Regularization Methods for System Identification

Hyperparameter Estimation

Biqiang MU
A joint work with Tianshi Chen and Lennart Ljung

Septermber 16 2019

Academy of Mathematics and Systems Science
Chinese Academy of Sciences

Table of contents

Introduction

Regularization methods

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

Regularization methods

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

A general framework

$$\widehat{\theta} \stackrel{\triangle}{=} \arg\min_{\theta \in \mathcal{M}} \left(\text{Fit} + \text{Complexity penalty}\right)$$

Regularization methods

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

A general framework

$$\widehat{\theta} \stackrel{\triangle}{=} \arg\min_{\theta \in \mathcal{M}} \left(\text{Fit} + \text{Complexity penalty} \right)$$

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

$$\widehat{\theta} \stackrel{\triangle}{=} \arg\min_{\theta \in \mathcal{M}} (\text{Fit} + \text{Complexity penalty})$$

The bias/variance tradeoff is at the heart of identification

Suppose that

$$\theta_0 - \text{True parameter} \qquad \widehat{\theta} - \text{Estimate}$$

Bias-variance tradeoff

$$\underbrace{E\|\widehat{\theta} - \theta_0\|^2}_{\text{MSE}} = \underbrace{\|E\widehat{\theta} - \theta_0\|^2}_{\substack{\text{bias's square} \\ \text{deterministic}}} + \underbrace{E\|\widehat{\theta} - E\widehat{\theta}\|^2}_{\substack{\text{variance} \\ \text{random}}}$$

Regularization methods

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

A general framework

$$\widehat{\theta} \triangleq \arg\min_{\theta \in \mathcal{M}} \left(\text{Fit} + \text{Complexity penalty}\right)$$

The bias/variance tradeoff is at the heart of identification

Suppose that

$$\theta_0 - \text{True parameter} \qquad \widehat{\theta} - \text{Estimate}$$

Bias-variance tradeoff

$$\underbrace{E\|\widehat{\theta} - \theta_0\|^2}_{\text{MSE}} = \underbrace{\|E\widehat{\theta} - \theta_0\|^2}_{\substack{\text{bias's square} \\ \text{deterministic}}} + \underbrace{E\|\widehat{\theta} - E\widehat{\theta}\|^2}_{\substack{\text{variance} \\ \text{random}}}$$

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

Linear models

$$Y = \Phi\theta_0 + V$$

Regularization methods

Linear models

$$Y = \Phi\theta_0 + V$$

ℓ_1-norm regularization

$$\widehat{\theta}_1 \triangleq \arg\min_{\theta\in\mathcal{M}} \left(\|Y - \Phi\theta\|^2 + \lambda\|\theta\|_1\right)$$

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

Regularization methods

Linear models

$$Y = \Phi\theta_0 + V$$

ℓ_1-norm regularization

$$\widehat{\theta}_1 \triangleq \arg\min_{\theta \in \mathcal{M}} \left(\|Y - \Phi\theta\|^2 + \lambda \|\theta\|_1 \right)$$

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

ℓ_2-norm regularization

$$\widehat{\theta}_2 \triangleq \arg\min_{\theta \in \mathcal{M}} \left(\|Y - \Phi\theta\|^2 + \lambda \|\theta\|_2^2 \right)$$

The bias/variance tradeoff is at the heart of identification: regularization offers new techniques for robust smaller MSE

The main issue

Can regularization methods bring forth some benefits for system identification?

Yes!

Regularization methods for linear system identification

Impulse response identification

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

Impulse response identification

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

Output error (OE) systems

$$y(t) = \sum_{k=1}^{\infty} g_k^0 u(t-k) + v(t), \; t = 1, 2, \cdots$$

Impulse response identification

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

Output error (OE) systems

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

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

$$\{u(t), y(t)\}_{t=1}^{N}$$

Impulse response identification

The impulse response identification could be ill-conditioned in practice since it involves to estimate an infinite number of parameters

Impulse response identification

The impulse response identification could be ill-conditioned in practice since it involves to estimate an infinite number of parameters

The identification is to make the ill-conditioned problem 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

  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

$$\text{stability}: \; g_k^0 \sim O(\tau^k) \text{ for some } 0 < \tau < 1$$

Linear regression form

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

where

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

$$Y = \begin{bmatrix} y(1) & y(2) & \ldots & y(N) \end{bmatrix}^T$$

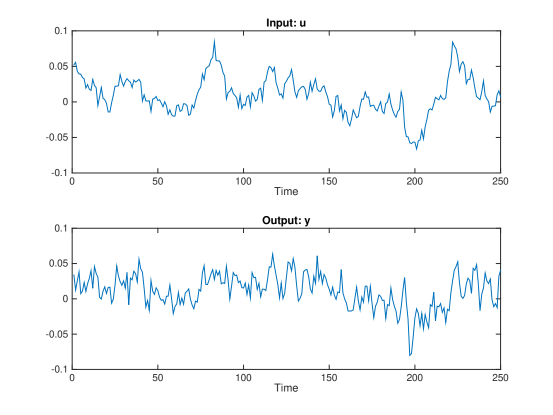
$$V = \begin{bmatrix} v(1) & v(2) & \ldots & v(N) \end{bmatrix}^T$$

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

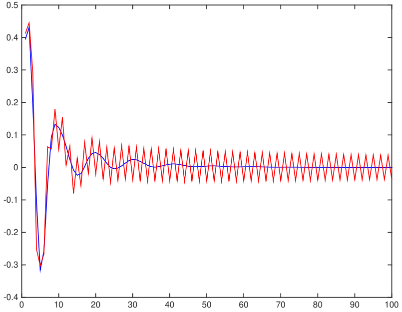
$$\text{Fit} = 100 \times \left(1 - \frac{\|\widehat{\theta}_{\text{im}} - \theta_0\|}{\|\theta_0 - \bar{\theta}_0\|}\right), \quad \bar{\theta}_0 = \frac{1}{n}\sum_{k=1}^{n} g_k^0$$

where $\widehat{\theta}_{\text{im}}$ is the corresponding first $n = 100$ impulse response of the estimate for $\widehat{\theta}$.

Estimation results

The OE-system of order 6 by CV

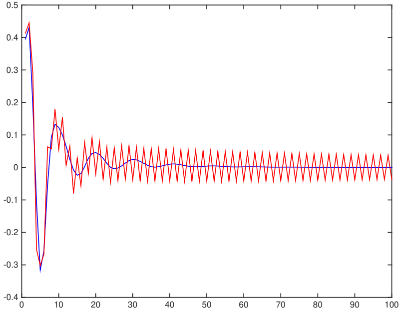
Fit = 36.78



Estimation results

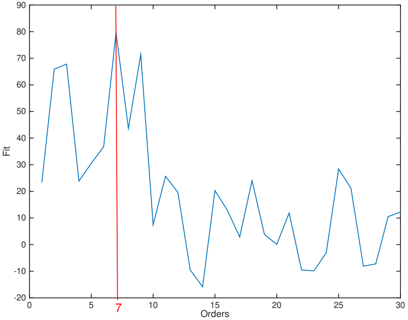
The OE-system of order 6 by CV

Fit = 36.78



The best OE system of the order 7

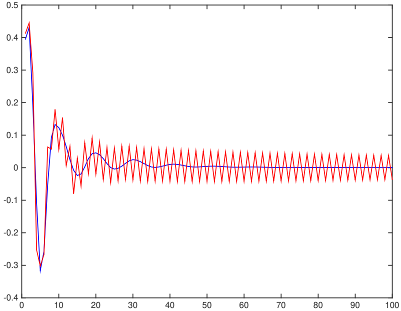
Fit = 79.63



Estimation results

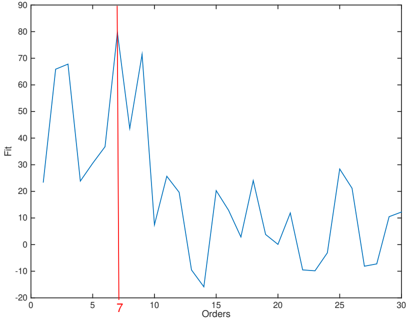
The OE-system of order 6 by CV

Fit = 36.78



The best OE system of the order 7

Fit = 79.63



The estimate is sensitive to the choice of model order

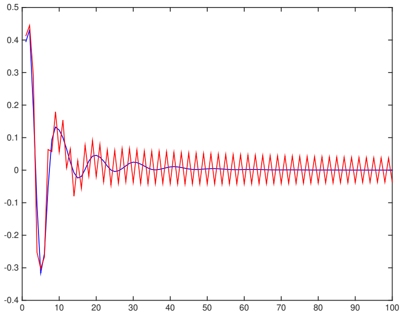
Estimation results

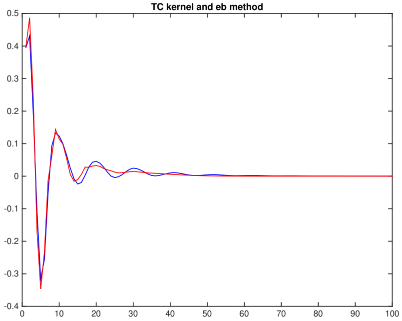
The OE-system of order 6 by CV

Fit = 36.78

Regularization methods

Fit = 83.40





Objective functions

$$\underbrace{\ell(Y, \Phi\theta)}_{\text{loss term}} + \underbrace{R(\theta)}_{\text{regularization term}}$$

Regularization

Objective functions

$$\underbrace{\ell(Y, \Phi\theta)}_{\text{loss term}} + \underbrace{R(\theta)}_{\text{regularization term}}$$

Loss term

- characterize the feature of the noise

Regularization term

- ill-posed problem
- encode prior knowledge

Regularization

Objective functions

$$\underbrace{\ell(Y, \Phi\theta)}_{\text{loss term}} + \underbrace{R(\theta)}_{\text{regularization term}}$$

Loss term

- characterize the feature of the noise

Regularization term

- ill-posed problem
- encode prior knowledge

Some examples

$$\|Y - \Phi\theta\|_p^p + \lambda\|\theta\|_q^q,\ p \geq 0, q \geq 0$$

Regularization

Recall that

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

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:

$$\widehat{\theta}^{\mathrm{LS}} \triangleq \arg\min_{\theta} \|Y - \Phi^T\theta\|^2 = (\Phi^T\Phi)^{-1}\Phi^T Y$$

$$\mathrm{MSE}(\widehat{\theta}^{\mathrm{LS}}) = E\|\widehat{\theta}^{\mathrm{LS}} - \theta_0\|^2 = \sigma^2 \mathbf{Tr}\big((\Phi^T\Phi)^{-1}\big)$$

Too many parameters? Put them on leashes!

$$\widehat{\theta}^{\mathrm{R}} \triangleq \arg\min_{\theta \in \mathbb{R}^n} \|Y - \Phi\theta\|^2 + \sigma^2\theta^T K^{-1}\theta = (\Phi^T\Phi + \sigma^2 K^{-1})^{-1}\Phi^T Y$$

where K is a positive semidefinite matrix to be tuned by the data.

A frequentist perspective

The estimator:

$$\widehat{\theta}^{\mathrm{R}} \triangleq \arg\min_{\theta\in\mathbb{R}^n} \|Y - \Phi\theta\|^2 + \sigma^2\theta^T K^{-1}\theta = R^{-1}\Phi^T Y,\ R = \Phi^T\Phi + \sigma^2 K^{-1}$$

Bias

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

MSE

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

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

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

The optimal kernel matrix for any data length

$$K = \theta_0\theta_0^T$$

Bayesian Interpretation

Prior

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

Posterior

$$\theta_0 | Y \sim \mathcal{N}(\widehat{\theta}^{\mathrm{R}}, \widehat{K}^{\mathrm{R}})$$
$$\widehat{\theta}^{\mathrm{R}} = R^{-1}\Phi^T Y, \ \widehat{K}^{\mathrm{R}} = \sigma^2 R^{-1}$$
$$R = \Phi^T \Phi + \sigma^2 K^{-1}$$

This interpretation provides a clue to select K

Regularization in system identification

Regularization for handling ill-posed problems (Tikhonov & Arsenic, 1977)[1]

[1]A. N. Tikhonov and V. Y. Arsenic. Solutions of Ill-Posed Problems, New York: John Wiley, 1977.
[2]J. Sjöberg, T. McKelvey, and L. Ljung. On the use of regularization in system identification. Proceedings of the 12th IFAC World Congress: 381–386, Sydney, Australia.
[3]G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93, 2010.

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]

$$\begin{aligned}\widehat{\theta}^{\mathrm{R}} &= \arg\min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\|\theta\|^2 \\ &= (\Phi^T\Phi + \gamma I_n)^{-1}\Phi^T Y\end{aligned}$$

But no important progress until Pillonetto & De Nicolao (2010)[3]

---

[1] A. N. Tikhonov and V. Y. Arsenic. Solutions of Ill-Posed Problems, New York: John Wiley, 1977.
[2] J. Sjöberg, T. McKelvey, and L. Ljung. On the use of regularization in system identification. Proceedings of the 12th IFAC World Congress: 381–386, Sydney, Australia.
[3] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93, 2010.

How to tune a ”good” kernel K by the data

The estimator:

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

Two extrema

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

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

[1] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93, 2010.

Kernel design

Cubic spline kernels (Wahba, 1990)[1]

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

Prior: exponential decay

$$g_k^0 \sim O(\tau^k) \text{ for some } 0 < \tau < 1$$

Stable spline kernels (Pillonetto & De Nicolao, 2010)[2]

An exponential transform:

$$i \to \lambda^i$$

for some $0 < \lambda < 1$

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

[1] G. Wahba. Spline Models for Observational Data. New York: SIAM, 1990.

[2] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93, 2010.

The optimal kernel

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

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

Prior

$$g_k^0 \sim O(\tau^k) \text{ for some } 0 < \tau < 1$$

Heuristic methods

DI kernel

$$K(\eta) = c\,\mathrm{diag}([\lambda, \cdots, \lambda^n])$$
$$\eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}$$

Heuristic methods

DI kernel

$$K(\eta) = c\,\mathrm{diag}([\lambda, \cdots, \lambda^n])$$
$$\eta = [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}$$

DC kernel

$$K_{i,j}(\eta) = c\lambda^{(i+j)/2}\rho^{|i-j|}$$
$$K(\eta) = c\begin{bmatrix} \lambda & \lambda^{\frac{3}{2}}\rho & \cdots & \lambda^{\frac{n+1}{2}}\rho^{n-1} \\ \lambda^{\frac{3}{2}}\rho & \lambda^2 & \cdots & \lambda^{\frac{n+2}{2}}\rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \lambda^{\frac{n+1}{2}}\rho^{n-1} & \lambda^{\frac{n+2}{2}}\rho^{n-2} & \cdots & \lambda^n \end{bmatrix}$$

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 $\rho = \sqrt{\lambda}$.

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

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

[1] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes–Revisited. *Automatica, 48(8)*: 1525–1535, 2012.

Heuristic methods

Multiple kernels (Chen et al., 2014)[1]

Better capture complicated dynamics of the system

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

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

[1] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control, 59(11)*: 2933–2945, 2014.

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)

[1] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica, 50(3)*: 657–682, 2014.

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)

$$\mathrm{EB} : \widehat{\eta}_{\mathrm{EB}} = \arg\min_{\eta \in \Omega} Y^T Q^{-1} Y + \log\det(Q)$$

Stein's unbiased risk estimator (SURE)

MSE (for prediction ability):

$$\text{MSE}(K) = E\|\Phi(\widehat{\theta}^{\mathrm{R}} - \theta_0)\|^2$$

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

SURE method

- To construct an unbiased estimators of the MSE

$$\mathscr{F}_{\text{SURE}}(K) = \|Y - \Phi\widehat{\theta}^{\mathrm{R}}\|^2 + 2\sigma^2 \mathrm{Tr}(R^{-1}\Phi^T\Phi)$$
$$R = \Phi^T\Phi + \sigma^2 K^{-1}$$

- To estimate the hyperparameter η by

$$\text{SURE}: \ \widehat{\eta}_{\text{SURE}} = \arg\min_{\eta\in\Omega} \mathscr{F}_{\text{SURE}}(K(\eta))$$

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

$$|s| = k, \ \{1, \cdots, N\} = s \cup s^c$$

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

$$\mathrm{APE}_s = \frac{1}{k} \sum_{t \in s} \left(y(t) - \phi(t)^T \widehat{\theta}_{s^c} \right)^2 = \frac{1}{k} \| Y_s - \Phi_s \widehat{\theta}_{s^c} \|^2$$

Advantage: does not require to estimate the noise variance σ^2

Variants of CVs

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

$$\widehat{\eta}_{\text{LKOCV}} = \arg\min_{\eta\in\Omega} \frac{1}{\binom{N}{k}} \sum_{s} \mathrm{APE}_s \text{ (all choices)}$$

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

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

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

3. Generalized cross validation (GCV)

$$\widehat{\eta}_{\text{GCV}} = \arg\min_{\eta\in\Omega} \frac{1}{N} \frac{\sum_{t=1}^{N} \left(y(t) - \widehat{y}(t)\right)^2}{\left(1 - \mathrm{Tr}(H)/N\right)^2}$$

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]

$$\eta^* = \arg\min_{\eta\in\Omega} \theta_0^T K^{-1}\Sigma^{-1}K^{-1}\theta_0 - 2\mathrm{Tr}\left(\Sigma^{-1}K^{-1}\right)$$

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

[1] B. Mu, T. Chen and L. Ljung. On Asymptotic Properties of Hyperparameter Estimators for Kernel-based Regularization Methods. Automatica, 94: 381–395, 2018.

Asymptotically theoretical properties

Theorem

- $\widehat{\eta}_{\text{SURE}} \to \eta^*$
  $\widehat{\eta}_{\text{EB}} \to \arg\min_{\eta \in \Omega} \theta_0^T K^{-1} \theta_0 + \log\det(K)$ (Mu et al., 2018c) [1]
- $\widehat{\eta}_{\text{GCV}} \to \eta^*$ (Mu et al., 2018a) [2]
  $\widehat{\eta}_{\text{LOOCV}} \to \eta^*$ if the input is bounded
  $\widehat{\eta}_{\text{LKOCV}} \to \eta^*$ if $k/N \to 0$ and the input is bounded
  (Mu et al., 2018b) [3]

[1] B. Mu, T. Chen and L. Ljung. On Asymptotic Properties of Hyperparameter Estimators for Kernel-based Regularization Methods. Automatica, 94: 381–395, 2018.

[2] B. Mu, T. Chen and L. Ljung. Asymptotic Properties of Generalized Cross Validation Estimators for Regularized System Identification. Proceedings of the IFAC Symposium on System Identification, 203–205, 2018.

[3] B. Mu, T. Chen and L. Ljung. Asymptotic Properties of Hyperparameter Estimators by Using Cross-Validations for Regularized System Identification. Proceedings of the IEEE Conference on Decision and Control, 644–649, 2018.

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)

Results

Table 1: Average fits for 1000 test systems.

Inputs	Sizes	EB	LOOCV	GCV	SURE	MSE
IT1	500	86.16	86.24	86.24	86.03	87.02
	8000	96.44	96.60	96.60	96.60	96.67
IT2	500	39.03	-85.95	-84.84	-146.4	41.94
	8000	50.86	38.79	38.89	38.86	53.63
IT3	500		69.33	89.55	89.52	89.95
	8000		81.42	96.64	96.64	96.70

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?

References

Chen, T., Andersen, M. S., Ljung, L., Chiuso, A., & Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, *59*, 2933–2945.

Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and gaussian processes–revisited. *Automatica*, *48*, 1525–1535.

Mu, B., Chen, T., & Ljung, L. (2018a). Asymptotic properties of generalized cross validation estimators for regularized system identification. In *Proceedings of the IFAC Symposium on System Identification* (pp. 203–205). Stockholm, Sweden.

Mu, B., Chen, T., & Ljung, L. (2018b). Asymptotic properties of hyperparameter estimators by using cross-validations for regularized system identification. In *Proceedings of the 57th IEEE Conference on Decision and Control* (pp. 644–649).

Mu, B., Chen, T., & Ljung, L. (2018c). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica*, *94*, 381–395.

Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, *50*, 657–682.

Sjöberg, J., McKelvey, T., & Ljung, L. (1993). On the use of regularization in system identification. In *Proceedings of the 12th IFAC World Congress* (pp. 381–386). Sydney, Australia.

Tikhonov, A. N., & Arsenic, V. Y. (1977). *Solutions of Ill-Posed Problems*. New York: John Wiley.

Wahba, G. (1990). *Spline Models for Observational Data*. New York: SIAM.