasible set is valid with probability $Q(\alpha)$.

Lemma 1 The result of validation of (4.40), with $x_m = \frac{1}{N}||y - \hat{y}_m^N||_2^2$, for feasible g_ms provides the following upper and lower bound for g_m,

$$Lg_m \leq g_m \leq Ug_m, \tag{4.42}$$

- If $x_m \leq m_w - \alpha\sqrt{v_m}$, there in no valid g_m.
- If $m_w - \alpha\sqrt{v_m} \leq x_m \leq m_w + \alpha\sqrt{v_m}$,

$$Lg_m = 0, \tag{4.43}$$

$$Ug_m = x_m - m_w + \frac{2\alpha^2\sigma_w^2}{N} + \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_m - \frac{3}{2}m_w}. \tag{4.44}$$

- If $m_w + \alpha\sqrt{v_m} \leq x_m$,

$$Lg_m = x_m - m_w + \frac{2\alpha^2\sigma_w^2}{N} - \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_m - \frac{3}{2}m_w}. \tag{4.45}$$

$$Ug_m = x_m - m_w + \frac{2\alpha^2\sigma_w^2}{N} + \frac{2\alpha\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_m - \frac{3}{2}m_w}. \tag{4.46}$$

Proof In appendix C.3 $\diamondsuit$

Note that to avoid the first case we have to choose α large enough such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_m}{\sigma_w^2}\right). \tag{4.47}$$

4.3.2 Probabilistic Upper and Lower Bounds for SIRE and IRE

We suggest a method to find bounds for the quadratic form m_c, v_c, $||\Delta_m^N||^2$ in (4.32),(4.33),(4.34) by using the validation result of Lemma 1 . The unmodeled

dynamics effect in the output error is in form

$$g_m = (\Delta_m^N)^T D_4 \Delta_m^N \tag{4.48}$$

where $D_4 = \frac{1}{N} B_m^T G_m B_m$. Therefore, given that $Lg_m \leq g_m \leq Ug_m$, bounds L_{m_c}, U_{m_c}, L_{v_c}, U_{v_c}, L_{m_Δ} and U_{m_Δ} are calculated

$$\begin{aligned} Lm_c &\leq m_c \leq Um_c,\ Lv_c \leq v_c \leq Uv_c, \\ Lm_\Delta &\leq ||\Delta_m^N||^2 \leq Um_\Delta. \end{aligned} \tag{4.49}$$

This step is a deterministic procedure which solves a constrained quadratic optimization problem.

Theorem 4.3.2 Given the bounds on $g_m = \frac{1}{N}(B_m \Delta_m^N)^T G_m B_m \Delta_m^N$, $L_{gm} \leq g_m \leq U_{gm}$, for the two random variables SIRE, v_m, and IRE, z_m, with probability larger than $1 - \frac{1}{\beta^2}$ the following hold

$$\max\{0, Lm_c + tr(C_m)\sigma_w^2 - \beta J_m\} \leq V_m \leq Um_c + tr(C_m)\sigma_w^2 + \beta J_m, \tag{4.50}$$

$$\begin{aligned} \max\{0, Lm_c + Lm_\Delta + tr(C_m)\sigma_w^2 - \beta J_m\} &\leq \\ Z_m \leq Um_c &+ Um_\Delta + tr(C_m)\sigma_w^2 + \beta J_m, \end{aligned} \tag{4.51}$$

where $J_m = \sqrt{Uv_c + \mathrm{var}(w^T C_m w)}$.

Proof By using the Chebychev inequality for random variables V_m or Z_m and bounds on (4.49), the proof is done. $\diamondsuit$

Asymptotic Behavior of the Error Estimates and Conditions on α and β

In appendix C.2 we show that the noise related part of the output error, $v_m = \frac{1}{N} w^T G_m w$ in (4.13), is such that

$$\alpha\sqrt{v_m} = \alpha(1 - \frac{m}{N})\frac{\sqrt{2}(\sigma_w^2)}{\sqrt{N}}.$$

Therefore, by choosing α as a function of N, α_N, such that $\lim \frac{\alpha_N}{\sqrt{N}} = 0$, the upper and lower bounds of g_m, in Lemma 1, approach each other as N grows. To have the validation probability approaching one as N grows, we choose α_N such that $\lim_{N\to\infty} \alpha_N = \infty$, therefore $Q(\alpha_N) \to 1$. For example one candidate for α_N is $\log(N)$.

While as N grows, with proper choice of α_N, the lower and upper bounds of g_m (4.37) in lemma 1 approach each other, the upper and lower bounds of unmodeled dynamics related terms of SIRE and IRE in (4.49) might not approach each other as N grows. Note that since the input is quasi-stationary limits of these bounds exist and are finite. The asymptotic behavior of these bounds depends on the structure of the input. If the input is independent identically distributed random variable, the upper and lower bounds of the unmodeled part of SIRE and IRE also converge to each other as N grows.

In Theorem 4.3.2 to have the probabilistic bounds with probability which goes to one as N grows, we pick β as a function of N such that $\lim_{N\to\infty}\beta_N = \infty$. To have finite values for the upper and lower bounds of V_m and Z_m, the term $\beta\sqrt{Uv_c + \mathrm{var}(w^T C_m w)}$ in (4.50) and (4.51) has to be finite for all N.

Corollary 1 For the variance of $w^T C_m w$ we have

$$Uv_c + \mathrm{var}(w^T C_m w) \leq \frac{k}{N}, \tag{4.52}$$

for some finite number k.

proof In appendix C.4.2. $\diamondsuit$

Therefore, as long as the rate of growth of β_N is such that $\lim_{N\to\infty}\frac{\beta_N}{\sqrt{N}} = 0$, the upper and lower bounds in (4.50) are only functions of the bounds of the mean m_c and unmodeled dynamics norm $||\Delta_m^N||^2$ in (4.49). Hence, to have tight bounds on the errors with validation and region confidence probabilities which goes to one asymptotically, the necessary conditions for α and β are :

- $\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right)$,
- $\lim_{N\to\infty}\alpha_N = \infty$, $\lim\frac{\alpha_N}{\sqrt{N}} = 0$,
- $\lim_{N\to\infty}\beta_N = \infty$, $\lim_{N\to\infty}\frac{\beta_N}{\sqrt{N}} = 0$,

The first condition is from Lemma one.

4.4 Independent Identically Distributed Input

Consider a subset of quasi-stationary inputs, sequence of independent identically distributed(IID) random variables with unit variance and zero mean. An example of such input is a Bernoulli sequence of ± 1 which is commonly used in communications.

Theorem 4.4 If the input of the system is IID, the expected value and variance of SIRE and IRE in (4.26), (4.27) and (4.28) are

$$E(||\hat{H}_m^N - h_m^N||^2) = \sigma_w^2\frac{m}{N} + \frac{m}{N}g_m^N + O(\frac{1}{N}), \tag{4.53}$$

$$E(||\hat{H}_m^N - h^N||^2) = \sigma_w^2\frac{m}{N} + \frac{m}{N}g_m^N + ||\Delta_m^N||^2 + O(\frac{1}{N}), \tag{4.54}$$

$$\mathrm{var}(||\hat{H}_m^N - h^N||^2) = \mathrm{var}(||\hat{H}_m^N - h_m^N||^2) \tag{4.55}$$

$$\leq \sigma_w^4\frac{2m}{N^2} + \frac{m^2}{N^2}(g_m^N)^2 + O(\frac{1}{N^2}), \tag{4.56}$$

where g_m^N is

$$\begin{aligned} g_m^N &= \frac{1}{N}E||B_m\Delta_m||^2 \\ &= \frac{N-m}{N}h_{m+1}^2 + \frac{N-m+1}{N}h_{m+2}^2 + \cdots + \frac{1}{N}h_N^2 \\ &= \sum_{i=m+1}^{N}\frac{N-i+1}{N}||h_i||^2. \end{aligned} \tag{4.57}$$

proof See appendix D.2. $\diamondsuit$

4.4.1 Estimation of g_m^N

Theorem 4.4 shows that in oder to provide probabilistic bounds on IRE and SIRE we need to find estimates of g_m^N and $||\Delta_m^N||^2$. For the unmodeled related parts from (4.29), (4.30), we have

$$m_c \approx \frac{m}{N}g_m^N, \quad v_c \leq \frac{m}{N^2}(g_m^N)^2. \tag{4.58}$$

Following the proposed method in the previous section we first use the observed output error to find bounds on g_m, in (4.37). The goal is to use this estimate to find bounds on g_m^N in (4.57) and $||\Delta_m^N||^2$.

Lemma 2 If the input of the system is IID, g_m in (4.37) is also a random variable with the following expected value and variance

$$\begin{aligned} E(g_m) &= E\frac{1}{N}(B_m\Delta_m^N)^T G_m B_m \Delta_m^N \\ &= (1-\frac{m}{N})g_m^N + O(\frac{1}{N}), \end{aligned} \tag{4.59}$$

$$\text{var}g_m \leq k_m\frac{(g_m^N)^2}{N} + O(\frac{1}{N}), \tag{4.60}$$

where

$$k_m = l + m \tag{4.61}$$

and l is an upper bound for $\frac{||\Delta_m^N||_{H_\infty}}{||g_m^N||_2^2}$ ($||x||_{H_\infty}$ is the H_∞ norm of system with impulse response x).

proof In appendix D.1. $\diamondsuit$

By using Lemma one, we obtain upper and lower bound for random variable g_m with validation probability $Q(\alpha)$. Next we find the feasible set of g_m^Ns for which $L_{gm} \leq g_m \leq U_{gm}$ and g_m is within $\frac{\gamma}{\sqrt{k_m}}\sqrt{\text{var}(g_m)}$ distance of its mean

$$E(g_m) + \frac{\gamma}{\sqrt{k_m}}\sqrt{\text{var}(g_m)} \leq g_m \leq E(g_m) + \frac{\gamma}{\sqrt{k_m}}\sqrt{\text{var}(g_m)}, \tag{4.62}$$

$$g_m^N - \gamma\frac{g_m^N}{\sqrt{N}} \leq g_m \leq g_m^N + \gamma\frac{g_m^N}{\sqrt{N}}. \tag{4.63}$$

Therefore,

$$\frac{Lg_m}{1+\frac{\gamma}{\sqrt{N}}} \leq g_m^N \leq \frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}. \tag{4.64}$$

This provides upper and lower bounds on g_m^N with validation probability greater than $Q(\gamma/k_m)$. Note that practically k_m in (4.61) is not available. We can choose γ as a function of N such that

$$\lim_{N\to\infty} \gamma \to \infty, \quad \lim_{N\to\infty} \frac{\gamma}{\sqrt{N}} \to 0. \tag{4.65}$$

This conditions guarantees the convergence of the upper and lower bound of g_m^N to each other and at the same time the probability $Q(\gamma/k_m)$ approaches one as N grows. Note that since k_m is an unknown but bounded number, we know that $\gamma/\sqrt{k_m}$ grows as N grows. The rate of growth however is unknown.

4.4.2 Bounds on $||\Delta_m^N||^2$

For the expected value of IRE, bounds on $||\Delta_m^N||^2$ is needed. The estimate of g_m^N in (4.57) provides a lower bound since

$$g_m^N \leq ||\Delta_m^N||^2. \tag{4.66}$$

An upper bound for $||\Delta_m^N||^2$ can not be provided by using g_m^N, however, for $\sum_{i=m+1}^{M} ||h_i||^2$, when $M < N$, we have

$$\begin{aligned}\sum_{i=m+1}^{M} ||h_i||^2 &\leq \sum_{i=m+1}^{M} \frac{N-i+1}{N-M}||h_i||^2 \leq \frac{N}{N-M}\sum_{i=m+1}^{M}\frac{N-i+1}{N}||h_i||^2, \\ &\leq \frac{1}{1-\frac{M}{N}}g_m^N. \end{aligned} \tag{4.67}$$

We can choose M, as a function of N, such that

$$\lim_{N\to\infty} M(N) = \infty, \quad \lim_{N\to\infty} \frac{M(N)}{N} = 0 \tag{4.68}$$

and therefore

$$\lim_{N\to\infty} \sum_{i=m+1}^{M} ||h_i||^2 \to ||\Delta_m^N||^2. \tag{4.69}$$

As a result, for large N, g_m can be a tight estimate of $||\Delta_m^N||^2$.

Corollary 2 For IRE and SIRE, with validation probability of $Q(\alpha)$ and boundary probabilities greater than $Q(\gamma/k_m)$, with probability greater than $1-1/\beta^2$ the bounds in Theorem 4.3.2 are

$$||\hat{h}_m^N - h_m^N||_2^2 \leq \frac{m}{N}\sigma^2 + \frac{m}{N}\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}} + \frac{\beta\sqrt{m}}{N}\sqrt{2(\sigma^2)^2 + m\left(\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}\right)^2} \tag{4.70}$$

$$||\hat{h}_m^N - h_m^N||_2^2 \quad \geq \tag{4.71}$$
$$\max\left\{0, \frac{m}{N}\sigma^2 + \frac{m}{N}\frac{Lg_m}{1+\frac{\gamma}{\sqrt{N}}} - \frac{\beta\sqrt{m}}{N}\sqrt{2(\sigma^2)^2 + m\left(\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}\right)^2}\right\}$$

$$||\hat{h}_m^N - h^N||_2^2 \leq \frac{m}{N}\sigma^2 + (1+\frac{m}{N})\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}} + \frac{\beta\sqrt{m}}{N}\sqrt{2(\sigma^2)^2 + m\left((\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}\right)^2} \tag{4.72}$$

$$||\hat{h}_m^N - h^N||_2^2 \quad \geq \tag{4.73}$$
$$\max\left\{0, \frac{m}{N}\sigma^2 + (1+\frac{m}{N})\frac{Lg_m}{1+\frac{\gamma}{\sqrt{N}}} - \frac{\beta\sqrt{m}}{N}\sqrt{2(\sigma^2)^2 + m\left(\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}})\right)^2}\right\}$$

where Lg_m and Ug_m are lower and upper bounds on the unmodeled part of the output error obtained in lemma 1.

proof The proof follows from Theorem 4.3.2 and Theorem 4.4. $\diamondsuit$

Note that when m, the order of subspace S_m, is large enough both IRE and SIRE can be estimated by Gaussian random variables using the Central Limit Theorem. In this case the probability of $1-\frac{1}{\beta^2}$ is replaced by $Q(\beta)$.

Rate of Convergence If we choose α, β and γ as functions of N such that

$$\lim \alpha_N = \infty, \quad \lim \beta_N = \infty, \quad \lim \gamma_N = \infty, \tag{4.74}$$

the probabilities $Q(\alpha)$, $Q(\beta)$ (or $1-\frac{1}{\beta^2}$) and $Q(\gamma/k_m)$ go to one as N grows. Also if we have

$$\lim \frac{\alpha_N}{\sqrt{N}} = 0, \quad \lim \frac{\beta_N}{N} = 0, \quad \lim \frac{\gamma_N}{\sqrt{N}} = 0, \tag{4.75}$$

then Ug_m and Lg_m approach each other as N grows. With these conditions on α, β and γ inequalities (4.70),(4.71),(4.72),(4.73), give tight estimates for the rate of

convergence of the errors to their limits. In the limit Ug_m and Lg_m both converge to $\max\left\{0, \frac{1}{N}||y - \hat{y}_m^N||_2^2 - \sigma_w^2\right\}$. For SIRE and IRE in the limit we have

$$\lim_{N\to\infty} ||\hat{h}_m^N - h_m^N||_2^2 = 0, \tag{4.76}$$

$$\lim_{N\to\infty} ||\hat{h}_m^N - h^N||_2^2 = \max\left\{0, \frac{1}{N}||y - \hat{y}_m^N||_2^2 - \sigma_w^2\right\}. \tag{4.77}$$

4.5 Orthonormal Basis

In previous sections we examined estimation of an impulse response of a LTI system in time domain. Here we generalize the same approach to estimation of the system impulse response using basis functions. The problem of system identification implementing orthonormal basis is a subject that attracted number of researchers [25, 57]. The main advantage of the following approach, unlike the existing methods, is that it separates the unmodeled dynamics effects and noise effects in the identification process.

Consider basis functions s_is for representation of the impulse response of length N. The orthonormal basis functions s_i are such that

$$||s_i||^2 = 1, \quad s_i^T s_j = 0 \quad i \neq j. \tag{4.78}$$

The impulse response of a stable LTI system, T, is represented as follows

$$T = \sum_1^\infty h(i)s_i. \tag{4.79}$$

where h is the vector of coefficients of T in the basis. Finite, N, points of input and output, (y^N, u^N) of the system is available. The noisy output is in form of

$$y = T * u + w, \tag{4.80}$$

where w is a zero mean additive white Gaussian noise. Note that only the first N taps of T relate N taps of the input to N taps of the output. Therefore, we can at most estimate the first N taps of T, $T_1^N \in R^N$, using the finite data. We assume that the bases s_is are such that if we choose m of those basis i.e. $[s_1, \cdots, s_m]$, the first N elements of each s_i, $1 \leq i \leq m$ form a orthogonal basis and matrix $[(s_1)_1^N, \cdots, (s_m)_1^N]$ has full rank m. To span the space R^N, the rest of the bases are $[\bar{s}_{m+1}, \cdots, \bar{s}_N]$, such that the bases of this subspace are orthogonal to the elements of $[(s_1)_1^N, \cdots, (s_m)_1^N]$. Also

$$\lim_{N\to\infty} \bar{s}_i = s_i. \tag{4.81}$$

For the first N elements of T, T^N, we have

$$T^N = \sum_1^M h^N(i)(s_i)_1^N + \sum_{m+1}^N h^N(i)(\bar{s}_i)_1^N \tag{4.82}$$

and the assumption is that

$$\lim_{N\to\infty} T_1^N = T \tag{4.83}$$

$$\lim_{N\to\infty} h^N = h. \tag{4.84}$$

Therefore the input-output relationship of the system is as follows

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ u_2 & u_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_N & u_{N-1} & \cdots & u_1 \end{bmatrix} \times$$

$$[(s_1)_1^N \cdots (s_m)_1^N \ (\bar{s}_{m+1})_1^N \cdots (\bar{s}_N)_1^N] \begin{bmatrix} h^N(1) \\ h^N(2) \\ \vdots \\ h^N(N) \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}$$

$$= \begin{bmatrix} A_m(N) & B_m(N) \end{bmatrix} \begin{bmatrix} h_m^N \\ \\ \Delta_m^N \end{bmatrix} + w_1^N, \tag{4.85}$$

where $h_m^N = [h^N(1), \cdots, h^N(m)]^T$, $\Delta_m^N = [h^N(m+1), \cdots, h^N(N)]^T$ and

$$A_m(N) = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ u_2 & u_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_N & u_{N-1} & \cdots & u_1 \end{bmatrix} \begin{bmatrix} s_1(1) & s_2(1) & \cdots & s_m(1) \\ s_1(2) & s_2(2) & \cdots & s_m(2) \\ \vdots & \ddots & \ddots & \vdots \\ s_1(N) & s_2(N) & \cdots & s_m(N) \end{bmatrix} \tag{4.86}$$

$$B_m(N) = \begin{bmatrix} u_1 & 0 & \cdots & 0 \\ u_2 & u_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ u_N & u_{N-1} & \cdots & u_1 \end{bmatrix} \begin{bmatrix} \bar{s}_{m+1}(1) & \bar{s}_{m+2}(1) & \cdots & \bar{s}_N(1) \\ \bar{s}_{m+1}(2) & \bar{s}_{m+2}(2) & \cdots & \bar{s}_N(2) \\ \vdots & \ddots & \ddots & \vdots \\ \bar{s}_{m+1}(N) & \bar{s}_{m+2}(N) & \cdots & \bar{s}_N(N) \end{bmatrix} \tag{4.87}$$

To estimate T^N in the subspace with order m, i.e., the subspace represented by $(s_1)_1^N, \cdots, (s_m)_1^N$, use the conventional least-square method. We are interested in choosing the subspace which best represent the true system. The approach is identical to what is presented in Chapter 3. Here A_m and B_m in (4.4) are replaced by A_m and B_m in (4.86,4.87) and the rest of the procedure is the same.

4.6 Time Series Analysis

Let $y_1, \cdots, y_N$ be the finite sample of length N from a stationary process y, satisfying the recursion

$$y_n + \sum_{i=1}^{p_0} a_i y_{n-i} = \sigma[w_n + \sum_{i=1}^{q_0} b_i w_{n-i}] \tag{4.88}$$

where w_i is white noise with zero mean and unit variance. This is an ARMA(p_0, q_0) process and the model is assumed to be stable and minimum phase.

The identification problem is as follows. Given an observed data y^N find the estimates of p_0, q_0, σ, a_i and b_i. Here we state a similar problem. Since the model is minimum phase we can rewrite (4.88) in form of

$$y_n - \sum_{i=1}^{N} h_i y_{n-i} = w_n, \tag{4.89}$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} y_1 & 0 & \cdots & 0 \\ y_2 & y_1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ y_N & y_{N-1} & \cdots & y_1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}. \tag{4.90}$$

Then we suggest using the proposed estimation method in this chapter to find estimate of h_is. This part is called the denoising step of the estimation. The estimate of h is $\hat{h}_{S_m^*}$ for the subspace which IRE is minimized. To find an estimate of the parameters, p_0, q_0, σ, a_i and b_i, we suggest minimizing the distance between this estimate of the impulse response and elements of family of ARMA models in (4.88). This part of the procedure is the curve fitting step.

Note that another interesting solution to this problem is given in [37]. It will be interesting to compare the complexity of the algorithms solving the problem in these two methods.

4.7 Additive Colored Noise

Consider the problem of identification of a stable LTI system when the additive noise is colored,

$$y = H * u + G * e \tag{4.91}$$

where G is a minimum phase filter representing the auto-correlation of the noise, $w = G * e$, e is white noise with zero mean and unit variance. With the prior assumption that G or its estimate is available we can generalize the new identification and quality evaluation method. In practical problems the estimate of G is available

by experimenting the output of the system with no exogenous input. Note that, in this case, one method of identification of G is the identification method we described for the time series analysis in section 4.6.

If the estimate of G is available we suggest first to pass the output through G^{-1}. Therefore

$$\bar{y} = G^{-1} * y = T * (G^{-1} * u) + e \tag{4.92}$$

Now we can use the proposed method for finding bounds on the impulse error treating $\bar{y}$ as the output and $G^{-1} * u$ as the input.

4.7.1 Input Design

We can design the input such that the filtered input, by the noise related filter G^{-1}, is IID. In this case we are able to use the asymptotic result of the IID case. For this purpose the input can be designed in form of

$$u = G * v \tag{4.93}$$

where v is IID itself. Therefore the output of the system is

$$\begin{aligned} \bar{y} &= G^{-1} * y = T * (G^{-1} * G * v) + e && (4.94) \\ &= H * v + e && (4.95) \end{aligned}$$

and the identification method in section 4.4 is applicable for identification and quality evaluation of the system impulse response.

4.8 Zero-Pole Representation

One other application of the proposed method is for identification and order estimation of ARMAX model with the following structure

$$y_n + \sum_{i=1}^{p_0} a_i y_{n-i} = \sum_{i=0}^{q_0} b_i u_{n-i} + w_n \tag{4.96}$$

where the additive noise w is zero mean with variance σ_w^2. Here estimation of p_0 and q_0, which represent the number of zero and poles of the system, is as important as identification of a_i and b_i. The problem can be considered in form of

$$y^N = \begin{bmatrix} y_T^N & u_T^N \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + w. \tag{4.97}$$

where y_T^N and u_T^N are the Toeplits matrices generated by terms $\sum_{i=1}^{p_0} a_i y_{n-i}$ and $\sum_{i=0}^{q_0} b_i u_{n-i}$. Note that with this observation the sum of maximum length of a_i and b_i can not exceed N.

4.9 Algorithm for Choosing the Best Finite Number of Parameters

Assume that the length of impulse response, including a possible delay, is less than M where $M << N$. Using the proposed method we find the optimum length of the impulse response, which might be much less than M, and the delay associated with it. First compare all the estimates of the impulse responses of length one. This will need a search of order M. Next we search among subspaces of order two which include this element and choose the subspace which minimizes the impulse response error. This search is of order $M-1$. By continuing this method, the maximum number of searchs is of order $M+(M-1)+\cdots+1=\frac{M(M-1)}{2}$.

4.10 Comparison of the Quality Evaluation Methods

In this chapter our main goal was to find an estimate of the IRE which is the distance between the estimate in the subspace and the true impulse response. Here we compare the proposed quality evaluation method with the methods reviewed in Chapter 2. Note that all the methods in chapter 2, except one, concentrate on finding an estimate of the "subspace" IRE. The one method which focuses on the IRE, with probabilistic assumption on the unmodeled dynamics, fails to separate the effect of unmodeled dynamics and subspace impulse response estimate. The problems with the prior assumption on the unmodeled dynamics of that method reviewed in section 2.3.

In MUDP method, section 2.2.2, an upper bound for the expected value of SIRE is provided. The prior assumption is that there exists an upper bound on norm of $||\Delta_m^N||^2$, i.e., $||\Delta_m^N||^2 \leq \gamma$. With such assumption, for all the subspaces of different order m, m_c defined in (4.29) is upper bounded as follows

$$m_c \leq \gamma \sigma_{max}(D_1) \tag{4.98}$$

where $\sigma_{max}(D_1)$ is the maximum singular value of matrix D_1 defined in (4.32). As a result the provided upper bound of the expected value of SIRE is a decreasing function of m and does not provide a tool for comparison of the subspace estimated, i.e., the provided upper bound is minimized for the subspaces of order one!. In this chapter we proposed a method to use the output error and provide bounds on m_c and $||\Delta_m||^2$ for each subspace separately.

4.10.1 Set-membership Identifications

In review of set membership identification, in section 2.3, we concluded that none of the methods are able to provide "robust convergence" because of the conservative definition of noise.

The proposed method in this chapter is an H_2 identification method. Here we define the noise in a deterministic setting such that it satisfies more properties of a

practical additive noise. The method is proved to be robustly convergent for the IID inputs.

Deterministic Noise

In conventional set membership identification methods the additive noise belongs to a bounded norm set. However, in general more can be said about the additive noise by restricting the set with additional constraints on the correlation of the noise with the input or with itself. Paganini introduces such set descriptions in [35].

Here to bridge the gap between set description and stochastic additive noise, and inspired by the what Paganini suggests in [35], we check the richness of a set with the probability that a stochastic noise is a member of set. Lets assume that the additive noise in (4.1) belongs to the following set, W,

$$\begin{aligned} W = \{v| \ |\sum_{i=1}^{N} \frac{v_i a_i}{N}| \ &\leq \ \frac{\alpha}{\sqrt{N}} \sqrt{R_1^v(0)} \sqrt{R_1^a(0)}, \ |R_a^v(\tau)| \leq \frac{\alpha}{\sqrt{N}} R_1^v(0) \sqrt{R_1^a(0)} \quad (4.99) \\ \tau \neq 0, \ |R_a^v(0) \ &- \ R_1^v(0) \sum \frac{a_i}{N}| \leq \alpha \frac{1}{\sqrt{N}} \sqrt{R_1^a(0)} \sqrt{R_1^{v^2}(0) - (R_1^v(0))^2}\} \end{aligned}$$

where a is a bounded power sequence and

$$R_a^v(\tau) = \frac{1}{N} \sum_{i=1}^{N} v_i v_{i+\tau} a_i \quad (4.100)$$

$$R_1^v(\tau) = \frac{1}{N} \sum_{i=1}^{N} v_i v_{i+\tau}. \quad (4.101)$$

The method presented in this section provides the worst-case IRE in each subspace S_m

$$\sup_{w \in W} ||h^N - \hat{h}_m^N||^2, \ \inf_{w \in W} ||h^N - \hat{h}_m^N||^2. \quad (4.102)$$

As N grows, an additive white Gaussian noise is a member of W with probability $[Q(\alpha)]^{N+2}$ and α can be chosen as a function of N such that α_N goes to infinity as N grows [2]. Therefore the set W is rich enough since AWGN is a member of it asymptotically.

[2]Each of the $N+2$ conditions of the set W are satisfied by AWGN with probability of $Q(\alpha)$ and since each of the conditions are asymptotically Gaussian and uncorrelated, therefore they are also independent, so the probability of all the events is the product of the probability of each event $[Q(\alpha)]^{N+2}$

4.11 Simulation Results

Figure (4-6) shows the impulse response used in the simulation, $h(n) = .3(.5)^{n-1} + 3(n-1)(.8)^{n-1}$. The input is an IID Bernoulli sequence of ± 1 and the noise is additive, white and Gaussian. Figures (4-7), (4-8) show SIRE and IRE errors. The bounds on

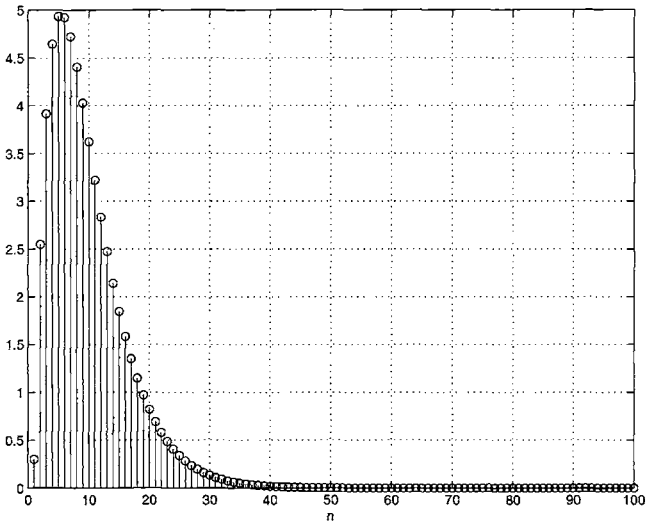

Figure 4-6: *The impulse response of the system.*

the error are calculated based on the upper and lower bounds given in (4.70),(4.71), (4.72), (4.73). The solid line in both figures is the estimate of expected value of the errors

$$E(||\hat{h}_m^N - h_m^N)||^2) \approx tr(C_m)\sigma_w^2 + \frac{m}{N}(\max\{0, \frac{1}{N}||y - \hat{y}_m^N||_2^2 - (1 - \frac{m}{N})\sigma_w^2\}) \quad (4.103)$$

$$E(||\hat{h}_m^N - h^N)||^2) \approx tr(C_m)\sigma_w^2 + \max\{0, (1 + \frac{m}{N})(\frac{1}{N}||y - \hat{y}_m^N||_2^2 - (1 - \frac{m}{N})\sigma_w^2)\} \quad (4.104)$$

Figure (4-9) shows the simulation results for inputs with different length and fixed noise variance, $\sigma_w = .02$. Figures (4-10) and (4-11) show the SIRE and IRE respectively for when $N = 400$, and for two different noise variances, $\sigma_w = .2$, $\sigma_w = .02$.

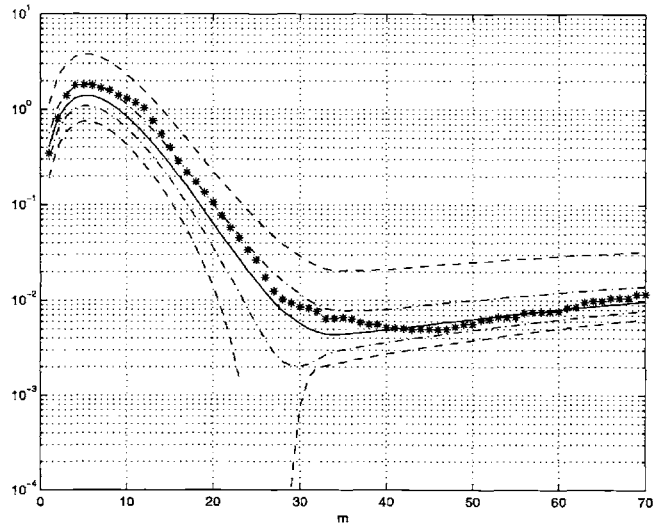

Figure 4-7: *'*': Subspace impulse response error, $||\hat{h}_m^N - h_m^N||_2^2$, for $\sigma_w = .2$ and N=400. '–': Bounds calculated for $\alpha^2 = \gamma^2 = \beta = log(N)$, '-.':Bounds calculated for $\alpha^2 = \gamma^2 = \beta = 2$. Solid line: Estimate of expected value of SIRE, $E||\hat{h}_m^N - h_m^N||_2^2$, in (4.103).*

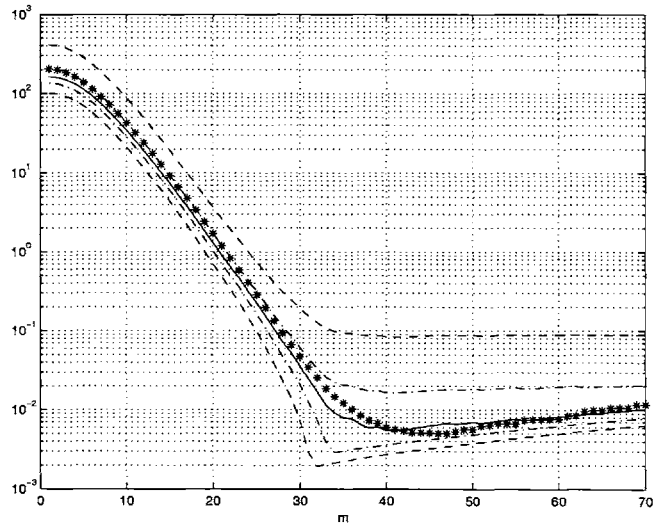

Figure 4-8: *'*': Impulse response error, $||\hat{h}_m^N - h^N||_2^2$, for $\sigma_w = .2$ and N=400. '–': Bounds calculated for $\alpha^2 = \gamma^2 = \beta = log(N)$. '-.': Bounds calculated for $\alpha^2 = \gamma^2 = \beta = 2$. Solid line: Estimate of expected value of IRE, in $E||\hat{h}_m^N - h^N||_2^2$ (4.104).*

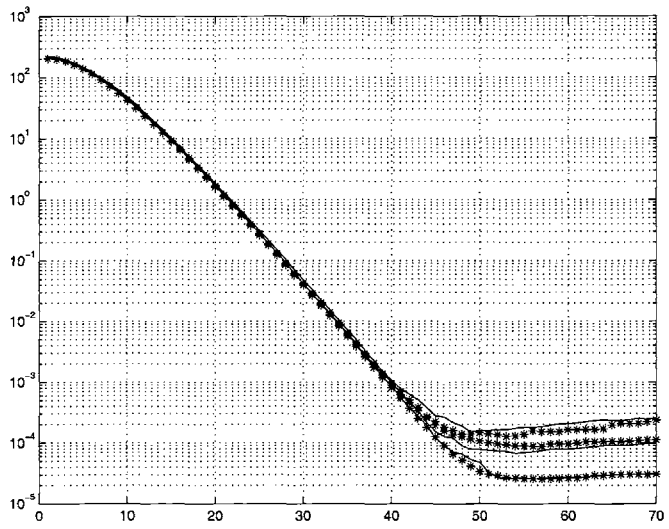

Figure 4-9: *'*' line: Impulse response error,* $||\hat{h}_1^m - h||_2^2$, *for* $\sigma_w = .02$ *and N=200,400,1000. Solid line:* $E||\hat{h}_1^m - h||_2^2$.

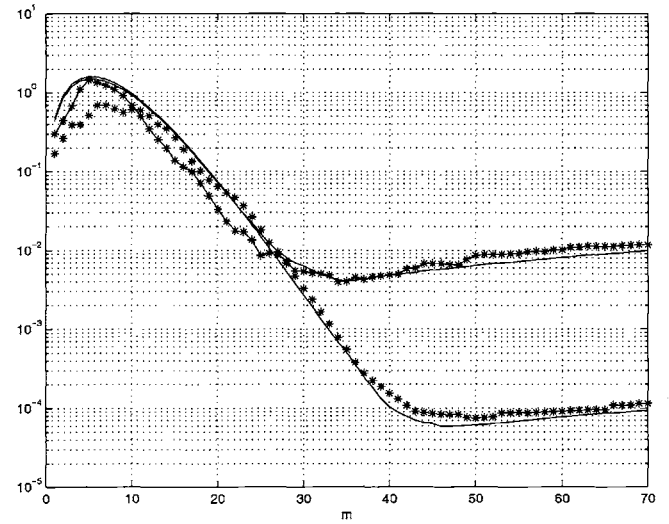

Figure 4-10: *'*' line: Subspace impulse response error,* $||\hat{h}_1^m - h_m||_2^2$, *for* $\sigma_w = .02$ *and* $\sigma_w = .2$ *N=400. Solid line:* $E||\hat{h}_1^m - h_m||_2^2$.

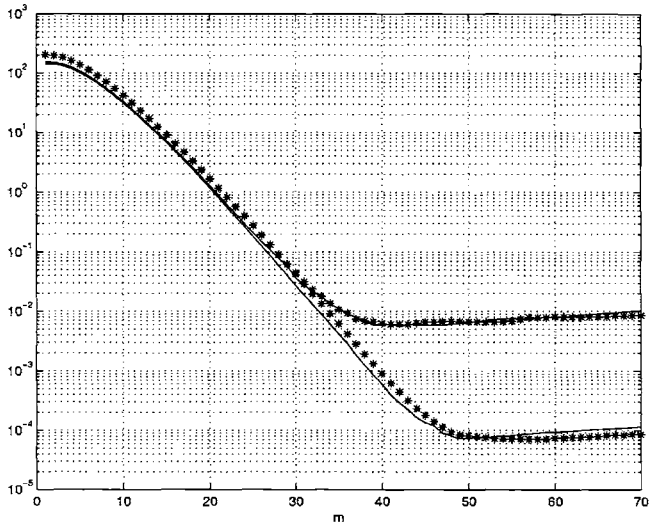

Figure 4-11: *'*' line: Impulse response error,* $||\hat{h}_1^m - h||_2^2$*, for* $\sigma_w = .02$ *and* $\sigma_w = .2$ *N=400. Solid line:* $E||\hat{h}_1^m - h||_2^2$.

Chapter 5

New Information Theoretic Approach to the Order Estimation Problem

In Chapter 3 we reviewed the well-known methods of order estimation. Here we present a new information theoretic approach to the order estimation problem. We compare the proposed method with the existing information theoretic approaches. The advantages of implementation of the new approach are illustrated in this chapter.

5.1 Minimum Description Complexity

An observed data y^N, of length N, is a sample of independent identically distributed random variable Y^N. The random variable Y^N is generated by a parametric probability distribution $p_\theta(Y^N)$ where θ is an element of a compact set Θ. Shannon shows that for any prefix code we have

$$E_\theta(L(Y^N)) - H_\theta(Y^N) \geq 0 \tag{5.1}$$

where $L(Y^N)$ is the corresponding codelength defined by the prefix code. Note that given any probability distribution $q(Y^N)$, for which $q(y^N) \neq 0$, for all elements of Y^N the code with codelength

$$L_q(Y^N) = -\log q(Y^N). \tag{5.2}$$

is prefix. Rissanen finds a nonzero lower bound for the distance $E_\theta(L(Y^N)) - H_\theta(Y^N))$ when θ is unknown: Assume that $q(Y^N)$ is defined with a parametric probability defined by a member of closed subset of Θ with dimension m, S_m. Also assume that there exists an estimator $\hat{\theta}_{S_m}(y^N)$ for θ in S_m such that the distribution of $\sqrt{N}(\hat{\theta}_{S_m}(y^N) - \theta)$ converges to a zero-mean normal distribution (Note that this assumption is equivalent to the prior assumption that θ is a member of S_m). Then Rissanen shows that for

$\theta = \theta_{S_m}$, a member of S_m,

$$E_{\theta_{S_m}}(L(Y^N)) - H_{\theta_{S_m}}(Y^N) \geq m(1-\epsilon)\log N, \tag{5.3}$$

The inequality is valid with a probability which is a function of ϵ and with a proper choice of ϵ the probability approaches one as N grows. In Theorem 3.4.2 we proved that $\log(N)$ in this inequality can be replaced with a family of functions of N.

In order estimation problem, as it is shown in figure 5-1, first $\hat{\theta}_{S_m}$, the estimate of θ in each subset S_m, is calculated. The estimate in all the information theoretic methods is obtained by choosing an element of S_m which minimizes the codelength corresponding to the observed data (ML estimate)

$$\hat{\theta}_{S_m} = \arg\min_{\theta_{S_m}} L_{\theta_{S_m}}(y^N). \tag{5.4}$$

Then the goal is to compare these estimates, of different subspaces, based on a proper criterion and choose the one which minimizes the criterion as the optimum estimate with optimum order. The main drawback of all the existing order estimation methods

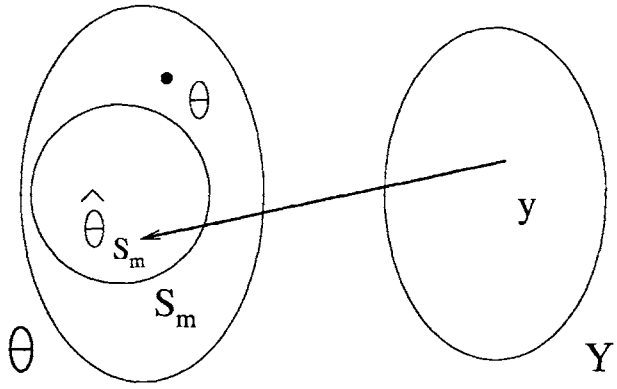

Figure 5-1: *Order estimation problem.*

is the prior assumption that θ is a member of S_m.

In the following we introduce a new method of order estimation. Define $\bar{\theta}_{S_m}$ in set S_m as

$$\bar{\theta}_{S_m} = \arg\min_{\theta_{S_m} \in S_m} E(L_{\theta_{S_m}}(Y^N)) - H_\theta(Y^N). \tag{5.5}$$

Therefore, the probability distribution generated by $\bar{\theta}_{S_m}$ minimizes the Kullback-Leibler distance of the true distribution and $p_{\theta_{S_m}}$. We suggest to use this distance as a criterion to compare the competing subspaces.

The entropy $H_\theta(Y^N)$ is a fixed number for all the subsets, therefore, for comparison of the model sets we suggest to estimate the description complexity(DC) of the random variable Y^N when $\bar{\theta}_{S_m}$ is used.

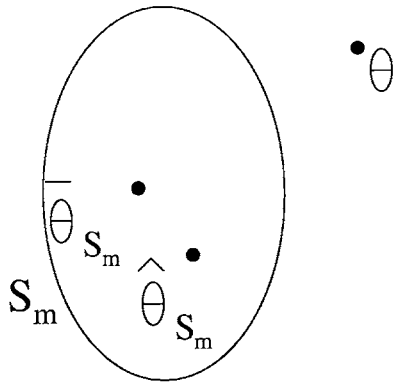

Figure 5-2: *Order estimation and the notion of information distance.*

Definition 5.1 The description complexity of Y imposed by any element of S_m is defined by

$$\mathrm{DC}(\theta, \theta_{S_m}) \triangleq \frac{1}{N} E_\theta(L_{\theta_{S_m}}(Y^N)) \tag{5.6}$$

The description complexity of the data imposed by subset S_m is then defined as

$$\begin{aligned} \mathrm{DC}(\theta, S_m) &\triangleq \min_{\theta_{S_m} \in S_m} \mathrm{DC}(\theta, \theta_{S_m}) && (5.7) \\ &= \mathrm{DC}(\theta, \bar{\theta}_{S_m}) && (5.8) \end{aligned}$$

Definition 5.2 The minimum description complexity(MDC) of random variable Y^N is attained at

$$S_m^* = \arg\min_{S_m} \mathrm{DC}(\theta, \bar{\theta}_{S_m}). \tag{5.9}$$

When the true parameter is not known and the only available data is an observed data, calculation of this criterion is complicated and might be impossible. The following theorem provides a tool to estimate $\bar{\theta}_{S_m}$ in this scenario.

Theorem 5.1.1 Asymptotically as the length of data grows,

$$\hat{\theta}_{S_m}(y^N) \to \bar{\theta}_{S_m} \tag{5.10}$$

where $\hat{\theta}_{S_m}(y^N)$ is the ML estimate of θ in subspace S_m (in 5.4).

Proof For the elements of the typical set of θ we have[1]

$$\begin{aligned} -\frac{1}{N}\log p_\theta(y^N) &= -\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(y_i) \\ &\to -E_\theta \log p_\theta(Y^N) \qquad (5.11) \\ &= \frac{1}{N}H_\theta(Y^N) \qquad (5.12) \end{aligned}$$

with probability which approaches one as N grows. Similarly for the elements of the typical set, with probability which approaches one as N grows

$$\begin{aligned} \frac{1}{N}L_{\theta_{S_m}}(y^N) &= \frac{1}{N}\sum_{i=1}^{N}\log p_{\theta_{S_m}}(y_i) \qquad (5.13) \\ &\to -E_\theta \log p_{\theta_{S_m}}(Y^N) \qquad (5.14) \\ &= \frac{1}{N}E_\theta(L_{\theta_{S_m}}(Y^N)). \qquad (5.15) \end{aligned}$$

Therefore, with probability approaching one as N grows

$$\min_{\theta_{S_m}} \frac{1}{N}L_{\theta_{S_m}}(y^N) \to \min_{\theta_{S_m}} \frac{1}{N}E(L_{\theta_{S_m}}(Y^N)) \qquad (5.16)$$

which proves the theorem. $\diamondsuit$

Therefore, the estimate of DC of Y^N is the DC of Y^N using the observed data which is defined as

$$\mathrm{DC}_{S_m}(y^N) \triangleq \mathrm{DC}(\theta, \hat{\theta}_{S_m}) \qquad (5.17)$$

Here we provide a method of calculation of this criterion for the following order estimation problem.

Consider the problem formulated in chapter 4. Finite length, N, output of an LTI system, corrupted by AWGN, is available. Subspace S_m of R^N represents one of the spaces of impulse responses of length m

$$\begin{aligned} y &= \bar{y} + w_y = h * u + w \qquad (5.18) \\ &= h_{S_m} * u + \Delta_{S_m} * u + w \\ &= A_{S_m} h_{S_m} + B_{S_m}\Delta_m + w \end{aligned}$$

[1]Typical set $A_\epsilon(\theta, N)$ is defined as

$$A_\epsilon(\theta, N) = \{y^N | 2^{-\log H_\theta(Y^N) - N\epsilon} \le p(y^N) \le 2^{-\log H_\theta(Y^N) + N\epsilon}\}.$$

The probability of this set when N is sufficiently large is

$$\Pr(A_\epsilon(\theta, N)) = 1 - \epsilon.$$

where A_{S_m} and B_{S_m} are functions of input u. In each subspace S_m, $\hat{h}_{S_m}$ is the estimate of h. Use the Shannon coding to describe a codelength for elements of Y^N using the estimate of h

$$\begin{aligned} L_{\hat{h}_{S_m}}(x) &= \log \frac{1}{f(x;\hat{h}_{S_m})} \\ &= \log \left(\sqrt{2\pi\sigma_w^2}\right)^N e^{\frac{||x-\hat{y}_{S_m}||^2}{2\sigma_w^2}}. \end{aligned} \tag{5.19}$$

The estimate of the impulse response in each subspace is the element of S_m which minimizes the codelength of the observed data y^N. Therefore the estimate of impulse response is the least-square solution

$$\hat{h}_{S_m} = \arg \min_{g_{S_m} \in S_m} ||y^N - y_{S_m}(g_{S_m})||^2 \tag{5.20}$$

where $y_{S_m}(g_{S_m}) = u * g_{S_m}$. The description complexity of random variable Y^N in (5.17), which is generated by h, using distribution $p_{\hat{h}_{S_m}}(Y^N)$ is

$$\begin{aligned} \mathrm{DC}(h, \hat{h}_{S_m}) &= \frac{1}{N} E_h(L_{\hat{h}_{S_m}}(Y^N)) \\ &= \log \sqrt{2\pi\sigma_w^2} + E\left(\frac{||Y^N - \hat{y}_{S_m}||^2}{2N\sigma_w^2}\right) \log e \end{aligned} \tag{5.21}$$

where $\hat{y}_{S_m} = u * \hat{h}_{S_m}$. This leads to the calculation and comparison of

$$\begin{aligned} \frac{1}{N} E_h(||Y^N - \hat{y}_{S_m}||^2) &= \frac{1}{N} E_h(||A_{S_m} h_{S_m} + B_{S_m}\Delta_m + w - A_{S_m}\hat{h}_{S_m}||^2) \\ &= \frac{1}{N}||A_{S_m}(h_{S_m} - \hat{h}_{S_m}) + B_{S_m}\Delta_m||^2 + \frac{\sigma_w^2}{N} \end{aligned} \tag{5.22}$$

for different subspaces. Therefore we conclude with the following theorem

Theorem 5.1.2 For the LTI model in (5.18), consider the compact set Θ which is the space of possible first N taps of the impulse response and S_m's which are the closed subspaces of order m of Θ. The MDC, defined in (5.17), is attained for

$$S_m^*(y) = \arg \min_{S_m} \frac{1}{N}||A_{S_m}(h_{S_m} - \hat{h}_{S_m}(y^N)) + B_{S_m}\Delta_m||^2 \tag{5.23}$$

where $\hat{h}_{S_m}$ is the least-square estimate of h in (5.20). Bounds on $||\frac{1}{N}A_{S_m}(h_{S_m} - \hat{h}_{S_m})||^2$ can be calculated similar to bounds on SIRE, $||h_{S_m} - \hat{h}_{S_m}||^2$ which is provided in section 4.3. Also estimation of $||B_{S_m}\Delta_m||^2$, using the output error, is provided in section 4.3.1.

Definition 5.3 If the input u of the LTI system in (5.18) is stochastic, the description complexity of Y^N imposed by elements of S_m is defined as

$$\begin{aligned} \mathrm{DC}_U(h, \hat{h}_{S_m}) &\triangleq E_U \mathrm{DC}(h, \hat{h}_{S_m}) && (5.24) \\ &= \frac{1}{N} E_U E_h(L_{\hat{h}_{S_m}}(Y^N)) && (5.25) \end{aligned}$$

For IID inputs, the properties of A_{S_m} and B_{S_m} are studied in section 4.4. In this case we have

$$\begin{aligned} \frac{1}{N} E_h(||Y^N - \hat{y}_{S_m}||^2) &= \frac{1}{N}||A_{S_m}(h_{S_m} - \hat{h}_{S_m})||^2(1 + O_1(\frac{1}{N})) + \\ &\quad \frac{1}{N}||B_{S_m}\Delta_m||^2(1 + O_2(\frac{1}{N})) + \frac{\sigma_w^2}{N} \end{aligned} \quad (5.26)$$

and the expected value with respect to the input, ignoring $O(\cdot)$ terms, is

$$\begin{aligned} \frac{1}{N} E_U E_h(||Y^N - \hat{y}_{S_m}||^2) &= ||h_{S_m} - \hat{h}_{S_m}||^2 + \frac{1}{N} E_U||B_{S_m}\Delta_m||^2 + \frac{\sigma_w^2}{N} \\ &= ||h - \hat{h}_{S_m}||^2(1 + O(\frac{1}{N})) + \frac{\sigma_w^2}{N} \end{aligned} \quad (5.27)$$

Which is calculated in chapter 4.

Theorem 5.1.3 For the LTI system in (5.18) with IID input of zero mean and unit variance the MDC, defined in (5.17), is attained for

$$\begin{aligned} S_m^*(y) &= \arg\min_{S_m} \frac{1}{N} E_U E_h(L_{\hat{\theta}_{S_m}}(Y)) && (5.28) \\ &= \arg\min_{S_m} ||h - \hat{h}_{S_m}||^2(1 + O(\frac{1}{N})) + \frac{\sigma_w^2}{N}. && (5.29) \end{aligned}$$

Therefore as we show in chapter 4 the rate of convergence of the distance error in (5.17) is not just a function of m and N but also a function of unmodeled dynamics which can be validated by the use of output error.

The proposed method in this section calculates the description complexity of Y^N for the family of Gaussian distributions and when the estimate of impulse response is provided by the least square method. However, the method can be generalized for calculation of the description complexity of Y^N even if the additive noise w is not Gaussian.

5.1.1 New Minimum Description Length

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

Occam's razor is a logical principle attributed to the medieval philosopher William of Occam. Applying the principal to the statement above, the main message is

"The simplest explanation is the best"

The principle states that one should not make more assumptions than the minimum needed. A computer scientific approach to this principle is manifested in Kolmogorov complexity. Let y be a finite binary string and let $\mathcal{U}$ be a universal computer. Let $l(y)$ denote the length of the string y. Let $\mathcal{U}(pg)$ denote the output of the computer U when presented with program pg. Then the Kolmogorov complexity $K_{\mathcal{U}}(y)$ of a string y with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(y) = \min_{pg\,:\,\mathcal{U}(pg)=y} l(pg) \tag{5.30}$$

The complexity of string y is called the minimum description length of y. For any other computer $\mathcal{A}$ we have

$$K_{\mathcal{U}}(y) \leq K_{\mathcal{A}}(y) + c_{\mathcal{A}} \tag{5.31}$$

where $c_{\mathcal{A}}$ does not depend on y. This inequality is known as universality of Kolmogorov complexity. The minimum description complexity method we described previously deals with averaging the description length of a set Y. Kolmogorov complexity is a modern notion of randomness dealing with the quantity of information in individual objects; that is "pointwise" randomness rather than average randomness produced by a random source.

Consider the order estimation problem for the LTI system which we discussed before. The output of the system is given in (5.18). Inspired by the Kolmogorov complexity and notion of minimal description length of string y, we want to search for the subspace which provides the minimum description length of the "data". In each subspace S_m the description length of y is described as the minimum codelength which can describe y by an element of S_m. For the codelength in this probabilistic setting the Shannon coding method is used, therefore

$$\begin{aligned} \mathrm{DL}_{S_m}(y) &= \min_{g \in S_m} -\log f(y;g) && (5.32)\\ &= -\log f(y;\hat{h}_{S_m}) && (5.33) \end{aligned}$$

Note that in this scenario the probability distribution defined by each θ_{S_m} is a Gaussian distribution with output of form

$$y_{S_m} = g * u + w. \tag{5.34}$$

where the mean of the random variable is $g * u$, $g \in S_m$ and variance of the additive noise is σ_w^2. Therefore the least square estimate of h in each subspace provides the DL of the output in that subspace

$$\mathrm{DL}_{S_m}(y) = \log\left(\sqrt{2\pi\sigma_w^2}\right)^N + \frac{||y - \hat{y}_{S_m}||^2}{2\sigma_w^2}\log e. \tag{5.35}$$

But comparison of this description length for different subspaces always leads to the choice of S_m with largest possible order, S_N for which the output error is zero!.

To avoid this problem, in two-stage MDL, Rissanen introduces the codelength which describes elements of S_m as well. Here the assumption is that the length of the code describing any element of subspace S_m is the same and is of order

$$\frac{m}{2}\log(N) \tag{5.36}$$

Therefore the total codelength describing y in subspace S_m is the codelength describing $\hat{h}_{S_m}$ in (5.36) plus the description length of the output given this estimate from (5.35)

$$\mathrm{DL}_{\hat{h}_{S_m}}(y) = \frac{m}{2}\log(N) + \log\frac{1}{f_{S_m}(y;\hat{h}_{S_m})}. \tag{5.37}$$

Choosing the description length of elements of S_m, θ_{S_m}, by codes of length $\frac{m}{2}\log(N)$ seems to be an ad-hoc method. Partitioning the subspace S_m can be done with any other descritization per dimension factor other than $\log(N)$. For this reason we believe that the codelength for all elements of S_N and its subspaces is the same.

Another method of achieving this description length is given in [38]. It is argued that the codelength in (5.37) is optimum since it can achieve the lower bound in the inequality given in (5.3). However, as we discussed before, $\log(N)$ in this inequality can be replaced by a family of functions of N. This implies that $\log(N)$ in describing the description length can be replaced by that family of functions of N.

The comparison of the codelength in (5.35) fails because of the following argument: Minimizing the description length in (5.32) is the same as

$$\arg\min_{g\in S_m} -\log f(y;g) = \arg\max_{g\in S_m} f(y;g) \tag{5.38}$$

which provides the ML estimate of h in each subspace. As we discussed in previous sections the ML estimation always points to a member of S_N, which has the highest possible, as a perfect candidate. For example in the Gaussian case minimizing the codelengths is equivalent to minimizing the output error $||y-\hat{y}_{S_m}||$ which happens for a member of S_N for when the error is zero. Therefore comparison of the codelength describing y itself in each subspace is not a proper tool for comparison of the estimates. Here y is not the string of "data", y is the data which is corrupted by an additive noise. Therefore we believe that the codelength of the "noiseless output" in each of these subspaces is the proper criterion.

To follow the Kolmogorov complexity is to compare the codelength which describes the noiseless output $\bar{y}$ in each subspace. Therefore the new description length is

Definition 5.4 The description length of "data" in subspace S_m is defined as

$$\mathrm{DL}_{\hat{h}_{S_m}}(y) \;=\; -\log f(\bar{y};\hat{h}_{S_m}) \tag{5.39}$$

$$= -\log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}_{S_m}||^2}{2\sigma_w^2}}. \quad (5.40)$$

Comparison of such description length for different subspaces leads to comparison of the reconstruction error

$$\frac{1}{N}||\hat{y}_{S_m} - \bar{y}||^2 = ||A_{S_m}\hat{h}_{S_m} - (A_{S_m}h_{S_m} + B_{S_m}\Delta_{S_m})||^2 \quad (5.41)$$

In chapter 4 the goal is to find the probabilistic bounds on the impulse response error $||\hat{h}_{S_m} - h^N||^2$, by using one sample of the output error $\frac{1}{N}||\hat{y}_{S_m} - y||^2$. We suggest to use the same approach for estimation of the new description length.

For IID inputs, the properties of A_{S_m} and B_{S_m} are studied in section 4.4. For such inputs

$$\frac{1}{N}||\hat{y}_{S_m} - \bar{y}||^2 = (1 + O(1))||\hat{h}_{S_m} - h^N||^2 \quad (5.42)$$

Therefore the impulse response error can be used in describing the code length of $\bar{y}$.

Note that comparison of the DC in (5.21) and the new DL in (5.40) in this scenario are the same.

5.2 Comparison of the Order Estimation Methods

In this section we compare the order estimation methods for when the input is IID. AIC, BIC and two-stage MDL for each model set S_m are given by

$$\mathrm{AIC}_{S_m}(y) = -\frac{1}{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{||\hat{y}_{S_m}-y||^2}{2\sigma_w^2}}\right) + \frac{m}{N}. \quad (5.43)$$

$$\frac{1}{N}\mathrm{BIC}_{S_m}(y) = \frac{1}{N}\mathrm{DL}_{S_m}(y) = -\frac{1}{N}\log\left(\frac{1}{(\sqrt{2\pi}\sigma_w)^N} e^{-\frac{||\hat{y}_{S_m}-y||^2}{2\sigma_w^2}}\right) + m\frac{\log N}{2N} \quad (5.44)$$

Using the new information theoretic approach, section 5.1, the description complexity of Y is

$$\mathrm{DC}_{S_m}(y) = -\frac{1}{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{||\hat{y}_{S_m}-\bar{y}||^2}{2\sigma_w^2}}\right). \quad (5.45)$$

We also showed that comparison of this description complexity for IID inputs is similar to comparison of

$$||h - \hat{h}_m^N||_2^2 = ||h_m^N - \hat{h}_m^N||_2^2 + ||\Delta_m^N||^2, \quad (5.46)$$

for different subspaces. In chapter 4 we provided bounds on this error using the output error. We showed that with validation probabilities $Q(\alpha)$, and $Q(\gamma/k_m)$[2] and with probability $Q(\beta)$ [3]

$$||\hat{h}_m^N - h^N||_2^2 \leq \frac{m}{N}\sigma_w^2 + (1+\frac{m}{N})\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}} + \frac{\beta\sqrt{m}}{N}\sqrt{2\sigma_w^4 + m\left(\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}\right)^2}, \tag{5.47}$$

$$||\hat{h}_m^N - h^N||_2^2 \geq \max\left\{0, \frac{m}{N}\sigma_w^2 + (1+\frac{m}{N})\frac{Lg_m}{1+\frac{\gamma}{\sqrt{N}}} - \frac{\beta\sqrt{m}}{N}\sqrt{2\sigma_w^4 + m\left(\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}})\right)^2}\right\}$$

where

$$Ug_m = x_m + (\frac{2\alpha^2}{N} - 1)m_w + \frac{2\alpha\sqrt{m_w}}{\sqrt{N}}\sqrt{\frac{\alpha^2 m_w}{N} + x_m - \frac{3}{2}m_w}. \tag{5.48}$$

and $x_m = \frac{1}{N}||y - \hat{y}_{S_m}||_2^2$ and $m_w = (1-\frac{m}{N})\sigma_w^2$. The conditions on α and β and γ are given as follows

$$\alpha_N \geq \sqrt{\frac{N}{2}}(1 - \frac{x_m}{(1-\frac{m}{N})\sigma_w^2}), \tag{5.49}$$

$$\lim_{N\to\infty}\alpha_N = \infty \quad , \quad \lim_{N\to\infty}\beta_N = \infty, \quad \lim_{N\to\infty}\gamma_N = \infty, \tag{5.50}$$

$$\lim\frac{\alpha_N}{\sqrt{N}} = 0 \quad , \quad \lim_{N\to\infty}\frac{\beta_N}{N} = 0, \quad \lim_{N\to\infty}\frac{\gamma_N}{\sqrt{N}} = 0. \tag{5.51}$$

The expected value of impulse response error(IRE), with validation probabilities $Q(\alpha)$ and $Q(\gamma/k_m)$, is

$$E||\hat{h}_m^N - h^N||_2^2 = \frac{m}{N}\sigma_w^2 + (1+\frac{m}{N})\frac{Ug_m}{1-\frac{\gamma}{\sqrt{N}}}, \tag{5.52}$$

In the following sections we elaborate the advantages of the new information theoretic approach over the existing ones regarding issues such as unknown additive noise, consistency and sensitivity to the signal to noise ratio.

5.2.1 Additive Noise Variance

In practical problems the variance of the additive noise is usually unknown. The existing order estimation methods MDL and AIC suggest to find the variance of noise for each subspace separately. Since the unmodeled dynamics effects are considered as

[2] k_m is defined in (4.61)

[3] Here we replaced $1-\frac{1}{\beta^2}$ with $Q(\beta)$. This is valid for when the Chi-squared distribution of the impulse response error can be approximated with a Gaussian distribution. The approximation is possible for large enough m.

a part of the additive noise effects, for each subspace the estimate of variance is

$$\hat{\sigma}_w^2(S_m) = \frac{1}{N}||\hat{y}_{S_m} - y||^2. \tag{5.53}$$

In this case the AIC and MDL are calculated in [32]

$$\mathrm{AIC_{S_m}(y)} \approx (1 + \frac{2m}{N})\frac{||\hat{y}_{S_m} - y||^2}{N}, \tag{5.54}$$

$$\mathrm{MDL_{S_m}(y)} \approx (1 + m\frac{\log N}{N})\frac{||\hat{y}_{S_m} - y||^2}{N}. \tag{5.55}$$

In [59] calculation of AIC and MDL with unknown variance is expanded for the multi-output problems. Extension of this criteria for when the additive noise is not white shows some drawbacks in [64] in practice.

In calculation of the new information theoretic methods, MDC or new MDL, since the effects of unmodeled dynamics is separated from the effects of the additive noise, one estimate of variance has to be used for all the subspaces. We suggest to use the variance estimate obtained for a model set with high enough order M, $M << N$. Therefore, the estimated variance for all model sets is

$$\hat{\sigma}_w^2 = \frac{1}{N}||\hat{y}_{S_M} - y||^2. \tag{5.56}$$

The choice of M is experimental and such that the minimum description complexity is minimized.

5.2.2 Consistency Issues

The order estimation method is consistent if as N grows the method chooses the true model. Assume that the impulse response has finite length M and the order estimation method picks $S_{m^*(N)}$. the method is consistent if

$$\lim_{N\to N} m^*(N) = M, \quad \lim_{N\to N} h_{S_{m^*(N)}} = h, \tag{5.57}$$

Given the structure of AIC, MDL and BIC in (5.43),(5.44) we can propose the following question. Consider the family of criteria in form of

$$\frac{||\hat{y}_{S_m} - y||^2}{N} + m\sigma_w^2 f(N). \tag{5.58}$$

For what family of $f(N)$s is this criterion consistent? We know that AIC is obtained with $f(N) = \frac{2}{N}$ which is not a consistent method. Also MDL and BIC are obtained with $f(N) = \frac{\log(N)}{N}$ which makes the criterion consistent. In [23] Hannan suggests to use $f(N) = \frac{\log\log N}{N}$, which decreases faster than $f(N)$ in MDL and still provides a consistent method.

The consistency of AIC and MDL has been investigated in a wide range of liter-

ature. The problem of overmodeling of these methods in some practical applications are shown in examples in [31]. Also the consistency issues for the multi-output systems (sensor array) in practical problems are investigated in [13].

Here we check the consistency from a new perspective. For large enough m if the unmodeled effects is almost zero, AIC provides an estimate of $E(||\hat{y}_{S_m} - y||^2)$. In this case we can show that MDL is estimating an upper bound for this error which contains the error variance as N grows. Based on our method of estimation of IRE, as N grows, this upper bound is valid with probability one.

In the new approach, we showed that the choice of α, β, γ plays a major role in defining an event. We are comparing not just the estimate of the expected values but comparing events in different subspaces. Therefore the consistency of the new method is guaranteed as long as the validation and confidence region probabilities go to one as N grows. The sufficient criteria is satisfied by choosing α, β and γ based on conditions given in (5.50) and (5.51). For example, if $\alpha = \beta = \gamma = 0$ the estimate of error (5.52) is ,

$$E(||\hat{h}_{S_m} - h^N||^2) \approx \frac{m}{N}\sigma_w^2 + (1 + \frac{m}{N})\left(\frac{1}{N}||\hat{y}_{S_m} - y||^2 - (1 - \frac{m}{N})\sigma_w^2\right) \quad (5.59)$$

$$\approx (1 + \frac{m}{N})\frac{1}{N}||\hat{y}_{S_m} - y||^2 + \frac{m}{N}\sigma_w^2 + \frac{m^2}{N^2}\sigma_w^2 + \sigma_w^2 \quad (5.60)$$

We know that this choice of the parameters result an inconsistent method since the variances of the random variables are completely ignored. When $m \geq M$, then $\frac{m}{N}(\frac{1}{N}||Y - \hat{Y}_m^N||_2^2 \approx \frac{m}{N}\sigma_w^2$ and

$$E(||\hat{h}_{S_m} - h^N||^2) \approx \frac{2m}{N}\sigma_w^2 + \frac{1}{N}||\hat{y}_{S_m} - y||^2 + \sigma_w^2. \quad (5.61)$$

Interestingly with this choice of α, β and γ this estimate of IRE is the same as AIC. On the other hand the choice of $\alpha^2 = \beta = \gamma^2 = \log\log N$ satisfies the sufficient conditions and provides a criterion which leads to a consistent method.

5.2.3 Thresholding

Information theoretic methods attempt to "determine" the length of the impulse response. In most practical problems, the impulse response does not have a finite length and we require to detect the minimum number of taps of the impulse response which represents the "significant part" of the impulse response. Implementing the MDL method in this situation, provides an estimate for the length of the impulse response which is very sensitive to the variation in signal to noise ratio(SNR) [4] and to the length of the output [30]. When the length of the true impulse response is infinite, the consistent methods, such as MDL and BIC, point to a higher and higher length for the impulse response as N and/or SNR grows. Some related practical

[4] SNR= $10\log_{10}\frac{||Y^N||_2^2}{N\sigma_w^2}$

problems of the information theoretic methods are addressed in [60] and [7].

To overcome the consistency problem for when impulse response length is infinite we propose implementing the new information theoretic method of order estimation. With this method we can avoid this consistency problem by using a threshold for the impulse response error. If a threshold ϵ is used for the minimum acceptable IRE, then we choose the smallest m for which $U_{S_m} \leq \epsilon$. An example of this approach is given in the simulation section.

MDL Thresholding

Can thresholding be used for two-stage MDL? In order to make the description length in (5.37) a valid codelength corresponding to a prefix code, which satisfies the Kraft's inequality, in [38] it is suggested to add a normalizing constant $C(N)$ to the description length

$$\mathrm{DL}_{S_m}(y) = m\log(N) + \log\frac{1}{f_{S_m}(y;\hat{h}_{S_m})} + C(N). \tag{5.62}$$

It is argued that as N grows $C(N)/N \to 0$.

However, note that as N grows the factor $m\frac{\log(N)}{N}$ also goes to zero. For any fixed N, $C(N)$ might be comparable with $m\frac{\log(N)}{N}$. Calculation of this normalizing factor is not trivial and it is not provided. Because of the structure of the DL we believe $C(N)$ is a function of the noiseless output $\bar{y}$. Since $C(N)$ is a fixed number in the comparison of different subspaces this term is ignored in the MDL method. However, since $C(N)$ might change for different order estimation settings, for example with the change of $\bar{y}$, implementation of threshold is meaningless for this criterion. Note that the problem of calculation of $C(N)$ is consistent in definition of other existing MDL methods.

In previous section we suggested to use thresholding for the comparison of the description complexity. Here we prove that the use of threshold also is meaningful for the new proposed MDL method. Assume that for any problem setting the descritization in output space Y is the same. We prove that the new description length itself is a codelength which is uniquely decodable. The new DL satisfies the Kraft's inequality by adding a normalized factor. The normalizing factor is not a function of y or $\bar{y}$ but a function of the order of subspace and descritization of Y.

Theorem 5.2.3 The new description length, defined in (5.40), satisfies the Kraft's inequality when a normalized factor $C(N)$ is considered

$$\mathrm{DL}_{S_m}(y) = \log\frac{1}{f_{S_m}(\bar{y};\hat{h}_{S_m})} + C(N). \tag{5.63}$$

where

$$C(N) = -\frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m\sqrt{2\pi\sigma_w^2}^{(N-m)}). \tag{5.64}$$

Note that although $C(N)$ is a function of m, with a proper choice of α, β and γ, $C(N)/N$ goes to zero much faster that the terms in the estimate of $\log \frac{1}{f_{S_m}(\bar{y};\hat{h}_{S_m})}$ and it can be ignored for large enough N.

Proof The codelength defined for the descritized version of y, $y^d(i)$ using $\hat{\theta}_{S_m}$ is defined as

$$\begin{aligned} \mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y}) &= -\log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}^d_{S_m}(i)||^2}{2\sigma_w^2}} \qquad (5.65) \\ &= -\frac{1}{\mathrm{Ln}(2)} \mathrm{Ln} \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}^d_{S_m}(i)||^2}{2\sigma_w^2}}. \qquad (5.66) \end{aligned}$$

To check the Kraft's inequality for each code word of length $\mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y})$ we have to show that

$$\sum_i D^{-\mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y})} \leq 1 \qquad (5.67)$$

where D is the size of alphabet resulted from descritizing the output space Y. Equivalently we can check the following inequality

$$\sum_i \left(e^{-\mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y})}\right)^{\mathrm{LnD}} \leq 1 \qquad (5.68)$$

we know that

$$\begin{aligned} \sum_i \left(e^{-\mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y})}\right)^{\mathrm{LnD}} &\leq \left(\sum_i e^{-\mathrm{DL}_{\hat{h}_{S_m}(y^d(i))}(\bar{y})}\right)^{\mathrm{LnD}} \\ &\leq \left(\sum_i \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}^d_{S_m}(i)||^2}{2\sigma_w^2}}\right)^{\frac{\mathrm{LnD}}{\mathrm{Ln2}}} \end{aligned}$$

Note that

$$\delta^m \sum_i \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m} e^{-\frac{||\bar{y}-\hat{y}^d_{S_m}(i)||^2}{2\sigma_w^2}} \approx \int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m} e^{-\frac{||\bar{y}-\hat{y}_{S_m}||^2}{2\sigma_w^2}} dy \qquad (5.69)$$

where δ is the precision per dimension in space Y, or equivalently in space of additive noise W. From chapter 4, the error $\bar{y} - \hat{y}_{S_m}$ is such that

$$\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^m} e^{-\frac{||\bar{y}-\hat{y}_{S_m}||^2}{2\sigma_w^2}} dy = 1. \qquad (5.70)$$

Therefore

$$\int \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\bar{y}-\hat{y}_{S_m}||^2}{2\sigma_w^2}} dy = \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^{N-m}} \tag{5.71}$$

Therefore

$$\sum_i \left(e^{-\left(\mathrm{DL}_{\hat{h}_{S_m}(y^{d(i)})}(\bar{y}) - \frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m \sqrt{2\pi\sigma_w^2}^{(N-m)})\right)} \right)^{\mathrm{LnD}} \leq 1 \tag{5.72}$$

Hence the normalizing factor is $\frac{1}{\mathrm{Ln}2}\mathrm{Ln}(\delta^m \sqrt{2\pi\sigma_w^2}^{(N-m)})$. $\diamondsuit$

5.3 Simulation Results

We use the microwave radio channel, *chan*10.*mat*, which is available at

http://spib.rice.edu/spib/microwave.html.

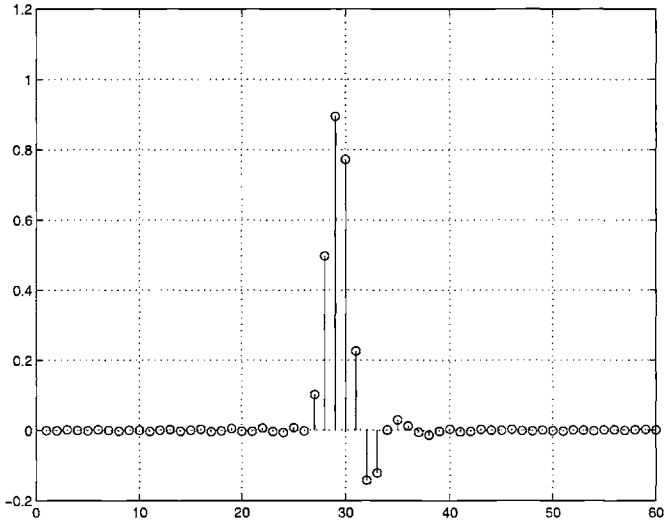

Figure 5-3: *Real part of the first 60 taps of a microwave radio channel impulse response.*

Figure (5-3) shows the real part of the first 60 taps of the impulse response. The simulation result for $N = 300$ and SNR=10db is as follows: $\hat{m}$(AIC)=34, $\hat{m}$(MDL)=32. The new proposed criterion selects $m^* = 33$. Figure(5-5) shows the upper and lower bound on IRE for N=800, SNR=90db. The bounds on the error are from (4.70), (4.72) with $\alpha = \sqrt{\beta} = \gamma = \log\log N$. The solid line is the estimate of expected value of I.R.E, $E(||\hat{h}_m^N - h^N)||^2)$, when $\alpha = \beta = \gamma = 0$.

In this case all the methods select an impulse response length which is larger than 130. With higher SNR and/or longer data sample, all the methods choose a larger and

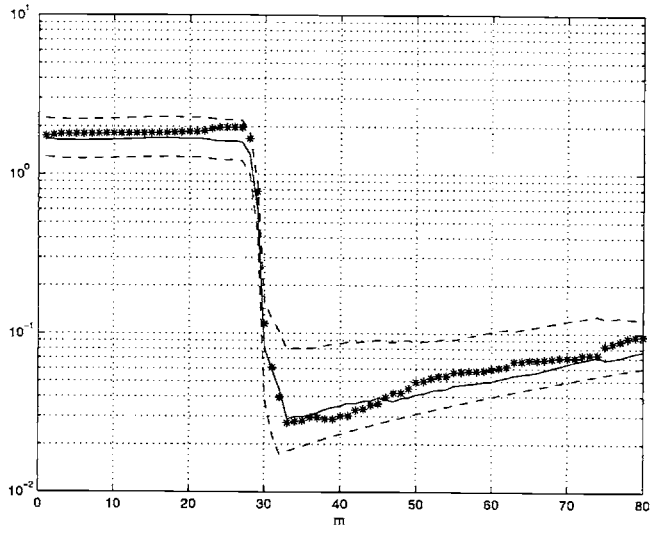

Figure 5-4: *'*' line: Impulse response error, $||\hat{h}_m^N - h||_2^2$, for SNR=10db, N=300. '-.': Upperbound and lowerbound of IRE. Solid line: Estimate of $E||\hat{h}_m^N - h||_2^2$.*

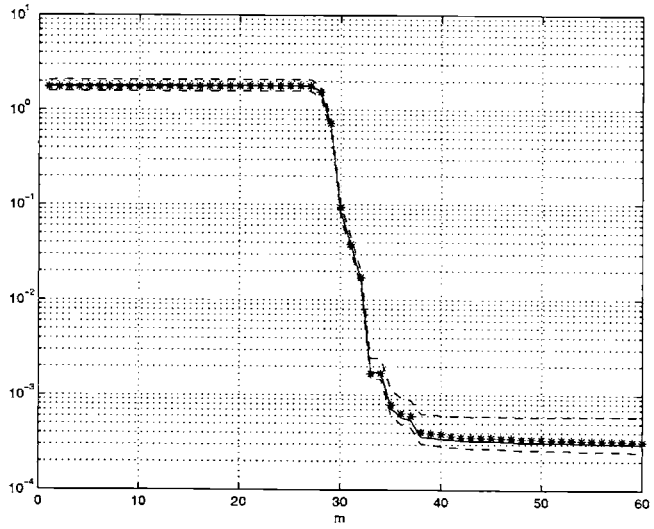

Figure 5-5: *'*' line: Impulse response error, $||\hat{h}_m^N - h||_2^2$, for SNR=90db, N=800. '-.': Upperbound and lowerbound of IRE. Solid line: Estimate of $E||\hat{h}_m^N - h||_2^2$.*

larger length for the impulse response estimate. However, if we choose a threshold for the IRE to be 10^{-3}, the new criterion selects $m^* = 35$. With this threshold $m^* \leq 35$ when SNR grows and/or the length of data gets larger. Counting for the delay of the system, with the same threshold, the proposed method chooses the 10 taps of the impulse response estimate from 27 to 36 for modeling the system.

Chapter 6

Signal Denoising

The problem of estimating an unknown signal embedded in Gaussian noise has received a great deal of attention in numerous studies. The denoising process is to separate an observed data sequence into a "meaningful" signal and a remaining noise. The choice of the denoising criterion depends on the properties of the additive noise, smoothness of the class of the underlying signal and the selected signal estimator.

In this chapter we review some of the important existing denoising methods: The pioneer thresholding method of wavelet denoising which was first formalized by Donoho and Johnstone [11], the thresholding method proposed in [29] and the normalized minimum description length approach [43]. We propose a new denoising method based on MDC. Calculation of MDC in this case is very similar to the calculation of MDC for system identification which was proposed in chapter 4. We illustrate the benefits of implementation of the new method over the existing methods.

6.1 Problem Formulation

Consider noisy data y of length N,

$$y(n) = \bar{y}(n) + w(n), \tag{6.1}$$

where $\bar{y}$ is the noiseless data and w is the additive white Gaussian noise with zero mean and variance σ_w^2. Data denoising is achieved by choosing an orthogonal basis which approximates the data with fewer nonzero coefficients than the length of the data. Consider the orthogonal basis of order N, S_N. The basis vectors $s_1, s_2, \cdots, s_N$ are such that $||s_i||_2^2 = N$. Any vector of length N can be represented with such basis. Therefore, there exists $h = [h_1, h_2, \cdots, h_N]^T$ such that $\bar{y}(n) = \sum_{i=1}^{N} s_i(n)h_i$. As a result the noisy data is

$$y(n) = \sum_{i=1}^{N} s_i(n)h_i + w(n). \tag{6.2}$$

The least square estimate of each basis coefficient is

$$\hat{h}_i = \frac{1}{N} s_i^T y^N = h_i + \frac{1}{N} s_i^T w \tag{6.3}$$

where $y^N = [y(1), y(2), \cdots, y(N)]$, the observed noisy data, is a sample of random variable Y^N. The benefit of using a proper basis is that $\frac{1}{N} s_i^T w$ is almost zero as N is assumed to be large enough and we hope that there exist a large number of basis vectors for which $h_i = 0$. Therefore, the estimation of the noisy signal on this basis has the advantage of noise elimination. For such reason conventional basis denoising methods suggest choosing a threshold, τ, for the coefficient estimates $\hat{h}_i$'s. The denoising process is to ignore the coefficient estimates smaller than the threshold

$$\begin{aligned} \hat{h}_i &= \frac{1}{N} s_i^T y^N, \text{ if } |\frac{1}{N} s_i^T y^N| \geq \tau \\ \hat{h}_i &= 0, \quad \text{if } |\frac{1}{N} s_i^T y^N| < \tau \end{aligned} \tag{6.4}$$

and the estimate of the noiseless signal is

$$\hat{y}^N(n) = \sum_{i=1}^{N} s_i(n) \hat{h}_i. \tag{6.5}$$

A very important factor in finding the optimum threshold is the behavior of the mean square reconstruction error

$$\frac{1}{N} E(||\bar{y}^N - \hat{Y}^N||_2^2). \tag{6.6}$$

6.1.1 Wavelet Thresholding

The classic paper [11] provides an upperbound for the mean square reconstruction error in wavelet denoising by solving a min-max problem. Assume that the wavelet coefficients of signal y are $\hat{h}_{i,k}$s,

$$\hat{h} = \mathcal{S} y \tag{6.7}$$

where wavelets s_{ik} denotes the (i,k)th row of $\mathcal{S}$. The vector $\hat{h}$ has $N = 2^{I+1}$ elements. The wavelet basis is s_{ik} which is generated by a mother wavelet ϕ. for $i_0 \leq i \leq I - i_1$

$$N^{\frac{1}{2}} s_{ik}(j) \approx 2^{i/2} \phi(2^i t - k) \quad t = j/N \tag{6.8}$$

This approximation improves with increasing N and increasing i_1. The orthogonal properties of the wavelet is such that $y = \mathcal{S}^T \hat{h}$. If $\bar{h}$ is the coefficients corresponding to the noiseless data $\bar{y}$, we have

$$\hat{h}_{i,k} = \bar{h}_{i,k} + z_{i,k} \tag{6.9}$$

where $z = \mathcal{S}w$ is additive white Gaussian noise with the same variance of w, σ_w^2. The main assumption is that very few wavelet coefficients contribute to the signal, i.e., the signal is piecewise smooth.

To find the optimum threshold the following procedure is suggested. Without loss of generality and to simplify the calculations, the wavelet notation is changed to the vector representation introduced in the problem formulation of the last section. Therefore, the estimate of a wavelet coefficient is

$$\hat{h}_i = \bar{h}_i + z_i \tag{6.10}$$

where z_i is additive white Gaussian noise with variance σ_w^2. The following soft thresholding method with threshold level τ is used

$$\hat{h}_i(\tau) = \eta(\hat{h}, \tau) = \text{sign}(\hat{h})(|\hat{h}_i| - \tau)_+ \tag{6.11}$$

The goal is to find an estimate for

$$E(||\hat{H}(\tau) - \bar{h}||_2^2 \tag{6.12}$$

A min-max solution to the problem is proposed as follows. In [11] it is shown that as $N \to \infty$

$$\inf_{\hat{h}(\tau)} \sup_{\bar{h}} \frac{E(||\hat{H}(\tau) - \bar{h}||_2^2)}{\sigma_w^2 + \sum_{i=1}^{N} \min(\bar{h}_i, \sigma_w^2)} \approx 2 \log N \tag{6.13}$$

where the infimum is obtained for the optimum threshold

$$\tau = \sigma_w \sqrt{2 \log N}. \tag{6.14}$$

Note that the method eventually is not able to provide an estimate for the reconstruction error. But it introduces a function of τ, $f(\tau) = \sigma_w^2 + \sum_{i=1}^{N} \min(\bar{h}_i, \sigma_w^2)$ and provides the τ for which the ratio of "reconstruction error/$f(\tau)$" is optimized in a min-max setting.

6.1.2 Estimation of Mean-square Reconstruction Error

In [29] an estimate of the mean square error as a function of a given threshold is provided heuristically. Consider an orthonormal basis with basis functions s_is. The estimate of coefficients are provided with the least-square algorithm in (6.3) and the thresholding method with level τ in (6.4) is used. An estimate of error is provided based on a heuristic method. It demonstrates that for a class of signals the proposed threshold by Donoho and Johnstone, $\sigma_w \sqrt{2 \log N}$, does not provide the optimal threshold.

6.1.3 MDL Denoising

Instead of focusing on finding a threshold one can compare the signal estimate in different subspaces of the basis. Choosing a subspace to estimate the data is equivalent to setting the coefficients of the basis vectors out of that subspace to zero without thresholding. MDL denoising is the first method which approaches the denoising problem with this idea. In each subspace it calculates the defined description length of the data and suggests to pick the subspace which minimizes this criterion.

Rissanen suggests using the normalized MDL for the denoising problem. We briefly reviewed the normalized MDL in section 3.3.4. In each subspace S_m, $\hat{h}_{S_m}$, the least-square estimate of h, is obtained. Also an estimate of the variance is calculated as follows

$$\hat{\sigma}_w^2 = \frac{1}{N}||y - \hat{y}_{S_m}||^2. \tag{6.15}$$

In each subspace of the basis, S_m, the criterion is defined as

$$\hat{f}(y^N; S_m) = \frac{f_{S_m}(y^N; \hat{h}_{S_m}(y^N), \hat{\sigma}_w^2(y^N))}{\int_{Z(R,\sigma_0^2)} f_{S_m}(z; \hat{h}_{S_m}(z), \hat{\sigma}_w^2(z))dz} \tag{6.16}$$

where

$$Z(r, \sigma_0^2) = \{z | \hat{h}'_{S_m}(z)\Sigma_{S_m}\hat{h}_{S_m}(z)/N \leq R, \hat{\sigma}_w^2 \geq \sigma_0^2\}, \tag{6.17}$$

and $\Sigma_{S_m} = A'_{S_m}A_{S_m}$ where A_{S_m} is the matrix whose columns are the bases of S_m. It is said that r and σ_0^2 are chosen such that the ML estimates fall within $S(r, \sigma_0^2)$.

The normalized description length in (6.16) is claimed to be the solution to the min-max problem

$$\min_q \max_x \mathrm{Ln} \frac{f_{S_m}(x; \hat{h}_{S_m}(x), \hat{\sigma}^2(x))}{q(x)} \tag{6.18}$$

The approach is considered as a result of the universal coding of a single message proposed in [47].

6.1.4 The New Approach

Similar to MDL denoising approach, we investigate on estimation of a criterion which is defined for the subspaces of the basis. For each subspace S_m, $\hat{h}_{S_m}$ denotes the estimate of the coefficients in that subspace. Our goal is to find an estimate of the coefficient estimation error in each subspace, $||h - \hat{h}_{S_m}||_2^2$. Note that, as a result of the Parseval's Theorem, this error is the same as the *reconstruction error* for each subspace

$$||h - \hat{h}_{S_m}||_2^2 = \frac{1}{N}||\bar{y}^N - \hat{y}_{S_m}^N||_2^2. \tag{6.19}$$

Because of the additive noise the coefficients error is also a random variable. The objection is to compare the worst case behavior of this error in different subspaces probabilistically. The best representative of the signal is then the signal estimate of the subspace which minimizes such criterion. In the following section we describe the method in detail. The first step is to probabilistically validate the error caused by the elimination of the basis vectors out of the subspace. Next we estimate both the mean-square and the variance of the coefficients error. The approach is similar to the quality evaluation method for impulse response estimate of the LTI system which is proposed in chapter 3.

6.2 New Denoising Method

Consider a subspace of order m of the orthogonal basis, S_m. We want to estimate the error of coefficient estimation in this subspace, $||h - \hat{h}_{S_m}||_2^2$. Given the noisy data in (6.1), we suggest the following procedure to estimate the error: For the subspace S_m, matrix A_{S_m} separates the basis vectors as follows

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} = \begin{bmatrix} A_{S_m} & B_{S_m} \end{bmatrix} \begin{bmatrix} h_{S_m} \\ \\ \Delta_{S_m} \end{bmatrix} + w \tag{6.20}$$

where columns of A_{S_m} are $s_i \in S_m$, columns of B_{S_m} are basis vectors which are not in S_m, $s_i \in \bar{S}_m$, and h_{S_m} is the coefficients of the noiseless data $\bar{y}^N = [\bar{y}(1), \cdots \bar{y}(N)]^T$ in S_m. The least square estimate of coefficients in each subspace using the noisy data is

$$\hat{h}_{S_m} = \frac{1}{N} A_{S_m}^T y^N = h_{S_m} + \frac{1}{N} A_{S_m}^T w. \tag{6.21}$$

Therefore, for the subspace error we have

$$||\hat{h}_{S_m} - h_{S_m}||_2^2 = \frac{1}{N} ||A_{S_m}^T w||^2 \tag{6.22}$$

$$||\hat{h}_{S_m} - h||_2^2 = ||\hat{h}_{S_m} - h_{S_m}||_2^2 + ||\Delta_{S_m}||_2^2. \tag{6.23}$$

The additive noise has a normal distribution of $N(0, \sigma_w^2)$. Therefore, y^N is an element of a Gaussian random variable Y^N and $\hat{h}_{S_m}$ is also an element of a Gaussian random variable $\hat{H}_{S_m}$. Both errors in (6.22) and (6.23) are Chi-square random variables. Expected value and variance of coefficient error $Z_{S_m} = ||\hat{H}_{S_m} - h||_2^2$ is

$$\mathrm{E}(Z_{S_m}) = \mathrm{E}||\hat{H}_{S_m} - h||_2^2 = \frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||^2 \tag{6.24}$$

$$\mathrm{var}(Z_{S_m}) = \mathrm{var}||\hat{H}_{S_m} - h||_2^2 = \frac{2m^2}{N^2}(\sigma_w^2)^2. \tag{6.25}$$

If the norm of the discarded vector coefficients in each subspace, $||\Delta_{S_m}||^2$, was known, how do we choose the subspace which best represents the data? The suggestion is to compare $||\hat{h}_{S_m} - h||_2^2$ of different subspaces. If we compare subspaces with the same order, m, the error random variable in each subspace has the same variance of $\frac{2m^2}{N^2}(\sigma_w^2)^2$. Therefore, we can only compare the expected values of the error and pick the subspace which has the minimum $||\Delta_{S_m}||_2^2$. The expected value of the error has two components, one caused by the noise and other by the ignored vector coefficients. The tradeoff between the noise related and the ignored coefficients related parts minimizes the expected value of the error for some m. This is called the bias-variance tradeoff method. Here we argue that ignoring the variance of the random variable can be problematic. For example, what if we are comparing two subspaces with different orders? Instead of comparing only the expected values, lets compare an *event* happening in each subspace with same probability. In this case both expected value and variance of the random variable might be involve in our decision. Assume that for a particular $\bar{m}$ the expected value of error is minimized. Therefore, the expected value for the subspace of order $\bar{m}-1$, $E_{S_{\bar{m}-1}}$, is larger than $E_{S_{\bar{m}}}$. However, the variance of the error in $S_{\bar{m}-1}$ is smaller than the variance in $S_{\bar{m}}$. Therefore, when we are comparing two events in these two spaces, which occur with the same probability, the worst case error might be smaller in space $S_{\bar{m}-1}$ than in $S_{\bar{m}}$.

The *event*, we consider in each subspace, is that the random variable $z_{S_m} = ||\hat{h}_{S_m} - h||_2^2$ is around its mean with a given probability $P1$

$$\Pr\{|Z_{S_m} - E(Z_{S_m})| \leq D_{S_m}\} = P1. \tag{6.26}$$

Therefore, D_{S_m} is a function of $||\Delta_{S_m}||$, σ_w, m and $P1$, and for each subspace S_m with probability $P1$ the error is between the following bounds

$$\frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||^2 \pm D_{S_m}(P1, \sigma_w, m, ||\Delta_{S_m}||). \tag{6.27}$$

To find the optimal subspace we suggest to choose the subspace which minimizes the worst case error with the probability $P1$,

$$\begin{aligned} S_m^* &= \arg\min_{S_m}\{E(Z_{S_m}) + D_{S_m}\} && (6.28)\\ &= \arg\min_{S_m}\{\frac{m}{N}\sigma_w^2 + ||\Delta_{S_m}||_2^2 + D_{S_m}(P1, \sigma_w, m, ||\Delta_{S_m}||)\}. && (6.29) \end{aligned}$$

In the example we discussed previously, the variance of error in $S_{\bar{m}-1}$ is lower than the variance in $S_{\bar{m}}$. Therefore, $D_{S_{\bar{m}-1}}$, which depends on the variance, might be less than $D_{S_{\bar{m}}}$ and the worst case error in $S_{\bar{m}-1}$ might be less than that of $S_{\bar{m}}$. It is important to mention that since the variance of error is of order $\frac{1}{N^2}$, for large enough N we are able to pick $P1$ close to one and still have a bounded number for D_{S_m}. We will discuss this issue later in detail.

So far the argument was with the assumption that $||\Delta_{S_m}||$ is known. In our problem setting, however, $||\Delta_{S_m}||$ is unknown. To use a similar approach, we next suggest a method to probabilistically validate $||\Delta_{S_m}||$ using the observed noisy data.

Estimation of $||\Delta_{S_m}||$

In each subspace the data representation error is

$$\begin{aligned} \frac{1}{N}||y^N - \hat{y}^N_{S_m}||_2^2 &= \frac{1}{N}||B_{S_m}\Delta_{S_m} + G_{S_m}w||^2 \\ &= ||(\Delta_{S_m} + v)||^2 \end{aligned} \tag{6.30}$$

where $\hat{y}^N_{S_m} = A_{S_m}\hat{h}_{S_m}$ and $G_{S_m} = (I - \frac{1}{N}A_{S_m}A^T_{S_m}) = \frac{1}{N}B_{S_m}B^T_{S_m}$ is a projection matrix. Therefore, $G_{S_m}w = v$ where v_i's are independent Gaussian random variables. Note that using the Parseval's theorem we already know that

$$\frac{1}{N}||y^N - \hat{y}^N_{S_m}||_2^2 = ||\frac{1}{N}B^T_{S_m}y^N||_2^2 = \sum ||\hat{h}_{\bar{S}_m}||^2. \tag{6.31}$$

The data error $X_{S_m} = \frac{1}{N}||Y - \hat{Y}_{S_m}||_2^2$ is also a Chi-square random variable for which

$$E(X_{S_m}) = (1 - \frac{m}{N})\sigma_w^2 + ||\Delta_{S_m}||^2 \tag{6.32}$$

and $\text{var}(X_{S_m})$ is

$$\frac{2}{N}(1 - \frac{m}{N})\sigma_w^2((1 - \frac{m}{N})\sigma_w^2 + 2||\Delta_{S_m}||^2). \tag{6.33}$$

Given the noisy data, one sample of this random variable is available. We call this observed error x_{S_m}. Note that the variance of the data error is of order $\frac{1}{N}$ of its expected value. Therefore, one method of estimating $||\Delta_{S_m}||$ is to assume that this one sample is a good estimate of its expected value,

$$||\hat{\Delta}_{S_m}||_2^2 \approx x_{S_m} - (1 - \frac{m}{N})\sigma_w^2. \tag{6.34}$$

This can be a convenient method of estimation of $||\Delta_{S_m}||$ when N is large enough. However, since we want to use the estimate to compare the different subspaces, we have to be more precise in the estimation process: Each X_{S_m} has a different variance and the confidence on the estimate is different for each of the subspaces even as N grows. So how "relatively" close we are to the estimate in each subspace is very important. As a result we suggest the following validation method for estimation and comparison of $||\Delta_{S_m}||_2^2$ in different subspaces.

The Chi-square probability distribution of the data error is a function of $||\Delta_{S_m}||$ and the noise variance, i.e., $f_{X_{S_m}}(x_{S_m}; m, \sigma_w, ||\Delta_{S_m}||)$. We suggest validating $||\Delta_{S_m}||$ such that X_m is in the neighborhood of its mean with probability $P2$, i.e., validate $f_{X_{S_m}}(x_{S_m}; m, \sigma_w, ||\Delta_{S_m}||)$, and therefore, $||\Delta_{S_m}||$, such that

$$\Pr(|X_{S_m} - E(X_{S_m})| \leq J_{S_m}) = P2. \tag{6.35}$$

The bound J_{S_m} is a function of $||\Delta_{S_m}||$, σ_w^2, m, and $P2$, $J_{S_m}(P2, \sigma_w^2, m, ||\Delta_{S_m}||)$. Therefore, for each subspace S_m, with validation probability $P2$, we find U_{S_m} and

L_{S_m}, the upper bound and lower bound on $||\Delta_{S_m}||$, $L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}$.

Subspace Comparison

Using the estimate of $||\Delta_{S_m}||$ from the previous section, we can estimate the worst case error criterion in (6.28). The validation part finds bounds on $||\Delta_{S_m}||_2^2$. Therefore, we suggest to pick m^* such that

$$\begin{aligned} S_m^* = \quad & \arg \min_{S_m} \max_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) \\ + \quad & D_{S_m}(P1, \sigma_w, m, ||\Delta_{S_m}||)\}. \end{aligned} \tag{6.36}$$

The worst case estimate in each subspace is given with confidence probability $P1$ and validation probability $P2$. The confidence region of error here is between

$$b_{S_m} = \max_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) + D_{S_m}\}, \tag{6.37}$$

and

$$a_{S_m} = \min\{0, \min_{||\Delta_{S_m}|| \in (L_{S_m}, U_{S_m})} \{E(Z_{S_m}) - D_{S_m}\}\}. \tag{6.38}$$

Note that one choice for J_{S_m} in (6.35) is $J_{S_m} = \beta \text{var} X_{S_m}$. In this case using the Chebychev inequality we have

$$P2 \geq 1 - \frac{1}{\beta^2} \text{ or } \beta \leq \sqrt{\frac{1}{1-P2}} \tag{6.39}$$

which shows how β and $P2$ are related. How close β is to $\sqrt{\frac{1}{1-P2}}$ depends on the distribution of the error in each subspace.

6.2.1 Gaussian Estimation

In both the probabilistic and validation part we use the table of Chi-square distribution. However, in this setting we can use the central limit theorem(CLT) to approximate the Chi-square distributions with Gaussian distributions. This gives us the advantage of finding a mathematical expression for the error bounds and worst case error (6.36),(6.37), (6.38) as a function of $P1, \sigma_w, m, P2$ and the observed noisy signal.

Data Error and Estimation of $||\Delta_{S_m}||$

The data error (6.30) is of form

$$\frac{1}{N}||y^N - \hat{y}_{S_m}^N||_2^2 \quad = \quad ||\Delta_{S_m} + v||^2$$

$$= \sum_{i=1}^{N-m} (\delta_i + v_i)^2 \tag{6.40}$$

where v_i's are zero mean white Gaussian random variables with variance $\frac{\sigma_w^2}{N}$. If $N-m$ is large enough we can estimate the Chi-square distribution of the data error with a Gaussian distribution. For a Gaussian random variable X with mean m_X and variance σ_X^2 we have

$$\Pr(m_X - \alpha\sigma_X < X < m_X + \alpha\sigma_X) = Q(\alpha), \tag{6.41}$$

where $Q(\alpha) = \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. For the data error, $X_{S_m} = \frac{1}{N}||Y - \hat{Y}_{S_m}||_2^2$,

$$\mathrm{E}(X_{S_m}) = m_w + m_\delta, \tag{6.42}$$

$$\mathrm{var}(X_{S_m}) = \frac{2m_w}{N}(m_w + 2m_\delta), \tag{6.43}$$

where $m_w = (1 - \frac{m}{N})\sigma_w^2$ and $m_\delta = ||\Delta_{S_m}||_2^2$. Using the one observed data error given the noisy data, x_{S_m}, with probability $Q(\alpha)$ we have

$$|x_{S_m} - (m_v + m_\delta)| \leq \alpha\sqrt{\frac{4m_w}{N}m_\delta + v_m}, \tag{6.44}$$

where $v_m = \frac{2}{N}m_w^2 = \frac{2}{N}(1 - \frac{m}{N})^2\sigma_w^4$.

Lemma 1 Validation of (6.44) for feasible $||\Delta_{S_m}||$s provides the following upper and lower bound for $||\Delta_{S_m}||_2^2$

$$L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}, \tag{6.45}$$

where

- If $x_{S_m} \leq (m_w - \alpha\sqrt{v_m})$, there is no valid $||\Delta_{S_m}||$ given the data.
- If $(m_w - \alpha\sqrt{v_m}) \leq x_{S_m} \leq (m_w + \alpha\sqrt{v_m})$,

$$L_{S_m} = 0 \tag{6.46}$$

$$U_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha). \tag{6.47}$$

  where

$$K_{S_m}(\alpha) = 2\alpha\frac{\sigma_w}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}. \tag{6.48}$$

- If $(m_w + \alpha\sqrt{v_m}) \leq x_{S_m}$,

$$L_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} - K_{S_m}(\alpha) \tag{6.49}$$

$$U_{S_m} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + K_{S_m}(\alpha). \tag{6.50}$$

Proof In Appendix C.3. $\diamondsuit$

Note that to avoid the first case α has to be large enough such that

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}}\left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right). \tag{6.51}$$

Comparison of Subspaces

For the error in each subspace S_m, $Z_{S_m} = ||\hat{H}_{S_m} - h||^2$, in (4.7), we have

$$||\hat{h}_{S_m} - h||^2 = ||\Delta_{S_m}||_2^2 + \sum_{i=1}^{m} u_i^2, \tag{6.52}$$

where u_is are zero mean white Gaussian noises with variance $\frac{\sigma_w^2}{N}$. If m is large enough we can estimate the Chi-square distribution of the error with a Gaussian distribution. Then the probabilistic bounds on this error are provided as following. With probability $Q(\beta)$ we have

$$|Z_{S_m} - E(Z_{S_m})| \leq \beta \text{var} Z_{S_m}. \tag{6.53}$$

The bounds on expected value and variance of Z_{S_m}, in (6.24) and (6.25), can be calculated by using the bounds from lemma one. Therefore, the worst case error bound in subspace S_m with probability $Q(\beta)$ and validation probability of $Q(\alpha)$ is

$$E(Z_{S_m}) + D_{S_m} = \frac{m}{N}\sigma_w^2 + U_{S_m} + \beta\frac{2m^2}{N^2}\sigma_w^4. \tag{6.54}$$

Theorem 6.2.1 If both random variables $X_{S_m} = ||\hat{Y}_{S_m} - y||^2$ and $Z_{S_m} = ||\hat{H}_{S_m} - h||^2$ are estimated by a Gaussian distribution, then the optimum subspace in (6.36), is provided by

$$S_m^* = \arg\min_{S_m}\{\frac{m}{N}\sigma_w^2 + U_{S_m} + \beta\frac{2m^2}{N^2}\sigma_w^4\}, \tag{6.55}$$

where U_{S_m} is provided by lemma 1. The optimality of S_{m^*} is valid with probability $Q(\beta)$ and validation probability $Q(\alpha)$.

Proper Choice of α and β

In order to have the probability of validation close to one, α and β should be as large as possible. Simultaneously, to have limited tight bounds, at both stages of finding bounds on $||\Delta_{S_m}||$ in lemma 1 and finding bounds for the subspace error in (6.53), we have to choose $\alpha/\sqrt{N}$ and β/N^2 small enough. Also as a result of the validation

stage in lemma one, another necessary condition is that α satisfies the inequality in (6.51). Note that with this proper choice of α and β, the upper and lower bounds provided in (6.37) and (6.38) can be used to evaluate the quality of estimate of each subspace.

6.3 Probabilistic Assumptions on the Noiseless Data

Theorem 6.3 Assume that

1. All the model sets, S_ms, have equal probability of being the best model set to fit the data.

2. The conditional probability density function(pdf) of the unmodeled dynamics effects $g_{S_m} = ||\Delta_{S_m}||^2$ in each subspace has a uniform distribution, i.e., $0 \leq g_{S_m} \leq G$ with a uniform distribution $f_{g_m}(g) = \frac{1}{G}$.

then with the available noisy data y as N grows

$$\Pr(L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}|y) \quad \geq \quad \frac{1}{G^2}Q(\alpha), \tag{6.56}$$

where L_{S_m} and U_{S_m} are provided by lemma one. Also as N grows, the bounds on the reconstruction error $z_{S_m} = ||h - \hat{h}_{S_m}||^2$ are valid with probability

$$\Pr(a_{S_m} < z_{S_m} < b_{S_m}|y) \quad \geq \quad \frac{1}{G^2}Q(\alpha)Q(\beta) \tag{6.57}$$

where a_{S_m} and b_{S_m} are defined in (6.38) and (6.37).

Proof Because of the given prior assumptions x_{S_m} also has a uniform distribution: the distribution of the additive noise effects can be ignored, since the variance of x_{S_m} is mainly effected by the uniform distribution of g_{S_m}. Therefore, if N is large enough and m is small enough then $f(x_{S_m}) \approx \frac{1}{G}$. Also the set of x_{S_m}s satisfying the second case in lemma 1 is a zero measure set as N grows. Therefore, the probability that $L_{S_m} \leq ||\Delta_{S_m}||^2 \leq U_{S_m}$, given the observed data, is

$$\begin{aligned}
\Pr(L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m} \quad | \quad x_{S_m}) &= \frac{1}{f(x_{S_m})}\int_{L_{S_m}}^{U_{S_m}} f_{X_{S_m}|g_{S_m}}(x_{S_m}|g) f_{g_{S_m}}(g)dg, \qquad (6.58)\\
&\approx \frac{1}{G^2}\int_{L_{S_m}}^{U_{S_m}} \frac{1}{\sqrt{2\pi}\sqrt{4\sigma_w^2\frac{g}{N} + v_m}} e^{-\frac{x_{S_m}-g}{8\sigma_w^2\frac{g}{N}+2v_m}} dg,\\
&\geq \frac{1}{G^2}\sqrt{\frac{4\sigma_w^2\frac{U_{S_m}}{N} + v_m}{v_m}}\int_{L_{S_m}}^{U_{S_m}} \frac{1}{\sqrt{2\pi}\sqrt{v_m}} e^{-\frac{x_{S_m}-g}{8\sigma_w^2\frac{L_{S_m}}{N}+2v_m}} dg,
\end{aligned}$$

$$\geq \frac{1}{G^2}\sqrt{\frac{4\sigma_w^2\frac{U_{S_m}}{N}+v_m}{4\sigma_w^2\frac{L_{S_m}}{N}+v_m}}Q\left(\frac{U_{S_m}-x_{S_m}}{\sqrt{4\sigma_w^2\frac{L_{S_m}}{N}+v_m}}\right).$$

Note that as N grows, the lower bound approaches $\frac{1}{G^2}Q(\alpha)$:

$$\Pr(L_{S_m} \leq ||\Delta_{S_m}||_2^2 \leq U_{S_m}|x_{S_m}) \geq \frac{1}{G^2}Q(\alpha), \tag{6.59}$$

Given any g_{S_m}, the probabilistic upper and lower bounds on z_{S_m}, the reconstruction error, are

$$f_1(g_{S_m}) = m_{z_{S_m}} - \beta\sigma_{z_{S_m}}, \quad f_2(g_{S_m}) = m_{z_{S_m}} + \beta\sigma_{z_{S_m}} \tag{6.60}$$

where $m_{z_{S_m}}$ is the mean and $\sigma_{z_{S_m}}$ is the variance of Z_{S_m}. Given a fixed $g_{S_m} = ||\Delta_{S_m}||_2^2$ we pick the confidence region such that

$$|z_{S_m} - m_{z_{S_m}}| \leq \beta\sigma_{z_{S_m}}. \tag{6.61}$$

Therefore, we have

$$\Pr\left(|z_{S_m} - m_{z_{S_m}}| \leq \beta\sigma_{z_{S_m}}\right) \approx Q(\beta). \tag{6.62}$$

Define a_{S_m} and b_{S_m} as follows

$$a_{S_m} = \min_{L_{S_m} \leq g_{S_m} \leq U_{S_m}} f_1(g_{S_m}) \tag{6.63}$$

$$b_{S_m} = \max_{L_{S_m} \leq g_{S_m} \leq U_{S_m}} f_2(g_{S_m}) \tag{6.64}$$

These are the bounds we provided in (6.37) and (6.38) for the reconstruction error. Therefore, using (6.59), we conclude that

$$\begin{aligned}\Pr(a_{S_m} < z_{S_m} < b_{S_m}|x_{S_m}) &= \Pr(a_{S_m} < z_{S_m} < b_{S_m}|x_{S_m}, L_{S_m} \leq g_{S_m} \leq U_{S_m}) \times \\ &\quad \Pr(L_{S_m} \leq g_{S_m} \leq U_{S_m}|x_{S_m}) \qquad (6.65)\\ &\geq \frac{1}{G^2}Q(\alpha)\int_{L_{S_m}}^{U_{S_m}} \Pr(a_{S_m} < z_{S_m} < b_{S_m}|g)dg \qquad (6.66)\end{aligned}$$

using (6.62)

$$\Pr(a_{S_m} < z_{S_m} < b_{S_m}|x_{S_m}) \geq \frac{1}{G^2}Q(\alpha)Q(\beta) \tag{6.67}$$

Note that the conditional probabilities in each subspace is found by using x_{S_m} which is provided by y. Therefore, x_{S_m} can be replaced by y in (6.59) and (6.67).

6.4 Discussion on Normalized MDL Denoising

The normalized MDL, described in section 6.1.3, is in form of

$$\hat{f}(y^N; S_m) = \frac{f_{S_m}(y^N; \hat{h}_{S_m}(y^N), \hat{\sigma}_w^2(y^N))}{\int_{Z(R,\sigma_0^2)} f_{S_m}(z; \hat{h}_{S_m}(z), \hat{\sigma}_w^2(z))dz} \tag{6.68}$$

where the information about the set $Z(R, \sigma_0^2)$ is given in section 6.1.3. This description length is claimed to be minimizing the following criterion

$$\min_q \max_x \mathrm{Ln} \frac{f_{S_m}(x; \hat{h}_{S_m}(x), \hat{\sigma}^2(x))}{q(x)} \tag{6.69}$$

which is defined based on the universal coding in [47]. The universal coding of a single message in [47] is as follows.

Universal Coding Theorem Let $\mathcal{A}$ be a describe alphabets, α_i of $m \geq 2$ letters. $\mathcal{A}^N$is the set of all m^N sequences $\alpha^N = \alpha_1, \cdots \alpha_N$ can be generated with any source $\theta \in \Theta$ with distribution $p(\alpha^k; \theta)$. We use a code denote by q with codelength of $L_q(\alpha^N)$ to code $\alpha^N \in \mathcal{A}^N$. The quantity $-\log p(\alpha^N; \theta)$ is naturally interpreted as the amount of information contained in the block α^N on the output of the source θ. Then

$$\rho(\alpha^N; L_q, \theta) = \frac{1}{N}(L_q(\alpha^N) + p(\alpha^N; \theta)) \tag{6.70}$$

is the *redundancy of coding the block* α^N *on the output of the source* θ *by code* q. For the elements of Θ define

$$\begin{aligned} \rho(\alpha^N; L_q, \Theta) &= \sup_\theta \rho(\alpha^N; L_q, \theta) \qquad (6.71) \\ &= \frac{1}{N}(L_q(\alpha^N) + p(\alpha^N; \Theta)) \end{aligned}$$

where

$$p(\alpha^N; \Theta) = \sup_\theta p(\alpha^N; \theta). \tag{6.72}$$

The goal is to find a code q for which the redundancy as a function of the set Θ is

$$\rho(\Theta) = \inf_q \max_{\alpha^N} \rho(\alpha^N; L_q, \Theta) \tag{6.73}$$

Note that this quantity is an upper bound for the "mean" redundancy defined as

$$\begin{aligned} r(\Theta) &= \inf_q \max_{\alpha^N} r(\alpha^N; L_q, \Theta) \qquad (6.74) \\ &= \inf_q \max_{\alpha^N} \sup_\theta E_\theta(\rho(\alpha^N, L_q, \theta)). \qquad (6.75) \end{aligned}$$

In [47] it is proved that for any source Θ

$$\rho(\Theta) \geq s(\theta) \tag{6.76}$$

where

$$s(\Theta) = \sum_{\alpha^N} p(\alpha^N; \Theta) \tag{6.77}$$

and there exists a uniquely decodable code q with codeword length

$$L_{q^*}(\alpha^N) = \lceil \log s(\Theta) - \log p(\alpha^N; \Theta) \rceil \tag{6.78}$$

with the corresponding probability distribution

$$q^*(\alpha^N; \Theta) = \frac{p(\alpha^N; \Theta)}{s(\Theta)} \tag{6.79}$$

such that

$$\max_{\alpha^N} r(\alpha^N; L_{q^*}, \Theta) \leq \frac{1}{N}(s(\Theta) + 1) \leq \rho(\Theta) + \frac{1}{N}. \tag{6.80}$$

Note that Shannon coding theorem results when Θ only has one element.$\diamondsuit$

How does the transition from this theorem to the normalized MDL happen? Here y^N is considered to be one of the α^Ns. Therefore, a form of descritization is considered such that the probabilities in the universal coding are replaced by the probability distributions. The min-max criterion in (6.69) has to be the redundancy function, (6.73), for this setting

$$\rho(S_m) = \min_q \max_x \sup_{h_{S_m} \in S_m} \mathrm{Ln} \frac{f_{S_m}(x; h_{S_m})}{q(x)} \tag{6.81}$$

where the minimum is attained for a code q^*. From (6.79), this code is defined in form of

$$q^*(y^N, S_m) = \hat{f}(y^N; S_m) \quad = \quad \frac{\sup_{h_{S_m} \in S_m} f_{S_m}(y^N; h_{S_m})}{s(S_m)}. \tag{6.82}$$

To obtain the normalized MDL in [43] the numerator is replaced by

$$\sup_{h_{S_m} \in S_m} f_{S_m}(y^N; h_{S_m}) = f_{S_m}(y^N; \hat{h}_{S_m}(y^N), \hat{\sigma}_w^2(y^N)) \tag{6.83}$$

where

$$\hat{\sigma}_w^2 = \frac{1}{N}||y - \hat{y}_{S_m}||^2 \tag{6.84}$$

and the denominator $s(S_m)$, which is defined in (6.77), is replaced by

$$s(S_m) = \int_{Z(R,\sigma_0^2)} f_{S_m}(z; \hat{h}_{S_m}(z), \hat{\sigma}_w^2(z))dz \tag{6.85}$$

where

$$Z(r, \sigma_0^2) = \{z | \hat{h}'_{S_m}(z)\Sigma_{S_m}\hat{h}_{S_m}(z)/N \leq r, \hat{\sigma}_w^2 \geq \sigma_0^2\}, \tag{6.86}$$

and $\Sigma_{S_m} = A'_{S_m}A_{S_m}$ where A_{S_m} is the matrix whose columns are the bases of S_m. It is said that r and σ_0^2 are chosen such that the ML estimates fall within $Z(r, \sigma_0^2)$.

It is not clear how $s(S_m)$ is estimated with (6.85). The numbers r and σ_0 are unknown numbers that through the calculation of the MDL are replaced by their asymptotic approximates and are eliminated by considering the asymptotic behavior of the integral in (6.85). For more details please see [43].

It is important to note that in this scenario (6.83) shows that the subspaces, S_ms, are not the same as what we considered in our problem formulation. While in our discussion S_m is the set of subspaces of order m describing the noiseless data with additive noise w with fixed variance for all subspaces, here in the subspace S_m the variance of the additive noise is also a variable which is estimated by the observed data.

If the variance of the noise is considered fixed σ_w^2 for all the subspaces, one important fact is that the calculation of $f_{S_m}(y|\hat{h}_{S_m})$ is meaningful only if y has been generated with an element of S_m. The conditional probability distribution function in S_m is then defined for elements which can be represented in form of

$$y_{S_m} = \sum_{s_i \in S_m} s_i h_i + w. \tag{6.87}$$

In normalized MDL, however, the ignored basis vectors effects, $\sum_{s_i \in \bar{S}_m} s_i h_i$, is considered as a part of the additive *zero* mean noise. If such effects are indeed nonzero, the new defined noise is not anymore zero mean and it contradicts the prior assumption on the noise to be zero mean. As a result of such approach the estimates of the noise variance in different subspaces are different, even though the ignored coefficient part only effects the mean and not the variance of estimates. This causes problem in the evaluation of the description length even if the true number of the basis vectors, which has generated the noiseless data $\bar{y}$, is finite. Consider an example for which only h_1 and h_3 are nonzero and the prior knowledge is that only two basis vectors are enough to represent noiseless data. With the prior assumption that the additive noise is zero mean, the description length can be calculated. However, except for one subspace of order two, $\{s_1, s_3\}$, such assumption is not valid and the mean of the noise is the effects of $h_1 s_1$ and/or $h_3 s_3$.

6.5 New MDL Denoising

The minimum description complexity and the new MDL criteria defined in chapter 5 are valid criteria for comparison of the subspaces for the purpose of denoising. Both of these criteria are applicable if similar to normalized MDL we assume that y is an output of a source which might continue sending messages and therefore, the notion of coding and/or averaging in each subspace is meaningful. In this case we are estimating a "model" represented by $\hat{h}_{S_m}$ which describes not only the observed available data but will possibly code other signals which are generated by h.

However, in denoising problem we just observe one set of data and there is no "system model" to be estimated. Although the estimate of coefficients in this chapter and the estimate of taps of the impulse response in chapter 4 both are denoted by vector h, there is a conceptual difference between what these two vectors represent. For example in the system identification problem we estimate a model which might be used later to decode the system input in communications. Therefore, the information theoretic methods deal with optimizing the codelength of the data generated by the true model set h, when an estimate of h is used. However, in the denoising problem the goal is to find the denoised version of the signal and the estimate of the coefficients $\hat{h}_{S_m}$s does not represent any model set to code any random variable further. Although the provided description length in chapter 5 can be used in this setting we here introduce another description length which is well suited to the denoising problem:

Here the only probability distribution to be considered is provided by the additive noise which generated the noisy data in space S_N. We already know that the minimum description length of the noiseless data $\bar{y}$ is provided by the codelength of the code which represent $\bar{y}$ with probability distribution defined by h. The description length for the denoised data of subspace S_m is also the codelength of the code which is defined by the same distribution: The prior assumption is that the noisy data is generated with h

$$y = \sum_{i=1}^{N} h_i s_i + w = \sum_{s_i \in S_m} s_i h_i + \sum_{s_i \in \bar{S}_m} s_i h_i + w. \tag{6.88}$$

The code length of any signal of length N with the distribution that generated y is of form

$$L(x) = \log \frac{1}{f_h(x|h)} \tag{6.89}$$

$$= -\log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{\|x-\bar{y}\|^2}{2\sigma_w^2}} \tag{6.90}$$

where noiseless data $\bar{y} = \sum_{i=1}^{N} s_i h_i$ is the expected value of random variable Y.

Definition 6.1 The description length of S_m as the code length of the estimate of data using the estimate of h in S_m, is defined as

$$\begin{aligned} \mathrm{DL}_h(y, S_m) &= \min_{y_{S_m}} \frac{1}{N} L(y_{S_m}) = \log \frac{1}{f_h(\hat{y}_{S_m}|h)} && (6.91) \\ &= \frac{1}{N} \log \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||\hat{y}_{S_m} - \bar{y}||^2}{2\sigma_w^2}}. && (6.92) \end{aligned}$$

where y_{S_m} is an estimate of y using the basis in S_m. Comparison of such description length for different subspaces leads to comparison of the reconstruction error

$$\frac{1}{N}||\hat{y}_{S_m} - \bar{y}||^2 = ||\hat{h}_{S_m} - h||^2 \tag{6.93}$$

which was calculated in previous sections of this chapter. Therefore, the "uniquely" decodable codelength for the estimate of signal $\bar{y}$, which is corrupted by additive white Gaussian noise with variance σ_w^2, is in form of

$$\mathrm{DL}_h(y, S_m) = -\log \frac{1}{\sqrt{2\pi\sigma_w^2}} + \frac{||\hat{h}_{S_m} - h||^2}{2\sigma_w^2} \log e. \tag{6.94}$$

Note that in definition of the description length in chapter 5 we calculate the description length of $\bar{y}$ when it is generated by a system defined in each subspace S_m. Here the description length of the denoised signal in S_m is the codelength of the code describing this signal with one model which is represented by h. Anther interesting observation is that in this case the MDC and the description length, in (6.94), are the same.

Figure 6-1 shows the behavior of the denoised version of y in each subspace S_m. While the data error is always a decreasing function of m (which is zero for $m = N$), the reconstruction error is minimized for some m^* depending on the length of data, noise variance and the basis.

6.5.1 Best Basis Search

For a given noisy data one might proceed to search for the basis which is the best representative of the data. We suggest to compare the new proposed MDL of different families of basis functions. The method leads to the choice of the basis which minimizes this criterion. Among the basis B_{si} pick the one for which the new MDL

$$\mathrm{MDL}_{B_{si}}(y) = \min_{S_m \subset B_{si}} \mathrm{DL}_h(y, S_m) \tag{6.95}$$

is minimized. Therefore, the best basis among B_{si}s is

$$B_s^* = \arg\min_{B_{si}} \mathrm{MDL}_{B_{si}}(y) \tag{6.96}$$

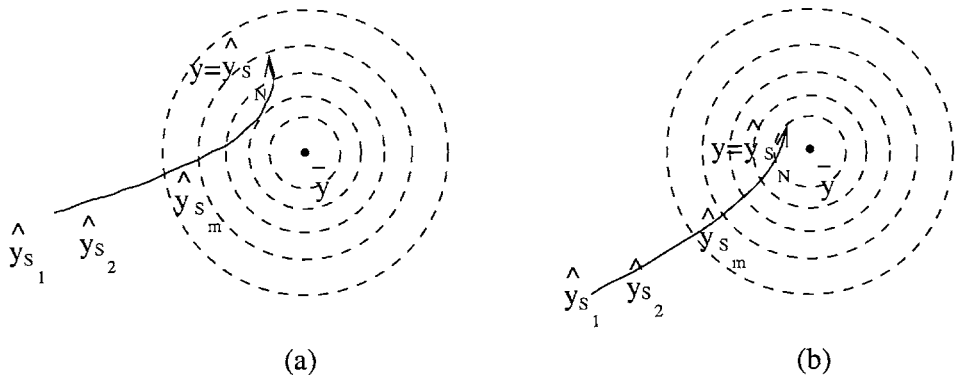

Figure 6-1: *Figures (a) and (b) show the behavior of the estimates of y for same $\bar{y}$ with different noise variances. The variance of noise in figure(a) is larger than the one in figure(b).*

6.5.2 Thresholding Denoising Methods

In threshold methods, a threshold τ, is provided before the calculation of the coefficients' estimates. It is not known that for this choice of threshold how many coefficient estimates, which are less than τ, are due to the additive noise only. In cases where we know a priori that there are few nonzero coefficients to represent the noiseless part of the data, it might be intuitive to pick the threshold only as a function of the variance of the noise and the length of the data, as it is suggested in [11]. But as [29] shows without such prior assumption it is not trivial to decide on the optimum threshold beforehand.

What we showed in our method is that the critical possible thresholds are the absolute values of the coefficient estimates. Lets sort the basis vectors based on the absolute value of the coefficient estimates. Our method is computing the estimation error for any of those absolute values as the threshold. We find the optimal of those thresholds comparing the error estimation of such thresholding. Depending on the tradeoff between the eliminated coefficients and the noise effects there is a subspace for which the estimation error is minimized.

Similar to our approach, MDL denoising suggests a criterion to be calculated for different subspaces. However, as we discussed previously, in this method the effect of the eliminated coefficients in each subspace is considered as a part of the additive noise. The comparison of the NML criterion for different subspaces asymptotically provides a threshold which is only a function of the variance and the length of the data. In [43] the normalized MDL threshold is shown to be

$$\tau \approx \sigma_w \sqrt{\log N} \tag{6.97}$$

which is half of what is suggested in wavelet thresholding. As we argued previously, in the proposed method, there is a distinction between the noise effects and the eliminated coefficients effects in different subspaces. Therefore, even the asymptotic results can not provide a threshold which is only a function of the noise variance and

the length of the data. The optimal threshold is sensitive to the coefficient estimates of all the basis vectors.

6.5.3 Unknown Noise variance

If the variance of the additive noise is unknown but bounded, we can estimate the variance as follows. Calculate the description length of the data as a function of σ_w:

$$\mathrm{DL}_h(y, S_m, \sigma_w) = -\log \frac{1}{\sqrt{2\pi\sigma_w^2}} + \frac{||\hat{h}_{S_m} - h||^2}{2\sigma_w^2} \log e. \tag{6.98}$$

Therefore, the MDL (y, σ_w) is

$$\mathrm{MDL}(y, \sigma_w) = \min_{S_m} \mathrm{DL}_h(y, S_m, \sigma_w^2) \tag{6.99}$$

Choose the variance such that

$$\sigma_w^* = \arg\min_{\sigma_w} \mathrm{MDL}(y, \sigma) \tag{6.100}$$

We illustrate the application of this method in an example in the simulation section.

6.6 Application: Blind Channel Identification

One potential application of the new denoising method is in blind channel identification. An important application of the blind channel identification is in wireless communications. The time varying system in this case is identified by the use of training signals which are send periodically between the communication signal. Main methods of blind channel estimation are surveyed in [52]. Various existing algorithms are classified into the moment-based and the maximum likelihood(ML) methods. The ML methods are usually optimal for large data records and unlike moment-based methods cannot be obtained in closed form. In method of moments the focus is on matching the second-order moment of the noisy data to that of the available trained signal. Most of the available methods assume a prior knowledge of the length of the impulse response.

Second order statistical subspace method is one of the important methods of moments. In this method first the length of the impulse response is estimated using the MDL estimate in [59]. Then the noisy moment provided by the output is denoised using the prior assumption on the noise variance, and the obtained length of the impulse response. Therefore, the denoising step is sensitive to assumption on the length of the impulse response. The last step is a curve fitting algorithm which uses the denoised moment to find the best estimate of the impulse response. Here we suggest to combine the first two steps, the order estimation and the denoising step, by using the proposed MDL denoising method.

6.7 Simulation

The unit-power signal shown in Figure (6-2) used to illustrate the performance of the MDC/MDL denoising method. Figure (6-3) shows the absolute value of the discrete Fourier transform of the noiseless signal. Figure (6-4) shows the subspace error of

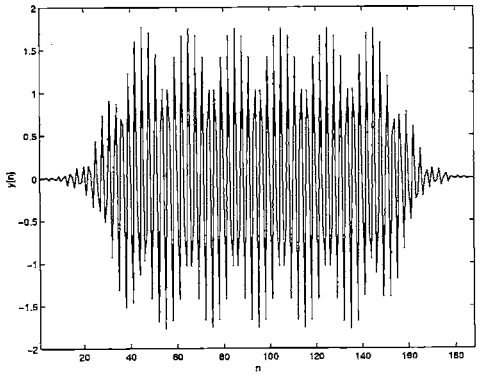

Figure 6-2: *Noiseless unit-power signal of length 188.*

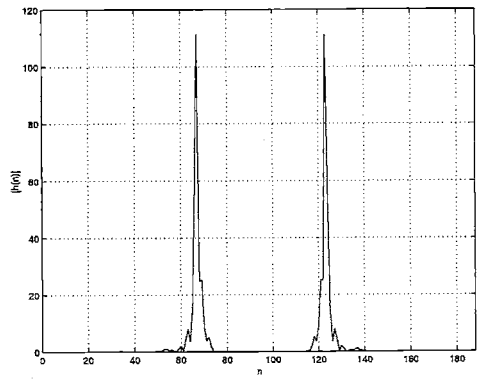

Figure 6-3: *188 points discrete Fourier transform of the noiseless signal.*

the noiseless data. The subspace of order m is the one among the subspaces of same order which minimizes the error. As we expect such error decreases as the subspace order increases. Figure (6-5) shows the subspace error in presence of additive noise with variance 0.25. It shows that the subspace error in this scenario is minimum for S_7 and our method also picks S_7. If the variance of additive noise is very small we can use a threshold on the error (or the description length). Note that since the description length proposed in this chapter is a valid uniquely decodable code, defining a threshold on this error is valid. Note that as we showed in 5.2.3, the normalized MDL is ignoring a normalizing constant and a threshold used in one setting is not comparable with a denoising problem with a different $\bar{y}$. Figure(6-6 shows the simulation results of the same noiseless data when it is corrupted by a noise with very small variance. The minimum is attained for S_{25}, however, using a threshold either for the reconstruction error or for the description length can provide a subspace with less order.

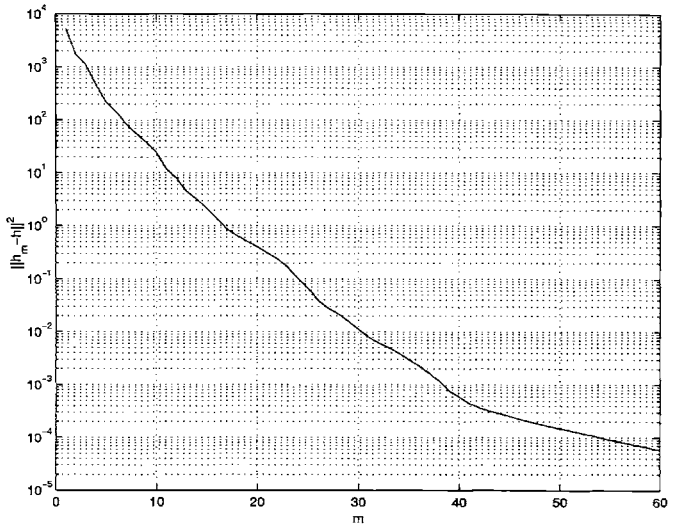

Figure 6-4: *Subspace error for the noiseless signal with subspaces of order m. The subspace with order m is the one among all the subspaces with same order which minimizes the error.*

6.7.1 Unknown Noise Variance

In an example here we show how to use the new minimum description length (6.98) to estimate the noise variance. Assume that the data corrupted with AWGN which has variance of 0.25. First step is to find the valid αs for which the upper bound can be calculated. When the variance of noise is known the lower bound for α can be found using the condition from lemma one

$$\alpha \geq \frac{N}{\sqrt{2(N-m)}} \left(1 - \frac{m}{N} - \frac{x_{S_m}}{\sigma_w^2}\right). \tag{6.101}$$

If variance of noise is .25 the available data shows that any α greater than .5 is valid. Here we check for proper choice of α by using the MDL. Figure (6-7) shows the MDL for variable variances when α varies. The minimum valid α is the one for which the minimum description length still is a positive number. In this case, for $\beta = 1$, as the simulation shows the lower bound for α is .64. The lower bound is obtained through validation the condition in lemma one and the MDL defined in (6.98). Next we choose a valid α and choose the variance for which the MDL is minimized for that α. As figure(6-7) shows for $\alpha = 1$ the optimum variance is .27. Figure(6-8) shows the description length of the data with variance .27. In this case S_8 is chosen as the best subspace. The same figure also shows the true description length of the data with the known variance. In this case the validation probability and the confidence probability are both $Q(1) = .68$. Note that the simulations shows that the algorithm is very robust on the choice of α. For example for $\alpha = 2$ the optimum variance as figure (6-7) shows is $\sigma^* = .6$, ($\sigma_w = .36$) and in this case still S_8 is the optimum subspace.

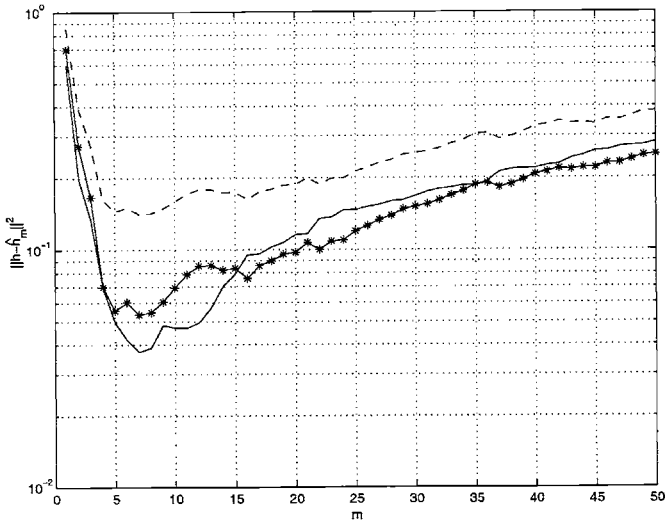

Figure 6-5: *Subspace error for the noisy signal with subspaces with order m. The subspace with order m is the one among all the subspaces with same order which minimizes the error. Noise variance is* $\sigma_w^2 = .25$. *The solid line is the subspace error. The line with "*" is the estimate of the expected value of subspace error using the proposed method. The dashed line is the error's upper bound,* U_{S_m} *in (6.54), with* $\alpha = \log(N)/2$ *and* $\beta = \log(N)$.

6.7.2 Search for the Optimum Finite Number of Coefficients

For the search of optimum subspace we suggest the following algorithm. Search among the subspaces of order one to minimize the DL, this search is of order N. Then search among the subspaces of order 2 which includes the one basis function provided in first step of the search, this search is of order $N-1$. Continue the search for higher order basis and the total search is of order

$$N + (N-1) + \cdots + 1 = \frac{N(N-1)}{2} \tag{6.102}$$

Another search method is to first sort the basis functions, in decreasing order, based on the absolute value of its estimated coefficients. The best S_1 is then the one which has the first basis of the sorting result. The best S_2 has the first two elements of the sorting process. This search, therefore, is of order one!. Figure(6-9) shows the result of using both the algorithms. As it is expected the second method provides a higher error. Note that the examples given in this section are provided with the first search method. In the second method the additive distribution of the noise is no more independent. The dependence is caused by sorting the coefficients. For example by having the first n elements of the sorted error, we have an upper bound on the next coefficient error and this is an information imposed on the next additive Gaussian distribution of the $n+1$st element.

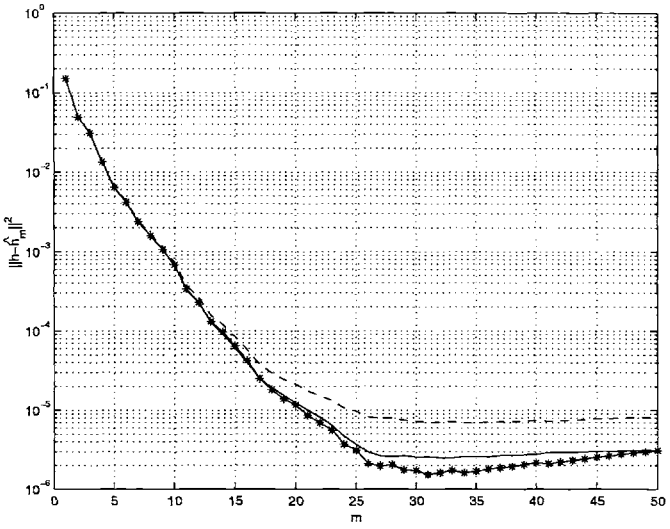

Figure 6-6: *Subspace error for the noisy signal for subspaces with order m. The subspace with order m is the one among all the subspaces with same order which minimizes the error. Noise variance is* $\sigma_w^2 = 6.25 \times 10^{-6}$. *The solid line is the subspace error. The line with "*" is the estimate of the expected value of subspace error using the proposed method. The dashed line is the error's upper bound,* U_{S_m} *in (6.54), with* $\alpha = \log(N)$ *and* $\beta = \log(N)$.

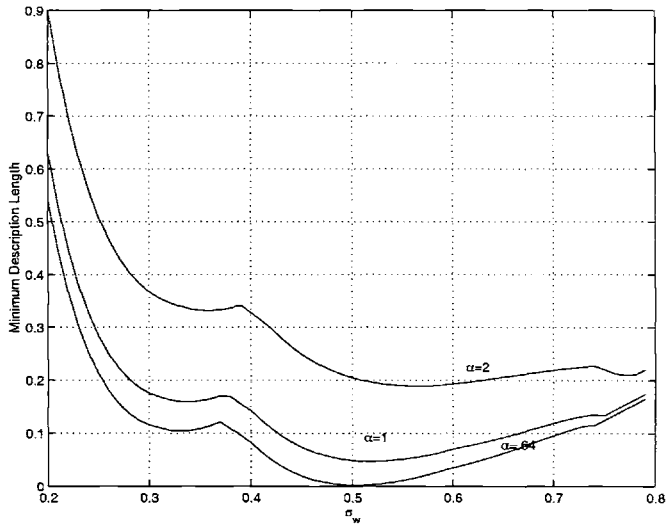

Figure 6-7: *MDL for variable standard deviations from .2 to .8 and for different* αs *with* $\beta = 1$.

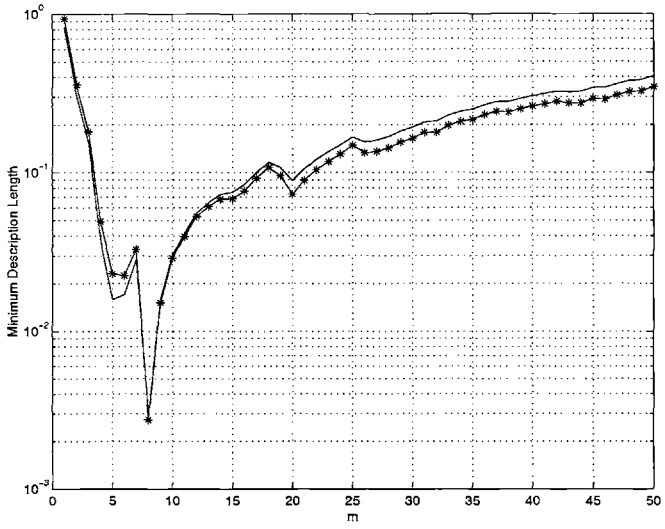

Figure 6-8: *For $\beta = 1$ and $\alpha = 1$: Solid line is the description length with the true variance $\sigma_w^2 = .25$ ($\sigma_w = .5$). Line with '*' is the DL provided with the estimated variance $\sigma_w^2 = .27$ ($\sigma_w = .64$).*

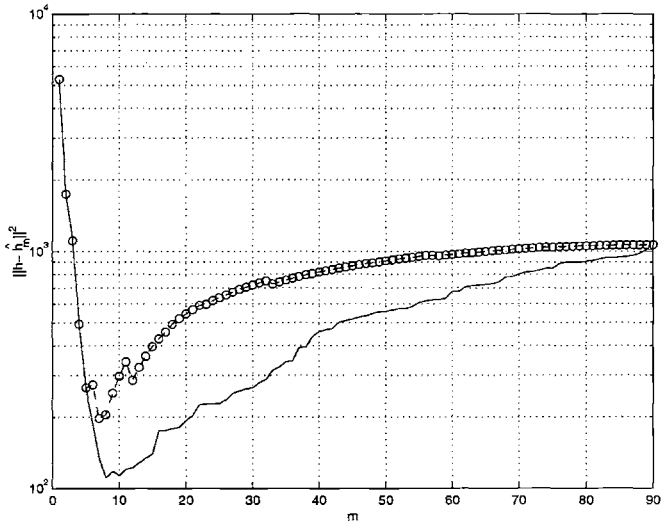

Figure 6-9: *Solid line is the subspace error resulted with search of order N^2 and line with "o" is the subspace error resulted by sorting the coefficients estimate with search of order 1.*

Chapter 7

Conclusion and Future Work

In this thesis we introduced a new method of parametric model selection. The method was proposed after extensive study of the theory of the existing order estimation methods such as AIC, MDL and NMDL. The new approach, minimum description complexity method, is based on the Kullback-Leibler information distance. The key advantage of the proposed method over the existing methods is that, unlike the prior asumption in calculation of existing methods, no assumption on the true model belonging to all the competing sets is needed. We provided a probabilistic method of MDC estimation for a class of parametric model sets. The main focus in this calculation was how to extract the most information from the given data. While the obtained MDC for this class of parametric models covers a wide range of applications in different areas, one challenge for future research is to calculate MDC for more general parametric sets.

We explored the application of MDC in two fields, system identification and signal denoising. We will close with more future directions in these application areas.

7.1 System Identification and Quality Evaluation

The problem of identification and quality evaluation of stable LTI systems was investigated. We thoroughly studied the existing methods of both deterministic and stochastic approaches to this problem. The methods seem to lack a proper prior assumption on the model structure and/or the description of the additive noise.

One interesting observation was that all the quality evaluations methods in system identification focus on calculation of the impulse response error without a complete use of the observed data, while the information theoretic approaches implement the observed data efficiently when the impulse response is an element of the competing model set. Here we presented a powerful method which uses the information theory tools to extract the observed data's information in order to estimate the impulse response error.

We were able to bridge between set-membership and stochastic identification approaches by introducing a proper deterministic noise definition. The definition is not as conservative as the existing bounded norm definitions and satisfies the sufficient

assumptions of the second order stochastic method of MDC estimation.

We examined the practical advantages of implementation of MDC over AIC and MDL. In practical problems, and due to the method of calculation of AIC and MDL, these methods are very sensitive to the length of the observed data and the signal to noise(SNR) ratio. We illustrated how implementation of MDC can avoid these problems.

More related practical problems such as zero-pole estimation and time series analysis were discussed briefly and MDC for these cases were provided theoretically. The application of MDC and comparison of the existing methods with MDC in these areas is another subject which clearly deserves more attention in future research. Also, for practical purposes it is beneficial to develop a recursive method of MDC calculation based on the provided estimation method. One other potential field to explore is off-line nonlinear identification. In this case, with a proper input design, MDC can be implemented to provide a piece-wise linear estimate of the system.

7.2 Signal Denoising and Data Representation

Another interesting problem we reviewed is data denoising and signal representation. We briefly reviewed the thresholding methods which are the well-known denoising approaches. We suggested to choose a new algorithm based on a subspace comparison. MDC was proposed and implemented as a subspace comparison method. We elaborated the theoretical superiority of MDC over the existing thresholding denoising methods. Application of MDC for this denoising problem also provides a new method for best basis search. The application of MDC in signal denoising promises to cover a broad set of applications in a variety of fields. For example one potential application is in blind channel identification. In available methods of blind channel identification first an estimate of the length of impulse response is calculated then the available noisy second order statistics of the data is denoised. MDC is able to combine these two steps and provide the estimate of the impulse response length and denoise the data's second order statistics in one step.

Appendix A

AIC and BIC

A.1 Estimation of Cost Function

For the cost function $V_m^N(\theta_m, Z^N)$ we have

$$\bar{V}_m(\theta_m) := \lim_{N\to\infty} EV_m^N(\theta_m, z^N) \tag{A.1}$$

where the expectation is taken over the true probability distribution of data $f(z^N, \theta)$. The parameter which minimizes this criterion is θ_m^*

$$\theta_m^* = \arg\min_{\theta_m} \bar{V}_m(\theta_m) \tag{A.2}$$

As N goes to infinity $\hat{\theta}_m^N$ converges to θ_m^* and therefore the Taylor expansion of $\bar{V}_m(\hat{\theta}_m^N)$, ignoring the terms with order higher than two, is

$$\bar{V}(\hat{\theta}_m^N) \approx \bar{V}_m(\theta_m^*) + \frac{1}{2}(\hat{\theta}_m^N - \theta_m^*)^T \bar{V}_m''(\xi_{1N})(\hat{\theta}_m^N - \theta_m^*). \tag{A.3}$$

Therefore the expected value of this random variable is

$$E\bar{V}(\hat{\theta}_m^N) \approx \bar{V}_m(\theta_m^*) + E\frac{1}{2}(\hat{\theta}_m^N - \theta_m^*)^T \bar{V}_m''(\xi_{1N})(\hat{\theta}_m^N - \theta_m^*). \tag{A.4}$$

Also assuming that N is large enough, functions $\bar{V}_m(\theta_m)$ and $EV_m^N(\theta_m, z^N)$ are close. Therefore the Taylor expansion of $EV_m^N(\theta_m, z^N)$ is

$$EV_m^N(\hat{\theta}_m^N, Z^N) \approx \bar{V}_m(\theta_m^*) - E\frac{1}{2}(\hat{\theta}_m^N - \theta_m^*)^T \bar{V}_m''(\xi_{2N})(\hat{\theta}_m^N - \theta_m^*) \tag{A.5}$$

where ξ_{1N} and ξ_{2N} are some intermediate points between θ_m^* and $\hat{\theta}_m^N$. Therefore $\bar{V}_m''(\xi_{iN}) \to \bar{V}_m''(\theta_m^*)$ as N goes to ∞ with probability one and

$$E\frac{1}{2}(\hat{\theta}_m^N - \theta_m^*)^T \bar{V}_m''(\xi_{iN})(\hat{\theta}_m^N - \theta^*) \;=\; \frac{1}{2}Etr\{\bar{V}_m''(\xi_{iN})(\hat{\theta}_m^N - \theta_m^*)(\hat{\theta}_m^N - \theta_m^*)^T\} \tag{A.6}$$

$$\approx \frac{1}{2} tr \bar{V}_m''(\theta_m^*) P_m^N \tag{A.7}$$

with $P_m^N = P_{\theta_m^*}^m / N$ and $\sqrt{N}(\hat{\theta}_m^N - \theta_m^*) \to N(0, P_{\theta_m^*}^m)$ in distribution. Therefore

$$E\bar{V}_m(\hat{\theta}_m^N) \approx \bar{V}_m(\theta_m^*) + \frac{1}{2} tr \bar{V}_m''(\theta_m^*) P_m^N \tag{A.8}$$

$$E(V_m^N(\hat{\theta}_m^N, z^N) \approx \bar{V}_m(\theta_m^*) - \frac{1}{2} tr \bar{V}_m''(\theta_m^*) P_m^N. \tag{A.9}$$

Subtracting (A.9) from (A.8), we conclude

$$J(S_m) = E\bar{V}_m(\hat{\theta}_m^N) \approx EV_m^N(\hat{\theta}_m^N, Z^N) + tr \bar{V}_m''(\theta_m^*) P_m^N. \tag{A.10}$$

A.2 Calculation of BIC

$$\theta_m' \cdot l(y^N) - b(\theta_m) = C_m - \lambda ||\theta_m - \hat{\theta}_m||^2 \tag{A.11}$$

and $C_m = \hat{\theta}_m' \cdot l(y^N) - b(\hat{\theta}_m) = \max_{\theta_m \in S_M} \theta_m' \cdot l(y^N) - b(\theta_m)$ for some $\hat{\theta}_m = (\hat{\theta}_{m1}, \cdots, \hat{\theta}_{mm}) \in S_m$.

The density of $\theta \in S_m$ is obtained by the Lebesgue measure on S_m, define $\mu(S_m) = \mu_m$. Then we have a uniform density for θ_m in S_m and $\mathrm{d}\mu_\mathrm{m}(\theta_\mathrm{m}) = \frac{\mathrm{d}\theta_\mathrm{m}}{\mu_\mathrm{m}}$. Hence (3.20) is in form

$$\begin{aligned} S(y^N, S_m) &= \log \alpha_m + \log \frac{1}{\mu_m} \int_{S_m} e^{(C_m - \lambda ||\theta_m - \hat{\theta}_m||^2)N} d\theta_m \\ &= \log \alpha_m + NC_m - \log \mu_m + \log \int_{S_m} e^{(-\lambda ||\theta_m - \hat{\theta}_m||^2)N} d\theta_m. \end{aligned} \tag{A.12}$$

We have

$$\int_{S_m} e^{(-\lambda ||\theta_m - \hat{\theta}_m||^2)N} d\theta_m = \int_{\theta'} e^{(-\lambda ||\theta' - \hat{\theta}'||^2)N} \left(\int_{l(\theta'}^{u(\theta')} e^{-\lambda(\theta_{mm} - \hat{\theta}_{mm})^2 N} d\theta_{mm} \right) d\theta'. \tag{A.13}$$

where $\theta' = (\theta_{m1}, \cdots, \theta_{m(m-1)})$ and $l(\theta')$ and $u(\theta')$ are the upper and lower bounds for the θ_{mm}. Since this integral is bounded for any N we have

$$\begin{aligned} &\lim_{N \to \infty} \int_{S_m} e^{(-\lambda ||\theta_m - \hat{\theta}_m||^2)N} d\theta = \\ &\lim_{N \to \infty} \int_{\theta'} e^{(-\lambda ||\theta' - \hat{\theta}'||^2)N} \left(\lim_{N \to \infty} \int_{l(\theta'}^{u(\theta')} e^{-\lambda(\theta_{mm} - \hat{\theta}_{mm})^2 N} d\theta_{mm} \right) d\theta'. \end{aligned} \tag{A.14}$$

For a Gaussian distribution w with mean w_0 and variance σ^2, if $w_0 \in [a, b]$

$$\lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{\frac{(w-w_0)^2}{2\sigma}} \to 1 \tag{A.15}$$

and if w_0 is not in $[a, b]$ the integral goes to 0 as σ goes to 0. By applying this fact for $w = \theta_{mi}$s, $w_0 = \hat{\theta}_{mi}$, $\sigma^2 = 1/2N\lambda$, the inner integral in (A.14) is

$$\int_{l(\theta')}^{u(\theta')} e^{-\lambda(\theta_{mm} - \hat{\theta}_{mm})^2 N} d\theta_{mm} \to \sqrt{2\pi\sigma^2} = \sqrt{\frac{\pi}{N\lambda}} \tag{A.16}$$

if $\theta_m m \in [l(\theta'), u(\theta']$, otherwise the integral is zero. Note that this assumption requires that θ be an element of the model set S_m. Therefore

$$\int_{S_m} e^{(-\lambda||\theta_m - \hat{\theta}_m||^2)N} d\theta = \sqrt{\frac{\pi}{N\lambda}} \int_{\theta'} e^{(-\lambda||\theta' - \hat{\theta}'||^2)N} d\theta'. \tag{A.17}$$

By repeating the procedure for each parameter θ_{mi}, and with the assumption that $\theta_m \in S_m$, we conclude

$$S(y^N, S_m) = \log \alpha_m + NC_m - \log \mu_m + \log \left(\sqrt{\frac{\pi}{n\lambda}} \right)^m \tag{A.18}$$

Appendix B

Proof of Theorem 3.4.2

Our approach is similar to the proof Rissanen has for his first theorem in [41]. At some point of the proof Rissanen uses function $\log(N)$ and we generalize the proof to a family of functions which satisfy the sufficient condition needed for the proof. Similar to the proof in [41], the proof here is presented in three steps. The first and last steps are the same as the existing proof. The second step however is modified.

For each θ in Ω, let $E_N(\theta)$ be a neighborhood of radius $r_N = c/\sqrt{N}$ with θ as the center. A θ-typical string of length N is

$$Y_N(\theta) = \{y^N | \hat{\theta}(y^N) \in E_N(\theta)\}. \tag{B.1}$$

Define $p_N(\theta) = P(Y_N(\theta))$, then

$$P(Y_N(\theta)) \equiv P(|\hat{\theta}(y^N) - \theta| \leq c/\sqrt{N}). \tag{B.2}$$

Because of the central limit theorem c can be picked large enough such that

$$P(Y_N(\theta)) \geq 1 - \delta(c) \tag{B.3}$$

where $\delta(c)$ is some function of c which goes to zero as c grows.

The distribution $Q(y^N)$ is defined by $L(y^N)$

$$Q(y^N) = \frac{2^{-L(y^N)}}{\sum_{y^N \in Y^N} 2^{-L(y^N)}}, \tag{B.4}$$

then

$$E(L) - H(\theta) = \sum p(y^N) \log \frac{p(y^N)}{q(y^N)} - \log \sum 2^{-L(y^N)}. \tag{B.5}$$

where $q(y^N) = 2^{-L(y^N)}$. The Krafts inequality results $-\log \sum 2^{-L(Y^N)} \geq 0$, therefore

$$E(L) - H(\theta) \quad \geq \quad \sum p(y^N) \frac{\log p(y^N)}{\log q(y^N)}$$

$$\geq \sum_{Y_N} p(y^N) \log \frac{p(y^N)}{q(y^N)} + \sum_{Y_N^c} p(y^N) \log \frac{p(y^N)}{q(y^N)} \tag{B.6}$$

The main goal of the proof is to find a lower bound for the right side of the above inequality. The lower bound is attained in the following three steps.

Step1 In first step, a lower bound for the first part of the right hand side of inequality (B.6) is obtained. We show that

$$\sum_{Y_N} p(y^N) \log \frac{p(y^N)}{q(y^N)} \geq p_N(\theta) \log \frac{p_N(\theta)}{q_N(\theta)} \tag{B.7}$$

where $q_N(\theta) = \sum_{Y_N} q(y^N)$

To prove the above inequality it is enough to show that the inequality holds for the following case. Assume p_1, p_2, q_1, q_2, numbers between zero and one such that $p_1 + p_2 \leq 1$, $q_1 + q_2 \leq 1$ then

$$p_1 \frac{\log p_1}{\log q_1} + p_2 \frac{\log p_2}{\log q_2} \geq (p_1 + p_2) \log \frac{p_1 + p_2}{q_1 + q_2}.$$

Since Kullback distance of the two distributions $(p_1', p_2') = (\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$ and $(q_1', q_2') = (\frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2})$ is always positive

$$p_1' \log \frac{p_1'}{q_1'} + p_2' \log \frac{p_2'}{q_2'} \geq 0$$

which results the desired inequality. Therefore by induction (B.7) is proved.

Step 2 In this step a lower bound for the right side of inequality in (B.7) is obtained. We replace the function $N^{m/2}$, in Risannen's proof, with a function with properties sufficient for the proof in this step, $\beta(N)$, to generalize the proof. Define the set $D_\epsilon(N)$ as following

$$D_\epsilon(N) = \{\theta | T_N(\theta) = p(\theta) \log \frac{p(\theta)}{q(\theta)} < (1 - \epsilon) \log \beta(N)\} \tag{B.8}$$

where $\beta(N)$ satisfies the following conditions:
1- $\lim_{N\to\infty} \beta(N) = \infty$
2- $\lim_{N\to\infty} \beta(N)^{\alpha(N)} N^{-m/2} = 0$ and $\alpha(N)$ is defined as following

$$\alpha(N) = 1 - \frac{\epsilon/2}{1 - \epsilon/2} - \frac{\log(1 - \epsilon/2)}{\log \beta(N)}, \tag{B.9}$$

For $0 \leq \epsilon < 1$. We chose $\beta(N)$ with such sufficient properties so that we can show that the Lebesgue measure of $D_\epsilon(N)$ goes to zero as N grows and therefore for almost all θ, $T_N(\theta)$ is lower bounded by $(1 - \epsilon) \log \beta(N)$. However the main proof in [42] uses

$\beta(N) = N^{m/2}$. We note that the second condition with given $\alpha(N)$ can be simplified to

$$\lim_{N\to\infty} \frac{1}{\sqrt{N}^m(1-\epsilon/2)}\beta(N)^{(1-\frac{\epsilon/2}{1-\epsilon/2})} = 0 \tag{B.10}$$

For elements of $D_\epsilon(N)$ we have

$$p_N(\theta)\log\frac{p_N(\theta)}{q_N(\theta)} \leq (1-\epsilon)\log\beta(N), \tag{B.11}$$

therefore,

$$-\log q(\theta) < [\frac{1-\epsilon}{p(\theta)} - \frac{\log p(\theta)}{\log\beta(N)}]\log\beta(N). \tag{B.12}$$

The assumption is that $0 \leq \epsilon < 1$ and we pick $\delta(c) \leq \epsilon/2$ then $p(\theta) \geq 1-\epsilon/2$ and

$$-\log q(\theta) < [1 - \frac{\epsilon/2}{1-\epsilon/2} - \frac{\log(1-\epsilon/2)}{\log\beta(N)}]\log\beta(N). \tag{B.13}$$

Therefore

$$q(\theta) > \beta(N)^{-\alpha(N)}. \tag{B.14}$$

Let $B(N)$ be the smallest set of the centers of the hypercubes with volume $(c/\sqrt{N})^m$ which cover $D_\epsilon(N)$. If v_N is the number of elements of $B(N)$ then V_N the volume of $D_\epsilon(N)$ is bounded by

$$V_N \leq v_N(c/\sqrt{N})^m. \tag{B.15}$$

Since $q(\theta)$ is the probability assigned for $Y_N(\theta)$ defined in (B.1) and sets $B(N)$ are disjoint, we have

$$1 \geq \sum_{B(N)} q(\theta) \geq v_N\beta(N)^{-\alpha(N)} \tag{B.16}$$

which gives an upperbound for v_N and from (B.15) we have

$$V_N \leq c^k\beta(N)^{\alpha(N)}N^{-m/2}. \tag{B.17}$$

Using the first condition on $\beta(N)$, for all ϵ, there exist N_ϵ such that for all $N \geq N_\epsilon$, α is bounded $0 < \alpha(N) < 1$. Using the second condition on $\beta(N)$, as N grows this upperbound goes to zero and $D_\epsilon(N)$ has almost zero measure.

One candidate for $\beta(N)$ is $N^{m/2}$ which makes the upperbound to be $N^{(\alpha-1)m/2}$ and therefore

$$V_N \leq c^m N^{(\alpha-m)/2}. \tag{B.18}$$

The inequality (B.11), then takes the following form

$$p_N(\theta) \log \frac{p_N(\theta)}{q_N(\theta)} \leq (1-\epsilon)\frac{m}{2} \log N. \tag{B.19}$$

Therefore for almost all θ, except a zero-measure set, the inverse of the above inequality holds which is the upperbound Rissanen uses in the proof.

Step 3 In previous step a lower bound for one component of the right hand side of the inequality in (B.6) is provided (in (B.7)). Therefore for all θ except a zero-measure set, for the inequality we have

$$E(L) - H(\theta) \quad \geq \quad (1-\epsilon) \log N^{m/2} + \sum_{Y_N^c} p(y^N) \log \frac{p(y^N)}{q(y^N)} \tag{B.20}$$

The lower bound for the second part is calculated as follows. For all positive z, $\log_{10} z \geq 1 - 1/z$, therefore

$$\sum_{Y_N^c} p(y^N) \log \frac{p(y^N)}{q(y^N)} \geq \log 10 \sum_{Y_N^c} p(y^n)(1 - \frac{q(y^N)}{p(y^N)}) \tag{B.21}$$

where $z = \frac{p(y^N)}{q(y^N)}$, hence

$$\sum_{Y_N^c} p(y^N) \log \frac{p(y^N)}{q(y^N)} \geq \log 10(1 - p(\theta) - (1 - q(\theta)) \geq \log 10\{q(\theta) - p(\theta)\}) \geq -\log 10 \tag{B.22}$$

Therefore from (B.20), (B.22) we have

$$E(L) - H(\theta) \geq (1-\epsilon) \log \beta(N) - \log 10 \tag{B.23}$$

The above is a corrected version of Rissanen's proof in his book [42]. The inequality $\log z \geq 1 - 1/z$ is not true if a log is based 2, as Rissanen assumes. The proof in the book misses log 10 in the expression for the lower bound.

Appendix C

Output Error and Impulse Response Error

C.1 Impulse Response Error

From (4.6), we have

$$\begin{aligned} E(||\hat{h}_m^N - h_m^N||^2) &= E(w^T C_m w) + E((B_m \Delta_m^N)^T C_m B_m) + E(2w^T C_m B_m \Delta_m^N) \\ &= tr(C_m)\sigma_w^2 + (B_m \Delta_m^N)^T C_m B_m \Delta_m^N. \end{aligned} \tag{C.1}$$

Similarly from (4.7)

$$\begin{aligned} E(||\hat{h}_m^N - h^N||^2) &= E(||\hat{h}_m^N - h_m^N)||^2) + ||\Delta_m^N||^2 \\ &= tr(C_m)\sigma_w^2 + (B_m \Delta_m^N)^T C_m B_m \Delta_m^N + ||\Delta_m^N||^2, \end{aligned} \tag{C.2}$$

and the variances are

$$\begin{aligned} \text{var}(||\hat{h}_m^N - h_m^N||^2) &= \text{var}((B_m \Delta_m^N)^T C_m B_m \Delta_m^N + w^T C_m w + 2w^T C_m B_m \Delta_m^N) & \text{(C.3)} \\ &= \text{var}(w^T C_m w + 2w^T C_m B_m \Delta_m^N) & \text{(C.4)} \\ &= \text{var}(w^T C_m w) + 4(B_m \Delta_m^N)^T C_m^2 B_m \Delta_m^N \sigma_w^2, & \text{(C.5)} \end{aligned}$$

Note that from (C.3) to (C.4) the variance of the deterministic part, $(B_m \Delta_m^N)^T C_m B_m \Delta_m^N$, is zero and $E(w^T C_m w)(2w^T C_m B_m \Delta_m^N) = 0$ since the additive noise is white.

$$\begin{aligned} \text{var}(||\hat{h}_m^N - h^N||^2) &= \text{var}(||\hat{h}_m^N - h_m^N)||^2 + ||\Delta_m^N||^2) = \text{var}(||\hat{h}_m^N - h_m^N)||^2) & \text{(C.6)} \\ &= \text{var}(w^T C_m w) + 4(B_m \Delta_m^N)^T C_m^2 B_m \Delta_m^N \sigma_w^2 & \text{(C.7)} \end{aligned}$$

C.2 Output error

A Chi-square distribution of order n is of form $x = \sum_{i=1}^{n} (v_i)^2$ where v_is are independent Gaussian random variables. If each v_i has a mean of m_i and variance of σ_i

then

$$E(X) = \sum_{i=1}^{n}(\sigma_i^2 + m_i^2) \quad \text{var}(X) = \sum_{i=1}^{n}(2\sigma_i^4 + 4\sigma_i^2 m_i^2). \tag{C.8}$$

The output error, X_m in (4.13), has a Chi-square distribution. Since G_m is a projection matrix of rank $N - m$ we have

$$\frac{1}{N}(w + B\Delta_m)^T G_m (w + B\Delta_m) = \frac{1}{N}\sum_{i=1}^{N-m}(v_i + m_i)^2 \tag{C.9}$$

where v_is are zero mean, independent Gaussian random variables with variance σ_w^2 and

$$\frac{1}{N}\sum_{1}^{N-m} m_i^2 = \frac{1}{N}(B\Delta_m)^T G_m (B\Delta_m) = g_m. \tag{C.10}$$

The expected value and variance of such random variable is

$$E(X_m) = (1 - \frac{m}{N})\sigma_w^2 + g_m \tag{C.11}$$

$$\text{var}(X_m) = = (1 - \frac{m}{N})\frac{2\sigma_w^4}{N} + \frac{4\sigma_w^2}{N} g_m \tag{C.12}$$

By using the Tchebyshev's Central Limit Theorem, the output error asymptotically behaves like a random variable with Gaussian distribution.

Tchebyshev's Central Limit Theorem : [36] Let $X_1, X_2, \cdots$ have zero mean and finite moments of all orders, and let variance of each X_i to be σ_i^2. Further suppose that

(i) $\lim_{n\to\infty}(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)/n$ exists and is finite,

(ii) $|E(X_i^r)| < A_r < \infty$, $r = 2, 3, \cdots$

Then

$$\lim_{n\to\infty} \Pr\{(a \le S_n/(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)^{1/2} \le b\} = \Phi(b) - \Phi(a) \tag{C.13}$$

where $\Phi(x)$ is the Gaussian cumulative distribution function(cdf). $\diamondsuit$

Next theorem is on the rate of convergence of a distribution to a Gaussian distribution.

The Berry-Esseen Theorem [36]: Let X_i be an IID sequence of random variables such that $E(X_i) = \mu_i$, $Var(X_i) = \sigma_i^2$ and has a finite absolute third moment $v_i = E(|X_i - E(X_i)|^3)$. Then for some positive constant C, $.41 < C < .8$ and G_n the cdf of $(X_1 + X_2 + \cdots + X_n - \sum \mu_i)/(\sqrt{n}\sigma)$. Further let $\rho_n = \sum v_i/(\sum \sigma_i^2)^{3/2}$. Then for all x, we have

$$\sqrt{n}|G_n(x) - \Phi(x)| \le C\rho_n . \diamondsuit \tag{C.14}$$

Therefore, with a rate of convergence of order $\frac{1}{\sqrt{N}}$ the cdf of the output error approaches the cdf of a Gaussian distribution.

C.3 Proof of Lemma 1

Define $\bar{x}_{S_m} = x_{S_m} - (1 - \frac{m}{N})\sigma_w^2$, we want to validate $m_\delta = g_m$ for which

$$m_\delta - \alpha\sqrt{\frac{4m_w}{N}m_\delta + v_m} \leq \bar{x}_{S_m} \leq m_\delta + \alpha\sqrt{\frac{4m_w}{N}m_\delta + v_m}$$

where $m_w = (1 - \frac{m}{N})\sigma_w^2$, and $v_m = \frac{2}{N}(1 - \frac{m}{N})\sigma_w^4$.

Lower Bound on m_δ

$$\bar{x}_{S_m} - m_\delta < \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \tag{C.15}$$

If $\bar{x}_{S_m} \leq \alpha\sqrt{v_m}$, then the inequality holds for $m_\delta > 0$.

If $\bar{x}_{S_m} \geq \alpha\sqrt{v_m}$, then the lower bound for m_δ is the smallest root of the following equation

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2(\frac{4\sigma_w^2}{N}m_\delta + v_m) \tag{C.16}$$

which is

$$L_{m_\delta} = x_{S_m} - m_w + \frac{2\sigma_w^2\alpha^2}{N} - \frac{2\alpha\sigma_w^2}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w} \tag{C.17}$$

Note that $L_{S_m} \leq \bar{x}_{S_m}$.

Upper Bound on m_δ

$$m_\delta - \bar{x}_{S_m} > \alpha\sqrt{\frac{4\sigma_w^2}{N}m_\delta + v_m} \tag{C.18}$$

If $\bar{x}_{S_m} \leq -\alpha\sqrt{v_m}$, then the inequality does not hold for any m_δ.

If $\bar{x}_{S_m} \geq -\alpha\sqrt{v_m}$, then the upper bound is the largest root of equation

$$(\bar{x}_{S_m} - m_\delta)^2 = \alpha^2(\frac{4\sigma_w^2}{N}m_\delta + v_m) \tag{C.19}$$

which is

$$U_{m_\delta} = x_{S_m} - m_w + \frac{2\alpha^2\sigma_w^2}{N} + \frac{2\alpha\sigma_w^2}{\sqrt{N}}\sqrt{\frac{\alpha^2\sigma_w^2}{N} + x_{S_m} - \frac{1}{2}m_w}.$$

C.4 White Noise

If w is a random vector for which each element is a white noise with zero mean and variance σ_w^2, then for any $N \times N$ matrix F

$$\begin{aligned} E_w(w^T F w) &= \mathrm{tr}(F)\sigma_w^2 \\ \mathrm{Var}(w^T F w) &= [E(w^4) - (\sigma_w^2)^2] \sum_{i=1}^{N} f_{ii}^2 + (\sigma_w^2)^2 \sum_{i,j,i\neq j}^{N} f_{ij}^2 \end{aligned} \tag{C.20}$$

For the impulse response error in each space S_m, $F = C_m$ which is defined in (4.8).

C.4.1 Output Error

Since G_m is a projection matrix of rank $N - m$, we have

$$\frac{1}{N} w^T G_m w = \frac{1}{N} \sum_{i=1}^{N-m} (v_i)^2 \tag{C.21}$$

where v_is are zero mean, independent Gaussian random variables with variance σ_w^2. Therefore, the expected value and variance of this random variable, with a Chi-square distribution, is

$$E(\frac{1}{N} w^T G_m w) = (1 - \frac{m}{N})\sigma_w^2 \tag{C.22}$$

$$\mathrm{var}(\frac{1}{N} w^T G_m w) = (1 - \frac{m}{N})\frac{2\sigma_w^4}{N} \tag{C.23}$$

C.4.2 Bounds on the Expected Value and Variance of $w^T C_m w$

Elements of C_m are

$$c_{ij} = \frac{1}{N^2} v_i^T (\frac{A_m^T A_m}{N})^{-1} (\frac{A_m^T A_m}{N})^{-1} v_j \tag{C.24}$$

where v_i is defined as rows of A_m

$$A_m = \begin{bmatrix} v_1^T \\ \vdots \\ v_N^T \end{bmatrix}. \tag{C.25}$$

where A_m is a $N \times m$ matrix defined in (4.4). Matrix $(\frac{A_m^T A_m}{N})^{-1}$ is $m \times m$ and since the input is quasi-stationary there exist finite numbers c_{min} and c_{max} such that

$$c_{min} \leq v_i^T (\frac{A_m^T A_m}{N})^{-1} (\frac{A_m^T A_m}{N})^{-1} v_j \leq c_{max}. \tag{C.26}$$

Therefore

$$0 < \frac{c_{min}}{N^2} < c_{ii} < \frac{c_{max}}{N^2} \tag{C.27}$$

and $|c_{ij}| < \frac{cc}{N^2}$ and we have

$$\begin{aligned} 0 < \frac{c_{min}}{N} &< E(w^T C_m w) = \sum_{i=1}^{N} c_{ii} < \frac{c_{max}}{N} \qquad \text{(C.28)} \\ \text{var}(w^T C_m w) &= [E(w^4) - (\sigma_w^2)^2] \sum c_{ii}^2 + (\sigma_w^2)^2 \sum c_{ij}^2 \\ &\leq \frac{cc}{N^2}(\frac{[E(w^4) - (\sigma_w^2)^2]}{N^2} + (\sigma_w^2)^2) \end{aligned}$$

For the cross term $w^T C_m B_m \Delta_m^N$ we have

$$E(2w^T C_m B_m \Delta_m^N) = 0 \tag{C.29}$$

$$\text{var}(2w^T C_m B_m \Delta_m^N) = 4\sigma_w^2 (\Delta_m^N)^T B_m^T C_m^T C_m B_m \Delta_m^N \tag{C.30}$$

We used bounds on absolute value of element of matrix C_m to provide the bounds. With a similar method we can show that the absolute value of each element of matrix $C_m^T C_m = (C_m)^2$ is less than or equal $\frac{cc}{N^3}$ and we have

$$\text{var}(2w^T C_m B_m \Delta_m^N) = 4\sigma_w^2 (\Delta_m^N)^T B_m^T C_m^T C_m B_m \Delta_m^N \leq 4\sigma_w^2 \frac{cc}{N} || \frac{B_m \Delta_m^N}{N} ||^2 \tag{C.31}$$

Since the input is a bounded power signal and the system has a bounded power/power gain, then $|| \frac{B_m \Delta_m^N}{N} ||^2$ which is less than the power of output of the system is a finite number.

Appendix D

Independent Identically Distributed (IID) Input

The following lemma is used in the proves of this section.

Lemma C1

With probability greater than $[Q(\frac{1}{N^{\mu-.5}})]^{m^2}$ and for any μ, $0 < \mu < .5$ the following equality holds

$$v^T(\frac{A_m^T A_m}{N})^{-1}v = ||v||^2(1 + \frac{c}{N^\mu}), \tag{D.1}$$

where v is a vector of length m and c is a finite number.

Proof of Lemma C1

If matrix P is an $m \times m$ matrix(m finite) and $|P_{ii} - 1| \leq K$, $|P_{ij}| \leq K$, then

$$||v||^2[1 - (2m+1)K] \leq v^T P v \leq ||v||^2[1 + (2m+1)K]. \tag{D.2}$$

We show that with probability greater than $[Q(\frac{1}{N^{\mu-.5}})]^{m^2}$, $\mu < .5$, $[\frac{1}{N}(A^T A)]^{-1}$ has the properties of P with $K = \frac{c}{N^\mu}$. In the next lemma we first show some properties of the inverse of that matrix using the properties of IID input.

Lemma C1.1 With probability greater than $[Q(\frac{1}{N^{\mu-.5}})]^{m^2}$, $\mu < .5$, for matrix $R_m = \frac{1}{N}A_m^T A_m$ the following holds

$$|R_m(i,i) - 1| \leq \frac{1}{N^\mu}, \quad |R_m(i,j)| \leq \frac{1}{N^\mu}, \quad j \neq i. \tag{D.3}$$

Proof of Lemma C1.1 The elements of R_m are in the form of

$$R_m(i,j) = \frac{1}{N}\sum_{k=1}^{N-\max(i,j)+1} u_1 u_{1+|i-j|}. \tag{D.4}$$

Each of these elements are random variables with

$$E(R(i,j)R(k,l)) = 0, i \neq k \text{ or } j \neq l. \tag{D.5}$$

If $m << N$ by using the law of large number each of these elements are asymptotically a Gaussian random variable, which are independent from each other. The expected value and variance of these random variables are

$$E(R_m(i,i)) = \frac{N-i+1}{N} \approx 1 \quad , \quad var(R_m(i,i)) = var(u^2)\frac{N-i}{N^2} \leq \frac{var(u^2)}{N}$$
$$E(R_m(i,j)) = 0 \quad , \quad var(R_m(i,j)) = \frac{N-\max(i,j)}{N^2} \leq \frac{1}{N} \quad j \neq i \tag{D.6}$$

Variance of the elements of input squared, u^2, is a finite number and for simplicity we assume it is one. By using the Chebychev inequality we have

$$\begin{aligned} \text{Prob}\left(|R_m(i,i)-1| \leq \frac{1}{N^\mu}\right) &\geq Q\left(\frac{1}{N^{\mu-.5}}\right) \\ \text{Prob}\left(|R_m(i,j))| \leq \frac{1}{N^\mu}\right) &\geq Q\left(\frac{1}{N^{\mu-.5}}\right), \quad i \neq j. \end{aligned} \tag{D.7}$$

Since these events are independent, the probability that the inequality holds for all the elements of R_m is the product of the probability of each event. The $m \times m$ matrix $\frac{A^T A}{N}$ has m^2 elements and the proof of *Lemma C1.1* is completed. $\diamondsuit$

Lemma C1.2 Next we claim that if $m \times m$ matrix P^{-1} is such that $|P^{-1}(i,j) - I(i,j)| \leq \epsilon$, where I is the identity matrix, then for the elements of the inverse of P^{-1}, P, we have

$$|P(i,i)-1| \leq \epsilon f_1(\epsilon,m), \quad |P(i,j)| \leq \epsilon f_2(\epsilon,m), \tag{D.8}$$

where $f_1(\epsilon,m)$, $f_2(\epsilon,m)$ are finite functions of ϵ.

Proof of Lemma C1.2 For each element of P

$$P(i,j) = \frac{\text{Det}P_{ij}^{-1}}{\text{Det}P^{-1}}, \tag{D.9}$$

where matrix P_{ij}^{-1} is matrix P^{-1} without ith column and jth row. Therefore

$$P(i,j) = \frac{\text{Det}P_{ij}^{-1}}{\text{Det}P^{-1}}. \tag{D.10}$$

If $|P^{-1}(i,j) - I(i,j)| \leq \epsilon$, then

$$\begin{aligned} (1-\epsilon)^m - m!\epsilon(1+\epsilon)^{m-1} &\leq \text{Det}P^{-1} \leq (1+\epsilon)^m + m!\epsilon(1+\epsilon)^{m-1} \\ P_{ij}^{-1} &\leq \epsilon((1+\epsilon)^{m-2} + m!(1+\epsilon)^{m-2}. \end{aligned} \tag{D.11}$$

Therefore

$$|P(i,i)-1| \leq \epsilon f_1(\epsilon,m), \quad |P(i,j)| \leq \epsilon f_2(\epsilon,m), \tag{D.12}$$

where $f_1(\epsilon,m)$, $f_2(\epsilon,m)$ are finite functions of ϵ. $\diamondsuit$

Using *Lemma C1.1* and *Lemma C1.2*, the proof of *Lemma C1* is completed. $\diamondsuit$

D.1 Output Error

From the first step we obtain bounds for $g_m = \frac{1}{N}(B_m\Delta_m^N)^T G_m B_m \Delta_m^N$, the unmodeled part of the output error. Since the input is IID, g_m itself is a random variable and we have

$$\begin{aligned} E(g_m) &= E\frac{1}{N}(B_m\Delta_m^N)^T G_m B_m \Delta_m^N \\ &= E\frac{1}{N}(B_m\Delta_m^N)^T B_m \Delta_m^N - E\frac{1}{N^2}(A_m^T B_m \Delta_m^N)^T (\frac{A_m^T A_m}{N})^{-1} A_m^T B_m \Delta_m^N \end{aligned} \quad \text{(D.13)}$$

Next we find the estimate of the two components of $E(g_m)$ in (D.13).

D.1.1 Estimation of $E\frac{1}{N}(B_m\Delta_m^N)^T B_m\Delta_m^N$

$$E\frac{1}{N}(B_m\Delta_m^N)^T B_m\Delta_m^N = E\frac{1}{N}(b_m^T\Delta_\delta^T\Delta_\delta b_m), \quad \text{(D.14)}$$

where

$$b_m = \begin{bmatrix} u_1 \\ \vdots \\ u_N \end{bmatrix}, \quad \Delta_\delta = \begin{bmatrix} h_{m+1} & 0 & \cdots & 0 \\ h_{m+2} & h_{m+1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h_N & h_{N-1} & \cdots & h_{m+1} \end{bmatrix}, \quad \text{(D.15)}$$

and b_m is a vector with IID elements. By using (2.80)

$$E\frac{1}{N}(B_m\Delta_m^N)^T B_m\Delta_m^N = E\frac{1}{N}(b_m^T\Delta_\delta^T\Delta_\delta b_m) = \frac{1}{N}Trace(\Delta_\delta^T\Delta_\delta) = g_m^N, \quad \text{(D.16)}$$

where

$$g_m^N = \frac{N-m}{N}h_{m+1}^2 + \frac{N-m+1}{N}h_{m+2}^2 + \cdots + \frac{1}{N}h_N^2 = \sum_{i=m+1}^{N}\frac{N-i+1}{N}|h_i|^2. \quad \text{(D.17)}$$

D.1.2 Estimation of $E\frac{1}{N^2}(A_m^T B_m\Delta_m^N)^T(\frac{A_m^T A_m}{N})^{-1}A_m^T B_m\Delta_m^N$

By using *Lemma C1*, (D.1), we have

$$\frac{1}{N^2}(A_m^T B_m\Delta_m^N)^T(\frac{A_m^T A_m}{N})^{-1}A_m^T B_m\Delta_m^N = \frac{1}{N^2}(A_m^T B_m\Delta_m^N)^T A_m^T B_m\Delta_m^N(1+O(1)) \quad \text{(D.18)}$$

where

$$\frac{1}{N}A_m^T B_m \Delta_m^N = \begin{bmatrix} r_{u1}(m+1) & r_{u2}(m+2) & \cdots & r_{uN}(N) \\ r_{u1}(m) & r_{u2}(m+1) & \cdots & r_{uN}(N-1) \\ \vdots & \ddots & \ddots & \vdots \\ r_{u1}(1) & r_{u2}(2) & \cdots & r_{uN}(N-m) \end{bmatrix} \begin{bmatrix} h_{m+1} \\ \vdots \\ h_N \end{bmatrix}$$

$$= \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}, \tag{D.19}$$

and

$$r_{ui}(j) = \frac{1}{N} \sum_{k=1}^{N-m-1+i} u(k)u(k+j). \tag{D.20}$$

For each v_i in (D.19) we have

$$E(v_i^2) = \frac{1}{N}(\frac{N-m}{N}h_{m+1}^2 + \frac{N-m+1}{N}h_{m+2}^2 + \cdots + \frac{1}{N}h_N^2) = \frac{1}{N}g_m^N. \tag{D.21}$$

Therefore,

$$E\frac{1}{N^2}(A_m^T B_m \Delta_m^N)^T A_m^T B_m \Delta_m^N = \sum_{i=1}^{m} E(v_i)^2$$
$$= \frac{m}{N}g_m^N, \tag{D.22}$$

where g_m^N is defined in (D.17).

From (D.16),(D.22), for $E(g_m)$ we have

$$E(g_m) = E\frac{1}{N}(B_m \Delta_m^N)^T G_m B_m \Delta_m^N = (1 - \frac{m}{N})g_m^N + O(\frac{1}{N}), \tag{D.23}$$

where $O(\frac{1}{N})$ is the result of using *lemma C1* (D.1).

D.1.3 Variance of g_m

The variance of g_m is

$$\text{var}\frac{1}{N}(B_m \Delta_m^N)^T G_m B_m \Delta_m^N = E((\frac{1}{N}(B_m \Delta_m^N)^T G_m B_m \Delta_m^N)^2) \tag{D.24}$$
$$-(E(\frac{1}{N}(B_m \Delta_m^N)^T G_m B_m \Delta_m^N))^2$$
$$= \text{var}\frac{1}{N}(||B_m \Delta_m^N||^2) + \text{var}\frac{1}{N^2}(||A_m^T B_m \Delta_m^N||^2)$$
$$+E(\frac{1}{N}(||B_m \Delta_m^N||^2))E(\frac{1}{N^2}(||A_m^T B_m \Delta_m^N||^2))$$

$$-E(\frac{1}{N}(||B_m\Delta_m^N||^2\frac{1}{N^2}(||A_m^TB_m\Delta_m^N||^2)). \quad \text{(D.25)}$$

Lemma C2

$$\text{var}\frac{1}{N}(B_m\Delta_m^N)^TG_mB_m\Delta_m^N \leq k_m\frac{(g_m^N)^2}{N} + O(\frac{1}{N}) \quad \text{(D.26)}$$

where $k_m = l_m + m$ and l_m is an upper bound for $\frac{\sigma_{max}(\Delta_m^N)}{||g_m^N||_2^2}$

Proof of Lemma C2 We find upper bounds for the four elements of the variance of g_m in (D.25).

Lemma C2.1 First we show that

$$\text{var}\frac{1}{N}(||B_m\Delta_m^N||^2) \quad \leq \quad l_m\frac{(g_m^N)^2}{N} \quad \text{(D.27)}$$

where l_m is an upper bound of $\frac{\sigma_{max}(\Delta_m^N)}{||g_m^N||_2^2}$

Proof of Lemma C2.1 By using (2.80) with $z = \Delta_\delta^T\Delta_\delta$ we have

$$\text{var}\frac{1}{N}(||B_m\Delta_m^N||^2) = \frac{1}{N^2}([E(u^4) - (\sigma_u^2)^2]\sum_i^N z_{ii}^2 + (\sigma_u^2)^2\sum_{i,j,i\neq j}^N z_{ij}^2) \quad \text{(D.28)}$$

For each fixed i,

$$[E(u^4) - (\sigma_u^2)^2]z_{ii}^2 + (\sigma_u^2)^2\sum_{j\neq i}^N z_{ij}^2 \leq \alpha_z(\sigma_u^2)^2\sum_j^N z_{ij}^2. \quad \text{(D.29)}$$

where α_z is a function of the random variable u. When u is a Bernoulli input of unit variance, then $[E(u^4) - (\sigma_u^2)^2] = 0$ and $\alpha_z = 0$. For a fixed i then $\sum_j^N z_{ij}$ is the l_2 norm of some elements of the correlation of the ith column of Δ_δ with Δ_m^N. If k_i is the correlation of $f_i = \Delta_\delta(:,i)$ and Δ_m^N then

$$\sum_j^{N-m} z_{ij}^2 \quad \leq \quad \sum_n |k_i(n)|^2 \quad \text{(D.30)}$$

$$\leq \quad \sigma_{max}(\Delta_m^N)\sum_{n=1}^{N-i-m+1}|h_{m+n}|^2 \quad \text{(D.31)}$$

where σ_{max} is the H_∞ norm of Δ_m^N. From (D.28) and (D.30) we conclude

$$\text{var}\frac{1}{N}(||B_m\Delta_m^N||^2) \quad \leq \quad \frac{1}{N^2}\sigma_{max}(\Delta_m^N)\sum_{i=1}^{N-m}(\sum_j z_{ij}^2) \quad \text{(D.32)}$$

$$\leq \quad \frac{1}{N}\sigma_{max}(\Delta_m^N)g_m^N \quad \text{(D.33)}$$

Since Δ_m^n has a bounded σ_{max} there exists a bounded number l_m such that $\sigma_{max}(\Delta_m^N) \leq l_m g_m^N$, hence

$$\mathrm{var}\frac{1}{N}(||B_m\Delta_m^N||^2) \leq \frac{l_m}{N}(g_m^N)^2, \tag{D.34}$$

$\diamondsuit$

Lemma C2.2

$$\mathrm{var}\frac{1}{N^2}(||A_m^T B_m\Delta_m^N||^2) \leq \frac{m^2(g_m^N)^2}{N^2}, \tag{D.35}$$

Proof of Lemma C2.2 By using v_i from (D.19) we have

$$v_i = \sum_{j=1}^{N-m} r_{ui}(j)h_{j+m}. \tag{D.36}$$

Also by implementing the law of large numbers, each $r_{ui}(j)h_{j+m} = d_j$ is an independent Gaussian random variable with zero mean and variance of $|h_{j+m}|^2\frac{N-j}{N^2}$. Therefore v_i is also a Gaussian random variable with variance $\sum \mathrm{var}d_j = \frac{1}{N}g_m^N$, variance of v_i^2 then is

$$\begin{aligned}\mathrm{var}v_i^2 &= E(v_i^4) - (E(v_i^2))^2 = 3(\mathrm{var}v_i)^2 - (\mathrm{var}v_i)^2 \\ &= 2\frac{(g_m^N)^2}{N^2}.\end{aligned} \tag{D.37}$$

Combining (D.19) and (D.37)

$$\begin{aligned}\mathrm{var}\frac{1}{N^2}(||A_m^T B_m\Delta_m^N||^2) &= \mathrm{var}(\sum_{i=1}^m v_i^2) = \sum_{i=1}^m \mathrm{var}v_i^2 + \sum_{i\neq j} E(v_i^2v_j^2) - E(v_i^2)E(v_j^2) \\ &\leq \sum_{i=1}^m \mathrm{var}v_i^2 + \sum_{i\neq j}\sqrt{\mathrm{var}v_i^2\mathrm{var}v_j^2} \\ &\leq m^2\mathrm{var}v_i^2 \leq \frac{m^2(g_m^N)^2}{N^2}\end{aligned} \tag{D.38}$$

$\diamondsuit$

For the third element in (D.25) we use (D.16), (D.22)

$$E(\frac{1}{N}(||B_m\Delta_m^N||^2))E(\frac{1}{N^2}(||A_m^T B_m\Delta_m^N||^2)) = g_m^N \times \frac{m}{N}g_m^N = \frac{(mg_m^N)^2}{N} \tag{D.39}$$

And for the last element in (D.25), we use the upper bounds on the variances in *Lemmas C2.1, C2.2* (D.27), (D.35)

$$E(\frac{1}{N}||B_m\Delta_m^N||^2\frac{1}{N^2}||A_m^T B_m\Delta_m^N||^2) \leq \sqrt{\mathrm{var}(\frac{1}{N}||B_m\Delta_m^N||^2)\mathrm{var}\frac{1}{N^2}(||A_m^T B_m\Delta_m^N||^2)}$$

$$\leq \sqrt{\frac{l_m(g_m^N)^2}{N}\frac{m^2(g_m^N)^2}{N^2}} \tag{D.40}$$

from (D.27)(D.35)(D.39),(D.40)

$$\text{var}\frac{1}{N}(B_m\Delta_m^N)^T G_m B_m \Delta_m^N \leq k_m \frac{(g_m^N)^2}{N} + O(\frac{1}{N}) \tag{D.41}$$

The proof of Lemma C2 is completed $\diamondsuit$

D.2 Estimates of IRE and SIRE

In this section we find estimates of the expected value and variance of IRE and SIRE for IID inputs. From (4.26),(4.27) and(4.28) we have

$$\begin{aligned} E(||\hat{h}_m^N - h_m^N||^2) &= tr(C_m)\sigma_w^2 + (B_m\Delta_m^N)^T C_m B_m \Delta_m^N, & \text{(D.42)} \\ E(||\hat{h}_m^N - h^N||^2) &= tr(C_m)\sigma_w^2 + (B_m\Delta_m^N)^T C_m B_m \Delta_m^N + ||\Delta_m^N||^2, & \text{(D.43)} \end{aligned}$$

and here the input is also a random variable so the variance in (C.3) is changed to

$$\begin{aligned} \text{var} \quad (\quad ||\hat{h}_m^N - h_m^N||^2) &= \text{var}(||\hat{h}_m^N - h^N||^2) & \text{(D.44)} \\ = \quad \text{var}(w^T C_m w) &+ \text{var}((B_m\Delta_m^N)^T C_m B_m \Delta_m^N) + 4E(B_m\Delta_m^N)^T C_m^2 B_m \Delta_m^N \sigma_w^2) & \text{(D.45)} \end{aligned}$$

In section D.2.1 we show that for the unmodeled related parts of the expected values and variance we have

$$\begin{aligned} E((B_m\Delta_m^N)^T C_m B_m \Delta_m^N) &= \frac{m}{N} g_m^N + O(\frac{1}{N}) & \text{(D.46)} \\ \text{var}((B_m\Delta_m^N)^T C_m B_m \Delta_m^N) &\leq \frac{m^2(g_m^N)^2}{N^2} + O(\frac{1}{N^2}), & \text{(D.47)} \\ E(4(B_m\Delta_m^N)^T C_m^2 B_m \Delta_m^N \sigma_w^2) &= 4\sigma_w^2 \frac{m}{N^3} g_m^N + O(\frac{1}{N^3}) & \text{(D.48)} \end{aligned}$$

and in section D.2.2 we show that for the noise related parts we have

$$\begin{aligned} tr(C_m)\sigma_w^2 &= \sigma_w^2 \frac{m}{N} + O(\frac{1}{N}) & \text{(D.49)} \\ \text{var}(w^T C_m w) &= (\sigma_w^2)^2 \frac{2m}{N^2} + O(\frac{1}{N^2}) & \text{(D.50)} \end{aligned}$$

D.2.1 Unmodeled Dynamics Effects

The unmodeled part of the subspace impulse error also is a random variable. Here we have

$$C_m = \frac{1}{N^2} A_m (\frac{A_m^T A_m}{N})^{-1} (\frac{A_m^T A_m}{N})^{-1} A_m^T. \tag{D.51}$$

Using the same argument we had for $(\frac{A^tA}{N})^{-1}$ in Lemma C1, (D.2),(D.1), here for $(\frac{A^tA}{N})^{-1}(\frac{A^tA}{N})^{-1}$ we have

$$v^T(\frac{A_m^TA_m}{N})^{-1}(\frac{A_m^TA_m}{N})^{-1}v = ||v||^2 + O(||v||^2). \tag{D.52}$$

Therefore, estimating the variance and expected value of $(B_m\Delta)^TC_mB_m\Delta$, is similar to estimating the variance and expected value of $\frac{1}{N^2}(B_m\Delta)^TA_mA_m^TB_m\Delta$ which is provided in D.1.2. Hence we have

$$E((B_m\Delta_m^N)^TC_mB_m\Delta_m^N) = \frac{m}{N}g_m^N + O(\frac{1}{N}), \tag{D.53}$$

$$4\sigma_w^2E((B_m\Delta_m^N)^TC_m^2B_m\Delta_m^N) = 4\sigma_w^2\frac{m}{N^3}g_m^N + O(\frac{1}{N^3}) \tag{D.54}$$

With similar argument we can use lemma C2.2 (D.35),

$$\text{var}(B_m\Delta_m^N)^TC_mB_m\Delta_m^N) \leq \frac{m^2(g_m^N)^2}{N^2} + O(\frac{1}{N^2}) \tag{D.55}$$

D.2.2 Noise Components

Here we find estimates for the expected value and the variance of $\frac{1}{N}w^TC_mw$

$\underline{E(\frac{1}{N}w^TC_mw)}$

$$E(\frac{1}{N}w^TC_mw) = E(\frac{w^TA_m}{N}(\frac{A_m^TA_m}{N})^{-1}(\frac{A_m^TA_m}{N})^{-1}\frac{A_m^Tw}{N}) \tag{D.56}$$

$$= E(\frac{w^TA_mA_m^Tw}{N^2}) + O(E((\frac{w^TA_mA_m^Tw}{N^2})) \tag{D.57}$$

where from (D.56) to (D.57) we used (D.52).

Here $\frac{A_m^Tw}{N}$ is a vector of length m,

$$\frac{A_m^Tw}{N} = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}, \tag{D.58}$$

where v_i's are independent Gaussian random variables, $E(||v_i||^2) = \frac{N-i}{N^2}\sigma_w^2$. Assume $m << N$, then

$$E(\frac{w^TA_mA_m^Tw}{N^2}) = \sum_{i=1} mE(||v_i||^2) \approx \frac{m}{N}\sigma_w^2. \tag{D.59}$$

Therefore, with (D.57), (D.59) we conclude

$$E(\frac{1}{N}w^TC_mw) = \frac{1}{N}tr(C_m)\sigma_w^2 = \sigma_w^2\frac{m}{N} + O(\frac{1}{N}) \tag{D.60}$$

$\underline{\text{var}\frac{1}{N}(w^T C_m w)}$

$$\text{var}\frac{1}{N}(w^T C_m w) = E((\frac{1}{N}(w^T C_m w))^2) - (E(\frac{1}{N}(w^T C_m w)))^2 \quad \text{(D.61)}$$

$$\approx \text{var}(\frac{w^T A_m A_m^T w}{N^2}) \quad \text{(D.62)}$$

Note that here we ignored the $O(\cdot)$ terms which are results of using (D.52)

Next we calculate $\text{var}(\frac{w^T A_m A_m^T w}{N^2})$.

If random variable $f = \sum_{i=1}^{k} d_i^2$ where d_is are IID random variables then the random variable has Chi-square distribution with

$$\begin{aligned} E(f) &= k\sigma_d^2 \\ \text{var}f &= k\text{var}(d^2). \end{aligned} \quad \text{(D.63)}$$

If d_i is Gaussian then $E(d^4) = 3(\sigma_d^2)^2$,

$$\text{var}(d^2) = E(d^4) - (E(d^2))^2 = 3(\sigma_d^2)^2 - (\sigma_d^2)^2 = 2(\sigma_d^2)^2. \quad \text{(D.64)}$$

Here $\frac{w^T w}{N} = \frac{1}{N}\sum_1^N w_i^2$ and $\frac{w^T A_m A_m^T w}{N^2} = \sum_1^m v_i^2$, where v_i defined in (D.58). Therefore, from (D.63),(D.64) we have

$$\text{var}(\frac{w^T A_m A_m^T w}{N^2}) = \frac{2m\sigma_w^2}{N^2}, \quad \text{(D.65)}$$

and from (D.62),(D.65) we conclude

$$\text{var}\frac{1}{N}(w^T C_m w) = \frac{2m(\sigma_w^2)^2}{N^2} + O(\frac{1}{N^2}), \quad \text{(D.66)}$$

where $O(\frac{1}{N^2})$ is the result of using (D.52).

Bibliography

[1] H. Akaike. Statistical Predictor Identification. *Ann. Inst. Statist. Math 22:203–217, , 1970.*

[2] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, vol.AC-19, pp.716-723, 1974.

[3] Y. Barron, J. Rissanen, Bin Yu. The Minimum Description Length Principle in Coding and Modeling. *IEEE Trans. on Information theory,* vol 44, pp.2743-2760, Oct. 1998.

[4] S. Beheshti and M.A. Dahleh. On model quality evaluation of stable LTI systems. *Proceedings of the 39th IEEE Conference on Decision and Control,* pp.2716-2721, 2000.

[5] S.P. Boyd R.L Kosut, A.K. lau. Set-Membership Identification of Systems with Parametric and Non-parametric Uncertainty. *IEEE Trans. on Automatic Control,* vol.37, pp.929-941, 1992.

[6] P. M. T. Broersen. Finite Sample Criteria for Autoregressive Order Selection. *IEEE Trans. on Signal Processing,* vol.48, pp.3550-3558, 2000.

[7] W. Chen, K.M.Wong, and J. Reilly Detection of the Number of Signals: A Predicted Eigen-Threshold Approach. *IEEE Trans on Signal processing,* vol.39, pp.1089-1098, 1991

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* New York: Wiley, 1991.

[9] I. Csiszar. *Class notes: Information Theoretical Methods in Statistics.* University of Maryland, College park, MD, Spring, 1990.

[10] D. Davisson. Universal noiseless coding. *IEEE transaction on Information Theory,* pp. 783-795, 1973.

[11] D. Donoho, I. M. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, pp. 425-455, 1994.

[12] D. Donoho. De-noising by Soft Thresholding. *IEEE Trans. on Information Theory,* vol.41, pp. 613-627, 1995.

[13] E. Fishler, M. Grosman and H. Messer. Detection of Signals by Information Theoretic Criteria: General Asymptotic Performance Analysis. *IEEE Trans. on Signal Processing,* vol.50, pp.1027-1036, 2002.

[14] A. Garulli. Tight error bounds for projection algorithms in conditional set membership estimation. *Systems & Control Letters*,vol 37, pp.293-300, 1999.

[15] L. Giarre, B.Z. Kacewicz and M. Milanese. Model quality evaluation in set-membership identification. *Automatica*, vol.33, pp.1133-1139, 1997.

[16] L. Giarre and M. Milanese. Model quality evaluation in H_2 identification. *IEEE Trans. on Automatic Control*, vol.42, pp.691-698, 1997.

[17] L. Giarre and M. Milanese. SM Identification of Approximating Models for H_∞ Robust Control. *International Journal of Robust and Nonlinear Control,* vol.9, pp.319-332, 1999.

[18] L. Giarre, M. Milanese, and M. Taragna. H_∞ Identification and Model Quality Evaluation. *IEEE Trans. on Automatic Control,* vol. 42, pp.188-199, 1997.

[19] A. Garulli, A. Vicino and G. Zappa. Worst-case conditional system identification in a general class of norms. *IEEE Trans. on Automatic Control*, vol.45, pp.14-23, 2000.

[20] G. Goodwin, M Gevers, B. Ninness Quantifying the Error in Estimated Transfer Functions with Application to Model Order Selection. *IEEE Trans. on Automatic Control,* vol.37, No. 7, pp.913-927, 1992.

[21] F. Gustafsson and H. Hjalmarsson. Twenty-one ML Estimator for Model Selection. *Automatica,* vol. 31, pp.1377-1391, 1995.

[22] T.K. Gustafsson P.M. Makila, J.R. Partington. Worst-case Control-relevant Identification. Automatica, vol. 31, No. 12, pp.1799-1819, 1995.

[23] E. Hannan. The Determination of the order of an Auto-regression. *Journal of Royal Statistics Society,* vol.B-41, pp.190-195, 1979.

[24] M. Hansen and B. Yu. Model Selection and The Principle of Minimum Description Length. Technical Memorandum, Bell Labs, Murray Hill, N.J. 1998.

[25] P.S.C. Heuberger, P.M. Van den Hof, and O.H. Bosgra. A generalized orthonormal basis for linear dynamical systems. *IEEE Trans. on Automatic Control*, vol.40, pp.451-465 , 1995.

[26] H. Hjalmarsson and L. Ljung. A unifying view of disturbances in identification. *10th IFAC Symposium on System Identification,* vol. 2, pp.73-78, 1994.

[27] C.A. Jacobson Helmicki, A. and C.N. Nett. Control oriented system identification: a worst case/deterministic approach in H_∞. *IEEE Trans. Automat. Control,* AC-36, pp.1163-1176, 1991.

[28] B. Kacewicz. Worst-case conditional system identification in a general class of norms. *Automatica*, pp.1049-1058, 1999.

[29] H. Krim, D. Tucker, S. Mallat, D. Donoho. On Denoising and Best Signal Representation. *IEEE Trans. on Information Theory,* vol.45, pp. 2225-2238, 1999

[30] A.P. Liavas, P.A. Regalia, and J. Delmas. Blind channel approximation: effective channel order estimation. *IEEE Trans. on Signal Processing,* vol.47, pp.3336-3344, 1999.

[31] A.P. Liavas, P.A. Regalia, and J. Delmas. On the Behavior of Information Theoretic Criteria for Model Order Selection. *IEEE Trans. on Signal Processing,* vol.49, pp.1689-1695, 2001.

[32] L. Ljung. *System Identification: Theory for the user.* NJ: Prentice-Hall, 1987.

[33] L. Ljung B. Wahlber and H. Hjalmarss. Model quality: The role of prior knowledge and data information. *30th CDC IEEE Conference on Decision and Control,* pp.273-278, 1991.

[34] B. Ninness and G.C. Goodwin. Estimation of model quality. *Automatica*, vol.31, pp.1771-1797, 1995.

[35] F. Paganini. Set descriptions of white noise and worst induced norms. *Proceedings of the 32nd Conference on Decision and Control*, pp.3658-3663, December, 1993.

[36] C. B. Read J. K. Patel. *Handbook of the Normal Distribution.* Marcel Dekker, INC. New York, 1996.

[37] M. I. Ribeiro and J. M. F. Moura. LD^2-ARMA Identification Algorithm. *IEEE Trans. on Signal Processing,* vol.39, pp.1822-1834 , 1991.

[38] J. Rissanen. Universal Coding, Information, Prediction, and Estimation. *IEEE Trans. on Information Theory*, vol.IT30, pp. 629-636, 1984.

[39] J. Rissanen. Modeling by Shortest Data description. *Automatica,*14:PP.465–471 1978.

[40] J. Rissanen. A Predictive Least Squares Principle. *IMA Journal of Mathematical Control and Information,*, pp.211–222.

[41] J. Rissanen. Stochastic Complexity and Modeling. *The Annals of Statistics,* pp.1080-1100, 1986.

[42] J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific, New Jersey, second edition 1998.

[43] J. Rissanen. MDL Denoising. *IEEE Trans. on Information Theory,* vol.46, pp. 2537-2543, 2000.

[44] G. Schwarz. Estimating The Dimension of a Model. *The Annals of Statistics,* vol.6, pp.461-464, 1978.

[45] R. Shibata. An Optimal Selection of Regression Variables. *Biometrika,* 68 pp.45-54, 1981.

[46] R. Shibata. Asymptotic Mean Efficiency of a Selection of Regression Variables. *Ann. Inst. Statist. Math,* pp.415 -423 1983.

[47] Y. M. Shtarkov Universal sequential coding of single messages. *Probl. Inform. Transm.,* vol.23, pp. 175-186, 1987.

[48] T.P. Speed and B. Yu. Model Selection and Prediction: Normal Regression. *J. Inst. Statist. Math.* 45, pp.35-54, 1993.

[49] P. Stoica T. Soderstrom. *System Identification.* Prentice-Hall, 1989.

[50] E. Telatar. *Universal Coding: Three Approaches.* Area Exam, 1990.

[51] R. Tempo A. Vicino B.Z. Kacewicz, M. Milanese. Optimality of central and projection algorithms for bounded uncertainty. *Systems & control letters,* No. 8, pp.161-171, 1986.

[52] L. Tong, S. Perreau. Multichannel Blind Identification:From Subspace to Maximum Likelihood Methods. *Proceedings of the IEEE,* vol.86, pp.1951-1968, 1998.

[53] D.N.C. Tse M.A. Dahleh and J.N. Tsitsiklis. Optimal Asymptotic Identification under Bounded Disturbances. *IEEE Trans. on Automatic Control,* vol.38, pp.1176-1190, 1993.

[54] S.R. Venkatesh. *System Identification for Complex Systems.* Thesis MIT, 1997.

[55] S.R. Venkatesh and M.A. Dahleh. Identification in presence of unmodeled dynamics and noise. *IEEE Trans. on Automatic Control,* vol.42, pp.1620 -1635, 1997.

[56] S.R. Venkatesh and M.A. Dahleh. On system identification of complex systems from finite data. *IEEE Trans. on Automatic Control,* vol.46, pp.235-357, 2001.

[57] D.K. de Vries and P.M.J. Van den Hof. Frequency domain identification with generalized orthonormal Basis Functions. *IEEE Trans. on Automatic Control,* vol.43, pp.656-669, 1998.

[58] L.Y. Wang and G.G. Yin. Persistent identification of systems with unmodeled dynamics and exogenous disturbances. *IEEE Trans. on Automatic Control,* vol.45, pp.1246-1256, 2000.

[59] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol.ASSP-33, pp.387-392, 1985.

[60] K.M.Wong, Q.T. Zhang, J. Reilly and P.C. Yip On Information Theoretic Criteria for Determining the Number of Signals in High Resolution Array Processing. *IEEE Trans. Acoust. Speech, Signal processing,* vol.38, pp.1959-1971, 1990

[61] R. Younce and C.E. Rohrs. *Identification with Parametric and Non-parametric Uncertainty.* IEEE Trans. Autom. Control, AC-37, pp.715-728, 1992.

[62] G. Zames. On the metric complexity of causal linear systems: estimates of ϵ-entropy and ϵ-dimension for continuous time. *IEEE Trans. on Automatic Control,* vol.24, pp.222-230, 1979.

[63] A. Zellner. On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distribution. *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti,* 1986.

[64] Q.T. Zhang, K.M. Wong, Information Theoretic Criteria for the Determination of the Number of Signals Spatially Correlated Noise . *IEEE Trans on Signal processing,* vol.39, pp.1652-1663, 1993