Sparse Bayesian Identification of Nonlinear State-Space Systems via Stein Approximation Approach

Limin Zhang (✉ limin_zhang@yeah.net)

Hengshui University

Junpeng Li

Yanshan University School of Electrical Engineering

Wenting Zhang

Hengshui University

Research Article

Keywords: Sparse Bayesian identification, convex optimisation, Stein approximation method, ML method

Posted Date: February 8th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1313035/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Sparse Bayesian identification of nonlinear State-Space Systems via Stein approximation approach

Limin Zhang^a, Junpeng Li^b, Wenting Zhang^a

^a*Department of Mathematics and Computer Science, Hengshui University, Hengshui City, 053000, China*

^b*Institute of Electrical Engineering, Yanshan University, Qinhuangdao City, 066004, China*

Abstract

This paper concerns with the parameter estimation of nonlinear discrete-time systems from noisy state measurements in state-space form. A sparse Bayesian convex optimisation algorithm was proposed for the parameter estimation and prediction. The proposed method takes full account of the data correlation, parameter priori, constraints, which is explicitly modeled by Bayesian sparse learning and optimisation. The only assumption about the structure of the model is that there are only a few important terms that govern the dynamics, so that the equations are sparse in the space of possible functions and the number of basis function is generated randomly. The main identification challenge resides in two aspects: first, a new objective function can be obtained by Stein approximation method in the convex optimization problem. Second, a recursive least squares estimation with L1-regularization is developed, wherein the regularization parameter is selected from the point of optimization. Compared with the function maximum likelihood(ML) method only, it usually captures more information about the dependence of the data indicators. Three simulation examples are given to demonstrate the proposed algorithm's effectiveness. Furthermore, the performances of these approaches are analyzed, including parameter estimation of RMSE, parameter sparsity and prediction of state and output result.

Keywords: Sparse Bayesian identification, convex optimisation, Stein approximation method, ML method

1. Introduction

Estimating nonlinear systems is widely acknowledged for its importance and difficulties [1, 2]. As a result, there is a significant and active study effort focused on the issue. One of the most important aspects of this activity is that it concentrates on certain system classes, such as neural networks [3], nonlinear ARMAX (NARMAX) [4], and Hammerstein-Wiener [5] models and so on. Sums of terms from a set of parameterized functions can be used to expand nonlinear functional forms (see Sec.2 and [6]), a common method for identifying nonlinear state-space models is to look for a concise description that is consistent with the current data among a collection of probable nonlinear terms [7]. Classic functional decomposition methods such as Volterra expansion, Taylor polynomial expansion, or Fourier series [6, 8] provide a few

Email address: limin_zhang@yeah.net (Limin Zhang)

options for basis functions. These methods are founded on the idea that there are a finite set of fundamental functions whose linear combination can be utilized to describe the dynamics of a system.

Maximum likelihood(ML) approach is a common method for nonlinear systems estimating, which is also used in linear dynamic systems. To compute estimates, it is common to use a gradient-based search strategy like the damped Gauss-Newton method [9]. Employing the Expectation Maximisation method [10] for the production of ML estimates is an approach recently investigated in [11] in the context of bilinear systems. Various approximation EM-based techniques have also been studied in the past. In [12], the authors have examined applying mediocre results to the underlying nonlinear smoothing issue, often Kalman filter method. A GP state space model can also be used to describe the nonlinear model. In [13, 14], both a Bayesian and a maximum likelihood estimation strategy are employed in addition to a competitive GP approximation method. In [15], for learning, a Monte Carlo technique is utilized, which includes variational method using the same GP approximation.

However, too many unknown parameters increase the complexity of the model. There are just a few papers that deal with the sparse representation identification issue and multiple constraints on the system's parameters [17, 25]. Different from the previous work, our suggested framework, on the other hand, allows us to include more constraints on the corresponding model parameters, e.g., inequality constraint, priori information and stein discrepancy constraints imposed among parameters. The dynamical system discovery issue is reimaged in this paper from the perspectives of sparse regression [19, 20, 29], compressed sensing [21–23], stein approximation theory [24], convex optimization method [25]. The use of sparsity approaches in dynamical systems is relatively new [26–28]. We take use of the fact that most nonlinear systems have only a few important dynamics elements, resulting in sparse parameter set in a high-dimensional nonlinear function space.

Noise is an important factor affecting parameter estimation. When the input data of state measurements or their temporal derivatives is distorted by noise, the identification of reliable models is a key difficulty in model recovery via sparse identification of nonlinear dynamics (SINDy). Noisy measurements might lead to erroneous basis terms being identified and poor parameter estimates in the model. The essential insight is that the function $\mathbf{f}$ has only a few terms for many of the systems of interest, making it sparse in the range of potential functions. For model recovery, Least absolute shrinkage and selection operator (LASSO) [29], least angle regression (LARS) [30], sequentially thresholded least squares (STLS) [31], and basis pursuit denoising (BPDN) [32] are parameter sparsity algorithms that can be applied. Recent advancements in sparse methods, such as L0 and L1 norms, make this sparsity viewpoint attractive, because it is now feasible to detect which parameters are nonzero without undertaking a combinatorially arduous and time-consuming search. The sparse model identification that results strikes a natural balance between model sparsity and precision, preventing the model from being overfit to the data. In nonlinear dynamics system identification, to reduce noise and anticipate disasters, we use sparse regression.

We will discuss how a Bayesian technology, optimization method and stein approximation strategy might mitigate the difficulties of large correlations in the state matrix in this technical note, from a statistical perspective. The following are some of the technical note's most important contributions:

- (1) A sparse Bayesian nonconvex optimisation algorithm was proposed for the parameter

estimation and prediction. The proposed method takes full account of the data correlation, parameter priori, constraints, which is explicitly modeled by Bayesian sparse learning and optimisation. Moreover, a new objective function can be obtained by stein method. Compared with the function ML only, it usually captures more information about the dependence of the data indicators.

(2) Construct a nonconvex optimisation problem using a sparse Bayesian version of the nonlinear system identification issue with additive noise.

(3) Compared with other related methods, we consider the stein discrepancy of parameters as an objective function, which is minimized in the nonconvex optimization problem. It can captures more important information about the dependence of the data.

The rest of this paper is organized as follows. Section 2 describes problem statement and background. In section 3, constructing the model in Bayesian framework. In Section 4, non-convex optimisation with Stein constrain for identification is introduced. Three numerical illustration including Narendra-Li Model, NARX model, Kernel state-space models(KSSM) are presented in Section 5. Finally, in Section 6, we give some closing remarks.

2. Problem statement and background

This research looks at the difficulty of determining the parameters θ for certain categories of the nonlinear state-space model. Discrete-time dynamics is a nonlinear state-space model that is defined as follows:

$$\begin{aligned}\pi(\mathbf{x}_{ik}) &= \mathbf{f}_i(\mathbf{x}_k, \mathbf{u}_k) + \nu_{ik} \\ &= \sum_{r=1}^N \theta_{ir} f_{ir}(\mathbf{x}_k, \mathbf{u}_k) + \nu_{ik} \\ o_{ik} &= \mathbf{g}_i(\mathbf{x}_k, \mathbf{u}_k) + e_{ik}\end{aligned}\tag{1}$$

where $\mathbf{x}_k = [x_{1k}, \dots, x_{ik}, \dots, x_{n_x, k}]$ is the state variable in time k , $\mathbf{u}_k$ is the external control input. When system is time continuous, $\pi(x_{ik}) = \dot{x}_{ik}$; When system is time discrete, $\pi(x_{ik}) = x_{ik}$ or $x_{ik} - x_{i, k-1}$; $\nu_{ik} \sim \mathcal{N}(0, \sigma_{ik}^2)$ is noise($i \neq k, \sigma_{ik} = 0$), which is set to be i.i.d. Gaussian distribution. $f_{ir}(\mathbf{x}_k, \mathbf{u}_k) : \mathbb{R}^{n_x + n_u} \rightarrow \mathbb{R}$ is Lipschitz continuous and n_x and n_u are the dimensions of $\mathbf{x}_k$ and $\mathbf{u}_k$, respectively, while θ_{ir} are the weight of basis functions. Together, θ_{ir} and f_{ir} determine the dynamics. It's worth emphasizing that we make no assumptions regarding the nonlinear functions on the right-hand side of the equations' form (1). The output is o_{ik} .

If the system of interest can provide M data samples that meet (1), the system in (1) can be represented as

$$\mathbf{Y}_i = \mathbf{F}_i \theta_i + \nu_i, i = 1, \dots, n_x, \tag{2}$$

where $\mathbf{Y}_i = [\pi(x_{i1}), \dots, \pi(x_{iM})]^\top \in \mathbb{R}^{M \times 1}$, $\theta_{iT} \triangleq [\theta_{i1}, \dots, \theta_{iN_i}]^\top \in \mathbb{R}^{N_i \times 1}$, $\nu_i \triangleq [\nu_{i0}, \dots, \nu_{i, M-1}]^\top \in \mathbb{R}^{M \times 1}$, and $\mathbf{F}_i \in \mathbb{R}^{M \times N_i}$ is called dictionary matrix. Its j -th column is written as

$$[\mathbf{f}_{ij}(\mathbf{x}_0, \mathbf{u}_0), \dots, \mathbf{f}_{ij}(\mathbf{x}_{M-1}, \mathbf{u}_{M-1})]^\top.$$

The identification task in this framework is to find θ_i given the measured data in $\mathbf{Y}_i$. This leads to the solution of a linear regression issue, in which the least square(LS) method can be used if some of model's nonlinear part is understood, i.e., $\mathbf{F}_i$ is known. In (1), we merely discuss the identification problem, as we did in $\mathbf{g}_i(\mathbf{x}_k, \mathbf{u}_k)$. Only a few frequently-used nonlinear dynamical model must be considered, depending on the field. Because of the potential insertion of non-relevant or non-independent dictionary functions in F_i , the solution θ_i to (2) is often sparse.

The subscript i used to index the state variable is removed for convenience of writing because the linear regression issues in (2) are independent.

$$\mathbf{Y} = \mathbf{F}\theta + \nu \quad (3)$$

3. Constructing the model in Bayesian framework

3.1. Priors selection for Parameter Sparsity

All unknowns in Bayesian modeling are evaluated as they were random variables with specific distributions[33]. For $\mathbf{Y} = \mathbf{F}\theta + \nu$ in (3), noisy variables ν is Gaussian independently identically distribution(i.i.d.), i.e., $\nu \sim \mathcal{N}(\mathbf{0}, \beta\mathbf{I})$, $\beta = \sigma^2$. We can obtain the likelihood of the data $P(\mathbf{Y} | \theta) = \mathcal{N}(\mathbf{Y} | \mathbf{F}\theta, \beta\mathbf{I}) \propto \exp\{-\frac{1}{2\beta}\|\mathbf{Y} - \mathbf{F}\theta\|_2^2\}$. $P(\theta)$ is a prior distribution that we define as $P(\theta) \propto \exp\{-\frac{1}{2\alpha}\theta^T\theta\} = \prod_j \exp\{-\frac{1}{2\alpha}\theta_j^T\theta_j\} = \prod_j P(\theta_j)$. To impose a sparseness on θ , the function $-\frac{1}{2\alpha}\theta^T\theta$ is selected as a concave, non-decreasing function of $|\theta_j|$, which includes Gaussian priors and t-distribution priors (see [34] for details).

Parameter θ is identified based on Y , which is not fixed. We substitute the expected value of θ for its true value. The posterior $P(\theta | Y)$ is heavily linked and non-Gaussian, so computing the posterior mean $E(\theta | Y)$ is often difficult. To solve this problem, take $P(\theta | Y)$ as an approximation of Gaussian distribution. In [33], effective posterior computation algorithms is used in the computation.

Another method is to use super Gaussian priors, in which the priors $P(\theta_j)$ is computed with the variational EM algorithms [34]. We define hyperparameters $\alpha \triangleq [\alpha_1, \dots, \alpha_N]^T \in \mathbb{R}_+^N$. The priors of θ can be written as: $P(\theta) = \prod_{j=1}^n P(\theta_j)$, $P(\theta_j) = \mathcal{N}(\theta_j | 0, \alpha_j) \varphi(\alpha_j)$, where $\varphi(\alpha_j) \geq 0$. $P(\theta, \alpha) = \prod_j \mathcal{N}(\theta_j | 0, \alpha_j) \varphi(\alpha_j) = P(\theta | \alpha)P(\alpha) \leq P(\theta)$, where $P(\theta | \alpha) \triangleq \prod_j \mathcal{N}(\theta_j | 0, \alpha_j)$, $P(\alpha) \triangleq \prod_j \varphi(\alpha_j)$. Considering the data $\mathbf{Y}$, the posterior probability of θ can be represented as $P(\theta | \mathbf{Y}, \alpha) = \frac{P(\mathbf{Y}|\theta)P(\theta;\alpha)}{\int P(\mathbf{Y}|\theta)P(\theta;\alpha)d\theta} = \mathcal{N}(\mathbf{m}_\theta, \Sigma_\theta)$. From [35], the posterior mean $\mathbf{m}_\theta$ and covariance Σ_θ are given by:

$$\begin{aligned} \mathbf{m}_\theta &= \Lambda \mathbf{F}^T (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^T)^{-1} \mathbf{Y} \\ \Sigma_\theta &= \Lambda - \Lambda \mathbf{F}^T (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^T)^{-1} \mathbf{F} \end{aligned} \quad (4)$$

where Λ is a diagonal matrix written as $\text{diag}[\alpha]$. It is obvious that to maximize $P(\theta, \alpha | \mathbf{Y})$, the most important question is how to select the best $\hat{\alpha}$. $P(\mathbf{Y} | \theta)$ and $P(\theta; \alpha)$ can be taken as prior information, so we consider $\int P(\mathbf{Y} | \theta)P(\theta; \alpha)d\theta$ only. Using type-II maximum likelihood [35], the marginal likelihood $\int P(\mathbf{Y} | \theta)P(\theta; \alpha)d\theta$ can be maximised, the selected $\hat{\alpha}$ is written as

$$\hat{\alpha} = \underset{\alpha \geq 0}{\operatorname{argmax}} \int P(\mathbf{Y} \mid \theta) \prod_{j=1}^n \mathcal{N}(\theta_j \mid 0, \alpha_j) \varphi(\alpha_j) d\theta \quad (5)$$

After $\hat{\alpha}$ is computed in (5), the estimation of θ can be obtained as $\hat{\theta} = \mathbb{E}(\theta \mid \mathbf{Y}; \hat{\alpha}) = \hat{\Lambda} \mathbf{F}^\top (\lambda \mathbf{I} + \mathbf{F} \hat{\Lambda} \mathbf{F}^\top)^{-1} \mathbf{Y}$, with $\hat{\Lambda} \triangleq \operatorname{diag}[\hat{\alpha}]$. It indicates that by picking the most likely hyper-parameters $\hat{\alpha}$, which is capable of explaining the data $\mathbf{Y}$.

4. Nonconvex optimisation with Stein method for identification

4.1. Stein operators selection and Stein constrain design

The approach can be sketched as follows for a target distribution P with support $\mathcal{I}$. Find a suitable operator $\mathcal{B} := \mathcal{B}_P$ (referred to as the Stein operator) and a large class of functions $\mathcal{F}_\mathcal{B} = \mathcal{F}(\mathcal{B}_P)$ (referred to as the Stein class) such that Z has distribution P , denoted $Z \sim P$, if and only if we have $f \in \mathcal{F}_\mathcal{B}$ for all functions

$$\mathbb{E}[\mathcal{B}f(Z)] = 0$$

A stein operator can be designed in a variety of ways [24]. In our framework, stein's identity and kernelized stein difference are crucial. $P(\theta)$ is probability density function, which is continuous and differentiable on $\theta \subseteq R^d$. vector functions $\phi(\theta) = [\phi_1(\theta), \dots, \phi_d(\theta)]^\top$, $Q(\theta) = [Q_1(\theta), \dots, Q_d(\theta)]^\top$ are smooth. According to Stein's theory, a suitable $\phi(\theta)$ and $Q(\theta)$ can be obtained

$$E_{\theta \sim P} [\mathcal{B}_P \phi(\theta) Q(\theta)] = 0 \quad (6)$$

where

$$\begin{aligned} \mathcal{B}_P \phi(\theta) Q(\theta) &= \phi(\theta) Q(\theta) \nabla_\theta \log P(\theta)^\top \\ &+ \nabla_\theta \phi(\theta) Q(\theta) + \nabla_\theta Q(\theta) \phi(\theta) \end{aligned}$$

The stein operator, abbreviated $\mathcal{B}_P$, operates on the function $\phi(\theta)Q(\theta)$ and produces a zero mean function $\mathcal{B}_P \phi(\theta)Q(\theta)$ under $\theta \sim P$. $Q(\theta)$ is a different smooth density, which is a given function related with $P(\theta)$.

Assuming mild zero boundary conditions on $\phi(\theta)Q(\theta)$. $\forall \theta_i$, when θ is compact, $\theta \subseteq R^d$, $P(\theta)\phi(\theta) \approx 0$, then $\lim_{\|\theta\| \rightarrow \infty} \phi(\theta)P(\theta)Q(\theta) = 0$. The expectation of $\mathcal{B}_P \phi(\theta)Q(\theta)$ under $\theta \sim Q$ are no longer equal zero. The magnitude of $\mathbb{E}_{\theta \sim Q} [\mathcal{B}_P \phi(\theta)Q(\theta)]$ is related with the difference between P and Q . The probability distances for $P(\theta)$ and $Q(\theta)$ in some proper function set $\mathcal{F}_\mathcal{B}$ are defined as the expectations above.

$$S(Q, P) = \max_{\phi \in \mathcal{F}_\mathcal{B}} \{ [\mathbb{E}_{\theta \sim Q} \operatorname{trace} (\mathcal{B}_P \phi(\theta) Q(\theta))]^2 \} \quad (7)$$

The discriminative power and computational tractability of the stein discrepancy are determined by the function set $\mathcal{F}_\mathcal{B}$ chosen. Traditional definitions of $\mathcal{F}_\mathcal{B}$ include sets of functions with bounded Lipschitz norms, which creates a difficult functional optimization problem that is either computationally intractable or necessitates special considerations.

By choosing proper $Q(\theta)$ and optimizing ϕ in the unit sphere of a reproducing kernel Hilbert space (RKHS), kernelized stein discrepancy(KSD) avoids this challenge. The term KSD is described as

$$\begin{aligned} S(Q, P) &= \max_{\phi \in \mathcal{H}^d} [E_{x \sim Q}(\text{trace}(\mathcal{B}_P \phi(\theta) Q(\theta)))]^2, \\ \text{s.t.} \quad &|\phi(\theta) Q(\theta)|_{\mathcal{H}^d} \leq 1 \end{aligned} \quad (8)$$

The optimal solution of (8), which is $\psi(\theta) = [\phi_{Q,P}^*(\theta) Q(\theta)] / |\phi_{Q,P}^* Q(\theta)|_{\mathcal{H}^d}$ where

$$\psi_{Q,P}^*(\cdot) = \mathbb{E}_{\theta \sim Q} [\mathcal{B}_P \mathbf{k}(\theta, \cdot) Q(\theta)] \quad (9)$$

which we are responsible for

$$S(Q, P) = |\psi_{Q,P}^{*2}|_{\mathcal{H}^d} \quad (10)$$

For θ for any fixed θ' , kernel fuction $\mathbf{k}(\theta, \theta')$ belongs to set $\mathcal{F}_{\mathcal{B}}$. $S(Q, P) = 0$, that is to say $\psi_{Q,P}^*(\theta) \equiv 0$ only if $P(\theta) = Q(\theta)$, based on the information presented above. RBF kernel $\mathbf{k}(\theta, \theta') = \exp(-\frac{1}{h} \|\theta - \theta'\|_2^2)$ is purely positive definite in a strict sense. Because of its decaying property, the RBF kernel belongs to the stein class of accepted smooth densities in $\theta \subseteq R^d$. Both the stein operator and the KSD are only dependent on $P(\theta)$ via the score function $Q(\theta) \nabla_{\theta} \log P(\theta)$, the coefficient of which does not affect the overall structure of the model, so not be considered temporarily. As a result of this property, stein's identity can be used to handle unnormalized distributions effectively.

Define $T_j = \psi_{Q,P}^*(\cdot)$, $Q(\theta_j)$ is selected as $P(\theta_j | \tau_j)$

$$P(\theta_j | \tau_j) \propto \exp \left[-\frac{1}{2\tau_j} \theta_j^T \theta_j \right]$$

Let $K(\theta_j, \cdot) = \mathbf{k}(\theta_j, \cdot) Q(\theta_j)$. From section 2, prior distribution of θ_j can be written as

$$P(\theta_j | \alpha_j) \propto \exp \left[-\frac{1}{2\alpha_j} \theta_j^T \theta_j \right]$$

we have

$$K(\theta_j, \cdot) \propto \exp \left[-\frac{1}{h} (\theta_j - \theta_0)^T (\theta_j - \theta_0) \right]$$

$$\nabla_{\theta} \log P(\theta_j | \alpha_j) \propto -\frac{1}{\alpha_j} \theta_j$$

$$\nabla_{\theta} K(\theta_j, \cdot) \propto -\frac{2}{h} (\theta_j - \theta_0) \exp \left[-\frac{1}{h} (\theta_j - \theta_0)^T (\theta_j - \theta_0) \right]$$

We can then re-express T_j as

$$\begin{aligned}
T_j &= K(\theta_j, \cdot) \cdot \nabla_{\theta_j} \log P(\theta_j | \alpha_j) + \nabla_{\theta} K(\theta_j, \cdot) \\
&= -\frac{1}{\alpha_j} \theta_j \exp \left[-\frac{1}{h} (\theta_j - \theta_0)^T (\theta_j - \theta_0) \right] \\
&\quad - \frac{2}{h} (\theta_j - \theta_0) \exp \left[-\frac{1}{h} (\theta_j - \theta_0)^T (\theta_j - \theta_0) \right] \\
&\leq -\frac{1}{\alpha_j} \theta_j \exp \left[-\frac{1}{h} \theta_j^T \theta_j \right] - \frac{2}{h} \theta_j \exp \left[-\frac{1}{h} (\theta_j - \theta_0)^T (\theta_j - \theta_0) \right] \\
&= -\left(\frac{1}{\alpha_j} + \frac{2}{h} \right) \theta_j \exp \left[-\frac{1}{h} \theta_j^T \theta_j \right]
\end{aligned}$$

The expectation of T_j is

$$\begin{aligned}
\mathbb{E}[T_j] &= \int -\left(\frac{1}{\alpha_j} + \frac{2}{h} \right) \theta_j \exp \left[-\frac{1}{h} \theta_j^T \theta_j \right] \exp \left[-\frac{1}{2\tau_j} \theta_j^T \theta_j \right] d\theta_j \\
&= \int -\left(\frac{1}{\alpha_j} + \frac{2}{h} \right) \theta_j \exp \left[-\left(\frac{1}{h} + \frac{1}{2\tau_j} \right) \theta_j^T \theta_j \right] d\theta_j \\
&= \frac{\frac{1}{\alpha_j} + \frac{2}{h}}{\frac{1}{\tau_j} + \frac{2}{h}} \exp \left[-\left(\frac{1}{h} + \frac{1}{2\tau_j} \right) \theta_j^T \theta_j \right]
\end{aligned}$$

let $\xi(\alpha_j) = \frac{\frac{1}{\alpha_j} + \frac{2}{h}}{\frac{1}{\tau_j} + \frac{2}{h}}$, then

$$\mathbb{E}[T_j] = \xi(\alpha_j) \exp \left[-\left(\frac{1}{h} + \frac{1}{2\tau_j} \right) \theta_j^T \theta_j \right] \quad (11)$$

$$\mathbb{E}[T] = \prod \mathbb{E}[T_j]$$

Another technique to pick $\hat{\alpha}$ is to make it the minimizer of $\mathbb{E}[T]$.

$$\hat{\alpha} = \underset{\alpha \geq 0}{\operatorname{argmin}} \mathbb{E}[T] \quad (12)$$

4.2. Parameter Sparse Identification of Constraints from Data

In (12), $\mathbb{E}[T]$ is another objective function, which make the model parameters less dependent on noisy data and to prevent overfitting. The problem of system identification with convex constraints is given a sparse Bayesian formulation, which is then handled as a nonconvex optimisation problem in this section. To obtain a better parameter α , the new objective function is constructed as

$$\hat{\alpha} = \underset{\alpha \geq 0}{\operatorname{argmax}} (1/\mathbb{E}[T]) \int P(\mathbf{Y} | \theta) \prod_{j=1}^n \mathcal{N}(\theta_j | 0, \alpha_j) \varphi(\alpha_j) d\theta \quad (13)$$

4.2.1. Objective Function in parameter identification

Theorem 1: Use the notation $\mathcal{J}_\alpha(\alpha)$ as the objective function.

$$\begin{aligned}\mathcal{J}_\alpha(\alpha) &= \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| \\ &+ \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} + \sum_{j=1}^N p(\alpha_j) \\ &- 2 \sum_{j=1}^N \left(\frac{1}{h} + \frac{1}{2\tau_j} \right) \theta_j^T \theta_j - 2 \log \xi(\alpha_j)\end{aligned}\tag{14}$$

By minimising $\mathcal{J}_\alpha(\alpha)$, the optimal hyperparameters $\hat{\alpha}$ in (13) can be derived, where $p(\alpha_j) = -2 \log \varphi(\alpha_j)$. The mean of θ is calculated and represented as $\hat{\theta} = \hat{\Lambda} \mathbf{F}^\top (\lambda \mathbf{I} + \mathbf{F} \hat{\Lambda} \mathbf{F}^\top)^{-1} \mathbf{Y}$.

Proof 1:

Using the Woodbury inversion identity, re-express m_θ and Σ_θ in (4):

$$m_\theta = \Lambda \mathbf{F}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} = \frac{1}{\lambda} \Sigma_\theta^\top \mathbf{F} \mathbf{Y} \tag{15}$$

$$\begin{aligned}\Sigma_\theta &= \Lambda - \Lambda \mathbf{F}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{F} \Lambda \\ &= \left(\Lambda^{-1} + \frac{1}{\lambda} \mathbf{F}^\top \mathbf{F} \right)^{-1}\end{aligned}\tag{16}$$

We can express the integral in (13) as follow because the data likelihood $P(\mathbf{Y} \mid \theta)$ is Gaussian.

$$\begin{aligned}& (1/\mathbb{E}[T]) \int \mathcal{N}(\mathbf{Y} \mid \mathbf{F}\theta, \lambda \mathbf{I}) \prod_{j=1}^N \mathcal{N}(\theta_j \mid 0, \alpha_j) \varphi(\alpha_j) d\theta \\ &= (1/\mathbb{E}[T]) \left(\frac{1}{2\pi\lambda} \right)^{M/2} \left(\frac{1}{2\pi} \right)^{N/2} \int \exp\{-E(\theta)\} d\theta \prod_{j=1}^N \frac{\varphi(\alpha_j)}{\sqrt{\alpha_j}},\end{aligned}\tag{17}$$

where

$$\begin{aligned}E(\theta) &= \frac{1}{2\lambda} \|\mathbf{Y} - \mathbf{F}\theta\|^2 + \frac{1}{2} \theta^\top \Lambda^{-1} \theta \\ &= \frac{1}{2} (\theta - \mathbf{m}_\theta)^\top \Sigma_\theta^{-1} (\theta - \mathbf{m}_\theta) + E(\mathbf{Y})\end{aligned}\tag{18}$$

We get $E(\mathbf{Y})$ using the Woodbury inversion identity.

$$\begin{aligned}E(\mathbf{Y}) &= \frac{1}{2} \left(\frac{1}{\lambda} \mathbf{Y}^\top \mathbf{Y} - \frac{1}{\lambda} \mathbf{Y}^\top \mathbf{F} \Sigma_\theta \Sigma_\theta^{-1} \Sigma_\theta \mathbf{F}^\top \mathbf{Y} \frac{1}{\lambda} \right) \\ &= \frac{1}{2} \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y}\end{aligned}\tag{19}$$

Just for the sake of calculation, we can evaluate the integral of $\exp\{-E(\theta)\}$

$$\int \exp\{-E(\theta)\} d\theta = \exp\{-E(\mathbf{Y})\} (2\pi)^{N/2} |\Sigma_\theta|^{1/2}$$

Then applying a $-2\log(\cdot)$ transformation to (17), we have

$$\begin{aligned} & -2\log \int P(\mathbf{Y} | \theta) \prod_{j=1}^n \mathcal{N}(\theta_j | 0, \alpha_j) \varphi(\alpha_j) d\theta - 2\log \mathbb{E}[T] \\ &= -\log |\Sigma_\theta| + M \log 2\pi\lambda + \log |\Lambda| + \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} \\ & \quad + \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N \left(\frac{1}{h} + \frac{1}{2\tau_j}\right) \theta_j^T \theta_j - 2 \sum_{j=1}^N \log \xi(\alpha_j) \\ &= \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| + M \log 2\pi + \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} \\ & \quad + \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N \left(\frac{1}{h} + \frac{1}{2\tau_j}\right) \theta_j^T \theta_j - 2 \sum_{j=1}^N \log \xi(\alpha_j) \end{aligned}$$

where $|\Lambda^{-1} \lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| = |\lambda \mathbf{I}| |\Lambda^{-1} + \frac{1}{\lambda} \mathbf{F}^\top \mathbf{F}|$, $\log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| = -\log |\Sigma_\theta| + M \log \lambda + \log |\Lambda|$.

From (13), we then obtain

$$\begin{aligned} \hat{\alpha} &= \arg \min_{\alpha \geq 0} (\log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| + \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} \\ & \quad + \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N \left(\frac{1}{h} + \frac{1}{2\tau_j}\right) \theta_j^T \theta_j - 2 \log \xi_j(\alpha_j)) \end{aligned} \quad (20)$$

To acquire an approximation of θ , we compute the posterior mean $\theta : \hat{\theta} = \mathbb{E}(\theta | \mathbf{Y}; \hat{\alpha}) = \Lambda \mathbf{F}^\top (\lambda \mathbf{I} + \mathbf{F} \hat{\Lambda} \mathbf{F}^\top)^{-1} \mathbf{Y}$. **Theorem 1 end**

Lemma 1: $\mathcal{J}_\alpha(\alpha)$ in (14) is nonconvex function.

Proof of Lemma 1:

The data-dependent term $\mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y}$ in (14) is studied. Using (15), (16), it can be transformed as

$$\begin{aligned} & \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} \\ &= \frac{1}{\lambda} \mathbf{Y}^\top \mathbf{Y} - \frac{1}{\lambda} \mathbf{Y}^\top \mathbf{F} \Sigma_\theta \mathbf{F}^\top \frac{1}{\lambda} \mathbf{Y} \\ &= \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F} \mathbf{m}_\theta\|_2^2 + \mathbf{m}_\theta^\top \Lambda^{-1} \mathbf{m}_\theta \\ &= \min_{\mathbf{x}} \left\{ \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F} \mathbf{x}\|_2^2 + \mathbf{x}^\top \Lambda^{-1} \mathbf{x} \right\} \end{aligned} \quad (21)$$

The minimisation issue is simply demonstrated to be convex in θ and α dimensions. Define $\rho(\alpha) \triangleq \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| + \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N \log \xi(\alpha_j)$. $\log |x|$ is concave function. Furthermore, $\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top$ is an affine function of α . When $\alpha \geq 0$, it is positive semidefinite. This means $\log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top|$ is a concave, nondecreasing function of α . We can see that $\rho(\alpha)$ is a concave function with respect to α .

4.2.2. Modified objective Function in θ estimation

We use the modified objective function of θ with a penalty function. By analyzing the corresponding objective function of (14) in the α space, the analogous objective function is subsequently shown to be nonconvex as well.

Theorem 2: Solving the optimisation problem below yields the estimated value for θ given restrictions.

$$\min_{\theta} \|\mathbf{Y} - \mathbf{F}\theta\|_2^2 + \lambda r(\theta)$$

where penalty function $r(\theta) = \min_{\alpha \geq 0} \theta^\top \Lambda^{-1} \theta + \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| + \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N (\frac{1}{h} + \frac{1}{2\alpha_j}) \theta_j^T \theta_j - 2 \sum_{j=1}^N \log \xi(\alpha_j)$.

Proof of Theorem 2

Using the data-dependent term in (21) and $\mathcal{J}_\alpha(\alpha)$ in (14), a stringent upper boundary auxiliary function can be created on $\mathcal{J}_\alpha(\alpha)$ as

$$\begin{aligned} \mathcal{J}_{\alpha, \theta}(\alpha, \theta) &= \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F}\theta\|_2^2 + \theta^\top \Lambda^{-1} \theta + \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| \\ &+ \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N (\frac{1}{h} + \frac{1}{2\tau_j}) \theta_j^T \theta_j - 2 \sum_{j=1}^N \log \xi(\alpha_j). \end{aligned}$$

When we minimise $\mathcal{J}_{\alpha, \theta}(\alpha, \theta)$ over α and obtain

$$\begin{aligned} \mathcal{J}_\theta(\theta) &\triangleq \min_{\alpha \geq 0} \mathcal{J}_{\alpha, \theta}(\alpha, \theta) \\ &= \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F}\theta\|_2^2 + \min_{\alpha \geq 0} (\theta^\top \Lambda^{-1} \theta + \log |\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top| \\ &+ \sum_{j=1}^N p(\alpha_j) - 2 \sum_{j=1}^N (\frac{1}{h} + \frac{1}{2\tau_j}) \theta_j^T \theta_j - 2 \sum_{j=1}^N \log \xi(\alpha_j)) \end{aligned} \tag{22}$$

We can see from the derivations in (21) that the poseterior mean $\mathbf{m}_\theta$ is the estimate of the parameter θ .

Lemma 2 In Theorem2, the penalty function $r(\theta)$ promotes sparsity on the weights by being a non-decreasing, concave function of θ .

Proof of Lemma 2:

It is obvious that $\rho(\alpha)$ in Lemma 1 is concave. Using the duality lemma (see Sec. 4.2 in [35]), $\rho(\alpha)$ is denoted as $\min_{\alpha^* \geq 0} \langle \alpha^*, \alpha \rangle - \rho^*(\alpha^*)$, where $\rho^*(\alpha^*)$ is defined as the concave conjugate of $\rho(\alpha)$ and $\rho^*(\alpha^*) = \min_{\alpha \geq 0} \langle \alpha^*, \alpha \rangle - \rho(\alpha)$. Using (21), function $\mathcal{J}_{\alpha, \theta}(\alpha, \theta)$ can be re-written as

$$\begin{aligned} \mathcal{J}_{\alpha, \theta}(\alpha, \theta) &\triangleq \langle \alpha^*, \alpha \rangle - \rho^*(\alpha^*) + \mathbf{Y}^\top (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^\top)^{-1} \mathbf{Y} \\ &= \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F} \theta\|_2^2 + \sum_j \left(\frac{\theta_j^2}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}} + \alpha_j^* \alpha_j \right) - \rho^*(\alpha^*) \end{aligned} \quad (23)$$

$r(\theta)$ is re-expressed as

$$r(\theta) = \min_{\alpha, \alpha^* \geq 0} \left\{ \sum_j \left(\frac{\theta_j^2}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}} + \alpha_j^* \alpha_j \right) - \rho^*(\alpha^*) \right\} \quad (25)$$

It's easy to see that when $\alpha_j = \frac{1}{2} \sqrt{(\frac{2}{h} + \frac{1}{\tau_j})^2 + \frac{4\theta_j^2}{\alpha_j^*}} + (\frac{1}{h} + \frac{2}{\tau_j})$, $r(\theta)$ reaches the minimum over α . Substituting α_j into $r(\theta)$

$$r(\theta) = \min_{\alpha^* \geq 0} \left\{ \sum_j 2 \sqrt{\frac{\alpha_j}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}}} \alpha_j^* |\theta_j| - \rho^*(\alpha^*) \right\} \quad (26)$$

When $r(\theta)$ is minimum, θ is much more sparse. From (26), $r(\theta)$ is non-decreasing, concave function of θ .

4.2.3. Parameter estimation with sparse method

From (23), we can see that $\rho^*(\alpha^*)$ does not affect the estimation of parameters α , so $\mathcal{J}_{\alpha, \theta}(\alpha, \theta)$ is redefined as

$$\mathcal{J}_{\alpha^*}(\alpha, \theta) \triangleq \frac{1}{\lambda} \|\mathbf{Y} - \mathbf{F} \theta\|_2^2 + \sum_j \left(\frac{\theta_j^2}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}} + \alpha_j^* \alpha_j \right) \quad (27)$$

For a fixed α^* , we notice that $\mathcal{J}_{\alpha^*}(\alpha, \theta)$ is jointly convex in θ and α . In (27), we can obtain that $\frac{\theta_j^2}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}} + \alpha_j^* \alpha_j \geq 2 \sqrt{\frac{\alpha_j}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}}} \alpha_j^* |\theta_j|$. When $\alpha_j = \frac{1}{2} \sqrt{(\frac{2}{h} + \frac{1}{\tau_j})^2 + \frac{4\theta_j^2}{\alpha_j^*}} + (\frac{1}{h} + \frac{2}{\tau_j})$, $\mathcal{J}_{\alpha^*}(\alpha, \theta)$ is minimized for any θ . Substitute the α_j into $\mathcal{J}_{\alpha^*}(\alpha, \theta)$, $\hat{\theta}$ can be obtained as follow

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{F} \theta\|_2^2 + 2\lambda \sum_{j=1}^N \sqrt{\frac{\alpha_j}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}}} \alpha_j^* |\theta_j| \right\} \quad (28)$$

Due to the concavity of $r(\theta)$, the objective function in **Theorem 2** can be optimised using a re-weighted L_1 -minimisation in a similar k th way as was considered in (27).

In order to obtain more stable and accurate parameter $\hat{\theta}$, the re-estimated method is put forward (Algorithm 1). At k -th iteration, the modified weight is then supplied by $u_j^{(k)} \triangleq \frac{\partial r(\theta)}{2\partial|\theta_j|_{\theta=\theta^{(k)}}} = \sqrt{\frac{\alpha_j}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}}} \cdot \sqrt{\alpha_j^*}$. On the basis of the aforementioned, we can now describe how the parameters can be updated. To begin, we set the iteration count k to zero, $u_j^{(0)} = 1$ and initialise $\eta_j^{(k+1)} = \sqrt{\frac{\alpha_j^{(k)}}{\alpha_j^{(k)} - \frac{2}{h} - \frac{1}{\tau_j^{(k)}}}}$. We get $u_j^{(k)} = \eta_j^{(k)} \cdot \sqrt{\alpha_j^{*(k)}}$. Consider $\mathcal{J}_{\alpha,\theta}(\alpha, \theta)$ again. For any fixed α and θ , the tightest bound can be obtained by minimising over α^* . α^* is estimated, which equals the gradient of the function $\rho(\alpha)$ in **Lemma 1**. The estimation of α^* is computed as

$$\begin{aligned}\hat{\alpha}^* &= \nabla_{\alpha} \rho(\alpha) \\ &= \text{diag} \left[\mathbf{F}^{\top} (\lambda \mathbf{I} + \mathbf{F} \Lambda \mathbf{F}^{\top})^{-1} \mathbf{F} \right] + p'(\alpha) - 2\zeta(\alpha),\end{aligned}\tag{29}$$

where $p'(\alpha) = [p'(\alpha_1), \dots, p'(\alpha_N)]^{\top}$, $\zeta(\alpha) = \left[\frac{\xi'(\alpha_1)}{\xi'(\alpha_1)}, \dots, \frac{\xi'(\alpha_N)}{\xi'(\alpha_N)} \right]^{\top}$. The optimal $\alpha^{*(k+1)}$ can then be replaced by $\hat{\alpha}^{*(k+1)} = \text{diag} \left[\mathbf{F}^{\top} (\lambda \mathbf{I} + \mathbf{F} \Lambda^{(k)} \mathbf{F}^{\top})^{-1} \mathbf{F} \right] + p'(\alpha^{(k)}) - 2\zeta(\alpha)$. After computing the estimation of

$$\alpha_j = \frac{1}{2} \sqrt{\left(\frac{2}{h} + \frac{1}{\tau_j}\right)^2 + \frac{4\theta_j^2}{\alpha_j^*}} + \left(\frac{1}{h} + \frac{2}{\tau_j}\right),\tag{30}$$

we can compute $\alpha_j^{*(k+1)}$, which gives

$$\alpha_j^{*(k+1)} = \mathbf{F}_j^{\top} (\lambda \mathbf{I} + \mathbf{F}_j \Lambda^{(k)} \mathbf{F}_j^{\top})^{-1} \mathbf{F}_j + p'(\alpha_j^{(k)}) - 2\zeta(\alpha_j^{(k)})\tag{31}$$

$\alpha_j^{(k+1)}$ can then define

$$\alpha_j^{(k+1)} = \frac{1}{2} \sqrt{\left(\frac{2}{h} + \frac{1}{\tau_j^{(k)}}\right)^2 + \frac{4(\theta_j^{(k)})^2}{\alpha_j^{*(k)}}} + \left(\frac{1}{h} + \frac{2}{\tau_j^{(k)}}\right)$$

Algorithm 1 summarizes the above-mentioned procedure.

5. Numerical example

We will give three numerical examples: Narendra-Li Model[38], NARX model[37], and Kernel state-space models [39]. The utility and performance of Algorithm 1 are proven on three simulation cases in this section. The performance of Algorithm 1 on examples involving a well studied and challenging nonlinear system is then illustrated. The root mean squared error (RMSE) criterion will be utilized to demonstrate the performance of the suggested identification

Algorithm 1 Nonlinear Identification Algorithm

Input:

- 1: Generate time series data from the system of discrete-time dynamics characterized by (1)
- 2: Choose the dictionary functions that will be used to build the dictionary matrix mentioned in Section 2;

Initialization:

- $u_j^0 = 1, \alpha_j^{(k)}, h, \tau_j^{(k)}$
- 3: **for** $K = 0, 1$, **do**
- 4: Solve the minimisation problem with L1-regularization and optimization method on θ .

$$\underset{\theta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{F}\theta\|_2^2 + 2\lambda \sum_{j=1}^N \sqrt{\frac{\alpha_j}{\alpha_j - \frac{2}{h} - \frac{1}{\tau_j}}} \alpha_j^* |\theta_j| \right\}$$

- 5: Update parameter $\eta_j^{(k+1)}$ and $u_j^{(k+1)}$.
- 6: $u_j^{(k+1)} \triangleq \eta_j^{(k+1)} \cdot \sqrt{\alpha_j^{*(k+1)}}$
- 7: Update parameter $\alpha_j^{(k+1)} = \frac{1}{2} \sqrt{\left(\frac{2}{h} + \frac{1}{\tau_j^{(k)}}\right)^2 + \frac{4(\theta_j^{(k)})^2}{\alpha_j^{*(k)}}} + \left(\frac{1}{h} + \frac{2}{\tau_j^{(k)}}\right)$
- 8: **end for**
- 9: **if** a stopping criterion is satisfied **then**
- 10: Break
- 11: **end if**

Output:

The sparse weight set of θ .

approach against noise perturbation. The i th estimate of parameter θ is denoted by $\hat{\theta}_i$ at the Monte-Carlo experiment. The RMSE at the experiment is defined as

$$RMSE = \sqrt{\frac{(\hat{\theta}_i - \theta^*)^2}{n}}$$

where θ^* represents the true system parameter vector, n is the trials. To validate the theoretical results, the identification of the structured state-space model in cases will be simulated in this part.

5.1. Example 1: Parameters identification general Narendra-Li Model

Consider the state space representation of a nonlinear system.

$$x_{t+1}^1 = \left(\frac{\alpha_1 x_t^1}{1 + (x_t^1)^2} + \alpha_2 \right) \sin(x_t^2) + \xi_1(t) \quad (32)$$

$$\begin{aligned} x_{t+1}^2 = & \beta_1 x_t^2 \cos(x_t^2) + \beta_2 x_t^1 \exp\left(-\frac{(x_t^1)^2 + (x_t^2)^2}{8}\right) \\ & + \frac{\beta_3 (u_t)^3}{1 + (u_t)^2 + 0.5 \cos(x_t^1 + x_t^2)} + \xi_2(t) \\ y_t = & \frac{x_t^1}{1 + 0.5 \sin(x_t^2)} + \frac{x_t^2}{1 + 0.5 \sin(x_t^1)} \end{aligned} \quad (33)$$

where the state variable $x_t = [x_t^1, x_t^2]^\top$. $\xi_i(t)$ is Gaussian white noise. To generate the estimation data, the system is excited with a uniformly distributed random input signal $u(t) \in [-2.5, 2.5]$ with $1 \leq t \leq 500$. The validation data set is generated with the input

$$u(t) = \sin \frac{2\pi t}{10} + \sin \frac{2\pi t}{25}, \quad t = 1, \dots, 500$$

let $\Phi_1^t = \frac{x_t^1}{1 + (x_t^1)^2} \sin(x_t^2)$, $\Phi_2^t = \sin(x_t^2)$, $\Phi_3^t = x_t^2 \cos(x_t^2)$, $\Phi_4^t = x_t^1 \exp\left(-\frac{(x_t^1)^2 + (x_t^2)^2}{8}\right)$, $\Phi_5^t = \frac{\beta_3 (u_t)^3}{1 + (u_t)^2 + 0.5 \cos(x_t^1 + x_t^2)}$. The dictionary matrix Φ can be built as follows because there are two state variables.

$$\Phi = \begin{bmatrix} \Phi_1^t & \Phi_2^t & \Phi_3^t & \Phi_4^t & \Phi_5^t \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Phi_1^{t+M-1} & \Phi_2^{t+M-1} & \Phi_3^{t+M-1} & \Phi_4^{t+M-1} & \Phi_5^{t+M-1} \end{bmatrix} \quad (34)$$

Then the output can be defined as

$$\mathbf{y}_i \triangleq [x_{t+1}^1, \dots, x_{t+M}^1]^\top \in \mathbb{R}^{M \times 1}, i = 1, 2$$

Using the dictionary matrix Φ in (34), the true value of parameter θ_i for the model in (35) should be as follows:

$$\begin{aligned} \theta_{true} &= [\theta^{(1)}, \theta^{(2)}] \\ &= \begin{bmatrix} \alpha_1 (= 4) & 0 \\ \alpha_2 (= 3) & 0 \\ 0 & \beta_1 (= 1.4) \\ 0 & \beta_2 (= 1.5) \\ 0 & \beta_3 (= 1.6) \end{bmatrix} \end{aligned} \quad (35)$$

The values in brackets are the correct parameter values. In our study, we used $T = 1000$ data samples for learning and added a white Gaussian measurement noise of $\xi_i(t) = 0.1$ to the

Figure 1: RMSE of state x_1 and x_2

Figure 2: Output of the state model: (a) $\xi_1 = 0.1$ and (b) $\xi_1 = 0.3$

Table 1: Accuracy comparison of different method

Method	RMSE	data
Our paper	0.039	1000
Bayesian Learning [38]	0.06	2000
DWO[40]	0.43	50000
MOD[41]	0.46	50000
AHH[42]	0.31	2000
MARS[42]	0.49	2000

training data. We apply Algorithm 1 in the (32) and (33). The coefficients $\theta^{(j)}$ to learn is in the (35), with prior data. RMSE of θ is computed in the simulation, the result of which is 0.039. Despite using 1000 times as many data points, our finding is substantially better than the significantly larger simulation errors in [38]. We acquire a result that is nearly identical to the noise-free data in the more interesting scenario of measurement noise in the training data. Compare to some previous results reported in the literature [38, 40–42] with our method. In this experiment, the Algorithm 1 was executed 8 times. The average of the RMSE is computed. The RMSE of y_t is 0.06. From the Table 1, we can see that our method perform the best.

In this paper, we look examine how well the method performs for various $\xi_i(t) = 0.1, 0.3$, and see that Algorithm 1 is robust, the RMSE is not increased fast from Fig.2. In our experiment, 60 testing data is selected for testing and the predicting result is compared with the true value. It turns out this will give a satisfactory performance. It is clear that the proposed model is capable enough to well describe the system behavior.

5.2. Example 2: Application to a NARX model

We analyze the following polynomial terms model for a single-input single-output (SISO) nonlinear autoregressive system with exogenous (NARX) input in this example[37].

$$y(t_{k+1}) = 0.7y(t_{k-1}) - 0.5y(t_{k-2}) + 0.6u^2(t_{k-2}) - 0.7y(t_{k-2})u(t_{k-1}) + \xi(t_k) \quad (36)$$

with $y, u, \xi \in \mathbb{R}$. In expanded form, we may write (36) as:

$$\begin{aligned} y(t_{k+1}) &= w_1 + w_2y(t_k) + \dots + w_{m_x+2}y(t_{k-m_x}) + \dots + w_Ny^{d_x}(t_{k-m_x})u^{d_u}(t_{k-m_u}) + \xi(t_k) \\ &= \mathbf{w}^\top \mathbf{f}(y(t_k), \dots, y(t_{k-m_x}), u(t_k), \dots, u(t_{k-m_u})) + \xi(t_k) \end{aligned} \quad (37)$$

Model (36) is the general form of (37). d_x and d_u are the degree of the output and input; m_y and m_u is the given memory order of the output and input; $\mathbf{w}^\top = [w_1, \dots, w_N] \in \mathbb{R}^N$ is the weight vector; and $\mathbf{f}(y(t_k), \dots, y(t_{k-m_y}), u(t_k), \dots, u(t_{k-m_u})) = [f_1(\cdot), \dots, f_N(\cdot)]^\top \in \mathbb{R}^N$ is the functions vector. Taking the NARX model (36) as a example, we set that $d_y = 1$, $d_u = 2$, $m_y = 2$, $m_u = 2$. This yields $\mathbf{f}(\cdot) \in \mathbb{R}^{28}$ and thus $\mathbf{w} \in \mathbb{R}^{28}$. Since $\mathbf{w} \in \mathbb{R}^4$, only four of the twenty-eight linked weights w_i are nonzero. The estimated parameter w agrees with the true value, as shown in Fig. 3. The predicting performance of Algorithm 1 is in the Fig. 4 From the figure, the predicted and exact trajectories match well with different $\xi_i(t) = 0.1, 0.3$. The RMSE is 0.021 and 0.074.

Figure 3: The ditribution of w : the above is true w , the below is the estimated w model

Figure 4: Output of the state model: (a) $\xi = 0.1$ and (b) $\xi = 0.3$

Figure 5: Compare of sparse w with True value

5.3. Example 3: Kernel state-space models(KSSM) for Autoregressive Modeling

Kernel state-space models(KSSM) is autoregressive model, which satisfy the τ -order difference equation. As seen below, the model may be described as a first-order multivariate process.

$$\bar{\mathbf{x}}_{t+1} = F_t(\bar{\mathbf{x}}_t) + V_t \quad (38)$$

where $\bar{\mathbf{x}}_t = [\mathbf{x}_t, \dots, \mathbf{x}_{t-\tau+1}]^T$, $F_t(\bar{\mathbf{x}}_t) = [f_t(\mathbf{x}_t, \dots, \mathbf{x}_{t-\tau+1}), \mathbf{x}_t, \dots, \mathbf{x}_{t-\tau+2}]^T$, and $V_t = [\xi_t, 0, \dots, 0]^T$. The hidden state of an SSM can then be viewed as the process $\bar{\mathbf{x}}_t$, producing an SSM formulation of a complex autoregressive model with noisy. By using nonlinear autoregressive modeling with a fixed number of delayed samples, the model can be utilized to predict time series. In addition, if the state-transition function f_t is defined using kernels(39), we derive the suggested KSSM suited for autoregressive time series.

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{x}_t \\ \vdots \\ \mathbf{x}_{t-\tau+2} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N w_i k_i(x_i(t_k)) \\ \mathbf{x}_t \\ \vdots \\ \mathbf{x}_{t-\tau+2} \end{bmatrix} + \begin{bmatrix} \xi_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (39)$$

$$Y_t = h(\bar{\mathbf{x}}_t) + V_t$$

where Y_t is the observed process, $h(\bar{\mathbf{x}}_t)$ is the observation function, V_t is observation noise, and $w = [w_1, \dots, w_N]$ is the weight. Periodic time series is widely used in physics, engineering and biology. We take fourier kernel function in the KSSM. Consider 5 candidate kernel functions for $k_i(\cdot)$: $\sin x_i$, $\cos x_i$, x_i , $\sin 2x_i$ and $\cos 2x_i$.

The estimation data in the experiment has 500 sample points, and Fig.5 shows the simulated outputs of the two processes evaluated on the validation set. In the figure, We compare the true and estimated value w using the probability distribution and dispersoid distribution. The RMSE of w is 0.16, which is a satisfactory result.

6. Conclusion

The parameter estimation of nonlinear discrete-time systems using noisy state data in state-space form is the subject of this work. For parameter estimation and prediction, a sparse Bayesian convex optimisation technique was presented. The suggested method considers data correlation, parameter priori, and constraints, all of which are explicitly modeled using Bayesian sparse learning and optimization. The sole assumption made regarding the model's structure is that there are just a few key variables governing the dynamics, resulting in sparse equations in the space of potential functions. The number of basis functions is produced at random. The fundamental problem with identification is that it is divided into two parts: The first step is to create a reweighted L1-regularized least squares solver, with the regularization value chosen from the optimization point. The second step, the stein approach can be used to create a new goal function. It usually captures more information about the reliance of the data indicators than the function maximum likelihood(ML) technique alone.

References

- [1] Ljung L. Perspectives on system identification[J]. Annual Reviews in Control, 2010, 34(1): 1-12.
- [2] Quaranta G, Lacarbonara W, Masri S F. A review on computational intelligence for identification of nonlinear dynamical systems[J]. Nonlinear Dynamics, 2020, 99(2): 1709-1761.
- [3] Kumpati S N, Kannan P. Identification and control of dynamical systems using neural networks[J]. IEEE Transactions on neural networks, 1990, 1(1): 4-27.
- [4] Leontaritis I J, Billings S A. Input-output parametric models for non-linear systems part II: stochastic non-linear systems[J]. International journal of control, 1985, 41(2): 329-344.
- [5] Rangan S, Wolodkin G, Poolla K. New results for Hammerstein system identification[C]// Proceedings of 1995 34th IEEE Conference on Decision and Control. IEEE, 1995, 1: 697-702.
- [6] Billings S A. Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains[M]. John Wiley and Sons, 2013.
- [7] Haber R, Unbehauen H. Structure identification of nonlinear dynamic systems—a survey on input-output approaches[J]. Automatica, 1990, 26(4): 651-677.
- [8] Barahona M, Poon C S. Detection of nonlinear dynamics in short, noisy time series[J]. Nature, 1996, 381(6579): 215-217.
- [9] Ljung L. Theory for the User[J]. System Identification, 1987.
- [10] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1977, 39(1): 1-22.
- [11] Gibson S, Wills A, Ninness B. Maximum-likelihood parameter estimation of bilinear systems[J]. IEEE Transactions on Automatic Control, 2005, 50(10): 1581-1596.

- [12] Duncan S, Gyongy M. Using the EM algorithm to estimate the disease parameters for smallpox in 17th century London[C]//2006 IEEE International Symposium on Intelligent Control. IEEE, 2006: 3312-3317.
- [13] Solin A, Sarkka S. Hilbert space methods for reduced-rank Gaussian process regression[J]. *Statistics and Computing*, 2020, 30(2): 419-446.
- [14] Svensson A, Schon T B. A flexible state-space model for learning nonlinear dynamical systems[J]. *Automatica*, 2017, 80: 189-199.
- [15] Frigola R. Bayesian time series learning with Gaussian processes[D]. University of Cambridge, 2015.
- [16] Cerone V, Piga D, Regruto D. Enforcing stability constraints in set-membership identification of linear dynamic systems[J]. *Automatica*, 2011, 47(11): 2488-2494.
- [17] Zavlanos M M, Julius A A, Boyd S P, et al. Inferring stable genetic networks from steady-state data[J]. *Automatica*, 2011, 47(6): 1113-1122.
- [18] Tibshirani R. Regression shrinkage and selection via the lasso[J]. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58(1): 267-288.
- [19] Hastie T, Tibshirani R, Friedman J. The elements of statistical learnin[J]. Cited on, 2009: 33.
- [20] James G, Witten D, Hastie T, et al. An introduction to statistical learning[M]. New York: springer, 2013.
- [21] Donoho D L. Compressed sensing[J]. *IEEE Transactions on information theory*, 2006, 52(4): 1289-1306.
- [22] Candes E J, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information[J]. *IEEE Transactions on information theory*, 2006, 52(2): 489-509.
- [23] Tropp J A, Gilbert A C. Signal recovery from random measurements via orthogonal matching pursuit[J]. *IEEE Transactions on information theory*, 2007, 53(12): 4655-4666.
- [24] Stein C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables[C]//Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2:Probability Theory. The Regents of the University of California, 1972.
- [25] Cerone V, Piga D, Regruto D. Enforcing stability constraints in set-membership identification of linear dynamic systems[J]. *Automatica*, 2011, 47(11): 2488-2494.
- [26] Brunton S L, Tu J H, Bright I, et al. Compressive sensing and low-rank libraries for classification of bifurcation regimes in nonlinear dynamical systems[J]. *SIAM Journal on Applied Dynamical Systems*, 2014, 13(4): 1716-1732.
- [27] Bai Z, Wimalajeewa T, Berger Z, et al. Low-dimensional approach for reconstruction of airfoil data via compressive sensing[J]. *AIAA journal*, 2015, 53(4): 920-933.

- [28] Arnaldo I, O'Reilly U M, Veeramachaneni K. Building predictive models via feature synthesis[C] //Proceedings of the 2015 annual conference on genetic and evolutionary computation. 2015: 983-990.
- [29] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.
- [30] Efron B, Hastie T, Johnstone I, et al. Least angle regression[J]. The Annals of statistics, 2004, 32(2): 407-499.
- [31] Brunton S L, Proctor J L, Kutz J N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems[J]. Proceedings of the national academy of sciences, 2016, 113(15): 3932-3937.
- [32] Chen S S, Donoho D L, Saunders M A. Atomic decomposition by basis pursuit[J]. SIAM review, 2001, 43(1): 129-159.
- [33] Bishop C M. Pattern recognition[J]. Machine learning, 2006, 128(9).
- [34] Ma Z, Lai Y, Kleijn W B, et al. Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling[J]. IEEE transactions on neural networks and learning systems, 2018, 30(2): 449-463.
- [35] Tipping M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of machine learning research, 2001, 1(Jun): 211-244.
- [36] Svensson A, Schon T B. A flexible state-space model for learning nonlinear dynamical systems[J]. Automatica, 2017, 80: 189-199.
- [37] Pan W, Yuan Y, Goncalves J, et al. A sparse Bayesian approach to the identification of nonlinear state-space systems[J]. IEEE Transactions on Automatic Control, 2015, 61(1): 182-187.
- [38] Svensson A, Schon T B. A flexible state-space model for learning nonlinear dynamical systems[J]. Automatica, 2017, 80: 189-199.
- [39] Tobar F, Djuric P M, Mandic D P. Unsupervised state-space modeling using reproducing kernels[J]. IEEE Transactions on Signal Processing, 2015, 63(19): 5210-5221.
- [40] Roll J, Nazin A, Ljung L. Nonlinear system identification via direct weight optimization[J]. Automatica, 2005, 41(3): 475-490.
- [41] Stenman A. Model on demand: Algorithms, analysis and applications[M]. Department of Electrical Engineering, 1999.
- [42] Xu J, Huang X, Wang S. Adaptive hinging hyperplanes and its applications in dynamic system identification[J]. Automatica, 2009, 45(10): 2325-2332.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DataAvailabilityStatements.pdf](#)
- [DeclarationofInterestStatement20210131.pdf](#)