UNIVERSITY OF CALIFORNIA SAN DIEGO

Mutual Information Based Evaluation Of Data-set Quality

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Engineering Sciences (Mechanical Engineering)

by

Abhijeet J. Patil

Committee in charge:

Professor Robert Bitmead, Chair
Professor Thomas Bewley
Professor Mauricio De Oliveira

2020

The Thesis of Abhijeet J. Patil is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

For my family and teachers, without their direction I'd be lost.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my advisor, Robert Bitmead, without whose investment in me this thesis would not have been possible. His persistent guidance, patience, and feedback of my work is the foundation of this research. I am also grateful for the academic instruction from my committee members Thomas Bewley and Mauricio De Oliveira.

I would also like to convey my eternal appreciation towards my parents, who supported me and firmly believed in my potential.

Lastly, my endless appreciation to all the people who have contributed to my growth into the person I am today.

VITA

2017	Bachelor of Technology, Indian Institute of Technology Indore (IIT Indore)
2020	Master of Science, University of California San Diego

FIELDS OF STUDY

Major Field: Engineering Sciences (Mechanical Engineering)

Specialization in Dynamic Systems & Control

ABSTRACT OF THE THESIS

Mutual Information Based Evaluation Of Data-set Quality

by

Abhijeet J. Patil

Master of Science in Engineering Sciences (Mechanical Engineering)

University of California San Diego, 2020

Professor Robert Bitmead, Chair

In this study, we explore the possibility of using the information-theoretic concept of mutual information, between the output signal and the regressor of an ARX system model as a criterion to effectively select the most informative data-sets from a collection of experimental records. We derive an expression connecting the mutual information to the signal properties to help with our analysis. We then use the expression to check whether using mutual information as a design criterion to synthesize the input signal leads to any meaningful connections to identifiability. Our findings indicate that mutual information does not serve as a suitable criterion for experiment design.

Chapter 1

Introduction

System identification as a branch of control systems studies the different methods of parameter estimation based on the measurements arriving from the system. These estimated parameters are highly influenced by the nature of the measurement data we work with. Naturally an important aspect of system identification is the generation of measurement data useful for parameter estimation under some predefined criterion. Experiment design deals with the problem of generating an optimal input signal for parameter estimation.

In general an estimator algorithm works by first choosing a meaningful criterion based on the measurement data and then computing the system parameters by either maximizing or minimizing the said criterion. To mention one such criterion, the Least Squares Method, minimizes the mean square error between the measurements and predictions to estimate the parameters.

One way of comparing the results of different estimation algorithms is to compare the variances of the estimated parameters. For an unbiased estimator of a parameter, the Cramér–Rao bound signifies the best performance in terms of the least estimate variance an estimator could achieve. The Cramér–Rao bound states that the lower bound for the covariance of the parameter estimate ($\mathrm{cov}(\hat{\theta}(Y))$) is equal to the inverse of the Fisher

information matrix (M_θ).

$$\text{cov}(\hat{\theta}(Y)) \geq M_\theta^{-1}$$

$$\text{where,} \qquad M_\theta = E_{Y|\theta}\{[\frac{\partial \log p(Y|\theta)}{\partial \theta}]^T[\frac{\partial \log p(Y|\theta)}{\partial \theta}]\}$$

$\hat{\theta}(Y)$ is the parameter estimate computed from the measurement

$p(Y|\theta)$ is the conditional probability density of the measurement given the parameter θ

The Fisher information matrix, which is computed using the measured data, is a measure of the amount of information a measured signal contains about the unknown system parameters. It therefore follows that a criterion based on the Fisher information matrix would be an useful tool for parameter estimation. In fact different criteria such as A-Optimal (minimizing the trace of the inverse of the information matrix) and D-Optimal (maximizing the determinant of the information matrix) have been developed. Goodwin and Payne [1] delve deeper into this topic and explore optimal input design based on the Fisher information matrix criteria.

In this study we aim to probe the potential of an information theoretic concept called the mutual information, to serve as an indicator for the amount of information in the provided data-set. This in turn translates to the adequacy of mutual information to serve as a criterion for parameter estimation. Our motivation to explore this connection is derived from the recent works of Andrew Liu and Robert Bitmead [2], where they discuss an entropy based approach to understanding observability in a non-linear context. Another connection that motivates us is the one between observability and identifiability [3], where identifiability refers to the problem of estimating the system parameters from the given data. Hamed Ebrahimian, Rodrigo Astroza, Joel Conte and Robert Bitmead have also studied the applicability of information theoretic approach in identifiability of nonlinear

models [4]. Assuming that we have access to both the input and output data from a system, our goal is to quantify the information content within each data-set in terms of mutual information, thus enabling us to segregate the informative data from the uninformative ones.

In Section 2 we investigate if signal properties such as entropy and conditional entropy can be empirically calculated from the provided data-set. We discuss the different factors influencing these empirical calculations such as binning of the data-set and length of the measurements, which provides us with some insight regarding the deviation of the empirical values from the theoretical ones, and its dependence on the size of the data-set.

In Section 3, we explore the idea of using mutual information between the measured signals as a viable criterion for designing an input signal for the purpose of parameter estimation. In the vein of this discussion we make some assumptions to facilitate our analysis such as focusing only on ARX system models, with all the signals involved following a Gaussian distribution. As there already exists some theoretical framework connecting the entropy ($H(X)$) and therefore mutual information of jointly Gaussian signals with their probability distribution parameters, we borrow these ideas to come up with the optimal input signal.

$$H(X) = \frac{1}{2}\log_2\Big((2\pi e)^k \det(\Sigma)\Big)$$

$$\text{where,} \qquad X \sim \mathcal{N}(m, \Sigma)$$

$$\text{and,} \qquad X = [X_1, X_2, ..., X_k]^T$$

We then discuss some case studies where we show that by forcing input constraint of unit power, the optimal input maximizing the mutual information criterion in an ARX system model turns out to be a single sinusoid at a particular frequency, and that by further constraining the input by imposing additional structure to it, we arrive at the optimal

input as a coloured noise signal. Towards the end of this section we compare how mutual information fare as a criterion for input synthesis, with some other pre-established well known criterion based on the Fisher information matrix [1]. We find mutual information to be problematic in this context.

We end this study by noting that this attempt at studying the implementation of information theoretic approach for experiment design is merely scratching the surface, and that there still remains a plethora of research to be done.

Chapter 2

Empirical Entropy and Mutual Information Calculation

Before we dive in any deeper into the topic of experiment design we first need to define and understand the criterion we will be focusing on. In Section 2.1 we introduce the concepts of entropy, conditional entropy, and mutual information.

In the later sections we explore the question of how reliably can we empirically compute these said quantities from the measurement data, and the effect of variables such as measurement length and bin sizes on the deviation of empirical value from theoretical ones.

2.1 Entropy and Mutual Information

Definition 1. *(Entropy) The entropy $H(X)$ expressed in bits, of a discrete random variable is defined as the degree of uncertainty associated with the random variable [5], and is given*

by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \tag{2.1}$$

where for a discrete random variable X taking values in $\mathcal{X}$, $p(x)$ is the probability mass function (pmf) of random variable X.

One way of interpreting entropy is to think of it as the number of binary yes or no question one needs to ask on average to ascertain the outcome of a random variable.

A similar concepts of entropy exist for a continuous random variable. For a continuous random variable X taking values in $\mathcal{X}$, with probability density function (pdf) $p(x)$, entropy $H(X)$ is given by

$$H(X) = -\int_{\mathcal{X}} p(x) \log_2 p(x)\, dx \tag{2.2}$$

A continuous random variable can be discretized by binning its range and grouping its possible outcome, and can therefore be approximated by a discrete random variable. The entropy of a continuous random variable can then be related to the entropy of its discretized version. Thomas and Cover in their book [5], show that when a continuous random variable X is approximated by its discretized version X^{Δ}, by dividing the range of X into bins of equal length Δ, the entropy of the continuous random variable $H(X)$ can be approximated using the entropy of its discretized version $H(X^{\Delta})$ using the following expression.

$$H(X) \approx H(X^{\Delta}) + \log_2 \Delta$$

and additonally,

$$H(X^{\Delta}) + \log_2 \Delta \to H(X), \text{ as } \Delta \to 0$$

One observation to make here is that in the above expression as Δ tends to 0, the $\log_2 \Delta$ term approaches negative infinity, while at the same time the term $H(X^\Delta)$ grows as increasingly large number of summations are required to calculate $H(X^\Delta)$, assuming that we have enough data points to adequately compute the pmf for X^Δ. However when these two terms are added and then the bin size is decreased, they have the apparent effect of cancelling out the of change in each other, while their sum approaches the theoretical value of $H(X)$.

Definition 2. *(Conditional Entropy $H(Y|X)$) For jointly distributed random variables conditional entropy is defined as the average uncertainty in a random variable given the knowledge of another [5].*

For discrete random variables Y and X taking values in $\mathcal{Y}$ and $\mathcal{X}$ with marginal probability mass functions $p(y)$ and $p(x)$ respectively, and joint probability mass function $p(x,y)$ and conditional probability mass function $p(y|x)$ of Y given X, the conditional entropy $H(Y|X)$ of Y given X is given by

$$\begin{aligned} H(Y|X) &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)\log_2 p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(y|x) \\ &= H(X,Y) - H(X) \end{aligned} \tag{2.3}$$

where $H(X,Y)$ is the entropy of the joint random variable (X,Y)

Definition 3. *(Mutual Information) The mutual information $\mathcal{I}(X;Y)$ between the random variables X and Y is the difference between the entropy of Y and the conditional entropy of Y given X. It can be understood as the reduction in uncertainty, or equivalently a gain in*

information about of a random variable due to knowledge of another.

$$\begin{aligned}\mathcal{I}(X;Y) &= H(Y) - H(Y|X) = H(X) - H(X|Y) \\ &= H(Y) + H(X) - H(X,Y) \end{aligned} \tag{2.4}$$

As was the case for entropy, the mutual information between two continuous random variables can also be approximated by the mutual information between their discretized versions.

$$\begin{aligned}\mathcal{I}(X;Y) &= H(Y) - H(Y|X) \\ &\approx H(Y^{\Delta}) + \log_2 \Delta - (H(Y^{\Delta}|X^{\Delta}) + \log_2 \Delta) \\ &\approx H(Y^{\Delta}) - H(Y^{\Delta}|X^{\Delta}) \\ &\approx \mathcal{I}(X^{\Delta};Y^{\Delta}) \end{aligned} \tag{2.5}$$

2.2 Simulation Setup

If we intend to use information theoretic concepts such as mutual information for experiment design, it follows that we should investigate how reliably can these quantities be computed from the measurement data.

As an exercise in how accurately one can expect to empirically compute entropy and mutual information between continuous random variables, we set up a simulation where we obtain the empirical values for the entropy and mutual information using their discretized version, and compare it against their theoretical counterparts calculated using their respective mathematical formula.

Simulation steps for empirical entropy calculation

i.) We begin by choosing an i.i.d. continuous random variable X with an underlying Gaussian pdf, i.e. $X \sim \mathcal{N}(0,1)$. With the knowledge of these distribution parameters, the entropy of X can be computed using the formula [5].

$$H(X) = \frac{1}{2}\log_2(2\pi e\sigma^2) = \frac{1}{2}\log_2(2\pi e1) = 2.0471 \tag{2.6}$$

We then simulate a series of outcome of random variable X to obtain measurement of length N.

ii.) Next we pick a bin size Δ for the purpose of discretization.

iii.) Another important decision to make here is that of the number of bins n to use. Lets say we pick a range [-r, r] and divide it into bins of equal size Δ. Then the number of bins n_r in this range is

$$n_r = \frac{r-(-r)}{\Delta} = \frac{2r}{\Delta}$$

It is obvious that only two the variables can be selected simultaneously. We go with the choice of the bin size Δ, and the number of bins n_r as our variables.
To accommodate for the measurement values that falls outside of this range of [-r, r], we construct two additional bins, ranging from $(-\infty, -r)$ and (∞, r). Thus the total number of bins in our simulation is n = $(n_r + 2)$ bins.

iv.) Now that all the choices have been made, we can use them to bin the measurement data accordingly. By dividing the number of data points in each bin by the total length of measurement N, we compute the probability mass function for our discretized random variable X^{Δ}. Using this pmf the entropy of X^{Δ} can be computed.

v.) After adding the factor of $\log_2 \Delta$ to the the entropy of the discretized random variable

$H(X^{\Delta})$ we get the empirical value of entropy of X, which can be compared against the theoretical entropy of X

vi.) For this simulations we choose the following values for the variables N, n, and Δ.

$$N \in \{1000,\ 10000,\ 100000,\ 1000000,\ 10000000\}$$

$$n \in \{3,\ 102,\ 502,\ 1002,\ 2002\}$$

$$\Delta \in \{1,\ 0.1,\ 0.01,\ 0.001,\ 0.0001\}$$

Simulation steps for empirical mutual information calculation

i.) For simulation of empirical mutual information calculation, most of our simulation framework is similar to our previous exercise. The key differences being that instead of just simulating one continuous random variable X, we simulate another continuous random variable Y along with it.

ii.) We obtain the random variable Y by adding an i.i.d. noise signal with Gaussian distribution properties $\mathcal{N}(0,0.3)$, to the continuous random variable X.

iii.) Once we have the measurement information for both X and Y, we can compute the marginal pmf and joint pmf functions for the discretized random variable X^{Δ}, Y^{Δ}, $[X^{\Delta}, Y^{\Delta}]^T$

iv.) Proceeding in a similar fashion as before, we compute the empirical entropy for X^{Δ} and Y^{Δ}, and along with it also compute the joint entropy for the joint distribution $(X^{\Delta}, Y^{\Delta^T})$

v.) Using equation (2.4), we calculate the empirical mutual information between the random variable (X, Y), and then compare it's result with the theoretical value of

the mutual information between (X, Y)

By varying the values of the variable N, n, and Δ separately, their individual effect on the empirical value of entropy of X, and the mutual information between (X, Y) can be studied.

2.3 Simulation Results

In this section we discuss the results of the simulation from the previous section. In Figure 2.1 we have tabulated the absolute error between the empirical and the theoretical entropy values of the simulated random variable X.

One observation to make here is that for the same bin size Δ and number of bins n there seems to be a general trend that as we increase the measurement length N the error between the empirical value and theoretical value of entropy decreases. This is in line with what we would expect, that as we increase the number of data points, we can plot better probability mass function of the discretized random variable X^{Δ} which is closer to its true underlying distribution, therefore giving us a better approximation of the $H(X^{\Delta})$, and subsequently $H(X)$.

Another observation to make is that, in general, for a fixed bin size Δ and measurement length N, increasing the number of bins n leads to a better approximation for the entropy $H(X)$.

The parameter that seems to have the most impact on the accuracy of our calculation is the choice of bin size Δ. Choosing a larger value of bin size Δ can reduce the resolution of the pmf of X^{Δ} obtained from the data. On the other hand, if we reduce the bin size Δ too small, so that each bin corresponds to one measurement value, we might not have enough data points fall into each bin to reflect the true underlying distribution of X^{Δ}.

Usually, one only has a limited length of measurement data to work with. Therefore

in practice depending on the number of data points available to us we would needs to strike a balance between the number of bins n and the bin size Δ in such a way that helps us construct the probability mass function of X^{Δ} as close to it's actual underlying distribution as possible.

In our simulations we had selected a continuous random variable X whose entropy was:

$$H(X) = \frac{1}{2}\log_2\left(2\pi e(1)^2\right) = 2.0471$$

And with suitable choice of N, n and Δ we were able to achieve an absolute error on the order of 10^{-3}, which corresponds to a relative error of 0.048% between the theoretical and the empirically calculated value of the entropy of X.

Therefore it can be stated that, for a suitable choices of number of bin n, measurement length N, and bins size Δ one could possibly empirically compute the entropy of a continuous random variable with minor deviations.

Similar strategy needs to be applied while empirically calculating the mutual information from the data. Selecting a suitable bin size Δ and number of bins n, which provide us with an adequate reconstruction of the pmf of the discretized random variables can lead to a sufficiently accurate empirical calculation of the mutual information between two random variables.

Now that we have the confirmation that it is indeed possible to reliably empirically calculate the mutual information from the measurement data, we can move on to the next question and ask whether mutual information is an useful tool for experiment design.

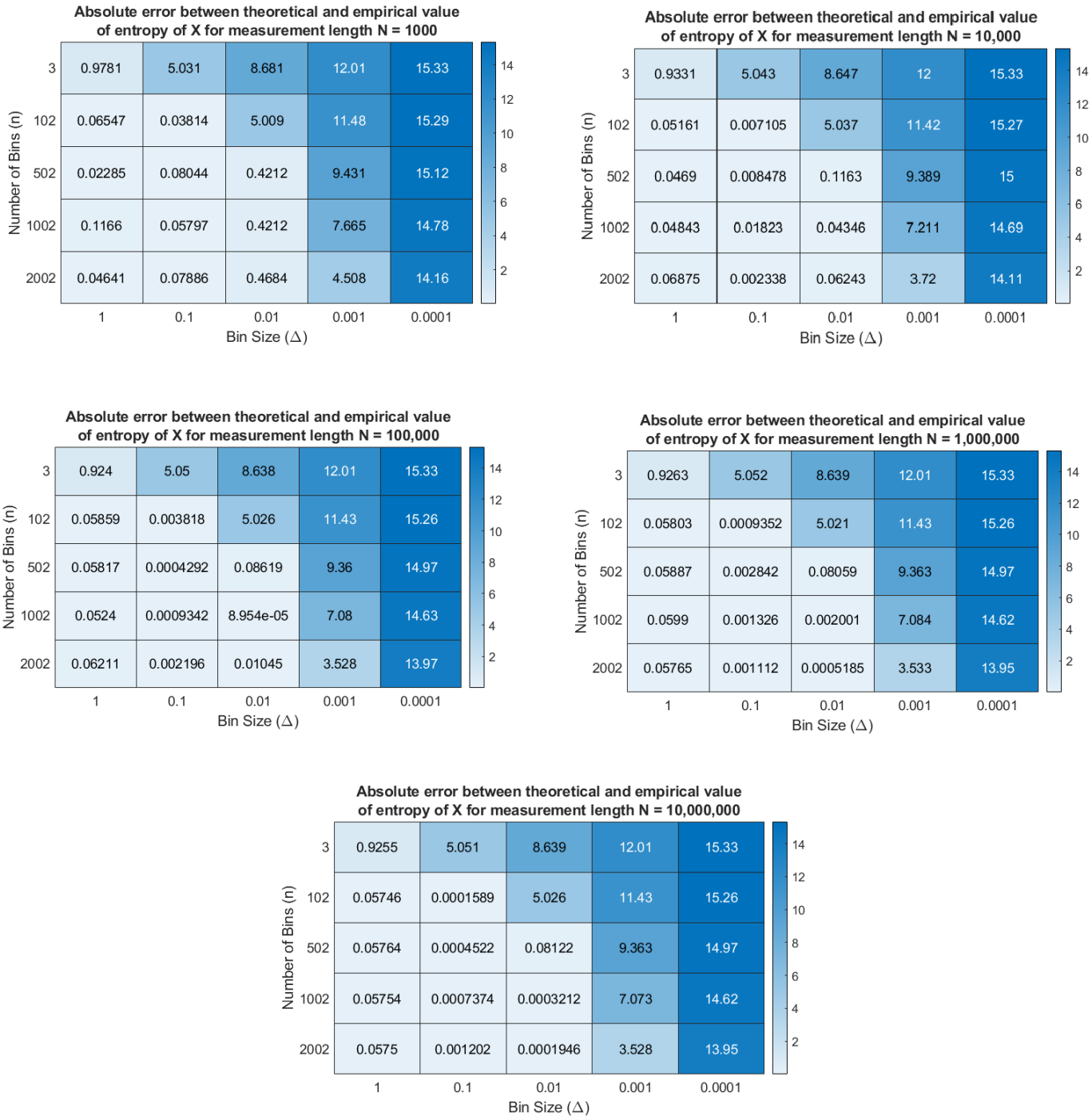

Figure 2.1: Absolute error between the empirically computed entropy and the theoretically calculated entropy of the simulated continuous random variable $X \sim (\mathcal{N}(0,1))$ for different values of measurement length N.

Chapter 3

Mutual Information and Identifiability

The preceding section dealt with the empirical evaluation of the quantities such as entropy of signal and the mutual information between two signals. Next we try and find how mutual information relates to a system model.

This section focuses on exploring the possibility of using mutual information as a criterion for identifiability and separating informative data-set from the rest. Our approach is to first select a system model to go along with our criterion, consider all the assumptions necessary, then analyze if maximizing the mutual information leads us to any conclusions about identifiability for the selected system model from the given data.

3.1 Model Setup

As stated before, we choose an ARX system model to accompany our mutual information criterion. A general ARX system is described by the following equivalent

discrete time equations.

$$y(t) = \frac{\alpha_0 + \alpha_1 q^{-1} + ... + \alpha_n q^{-n}}{1 - \beta_1 q^{-1} - \beta_2 q^{-2} - ... - \beta_m q^{-m}} x(t) + \frac{1}{1 - \beta_1 q^{-1} - \beta_2 q^{-2} - ... - \beta_m q^{-m}} e(t) \quad (3.1)$$

$$\begin{aligned} y(t) = {} & \alpha_0 x(t) + \alpha_1 x(t-1) + ... + \alpha_n x(t-n) + \beta_1 y(t-1) + \beta_2 y(t-2) + \\ & ... + \beta_m y(t-m) + e(t) \end{aligned}$$

$$y(t) = \begin{bmatrix} \alpha_0 & \alpha_1 & ... & \alpha_n & \beta_1 & \beta_2 & ... & \beta_m \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-1) \\ : \\ x(t-n) \\ y(t-1) \\ y(t-2) \\ : \\ y(t-m) \end{bmatrix} + e(t)$$

$$y(t) = \theta^T \phi(t) + e(t)$$

where θ is the vector of parameters and $\phi(t)$ is the regressor vector,

$$\theta^T = \begin{bmatrix} \alpha_0 & \alpha_1 & ... & \alpha_n & \beta_1 & \beta_2 & ... & \beta_m \end{bmatrix}$$

$$\phi(t)^T = \begin{bmatrix} x(t) & x(t-1) & ... & x(t-n) & y(t-1) & y(t-2) & ... & y(t-m) \end{bmatrix}$$

Throughout this entire chapter we will be making some key assumptions which we mention below.

Assumption 1. *The assumptions we make here are as follows,*

i.) The discrete time ARX system model is stable, i.e. all the poles of the transfer function of ARX system model lie inside of the unit circle.

ii.) The input, output, and noise signal are jointly Quasi-Stationary [3], i.e. for any signals s(t) and w(t)

$$E\{s(t)\} = m_s(t) \qquad |m_s(t)| \leq C \qquad \forall t$$
$$E\{s(t)w(r)\} = R_{sw}(t,r) \qquad |R_{sw}(t,r)| \leq C \qquad \forall t, r$$
$$\lim_{N\to\infty} \frac{1}{N} \sum_{\tau=1}^{N} R_{sw}(t, t-\tau) = R_{sw}(\tau) \quad \forall \quad \tau$$

iii.) The input signal x(t) is a zero-mean Gaussian signal, i.e. $x(t) \sim \mathcal{N}(0, R_x(0))$.

iv.) The noise signal e(t) is an i.i.d. zero-mean Gaussian signal, i.e. $e(t) \sim \mathcal{N}(0, R_e(0))$.

v.) The input signal is uncorrelated with the noise signal, i.e. E{x(t)e(r)} = 0 for all integer t and r.

3.2 System Analysis

In this section we try to come up with a mathematical expression for the mutual information between the output signal and the regressor signal for an ARX system model and use it to draw conclusions.

Conjecture 1. *For an ARX system (3.1) of any given order, subject to Assumption 1, the mutual information between the output signal and the regressor signal is half the logarithm*

to the base 2 of the output signal to noise ratio.

$$\mathcal{I}(y(t);\phi(t)) = \frac{1}{2}\log_2\left(\frac{R_y(0)}{R_e(0)}\right) \tag{3.2}$$

Next we demonstrate that the above conjecture holds true a first-order ARX system.

Theorem 1. *For a first-order ARX system, subject to Assumption 1, the mutual information between the output signal and the regressor signal is half the logarithm to the base 2 of the output signal to noise ratio.*

$$\mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2\left(\frac{R_y(0)}{R_e(0)}\right) \tag{3.3}$$

Proof. A first-order ARX system is described by the following discrete time equation:

$$\begin{aligned} y(t) &= \alpha_0 x(t) + \beta_1 y(t-1) + e(t) \\ y(t) &= \begin{bmatrix} \alpha_0 & \beta_1 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix} + e(t) \end{aligned} \tag{3.4}$$

Where mutual information between output y(t) and regressor vector $\begin{bmatrix} x(t) & y(t-1) \end{bmatrix}^T$ is given by:

$$\mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = H(\begin{bmatrix} y(t) \end{bmatrix}) + H(\begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) - H(\begin{bmatrix} y(t) \\ x(t) \\ y(t-1) \end{bmatrix}) \tag{3.5}$$

As all the signals in the above equation are Gaussian, we can express their respective

entropy using the formula for entropy of a Gaussian distribution [5].

$$H(\Big[y(t)\Big]) = \frac{1}{2}\log_2\left(2\pi e R_y(0)\right)$$

$$H(\begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2\left((2\pi e)^2 \det(\begin{bmatrix} R_x(0) & R_{yx}(-1) \\ R_{yx}(-1) & R_y(0) \end{bmatrix})\right)$$

$$H(\begin{bmatrix} y(t) \\ x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2\left((2\pi e)^3 \det(\begin{bmatrix} R_y(0) & R_{yx}(0) & R_y(1) \\ R_{yx}(0) & R_x(0) & R_{yx}(-1) \\ R_y(1) & R_{yx}(-1) & R_y(0) \end{bmatrix})\right)$$

where,

$$\begin{aligned} R_{yx}(0) &= \alpha_0 R_x(0) + \beta_1 R_{yx}(-1) \\ R_y(1) &= \alpha_0 R_{yx}(-1) + \beta_1 R_y(0) \\ R_{ye}(0) &= R_e(0) \\ R_{yx}(-1) &= \frac{(1-\beta_1^2)R_y(0) - \alpha_0^2 R_x(0) - R_e(0)}{2\alpha_0\beta_1} \end{aligned}$$

Making the above substitutions in equation (3.5)

$$\mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2 \left(\frac{R_y(0) * det(\begin{bmatrix} R_x(0) & R_{yx}(-1) \\ R_{yx}(-1) & R_y(0) \end{bmatrix})}{det(\begin{bmatrix} R_y(0) & R_{yx}(0) & R_y(1) \\ R_{yx}(0) & R_x(0) & R_{yx}(-1) \\ R_y(1) & R_{yx}(-1) & R_y(0) \end{bmatrix})} \right)$$

$$\mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2 \left(\frac{R_y(0)}{R_e(0)} \right)$$

The proof for a second-order ARX system in provided in the Appendix (A) and is fully parallel to the proof for first-order ARX system and provides a schema for higher order proofs. ■

From the above analysis it is easy to see that the mutual information between the output signal and the regressor signal for an ARX system is proportional to the output signal to noise ratio, and that if one intends to maximize the mutual information as a criterion, one must maximize the output signal power.

3.3 Maximizing Mutual Information: Case Study 1

Now that we have conjectured a mathematical expression for the mutual information between the output signal and the regressor we can check if it fits the role of a suitable criterion for input synthesis for system identification. We approach this problem by attempting to maximize the mutual information under the constraint of unit power for input signal.

Lemma 1. *Assuming Conjecture 1 holds true, for a given ARX system (3.1), under Assumption 1 and unit input signal power constraint, the choice of input that maximizes the mutual information between the output signal and regressor signal is a sinusoid at the frequency that maximizes the magnitude of the frequency response of the ARX system transfer function.*

Proof. For a general ARX system (3.1) denoted in transfer function notation,

$$y(t) = G(q)x(t) + H(q)e(t) \tag{3.6}$$

the output signal power σ_y^2 is given by

$$\begin{aligned}\sigma_y^2 = R_y(0) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_y(\omega)\, d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(|G(\omega)|^2\Phi_x(\omega) + |H(\omega)|^2\Phi_e(\omega)\right) d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} |G(\omega)|^2\Phi_x(\omega)\, d\omega + constant\end{aligned}$$

where, $\Phi_y(\omega)$ is the spectrum of the output signal,
and is equal to the Discrete Time Fourier Transform of the correlation function $R_y(\tau)$
For a unit input signal power, i.e.

$$\sigma_x^2 = R_x(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_x(\omega)\, d\omega = 1 \tag{3.7}$$

the maximum output signal power is achieved for

$$\begin{aligned}\Phi_x(\omega) &= \frac{1}{2}(2\pi)(\delta(\omega-\omega_o) + \delta(\omega+\omega_o)) \\ R_x(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_x(\omega)e^{j\omega\tau}\, d\omega = \cos(\omega_o\tau)\end{aligned}$$

where ω_o is the frequency that maximizes the magnitude of the frequency response of the system transfer function.

The input signal satisfying the above conditions is given by

$$x(t) = \sqrt{2}\cos(\omega_o t) \qquad \blacksquare$$

Now to discuss the findings, let us consider a simple case of an ARX system model with 3 system parameters to estimate. If we were to use mutual information as a criterion to design an input signal with unit power for parameter estimation, the above result instructs us to use a sinusoid at the frequency corresponding to maximum of the magnitude of the frequency response of the transfer function of the system.

However on the other hand, one can only estimate 2 system parameters from the data-set acquired through a single sinusoid at input. This means that no matter what we do, we will never be able to estimate all of the 3 system parameters, and therefore the ARX system model becomes unidentifiable with the data acquired under the criterion of maximizing the mutual information. And the situation only gets worse once we move to higher order ARX model, as even then the input is only suitable for estimating 2 system parameters.

Corollary 1. *Assuming Conjecture 1 holds true, for any ARX system model, subject to the Assumption 1 and unit input signal power constraint, the sinusoid input signal formulated by maximizing the mutual information criterion, leads to an unidentifiable ARX system model.*

To further inspect how mutual information fares as a criterion, we compare the results from the above study with some of the already established results from other methods. As mentioned in the introduction, criteria based on the Fisher information matrix are regularly used for experiment design.

Goodwin and Payne [1], use a criterion based on the determinant of Fisher informa-

tion matrix to design an input with unit signal power for an FIR model. They show that for an FIR system model the Fisher information matrix based criterion is maximized by a white noise input signal. This contradicts the mutual information optimizing input of a sinusoid, which lead to unidentifiability if there are 3 or more system parameters to be estimated.

The contrast between the two results is clear as both the designed input are as far removed from one another as possible. Where the Fisher information matrix criterion tells us to spread out the input spectrum over all the available frequencies, the mutual information criterion is recommending that we do the exact opposite and instructs us to concentrate the entire input power at one frequency.

This discussion leads us to conclude that mutual information is not a suitable criterion for experiment design with an ARX system model as it leads to unidentifiability in this simple case. We revisit the same problem in the next section, but this time with some stricter constraints while formulating the input signal to avoid running into the pitfall of designing a singular sinusoidal input.

It can also be concluded that maximizing mutual information between the output signal and the regressor signal of an ARX system model does not correspond to maximizing a measure of identifiability in any meaningful way, and that the data-sets segregated based on their empirically calculated mutual information values offer no visible merit over the rest.

3.4 Maximizing Mutual Information: Case Study 2

As another attempt at checking the viability of mutual information as a criterion, we attempt to improvise on the results from the previous discussion. The trouble with the solution from previous Section 3.3 was that it tends to invest the entire input signal power

at a single frequency, thus limiting our possibility of attaining identifiability.

Our approach in this section will be to address the aforementioned issue by providing additional structure to the input signal, and probing if maximizing the mutual information as a criterion still leads to unidentifiability.

We restrict the input signal to a particular class of functions, given by:

$$\textbf{Structure of input functions:} \quad x(t) = \sin(\zeta) * w(t) + \cos(\zeta) * w(t-1) \tag{3.8}$$

where $w(t)$ is an i.i.d zero-mean Gaussian white noise signal i.e. $w(t) \sim \mathcal{N}(0, R_w(0))$, uncorrelated with the noise signal e(t). Imposing the unit input signal power constraint

$$R_x(0) = R_w(0) = 1$$

Additionally,

$$R_x(1) = \frac{\sin(2\zeta)}{2}; \qquad R_x(\tau) = 0 \ \forall \ \tau \in \mathbb{Z}^+ - \{0, 1\}$$

Theorem 2. *Assuming Conjecture 1 and Assumption 1 holds true, for a first-order ARX system (3.4), with the input belonging to the structure (3.8), the value of ζ that maximizes the mutual information between the output signal and the regressor signal, depends on the sign of the system parameter β_1, and is given by*

$$\begin{aligned} \arg\max_{\zeta} \mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) &= \frac{\pi}{4} \quad \textit{if} \quad \beta_1 \in [0, 1) \\ &= \frac{3\pi}{4} \quad \textit{if} \quad \beta_1 \in (-1, 0] \end{aligned} \tag{3.9}$$

Proof. As mentioned in the theorem we focus our attention to a first-order ARX system

model, given by

$$y(t) = \alpha_0 x(t) + \beta_1 y(t-1) + e(t)$$

Using the result of Conjecture 1, we can write the Mutual Information between the output signal and the regressor for the ARX system as

$$\mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2\left(\frac{R_y(0)}{R_e(0)}\right)$$

Substitution for $R_y(0)$ for our first-order ARX system we get

$$R_y(0) = \frac{\alpha_0^2 R_x(0) + \alpha_0^2 \beta_1 \sin(2\zeta) R_x(0) + R_e(0)}{(1-\beta_1^2)}$$

$$\mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) = \frac{1}{2}\log_2\left(\frac{\alpha_0^2 R_x(0) + \alpha_0^2 \beta_1 \sin(2\zeta) R_x(0) + R_e(0)}{(1-\beta_1^2) * R_e(0)}\right)$$

By differentiating the above equation and equating it to zero, we arrive at

$$\begin{aligned} \arg\max_{\zeta} \mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}) &= \frac{\pi}{4} \quad \text{if} \quad \beta_1 \in [0,1) \\ &= \frac{3\pi}{4} \quad \text{if} \quad \beta_1 \in (-1,0] \end{aligned}$$

■

It can be easily verified that these results correspond to a maximum by checking the double derivative at these ζ values.

We arrive at two separate input signals that maximize the mutual information for

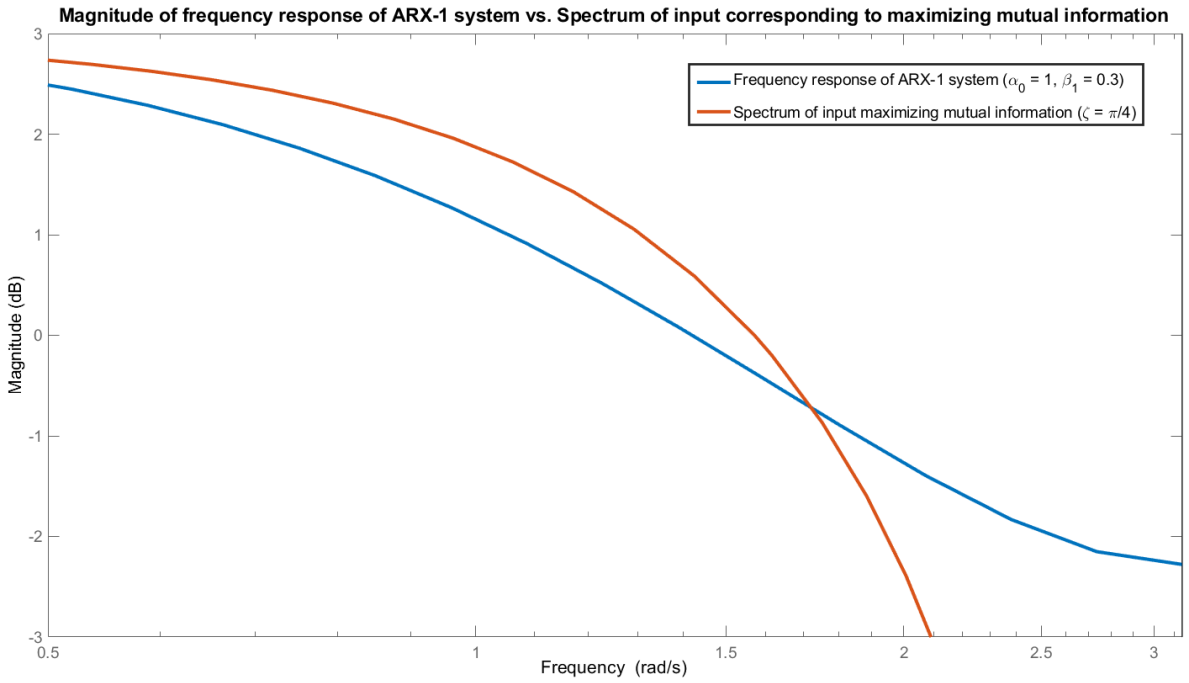

Figure 3.1: Magnitude curve of frequency response of the transfer function of ARX-1 system with parameters $\alpha_0 = 1$ and $\beta_1 = 0.3$ and spectrum of input signal maximizing mutual information.

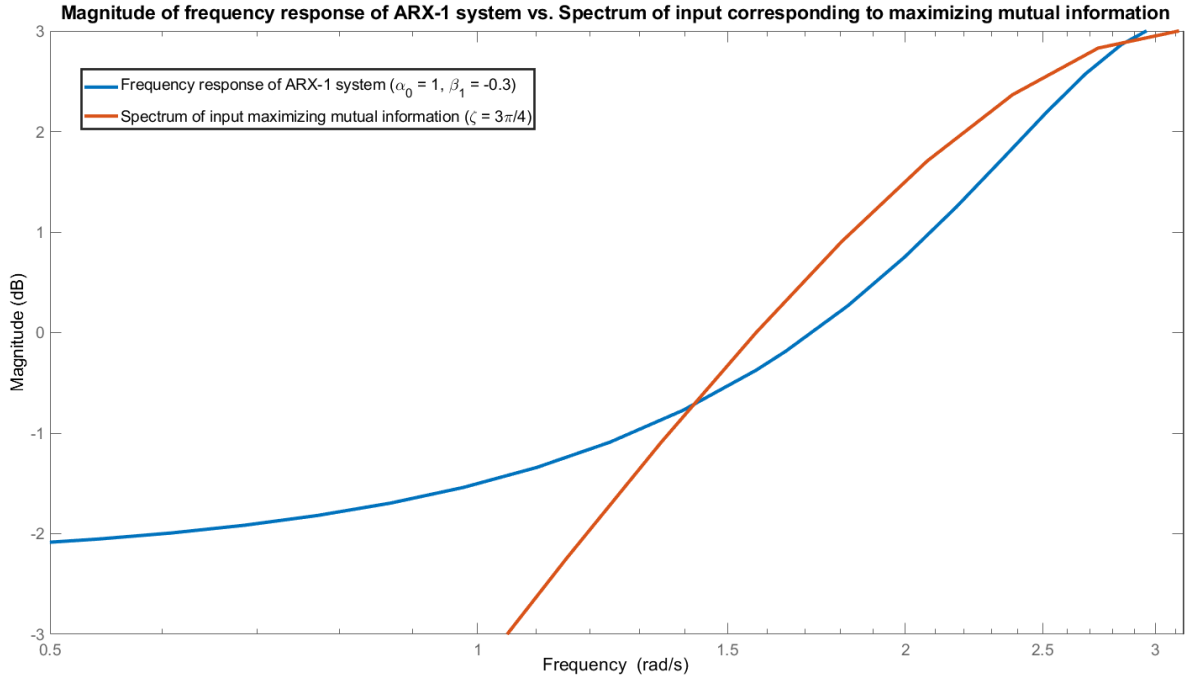

Figure 3.2: Magnitude curve of frequency response of the transfer function of ARX-1 system with parameters $\alpha_0 = 1$ and $\beta_1 = -0.3$ and spectrum of input signal maximizing mutual information.

the first-order ARX system depending upon the value of the system parameters.

For the case of $\beta_1 \in (-1, 0]$, the first order ARX system is a high-pass filter, and the input signal following the structure (3.8) that maximizes the mutual information for this type of system is a colored noise rich in high frequency content.

On the contrary, for $\beta_1 \in [0, 1)$, the first order ARX system is a low-pass filter, and the input signal following the structure (3.8) that maximizes the mutual information for this type of system is a colored noise rich in low frequency content.

Taking a look figure (3.1) and (3.2), it can be concluded that trying to maximize the mutual information for a first order ARX system leads to an input signal that tries to maximizes the overlap between its spectrum and magnitude curve of the frequency response of the transfer function of the first-order ARX system.

Comparing the results of this Section 3.4 with that of the previous Section 3.3, we can see that there is definitely some improvement in our design of input signal in terms of presence of identifiability. As noted in Section 3.3, when trying to maximize the mutual information by itself we arrived at a solution which almost always guarantees unidentifiability for our ARX system model, as a consequence of suggesting a pure sinusoid for the input signal.

As seen here, when we restrict the input signal to a particular class of functions for design of input signal based on mutual information criterion, we avoid the said problem of arriving at a singular sinusoid. In addition to that, the designed input signal also turns out to be a colored noise signal which fares better than a singular sinusoid for the purpose of parameter estimation.

We conclude this chapter by acknowledging that while by itself mutual information is not a suitable criterion for experiment design, parameter estimation, and qualitative sorting of data-sets, when augmented with appropriate input structure, or some other yet unknown constraints it does show signs of potential to be an effective design criterion.

Chapter 4

Discussion

Throughout this study we have attempted to explore the possibility of using mutual information as a criterion for separating informative data-sets from the rest. While we recognise that we were successful at conjecturing an expression connecting the mutual information between the output signal and the regressor of an ARX system model, we must stay conscious of all the assumptions under which the results are applicable.

We further found that, with just the signal power as a constraint on the input signal, the mutual information criterion leads to an objectively worse input design. Therefore an attempt at sorting informative data-set from the collection of experimental records based solely on empirically calculated mutual information values would not help us in any meaningful ways.

However there still remains a possibility for using mutual information as a criterion for experiment design, if additional structure is provided to the input.

From the results of the previous chapter it is evident that to use mutual information as a criterion we need to further refine the process of adequately setting up the design problem, as the assumptions and structure of the input plays a pivotal role on how viable the criteria is for experiment design. However still, these results serves as an insight that

this topic remains worth pursuing.

Appendix A

Derivation of Conjecture 1 for second-order ARX system

We utilize this appendix to show that the Conjecture 1 holds true even for a second-order ARX system.

Theorem 3. *For a second-order ARX system, subject to Assumption 1, the mutual information between the output signal and the regressor signal is half the logarithm to the base 2 of the output signal to noise ratio.*

$$\mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = \frac{1}{2}\log_2\left(\frac{R_y(0)}{R_e(0)}\right) \tag{A.1}$$

Proof. A second-order ARX system is described by the following discrete time equation:

$$y(t) = \alpha_0 x(t) + \alpha_1 x(t-1) + \beta_1 y(t-1) + \beta_2 y(t-2) + e(t) \tag{A.2}$$

$$y(t) = \begin{bmatrix} \alpha_0 & \alpha_1 & \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix} + e(t)$$

Where mutual information between output y(t) and regressor $\begin{bmatrix} x(t) & x(t-1) & y(t-1) & y(t-2) \end{bmatrix}^T$ is given by:

$$\mathcal{I}(\begin{bmatrix} y(t) \end{bmatrix}; \begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = H(\begin{bmatrix} y(t) \end{bmatrix}) + H(\begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) - H(\begin{bmatrix} y(t) \\ x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) \tag{A.3}$$

As all the signals in the above equation are Gaussian, we can express their respective entropy using the formula for entropy of a Gaussian distribution [5].

$$H(\begin{bmatrix} y(t) \end{bmatrix}) = \frac{1}{2} \log_2 (2\pi e R_y(0))$$

$$H(\begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = \frac{1}{2}\log_2 \left((2\pi e)^4 \det(\begin{bmatrix} R_x(0) & R_x(1) & R_{yx}(-1) & R_{yx}(-2) \\ R_x(1) & R_x(0) & R_{yx}(0) & R_{yx}(-1) \\ R_{yx}(-1) & R_{yx}(0) & R_y(0) & R_y(1) \\ R_{yx}(-2) & R_{yx}(-1) & R_y(1) & R_y(0) \end{bmatrix}) \right)$$

$$H(\begin{bmatrix} y(t) \\ x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = \frac{1}{2}\log_2 \left((2\pi e)^5 \det(\begin{bmatrix} R_y(0) & R_{yx}(0) & R_{yx}(1) & R_y(1) & R_y(2) \\ R_{yx}(0) & R_x(0) & R_x(1) & R_{yx}(-1) & R_{yx}(-2) \\ R_{yx}(1) & R_x(1) & R_x(0) & R_{yx}(0) & R_{yx}(-1) \\ R_y(1) & R_{yx}(-1) & R_{yx}(0) & R_y(0) & R_y(1) \\ R_y(2) & R_{yx}(-2) & R_{yx}(-1) & R_y(1) & R_y(0) \end{bmatrix}) \right)$$

Where,

$$R_{yx}(0) = \alpha_0 R_x(0) + \alpha_1 R_x(1) + \beta_1 R_{yx}(-1) + \beta_2 R_{yx}(-2)$$

$$R_{yx}(1) = \alpha_0 R_x(1) + \alpha_1 R_x(0) + \beta_1 R_{yx}(0) + \beta_2 R_{yx}(-1)$$

$$R_y(1) = \alpha_0 R_{yx}(-1) + \alpha_1 R_{yx}(0) + \beta_1 R_y(0) + \beta_2 R_y(1)$$

$$R_y(2) = \alpha_0 R_{yx}(-2) + \alpha_1 R_{yx}(-1) + \beta_1 R_y(1) + \beta_2 R_y(0)$$

$$R_y(0) = \frac{(\alpha_0^2 + \alpha_1^2) R_x(0) + 2\alpha_0\alpha_1 R_x(1) + 2\beta_1\beta_2 R_y(1) + 2\alpha_1\beta_1 R_{yx}(0)}{1 - \beta_1^2 - \beta_2^2} + \frac{2(\alpha_0\beta_1 + \alpha_1\beta_2) R_{yx}(-1) + 2\alpha_0\beta_2 R_{yx}(-2) + R_e(0)}{1 - \beta_1^2 - \beta_2^2}$$

Making the above substitutions in equation (A.3)

$$\mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = \frac{1}{2}\log_2 \left(\frac{R_y(0) * det(\begin{bmatrix} R_x(0) & R_x(1) & R_{yx}(-1) & R_{yx}(-2) \\ R_x(1) & R_x(0) & R_{yx}(0) & R_{yx}(-1) \\ R_{yx}(-1) & R_{yx}(0) & R_y(0) & R_y(1) \\ R_{yx}(-2) & R_{yx}(-1) & R_y(1) & R_y(0) \end{bmatrix})}{det(\begin{bmatrix} R_y(0) & R_{yx}(0) & R_{yx}(1) & R_y(1) & R_y(2) \\ R_{yx}(0) & R_x(0) & R_x(1) & R_{yx}(-1) & R_{yx}(-2) \\ R_{yx}(1) & R_x(1) & R_x(0) & R_{yx}(0) & R_{yx}(-1) \\ R_y(1) & R_{yx}(-1) & R_{yx}(0) & R_y(0) & R_y(1) \\ R_y(2) & R_{yx}(-2) & R_{yx}(-1) & R_y(1) & R_y(0) \end{bmatrix})} \right)$$

$$\mathcal{I}(\Big[y(t)\Big]; \begin{bmatrix} x(t) \\ x(t-1) \\ y(t-1) \\ y(t-2) \end{bmatrix}) = \frac{1}{2}\log_2 \left(\frac{R_y(0)}{R_e(0)} \right)$$

■

The above exercise therefore shows that Conjecture 1 holds for a second-order ARX system as well. We believe that similar relation holds for all higher order ARX systems, and that they can be derived in a similar way.

Bibliography

[1] Graham C. Goodwin and Robert L. Payne. *Dynamic System Identifications: Experiment Design and Data Analysis*. Academic Press, 1977.

[2] Andrew R. Liu and Robert R.Bitmead. Stochastic observability in network state estimation and control. *Automatica*, 2010.

[3] Lennart Ljung. *System Identification Theory for User*. Prentice Hall, 1999.

[4] Hamed Ebrahimian, Rodrigo Astroza, Joel P. Conte, and Robert R. Bitmead. Information-theoretic approach for identifiability assessment of nonlinear structural finite-element models. *American Society of Civil Engineers*, 2019.

[5] Thomas M. Cover and Joy A. Thomas. *Elements Of Information Theory*. Wiley-Interscience, 2006.