Semi-deterministic finite interval estimation of linear system dynamics and output trajectory

Farhad Rahbarnia

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

June 2020

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering.

Abstract

An efficient approach adopting Reproducing Kernel Hilbert Space, RKHS, to estimate the parameters of Differential Equations from noisy realizations of the system's output is presented in this thesis. Initially, this thesis studies the previous works on parameter and state estimation using RKHS. This approach estimates the parameters, order n, the output trajectory and the derivatives of the system up to n-1, where n is the true order. The presented approach is able to handle error in the variable using local fitting and regularization. The suggested method uses Bayesian Information Criterion, BIC, to evaluate possible order for unknown systems. Lastly, to increase the accuracy and computational speed, the approach applies hyper-parameter search and cross-validation to tune its cost function's coefficients. The MATLAB software package has been implemented to evaluate the suggested approach.

Résumé

Une approche efficace qui adopte la reproduction de l'espace Kernel Hilbert, RKHS, pour estimer les paramètres des équations différentielles des réalisations bruyantes de la sortie du système est présentée dans cette thèse. Initialement, cette thèse étudie les travaux précédents sur le paramètre et l'estimation de l'état à l'aide de RKHS. Cette approche estime les paramètres, l'ordre n, la trajectoire de sortie et les dérivés du système jusqu'à n-1, où n est l'ordre réel. L'approche présentée est capable de traiter l'erreur dans la variable en utilisant l'ajustement local et la régularisation. La méthode suggérée utilise le critère d'information bayésien, BIC, pour évaluer l'ordre possible pour les systèmes inconnus. Enfin, pour augmenter la précision et la vitesse de calcul, l'approche applique la recherche hyper-paramètre et la validation croisée pour régler les coefficients de sa fonction de coût. Le logiciel MATLAB a été mis en œuvre pour évaluer l'approche suggérée.

Acknowledgments

I want to express my gratitude to people who have supported me constantly throughout my academic journey to produce this thesis. First and foremost, I like to thank my supervisor, Professor Hannah Michalska, who guided me to gain a more proper understanding on the topic of this research. I would like to express appreciation for her patience and constant guidance. I sincerely acknowledge her help in every aspect of my studies and regard her with great respect for the exceptional researcher and kind person that she is.

I would also like to thank my thesis examiner Professor Masoud Asgharian from the Department of Mathematics and Statistics at McGill University. Prof. Asgharian's thoughtful feedback provided interesting directions for further study.

I would like to thank my parents; Prof. Freydoon Rahbarnia and Mrs. Marjaneh Golzarghaeiny, for their constant support, love and guidance. I would like to thank friends and family who encouraged me during this time.

Foremost, I would like to thank God, who always works in mysteries ways.

Preface

This is to declare that this thesis is part of a collaborative work done by myself under the supervision of Professor Hannah Michalska. In addition, there has been a collaboration between Professor Michalska, Debarshi Patanjali Ghoshal, Ph.D. scholar and I during the development of the kernels. This work extends the work carried out in[41].

Contents

List of Figures

List of Tables

List of Acronyms

RKHS	Reproducing Kernel Hilbert Space
KS	Kolmogorov Smirnov
LTI	Linear Time Invariant
RMSD	Root Mean Square Difference
SISO	Single Input Single Output
LOWESS	Linear regression over a window
LOESS	Quadratic regression over a window
SNR	Signal to Noise Ratio
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion
MSE	Mean Squared Error
IV	Incremental Variables
SSE	the Sum of Squared Error
Var(x)	Variance of variable x

Chapter 1

1 Introduction

The Control theory is a well-studied area with many applications; [63]. Control signals are predominantly generated in closed-loop[74]. A general control system employing feedback [41] is represented below[54].

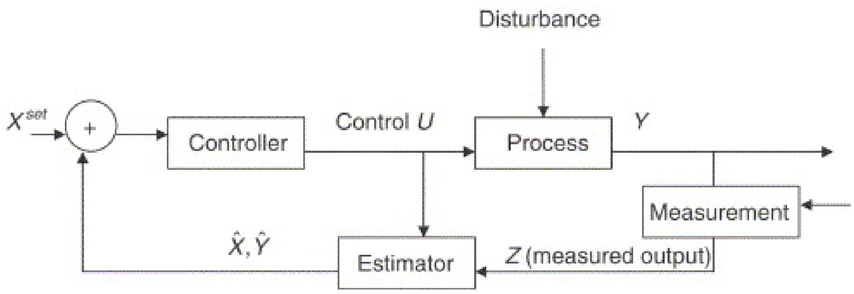

Figure 1: Feedback Controller with State Estimator

Figure 1 shows a simple feedback Controller with State Estimator. A simple example of feedback control is a system that monitor and adjust the temperature inside the living room. The *plant* is essentially the system whose output is to be controlled. For an electric room heating system the plant is the current carrying coils that would emit heat. The feedback element is a sensor that measures the output parameter that is to be controlled, in this case it is temperature. The sensor measurement is the feedback to the input and subtracted from the original signal i.e. the set point. This difference, also known as the error, is fed to the controller. The controller decides a control action that would reduce this error and drive the output to the desired set-point value. Choice of controller is crucial in designing a successful control system. Model based controller design hinges on the knowledge of the *plant* itself [74].

To design a model based controller, a parametric mathematical model with finite number of parameters of the *plant* is necessary to be known. The system is essentially described by differential equations. The knowledge of all its coefficients is needed to successfully design a model based controller. This causes the problem of parameter estimation to be of equal, if not more important, than state estimation.[74]

System identification is the process of determining the values of parameters that govern the output of a dynamical system.

An example of a parameter estimation problem would comprise of the following; consider a continuous-time (CT) linear time-invariant (LTI) system, whose order n is known, described by[12]:

$$a_n \frac{d^n x}{dt_n} + a_{n-1} \frac{d^{n-1} x}{dt_{n-1}} + \cdots + a_0 x(t) = b_m \frac{d^m u}{dt_m} + b_{m-1} \frac{d^{m-1} u}{dt_{m-1}} + \cdots + u(t) \tag{1}$$

A noise corrupted version of x is measured, $z(t_k) = x(t_k) + \epsilon(t_k), k = 1, 2, ..., M$ where, $\epsilon(t_k)$ is Gaussian noise with mean zero and variance σ^2, M is the number of samples taken. The goal of identification is to estimate the parameters $a_n, a_{n-1}, ..., a_1, a_0, b_m, b_{m-1}, ..., b_1$ from the input-output data i.e., $(u(t_k), z(t_k); k = 1, 2, ..., M)$. [79]

The parameter estimation of a homogeneous system can be viewed as the identification of a differential invariant $\mathcal{I}$ ($\mathcal{I} \equiv 0$)under the action of the flow of some closed loop system (such as its characteristic equation) [28]:

$$\mathcal{I}(t, y(t), y^{(1)}(t), \cdots, y^{(n)}(t)) \equiv y^{(n)}(t) + a_{n-1} y^{(n-1)}(t) + \cdots + a_0 y(t); t \geq 0 \tag{2}$$

The concept of state estimation is quite crucial in control theory. The figure below shows a closed loop control system with a state estimator.[12]

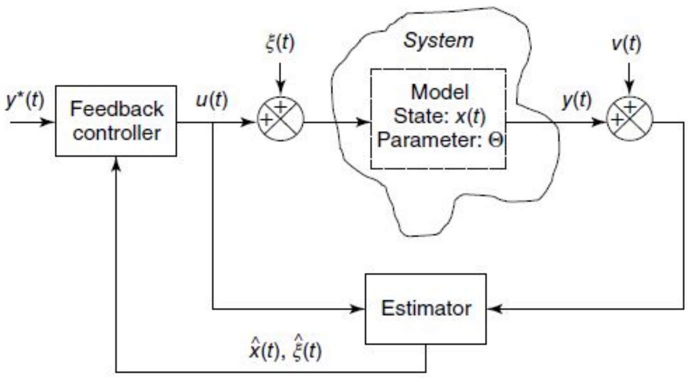

Figure 2: A closed loop control system with state estimator [74]

A state vector is a set of vectors that are used to describe the mathematical state of a dynamical system. The state of a system is a minimal set of variables/vectors (state variables/vectors) which together with the values of the input signals in the future, completely determine the future behavior of the system [74].

It can be seen in figure 2 that the states $x(t)$ and system parameters θ are required to produce a system output $y(t)$. For the closed loop feedback system to work, the controller needs all the states to generate a control action that would drive the output $y(t)$ to its set - point value $y^*(t)$. Also the system is fed with measurement noise $v(t)$ and input disturbances $\zeta(t)$. Hence, it is expected that the controller has to be less sensitive to these external signals.[12]

For a large system, all the state variables cannot be physically measured. So an estimator needs to be designed which can estimate the states $x(t)$ from the output $y(t)$ and input $u(t)$. Hence, extensive research on state observers and filters was conducted with seminal contributions of Kalman and Lueanberger; [45],[46],[48],[47],[55],[54],[14]

There are many important applications for system identification and filtering; for instance, [52], [43], [5]. Recursive approaches have been used extensively to estimate the system parameters. Kalman filter is one of the most famous examples of recursive algorithms, it is simple and yet has been used since 1960,[81]. Their low computational complexity and simple nature make them a powerful tool.

Recently, however, there is a growing interest in non-asymptotic estimation methods which is justified by modern developments in systems with rapidly switching dynamics [3], advanced nonlinear control methods based on differential flatness [21] and the need for powerful target tracking algorithms of superior performance in speed and precision [7]. Recently, new non-asymptotic and non-recursive approaches emerged for state and parameter estimation on finite time intervals[74].

Finite interval (algebraic) estimation relies on algebraic manipulation of output derivatives. Algebraic estimation methods are sensitive to measurement noise. Noise attenuation is attempted by iterated integration and shaping the annihilator functions used to *eliminate* the effect of the initial conditions.

The differential-algebraic theory of observability states that a system is algebraically observable if its states at any time instant can be expressed in terms of the input, the output and their respective time derivatives of finite order [16], [74].

When a system model, which is the mathematical description of a system, is defined in terms of variables and parameters, it is known as a parametric model [74]. Parameter estimation is the process of empirical determination of the parameter values that govern the behavior of the system from a noise-contaminated output signal [74].

The problem considered here is related to the joint parameter and state estimation over a finite interval of time. The length of the interval is constrained by the requirement of the identifiability of the system from a single realization of the stochastic measurement process. There is no restriction on the number of measured samples. Consider a noisy measurement signal as shown in Figure 3. This shows the true value of a system versus the noisy observed value. The behavior of a system is hard to observe, let alone predict. The problem is to identify an algebraic function that binds the time derivatives along the presumed trajectory and remains invariant under the dynamical flow of the system, irrespective of any noise.

Needless to remark that, as first appreciated in the seminal work of Wiener, [82], the above problem has other countless applications, all of which call for the highest estimation precision in the shortest possible time, often in the presence of high levels of measurement noise of unknown characteristics. To mention only the need for non-asymptotic estimation methods as justified by modern developments in systems with rapidly switching dynamics, [2], advanced nonlinear control design based on differential flatness, [20], and the demand for powerful target tracking algorithms of superior performance in speed and precision, [6].

The idea of constructing kernel representations of differential systems, as first presented in [32], and continued in [28], [29], [33], [75], [30], initially sprung from an attempt to improve on the algebraic signal differentiation approach proposed in [23], [74], as the principal tool required in the implementation of controllers for nonlinear systems that are differentially flat. The original algebraic differentiator was based on truncated Taylor series signal approximation. It attributed its properties to the introduction of an algebraic annihilator of the Taylor series coefficients up to any desired order, [23], [73], [58], and [13]. After iterated integration, it then delivered the values of the selected coefficients that escaped annihilation, effectively the values of the desired higher-order derivatives. The annihilator served yet a different purpose: that of shaping the noise rejection characteristic of the resulting filter, see [59]. As it was admitted, however, the method required frequent re-initialization when used forward in time, and its noise rejection properties were characterized as non-standard; see [18]. Despite the power of non-standard analysis methods, [15], used in [18] as well as the efforts to attenuate measurement noise by designing SNR-maximizing annihilators and pre-filters, it was finally admitted in [74] that algebraic estimation was still very sensitive to noise. It thus became clear, that any further improvements hinged on the application of more rigorous statistical and probabilistic analysis and framework.

The work of [41] has implemented the two-step Non-asymptotic Parameter and State Estimation. However, the method used in [41] is slow in computation and with large noise, it does not produce the correct values of parameters.

The previous works all had several drawbacks:

- The order of system was assumed to be known
- They all failed to handle noise with large SNR (-10dB)
- The processing time and power required was extremely large to the point that makes them none-practical

1.1 Thesis contributions

Here we build on the work [41] and improve on it. The contribution of this thesis can be summarized as follows;

- An approach is proposed to estimate the parameters, order n , the output trajectory and the derivatives of the system up to $n-1$, where n is the true order.
- An approach is proposed to handle error in the variable using local fitting, Regu-

larization, and instrumental variables to handle large noise in observation as well as over-fitting.

- A modified BIC, Bayesian Information Criterion, is proposed to evaluate possible order for unknown systems.
- An approach is proposed to increase the accuracy and computational speed using hyper-parameter search and cross-validation.
- A software package for MATLAB is developed which is more efficient and user-friendly than the previous works of such [41].

The method can easily be adapted to handle zero-mean coloured noise and applies to systems with or without input of arbitrary order, with very little a priori information available.

1.2 Thesis Objectives and Organization

The objectives of this thesis were stated as follows:

- Understand the mathematical background behind algebraic state and parameter estimation;
- Study and understand previous contributions of Professor Michalska on the non-asymptotic joint state and parameter estimation: [41] [64], [37], [27], [28]. Study the novel kernel representation of SISO LTI systems;
- Study and experiment with diverse statements of the finite interval simultaneous state and parameter estimation problem in reproducing kernel Hilbert spaces (RKHS).
- Understand the noise characteristics and methods that can handle large noise in measurements.
- Construct an estimator that is efficient and can handle noise in signal.
- Construct an estimator that automatically determines the order of the system.

The thesis is structured as follows:

Chapter 1 provides a introduction to algebraic state and parameter estimation. It states the objectives of this thesis.

Chapter 2 studies the works [41],[64], [37], [27] and [28] on derivation of the double sided kernels for homogeneous SISO LTI systems.

Chapter 3 provides a summery of previous works on parameter estimations. It states some of shortcomings of methods presented in [41],[64], [37], [27] and [28].

Chapter 4 introduces different methods to handle large noise in observations. It uses local smoothing as well as incremental variables, which were briefly introduced in [25], to tackle the error in the observation.

Chapter 5 introduces penalty term to the cost function to handle over-fitting. It uses hyper-parameter search to tune its cost function.

Chapter 6 introduces a simple, yet effective method to handle order selection. This chapter details methods to handle large noise, over-fitting, slow computation, and unrealistic assumption that the order of the system is known.

Chapter 7 presents diverse experiments that examine the proposed method in different settings and different noise levels.

Chapter 8 provides the conclusions and discusses directions for improvements of the method.

Chapter 2

2 A Double Sided Kernel in SISO LTI Representation

The double sided kernel approach was first developed and discussed in [36]. It employs forward and backward integration and the Cauchy formula for multiple integrals to convert a high order differential equation, that embodies a system invariant, into an integral form with no singularities at the boundaries of the observation window. In the approach of [36], the state equations are replaced with an output reproducing property on an arbitrary time $[t_a, t_b]$ which follows directly from the knowledge of the system characteristic equation. The behavioural model is derived from the differential in-variance which is characteristic of the system and eliminates the need of initial conditions and is in the form of a homogeneous Fredholm integral equation of the second kind with a Hilbert-Schmidt kernel [27]. The mathematical interpretation as a Reproducing Kernel Hilbert Space (RKHS) of the behavioural model helps us to extract signal and its time derivatives from noisy output measurements.

2.1 A finite interval estimation problem for an LTI system

Consider a general n-th order, strictly proper and minimal SISO LTI system no input signal in state space form evolving on a given finite time interval $[a, b] \subset \mathbb{R}$:

$$\begin{aligned} \dot{x} &= Ax \\ y &= C^T x \\ x &\in \mathbb{R}^n \end{aligned} \tag{3}$$

where the system matrix A is in canonical form,

$$\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-1} \end{bmatrix} \tag{4}$$

and,

$$C = \begin{bmatrix} 1 & 0 & 0 & \dots \end{bmatrix} \tag{5}$$

with matching dimensions of the system matrices and the characteristic equation,

$$\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0 = 0 \tag{6}$$

The input-output equation for system (3) becomes

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{7}$$

2.2 Problem Statement and Assumptions

The estimation problem is stated as follows. Given an arbitrary finite interval of time $[a, b]$, suppose that:

(1) The dimension of the state vector of the LTI system is not known a priori. However, a set of possible dimensions for the system is assumed.

(2) The system input function $u(t)$ is equal to zero (No input).

(3) The output of the system is observed as a single realization of a 'continuous' measurement process $y_M(t) := y(t) + \eta(t)$, $t \in [a, b]$ in which η denotes additive white Gaussian noise with unknown intensity (variance) σ^2.

Under the conditions (i)-(iv), it is required that:

(i) Identifiability of the system parameters $a_i, i = 0, \cdots, n-1$ from a single realization of the measurement process $y_M(t), t \in [a, b]$ be determined for the given observation horizon $[a, b]$;

(ii) Under identifiability condition, a parameter estimator for $a_i, i = 0, \cdots, n-1$ be proposed that is statistically consistent;

(iii) The true system output $y(t)$ and its $n-1$ time derivatives $y^{(i)}(t), i = 1, \ldots, n-1$ be reconstructed from the noisy observation $y_M(t)$ over $t \in [a, b]$;

(iv) The reconstruction error in the output function y and its derivatives $y^{(i)}, i = 1, \ldots, n-1$ converges to zero uniformly as the number of measurement samples of y_M increases freely.

An implementable version of assumption (iii) simply requires availability of an unrestricted number of output measurements over the observation horizon $[a, b]$.

For simplicity of exposition, the estimation approach is presented here with respect to the more interesting case of homogeneous systems in which persistent excitation cannot be rendered by way of the system input u. Without the loss of generality, the order of the system considered in an example demonstration of the validity and properties of the approach is taken to be $n = 4$.

The strategy to estimate parameters of ODE is a multi-stage method, which in the first stage estimates the function and its derivatives from noisy observations using data smoothing methods without considering differential equation models, and then in the following stages estimates of ODE parameters and the order of the system using statistical methods such as least squares,[84]. In the works of Liang and Wu ([51]) a two-stage method is developed for a general first order ODE model, using local polynomial regression in the first stage, with established asymptotic properties of the estimator. The two-stage methods are easy to implement, however, they might not be statistically efficient, due to the fact that derivatives cannot be estimated accurately from noisy data, especially higher order derivatives.

As mentioned before, a straightforward two-stage strategy, though easy to implement, has difficulty in estimating derivatives of the dynamic process accurately, leading to biased estimates of the ODE parameter. The method in this thesis uses multi-stage method that:

- First smooth the data using a local (quadratic) regression.
- It considers different candidate systems of orders to be (consider different orders to be fitted the noisy observation.).
- Then uses the kernel presented in Chapter 2 and the updated cost function to the estimate of the parameters for each of those system.
- It uses Hyper parameter tuning to choose the penalty term coefficients.
- It then estimate the output projection by the using the fundamental solution method for each candidate system.
- It then calculate the Bayesian Information Criterion (BIC) values for each candidate system.
- It selects the candidate system the lowest.
- The higher degree derivatives are calculated using kernels defined in Chapter 2

For the simplicity, we assume $n = 4$, a fourth order system.

2.3 A kernel representation of a system differential invariance

The unknown values of the parameters $a_i, i = 0, .., n-1$ need to be identified using noisy observations of the system's output $y(t)$ over a finite, but arbitrary, interval of time $t \in [0, T]$, $T > 0$.

The Kernel representation of the n-order SISO LTI system was introduced in [25]. The cornerstone of the finite interval estimation approach presented here is the integral representation of the controlled differential invariance of the system (7). For the homogeneous system the latter is given by a mapping J:

$$\begin{aligned} J(y, y^{(1)}, \cdots, y^{(n)}) &:= y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1 y^{(1)}(t) + a_0 y(t) \\ &\equiv const. = 0; \quad t \in [a, b] \end{aligned} \tag{8}$$

that remains constant under the action of the flow of the system in the absence of external forcing. The validity of (8) is seen to deliver additional measurement-noise independent information about the behaviour of the system beyond mere observation of the noisy y_M. To be useful, however, the 'zero-input response' characterization (8) has to be put in a form, which does not depend on the initial or boundary conditions of the system, and that does not involve any time derivatives of the output as the last are not available through direct measurement.[25] Such a representation of (8) has been presented for SISO LTI systems of arbitrary order in [30]. The full theorems along with their rigorous proofs is presented at [25] and attached here in appendix.

The following definition was introduced in [25]

Definition 2.1. A pair of smooth (class $\mathcal{C}^\infty$) functions (α_a, α_b), $\alpha_s : [a, b] \to \mathbb{R}$, $s = a$ or b, is an annihilator of the boundary conditions for system (3) if the functions α_s are non-negative, monotonic, vanish with their derivatives up to order $n-1$ at the respective ends of the interval $[a, b]$; i.e.

$$\alpha_s^{(i)}(s) = 0 \quad i = 0, \cdots, n-1; \quad s = a, b; \ \alpha_s^{(0)} \equiv \alpha_s \tag{9}$$

and such that their sum is strictly positive, i.e. that for some constant $c > 0$

$$\alpha_{ab}(t) := \alpha_a(t) + \alpha_b(t) > c \ ; \ t \in [a, b] \tag{10}$$

A simplest example of such an annihilator for (3) is the pair

$$\begin{aligned} &\alpha_a(t) := (t-a)^n, \quad \alpha_b(t) := (b-t)^n \ ; \ t \in [a, b] \\ &\alpha_{ab}(t) := \alpha_a(t) + \alpha_b(t) > 0 \ ; \ t \in [a, b] \\ &\alpha_{ab}(s) = (b-a)^n, \quad s = a, b \end{aligned} \tag{11}$$

Indeed, it is easy to see that (11) holds for all $n \geq 1$ because $\alpha_{ab}(a) = \alpha_{ab}(b) = (b-a)^n > 0$ and, for $n \geq 2$, α_{ab} has a unique stationary point $t^* = 0.5(a+b) \in [a, b]$ at which

$\alpha_{ab}(t^*) = (0.5)^{n-1}(b-a)^n$.

Employing this particular annihilator the integral representation for system (1) is rendered by the following:

Theorem 2.2. *[25] There exist Hilbert-Schmidt kernels $K_{DS,y}$, $K_{DS,u}$, such that the input and output functions u and y of (1) satisfy*

$$y(t) = \alpha_{ab}^{-1}(t)\left[\int_a^b K_{DS,y}(n,t,\tau)y(\tau)d\tau + \int_a^b K_{DS,u}(n,t,\tau)u(\tau)d\tau\right] \tag{12}$$

with

$$\alpha_{ab}^{-1}(t) := \frac{1}{(t-a)^n + (b-t)^n} \tag{13}$$

Hilbert-Schmidt double-sided kernels of (12) are square integrable functions on $L^2[a,b] \times L^2[a,b]$ and are expressed in terms of the 'forward' and 'backward' kernels given below:

$$\begin{aligned} K_{DS,y}(n,t,\tau) &\triangleq \begin{cases} K_{F,y}(n,t,\tau), & \text{for } \tau \leq t \\ K_{B,y}(n,t,\tau), & \text{for } \tau > t \end{cases} \\ K_{DS,u}(n,t,\tau) &\triangleq \begin{cases} K_{F,u}(n,t,\tau), & \text{for } \tau \leq t \\ K_{B,u}(n,t,\tau), & \text{for } \tau > t \end{cases} \end{aligned} \tag{14}$$

The kernel functions $K_{DS,y}, K_{DS,u}$ are $n-1$ times differentiable as functions of t.

The forward and backward kernels of (14) have the following expressions:

$$\begin{aligned} K_{F,y}(n,t,\tau) = \sum_{j=1}^{n}(-1)^{j+1}\binom{n}{j}\frac{n!(t-\tau)^{j-1}(\tau-a)^{n-j}}{(n-j)!(j-1)!} \\ + \sum_{i=0}^{n-1} a_i \sum_{j=0}^{i}(-1)^{j+1}\binom{i}{j}\frac{n!(t-\tau)^{n-i+j-1}(\tau-a)^{n-j}}{(n-j)!(n-i+j-1)!} \end{aligned}$$

$$\begin{aligned} K_{B,y}(n,t,\tau) = \sum_{j=1}^{n}\binom{n}{j}\frac{n!(t-\tau)^{j-1}(b-\tau)^{n-j}}{(n-j)!(j-1)!} \\ + \sum_{i=0}^{n-1} a_i \sum_{j=0}^{i}\binom{i}{j}\frac{n!(t-\tau)^{n-i+j-1}(b-\tau)^{n-j}}{(n-j)!(n-i+j-1)!} \end{aligned}$$

$$K_{F,u}(n,t,\tau) = \sum_{i=0}^{n-1} b_i \sum_{j=0}^{i} (-1)^{j+1} \binom{i}{j} \frac{n!(t-\tau)^{n-i+j-1}(\tau-a)^{n-j}}{(n-j)!(n-i+j-1)!}$$

$$K_{B,u}(n,t,\tau) = \sum_{i=0}^{n-1} b_i \sum_{j=0}^{i} \binom{i}{j} \frac{n!(t-\tau)^{n-i+j-1}(b-\tau)^{n-j}}{(n-j)!(n-i+j-1)!}$$

□

The proof is found in [25].

The proof of the representation Theorem 3.2 is essentially conducted by construction, involving the induction argument only at its final stage. Therefore, the only loss of information in the passage from the input-output equation of the system to the integral kernel representation in Theorem 3.2 is that of any pre-existing boundary conditions in (84) as the latter are annihilated during every integration operation by the presence of the annihilating factors α_a and α_b. The following conjecture is then quite obvious.

Corollary 2.3. *For any given input function u, the output function $y : [a,b] \to \mathbb{R}$ satisfies the system input-output equation (1) on the interval $[a,b]$ if an only if it satisfies the integral equation* (12) *regardless of any boundary conditions that may be imposed. The kernel representation of the system invariance provides a unique criterion whose reproducing property unambiguously characterizes all zero input solutions of the SISO LTI system. In particular all the fundamental solutions of the LTI system share the reproducing property* (12) *as they span a subspace of the RKHS of dimension n.*

□

Explicit kernel expressions for the derivatives of the output function

Due to the regularity properties of the kernel functions in Theorem 2.2, it is straightforward to obtain the corresponding recursive formulae for the time derivatives of the system output $y^{(i)}, i = 1, \cdots, n-1$.

Theorem 2.4. *[25] There exist Hilbert-Schmidt kernels $K_{F,k,y}$, $K_{F,k,u}$, $K_{B,k,y}$, $K_{B,k,u}$, $k = 1, \cdots, n-1$ such that the derivatives of the output function in* (1) *can be computed rercursively as follows:*

$$y^{(k)}(t) = \frac{1}{(t-a)^n + (b-t)^n}\Bigg[\sum_{i=1}^{k}(-1)^{i+1}\binom{p+i-1}{i}\frac{n!(t-a)^{n-i}y^{(k-i)}(t)}{(n-i)!}$$
$$+\sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}(-1)^{j+1}\binom{p+j-1}{j}\frac{n!(t-a)^{n-j}y^{(i-j-p)}(t)}{(n-j)!} + \int_a^t K_{F,k,y}(n,p,t,\tau)y(\tau)d\tau$$
$$+\sum_{i=p}^{n-1}b_i\sum_{j=0}^{i-p}(-1)^{j+1}\binom{p+j-1}{j}\frac{n!(t-a)^{n-j}u^{(i-j-p)}(t)}{(n-j)!} + \int_a^t K_{F,k,u}(n,p,t,\tau)u(\tau)d\tau$$

$$-\sum_{i=1}^{k}\binom{p+i-1}{i}\frac{n!(b-t)^{n-i}y^{(k-i)}(t)}{(n-i)!} - \sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}\binom{p+j-1}{j}\frac{n!(b-t)^{n-j}y^{(i-j-p)}(t)}{(n-j)!}$$
$$+\int_t^b K_{B,k,y}(n,p,t,\tau)y(\tau)d\tau - \sum_{i=p}^{n-1}b_i\sum_{j=0}^{i-p}\binom{p+j-1}{j}\frac{n!(b-t)^{n-j}u^{(i-j-p)}(t)}{(n-j)!}$$
$$+\int_t^b K_{B,k,u}(n,p,t,\tau)u(\tau)d\tau\Bigg]$$

where $\quad p = n-k$

$$K_{F,k,y}(n,p,t,\tau) = \sum_{j=1}^{p}(-1)^{j+n-p+1}\binom{n}{n-p+j}\frac{n!(t-\tau)^{j-1}(\tau-a)^{p-j}}{(p-j)!(j-1)!}$$
$$+\sum_{i=0}^{p-1}a_i\sum_{j=0}^{i}(-1)^{j+1}\binom{i}{j}\frac{n!(t-\tau)^{p-i+j-1}(\tau-a)^{n-j}}{(n-j)!(p-i+j-1)!}$$
$$+\sum_{i=p}^{n-1}a_i\sum_{j=1}^{p}(-1)^{j+i-p+1}\binom{i}{i-p+j}\frac{n!(t-\tau)^{j-1}(\tau-a)^{n-i+p-j}}{(n-i+p-j)!(j-1)!}$$

$$K_{F,k,u}(n,p,t,\tau) = \sum_{i=0}^{p-1}b_i\sum_{j=0}^{i}(-1)^{j+1}\binom{i}{j}\frac{n!(t-\tau)^{p-i+j-1}(\tau-a)^{n-j}}{(n-j)!(p-i+j-1)!}$$
$$+\sum_{i=p}^{n-1}b_i\sum_{j=1}^{p}(-1)^{j+i-p+1}\binom{i}{i-p+j}\frac{n!(t-\tau)^{j-1}(\tau-a)^{n-i+p-j}}{(n-i+p-j)!(j-1)!}$$

$$
\begin{aligned}
K_{B,k,y}(n,p,t,\tau) = & \sum_{j=1}^{p} \binom{n}{n-p+j} \frac{n!(t-\tau)^{j-1}(b-\tau)^{p-j}}{(p-j)!(j-1)!} \\
& + \sum_{i=0}^{p-1} a_i \sum_{j=0}^{i} \binom{i}{j} \frac{n!(t-\tau)^{p-i+j-1}(b-\tau)^{n-j}}{(n-j)!(p-i+j-1)!} \\
& + \sum_{i=p}^{n-1} a_i \sum_{j=1}^{p} \binom{i}{i-p+j} \frac{n!(t-\tau)^{j-1}(b-\tau)^{n-i+p-j}}{(n-i+p-j)!(j-1)!}
\end{aligned}
$$

$$
\begin{aligned}
K_{B,k,u}(n,p,t,\tau) = & \sum_{i=0}^{p-1} b_i \sum_{j=0}^{i} \binom{i}{j} \frac{n!(t-\tau)^{p-i+j-1}(b-\tau)^{n-j}}{(n-j)!(p-i+j-1)!} \\
& + \sum_{i=p}^{n-1} b_i \sum_{j=1}^{p} \binom{i}{i-p+j} \frac{n!(t-\tau)^{j-1}(b-\tau)^{n-i+p-j}}{(n-i+p-j)!(j-1)!}
\end{aligned}
$$

□

The proof from [25] is found in Appendix.

Remark 1 [62]

It is important to note that for $u \equiv 0$, the invariance representation (12) is in fact a continuous evaluation functional for the system output functions and hence induces a unique reproducing kernel Hilbert space (RKHS) with kernel

$$K_y(t_1, t_2) := \alpha_{ab}^{-2}(t) \int_a^b K_{DS,y}(t_1, \tau) K_{DS,y}(t_2, \tau) d\tau \tag{15}$$

$t_1, t_2 \in [a, b]$ where the dependence of the kernels on the system order n has been suppressed for brevity; see [71] for more details. The kernel (15) can in fact be regarded as a deterministic counterpart of a 'covariance structure' of a process whose trend is described in terms of the differential invariant. It then also follows that any bounded linear functional $L : y \to L(y)(t)$ on the RKHS space has a representer, which is the kernel $L(K_y(t, \cdot))$. This kernel trick and other properties of the RKHS have vast implications for estimation in RKHS, see [65], [66], followed by [42]. The properties of the kernels established by Theorem 3.2 also imply that approximation of functions in the induced RKHS (such as those needed in output signal reconstruction) can be carried out in some countable basis of feature splines:

$$\text{span}\{K_{DS,y}(t_i, \cdot);\; t_i \in [a, b], i \in \mathbb{I}\} \tag{16}$$

This approach is non-asymptotic in nature and the initial conditions need not be known. The parameter estimation of a homogeneous system can be viewed as the identification of a differential invariant $\mathcal{I}$ ($\mathcal{I} \equiv 0$) which does not change under the action of the system dynamics

$$\begin{aligned}&\mathcal{I}(t, y(t), y^{(1)}(t), \cdots, y^{(n)}(t)) \\ &\quad = y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_0 y(t) \ ; \ t \geq 0\end{aligned} \tag{17}$$

The problem with singularity at $t = 0$, encountered in [22] in the integral system representation is eliminated in [28] using two-sided forward-backward integration which leads to an alternative integral system representation of (17) in a reproducing kernel Hilbert space (RKHS).

The knowledge of the system characteristic equation enables one to replace the state equations with an output reproducing property on an arbitrary time interval $[a, b]$. The differential invariance, which is essentially the characteristic equation, is used to derive the behavioral model of the system[27]. The mathematical interpretation as a Reproducing Kernel Hilbert Space (RKHS) of the behavioral model allows us to extract signal and its time derivatives that confirm the system invariance from output measurement subject to noise[64].

2.4 Explicit kernel expressions for the derivatives of the output function

he explicit kernel expressions for the derivatives of the output function was derived in [25]. These are given as follows;

$$\begin{aligned} y^{(k)}(t) = \frac{1}{(t-a)^n + (b-t)^n} \Bigg[&\sum_{i=1}^{k} (-1)^{i+1} \binom{p+i-1}{i} \frac{n!}{(n-i)!} (t-a)^{n-i} y^{(k-i)}(t) \\ + \sum_{i=p}^{n-1} a_i \sum_{j=0}^{i-p} (-1)^{j+1} \binom{p+j-1}{j} &\frac{n!}{(n-j)!} (t-a)^{n-j} y^{(i-j-p)}(t) + \int_a^t K_{F,k,y}(n,p,t,\tau) y(\tau) d\tau \\ &- \sum_{i=1}^{k} \binom{p+i-1}{i} \frac{n!}{(n-i)!} (b-t)^{n-i} y^{(k-i)}(t) \\ - \sum_{i=p}^{n-1} a_i \sum_{j=0}^{i-p} \binom{p+j-1}{j} &\frac{n!}{(n-j)!} (b-t)^{n-j} y^{(i-j-p)}(t) + \int_t^b K_{B,k,y}(n,p,t,\tau) y(\tau) d\tau \Bigg] \\ &k = 1, ..., n-1 \end{aligned}$$

Chapter 3

3 Parameter and State Estimation of LTI SISO Systems on Finite Interval

3.1 Parameter estimation in the absence of noise

The works done by [28], [29], [33], [75], [30] and [41] show that the least squares method works well in the absence of noise or small additive noise. Here the properties of [41]'s method is explored.

3.1.1 Parameter Estimation [62], [41]

Consider a general fourth order system as below:

$$y^{(4)}(t) + a_3 y^{(3)}(t) + a_2 y^{(2)}(t) + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{18}$$

The goal is to estimate the parameter vector $a \triangleq (a_0, a_1, a_2, a_3)$ of the above system from the observation $y(\tau)$, $\tau \in [a, b]$ of the output function $y(\tau); \tau \in [a, b]$.

If exact matching is required only at a finite number n of discrete time points $t_j; j = 1, \cdots, n$ then the problem amounts to finding the optimal solution to the MSE method:

$$\min\{J(a) := \frac{1}{2n} \sum_{i=1}^{n} (y(t_i) - < y, K_{DS}(t_i, \cdot) >_2)^2 \mid \text{w.r.t. } \; a \in \mathbb{R}^3\} \tag{19}$$

$$= \min\{\frac{1}{2n} \sum_{i=1}^{n} \left[y(t_i) - \int_a^b K_{DS}(t_i, \tau) y(\tau) d\tau \right]^2 \mid \text{w.r.t. } \; a \in \mathbb{R}^3\}$$

A continuous time version of the above becomes

$$\min\{\frac{1}{2T} \int_a^b \left[y(t) - \int_a^b K_{DS}(t, \tau) y(\tau) d\tau \right]^2 dt \mid \text{w.r.t. } \; a \in \mathbb{R}^3\} \tag{20}$$

with $T := b - a$.

The cost function in (20) can be calculated as follows. $K_{DS}(t, \tau)$ is expressed as a scalar product of some partial kernels

$$K_{DS}(t, \tau) = K_v(t, \tau)^T a + k_{v4}(t, \tau) \quad \text{equivalently} \quad K_{DS}(t, \tau) = a^T K_v(t, \tau) + k_{v4}(t, \tau) \tag{21}$$

$$K_v(t, \tau)^T := [k_{v1}(t, \tau), k_{v2}(t, \tau), k_{v3}(t, \tau)]; \;\; a := [a_0, a_1, a_2, a_3]^T \tag{22}$$

Substituting the above into the cost of (20) yields (with $T := b - a$)

$$J(a) := \frac{1}{2T}\int_a^b \left[y(t) - \int_a^b K_{DS}(t,\tau)\ y(\tau)d\tau\right]^2 dt$$
$$= \frac{1}{T}\int_a^b \left[\frac{1}{2}y(t)^2 - y(t)\int_a^b K_{DS}(t,\tau)\ y(\tau)d\tau + \frac{1}{2}\left(\int_a^b K_{DS}(t,\tau)\ y(\tau)d\tau\right)^2\right] dt$$
$$= \frac{1}{T}\int_a^b \left[\frac{1}{2}y(t)^2 - y(t)\int_a^b K_{DS}(t,\tau)\ y(\tau)d\tau + \frac{1}{2}\int_a^b K_{DS}(t,\tau)\ y(\tau)d\tau \int_a^b K_{DS}(t,s)\ y(s)ds\right] dt$$
$$= \frac{1}{2T}\int_a^b y(t)^2 dt - \frac{1}{T}\int_a^b\int_a^b K_{DS}(t,\tau)\ y(\tau)y(t)\ d\tau\ dt$$
$$+ \frac{1}{2T}\int_a^b\int_a^b\int_a^b K_{DS}(t,\tau)K_{DS}(t,s)\ y(\tau)y(s)\ d\tau\ ds\ dt$$

Hence

$$J(a) =$$
$$= \frac{1}{2T}\int_a^b y(t)^2 dt - \frac{1}{T}\int_a^b\int_a^b [K_v(t,\tau)^T a + k_{v4}(t,\tau)]\ y(\tau)y(t)\ d\tau\ dt$$
$$+ \frac{1}{2T}\int_a^b\int_a^b\int_a^b [a^T K_v(t,\tau) + k_{v4}(t,\tau)][K_v(t,s)^T a + k_{v4}(t,s)]\ y(\tau)y(s)\ d\tau\ ds\ dt$$

$$= \frac{1}{2T}\int_a^b y(t)^2 dt - \frac{1}{T}\int_a^b\int_a^b K_v(t,\tau)^T\ y(\tau)y(t)\ d\tau\ dt\ a - \frac{1}{T}\int_a^b\int_a^b k_{v4}(t,\tau)\ y(\tau)y(t)\ d\tau\ dt$$
$$+ \frac{1}{2T}\ a^T\int_a^b\int_a^b\left[\int_a^b K_v(t,\tau)K_v(t,s)^T\ dt\right] y(\tau)y(s)\ d\tau\ ds\ a$$
$$+ \frac{1}{2T}\int_a^b\int_a^b\int_a^b k_{v4}(t,\tau)K_v(t,s)^T\ y(\tau)y(s)\ d\tau\ ds\ dt\ a$$
$$+ a^T\frac{1}{2T}\int_a^b\int_a^b\int_a^b K_v(t,\tau)k_{v4}(t,s)\ y(\tau)y(s)\ d\tau\ ds\ dt$$
$$+ \frac{1}{2T}\int_a^b\int_a^b\int_a^b k_{v4}(t,\tau)k_{v4}(t,s)]\ y(\tau)y(s)\ d\tau\ ds\ dt$$

Assembling terms

$$
\begin{aligned}
J(a) &= d + b^T a + \frac{1}{2} a^T C a \quad \text{with} \\
d &:= \left\{ \frac{1}{2T} \int_a^b y(t)^2 dt - \frac{1}{T} \int_a^b \int_a^b k_{v4}(t,\tau)\ y(\tau) y(t)\ d\tau\ dt \right. \\
&\qquad \left. + \frac{1}{2T} \int_a^b \int_a^b \int_a^b k_{v4}(t,\tau) k_{v4}(t,s)]\ y(\tau) y(s)\ d\tau\ ds\ dt \right\} \\
b^T &:= \left\{ -\frac{1}{T} \int_a^b \int_a^b K_v(t,\tau)^T\ y(\tau) y(t)\ d\tau\ dt \right. \\
&\qquad \left. + \frac{1}{2T} \int_a^b \int_a^b \int_a^b [K_v(t,s)^T k_{v4}(t,\tau) + K_v(t,\tau)^T k_{v4}(t,s)]\ y(\tau) y(s)\ d\tau\ ds\ dt \right\} \\
C &:= \frac{1}{T} \left\{ \int_a^b \int_a^b \left[\int_a^b K_v(t,\tau) K_v(t,s)^T\ dt \right] y(\tau) y(s)\ d\tau\ ds \right\}
\end{aligned}
$$

The standard quadratic cost yields a minimization problem w.r.t. parameter a that is solved globally and analytically;[61] :

$$
\begin{aligned}
& J(a) := d + b^T a + \frac{1}{2} a^T C a \\
& \min\{J(a) \mid a \in \mathbb{R}^3\} \quad \text{is attained globally and uniquely at} \\
& \hat{a} = -C^{-1} b; \quad \text{with minimum value } \ J(\hat{a}) = d - \frac{1}{2} b^T C^{-1} b
\end{aligned}
\tag{23}
$$

Also, it should be noted that the triple intergrals can be written as alternative integral products expressions which are easier to handle numerically

$$
\int_a^b \int_a^b \int_a^b k_{v4}(t,\tau) k_{v4}(t,s)]\ y(\tau) y(s)\ d\tau\ ds\ dt = \int_a^b \left(\int_a^b k_{v4}(t,\tau)\ y(\tau)\ d\tau \right)^2 dt
$$

$$
\begin{aligned}
&\int_a^b \int_a^b \int_a^b [K_v(t,s)^T k_{v4}(t,\tau) + K_v(t,\tau)^T k_{v4}(t,s)]\ y(\tau) y(s)\ d\tau\ ds\ dt \\
&\quad = \int_a^b \left(\int_a^b K_v(t,s)^T y(s)\ ds \right) \left(\int_a^b k_{v4}(t,\tau)\ y(\tau)\ d\tau \right)\ dt \\
&\qquad + \int_a^b \left(\int_a^b K_v(t,\tau)^T y(\tau)\ d\tau \right) \left(\int_a^b k_{v4}(t,s)\ y(s)\ ds \right)\ dt
\end{aligned}
$$

$$
\begin{aligned}
&\int_a^b \int_a^b \left[\int_a^b K_v(t,\tau) K_v(t,s)^T\ dt \right] y(\tau) y(s)\ d\tau\ ds \\
&\quad = \int_a^b \left[\int_a^b K_v(t,\tau)\ y(\tau)\ d\tau \right] \left[\int_a^b K_v(t,s)\ y(s)\ ds \right]^T\ dt
\end{aligned}
$$

The discrete cost (19) can be computed similarly, as follows:

$$
\begin{aligned}
J(a) &:= \frac{1}{2n}\sum_{i=1}^{n}\left[y(t_i) - \int_a^b K_{DS}(t_i,\tau)\ y(\tau)d\tau\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}y(t_i)^2 - y(t_i)\int_a^b K_{DS}(t_i,\tau)\ y(\tau)d\tau + \frac{1}{2}\left(\int_a^b K_{DS}(t_i,\tau)\ y(\tau)d\tau\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}y(t_i)^2 - y(t_i)\int_a^b K_{DS}(t_i,\tau)\ y(\tau)d\tau + \frac{1}{2}\int_a^b K_{DS}(t_i,\tau)\ y(\tau)d\tau\int_a^b K_{DS}(t_i,s)\ y(s)ds\right] \\
&= \frac{1}{2n}\sum_{i=1}^{n}y(t_i)^2 - \frac{1}{n}\sum_{i=1}^{n}y(t_i)\int_a^b K_{DS}(t_i,\tau)\ y(\tau)\ d\tau \\
&\qquad + \frac{1}{2n}\sum_{i=1}^{n}\int_a^b\int_a^b K_{DS}(t_i,\tau)K_{DS}(t_i,s)\ y(\tau)y(s)\ d\tau\ ds
\end{aligned}
$$

Expanding the kernels yields

$$
\begin{aligned}
J(a) &= \\
&= \frac{1}{2n}\sum_{i=1}^{n}y(t_i)^2 - \frac{1}{n}\sum_{i=1}^{n}\int_a^b [K_v(t_i,\tau)a + k_{v4}(t_i,\tau)]\ y(\tau)y(t_i)\ d\tau \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\int_a^b\int_a^b [a^T K_v(t_i,\tau) + k_{v4}(t_i,\tau)][K_v(t_i,s)^T a + k_{v4}(t_i,s)]\ y(\tau)y(s)\ d\tau\ ds \\
&= \frac{1}{2n}\sum_{i=1}^{n}y(t_i)^2 dt - \frac{1}{n}\sum_{i=1}^{n}\int_a^b K_v(t_i,\tau)^T\ y(\tau)y(t_i)\ d\tau\ a - \frac{1}{n}\sum_{i=1}^{n}\int_a^b k_{v4}(t_i,\tau)\ y(\tau)y(t_i)\ d\tau \\
&+ \frac{1}{2n}a^T\sum_{i=1}^{n}\int_a^b\int_a^b K_v(t_i,\tau)K_v(t_i,s)^T y(\tau)y(s)\ d\tau\ ds\ a \\
&+ \frac{1}{2n}\sum_{i=1}^{n}\int_a^b\int_a^b k_{v4}(t_i,\tau)K_v(t_i,s)^T\ y(\tau)y(s)\ d\tau\ ds\ a \\
&+ a^T\frac{1}{2n}\sum_{i=1}^{n}\int_a^b\int_a^b K_v(t_i,\tau)k_{v4}(t_i,s)\ y(\tau)y(s)\ d\tau\ ds \\
&+ \frac{1}{2n}\sum_{i=1}^{n}\int_a^b\int_a^b k_{v4}(t_i,\tau)k_{v4}(t_i,s)]\ y(\tau)y(s)\ d\tau\ ds
\end{aligned}
$$

Assembling terms again delivers a standard quadratic

$$J(a) = d + b^T a + \frac{1}{2} a^T C a \quad \text{with}$$

$$d := \left\{ \frac{1}{2n} \sum_{i=1}^{n} y(t_i)^2 - \frac{1}{n} \sum_{i=1}^{n} \int_a^b k_{v4}(t_i, \tau)\ y(\tau) y(t_i)\ d\tau \right.$$
$$\left. + \frac{1}{2n} \sum_{i=1}^{n} \int_a^b \int_a^b k_{v4}(t_i, \tau) k_{v4}(t_i, s)]\ y(\tau) y(s)\ d\tau\ ds \right\}$$

$$b^T := \left\{ -\frac{1}{n} \sum_{i=1}^{n} \int_a^b K_v(t_i, \tau)^T\ y(\tau) y(t_i)\ d\tau \right.$$
$$\left. + \frac{1}{2n} \sum_{i=1}^{n} \int_a^b \int_a^b [K_v(t_i, s)^T k_{v4}(t_i, \tau) + K_v(t_i, \tau)^T k_{v4}(t_i, s)]\ y(\tau) y(s)\ d\tau\ ds \right\}$$

$$C := \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_a^b \int_a^b K_v(t_i, \tau) K_v(t_i, s)^T\ y(\tau) y(s)\ d\tau\ ds \right\}$$

$$\begin{aligned} &\min\{J(a) \mid a \in \mathbb{R}^3\} \quad \text{is attained globally and uniquely at} \\ &\hat{a} = -C^{-1} b; \quad \text{with minimum value } \ J(\hat{a}) = d - \frac{1}{2} b^T C^{-1} b \end{aligned} \tag{24}$$

The double integrals above can again be written in terms of single integrals

$$\int_a^b \int_a^b k_{v4}(t_i, \tau) k_{v4}(t_i, s)]\ y(\tau) y(s)\ d\tau\ ds = \left(\int_a^b k_{v4}(t_i, \tau)\ y(\tau)\ d\tau \right)^2$$

$$\begin{aligned} \int_a^b \int_a^b &[K_v(t_i, s)^T k_{v4}(t_i, \tau) + K_v(t_i, \tau)^T k_{v4}(t_i, s)]\ y(\tau) y(s)\ d\tau\ ds \\ &= \left(\int_a^b K_v(t_i, s)^T y(s)\ ds \right) \left(\int_a^b k_{v4}(t_i, \tau)\ y(\tau)\ d\tau \right) \\ &+ \left(\int_a^b K_v(t_i, \tau)^T y(\tau)\ d\tau \right) \left(\int_a^b k_{v4}(t_i, s)\ y(s)\ ds \right) \end{aligned}$$

$$\begin{aligned} \int_a^b \int_a^b &K_v(t_i, \tau) K_v(t_i, s)^T y(\tau) y(s)\ d\tau\ ds \\ &= \left[\int_a^b K_v(t_i, \tau)\ y(\tau)\ d\tau \right] \left[\int_a^b K_v(t_i, s)\ y(s)\ ds \right]^T \end{aligned}$$

3.2 Output estimation by projection [41]

The system output after estimating the parameters can be reconstructed using

$$\tilde{y}(t) = \int_a^b K_{DS}(t_i, \tau) z(\tau) d\tau \tag{25}$$

where $z(\tau)$ is the measured signal.

However, as a more precise alternative to the above, the system output can be smoothed/reconstructed from a noisy measurement by direct orthogonal projection onto the subspace spanned by the fundamental solutions of the characteristic equation (7) of the system. This is because every solution of the characteristic equation with the already identified parameter vector a satisfies the reproducing property so, the projection onto the space of fundamental solutions will be the noise free trajectory of the system. The fundamental solutions of the characteristic equation are obtained by the direct integration of (7). This is performed as follows.

We select n independent vectors as initial conditions for the homogeneous LTI system in (7). These initial conditions can be taken as the vectors of the canonical basis in $\mathbb{R}^n$ i.e,

$$\begin{aligned} e_1 &= [1, 0, ..., 0] \\ e_2 &= [0, 1, ..., 0] \\ &\cdots \qquad (26) \\ e_n &= [0, 0, ..., 1] \qquad (27) \end{aligned}$$

The system equation (7) is then solved for each individual initial condition yielding solutions

$$\overline{y_i}(a) = e_i \qquad i = 1, \cdots, n.$$

It is an elementary fact from the theory of ordinary differential equations that any solution of the system (7) with any initial condition is a linear combination of such fundamental solutions $\overline{y_i}$. Hence we search for the coefficients of this linear combination so that the resulting function is the closest to the output measurement data. Closest solution is found in terms of an orthogonal projection onto span of S^a.

$$S^a = \text{span}\left\{\overline{y}_i(\cdot), i = 1, \cdots, n\right\}$$

This is best done by orthonormalizing the set of fundamental solutions.The projection of a measured noisy signal $z(\cdot) \in L^2[a, b]$ into S^a is given as,

$$y_E(\cdot) \triangleq \arg\min\left\{\|z - y\|_2^2 \,| y \in S^a\right\} \tag{28}$$

. We seek,

$$\tilde{y} = \sum_{i=1}^{n} \hat{c}_i \overline{y}_i \tag{29}$$

,
As $\hat{y}$ is a linear combination, the optimality conditions in (28) is achieved if and only if,

$$\left\langle z|\overline{y}_j\right\rangle_2 = \sum_{i=1}^{n} \hat{c}_i \left\langle \overline{y}_i|\overline{y}_j\right\rangle_2 j = 1, \cdots, n \tag{30}$$

which can be written in a matrix form as:

$$\begin{aligned} &v = G(\overline{y})\hat{c}; \quad G(\overline{y}) \triangleq \text{mat}\left\{\left\langle \overline{y}_i|\overline{y}_j\right\rangle_2\right\}_{i,j=1}^{n} \\ &v \triangleq \text{vec}\left\{\langle z|\overline{y}_i\rangle_2\right\}_{i=1}^{n}; \hat{c} \triangleq \text{vec}\left\{\hat{c}_i\right\}_{i=1}^{n} \end{aligned} \tag{31}$$

G is called the Gram matrix for vectors in span S^a and is invertible because it is known that all fundamental solutions are linearly independent, from the theory of differential equations.

$$\hat{c} = G^{-1}(\overline{y})v \tag{32}$$

$\hat{y}$, the estimated output is thus obtained from (29).

3.2.1 Reconstruction of the output derivatives

Once we obtain $\hat{y}$, the estimated output, the derivatives can be reconstructed using the formula[26],

$$y^{(i)}(t) = \int_a^b K_{DS}^i(t,\tau)\tilde{y}(\tau)\mathrm{d}\tau \tag{33}$$

where, K_{DS}^i are the kernel representation for output derivatives. In this thesis we consider $i = 1, 2$, and 3. The formulae for kernel representation of output derivatives are developed.

3.2.2 Shortcoming of Simple Least Square Method

However, [28], [29], [33], [75], [30], and [41] fail to work for low signal to noise ratio. This issue is handled by using more data points. Despite the increase in sample size, the parameter estimation is not accurate. In the experiment section, it has been shown that our method is much faster(by the factor of 10) and more accurate in a high noise environment. Lastly, the assumption that the dimensions of the system are known limits the usage of the previous works, [28], [29], [33], [75], [30] and [41].

Chapter 4

4 Parameter Estimation with Error in the Variable

The previous section summarized the parameter estimation carried out in [41]. The algorithm of [41] does not take into account, the error in independent variable when optimizing the cost function. The method suggested by [41] works well with large Signal to Noise Ratio, SNR, however, when the noise is large, the algorithm under-performs. The work of [41] tries to avoid this shortcoming by increasing the number of samples, which in return makes the estimation method slow. The algorithm of [41] does not consider the noise in y realization in the optimization method. This section first will try to explore the issue with the noise in the independent variable in the current kernel system. It hence tries to explore simple techniques to remove the noise from the independent variable. In the result section, we have compared our method with [41]'s method in terms of accuracy and computation speed. In both accuracy and computation speed, the current method which considers the error in the variable, outperforms the method of [41].

4.1 Identifiability of homogeneous LTI systems from a single realization of a measured output[25]

Referring to Definition and Theorem from [76], a homogeneous LTI system such as

$$\dot{x}(t) = Ax(t); \tag{34}$$

$$y(t) = Cx(t); \tag{35}$$

$$x \in \mathbb{R}^n$$

$$x(0) = b \tag{36}$$

is identifiable from a single noise-free realization of its output trajectory y on the interval $[0, \infty)$ under precisely stated conditions, which are, however, difficult to verify computationally. The definition of identifiability is stated in a form equivalent to that found in [76] as the following:

Definition: Model (34) *is globally identifiable from* b *if and only if the functional mapping* $A \mapsto y(\cdot; A, b)$ *is injective on* $\mathbb{R}^n$ *where* b *is fixed and* $y(\cdot; A, b)$ *denotes the output orbit of* (34) *through* b.

Theorem[25] Model (34) *is globally identifiable from* b *if and only if the evolution of the output orbit of* (34) *is not confined to a proper subspace of* $\mathbb{R}^n$.

The above criterion has limited use for two reasons: it assumes to infinite time horizons $[0, \infty)$ but, more importantly, it requires the output trajectory to be known exactly.

At this point the kernel system representation is again found helpful in that theoretical identifiability criterion from a single *noisy realization* y_M of the output trajectory on a finite interval $[a, b]$ can be stated in terms of linear independence of the functions involving the component kernel expressions in (25), namely functions; see [62]:

$$f_i(t) := \int_a^b K_{DS(i),y}(t,\tau) y_M(\tau) d\tau \ ; \ i = 1, \cdots, n \tag{37}$$

Linear independence of the above set can be readily checked by establishing non-singularity of the Wronskian matrix for (37) at some point in the interval $[a, b]$.

4.2 Parametric estimation as a least squares problem[25]

It was found in [25] that in the presence of large measurement noise, assumed to be white Gaussian and additive, the reproducing property of the regression equation (25) fails to hold along an inexact output trajectory. The work of [25] suggests that it must thus be suitably replaced leading to a stochastic regression problem. First, [25] assumes that the stochastic output measurement process on a suitable space such as $L^2(\Omega, \mathcal{F}, \mathbb{P})$ where y_M to be adapted to the natural filtration of the Wiener process W on $[a, b]$ is

$$y_M(t, \omega) = y_T(t) + \eta^\sigma(t, \omega) \ ; \quad t \in [a, b] \tag{38}$$

where η^σ signifies the white noise with constant variance σ^2 and where y is the true system output. Without adhering to any particular realization of the measurement process this yields a random kernel expression presented by equation (25) which is better written using the proper stochastic nomenclature as shown in [25]:

$$\int_a^b K_{DS,y}(t, \tau) y_M(\tau) \ d\tau = \int_a^b K_{DS,y}(t, \tau) y_T(\tau) \ d\tau + \int_a^b K_{DS,y}(t, \tau) \ dW^\sigma(\tau) \tag{39}$$

Here, W^σ is the Wiener process with intensity σ so that, informally, $\eta^\sigma(t)dt = \sigma dW(t)$ with W as the standard Brownian motion. It follows that the following equality is valid

$$y_M(t) = \int_a^b K_{DS,y}(t, \tau) y_M(\tau) \ d\tau + e(t) \tag{40}$$

$$\text{with} \ \ e(t) := \eta^\sigma(t) - \int_a^b K_{DS,y}(t, \tau) \ dW^\sigma(\tau) \tag{41}$$

It is noted that the random error variable e is dependent on the unknown system parameters $a_i, i = 0, \cdots, n-1$ while the stochastic regression equation

$$y_M(t) = \sum_{i=0}^{n} a_i \int_a^b K_{DS(i),y}(t, \tau) y_M(\tau) d\tau + e(t) \tag{42}$$

has the random regressor vector

$$\left[\int_a^b K_{DS(0),y}(t, \tau) y_M(\tau) d\tau, \cdots, \int_a^b K_{DS(n),y}(t, \tau) y_M(\tau) d\tau \right]^T \tag{43}$$

Thus [25] shows that the assumptions of the Gauss-Markov Theorem are violated in the linear regression problem (42) because the random regressor is correlated with a regression error. The above regression is thus a typical ‘error-in-the-variable’ problem. The Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists [68].

As the kernels of Theorem 2.2 are linear in the unknown system coefficients, the reproducing property (for homogeneous systems) is first re-written to bring out this fact while omitting the obvious dependence of the kernels on n. The term $\alpha_{ab}^{-1}(t)$ is subsumed in K_{DS} for convenience of notation.

$$y(t) = \int_a^b K_{DS,y}(t,\tau)y(\tau)\ d\tau \tag{44}$$

$$= \sum_{i=0}^{n} a_i \int_a^b K_{DS(i),y}(t,\tau)y(\tau)d\tau \tag{45}$$

Where the $K_{DS(i),y}; i = 0, \cdots, n$ are 'component kernels' of $K_{DS,y}$ that post-multiply the coefficients $a_i; i = 0, \cdots, n-1$, with $\beta_n = 1$ for convenience of notation. In a noise-free deterministic setting it can be solved using least squares error minimization provided adequate identifiability assumptions are met and the output is measured without error.

Recall equation (21), Kernels in Theorem 2.2 can be simplified if we assign $a_4 = 1$:

$$K_{DS}(t,\tau) = K_v(t,\tau)^T a \ \text{ equivalently } \ K_{DS}(t,\tau) = a^T K_v(t,\tau) \tag{46}$$

$$K_v(t,\tau)^T := [k_{v1}(t,\tau), k_{v2}(t,\tau), k_{v3}(t,\tau), k_{v4}(t,\tau)];\ a := [a_0, a_1, a_2, a_3, 1]^T \tag{47}$$

Hence, the equation (44) can be rewritten as:

$$y(t) = \int_a^b K_{DS,y}(t,\tau)y(\tau)\ d\tau \tag{48}$$

$$= \int_a^b K_v(t,\tau)^T a \cdot y(\tau)\ d\tau \tag{49}$$

Since the vector a with respect to τ is constant, it can be taken outside of the integral, thus the equation (49) is simplified to:

$$y(t) = \int_a^b K_v(t,\tau)^T y(\tau)\ d\tau \cdot a \tag{50}$$

For simplicity let us define

$$g(t) = \int_a^b K_v(t,\tau)^T y(\tau) d\tau \tag{51}$$

Hence the equation (50) can be rewritten as

$$y(t) = g(t) \cdot a \tag{52}$$

However, $y(t)$ is measured with noise, Let y_M be the measured output trajectory with noise.

$$y_M(t) = y(t) + \epsilon(t) \tag{53}$$

We assume the errors $\epsilon(t)$ is white noise with zero mean and finite variance $\text{var}(\epsilon(t)) = \sigma^2$.

Then the (52) can be written as:

$$y_M(t) = g(t) \cdot a + \epsilon(t) \tag{54}$$

In order to consider the noise in measurements, the definition of $g(t)$ needs to be updated to:

$$g(t) = \int_a^b K_v(t,\tau)^T y_M(\tau) d\tau \tag{55}$$

In the matrix form, equation (54) can be written as

$$Y_M = G \cdot a + \Sigma \tag{56}$$

where Y_M is the vector of all $y_M(t)$ observation , G is the vector of all $g(t)$, and Σ is the vector of all the noise terms $\epsilon(t)$, $\forall t \in [a,b]$,

$$Y_M = \begin{bmatrix} y_M(t=a) \\ \vdots \\ y_M(t=b) \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} g(t=a) \\ \vdots \\ g(t=b) \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \epsilon(t=a) \\ \vdots \\ \epsilon(t=b) \end{bmatrix} \tag{57}$$

Thus the least squares estimates of the coefficients can be found by using:

$$\widehat{a} = (G^T G)^{-1} G^T Y_M \tag{58}$$

However, G has been measured with errors, any estimation based on the standard assumption leads to inconsistent estimates. This means that the parameter estimates of a do not tend to the true values even in very large samples. This leads to an underestimate of the coefficient, known as the attenuation bias. Attenuation bias is the biasing of the regression slope towards zero caused by errors in the independent variable. This is clearly the case in [41] case, as shown in the experiment section here.

4.3 Error In The Variable

The work of [25] shows that the assumptions for Gauss–Markov theorem are violated.

It is well known that the presence of errors-in-variables induces an asymptotic bias in least squared regression estimates which is proportional to the signal-to-noise ratio in the observed regressand. In [25] the suggested method to eliminate estimation bias is to use *Instrumental Variables* (IV) (see [86]), in the normal equations that deliver the optimal estimates. The IV method has multiple applications; refer to [57], [83], [85], [10], [24], and [11].
However, unlike the work done by [25], we try to find the source of the noise and eliminate it early in the process. Here, a simple localized non-parametric filtering is applied to the signal to remove the white noise.

Considering the equation (54); the noise in the $g(t)$ term is the source of bias in the estimate. Consider the definition of the $g(t)$

$$g(t) = \int_a^b K_v(t,\tau)^T y_M(\tau) d\tau \tag{59}$$

If the y_M is observed with noise then the equation is as follows:

$$g(t) = \int_a^b K_v(t,\tau)^T (y(\tau) + \epsilon) d\tau \tag{60}$$

One of the most used method to handle the error in variable is to estimate $g(t)$ using local non-parametric methods. Consider $\hat{g}(t)$ to be the estimate of $g(t)$. To estimate $\hat{g}(t)$,the estimate of y, $\hat{y}$, has to be estimated and the $\hat{g}(t)$ can be calculated as follows.

$$\hat{g}(t) = \int_a^b K_v(t,\tau)^T \hat{y}(\tau) d\tau \tag{61}$$

To predict $\hat{y}$ from y_M a non-parametric local regression is considered.

4.4 Local Regression[39]

Local regression is used to model a relation between a predictor variable and response variable. Consider the fixed design model. The model has a form:

$$Y_i = f(x_i) + \varepsilon_i \tag{62}$$

where $f(x_i)$ is an unknown function and ε_i is an error term, representing random errors in the observations or variability from sources not included in the x_i.

We assume the errors ε_i are white noise with mean zero and finite variance $\text{var}(\varepsilon_i) = \sigma^2$.

We assume the function f is continuous, every continuous function defined on a closed interval [a, b] can be uniformly approximated as closely as desired by a polynomial function, according to Taylor's theorem.

A general ordinary differential equation model can be written as the following. The process $y(t)$ is usually measured with noise and we observe.

$$y_M(t) = y(t) + \epsilon(t) \tag{63}$$

Where $y_M(t)$ is the measured output while $y(t)$ is the true value of the system at time t. The $\epsilon(t)$ represents the white noise of the system at time t.
The non-parametric smoothing techniques were applied to observed process to avoid numerically solving the differential equations when estimating the true value of $y(t)$ from noisy measurement $y_M(t)$.

The goal is to estimate $y(t)$ from noisy $y_M(t)$. By inspecting the equation (63), it is evident that is it similar to (62). Thus, the estimate of true value of y can be found as follows:

$$y_M(t) = \hat{y}(t) + \epsilon(t) \tag{64}$$

Where $\hat{y}$ is the smooth value of y.

4.4.1 Locally Estimated Scatter-plot Smoothing(LOESS)[39]

Local polynomial regression is a generalization of moving average and polynomial regression. Here we consider LOESS method which is one of the most used methods of local polynomial regression. The advantages of LOESS are:

- It does not require the specification of a function to fit a model to all of the data in the sample.
- Few hyper parameters are required such as; the degree of the local polynomial.

- LOESS is very flexible, making it ideal for modeling complex processes for which no theoretical models exist.

The first step in LOESS is to define a weight function. For computational and theoretical purposes we will define this weight function so that only values within a *smoothing window* $[t-\delta, t+\delta]$, where δ is a small real number, will be considered in the estimate of $y(t)$. The δ, window size, is the hyper parameter which is re-calibrated during the experiments.

This is easily achieved by considering weight functions that are 0 outside of $[-1, 1]$.[39] For example Tukey's tri-weight function

$$W(u) = \begin{cases} (1-|u|^3)^3 & |u| \leq 1 \\ 0 & |u| > 1. \end{cases}$$

The weight sequence is then easily defined by

$$w_i(y(t)) = W\left(\frac{y(t_i) - y(t_0)}{\delta}\right)$$

Within the smoothing window, $y(t)$ is approximated by a quadratic polynomial.

$$y(t) \approx \beta_0 + \beta_1(y_M(t_i) - y_M(t_0)) + \frac{1}{2}\beta_2 y_M(t_i) - y_M(t_0))^2$$

for $t_i \in [t-\delta, t+\delta]$.

To obtain the local regression estimate $\hat{f}(x_0)$ we simply find the $\beta = (\beta_0, \beta_1, \beta_2)'$ that minimizes

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^3} \sum_{i=1}^{n} w_i(y_M(t_0))[y_M - \{\beta_0 + \beta_1(y_M(t) - y_M(t_0)) + \frac{1}{2}\beta_2(y_M(t) - y_M(t_0))\}]^2$$

and define

$$\hat{y}(t) = \hat{\beta}_0 + \hat{\beta}_1(y_M(t_i) - y_M(t_0)) + \frac{1}{2}\hat{\beta}_2 y_M(t_i) - y_M(t_0))^2$$

We apply this method throughout the observed signal to get $\hat{y}(t)\ \forall t \in [a, b]$

Chapter 5

5 Optimization with Penalty

In the previous work done by [41], the parameter are estimated by optimizing the cost function using Mean Squared Error (MSE). However, in the presence of large noise, or small signal-to-noise ratio, MSE tends to over-fit the data, given that our model is noisy, MSE may try to fit to the noise term. To overcome this issue, regularization can be used to penalty the estimates of the parameters when they tends to over-fit or when the cost function fitting tries to fit to the noise term.

Consider our developed equation:

$$y_M(t) = \hat{g}(t) \cdot a + \epsilon(t) \tag{65}$$

Where

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{66}$$

In order to estimate the values of a, originally the following cost function was studied by[41],

$$\hat{a} = \underset{a \in R^p}{\operatorname{argmin}} \; \|Y - \hat{G} \cdot a\|_2^2 \tag{67}$$

where Y and G are vectors defined at (57). The $\|\|$ refers to the Euclidean norm. Euclidean norm is strongly related with the Euclidean distance, and equals the square root of the inner product of a vector with itself. p is the order of the system.

As argued by[34], learning from data can be viewed as a multivariate function approximation from sparse data. The value of $\hat{g}(t)$ is estimated from highly noisy observations. It is well-known that such a problem is ill-posed as there exists an infinity of functions that pass perfectly through the data. One way to transform this problem into a well-posed one is to assume that the function$f$ presents some smoothness properties and therefor, the problem becomes a variational problem of finding the function $\hat{f}$ that minimizes the equation (68) [78]: Let define f as

$$f(t) = \hat{g}(t) \cdot a$$

$$L[f] = \frac{1}{b-a} \sum_{t=a}^{b} C(y(t) - f(t)) + \lambda \Omega[f] \tag{68}$$

where λ is a positive number, n is the sampling size, C is a cost function which determines how differences between $f(t)$ and $y(t)$ should be penalized and $\Omega[f]$ is a functional which denotes the prior information on the function f [69].

The hyper-parameter λ balances the trade-off between fitness of f to the data and smoothness of f. This regularization principle leads to different approximation schemes depending on the cost function C.

Using Lasso, Least absolute shrinkage and selection operator [77], The equation (68) leads to

$$\hat{a}^{\text{lasso}} = \underset{a \in R^p}{\text{argmin}} \; \|Y - \hat{G} \cdot a\|_2^2 + \lambda \sum_{j=1}^{p} |a_j| \tag{69}$$

$$= \underset{a \in R^p}{\text{argmin}} \; \underbrace{\|Y - \hat{G} \cdot a\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|a\|_1}_{\text{Penalty}} \tag{70}$$

where p is the order of a matrix. and λ is the a positive number.

LASSO has been developed as a tool to find sparse solutions of the linear regression problem. It has been used extensively in an expanding field of applications from statistics to estimation scenarios with remarkably good results. As a robust approximation of the well known Maximum Likelihood (ML) estimator, the LASSO can be realized robustly and efficiently.[40]

Since the regularization term in LASSO regression has absolute value operation, the well known programming libraries for optimization has a slower processing time. In order to increase the speed of the optimization step, the regularization term is updated to squared term so that the derivative of the term exists.

The original equation (68) leads to

$$\hat{a}^{\text{ridge}} = \underset{a \in R^p}{\text{argmin}} \; \|Y - \hat{G} \cdot a\|_2^2 + \lambda \sum_{j=1}^{p} (a_j)^2 \tag{71}$$

$$= \underset{a \in R^p}{\text{argmin}} \; \underbrace{\|Y - \hat{G} \cdot a\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|a\|_2^2}_{\text{Penalty}} \tag{72}$$

This lost function is called ridge Regression. In this work, the ridge Regression is preferred over LASSO due to the computational advantages it has. The value of λ is selected by K-fold Cross validation.

5.1 Cross validation[70]

The parameter λ in equation (68) can be seen as a hyper-parameter. Almost every machine learning algorithm comes with large number of settings that has to be done by researchers and practitioners. These parameters to tune, the so-called hyper-parameters, control the behavior of machine learning algorithms when optimizing for performance, finding the right balance between bias and variance. Hyper-parameter tuning for performance optimization is an art in itself, and there are no hard-and-fast rules that guarantee best performance on a given data set. This section focuses on different methods of cross-validation for model evaluation and model selection. It covers cross-validation techniques to rank models from several hyper-parameter configurations and estimate how well these generalize to independent future data points (estimates).

An example of a hyper-parameter is the value of a regularization parameter λ such as the lambda-term in L2-regularized, ridge, regression. Changing the hyper-parameter values when running a learning algorithm over a training set may result in different models. The process of finding the best-performing model from a set of models that were produced by different hyper-parameter settings is called *model selection*. The process of hyper-parameter tuning (or hyper-parameter optimization) and model selection can be regarded as a meta-optimization task. While the learning algorithm optimizes an objective function on the training set, hyper-parameter optimization is yet another task on top of it; the goal is to optimize a performance metric such as accuracy. After the tuning stage, selecting a model based on the test set performance seems to be a reasonable approach. However, reusing the test set multiple times would introduce a bias in the final performance estimate. It also results in overly optimistic estimates of the generalization performance. To avoid this problem, three-way split is used. It first divides the data points into a training, validation, and test data points. Having a training-validation pair for hyper-parameter tuning and model selections would make the test set *independent* for model evaluation.

Throughout the experiments, the MATLAB library for cross validation is used to select the best value for λ.

Chapter 6

6 Order Selection

Unlike the prior works [41] and [27] which include a strong assumption that the order of the system is known, this work makes no assumptions on the order of the system.

One of the challenges in differential equations estimation is ”guessing” the order of the system. Not knowing the dimensions of a system, makes the parameter estimation and forecasting of the system problematic. We investigate the methods to pre-guess the order of the system from the observed trajectory.

In the context of this thesis, we suggest using different potential kernels with various dimensions, e.g. kernels with order 4, 5 or 6. Subsequently using them, we estimate the parameters of same dimensions systems. Lastly, we compare the estimations generated by each kernels to each other and select the most appropriate estimate. To evaluate and compare them to each other, we initially considered Sum of Squared Error *SSE* and selecting the one with the least *SSE*. However, the *MSE* method's principle invariably leads to choosing the highest possible dimension. Therefore, it cannot be the right formalization of the intuitive notion of choosing the *right* dimension [72]. The higher order estimates tend to have a smaller SSE, as shown in the experiment. This issue can be handled by utilizing alternative method for comparing the kernels of different orders. One potential method of overcoming this problem is to use information criteria.

6.1 Akaike Information Criterion (AIC) [1] and the corrected (AICc)

Akaike [1] introduced the first information criterion, which is now known as the Akaike Information Criterion (AIC). The logic of information criteria is grounded in information theory. More specifically, AIC was based on Kullback-Leibler (K-L) distance [50], which connected information theory to random variable distributions. AIC is intended to be an approximately unbiased estimator of the Kullback-Leibler discrepancy between the candidate and true models. The AIC equation using residual sum of squares is:

$$AIC = n \cdot lg(SSE/n) + 2 \cdot k \tag{73}$$

Where n is the number of training cases, k is the number of parameters (weights and biases), and SSE is the Sum of Squared Errors for the training set. lg refers to binary logarithms or to base 2. For our problem the SSE is defined as:

$$SSE = \sum_{t=a}^{b} (y(t) - \tilde{y}(t))^2 \tag{74}$$

Given a set of models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit and also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages over-fitting, because increasing the number of parameters in the model almost always improves the goodness of the fit. For linear regression, AICc is exactly unbiased, assuming that the candidate family of models includes the true model. The AICc [38] is used when the sample size is small. In our case the sample size is large.

AIC is a number that is helpful for comparing models as it includes measures of both how well the model fits the data and how complex the model is.

We use AIC to select the order of our system. First we have an initial guess for the range of the order of the system. Thus we apply our three-step parameter estimation for all of the selected orders. Lastly we decide the order by comparing the AIC values of each model. We have shown the method in the experiment section.

6.2 Bayesian Information Criterion (BIC)

Another method for model (order) selection is the Bayesian information criterion (BIC) or Schwarz information criterion. In Bayesian model selection, a prior probability is set for each model, and prior distributions are also set for the nonzero coefficients in each model. If we assume that one and only one model, along with its associated priors, is true, we can use Bayes' Theorem to find the posterior probability of each model given the data[17].

The BIC equation using residual sum of squares is:

$$BIC = n \cdot lg(SSE/n) + k \cdot ln(n) \tag{75}$$

Where n is the number of training cases, k is the number of parameters (weights and biases), and SSE is the Sum of Squared Errors for the training set. lg refers to binary logarithms or to base 2.

Unlike AIC, BIC is consistent, which means that as the size of the sample increases, the criteria will select a true model of finite dimension as long as it is included in the set of candidate models [80]. BIC is sometimes preferred over AIC because BIC is "consistent."

6.3 Comparing models using information criteria

Once the set of candidate models are defined, model selection using information criteria is a three-step process. The candidate models here represent the different order of kernels that was fitted to the data.

The first step is to fit each candidate model to the same data, making sure that any the transformation to the outcome variable is maintained across all candidate models.

The second step is to obtain the desired information criteria for each model. For our experiments, BIC formula has been employed. In (75) each candidate model contributes uniquely to an information criterion value through its estimated log likelihood and the number of parameters used in the model. The sample size when computing sample size-dependent information criteria is the same across candidate models.

The third step is to compare the candidate models by ranking them based on the information criteria being used. The model with the lowest value is considered to be the "best" model [80].

Chapter 7

7 Results and Discussion

We employed our method to various scenarios. To begin, we examine the unstable system with different noise variances. Next we compare our method to the results of [41].

7.1 Case Study: Unstable System with an unknown order

An unstable 4-th order system is considered

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & -5 & -5 & 0 \end{bmatrix} x \quad ; \quad y = x_1 \quad ; \quad x(0) = [0, 0, 0, 1] \tag{76}$$

with characteristic equation

$$y^{(4)}(t) + a_3 y^{(3)}(t) + a_2 y^{(2)}(t) + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{77}$$

where the nominal values of parameters are

$$a_0 = 1 \quad , \quad a_1 = 5 \quad , \quad a_2 = 5 \quad , \quad a_3 = 0 \tag{78}$$

The measured realization of the output y_M is obtained by adding white noise to the true trajectory.

$$y_M = y_T + \epsilon \tag{79}$$

where y is the true value of system's output and ϵ is a white noise,

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2) \tag{80}$$

where $\mu = 0$ and the σ is a fixed constant. We have tested our MATLAB program for the following noise's variances(σ):

σ	signal-to-noise ratio(dB)
1	-8.9652
1.5	-12.5224
2	-15.00821

To begin, we will explore the case for which $\sigma = 1$ in detail. The results for other cases are attached.

7.1.1 Noise $\sigma = 1$

Figure 3 shows y_M versus y, where the true y is displayed in red and the noisy y in blue. The sample size is $N = 5000$. We assumed the order of the system is between 2 to 6. The results of estimation for each order is shown in the Table 2. Since the ridge optimization is used, the hyper-parameter tuning for λ is required. We take advantage of cross-validation function in MATLAB to handle this tuning. The tuning of λ is done for all potential models. For each of the *expected* system order, the parameters are estimated. Subsequently we use the fundamental solution method to reconstruct the output y. Thus using our order selection criteria, we then choose the best order, and the derivatives are then calculated and plotted. The standard Root Mean Square Deviation (RMSD) was employed to assess proximity of the nominal and estimated/reconstructed trajectories y and y_E. The y_E refer to the estimation of y using our estimation method.

Order	Values	a_0	a_1	a_2	a_3	a_4	a_5	RMSD	BIC
	True	1	5	5	0	-	-	-	-
2	Estimated	1.315	-0.32	-	-	-	-	9.63e-2	-3.0698e5
3	Estimated	0.113	6.104	-1.003	-	-	-	1.3e-3	-3.09618e5
4	Estimated	0.892	5.120	4.987	0	-	-	**1.6e-4**	**-3.0963e5**
5	Estimated	-1.000	0	5.629	4.688	-0.001	-	2.3e-4	-3.0962e5
6	Estimated	0	-39.737	0	0	3.778	0	3.2e-4	-3.0961e5

Table 2: Estimated parameter values with noise of 1 SD (SNR -8.9652 dB)

As the Table 2 shows; the smallest value for BIC is associated with the fourth order. It estimates the order correctly.

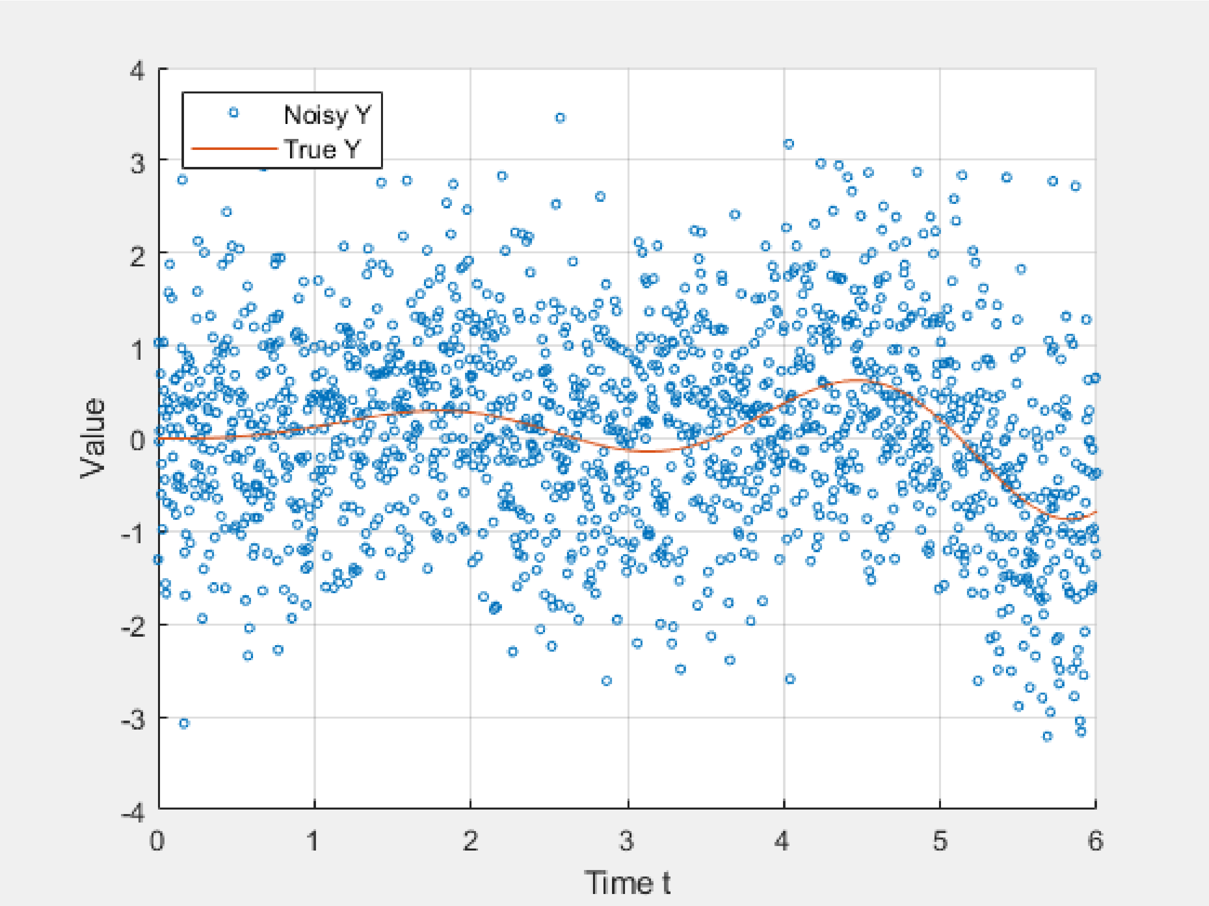

Figure 3: Noisy y_M vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

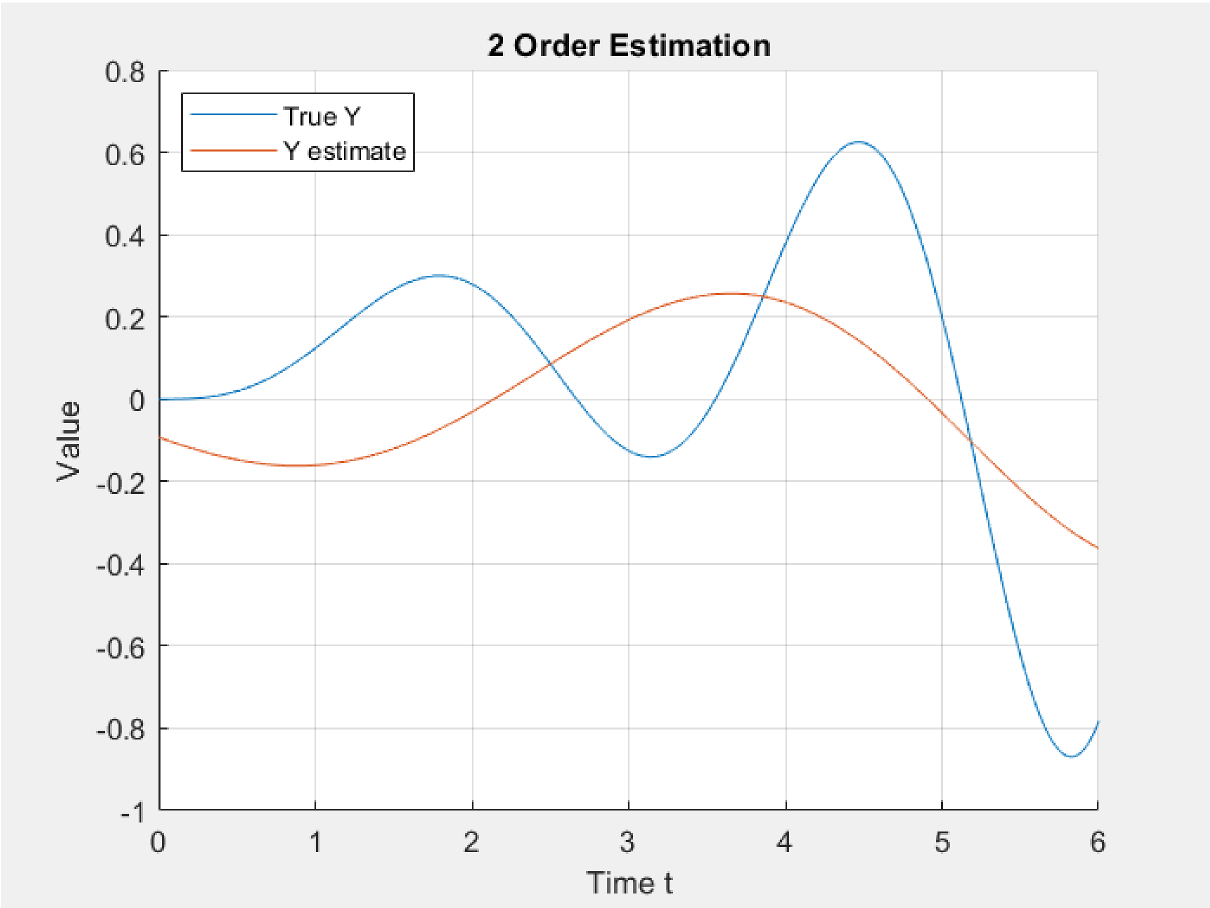

Figure 4: Estimate of y using second order vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

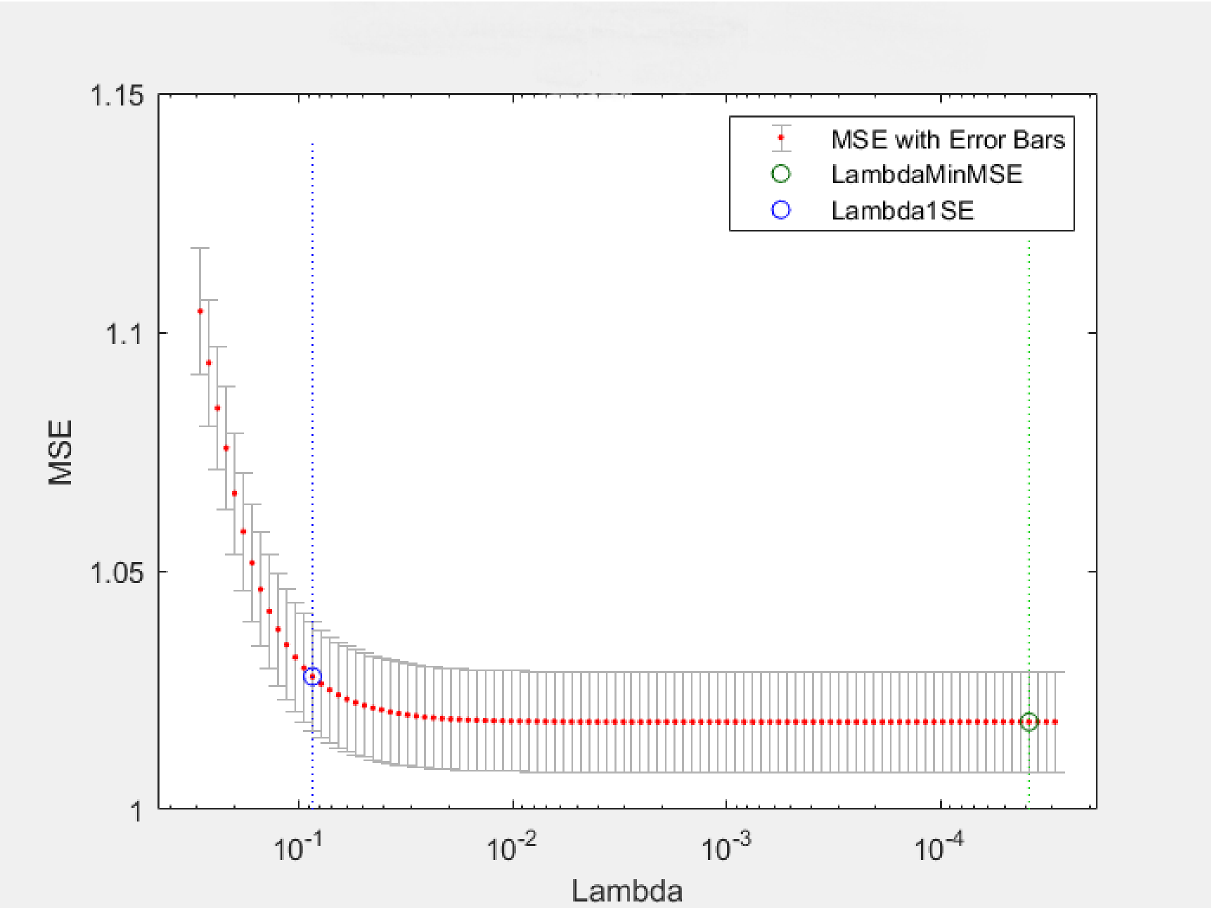

Figure 5: Cross-validation of MSE and λ for ridge regression of second order system for noise of 1 SD (SNR -8.9652 dB)

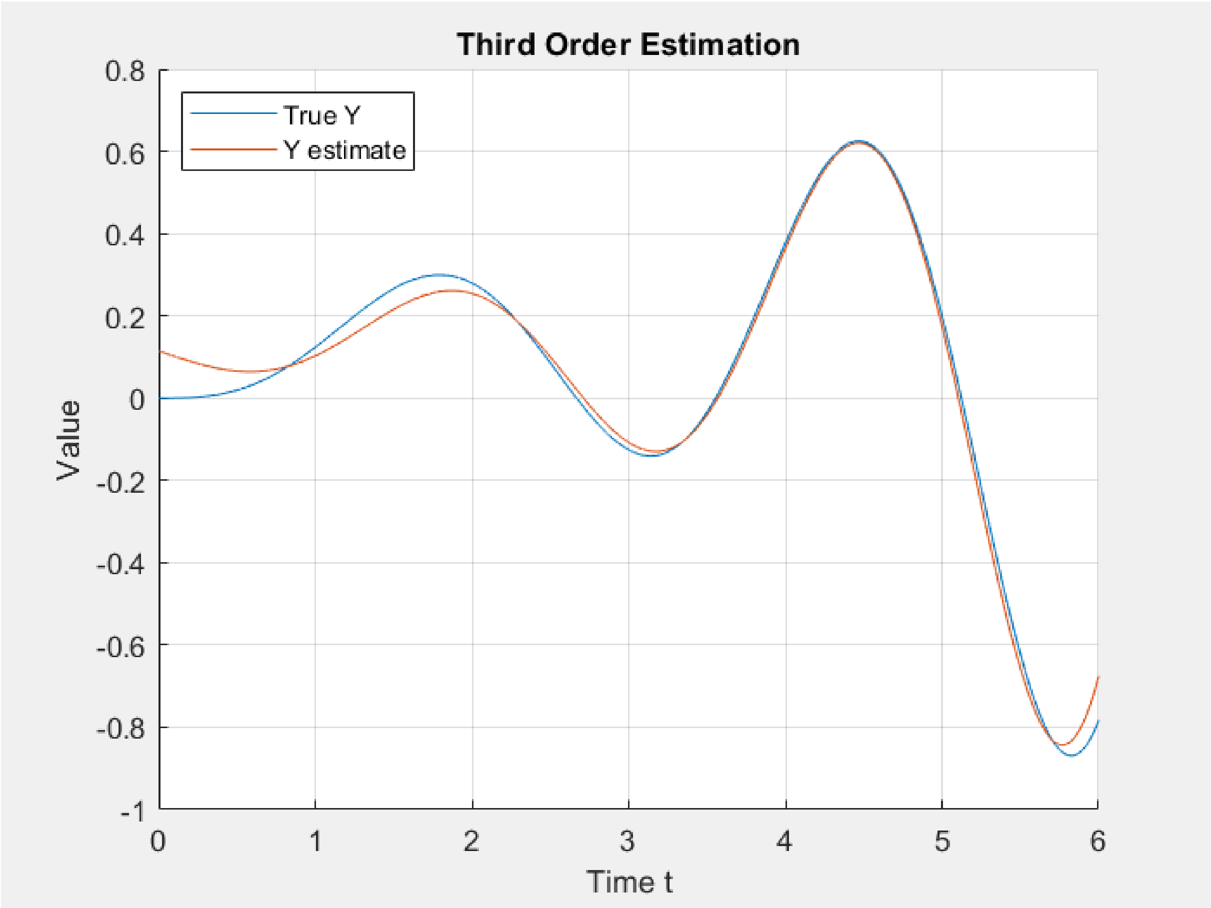

Figure 6: Estimate of y using third order vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

.

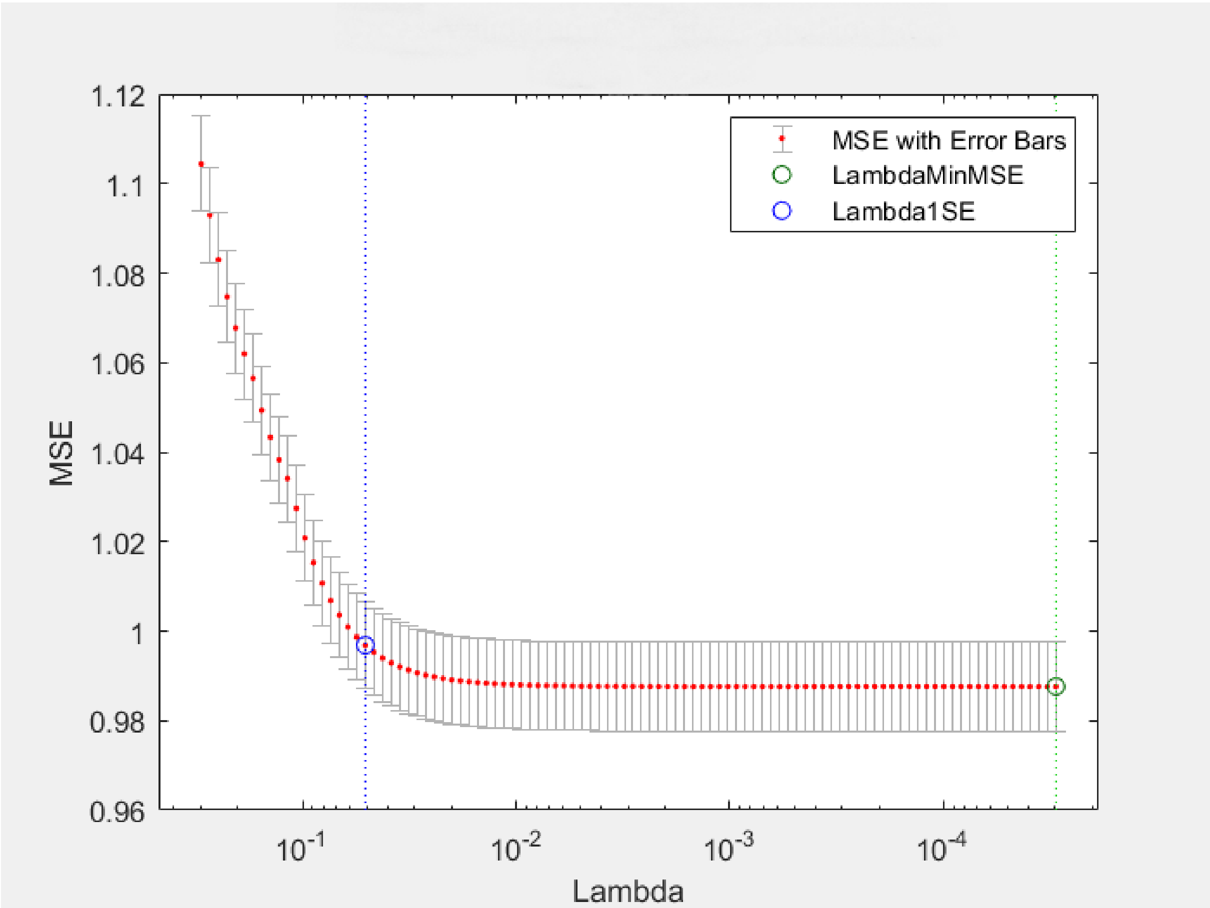

Figure 7: Cross-validation of MSE and λ for ridge regression of third order system for noise of 1 SD (SNR -8.9652 dB)

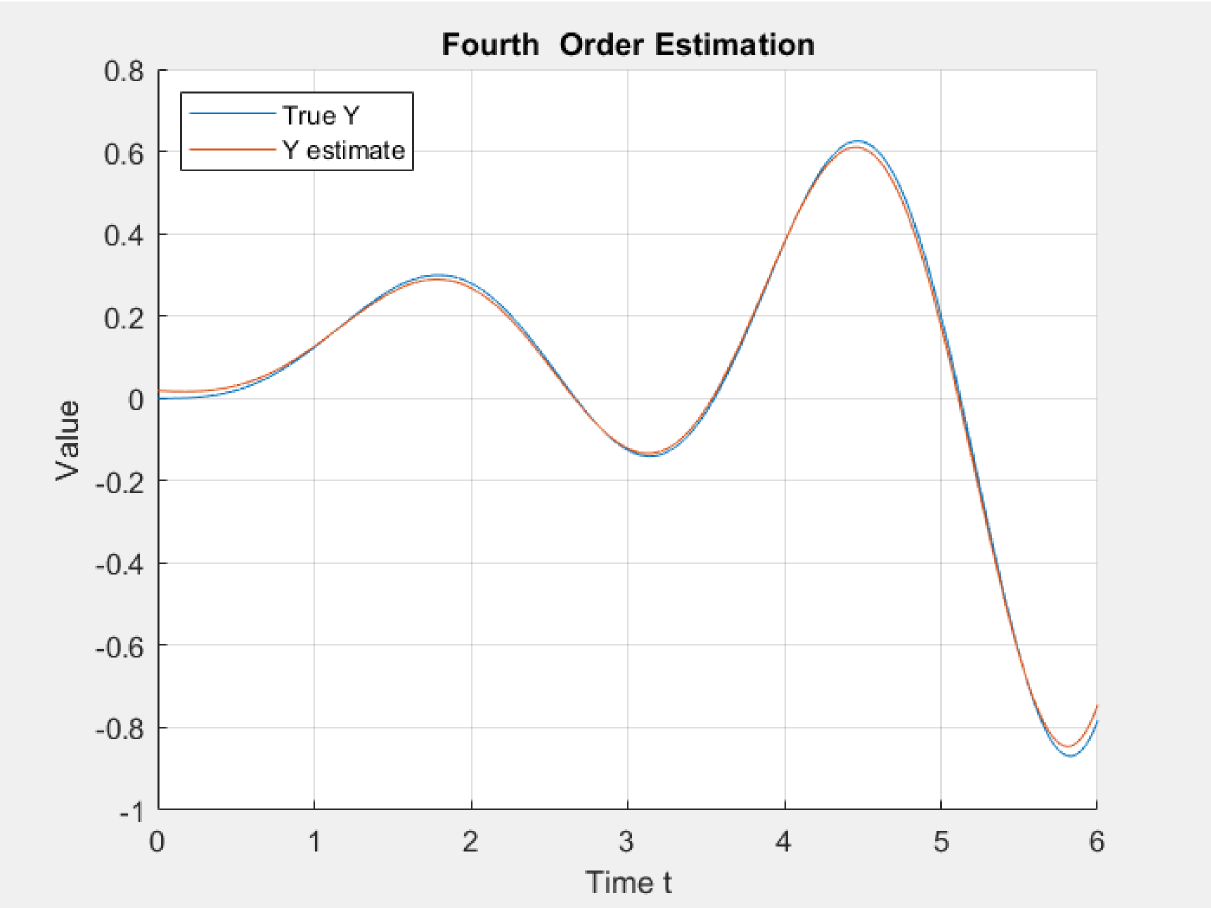

Figure 8: Estimate of y using forth order vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

.

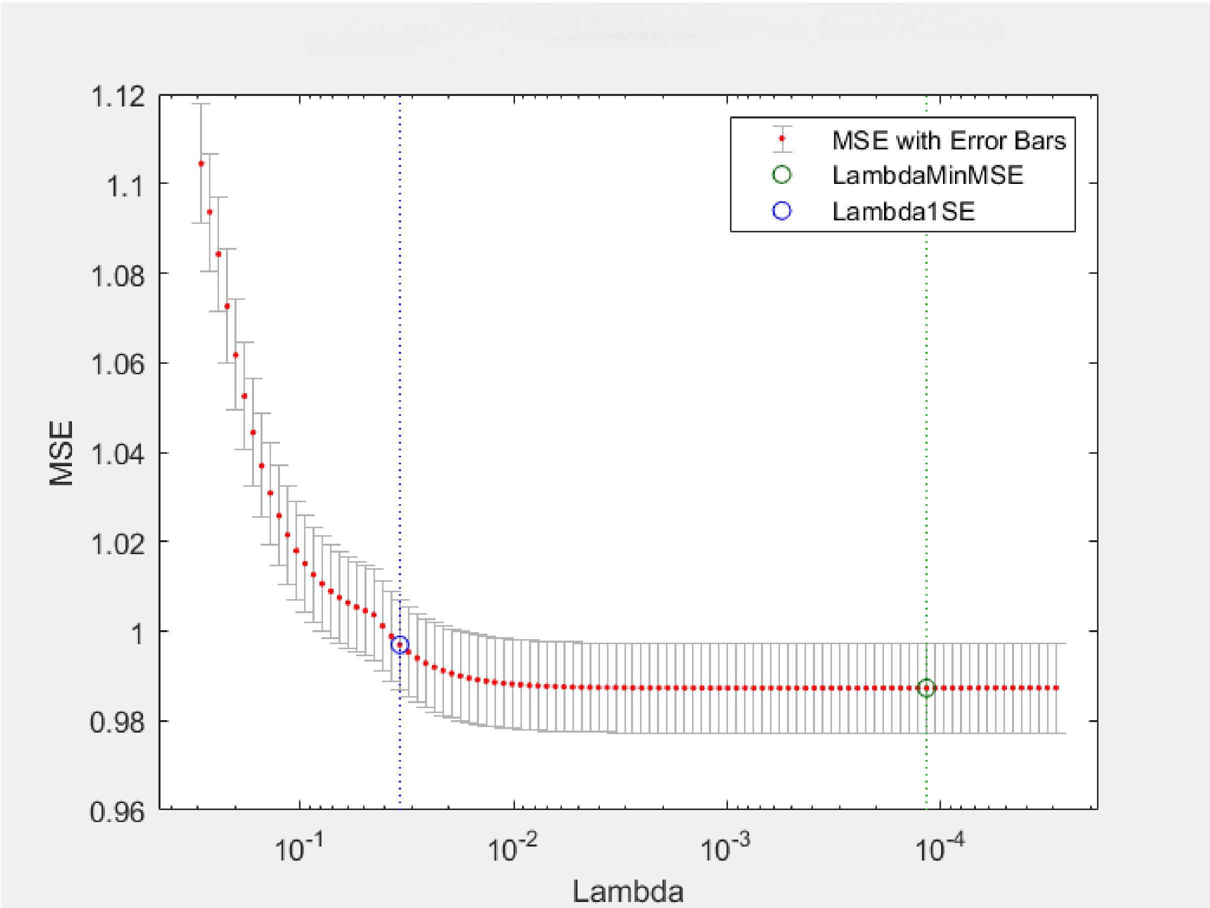

Figure 9: Cross-validation of MSE and λ for ridge regression of fourth order system for noise of 1 SD (SNR -8.9652 dB)

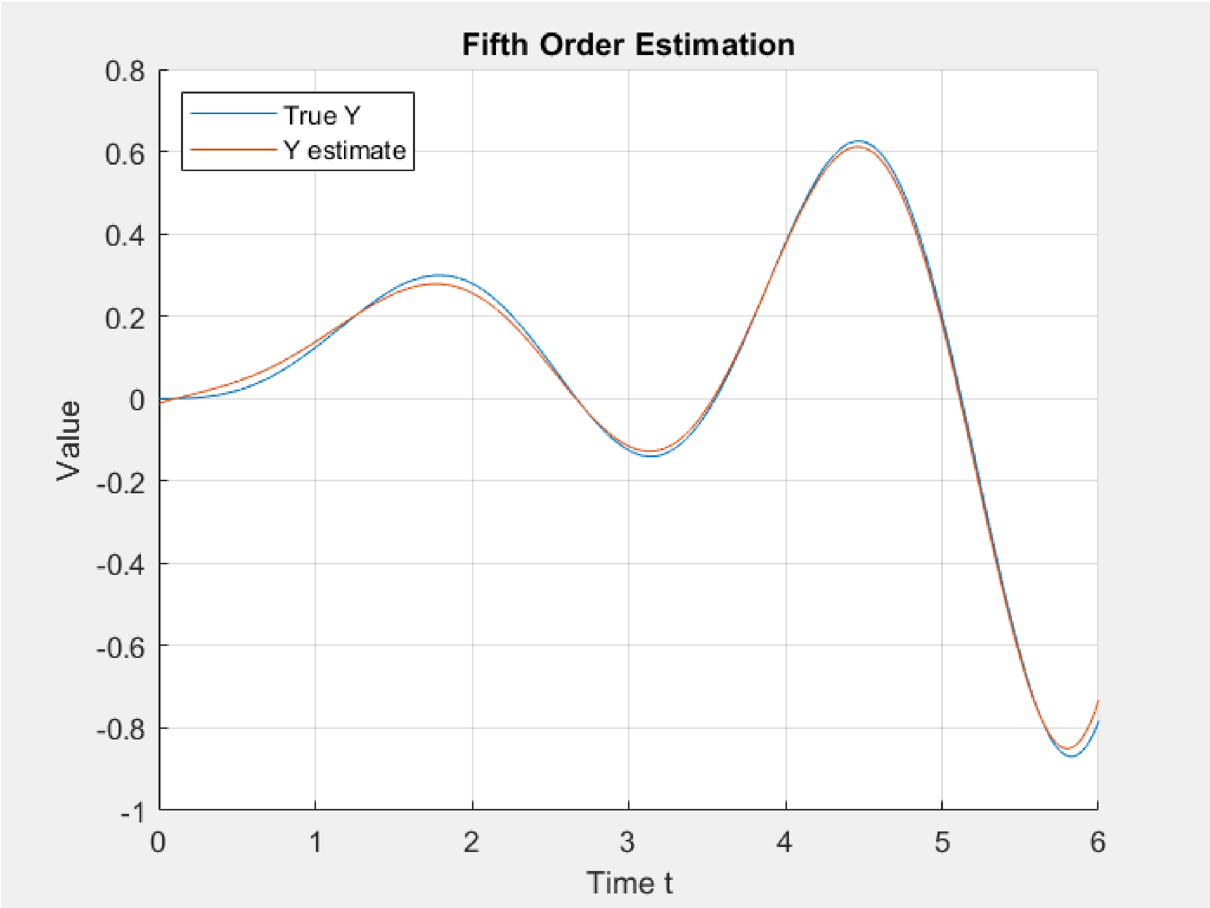

Figure 10: Estimate of y using fifth order vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

.

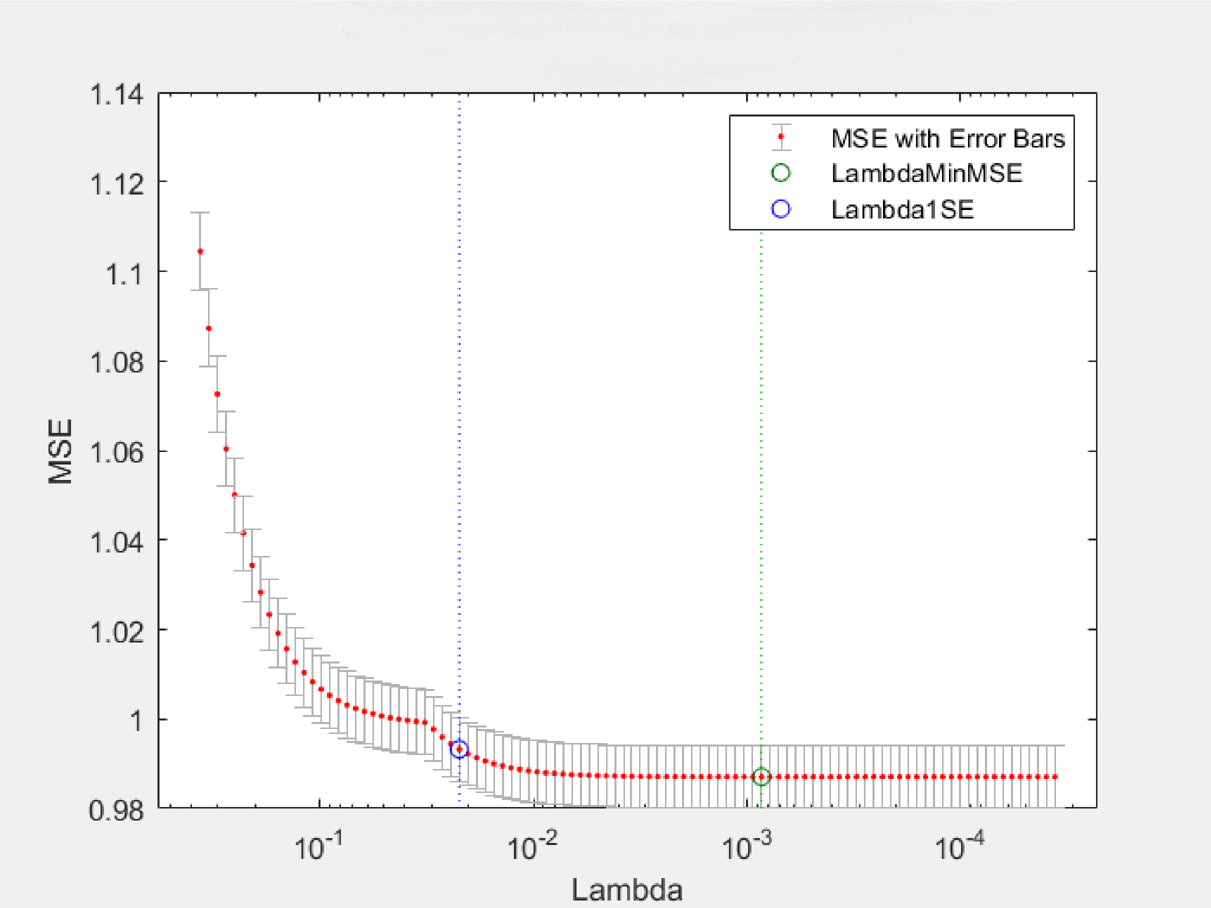

Figure 11: Cross-validation of MSE and λ for ridge regression of fifth order system for noise of 1 SD (SNR -8.9652 dB)

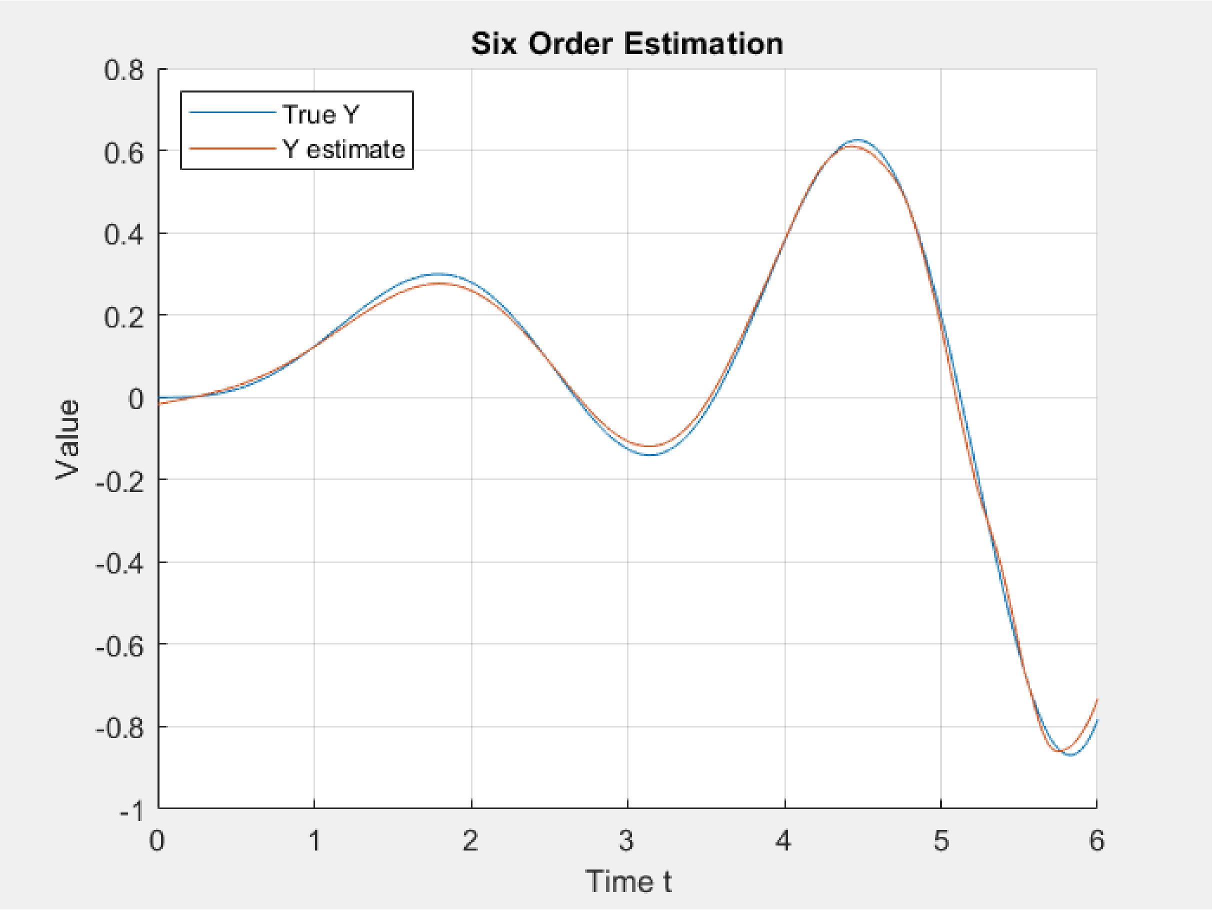

Figure 12: Estimate of y using sixth order vs. nominal y for noise of 1 SD (SNR -8.9652 dB)

.

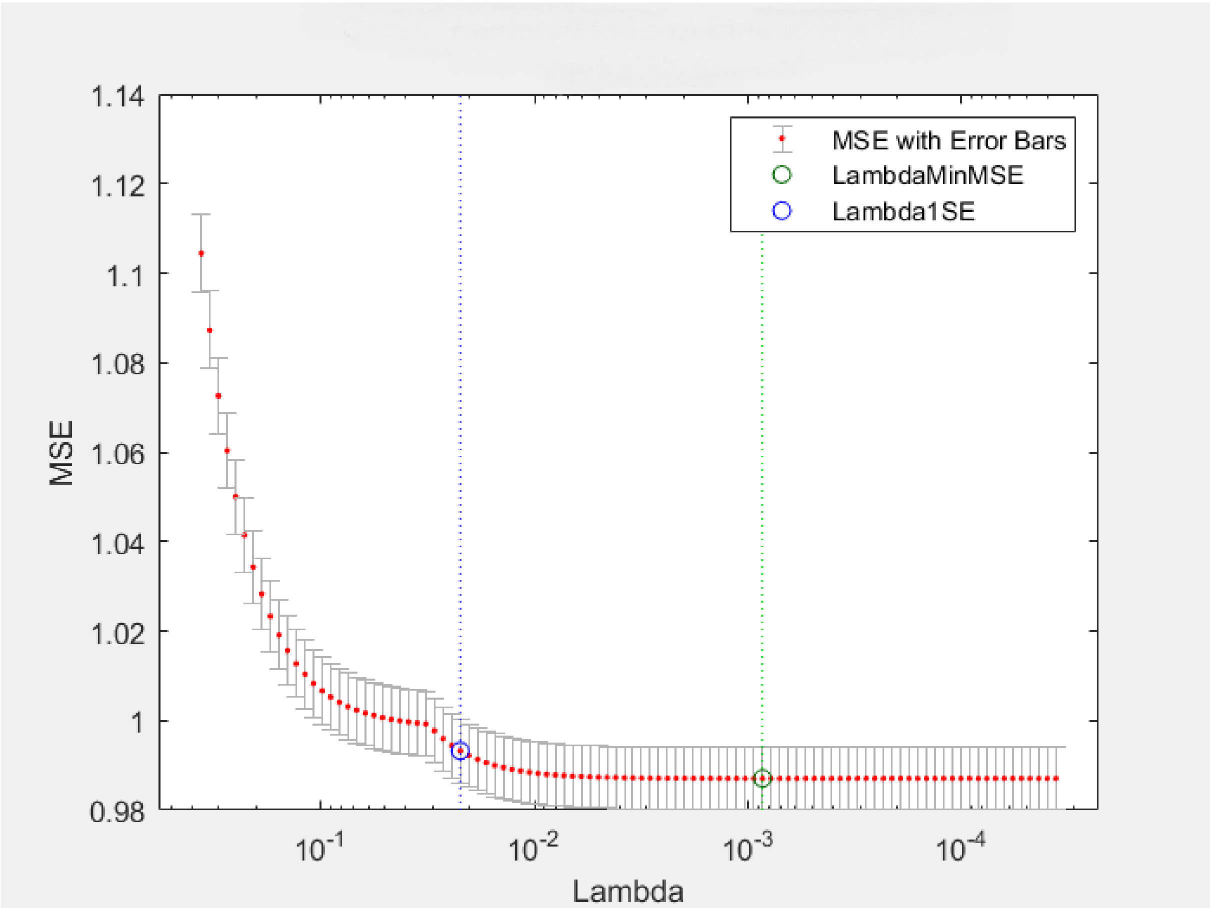

Figure 13: Cross-validation of MSE and λ for ridge regression of sixth order system for noise of 1 SD (SNR -8.9652 dB)

According to the Table 2, using the BIC values, the estimated order is the fourth order. As follows, the derivatives reconstructions are based on the fourth order system.

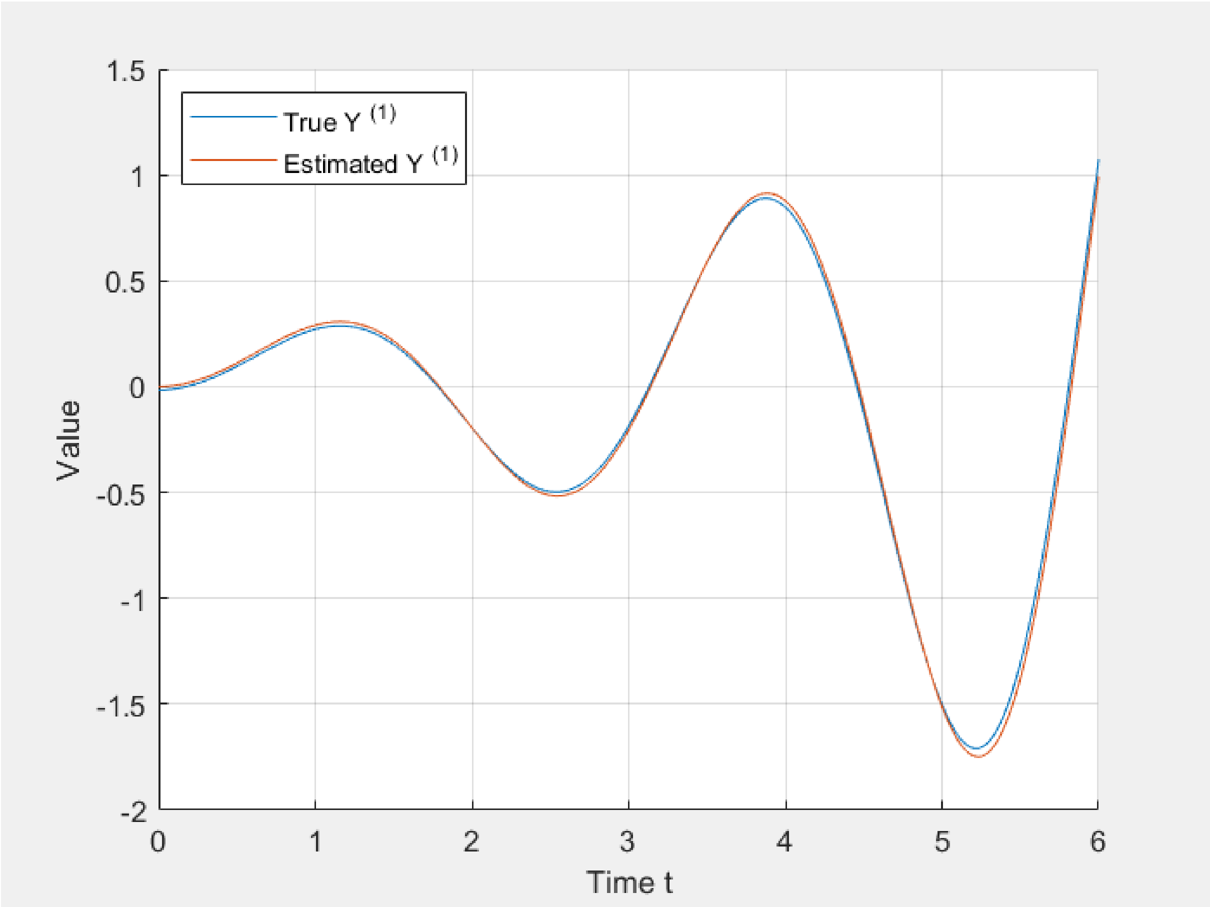

Figure 14: Estimate of $y^{(1)}$ vs. true $y^{(1)}$ for noise of 1 SD (SNR -8.9652 dB)

.

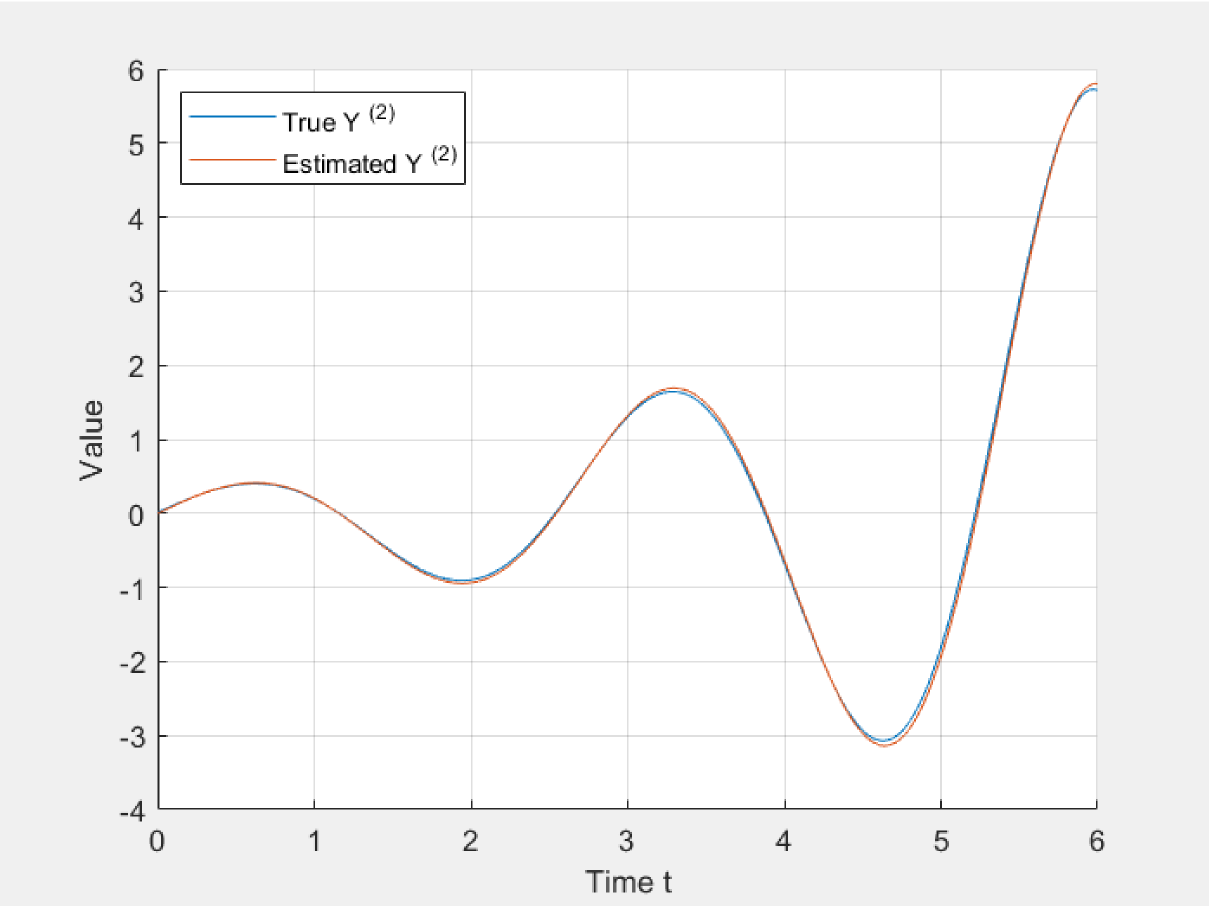

Figure 15: Estimate of $y^{(2)}$ vs. true $y^{(2)}$ for noise of 1 SD (SNR -8.9652 dB)
.

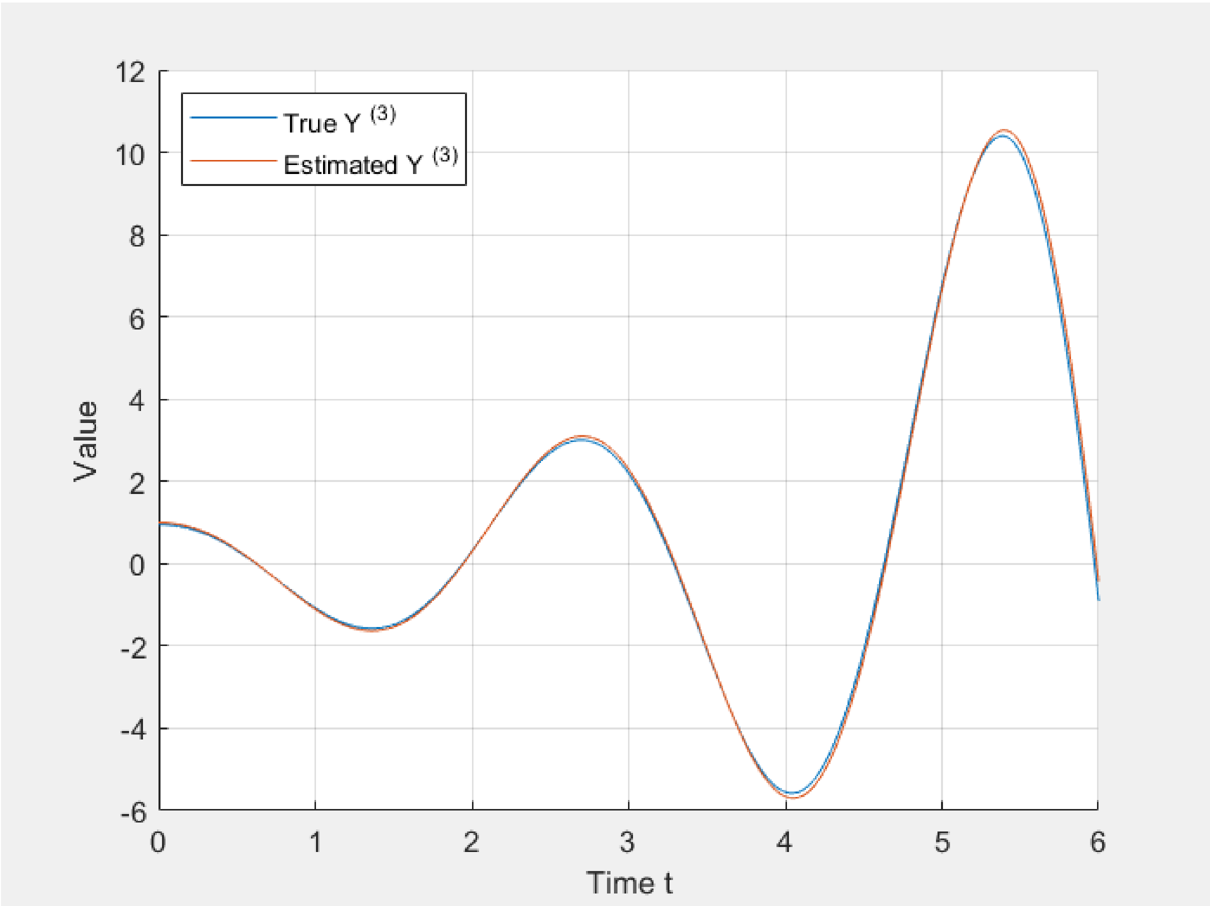

Figure 16: Estimate of $y^{(3)}$ vs. true $y^{(3)}$ for noise of 1 SD (SNR -8.9652 dB)
.

7.1.2 Noise $\sigma = 1.5$

Figure 17 shows y_M vs. y, where the true y is displayed in red and the noisy y in blue. The sample size is $N = 5000$. We assumed the order of the system is between 2 to 6. The Results of estimation for each order is shown in the 3.

Order	Values	a_0	a_1	a_2	a_3	a_4	a_5	RMSD	BIC
	True	1	5	5	0	-	-	-	-
2	Estimated	1.256	-0.331	-	-	-	-	9.60e-2	-2.83424e5
3	Estimated	0.119	6.194	-0.999	-	-	-	1.5e-3	-2.84631e5
4	Estimated	1.160	5.029	4.855	0	-	-	2.1e-4	**-2.84636e5**
5	Estimated	1.782	53.521	-2.442	13.476	0	-	8.4e-4	-2.84617e5
6	Estimated	20.414	99.302	127.147	0	29.323	-0.101	**1.9e-4**	-2.84617e5

Table 3: Estimated parameter values with noise of 1.5 SD (SNR -12.5224 dB)

As the Table 3 shows the least value for BIC is associated with the fourth order. It estimates the order correctly. In this situation, even though RMSD of the sixth order is less than fourth order, we will use BIC to decide the order.

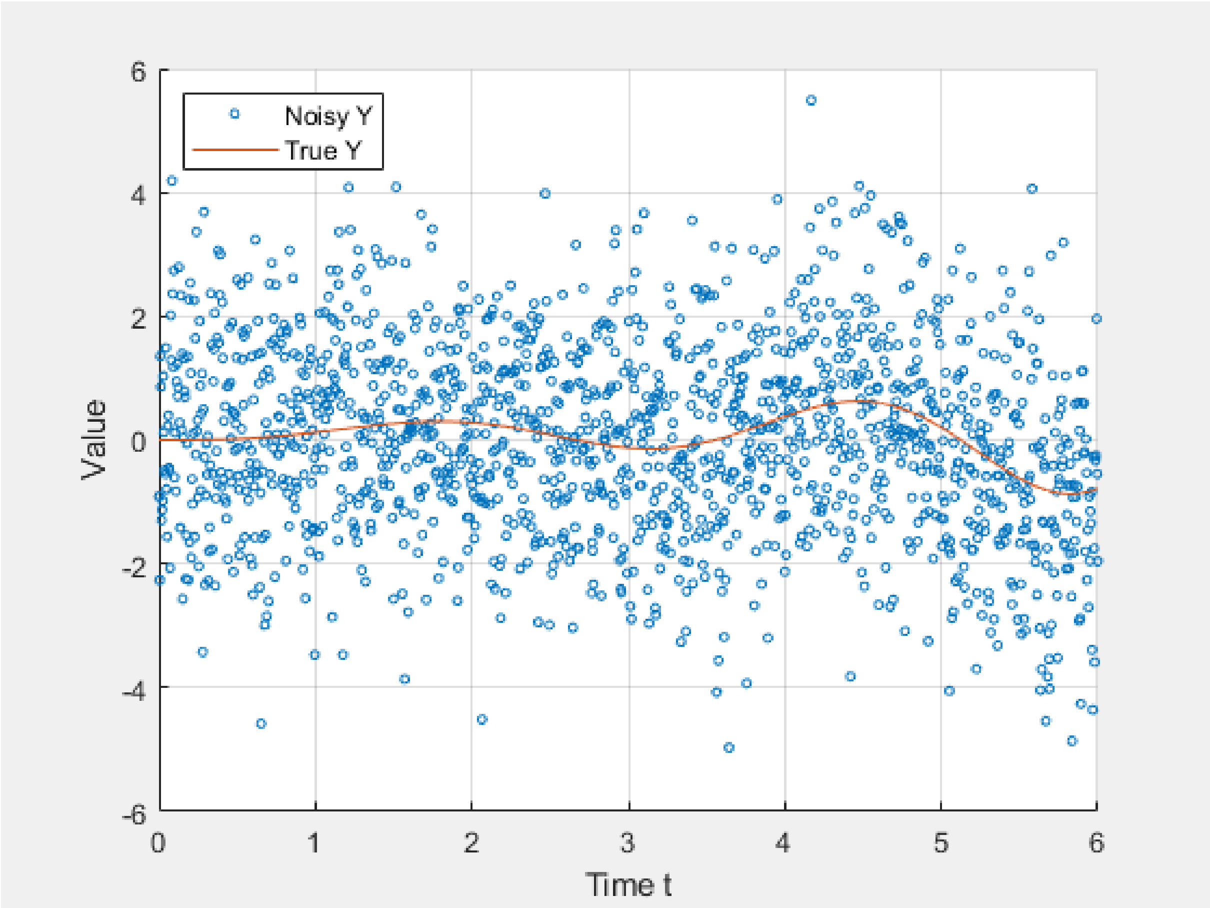

Figure 17: Noisy y_M vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

.

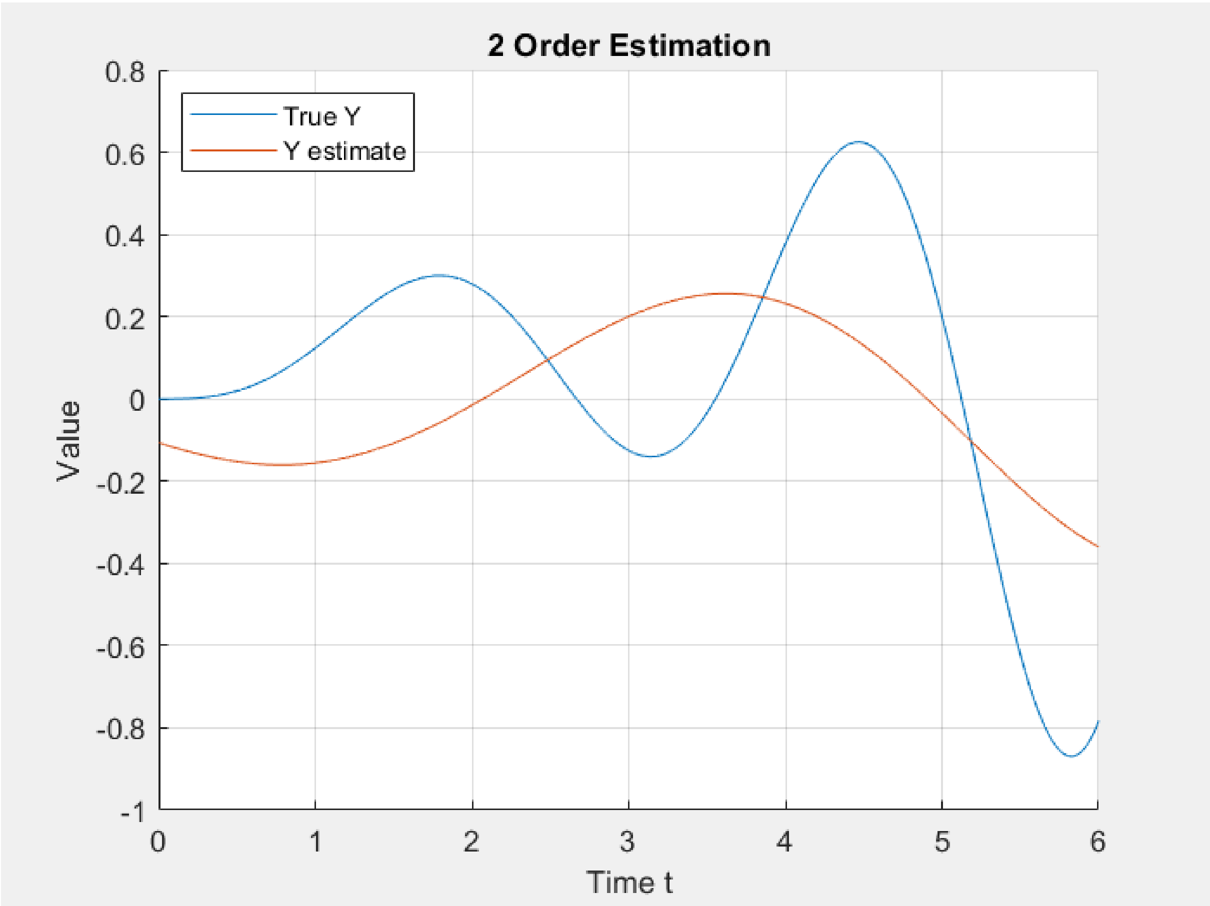

Figure 18: Estimate of y using second order vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

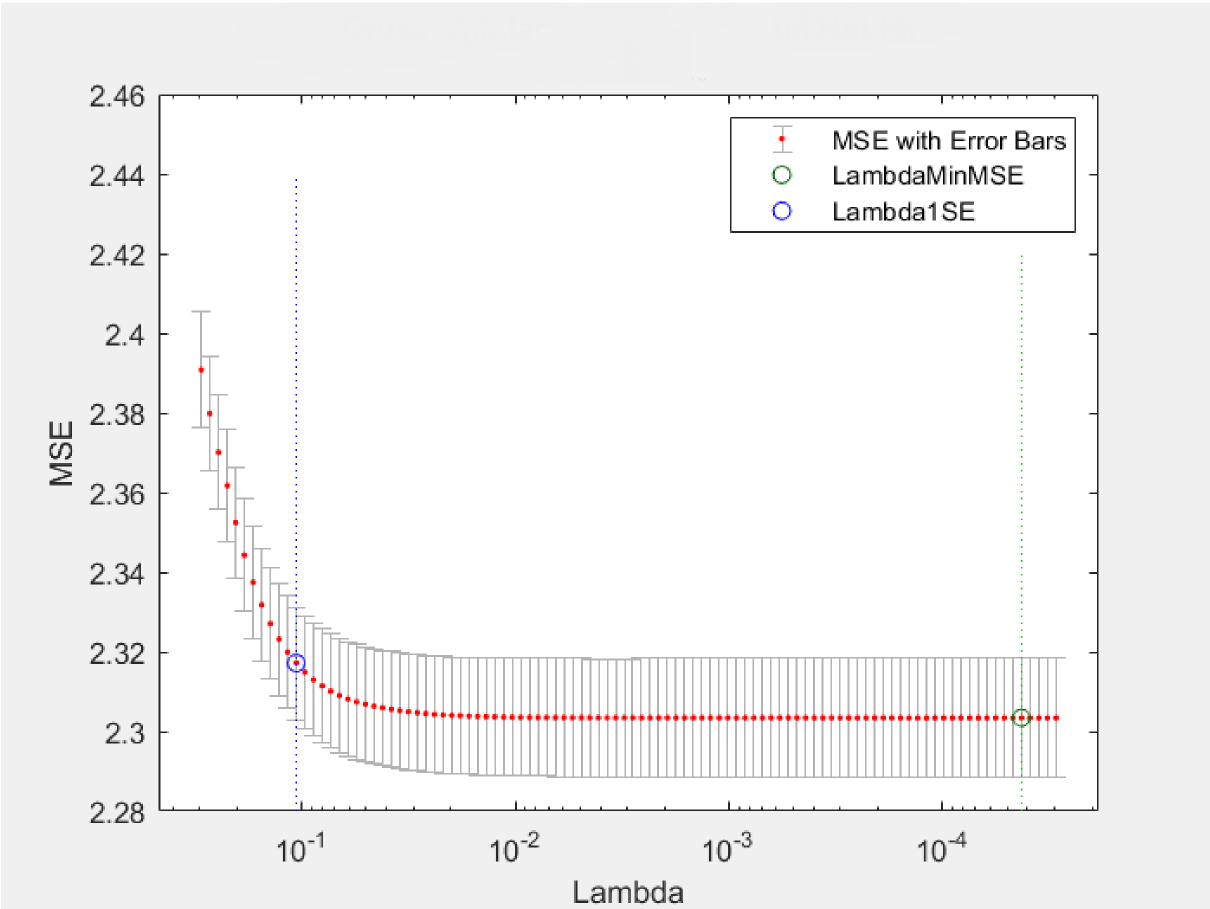

Figure 19: Cross-validation of MSE and λ for ridge regression of second order system for noise of 1.5 SD (SNR -12.5224 dB)

.

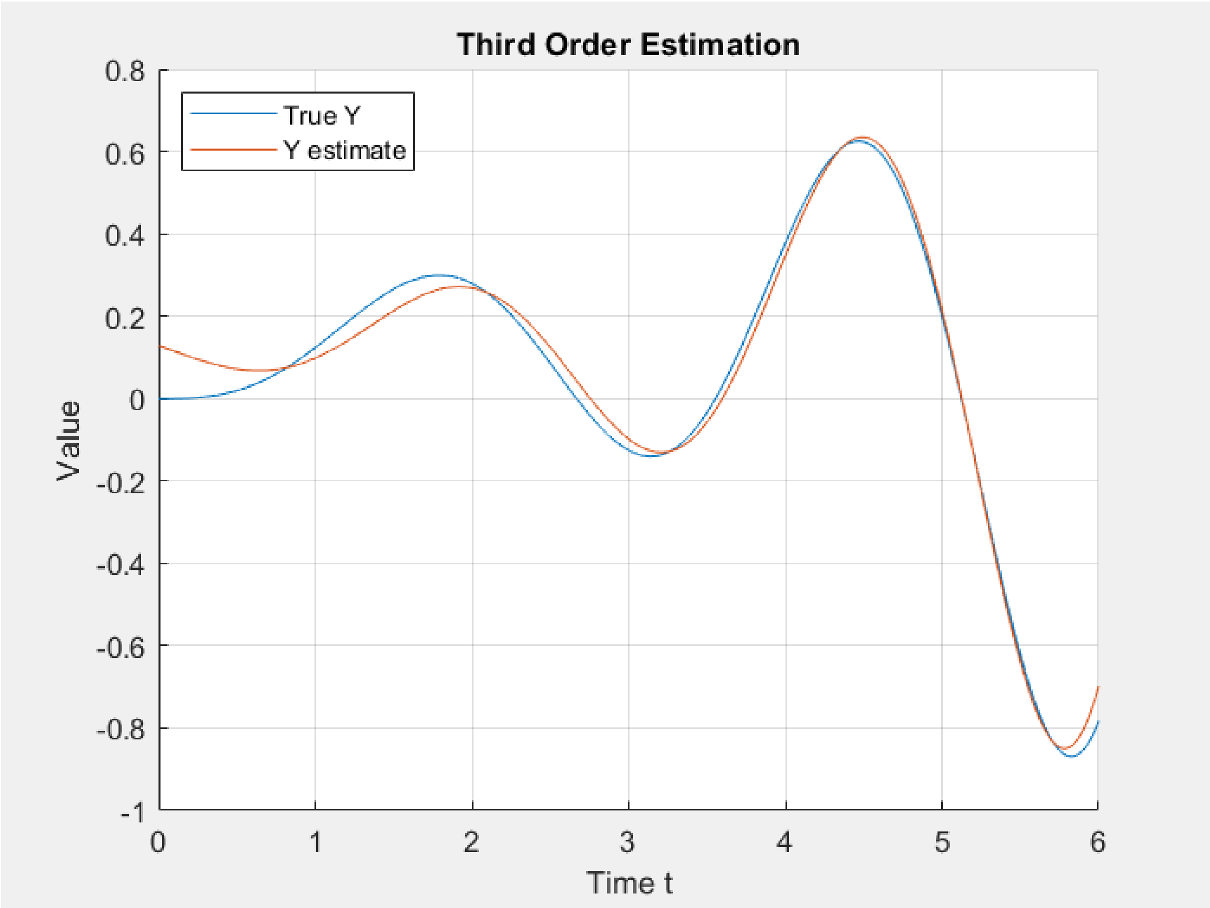

Figure 20: Estimate of y using third order vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

.

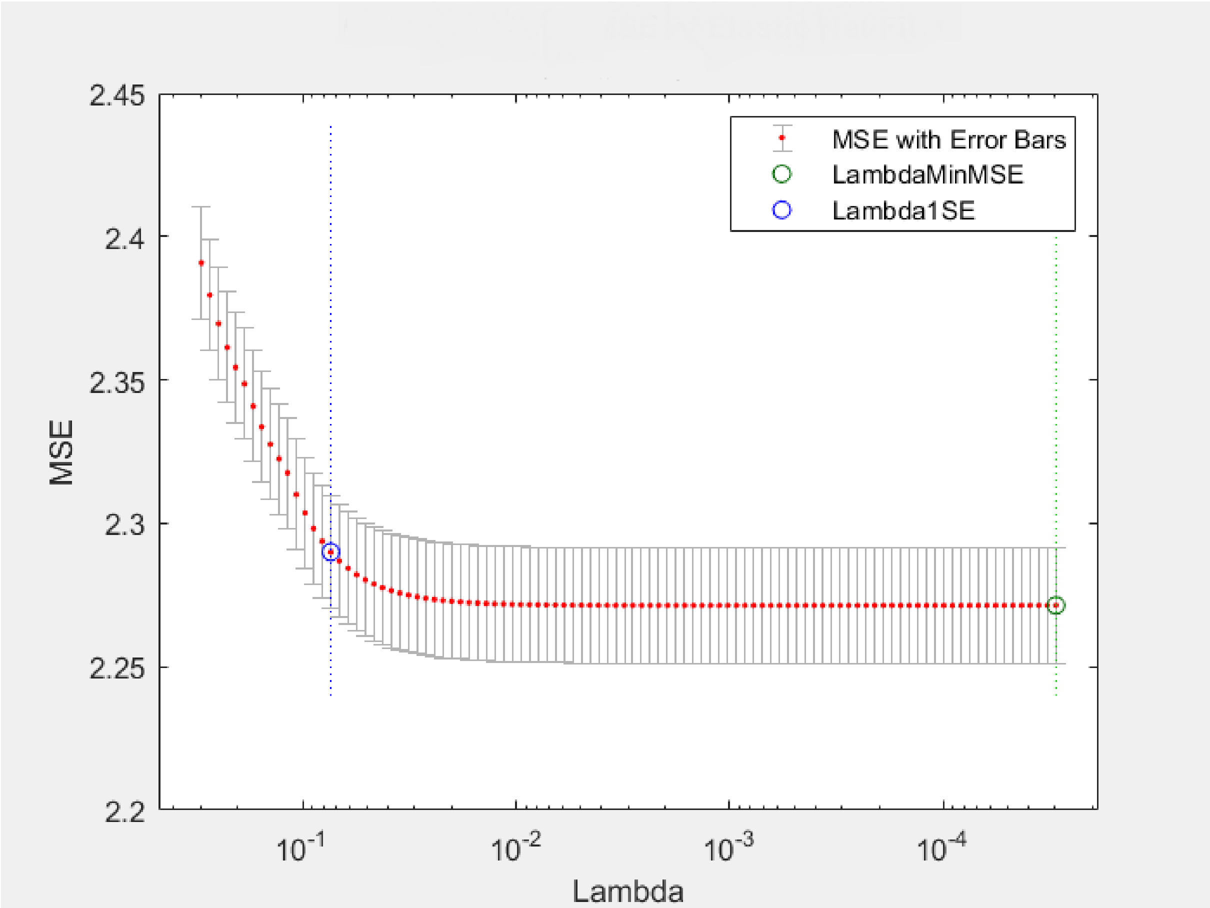

Figure 21: Cross-validation of MSE and λ for ridge regression of third order system for noise of 1.5 SD (SNR -12.5224 dB)

.

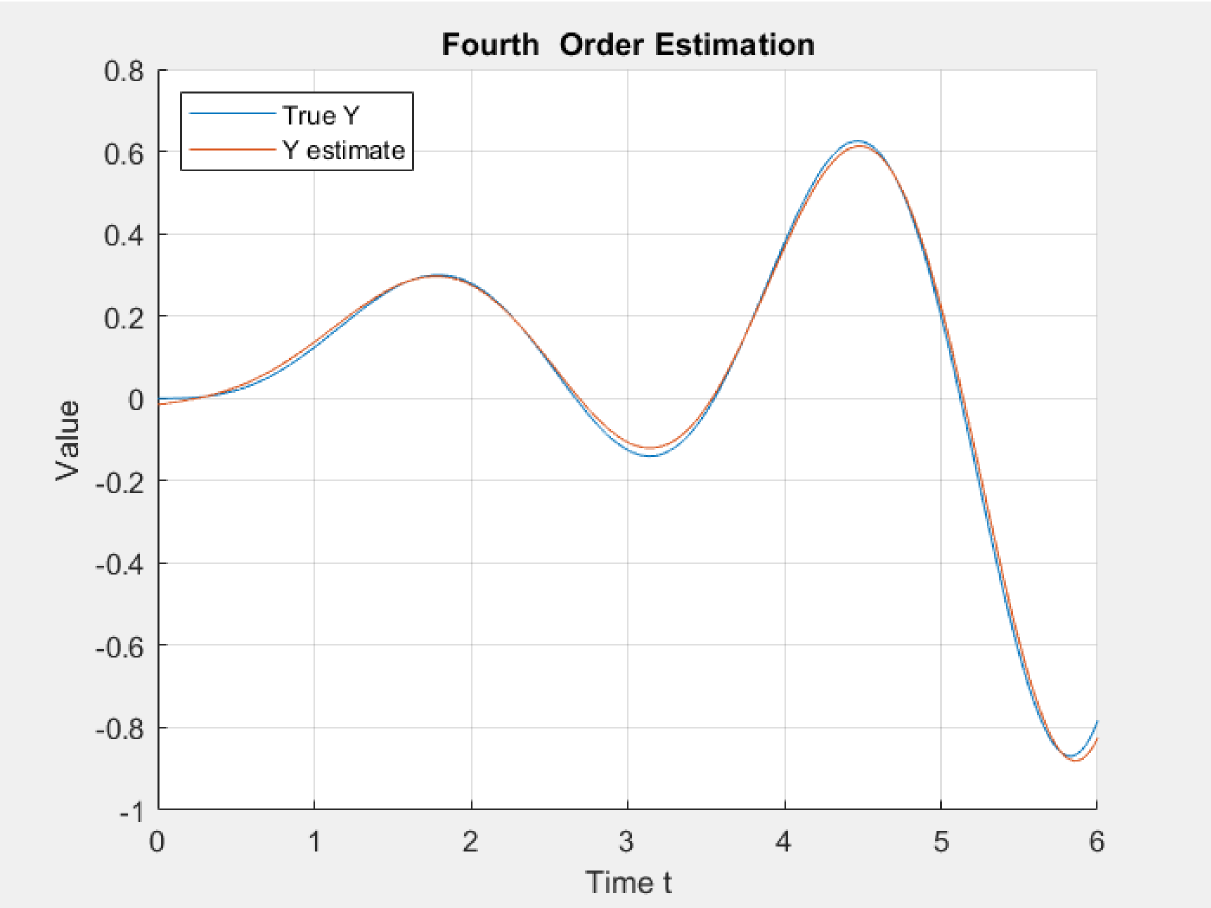

Figure 22: Estimate of y using fourth order vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

.

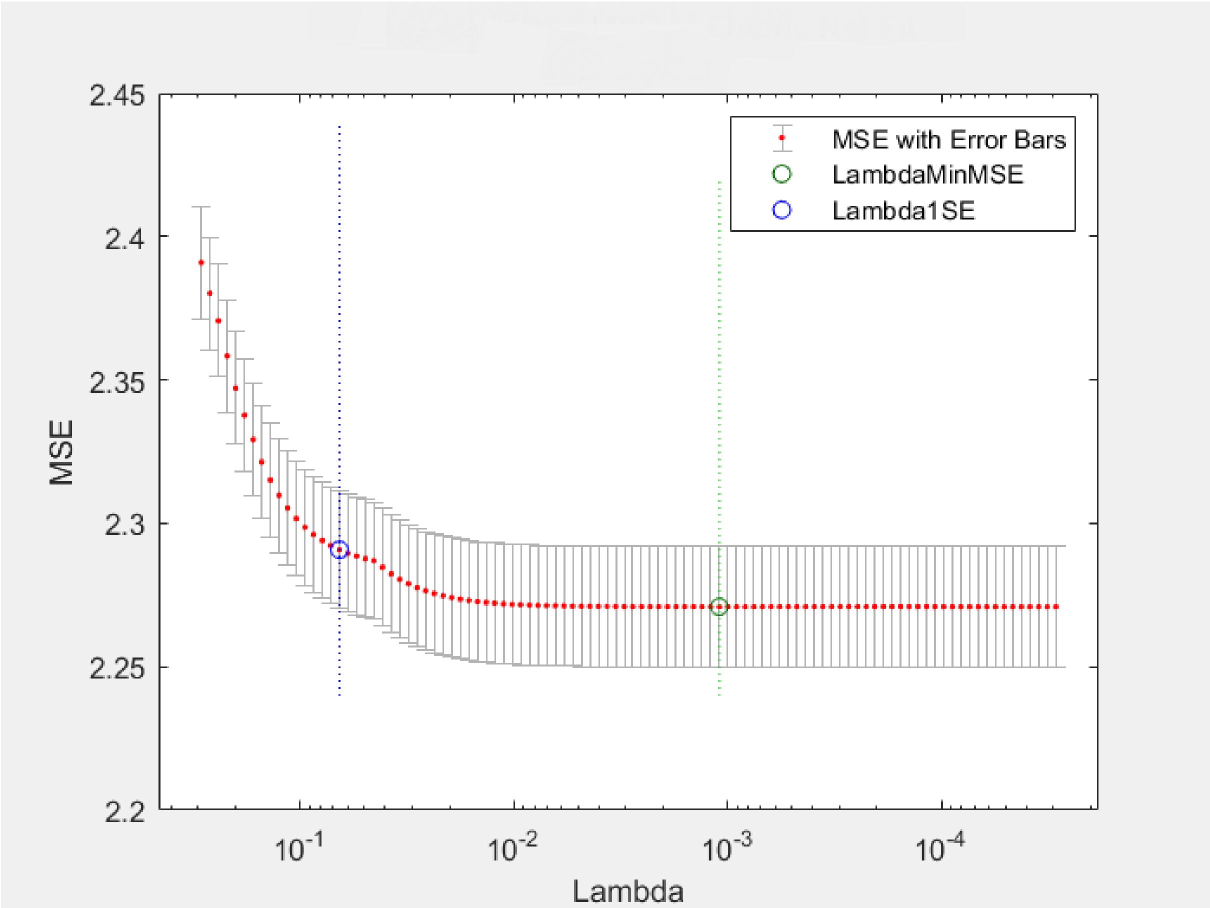

Figure 23: Cross-validation of MSE and λ for ridge regression of fourth order system for noise of 1.5 SD (SNR -12.5224 dB)

.

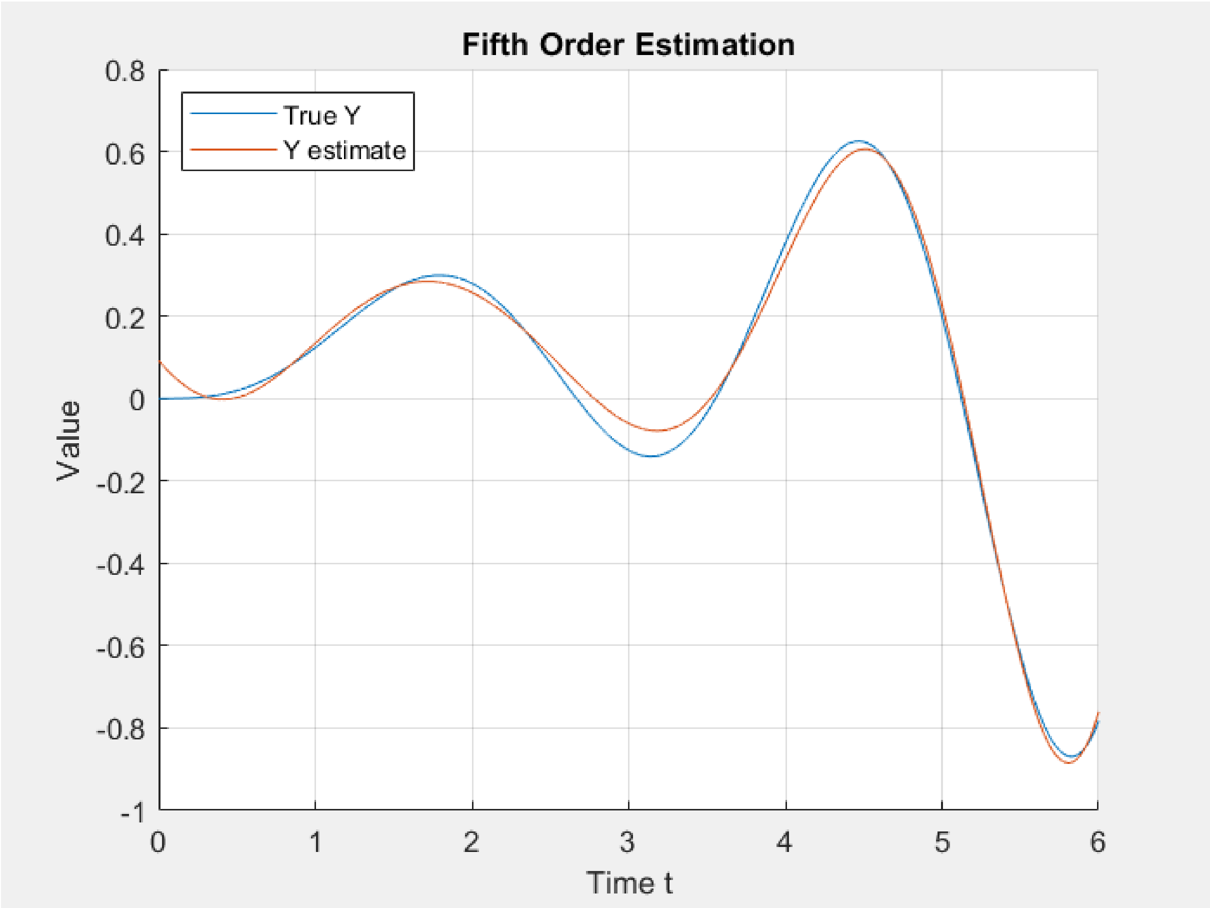

Figure 24: Estimate of y using fifth order vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

.

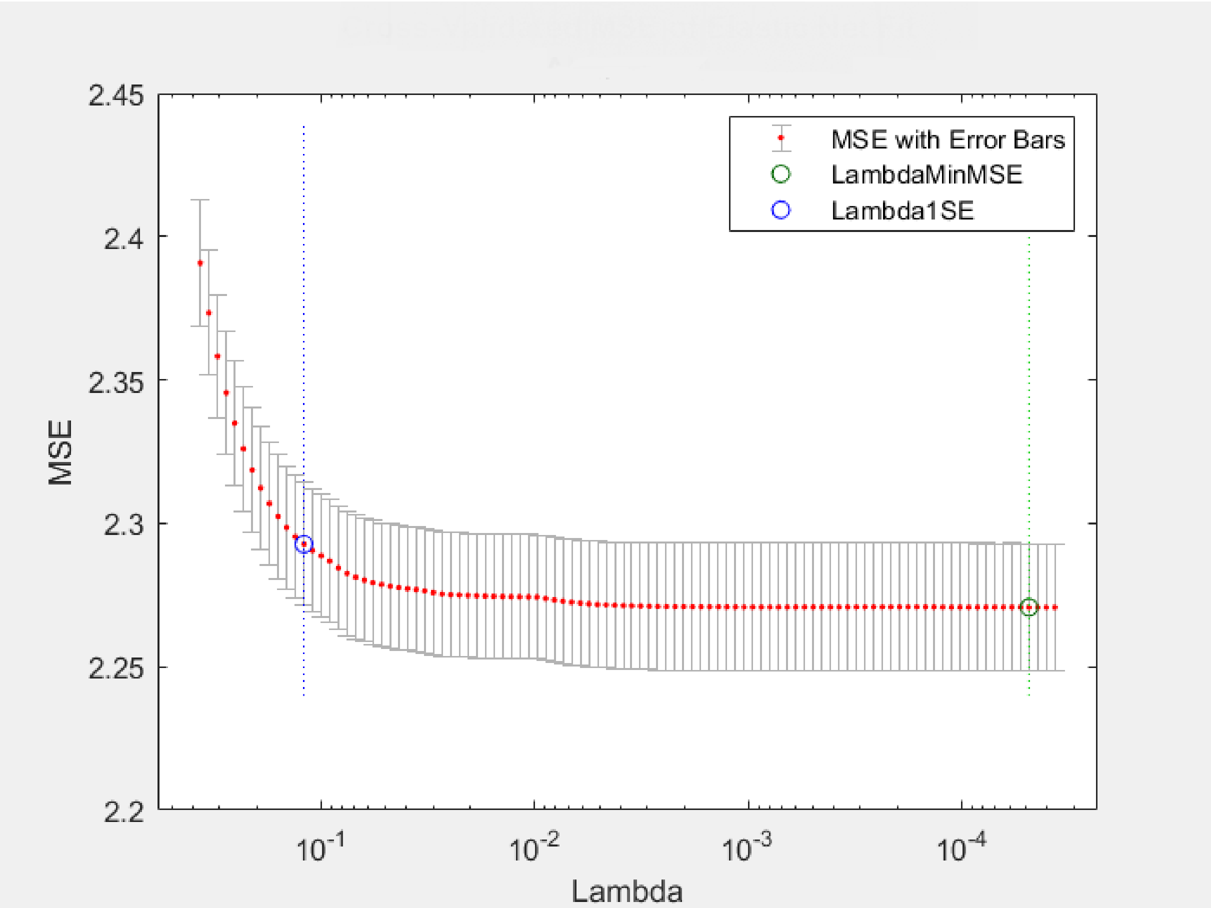

Figure 25: Cross-validation of MSE and λ for ridge regression of fifth order system for noise of 1.5 SD (SNR -12.5224 dB)

.

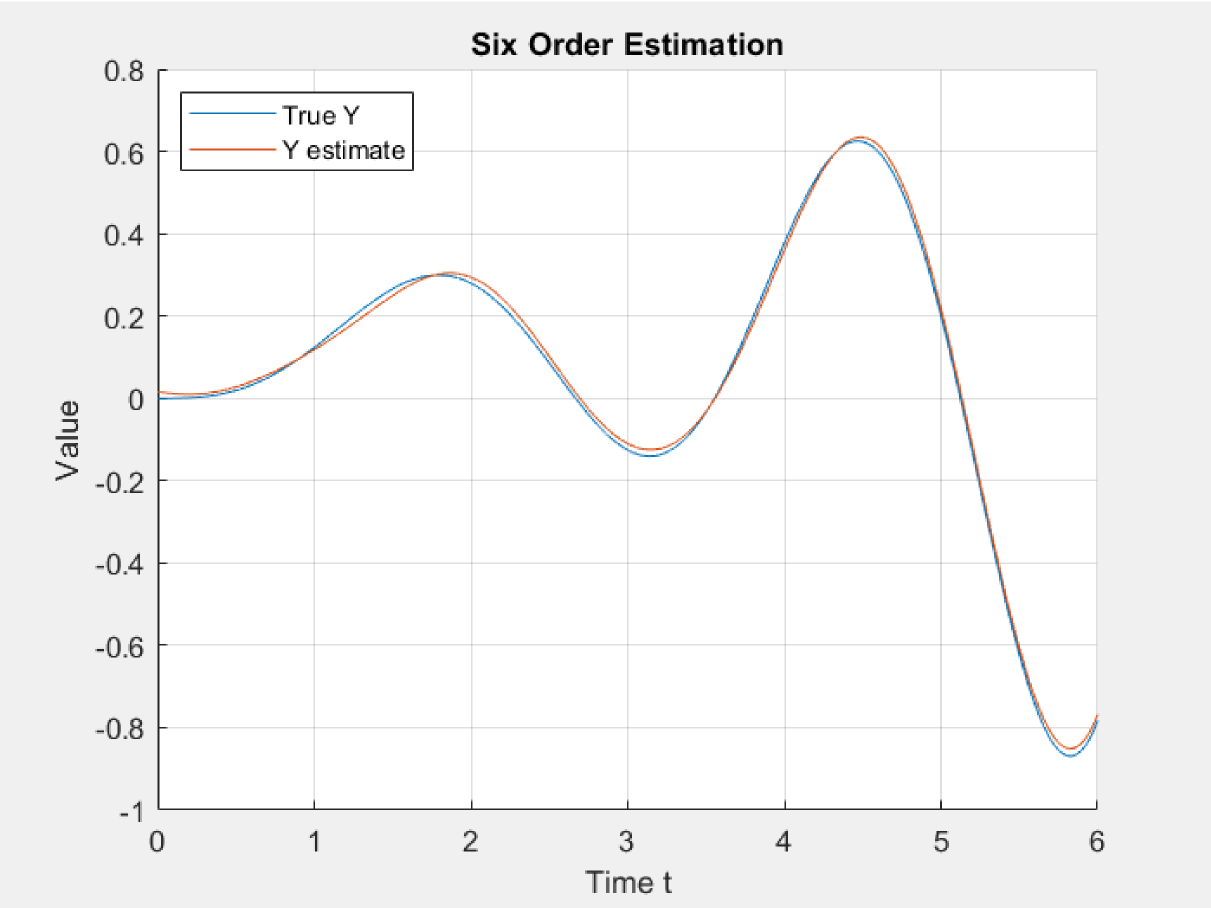

Figure 26: Estimate of y using sixth order vs. nominal y for noise of 1.5 SD (SNR -12.5224 dB)

.

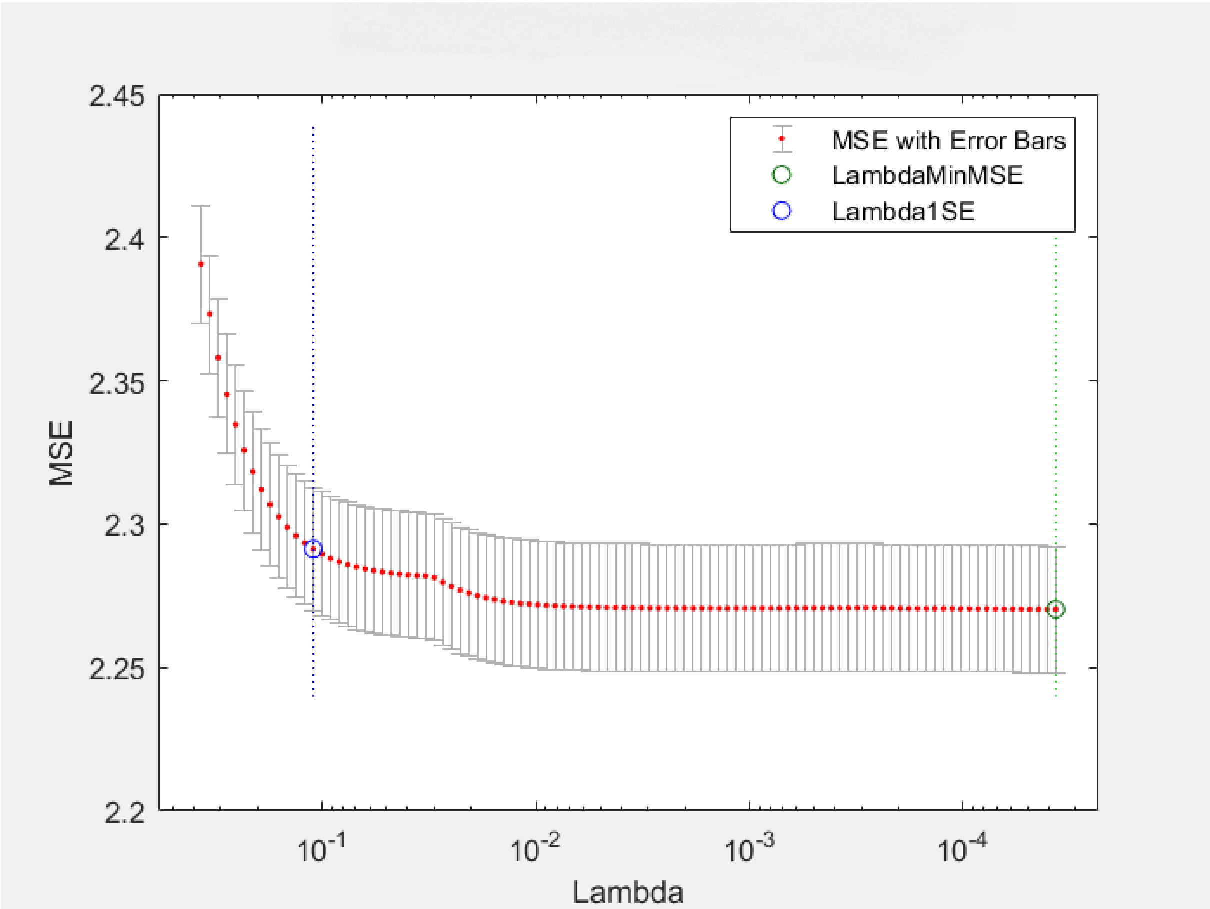

Figure 27: Cross-validation of MSE and λ for ridge regression of sixth order system for noise of 1.5 SD (SNR -12.5224 dB)

.

According to the table, using the BIC values, the estimated order is the fourth order order. As follows, the derivatives reconstructions are based on the fourth order order system.

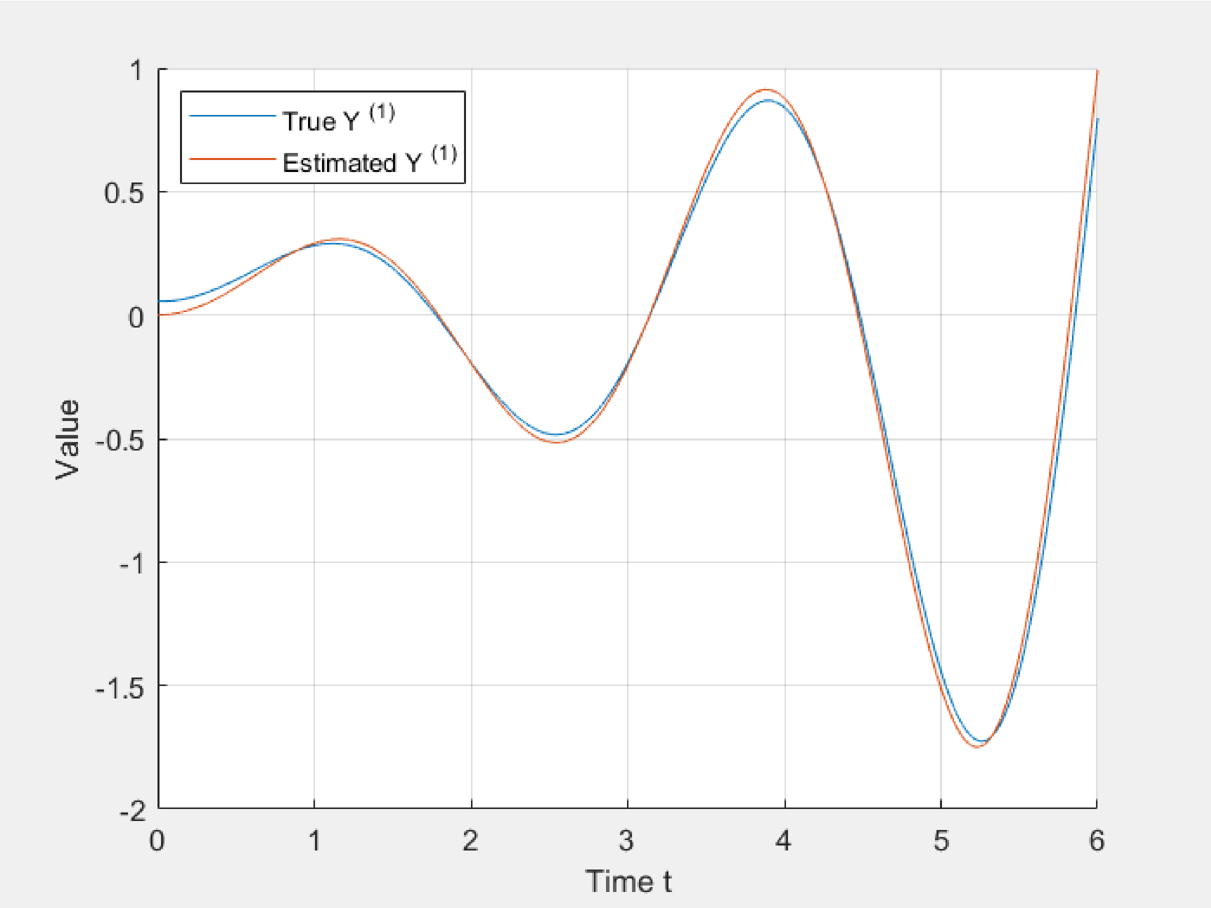

Figure 28: Estimate of $y^{(1)}$ vs. true $y^{(1)}$ for noise of 1.5 SD (SNR -12.5224 dB)

.

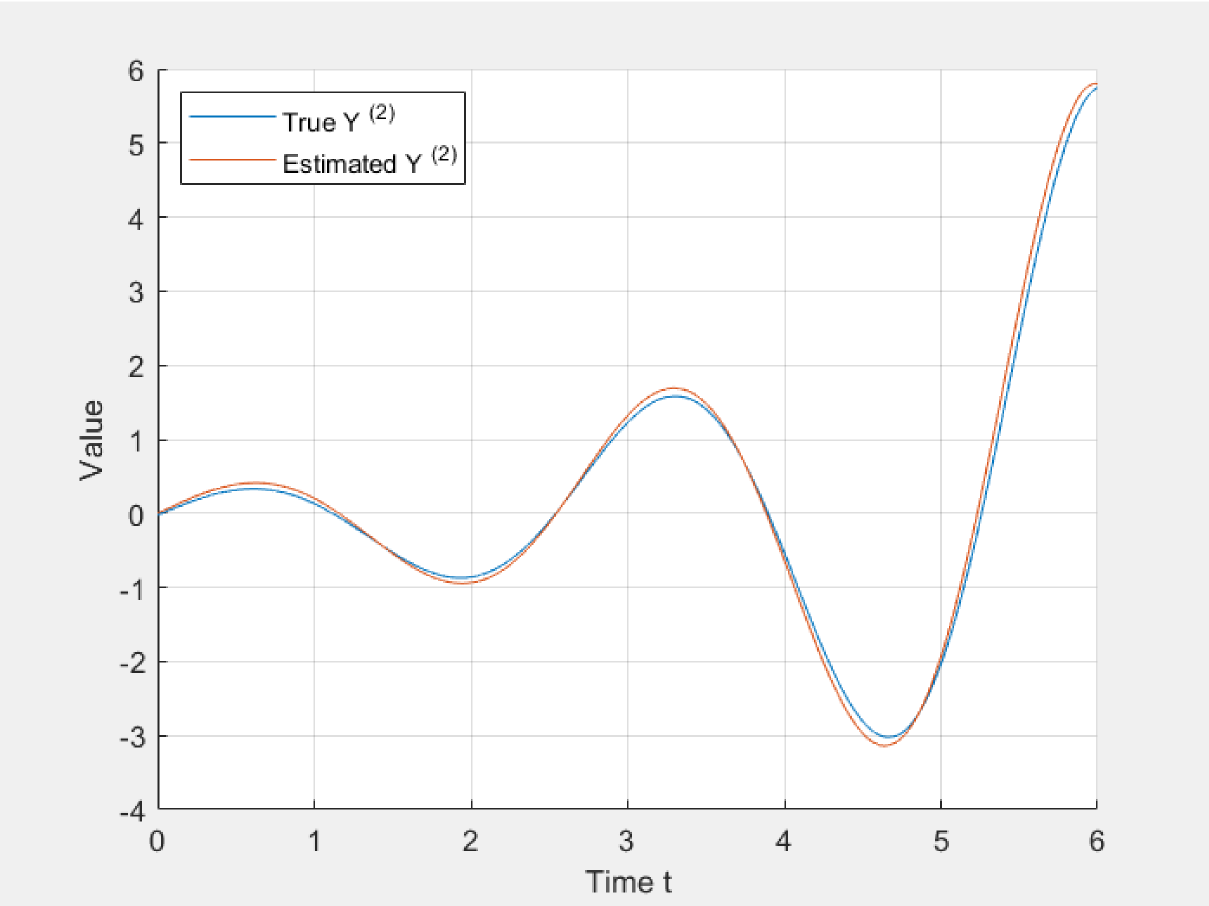

Figure 29: Estimate of $y^{(2)}$ vs. true $y^{(2)}$ for noise of 1.5 SD (SNR -12.5224 dB)

.

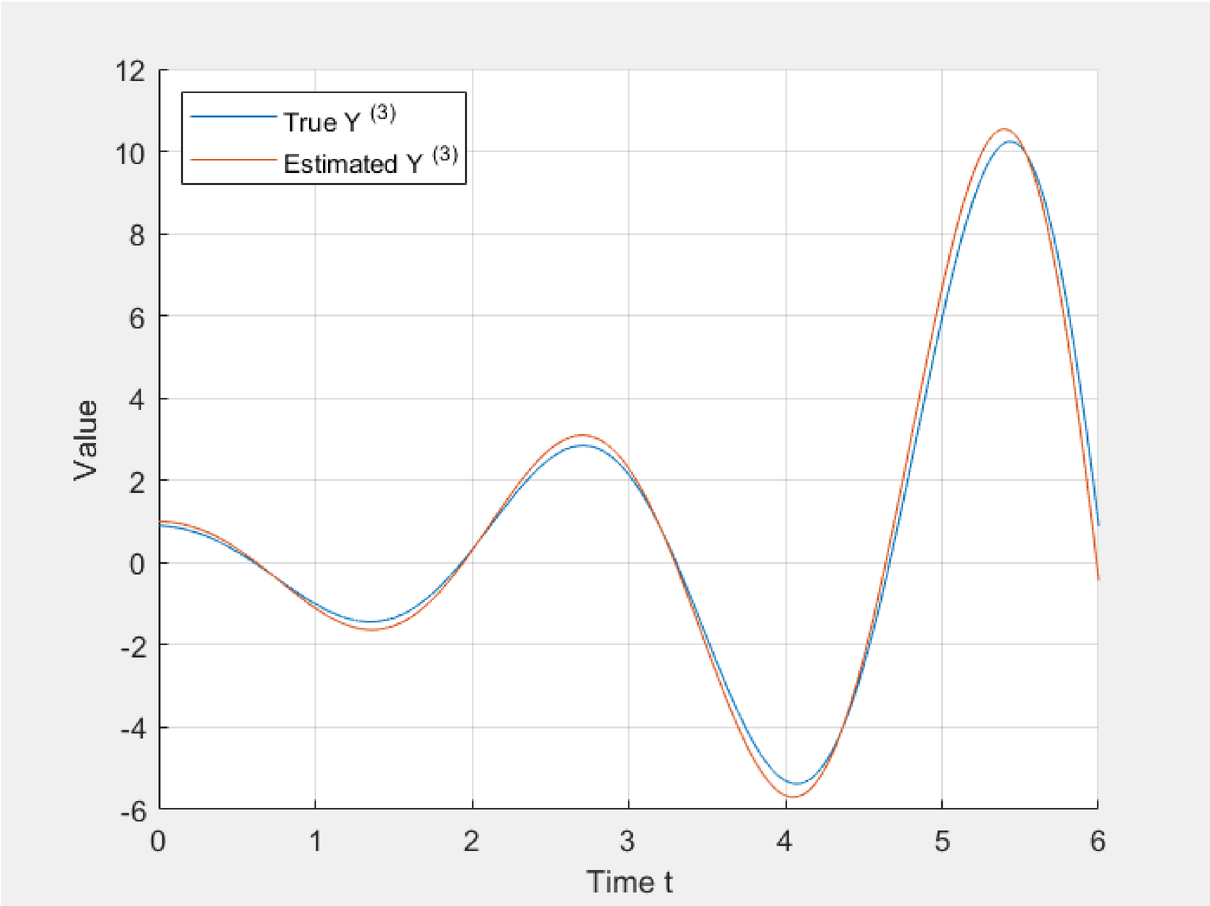

Figure 30: Estimate of $y^{(3)}$ vs. true $y^{(3)}$ for noise of 1.5 SD (SNR -12.5224 dB)

.

7.1.3 Noise $\sigma = 2$

Figure 31 shows y_M versus y, where the true y is displayed in red and the noisy y in blue. The sample size is $N = 10000$. We assumed the order of the system is between 2 to 6. The results of estimation for each order is shown in the Table 4.

Order	Values	a_0	a_1	a_2	a_3	a_4	a_5	RMSD	BIC
	True	1	5	5	0	-	-	-	-
2	Estimated	1.37	-0.45	-	-	-	-	8.7e-2	-5.76012e5
3	Estimated	0.23	6.24	-1.02	-	-	-	1.8e-3	-5.77252e5
4	Estimated	1.03	5.32	5.10	0	-	-	**1.9e-4**	**-5.77260e5**
5	Estimated	0.24	0	5.31	4.81	-0.0001	-	**1.7e-4**	-5.77250e5
6	Estimated	24.50	73.88	93.25	1.35	23.25	-0.16	5.1e-4	-5.77244e5

Table 4: Estimated parameter values with noise of 2 SD (SNR -15.00821 dB)

As the table shows the smallest value for BIC is associated with the fourth order order. It estimate the order correctly. Despite the fact that the RSMD of the fifth order is smaller, BIC for for fourth order order is selected.

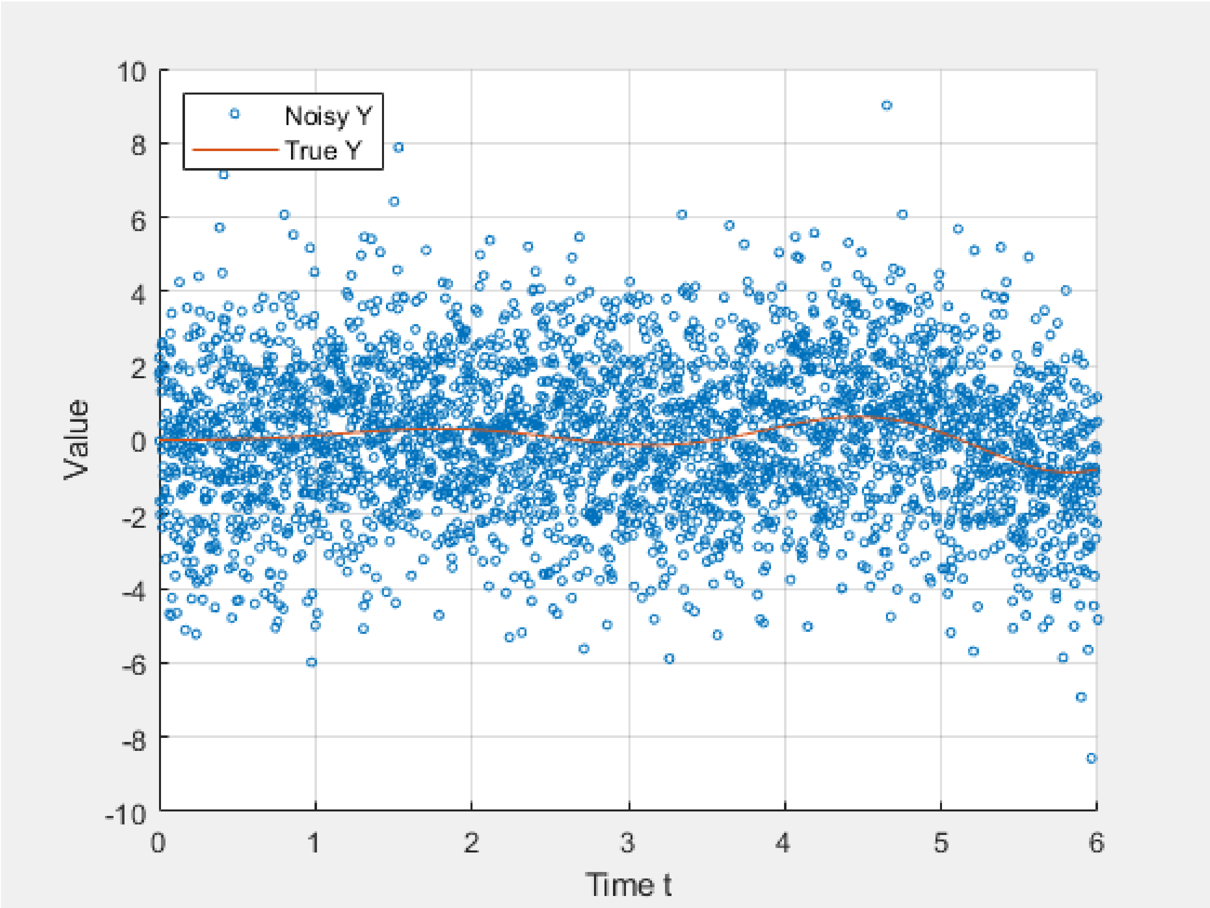

Figure 31: Noisy y_M vs. nominal y for noise of 2 SD (SNR -15.00821 dB)
.

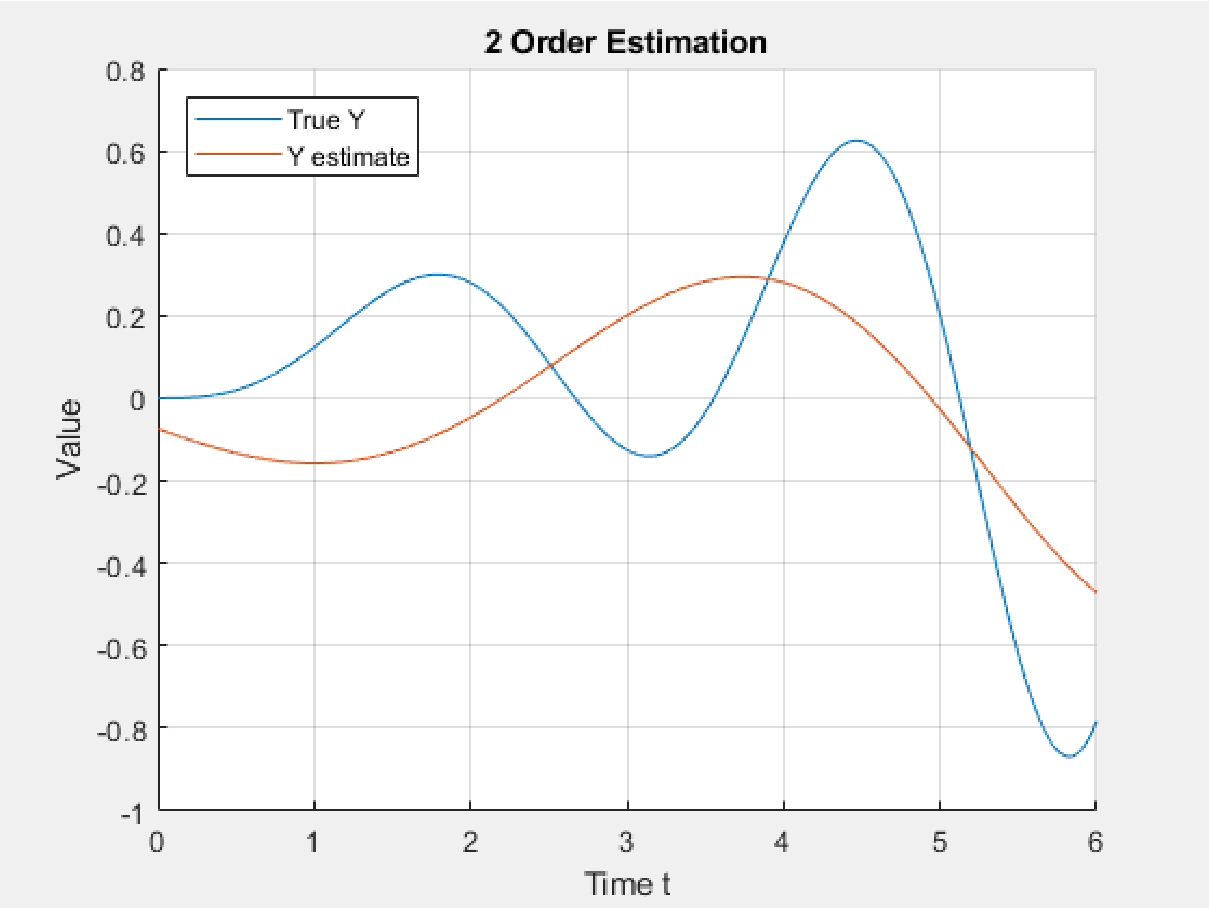

Figure 32: Estimate of y using second order vs. nominal y for noise of 2 SD (SNR -15.00821 dB)

.

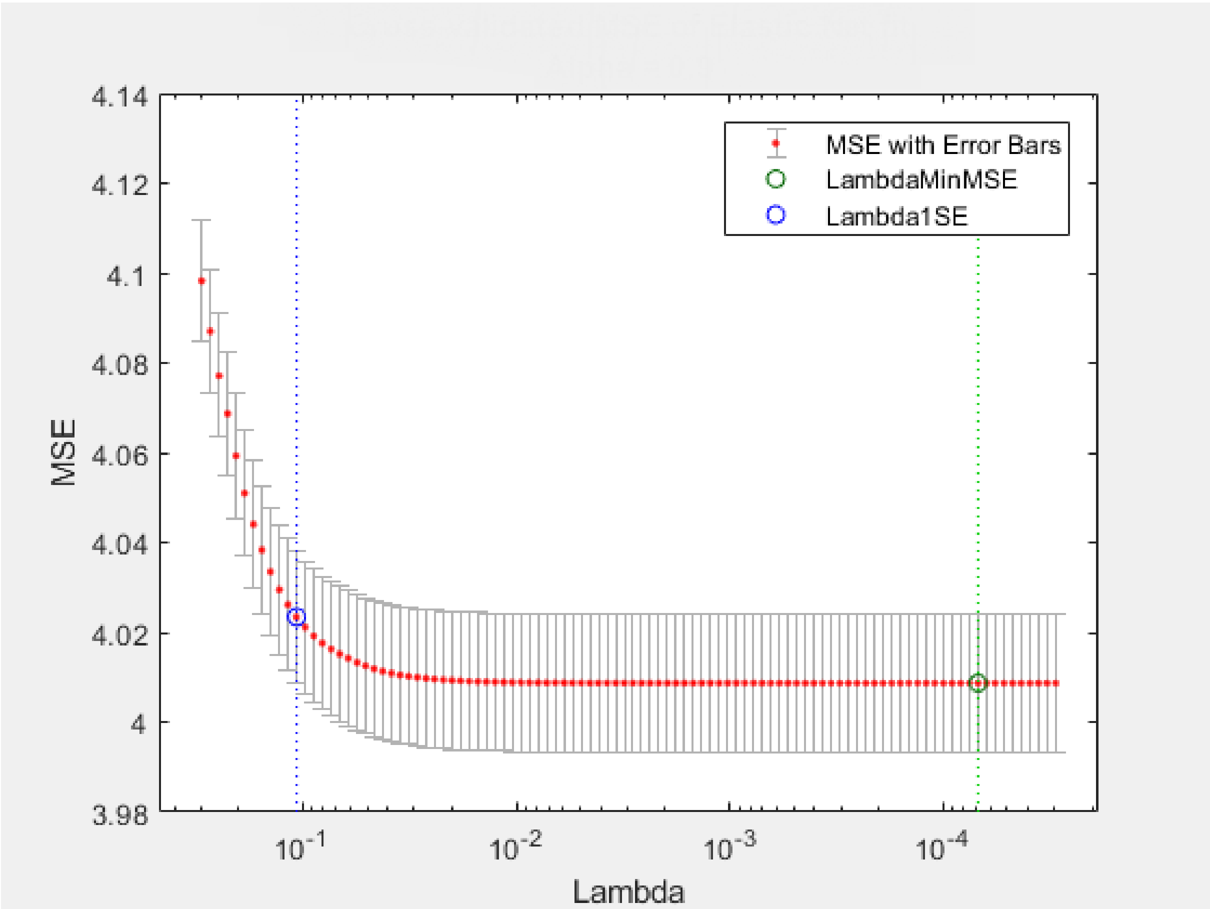

Figure 33: Cross-validation of MSE and λ for ridge regression of second order system for noise of 2 SD (SNR -15.00821 dB)

.

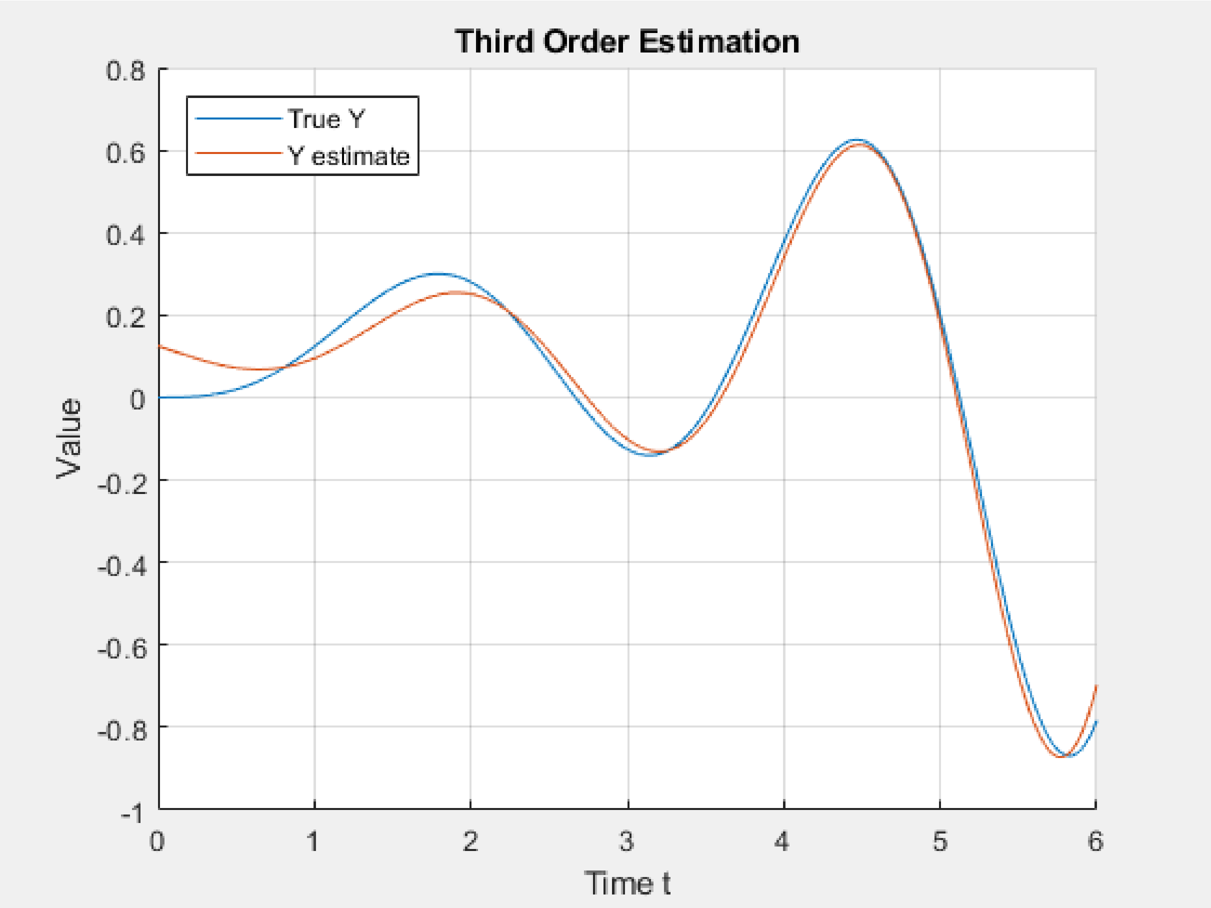

Figure 34: Estimate of y using third order vs. nominal y for noise of 2 SD (SNR -15.00821 dB)

.

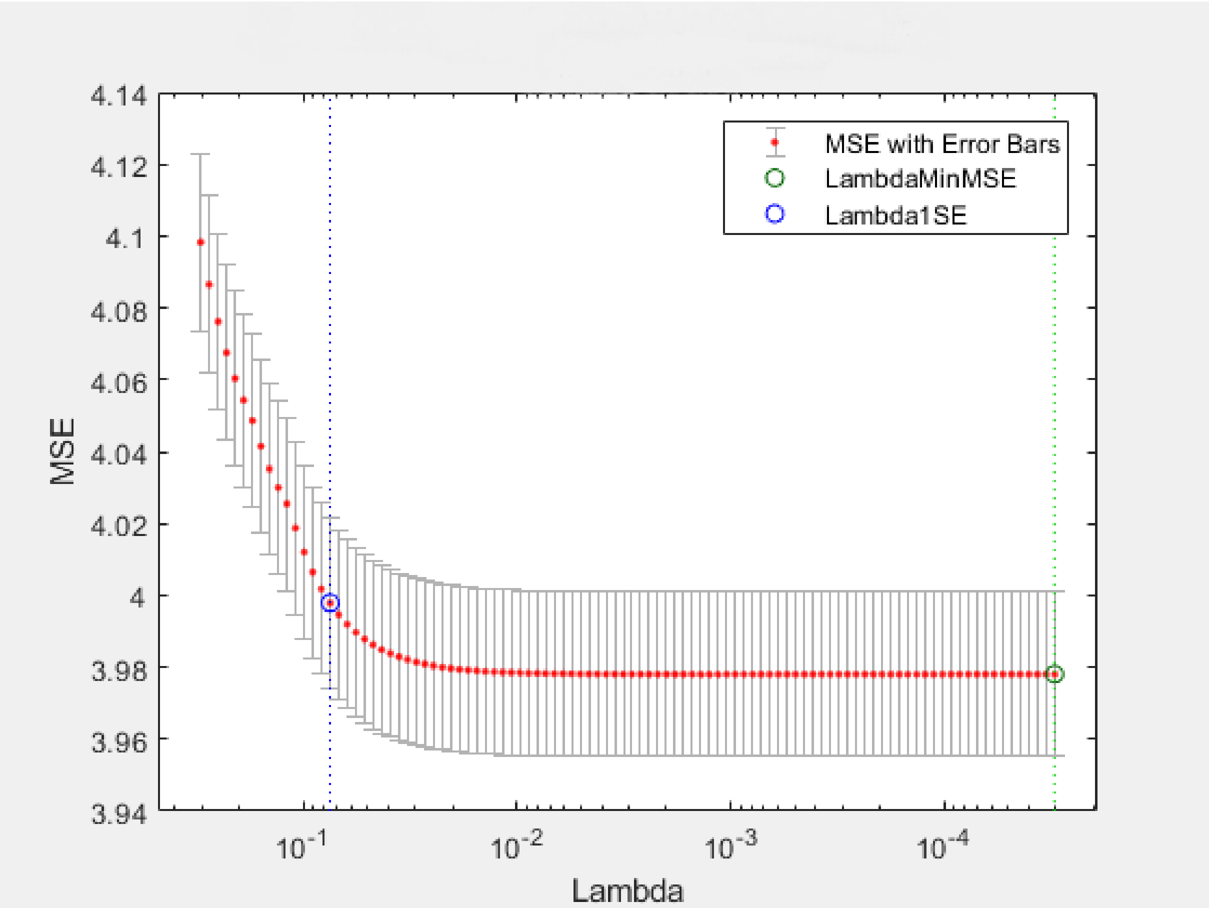

Figure 35: Cross-validation of MSE and λ for ridge regression of third order system for noise of 2 SD (SNR -15.00821 dB)

.

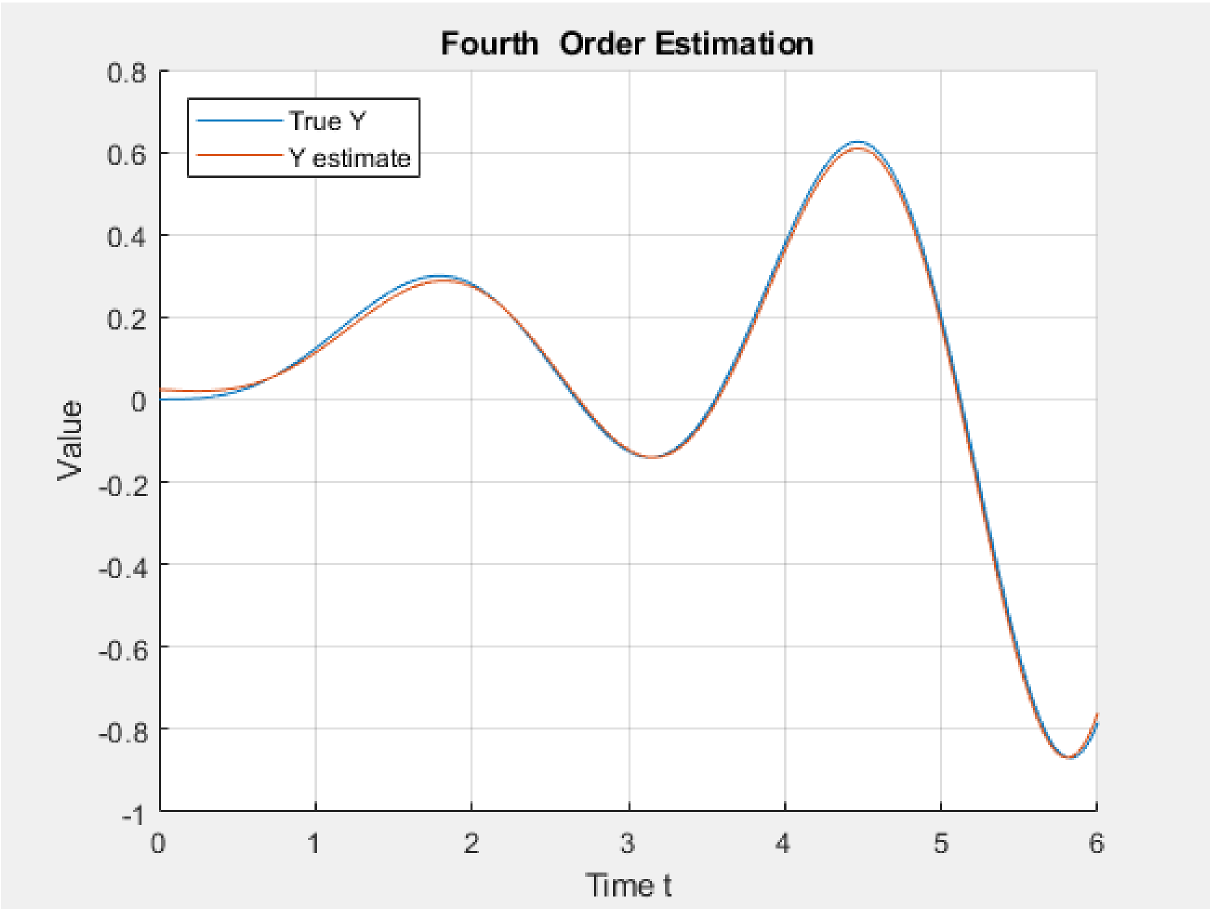

Figure 36: Estimate of y using fourth order vs. nominal y for noise of 2 SD (SNR -15.00821 dB)

.

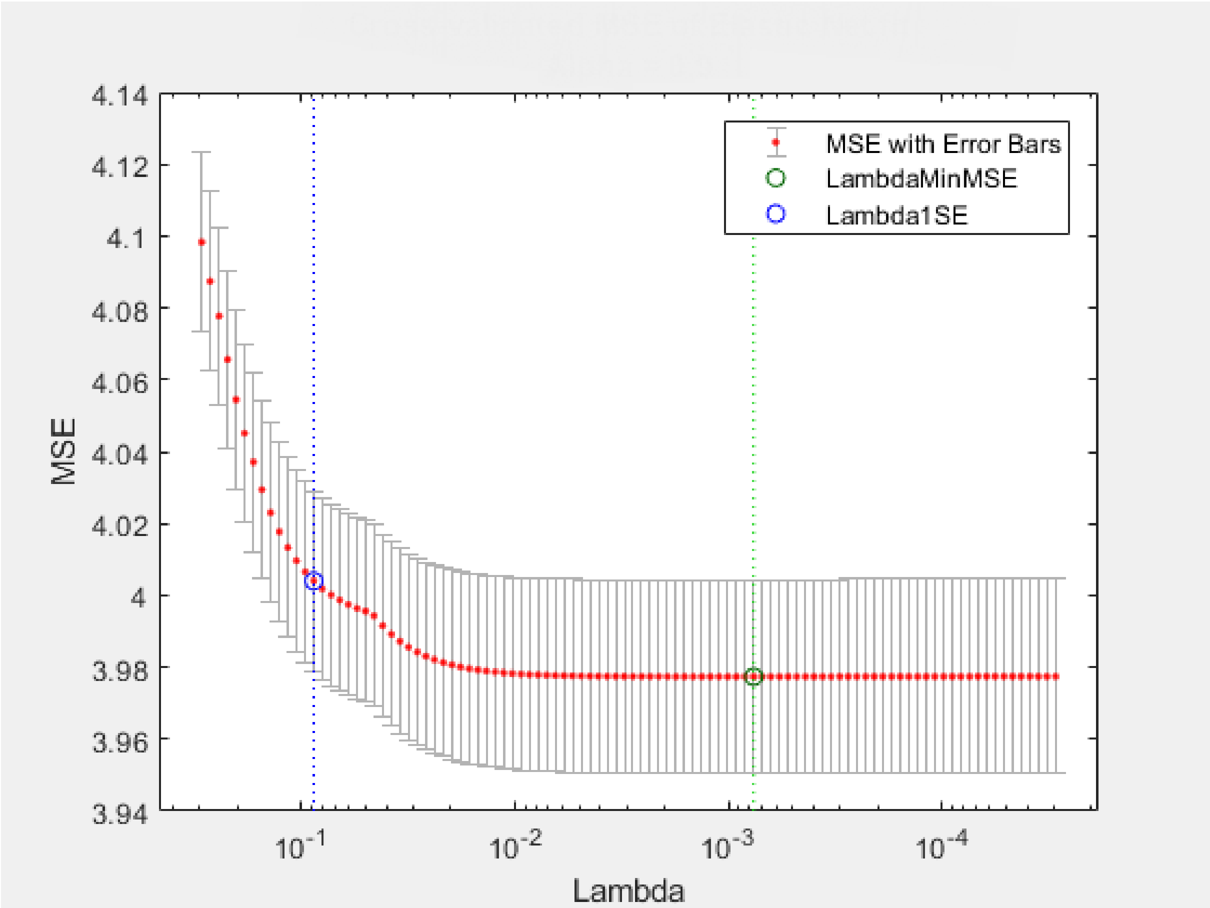

Figure 37: Cross-validation of MSE and λ for ridge regression of fourth order system for noise of 2 SD (SNR -15.00821 dB)

.

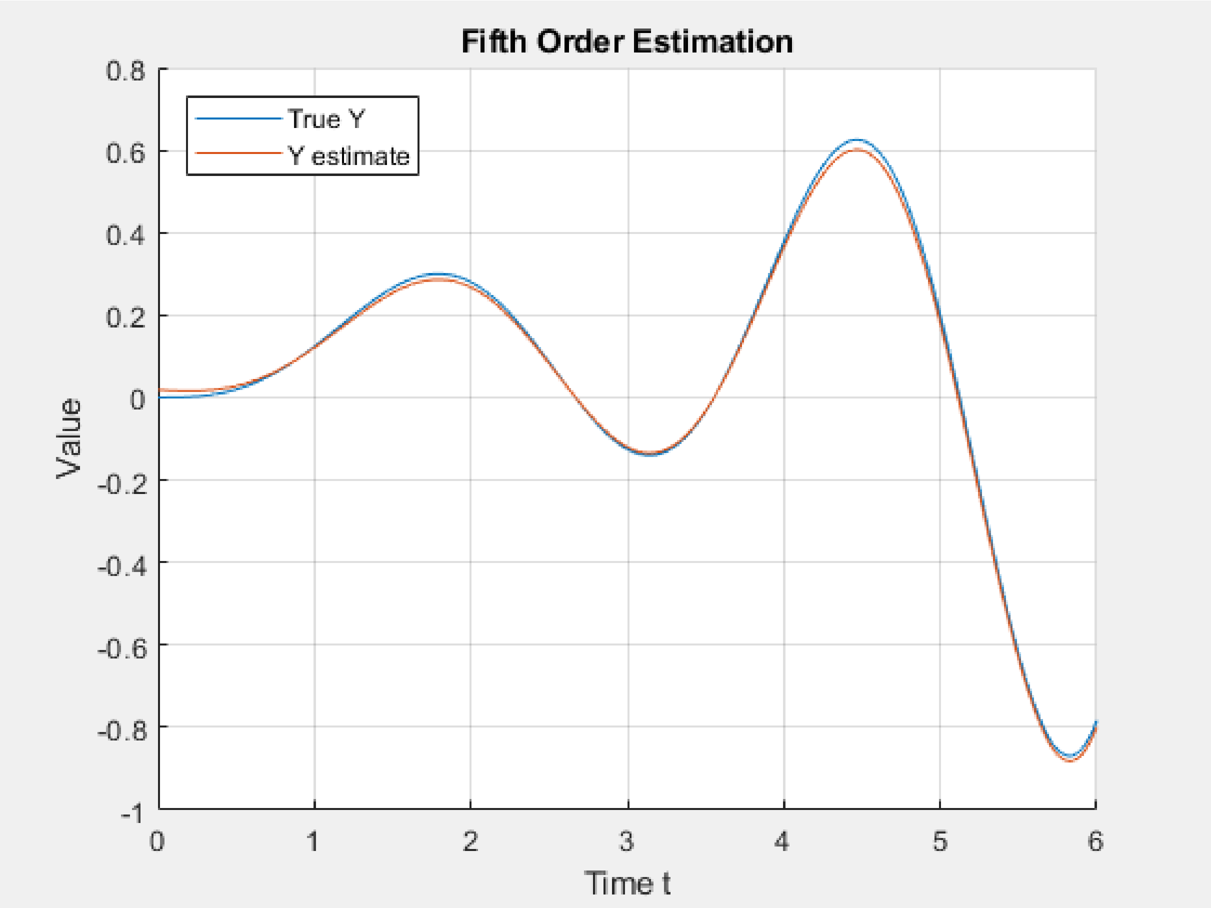

Figure 38: Estimate of y using fifth order vs. nominal y for noise of 2 SD (SNR -15.00821 dB)

.

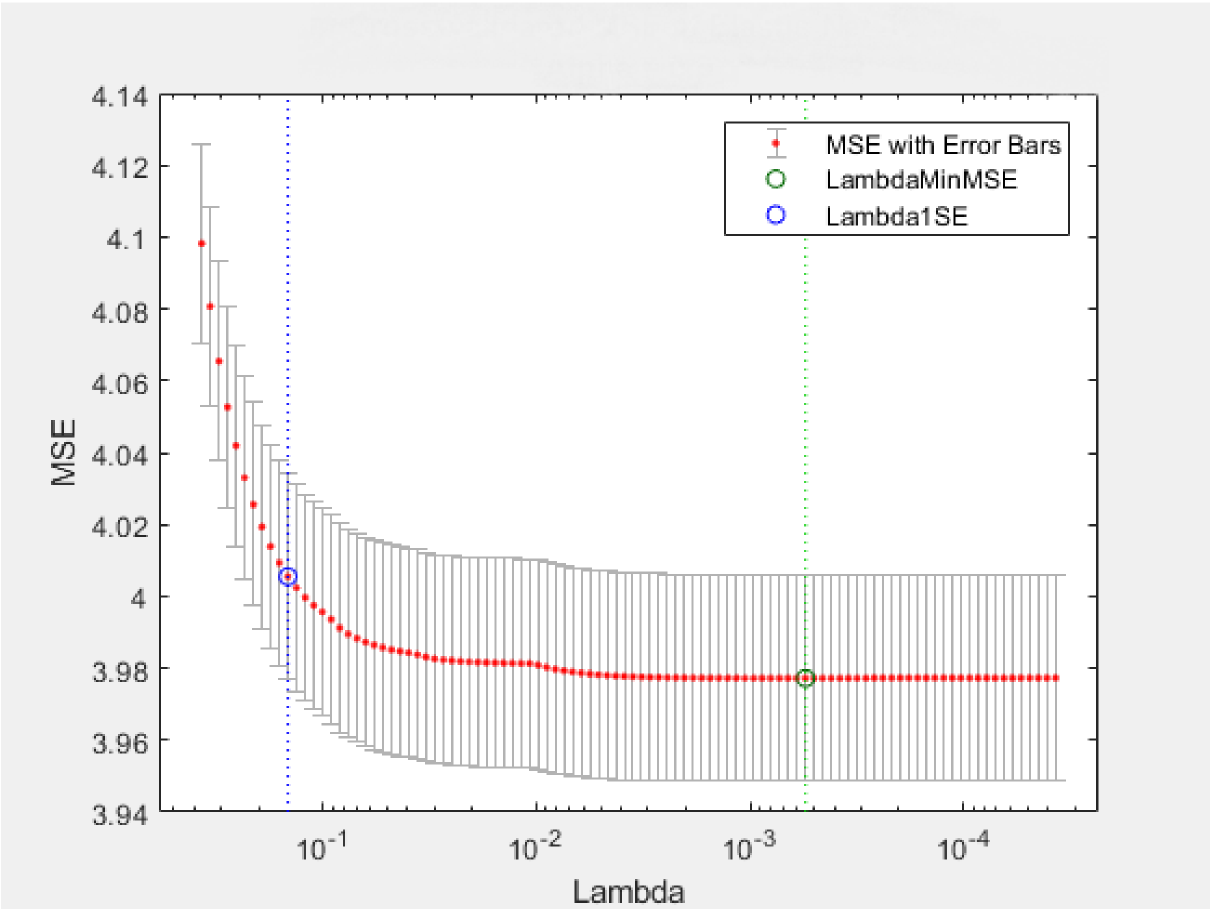

Figure 39: Cross-validation of MSE and λ for ridge regression of fifth order system for noise of 2 SD (SNR -15.00821 dB)

.

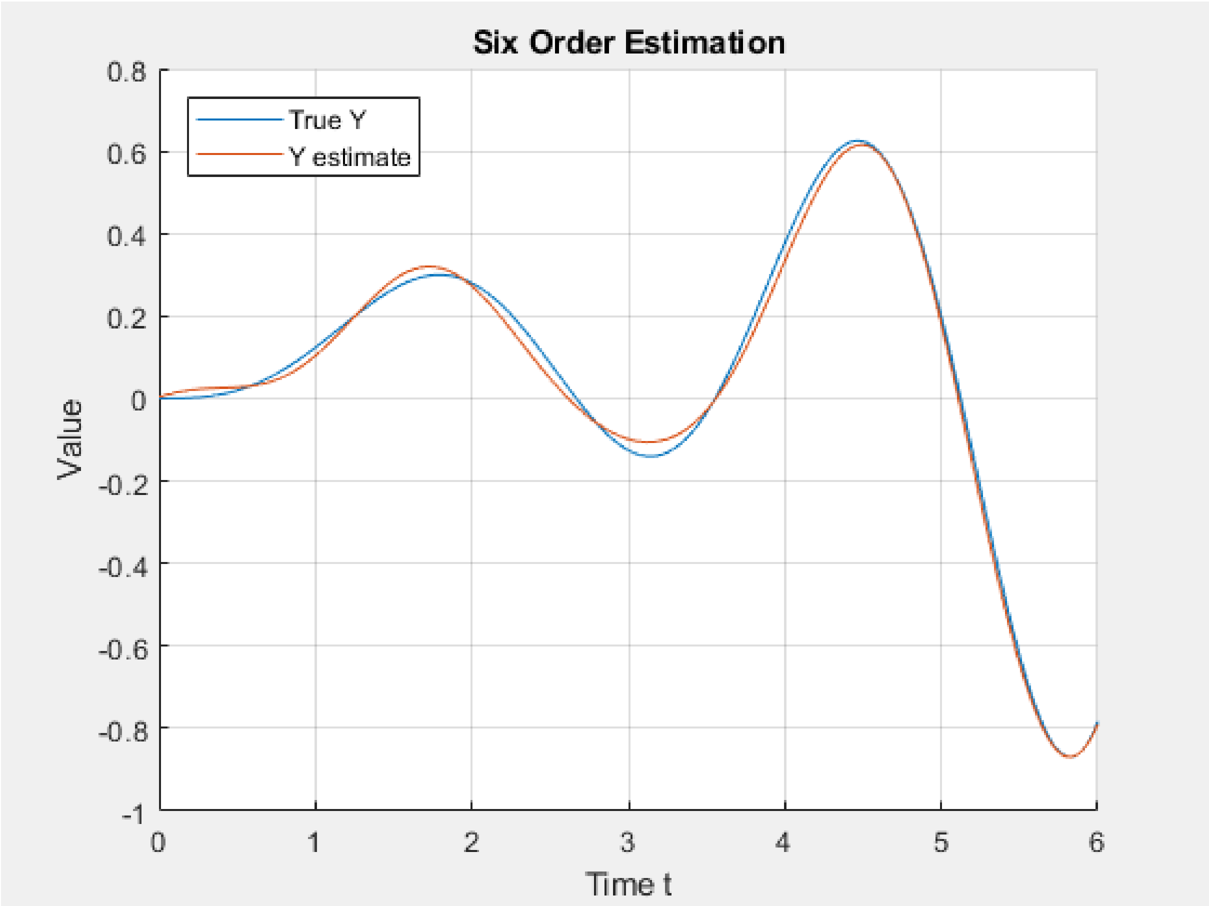

Figure 40: Estimate of y using sixth order vs. nominal y for noise of 2 SD (SNR -15.00821 dB)

.

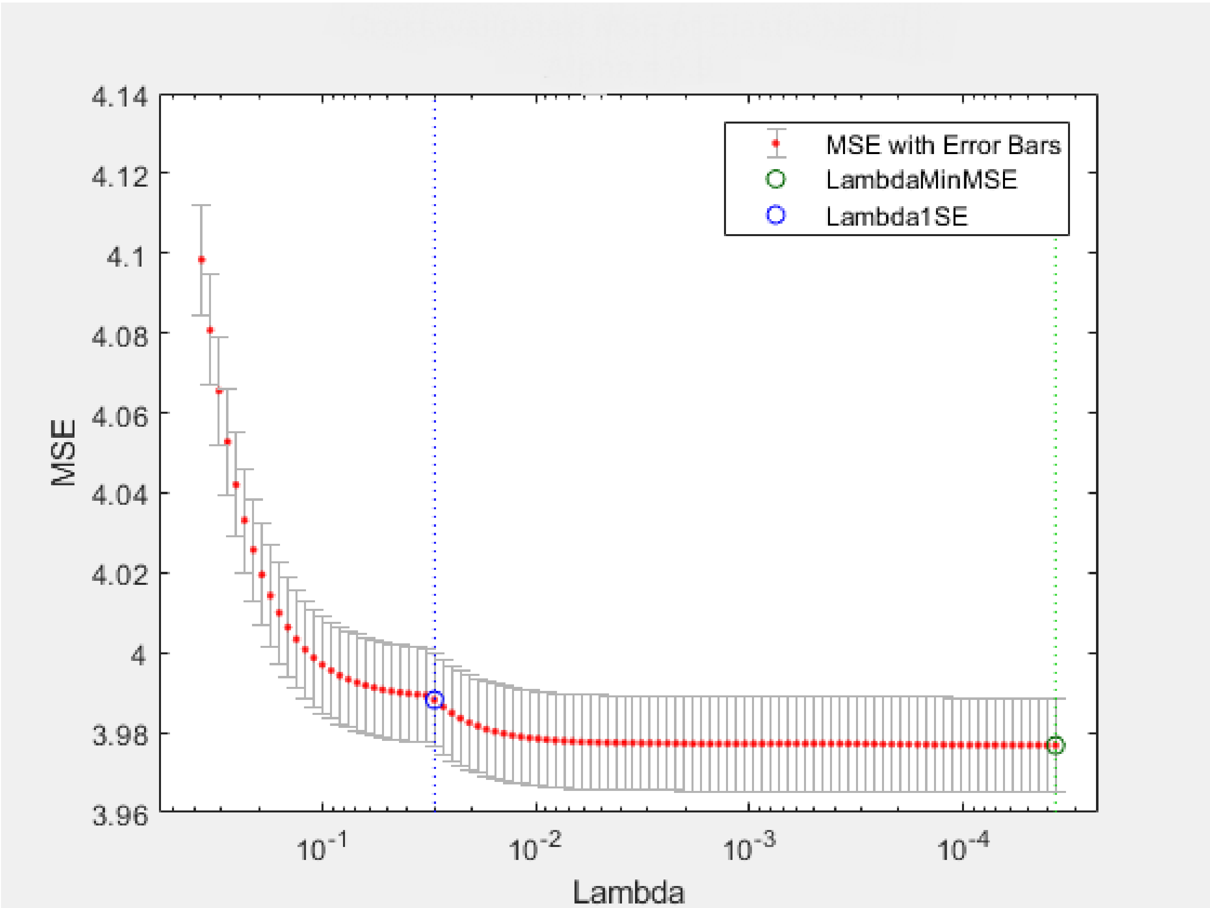

Figure 41: Cross-validation of MSE and λ for ridge regression of sixth order system for noise of 2 SD (SNR -15.00821 dB)

.

According to the table, using the BIC values, the estimated order is the fourth order order. As follows, the derivatives reconstructions are based on the fourth order order system.

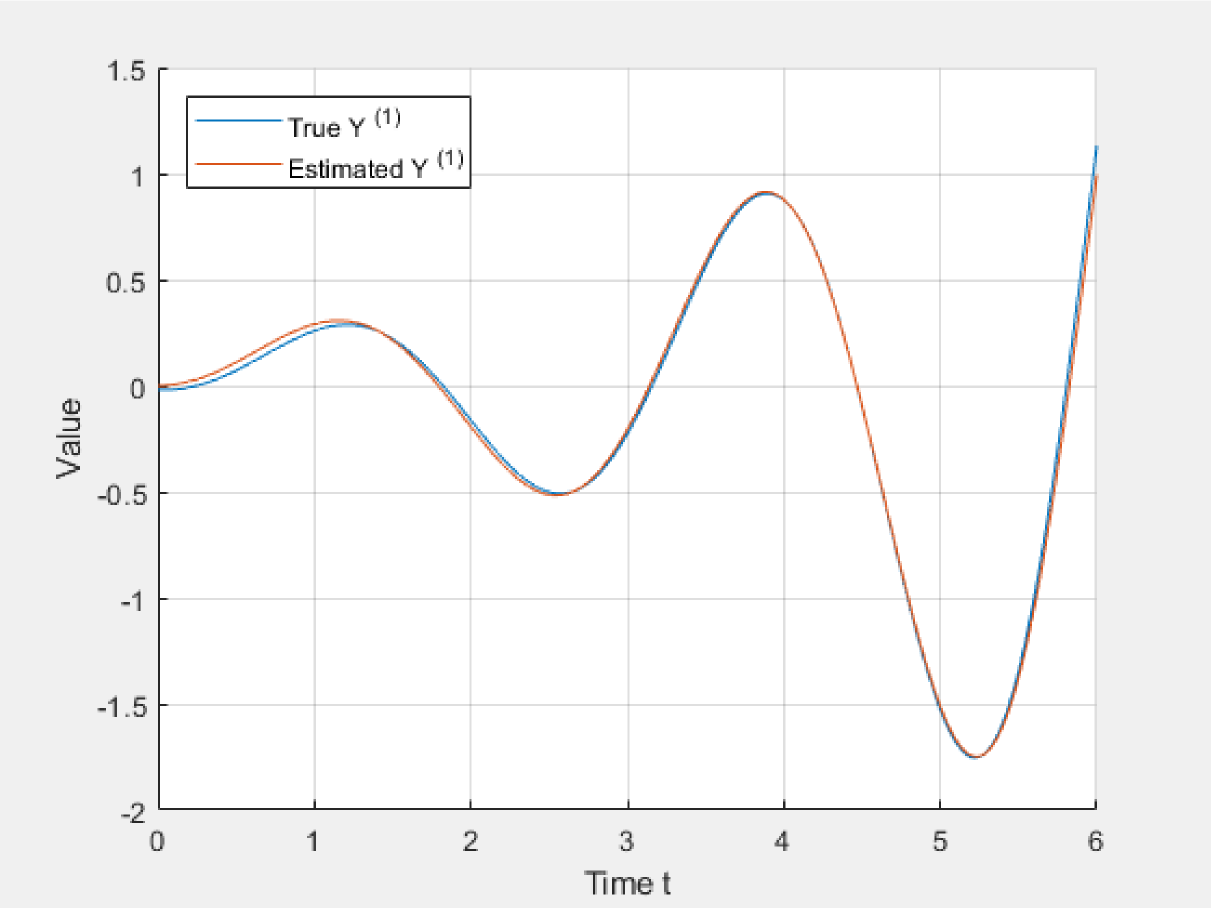

Figure 42: Estimate of $y^{(1)}$ vs. true $y^{(1)}$ for noise of 1 SD (SNR -15.00821 dB)

.

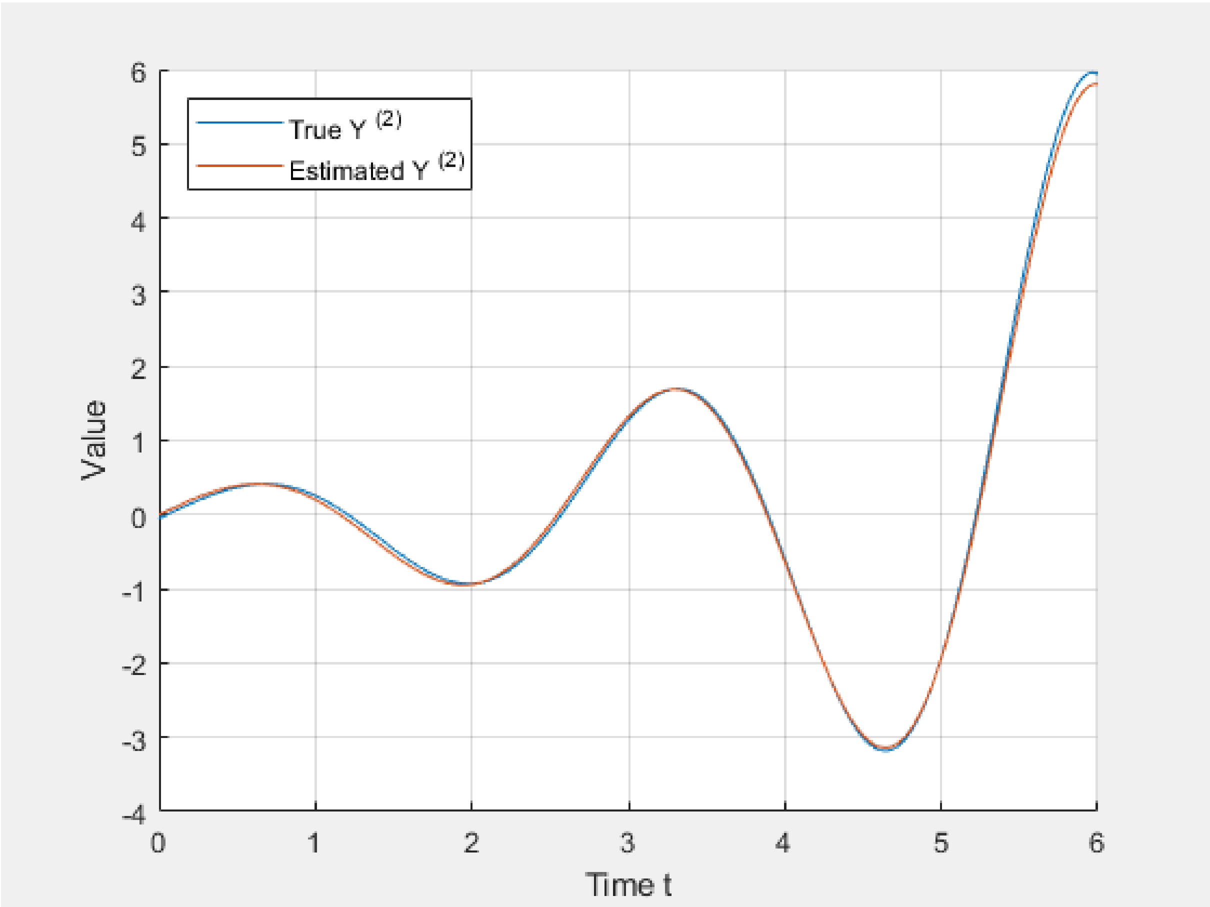

Figure 43: Estimate of $y^{(2)}$ vs. true $y^{(2)}$ for noise of 1 SD (SNR -15.00821 dB)

.

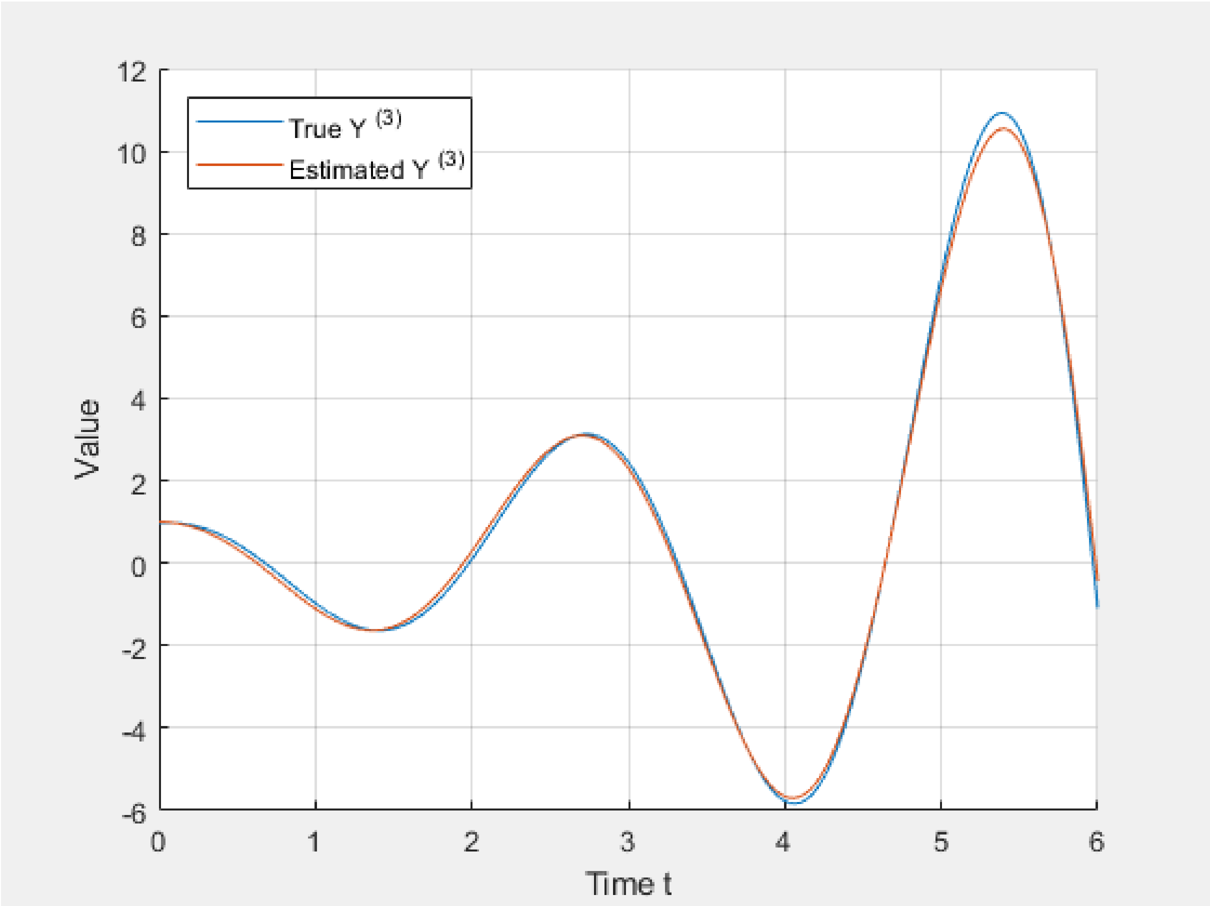

Figure 44: Estimate of $y^{(3)}$ vs. true $y^{(3)}$ for noise of 1 SD (SNR -15.00821 dB)

.

7.2 Compression with previous works

The method adopted in this thesis is compared with the *Recursive Least Square Algorithm* described in [26] and the *two step method* used by [41]. The system considered is:

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -150 & -125 & -31 & -5 \end{bmatrix} x \ ; y = x_1 \ ; \ x(0) = [1, 1, 1, 1] \tag{81}$$

with its corresponding characteristic equation

$$y^{(4)}(t) + a_3 y^{(3)}(t) + a_2 y^{(2)}(t) + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{82}$$

where,

$$a_0 = 150, a_1 = 125, a_2 = 31, a_3 = 5$$

with white noise of variance of 1 (high noise with SNR of -5.84dB) are added to the system. Below are the results obtained employing all three methods.

Variance	Method	a_0	a_1	a_2	a_3
1	True	150	125	30	5
	Ghosal et al.	45.376	51.867	23.366	2.218
	John [41]	72.322	71.5791	28.3529	2.8245
	Proposed method	**157**	**124**	**30.8**	**4.99**

Table 5: Comparative study with Ghoshal et al. and John

The method adopted in this thesis is proven to be more accurate. The most significant advantage of the proposed method is the time of computation. The proposed method is faster than both previous method.

Variance	Method	Time(sec)
1	Ghosal et al.	18000
	John [41]	23040
	Proposed method	**643**

Table 6: Comparison of the computational time with Ghoshal et al. and John

The result of employing our method with noise variance of 1 are represented here with the number of sample points $N = 100000$ which were selected to be equidistant in $[0, 5]$.

Once the parameter estimates are obtained, we reconstruct the output and its derivatives as mentioned earlier.

$$\text{MSE} = \frac{1}{N}\text{RSS} = \frac{1}{N}\sum (Y_{true}(t) - Y_{est}(t))^2 = 2.8e^{-4} \tag{83}$$

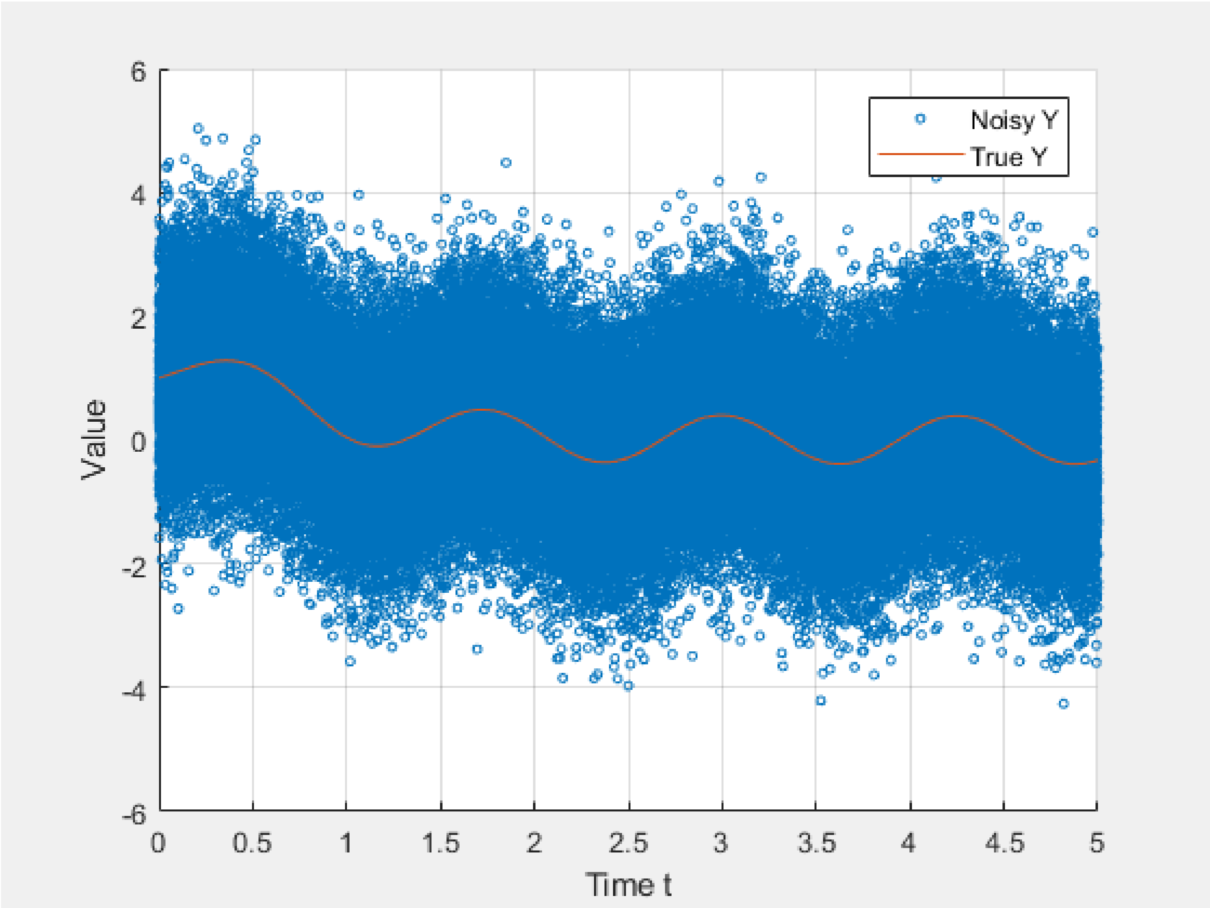

Figure 45: Noisy y_M vs. nominal y for noise of 1 SD (SNR -5.84 dB)
.

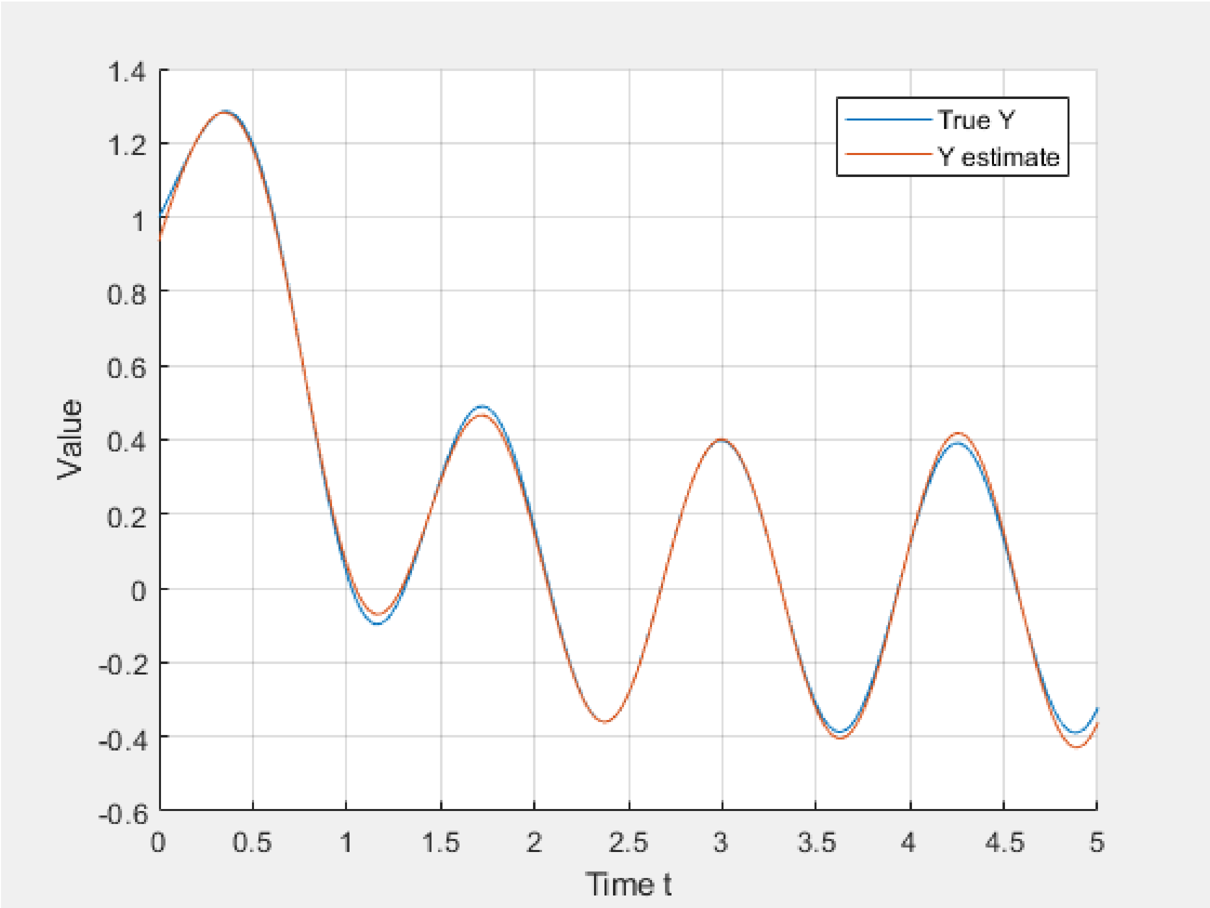

Figure 46: Estimate of y using fourth order vs. nominal y for noise of 1 SD (SNR -5.84 dB)

.

The derivatives:

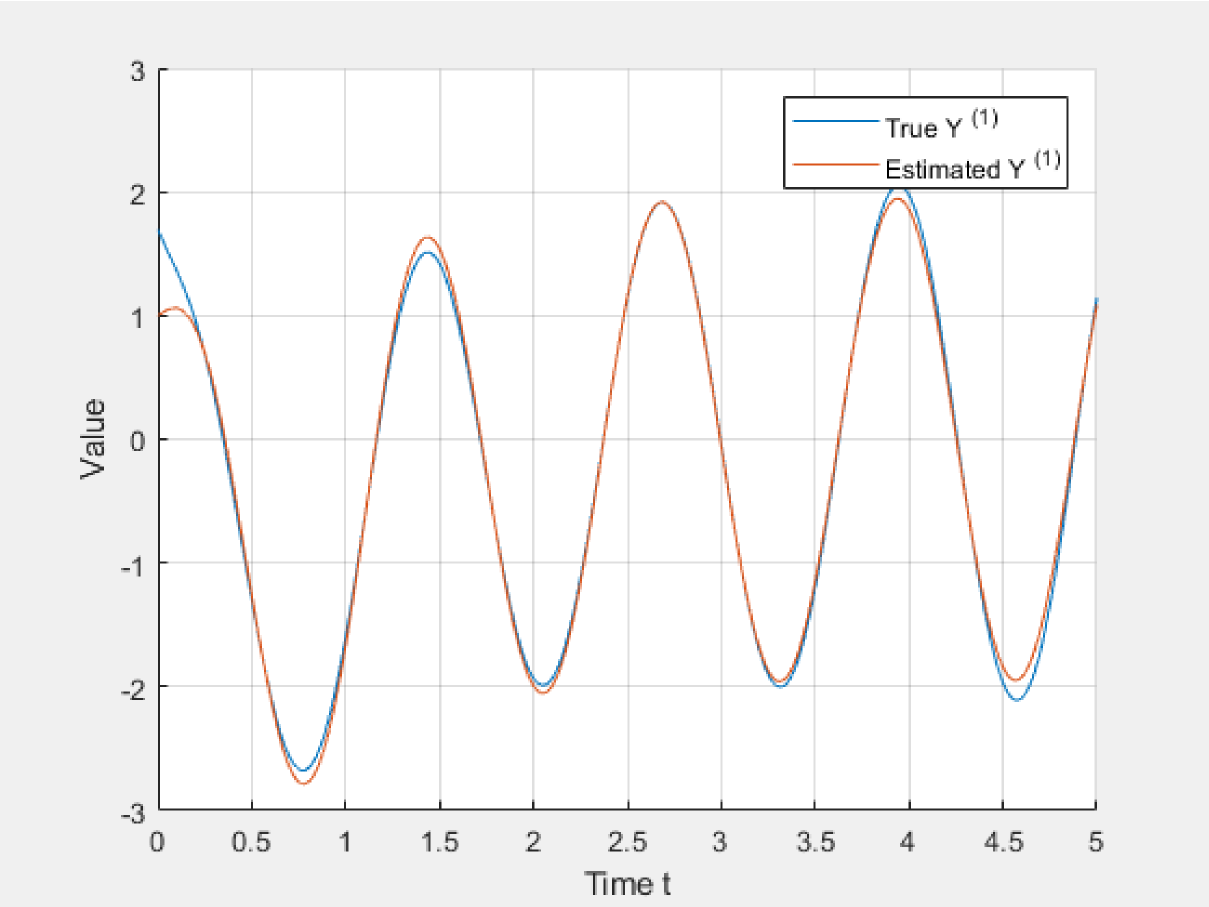

Figure 47: Estimate of $y^{(1)}$ vs. true $y^{(1)}$ for noise of 1 SD (SNR -5.84 dB)
.

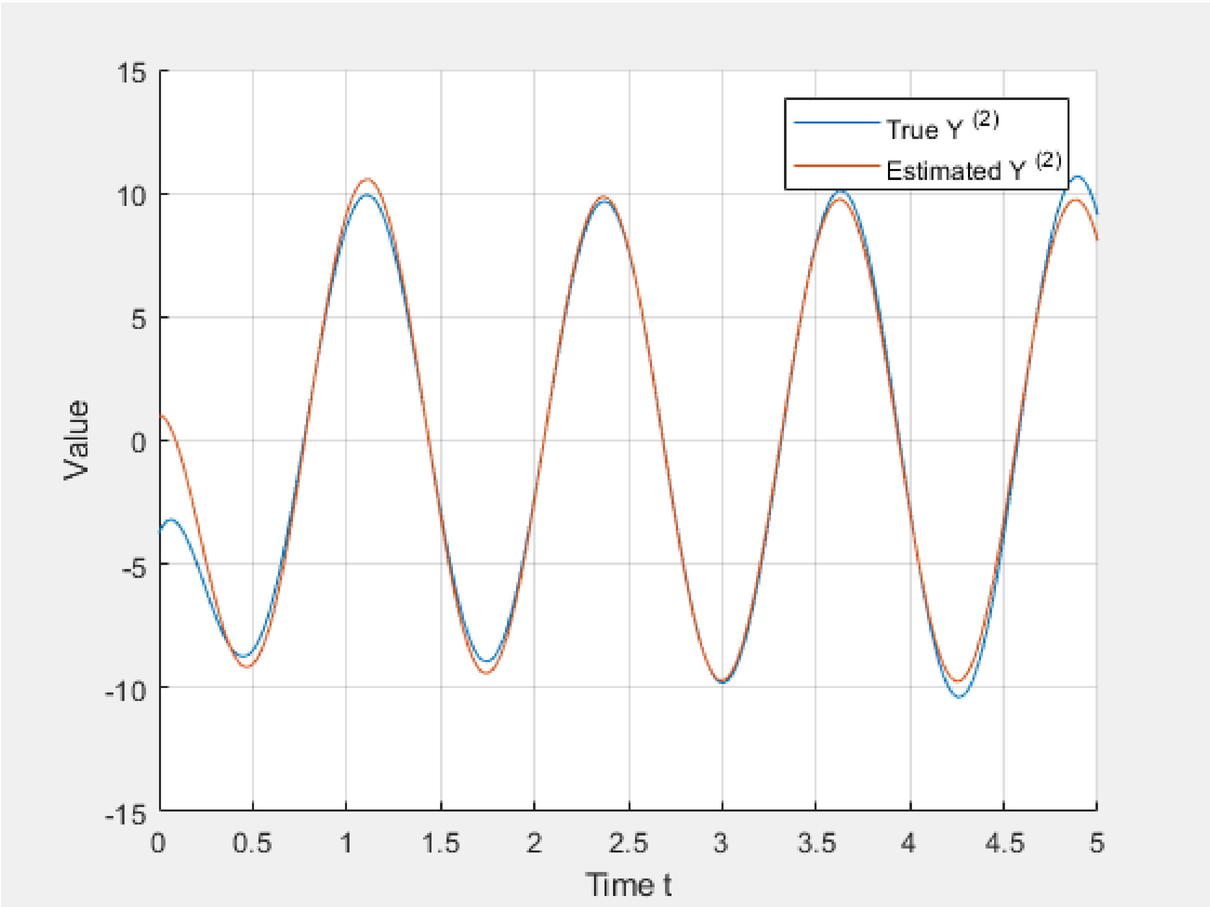

Figure 48: Estimate of $y^{(2)}$ vs. true $y^{(2)}$ for noise of 1 SD (SNR -5.84 dB)
.

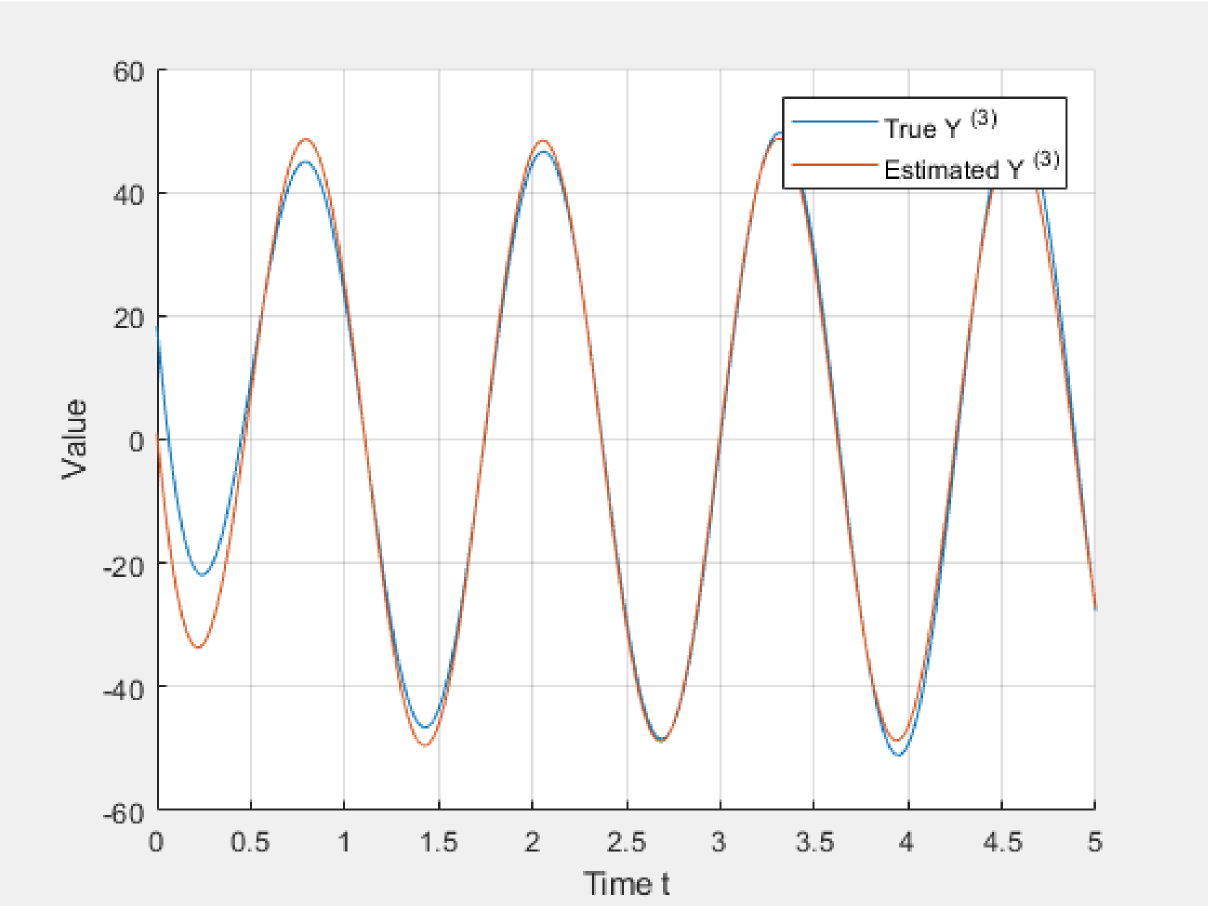

Figure 49: Estimate of $y^{(3)}$ vs. true $y^{(3)}$ for noise of 1 SD (SNR -5.84 dB)

.

7.3 Discussion

Several conclusions can be drawn from all the results presented.

(1) It is noted that the reconstructed output and derivative trajectories are virtually indistinguishable from their nominal counterparts for the rather noisy measurement process realization with large SNR -8.9652 dB, hence confirming the power of the estimation approach. Even with the measurement noise amounting to SNR -12.4797 dB, the estimates are remarkably close to their true counterparts.

(2) The suggested method of this thesis outperforms the exiting methods presented in [41] and [28], both in **accuracy** and **computing time**.

(3) The order does not need to be known for the system, however the range of possible orders is required for order estimation procedure to work.

(4) The parameter estimation procedure used here suffers from a clear deficiency. It makes no attempt to employ the derivative kernels *directly* in the cost function of the estimation process. It is rather remarkable that using the reproducing property as the only relation from which to discern the values of the components of a multi-dimensional parameter vector can deliver such good results. It is well known that identification of higher dimensional ARX time-series models must, albeit indirectly, employ information about the time derivatives of the system output. This fact is also clear from the algebraic identifiability conditions that explicitly employ output derivatives.[62]

(5) As the system is differentially flat (see [74]), the state vector x_E, for $t \in [a, b]$, can be faithfully recovered from estimated y_E and its derivatives.

(6) The estimation method can be immediately improved by considering the application of some version of Hilbert Kernels penalty term in conjunction with the cost used here by which the variance of the noise could be considered in one step in the estimation of parameters.

(7) In closing, it is worth pointing out the obvious: that the accuracy of estimation will deteriorate as the noise level increases without bound.

Chapter 8

8 Remarks and Future Work

State estimation, parameter estimation and filtering are important in many applications and various fields such as engineering, biology, psychology and finance. The body of work on system identification and filtering is extremely vast; the problem varying from classical methods [53], [4] and [35] of parameter estimations, methods involving data assimilation [49], recursive stochastic filtering [46] and [44]. Because of the ease and efficiency of noise removal, recursive stochastic algorithms are considered to be the best. However assumptions about the initial state and measurement noise characteristics, are the apparent drawbacks of these methods.

As previously mentioned, algebraic methods can overcome many of the drawbacks of the classical approach. However, they are computationally heavy, Thus the use of local smoothing can help with reducing the noise variance.

This thesis extends on the previous work of [41] which was based on algebraic methods of system identification developed in [19], [22], [74] and [59], and invariant observers [9] and [56]. The approach presented extends to time-varying systems, linear parameter varying systems, and multi-variate systems. In some aspects it is competitive with the continuous time Wiener filter without relying on frequency domain techniques. The method can easily be adapted to handle zero-mean coloured noise and applies to systems with and without input of arbitrary order, with very little a priori information available. There exist many improvements that can be done in future works.

To summarize the advantages of the proposed method:

- The order of the system is not required to be known and can be estimated.
- The method can handle large noise (SNR -12dB).
- The knowledge of initial conditions is not required.
- It can estimate the parameter of the system with accuracy in a timely manner.
- The method can reconstruct the output and derivative trajectories which are virtually indistinguishable from their nominal counterparts.

Future Work

Some directions for future improvements have already been indicated in our discussion of results. The system needs many data samples that makes the deployment of it challenging and not robust. The color noise has not been investigated here and can be possible direction. Finally, another topic to investigate is how to consider colored noise in a nonlinear setting.

Appendices

A A Kernel representation of the n-order SISO LTI system

This work is the summary of the work done by [41], [62] and [25].

A.1 Explicit Kernel representation of the n-order SISO LTI system

[25] The general characteristic equation of a n-th order system is written as below:

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1 y^{(1)}(t) + a_0 y(t) = 0 \tag{84}$$

Multiplying equation (84) by the term $(\xi - a)^n$, yields

$$(\xi - a)^n y^{(n)} + a_{n-1}(\xi - a)^n y^{(n-1)} + \cdots + a_1(\xi - a)^n y^{(1)} + a_0(\xi - a)^n y = 0 \tag{85}$$

This will be integrated multiple times on the interval $[a, a + \tau]$.

Multiplying equation (84) by the term $(b - \zeta)^n$, yields

$$(b - \zeta)^n y^{(n)} + a_{n-1}(b - \zeta)^n y^{(n-1)} + \cdots + a_1(b - \zeta)^n y^{(1)} + a_0(b - \zeta)^n y = 0 \tag{86}$$

This will be integrated multiple times on the interval $[b - \sigma, b]$.

To lower the degree of the derivatives, integration by parts is used. The result is then simplified algebraically.

Integration of the general term $(\xi - a)^n y^{(m)}$, (where $n \geq m$) over the interval $[a, a + \tau]$ yields the following expression:

$$\int_a^{a+\tau} (\xi - a)^n y^{(m)}(\xi) d\xi =$$
$$\sum_{i=0}^{m-1} (-1)^i \frac{n!}{(n-i)!} \tau^{n-i} y^{(m-i-1)}(a + \tau) + (-1)^m \frac{n!}{(n-m)!} \int_a^{a+\tau} (\xi - a)^{n-m} y(\xi) d\xi$$

(See [31].)

The kernel derivation for the output function y

An n-time integration of the general term $(\xi - a)^n y^{(m)}$ over the interval $[a, a+\tau]$ yields the following expressions:

For $n = m$,

$$\int^{(n)} (\xi - a)^n y^{(m)}(\xi) d\xi \ldots d\xi''^{\cdots} = \tau^n y + \sum_{j=1}^{n} (-1)^j \binom{n}{j} \frac{n!}{(n-j)!} \int^{(j)} (\xi - a)^{n-j} y(\xi) d\xi \ldots d\xi''^{\cdots}$$

For $n > m$,

$$\int^{(n)} (\xi - a)^n y^{(m)}(\xi) d\xi \ldots d\xi''^{\cdots} = \sum_{j=0}^{m} (-1)^j \binom{m}{j} \frac{n!}{(n-j)!} \int^{(n-m+j)} (\xi - a)^{n-j} y(\xi) d\xi \ldots d\xi''^{\cdots}$$

n-times integration of equation (85) on the interval $[a, a+\tau]$ gives

$$\tau^n y + \sum_{j=1}^{n} (-1)^j \binom{n}{j} \frac{n!}{(n-j)!} \int^{(j)} (\xi - a)^{n-j} y(\xi) d\xi \ldots d\xi''^{\cdots} + \sum_{i=o}^{n-1} a_i P_y(n,i) = 0$$

where

$$P_y(n,i) = \sum_{j=0}^{i} (-1)^j \binom{i}{j} \frac{n!}{(n-j)!} \int^{(n-i+j)} (\xi - a)^{n-j} y(\xi) d\xi \ldots d\xi''^{\cdots}$$

Applying the Cauchy formula for repeated integrals, while setting $\tau + a = t$,

$$(t-a)^n y + \sum_{j=1}^{n} (-1)^j \binom{n}{j} \frac{n!}{(n-j)!(j-1)!} \int_a^t (t-\xi)^{j-1} (\xi - a)^{n-j} y(\xi) d\xi + \sum_{i=0}^{n-1} a_i \overline{P}_y(n,i) = 0$$

where

$$\overline{P}_y(n,i) = \sum_{j=0}^{i} (-1)^j \binom{i}{j} \frac{n!}{(n-j)!(n-i+j-1)!} \int_a^t (t-\xi)^{n-i+j-1} (\xi - a)^{n-j} y(\xi) d\xi$$

Swapping $\int$ and $\sum$ yields the forward kernel after some algebraic manipulation.

Forward kernel for output y

$$y(t) = \frac{1}{(t-a)^n}\left[\int_a^t K_{F,y}(n,t,\tau)y(\tau)d\tau\right] \tag{87}$$

where

$$\begin{aligned}K_{F,y}(n,t,\tau) = &\sum_{j=1}^{n}(-1)^{j+1}\binom{n}{j}\frac{n!}{(n-j)!(j-1)!}(t-\tau)^{j-1}(\tau-a)^{n-j}\\ &+\sum_{i=0}^{n-1}a_i\sum_{j=0}^{i}(-1)^{j+1}\binom{i}{j}\frac{n!}{(n-j)!(n-i+j-1)!}(t-\tau)^{n-i+j-1}(\tau-a)^{n-j}\end{aligned}$$

The backward kernel is obtained similarly.

Backward kernel for output y

$$y(t) = \frac{1}{(b-t)^n}\left[\int_t^b K_{B,y}(n,t,\tau)y(\tau)d\tau\right] \tag{88}$$

where

$$\begin{aligned}K_{B,y}(n,t,\tau) = &\sum_{j=1}^{n}\binom{n}{j}\frac{n!}{(n-j)!(j-1)!}(t-\tau)^{j-1}(b-\tau)^{n-j}\\ &+\sum_{i=0}^{n-1}a_i\sum_{j=0}^{i}\binom{i}{j}\frac{n!}{(n-j)!(n-i+j-1)!}(t-\tau)^{n-i+j-1}(b-\tau)^{n-j}\end{aligned}$$

Adding equations (87) and (88), yields

$$y(t) = \frac{1}{(t-a)^n+(b-t)^n}\left[\int_a^b K_{DS,y}(n,t,\tau)y(\tau)d\tau\right]$$

where

$$K_{DS,y}(n,t,\tau) \triangleq \begin{cases}K_{F,y}(n,t,\tau), & \text{for } \tau \le t\\ K_{B,y}(n,t,\tau), & \text{for } \tau > t\end{cases}$$

A.2 Explicit kernel expressions for the derivatives of the output function[28]

It is straightforward to obtain the corresponding recursive formulae for the time derivatives of the system output

$y^{(i)}, i = 1, \cdots, n-1.$

There exist Hilbert-Schmidt kernels $K_{F,k,y}$, $K_{F,k,u}$, $k = 1, \cdots, n-1$ such that the derivatives of the output function in (3) can be computed rercursively as follows:

To get the expression for the kernel for $y^{(k)}$, the number of integrations to be performed is $(n-k)$ times.

Letting

$$p = n - k \ ; \ k \geq 1 \ ; \ p < n$$

and integrating p times the general term $(\xi - a)^n y^{(m)}$ over the interval $[a, a+\tau]$ yields the following expressions:

For $p \leq m$,

$$\int^{(p)} (\xi - a)^n y^{(m)}(\xi) d\xi \ldots d\xi''^{\cdots} = \sum_{i=0}^{m-p} (-1)^i \binom{p+i-1}{i} \frac{n!}{(n-i)!} \tau^{n-i} y^{(m-i-p)}(a+\tau)$$
$$+ \sum_{j=1}^{p} (-1)^{j+m-p} \binom{m}{m-p+j} \frac{n!}{(n-m+p-j)!} \int^{(j)} (\xi - a)^{n-m+p-j} y(\xi) d\xi \ldots d\xi''^{\cdots}$$

For $p > m$,

$$\int^{(p)} (\xi - a)^n y^{(m)}(\xi) d\xi \ldots d\xi''^{\cdots} = \sum_{j=o}^{m} (-1)^j \binom{m}{j} \frac{n!}{(n-j)!} \int^{(p-m+j)} (\xi - a)^{n-j} y(\xi) d\xi \ldots d\xi''^{\cdots}$$

A p-times integration of equation (85) on the interval $[a, a+\tau]$ gives

$$\sum_{i=0}^{n-p}(-1)^i\binom{p+i-1}{i}\frac{n!}{(n-i)!}\tau^{n-i}y^{(n-i-p)}(a+\tau)$$
$$+\sum_{j=1}^{p}(-1)^{j+n-p}\binom{n}{n-p+j}\frac{n!}{(p-j)!}\int^{(j)}(\xi-a)^{p-j}y(\xi)d\xi\ldots d\xi''\cdots$$
$$+\sum_{i=o}^{p-1}a_i\sum_{j=0}^{i}(-1)^j\binom{i}{j}\frac{n!}{(n-j)!}\int^{(p-i+j)}(\xi-a)^{n-j}y(\xi)d\xi\ldots d\xi''\cdots$$
$$+\sum_{i=p}^{n-1}a_i\left[\sum_{j=o}^{i-p}(-1)^j\binom{p+j-1}{j}\frac{n!}{(n-j)!}\tau^{n-j}y^{(i-j-p)}(a+\tau)\right.$$
$$\left.+\sum_{j=1}^{p}(-1)^{j+i-p}\binom{i}{i-p+j}\frac{n!}{(n-i+p-j)!}\int^{(j)}(\xi-a)^{n-i+p-j}y(\xi)d\xi\ldots d\xi''\cdots\right]$$

Applying the Cauchy formula for repeated integrals, and setting $\tau+a=t$,

$$\sum_{i=0}^{n-p}(-1)^i\binom{p+i-1}{i}\frac{n!}{(n-i)!}(t-a)^{n-i}y^{(n-i-p)}(t)$$
$$+\sum_{j=1}^{p}(-1)^{j+n-p}\binom{n}{n-p+j}\frac{n!}{(p-j)!(j-1)!}\int_a^t(t-\xi)^{j-1}(\xi-a)^{p-j}y(\xi)d\xi$$
$$+\sum_{i=0}^{p-1}a_i\sum_{j=0}^{i}(-1)^j\binom{i}{j}\frac{n!}{(n-j)!(p-i+j-1)!}\int_a^t(t-\xi)^{p-i+j-1}(\xi-a)^{n-j}y(\xi)d\xi$$
$$+\sum_{i=p}^{n-1}a_i\left[\sum_{j=0}^{i-p}(-1)^j\binom{p+j-1}{j}\frac{n!}{(n-j)!}(t-a)^{n-j}y^{(i-j-p)}(t)\right.$$
$$\left.+\sum_{j=1}^{p}(-1)^{j+i-p}\binom{i}{i-p+j}\frac{n!}{(n-i+p-j)!(j-1)!}\int_a^t(t-\xi)^{j-1}(\xi-a)^{n-i+p-j}y(\xi)d\xi\right]=0$$

Swapping $\int$ and $\sum$, and applying some algebraic manipulation, yields

Forward kernel for $y^{(k)}$

$$y^{(k)}(t) = \frac{1}{(t-a)^n}\left[\sum_{i=1}^{k}(-1)^{i+1}\binom{p+i-1}{i}\frac{n!}{(n-i)!}(t-a)^{n-i}y^{(k-i)}(t)\right.$$
$$\left.+\sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}(-1)^{j+1}\binom{p+j-1}{j}\frac{n!}{(n-j)!}(t-a)^{n-j}y^{(i-j-p)}(t)+\int_a^t K_{F,k,y}(n,p,t,\tau)y(\tau)d\tau\right] \tag{89}$$

where

$$K_{F,k,y}(n,p,t,\tau) = \sum_{j=1}^{p}(-1)^{j+n-p+1}\binom{n}{n-p+j}\frac{n!}{(p-j)!(j-1)!}(t-\tau)^{j-1}(\tau-a)^{p-j}$$
$$+\sum_{i=0}^{p-1}a_i\sum_{j=0}^{i}(-1)^{j+1}\binom{i}{j}\frac{n!}{(n-j)!(p-i+j-1)!}(t-\tau)^{p-i+j-1}(\tau-a)^{n-j}$$
$$+\sum_{i=p}^{n-1}a_i\sum_{j=1}^{p}(-1)^{j+i-p+1}\binom{i}{i-p+j}\frac{n!}{(n-i+p-j)!(j-1)!}(t-\tau)^{j-1}(\tau-a)^{n-i+p-j}$$

The backward kernel is obtained similarly

Backward kernel for $y^{(k)}$

$$y^{(k)}(t) = \frac{1}{(b-t)^n}\left[-\sum_{i=1}^{k}\binom{p+i-1}{i}\frac{n!}{(n-i)!}(b-t)^{n-i}y^{(k-i)}(t)\right.$$
$$\left.-\sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}\binom{p+j-1}{j}\frac{n!}{(n-j)!}(b-t)^{n-j}y^{(i-j-p)}(t)+\int_t^b K_{B,k,y}(n,p,t,\tau)y(\tau)d\tau\right] \tag{90}$$

where

$$K_{B,k,y}(n,p,t,\tau) = \sum_{j=1}^{p}\binom{n}{n-p+j}\frac{n!}{(p-j)!(j-1)!}(t-\tau)^{j-1}(b-\tau)^{p-j}$$
$$+\sum_{i=0}^{p-1}a_i\sum_{j=0}^{i}\binom{i}{j}\frac{n!}{(n-j)!(p-i+j-1)!}(t-\tau)^{p-i+j-1}(b-\tau)^{n-j}$$
$$+\sum_{i=p}^{n-1}a_i\sum_{j=1}^{p}\binom{i}{i-p+j}\frac{n!}{(n-i+p-j)!(j-1)!}(t-\tau)^{j-1}(b-\tau)^{n-i+p-j}$$

Adding equations (89) and (90) produces

$$y^{(k)}(t) = \frac{1}{(t-a)^n + (b-t)^n}\left[\sum_{i=1}^{k}(-1)^{i+1}\binom{p+i-1}{i}\frac{n!}{(n-i)!}(t-a)^{n-i}y^{(k-i)}(t)\right.$$
$$+\sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}(-1)^{j+1}\binom{p+j-1}{j}\frac{n!}{(n-j)!}(t-a)^{n-j}y^{(i-j-p)}(t) + \int_a^t K_{F,k,y}(n,p,t,\tau)y(\tau)d\tau$$
$$-\sum_{i=1}^{k}\binom{p+i-1}{i}\frac{n!}{(n-i)!}(b-t)^{n-i}y^{(k-i)}(t)$$
$$\left.-\sum_{i=p}^{n-1}a_i\sum_{j=0}^{i-p}\binom{p+j-1}{j}\frac{n!}{(n-j)!}(b-t)^{n-j}y^{(i-j-p)}(t) + \int_t^b K_{B,k,y}(n,p,t,\tau)y(\tau)d\tau\right]$$

B Reproducing Kernel Hilbert Spaces[60]

B.1 Introduction

Given a set X, if we equip the set of all functions from X to $\mathbb{F}$,$\mathcal{F}(X,\mathbb{F})$ with the usual operations of addition, $(f+g)(x) = f(x)+g(x)$, and scalar multiplication, $(\lambda\hat{A}\Delta f)(x) = \lambda\hat{A}\Delta(f(x))$, then $\mathcal{F}(X,\mathbb{F})$ is a vector space over $\mathbb{F}$ [67].

Definition 1. A Hilbert Space is an inner product space that is complete and separable with respect to the norm defined by the inner product.

Examples of Hilbert spaces include:

1. The vector space $\mathbb{R}^n$ with $\langle a,b\rangle = a'b$, the vector dot product of a and b.
2. The space l_2 of square summable sequences, with inner product $\langle x,y\rangle = \sum_{i=1}^{\infty} x_i y_i$
3. The space L^2 of square integrable functions (i.e., $\int_s f(x)^2 \mathrm{d}x < \infty$), with inner product $\langle f,g\rangle = \int_s f(x)g(x)\mathrm{d}x < \infty$) [8]

Definition 2. Given a set X, we will say that $\mathcal{H}$ is a reproducing kernel Hilbert space(RKHS) on X over $\mathbb{F}$, provided that:

1. $\mathcal{H}$ is a vector subspace of $\mathcal{F}(X,\mathbb{F})$
2. $\mathcal{H}$ is endowed with an inner product, $\langle .,.\rangle$ making it into a Hilbert space,
3. or every $y \in X$, the linear evaluation functional, $E_y : \mathcal{H} \mapsto \mathbb{F}$, defined by $E_y(f) = f(y)$, is bounded [67].

For instance, the L^2 space is a Hilbert space, but not an RKHS because the delta function which has the reproducing property:

$$f(x) = \int_s \delta(x-u)f(u)\mathrm{d}u \tag{91}$$

does not satisfy the square integrable condition, that is,

$$\int_s \delta(u)^2 \mathrm{d}u \not< \infty \tag{92}$$

thus the delta function is not in L^2 [8].

B.2 Reproducing Kernels for Modelling of LTI Systems [60]

The class of systems considered here comprises linear time invariant systems described in terms of their characteristic equations :

$$D^m y(t) = a_{m-1}D^{m-1}y(t) + a_{m-2}D^{m-2}y(t) + ... + a_1 D^1 y(t) + a_0 D^0 y(t), \tag{93}$$

with the usual definition of the differential operators :

$$D^0 y(t) = y(t); \quad D^1 y(t) = \frac{d}{dt}y(t); \quad D^k y(t) = \frac{d^k}{dt^k}y(t); k = 1, ..., m \tag{94}$$

that are satisfied by the system "output functions" y on some interval of time $t \in [a, b] \subset \mathbb{R}$, where $a_i \in \mathbb{R}, i = 1, .., m-1$, as real coefficients.

With a change of notation from K_{DS} to K^{DS}, it will be assumed that a double-sided kernel $K^{DS} : (t, \xi) \mapsto K^{DS}(t, \xi),\ (t, \xi) \in [a, b] \times [a, b]$, can be constructed such that every solution of (93) on an interval $[a, b]$ satisfies

$$y(p) = \int_a^b K^{DS}(p, \xi)y(\xi)d\xi \ ; \quad p \in [a, b] \tag{95}$$

Since the output variable y is assumed to be measured, and hence is likely to be corrupted by noise, it is convenient to regard it as a member of the Hilbert space $L^2[a, b]$. It follows from the construction the double-sided kernel that K^{DS} is also an L^2 function, so $K^{DS} \in L^2[a, b] \times L^2[a, b]$. For brevity it will be convenient to adopt the following shorthand notation:

$$K_p^{DS}(\xi) := K^{DS}(p, \xi); \quad K_p^{DS} \in L^2[a, b]; \text{ i.e. } K^{DS}(p, \cdot) \in L^2[a, b]; \text{ for all } p \in [a, b] \tag{96}$$

allowing to re-write (95) in terms of the scalar product on $L^2[a, b]$:

$$y(p) = \langle y, K_p^{DS} \rangle_2, \quad p \in [a, b] \tag{97}$$

$$\text{i.e. } y = \langle y, K^{DS} \rangle_2 \ ; \quad \text{where } K^{DS} : p \mapsto K_p^{DS} \in L^2[a, b] \ ; \ p \in [a, b] \tag{98}$$

where $\langle \cdot, \cdot \rangle_2$ is the scalar product in $L^2[a, b]$. It should be noted that with these definitions, the functions K_p^{DS}, although considered as members of $L^2[a, b]$ are point wise defined. The reproducing property (97) holds for any functions y satisfying (93) .
Since the reproducing property is linear in the reproduced signal y, the condition (98) already holds on a linear vector subspace of $L^2[a, b]$ comprising all solutions of the characteristic equation on the interval $[a, b]$ (regardless of their initial conditions).

References:

[1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.

[2] R. Alur, T. A. Henzinger, and E. D. Sontag. *Hybrid systems III: verification and control*, volume 3. Springer Science & Business Media, 1996.

[3] R. Alur, T. A. Henzinger, and E. D. Sontag. *Hybrid systems III: verification and control*, volume 3. Springer Science & Business Media, 1996.

[4] K. J. Åström and P. Eykhoff. System identification—a survey. *Automatica*, 7(2): 123–162, 1971.

[5] K. J. Åström and P. Eykhoff. System identification—a survey. *Automatica*, 7(2): 123–162, 1971.

[6] Y. Bar-Shalom. Multitarget-multisensor tracking: Applications and advances. volume iii. *Norwood, MA, Artech House, Inc., 2000*, 2000.

[7] Y. Bar-Shalom. Multitarget-multisensor tracking: Applications and advances. volume iii. norwood, ma, artech house. *Inc*, 2000.

[8] P. Bartlett. Instructor Notes: Reproducing Kernel Hilbert Spaces, 2008. URL `https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/7.pdf`.

[9] S. Bhattacharyya. Parameter invariant observers. *International Journal of Control*, 32(6):1127–1132, 1980.

[10] P. Caines. On the asymptotic normality of instrumental variable and least squares estimators. *IEEE Transactions on Automatic Control*, 21(4):598–600, 1976.

[11] P. E. Caines. *Linear stochastic systems*, volume 77. SIAM, 2018.

[12] A. Chatterjee. *Two-Stage Kernel-Based State and Parameter Estimation of LTI Systems*. PhD thesis, 2020.

[13] F. Courreges. Design of a discrete algebraic robust differentiation fir filter using an annihilator of the z-transform; frequency response analysis and parameter tuning. *Numerical Algorithms*, 68(4):867–901, 2015.

[14] G. David. An introduction to observers. *IEEE Transactions on automatic control*, 16(6):596–602, 1971.

[15] F. Diener and G. Reeb. *Analyse non standard*, volume 40. Editions Hermann, 1989.

[16] S. Diop and M. Fliess. Nonlinear observability, identifiability, and persistent trajectories. In *Decision and Control, 1991., Proceedings of the 30th IEEE Conference on*, pages 714–719. IEEE, 1991.

[17] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin. Sensitivity and specificity of information criteria. *bioRxiv*, page 449751, 2019.

[18] M. Fliess. Analyse non standard du bruit. *Comptes Rendus Mathematique*, 342(10): 797–802, 2006.

[19] M. Fliess and H. Sira-Ramírez. An algebraic framework for linear identification. *ESAIM: Control, Optimisation and Calculus of Variations*, 9:151–168, 2003.

[20] M. Fliess and H. Sira-Ramírez. Control via state estimations of some nonlinear systems. In *IFAC Symposium on Nonlinear Control Systtems (NOLCOS 2004)*, 2004.

[21] M. Fliess and H. Sira-Ramirez. Control via state estimations of some nonlinear systems. In *IFAC Symposium on Nonlinear Control Systtems (NOLCOS 2004)*, 2004.

[22] M. Fliess and H. Sira-Ramírez. Closed-loop parametric identification for continuous-time linear systems via new algebraic techniques. In *Identification of Continuous-time Models from sampled Data*, pages 363–391. Springer, 2008.

[23] M. Fliess, C. Join, and H. Sira-Ramírez. Non-linear estimation is easy. *International Journal of Modelling, Identification and Control*, 4(1):12–27, 2008.

[24] B. Friedlander. The overdetermined recursive instrumental variable method. *IEEE Transactions on Automatic Control*, 29(4):353–356, 1984.

[25] D. Ghoshal and H. Michalska. Forward-backward kernel-based state and parameter estimation forlinear systems of arbitrary order. 2019.

[26] D. Ghoshal and H. Michalska. Finite interval kernel-based identification and state estimation for lti systems with noisy output data. In *American Control Conference, 2019. ACC 2019*, 2019 forthcoming.

[27] D. Ghoshal, K. Gopalakrishnan, and H. Michalska. Using invariance to extract signal from noise - submitted. In *American Control Conference, 2017. ACC 2017*, 2017 forthcoming.

[28] D. Ghoshal, K. Gopalakrishnan, and H. Michalska. Algebraic parameter estimation using kernel representation of linear systems - submitted. In *IFAC World Congress, 2017. 20th World Congress*, 2017 forthcoming.

[29] D. Ghoshal, K. Gopalakrishnan, and H. Michalska. Parameter and state estimation in linear systems under arbitrary noise - submitted. In *IFAC World Congress, 2017. 20th World Congress*, 2017 forthcoming.

[30] D. P. Ghoshal and H. Michalska. Finite-interval kernel-based identification and state estimation for lti systems with noisy output data. In *2019 American Control Conference (ACC)*, pages 4982–4989. IEEE, 2019.

[31] D. P. Ghoshal and H. Michalska. Finite-interval kernel-based identification and state estimation for lti systems with noisy output data* this work was supported by the national science & engineering research council of canada. In *2019 American Control Conference (ACC)*, pages 4982–4989. IEEE, 2019.

[32] D. P. Ghoshal, K. Gopalakrishnan, and H. Michalska. Using invariance to extract signal from noise. In *American Control Conference (ACC), 2017*, pages 2588–2593. IEEE, 2017.

[33] D. P. Ghoshal, K. Gopalakrishnan, and H. Michalska. Kernel-based adaptive multiple model target tracking. In *IEEE Conference on Control Technology and Applications*, 2017.

[34] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[35] G. Goodwin and K. Sin. Adaptive filtering prediction and control. 1984. *Englewood Clifs: Prentice Ha lI.*

[36] K. Gopalakrishnan. *A New Method of Non-asymptotic Estimation for Linear Systems.* PhD thesis, McGill University Libraries, 2017.

[37] K. V. Gopalakrishnan. A new method of non-asymptotic estimation for linear systems. Master's thesis, McGill University, 2016.

[38] C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[39] R. Irizarry. URL `http://rafalab.github.io/pages/754/`.

[40] R. Jagannath and N. S. Upadhye. The lasso estimator: Distributional properties. *arXiv preprint arXiv:1605.03280*, 2016.

[41] A. John. PhD thesis, 2019.

[42] T. Kailath. Rkhs approach to detection and estimation problems–i: Deterministic signals in gaussian noise. *IEEE Transactions on Information Theory*, 17(5):530–549, 1971.

[43] T. Kailath. Lectures on wiener and kalman filtering. In *Lectures on Wiener and Kalman Filtering*, pages 1–143. Springer, 1981.

[44] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear estimation*, volume 1. Prentice Hall Upper Saddle River, NJ, 2000.

[45] R. Kalman. On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110, 1959.

[46] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[47] R. E. Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

[48] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Journal of basic engineering*, 83(1):95–108, 1961.

[49] E. Kalnay. *Atmospheric modeling, data assimilation and predictability.* Cambridge university press, 2003.

[50] S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[51] H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

[52] L. Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

[53] L. Ljung. System identification: Theory for the user, ptr prentice hall information and system sciences series. *ed: Prentice Hall, New Jersey*, 1999.

[54] D. Luenberger. Observers for multivariable systems. *IEEE Transactions on Automatic Control*, 11(2):190–197, 1966.

[55] D. G. Luenberger. Observing the state of a linear system. *IEEE transactions on military electronics*, 8(2):74–80, 1964.

[56] P. Martin and E. Salaun. Invariant observers for attitude and heading estimation from low-cost inertial and magnetic sensors. In *Decision and Control, 2007 46th IEEE Conference on*, pages 1039–1045. IEEE, 2007.

[57] D. Mayne. A method for estimating discrete time transfer functions. In *Advances in Computer Control, Second UKAC Control Convention.* The University of Bristol, 1967.

[58] M. Mboup, C. Join, and M. Fliess. A revised look at numerical differentiation with an application to nonlinear feedback control. In *The 15th Mediterrean Conference on Control and Automation-MED'2007*, 2007.

[59] M. Mboup, C. Join, and M. Fliess. Numerical differentiation with annihilators in noisy environment. *Numerical algorithms*, 50(4):439–467, 2009.

[60] H. Michalska. Instructor Notes: Reproducing Kernels for Modelling of Systems, 2016.

[61] H. Michalska. Instructor Notes: Parameter Estimation using Lsplines in RKHS, 2019.

[62] H. Michalska. Differential invariance representations in reproducing kernel hilbert spaces - towards a design of a 'differential french curve'. Technical report, Department of Electrical, Computer, and Software Engineering, McGill University, Canada., 2019.

[63] K. Ogata. *Modern Control Engineering,Fifth Edition*. Prentice Hall, Upper Saddle River, 1997.

[64] A. Pandey. Deterministic variational approach to system modelling by data assimilation. Master's thesis, McGill University, 2018.

[65] E. Parzen. Time series analysis for models of signal plus white noise. Technical report, Department of Statistics, Stanford University., 1966.

[66] E. Parzen. Statistical inference on time series by RKHS methods. Technical report, Department of Statistics, Stanford University., 1970.

[67] V. I. Paulsen and M. Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press, 2016. URL `http://dx.doi.org/10.1017/CBO9781316219232`.

[68] R. L. Plackett. Some theorems in least squares. *Biometrika*, 37(1/2):149–157, 1950.

[69] A. Rakotomamonjy and S. Canu. Frames, reproducing kernels, regularization and learning. *Journal of Machine Learning Research*, 6(Sep):1485–1515, 2005.

[70] S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

[71] S. Saitoh. *Theory of reproducing kernels and its applications*, volume 189. Longman, 1988.

[72] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6 (2):461–464, 1978.

[73] H. Sira-Ramírez and M. Fliess. An algebraic state estimation approach for the recovery of chaotically encrypted messages. *International Journal of Bifurcation and Chaos*, 16(02):295–309, 2006.

[74] H. Sira-Ramírez, C. G. Rodríguez, J. C. Romero, and A. L. Juárez. *Algebraic identification and estimation methods in feedback control systems*. John Wiley & Sons, 2014.

[75] D. Sridhar, D. P. Ghoshal, and H. Michalska. B-splines in joint parameter and state estimation in linear time-varying systems. In *2018 Annual American Control Conference (ACC)*, pages 3508–3513. IEEE, 2018.

[76] S. Stanhope, J. Rubin, and D. Swigon. Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory. *SIAM Journal on Applied Dynamical Systems*, 13(4):1792–1815, 2014.

[77] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[78] A. N. Tikhonov and V. Y. Arsenin. Solutions of ill-posed problems. *New York*, pages 1–30, 1977.

[79] S. K. Varanasi and P. Jampana. Parameter estimation and model order identification of lti systems. *IFAC-PapersOnLine*, 49(7):1002–1007, 2016.

[80] J. L. Waller and T. C. Smith. Sas® global forum. *Model Assisted Statistics and Applications*, 13(4):309–310, 2018. doi: 10.3233/mas-180440.

[81] G. Welch, G. Bishop, et al. An introduction to the kalman filter. 1995.

[82] N. Wiener. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT Press, 1950.

[83] K. Wong and E. Polak. Identification of linear discrete time systems using the instrumental variable method. *IEEE Transactions on Automatic Control*, 12(6):707–718, 1967.

[84] X. Xun, J. Cao, B. Mallick, A. Maity, and R. J. Carroll. Parameter estimation of partial differential equation models. *Journal of the American Statistical Association*, 108(503):1009–1020, 2013.

[85] P. C. Young. The use of linear regression and related procedures for the identification of dynamic processes. In *Seventh Symposium on Adaptive Processes*, pages 53–53. IEEE, 1968.

[86] P. C. Young. *Recursive estimation and time-series analysis: An introduction for the student and practitioner*. Springer Science & Business Media, 2011.