Regularization Methods for System Identification

Hyperparameter Estimation

Biqiang MU
A joint work with Tianshi Chen and Lennart Ljung
September 16 2019

Academy of Mathematics and Systems Science
Chinese Academy of Sciences
1. Introduction

2. Regularization methods for linear system identification

3. Conclusion
Introduction
Regularization methods have achieved a great success in statistics, machine learning, biometrics, etc, over the last two decades.
Regularization methods have achieved a great success in statistics, machine learning, biometrics, etc, over the last two decades.

A general framework

$$\hat{\theta} \triangleq \arg \min_{\theta \in \mathcal{M}} (\text{Fit} + \text{Complexity penalty})$$
Regularization methods

Regularization methods have achieved a great success in statistics, machine learning, biometrics, etc, over the last two decades.

A general framework

\[ \hat{\theta} \triangleq \arg \min_{\theta \in \mathcal{M}} (\text{Fit} + \text{Complexity penalty}) \]

The bias/variance tradeoff is at the heart of identification
Regularization methods

Regularization methods have achieved a great success in statistics, machine learning, biometrics, etc, over the last two decades.

A general framework

\[ \hat{\theta} \triangleq \arg \min_{\theta \in \mathcal{M}} (\text{Fit} + \text{Complexity penalty}) \]

The bias/variance tradeoff is at the heart of identification

Suppose that

\[ \theta_0 - \text{True parameter} \quad \hat{\theta} - \text{Estimate} \]

Bias-variance tradeoff

\[
\text{MSE} = \|E\hat{\theta} - \theta_0\|^2 + \|E\hat{\theta} - \theta_0\|^2 + E\|\hat{\theta} - E\hat{\theta}\|^2
\]

\( E\) deterministic

\( E\) random
Regularization methods have achieved a great success in statistics, machine learning, biometrics, etc, over the last two decades.

A general framework

\[ \hat{\theta} \triangleq \arg \min_{\theta \in \mathcal{M}} \left( \text{Fit} + \text{Complexity penalty} \right) \]

The bias/variance tradeoff is at the heart of identification

Suppose that

\[ \theta_0 - \text{True parameter} \quad \hat{\theta} - \text{Estimate} \]

Bias-variance tradeoff

\[
\text{MSE} = \mathbb{E} \| \hat{\theta} - \theta_0 \|^2 = \| E\hat{\theta} - \theta_0 \|^2 + E \| \hat{\theta} - E\hat{\theta} \|^2
\]

As complexity of \( \mathcal{M} \) increases, bias decreases but variance increases

To choose a proper complexity for the given data and to achieve a "good" bias/variance tradeoff
Regularization methods

Linear models

\[ Y = \Phi \theta_0 + V \]
Regularization methods

Linear models

\[ Y = \Phi \theta_0 + V \]

\( \ell_1 \)-norm regularization

\[ \hat{\theta}_1 \triangleq \arg \min_{\theta \in \mathcal{M}} (\|Y - \Phi \theta\|^2 + \lambda \|\theta\|_1) \]

To seek parsimonious models: regularization is a prime tool for sparsity
Regularization methods

Linear models

\[ Y = \Phi \theta_0 + V \]

\( \ell_1 \)-norm regularization

\[ \hat{\theta}_1 \triangleq \arg \min_{\theta \in \mathcal{M}} (\|Y - \Phi \theta\|^2 + \lambda \|\theta\|_1) \]

To seek parsimonious models: regularization is a prime tool for sparsity

\( \ell_2 \)-norm regularization

\[ \hat{\theta}_2 \triangleq \arg \min_{\theta \in \mathcal{M}} (\|Y - \Phi \theta\|^2 + \lambda \|\theta\|_2^2) \]

The bias/variance tradeoff is at the heart of identification: regularization offers new techniques for robust smaller MSE
Can regularization methods bring forth some benefits for system identification?

Yes!
Regularization methods for linear system identification
Linear time-invariant (LTI) system identification is a classical and fundamental problem.
Linear time-invariant (LTI) system identification is a classical and fundamental problem.

**Output error (OE) systems**

\[ y(t) = \sum_{k=1}^{\infty} g_k^0 u(t - k) + v(t), \quad t = 1, 2, \ldots \]
**Impulse response identification**

Linear time-invariant (LTI) system identification is a classical and fundamental problem.

**Output error (OE) systems**

\[
y(t) = \sum_{k=1}^{\infty} g_{0}^{k} u(t - k) + v(t), \ t = 1, 2, \cdots
\]

**The Goal**

To identify the impulse response sequence

\[
\theta_{0} = [g_{1}^{0}, g_{2}^{0}, \cdots]^{T} \text{ (infinite parameters)}
\]

as well as possible by a finite number of data

\[
\{u(t), y(t)\}_{t=1}^{N}
\]
The impulse response identification could be \textit{ill-conditioned} in practice since it involves to estimate an infinite number of parameters.
The impulse response identification could be \textit{ill-conditioned} in practice since it involves to estimate an infinite number of parameters.

The identification is to make the ill-conditioned problem \textit{well-conditioned}.

Two routes:

- Parametric methods (Classical methods: maximum likelihood, prediction error method, etc.):

\[
\sum_{k=1}^{\infty} g_k^0 q^{-k} = \frac{b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}}{1 + f_1 q^{-1} + \cdots + f_{n_f} q^{-n_f}}
\]

- Model class selection
- Model order selection: AIC, BIC, cross validation

\textbf{Asymptotic optimality}

- Nonparametric methods
Motivation

- Parametric methods are not as reliable as expected for short, ill-conditioned, low signal-to-noise ratio data
Nonparametric methods

Motivation

• Parametric methods are not as reliable as expected for short, ill-conditioned, low signal-to-noise ratio data

A high order finite impulse response (FIR) system, (e.g. $n = 100$)

$$y(t) = \sum_{k=1}^{n} g_k^0 u(t - k) + v(t)$$
Nonparametric methods

Motivation

- Parametric methods are not as reliable as expected for short, ill-conditioned, low signal-to-noise ratio data

A high order finite impulse response (FIR) system, (e.g. \( n = 100 \))

\[
y(t) = \sum_{k=1}^{n} g_k^0 u(t - k) + v(t)
\]

Prior

stability: \( g_k^0 \sim O(\tau^k) \) for some \( 0 < \tau < 1 \)
Nonparametric methods

Linear regression form

\[ Y = \Phi \theta_0 + V, \quad \theta_0 = [g_1^0, g_2^0, \ldots, g_n^0]^T \]

where

\[
\Phi = \begin{bmatrix}
    u(0) & u(-1) & \ldots & u(-n + 1) \\
    u(1) & u(0) & \ldots & u(-n + 2) \\
    \vdots & \vdots & \ddots & \vdots \\
    u(N - 1) & u(N - 2) & \ldots & u(N - n)
\end{bmatrix}
\]

\[
Y = \begin{bmatrix}
    y(1) \\
    y(2) \\
    \vdots \\
    y(N)
\end{bmatrix}^T
\]

\[
V = \begin{bmatrix}
    v(1) \\
    v(2) \\
    \vdots \\
    v(N)
\end{bmatrix}^T
\]
An example

Input-output data of a linear dynamic system:

- Data size: 250
- Input: a filtered white noise
- Noise: a white noise with the signal to noise ratio 5.45

To estimate the first 100 impulse response coefficients
Performance measure

Fit = 100 \times \left( 1 - \frac{\|\hat{\theta}_\text{im} - \theta_0\|}{\|\theta_0 - \bar{\theta}_0\|} \right), \quad \bar{\theta}_0 = \frac{1}{n} \sum_{k=1}^{n} g_k^0

where \(\hat{\theta}_\text{im}\) is the corresponding first \(n = 100\) impulse response of the estimate for \(\hat{\theta}\).
The OE-system of order 6 by CV

Fit = 36.78
Estimation results

The OE-system of order 6 by CV

Fit = 36.78

The best OE system of the order 7

Fit = 79.63

The estimate is sensitive to the choice of model order.
Estimation results

The OE-system of order 6 by CV

Fit = 36.78

The best OE system of the order 7

Fit = 79.63

The estimate is sensitive to the choice of model order.
Estimation results

The OE-system of order 6 by CV

Fit = 36.78

Regularization methods

Fit = 83.40
Regularization

Objective functions

\[ \ell(Y, \Phi \theta) + R(\theta) \]

- Loss term
- Regularization term

Characterize the feature of the noise

Ill-posed problem

Encode prior knowledge

Some examples

\[ \|Y\|_p^p + \|\|_q^q ; p_0 ; q_0 \]
Regularization

Objective functions

\[ \ell(Y, \Phi\theta) + R(\theta) \]

- **Loss term**
  - characterize the feature of the noise

- **Regularization term**
  - ill-posed problem
  - encode prior knowledge
Regularization

Objective functions

\[ \ell(Y, \Phi \theta) + R(\theta) \]

- **loss term**
- **regularization term**

Loss term

- characterize the feature of the noise

Regularization term

- ill-posed problem
- encode prior knowledge

Some examples

\[ \|Y - \Phi \theta\|_p^p + \lambda \|\theta\|_q^q, \ p \geq 0, \ q \geq 0 \]
Recall that
\[ \hat{\theta} \triangleq \arg \min_{\theta \in \mathcal{M}} (\text{Fit} + \text{Complexity penalty}) \]

Linear regression
\[ Y = \Phi \theta_0 + V, \quad \theta_0 = [g_1^0, g_2^0, \ldots, g_n^0]^T \]
\[ y(t) = \sum_{k=1}^{n} g_k^0 u(t - k) + v(t) \]

Least squares (LS) estimators:
\[ \hat{\theta}_{LS} \triangleq \arg \min_{\theta} \| Y - \Phi^T \theta \|^2 = (\Phi^T \Phi)^{-1} \Phi^T Y \]
\[ \text{MSE}(\hat{\theta}_{LS}) = E\| \hat{\theta}_{LS} - \theta_0 \|^2 = \sigma^2 \text{Tr}((\Phi^T \Phi)^{-1}) \]

Too many parameters? Put them on leashes!
\[ \hat{\theta}^R \triangleq \arg \min_{\theta \in \mathbb{R}^n} \| Y - \Phi \theta \|^2 + \sigma^2 \theta^T K^{-1} \theta = (\Phi^T \Phi + \sigma^2 K^{-1})^{-1} \Phi^T Y \]
where \( K \) is a positive semidefinite matrix to be tuned by the data.
A frequentist perspective

The estimator:

$$\hat{\theta}^R \triangleq \arg \min_{\theta} \| Y - \Phi \theta \|^2 + \sigma^2 \theta^T K^{-1} \theta = R^{-1} \Phi^T Y, \ R = \Phi^T \Phi + \sigma^2 K^{-1}$$

Bias

$$E\hat{\theta}^R - \theta_0 = \sigma^2 R^{-1} K^{-1} \theta_0 \neq 0$$

MSE

$$E\|\hat{\theta}^R - \theta_0\|^2 = \underbrace{\sigma^4 \theta_0^T P^{-1} R^{-1} R^{-1} K^{-1} \theta_0} + \underbrace{\sigma^2 \text{Tr}(R^{-1} \Phi^T \Phi R^{-1})}_{\text{bias's square}}$$

No regularization if $K^{-1} = 0$: Bias = 0 and Variance = $\sigma^2 (\Phi^T \Phi)^{-1}$

Proposition

If $\sigma^2 K^{-1} = \beta A$ and A is positive definite and fixed. Then we have

$$\text{MSE}(\hat{\theta}^R) \leq \text{MSE}(\hat{\theta}^{\text{LS}}), \ \text{when} \ 0 < \beta < 2\sigma^2 / (\theta_0^T A \theta_0)$$

The optimal kernel matrix for any data length

$$K = \theta_0 \theta_0^T$$
Bayesian Interpretation

Prior

$$\theta_0 \sim \mathcal{N}(0, K) \ (K : \text{Covariance/Kernel matrix})$$

Posterior

$$\theta_0|Y \sim \mathcal{N}(\hat{\theta}^R, \hat{K}^R)$$

$$\hat{\theta}^R = R^{-1}\Phi^TY, \quad \hat{K}^R = \sigma^2R^{-1}$$

$$R = \Phi^T\Phi + \sigma^2K^{-1}$$

This interpretation provides a clue to select $K$
Regularization for handling ill-posed problems (Tikhonov & Arsenic, 1977)\textsuperscript{1}

Regularization in system identification

Regularization for handling ill-posed problems (Tikhonov & Arsenic, 1977)\(^1\)

Regularization is not new in system identification

The first paper in system identification (Sjöberg et al., 1993)\(^2\)

\[
\hat{\theta}^R = \arg \min_{\theta} \| Y - \Phi \theta \|^2 + \gamma \| \theta \|^2
\]

\[
= (\Phi^T \Phi + \gamma I_n)^{-1} \Phi^T Y
\]

But no important progress until Pillonetto & De Nicolao (2010)\(^3\)


How to tune a "good" kernel $K$ by the data

The estimator:

$$\hat{\theta}^R = (\Phi^T \Phi + \sigma^2 K^{-1})^{-1} \Phi^T Y$$

Two extrema

$$\hat{\theta}^R = \begin{cases} 
0, & \text{if } K = 0 \\
\hat{\theta}^{LS}, & \text{if } K = \infty 
\end{cases}$$

How to tune a "good" kernel $K$ by the data
A two-step procedure

The seminal paper (Pillonetto & De Nicolao, 2010)\(^1\)

- Kernel design: determine the structure of \(K\) by using the prior knowledge

\[ K(\eta), \ \eta \text{ hyperparameter} \]

- Hyperparameter estimation: determine the hyperparameter by the data

---

Kernel design

**Cubic spline kernels** (Wahba, 1990)$^1$

\[
K_{CS}(i, j) = \begin{cases} 
  c \frac{i^2}{2} \left( j - \frac{i}{3} \right), & i \geq j \\
  c \frac{j^2}{2} \left( i - \frac{j}{3} \right), & i < j 
\end{cases}
\]

**Prior:** exponential decay

\[g^0_k \sim O(\tau^k) \text{ for some } 0 < \tau < 1\]

**Stable spline kernels** (Pillonetto & De Nicolao, 2010)$^2$

An exponential transform:

\[i \rightarrow \lambda^i\]

for some \(0 < \lambda < 1\)

\[
K_{SS}(i, j) = \begin{cases} 
  c \frac{\lambda^{2j}}{2} \left( \lambda^i - \frac{\lambda^i}{3} \right), & i \geq j \\
  c \frac{\lambda^{2j}}{2} \left( \lambda^j - \frac{\lambda^j}{3} \right), & i < j 
\end{cases}
\]

---


Heuristic methods

The optimal kernel

\[ K = \theta_0 \theta_0^T = \begin{bmatrix}
(g_1^0)^2 & g_1^0 g_2^0 & \cdots & g_1^0 g_n^0 \\
g_2^0 g_1^0 & (g_2^0)^2 & \cdots & g_2^0 g_n^0 \\
\vdots & \vdots & \ddots & \vdots \\
g_n^0 g_1^0 & g_n^0 g_2^0 & \cdots & (g_n^0)^2
\end{bmatrix} \]

\[ \theta_0 = [g_1^0, \cdots, g_n^0]^T \]

Prior

\[ g_k^0 \sim O(\tau^k) \text{ for some } 0 < \tau < 1 \]
Heuristic methods

DI kernel

\[ K(\eta) = c \text{diag}(\lambda, \ldots, \lambda^n) \]
\[ \eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\} \]
Heuristic methods

**DI kernel**

\[ K(\eta) = c \operatorname{diag}([\lambda, \cdots, \lambda^n]) \]

\[ \eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\} \]

**DC kernel**

\[ K_{i,j}(\eta) = c \lambda^{(i+j)/2} \rho^{|i-j|} \]

\[ K(\eta) = c \begin{bmatrix} \lambda & \lambda^{3/2} \rho & \cdots & \lambda^{n+1/2} \rho^{n-1} \\ \lambda^{3/2} \rho & \lambda^2 & \cdots & \lambda^{n+2/2} \rho^{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda^{n+1/2} \rho^{n-1} & \lambda^{n+2/2} \rho^{n-2} & \cdots & \lambda^n \end{bmatrix} \]

with hyperparameters \( \eta = [c, \lambda, \rho]^T \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1, |\rho| \leq 1\} \).
Heuristic methods

**TC kernel** (Chen et al., 2012) \(^1\)

A special case of DC kernel with \(\rho = \sqrt{\lambda}\).

\[
K_{k,j}(\eta) = c \min(\lambda^k, \lambda^j), \quad K(\eta) = c \begin{bmatrix}
\lambda & \lambda^2 & \cdots & \lambda^n \\
\lambda^2 & \lambda^2 & \cdots & \lambda^n \\
\vdots & \vdots & \ddots & \vdots \\
\lambda^n & \lambda^n & \cdots & \lambda^n
\end{bmatrix}
\]

with hyperparameters \(\eta = [c, \lambda]^T \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}\).

---

Multiple kernels (Chen et al., 2014)\(^1\)

Better capture complicated dynamics of the system

\[
K(\eta) = \sum_{i=1}^{m} \eta_i K_i, \quad \eta = [\eta_1, \ldots, \eta_m]
\]

where \(K_i\) has different dynamic behavior, e.g. decaying rate and magnitude.

---

Hyperparameter estimation

The goal

• To estimate the hyperparameters based on the data

The essence

• To tune model complexity in a continuous way

Some commonly used methods (Pillonetto et al., 2014)¹

1. Empirical Bayes (EB)
2. Stein’s unbiased risk estimator (SURE)
3. Cross validation (CV)

Empirical Bayes

Gaussian prior

\[
\theta \sim \mathcal{N}(0, K)
\]

\[
Y = \Phi \theta + V \sim \mathcal{N}(0, Q)
\]

\[
Q = \Phi K \Phi^T + \sigma^2 I_N
\]

Empirical Bayes (EB)

\[
\text{EB} : \hat{\eta}_{EB} = \arg \min_{\eta \in \Omega} Y^T Q^{-1} Y + \log \det(Q)
\]
Stein’s unbiased risk estimator (SURE)

**MSE** (for prediction ability):

\[ \text{MSE}(K) = E \| \Phi (\hat{\theta}^R - \theta_0) \|^2 \]

It is *intractable* to tune the hyperparameter by the MSE in practice.

SURE method

- To construct an *unbiased* estimator of the MSE

\[ \mathcal{F}_{\text{SURE}}(K) = \| Y - \Phi \hat{\theta}^R \|^2 + 2\sigma^2 \text{Tr}(R^{-1} \Phi^T \Phi) \]

\[ R = \Phi^T \Phi + \sigma^2 K^{-1} \]

- To estimate the hyperparameter \( \eta \) by

\[ \text{SURE} : \hat{\eta}_{\text{SURE}} = \arg \min_{\eta \in \Omega} \mathcal{F}_{\text{SURE}}(K(\eta)) \]
Cross-validation

Ideas

• divide the whole data into training data and validation data
• estimate on the training data
• evaluate on the validation data

Averaged prediction error

For each splitting way \( s \),

• \( s \) the index set of the validation data
• \( s^c \) the index set of the training data

where

\[ |s| = k, \quad \{1, \cdots, N\} = s \cup s^c \]

the averaged prediction error (APE) over the validation data is

\[
\text{APE}_s = \frac{1}{k} \sum_{t \in s} (y(t) - \phi(t)^T \hat{\theta}_{s^c})^2 = \frac{1}{k} \|Y_s - \Phi_s \hat{\theta}_{s^c}\|^2
\]

Advantage: does not require to estimate the noise variance \( \sigma^2 \)
Variants of CVs

1. Leave-\(k\)-out cross validation (LKOCV, intractable in general)

\[
\hat{\eta}_{\text{LKOCV}} = \arg \min_{\eta \in \Omega} \frac{1}{N} \sum_{s} \text{APE}_s \text{ (all choices)}
\]

2. Leave-one-out cross validation (LOOCV) \((k = 1)\)

\[
\hat{\eta}_{\text{LOOCV}} = \arg \min_{\eta \in \Omega} \frac{1}{N} \sum_{s} \text{APE}_s = \arg \min_{\eta \in \Omega} \frac{1}{N} \sum_{t=1}^{N} \left( \frac{y(t) - \hat{y}(t)}{1 - h_{tt}} \right)^2
\]

where \(H = \Phi(\Phi^T\Phi + \sigma^2 K^{-1})^{-1}\Phi^T\).

3. Generalized cross validation (GCV)

\[
\hat{\eta}_{\text{GCV}} = \arg \min_{\eta \in \Omega} \frac{1}{N} \sum_{t=1}^{N} \left( y(t) - \hat{y}(t) \right)^2 \\
\frac{1}{N} \left( \frac{\text{Tr}(H)}{N} \right)^2
\]
The key issue

How to choose a proper hyperparameter estimator for a given data?
Asymptotically theoretical properties

Suppose that

\[ \Phi^T \Phi / N \to \Sigma > 0 \text{ as } N \to \infty. \]

Then the **asymptotically optimal hyperparameter** in the MSE sense is (Mu et al., 2018c)\(^1\)

\[ \eta^* = \arg \min_{\eta \in \Omega} \theta_0^T K^{-1} \Sigma^{-1} K^{-1} \theta_0 - 2 \text{Tr}(\Sigma^{-1} K^{-1}) \]

depending on the true parameter, chosen kernel, and asymptotic covariance of the input

---

Theorem

- $\hat{\eta}_{\text{SURE}} \to \eta^*$
- $\hat{\eta}_{\text{EB}} \to \arg\min_{\eta \in \Omega} \theta_0^T K^{-1} \theta_0 + \log \det(K)$ (Mu et al., 2018c) \(^1\)

- $\hat{\eta}_{\text{GCV}} \to \eta^*$ (Mu et al., 2018a) \(^2\)
- $\hat{\eta}_{\text{LOOCV}} \to \eta^*$ if the input is bounded
- $\hat{\eta}_{\text{LKOCV}} \to \eta^*$ if $k/N \to 0$ and the input is bounded (Mu et al., 2018b) \(^3\)

---


Numerical illustrations

Systems: 1000 30th order OE test systems

3 Inputs:

- IT1, white Gaussian noise
- IT2, white Gaussian noise filtered by $1/(1 - 0.95q^{-1})^2$
- IT3, the impulsive input, $[\sqrt{N}, 0, \cdots, 0]$ (unbounded)

Noises: The SNR is uniformly distributed over $[1, 10]$

Sample sizes: $N = 500, 8000$

Kernel: TC kernel

Tuning methods:

- EB, LOOCV, GCV, SURE
- MSE for reference (optimal for any finite sample)
Table 1: Average fits for 1000 test systems.

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Sizes</th>
<th>EB</th>
<th>LOOCV</th>
<th>GCV</th>
<th>SURE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT1</td>
<td>500</td>
<td>86.16</td>
<td>86.24</td>
<td>86.24</td>
<td>86.03</td>
<td>87.02</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>96.44</td>
<td>96.60</td>
<td>96.60</td>
<td>96.60</td>
<td>96.67</td>
</tr>
<tr>
<td>IT2</td>
<td>500</td>
<td>39.03</td>
<td>-85.95</td>
<td>-84.84</td>
<td>-146.4</td>
<td>41.94</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>50.86</td>
<td>38.79</td>
<td>38.89</td>
<td>38.86</td>
<td>53.63</td>
</tr>
<tr>
<td>IT3</td>
<td>500</td>
<td>69.33</td>
<td>89.55</td>
<td>89.52</td>
<td>89.95</td>
<td>89.95</td>
</tr>
<tr>
<td></td>
<td>8000</td>
<td>81.42</td>
<td>96.64</td>
<td>96.64</td>
<td>96.70</td>
<td>96.70</td>
</tr>
</tbody>
</table>
Conclusion
Summary

- A brief introduction of regularization methods for impulse response identification of linear dynamic systems is given.
- Asymptotically theoretical properties of several hyperparameter estimation are shown.
Thanks for your listening
Questions?


