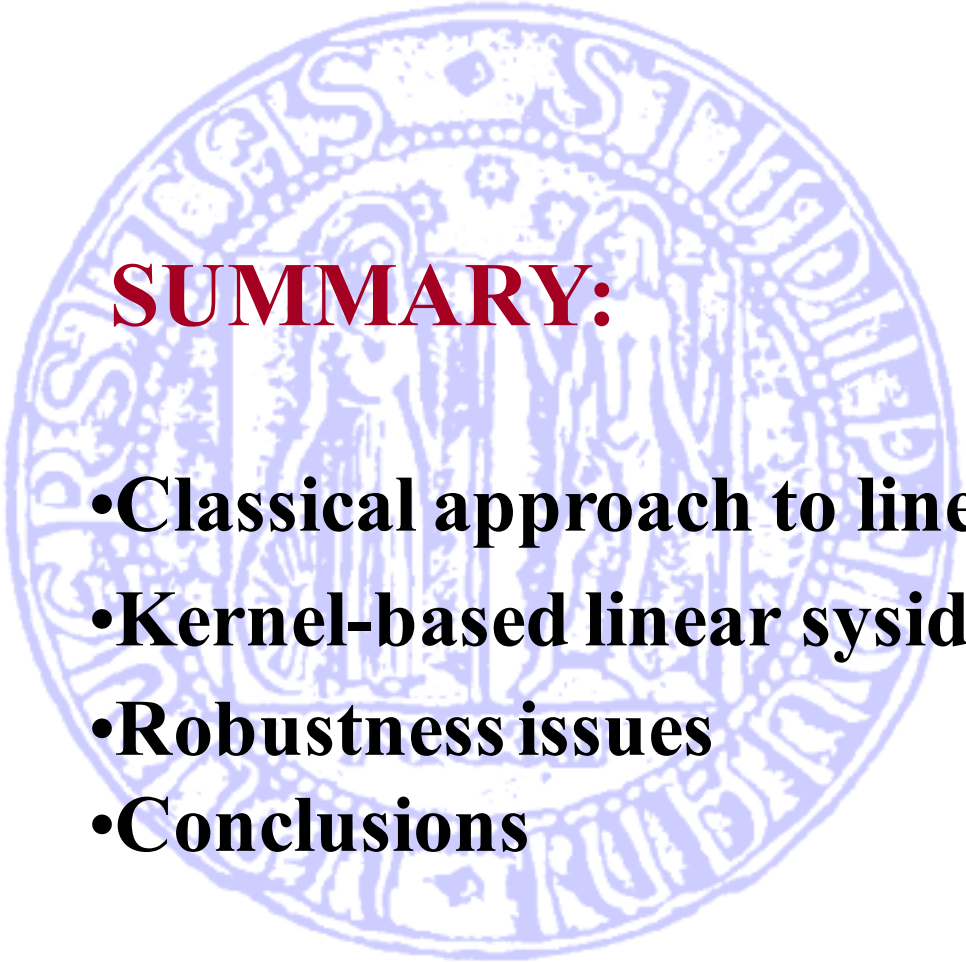




REGULARIZED KERNEL-BASED APPROACHES TO SYSTEM IDENTIFICATION

Gianluigi Pillonetto

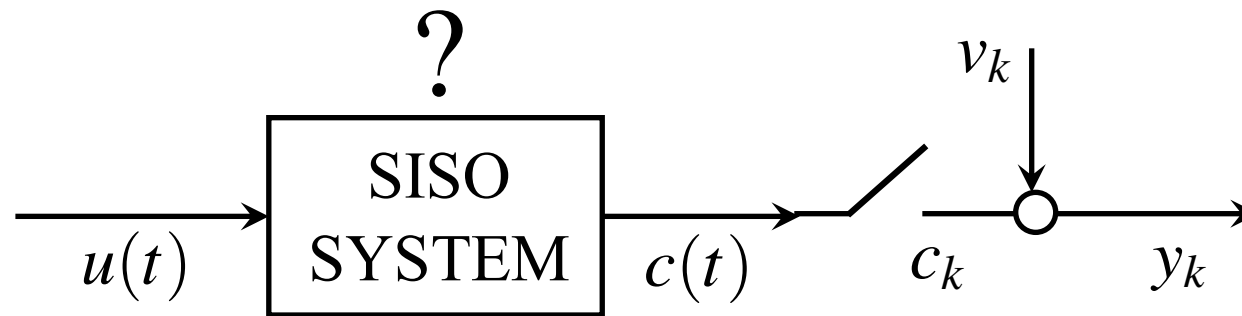
***Department of Information Engineering
University of Padova, Italy***



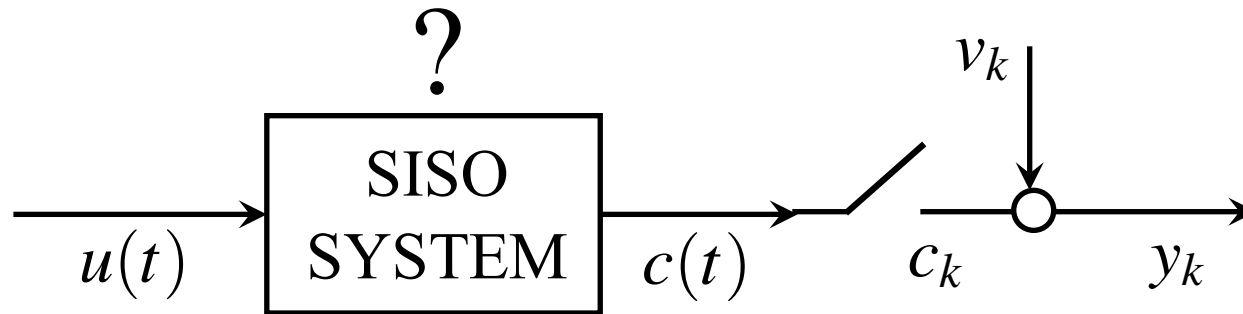
SUMMARY:

- **Classical approach to linear sysid**
- **Kernel-based linear sysid**
- **Robustness issues**
- **Conclusions**

Problem statement



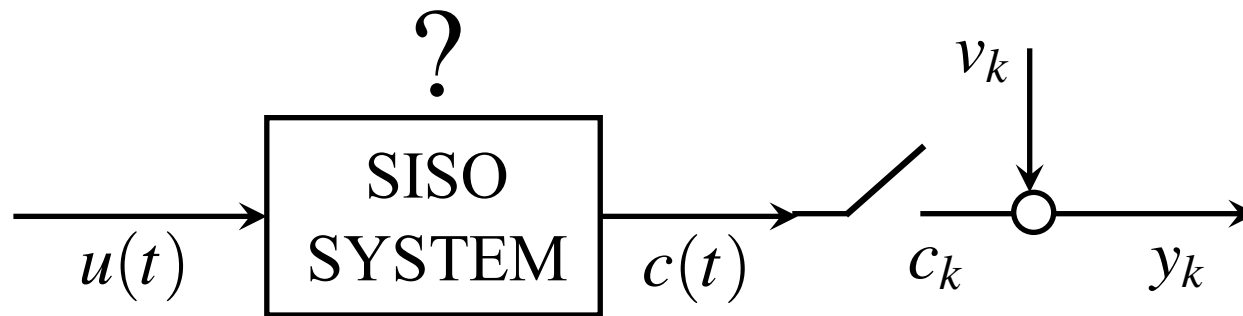
Problem statement



Assumptions:

- system is linear, time-invariant and BIBO-stable
- noise is white and Gaussian, independent of u

Problem statement



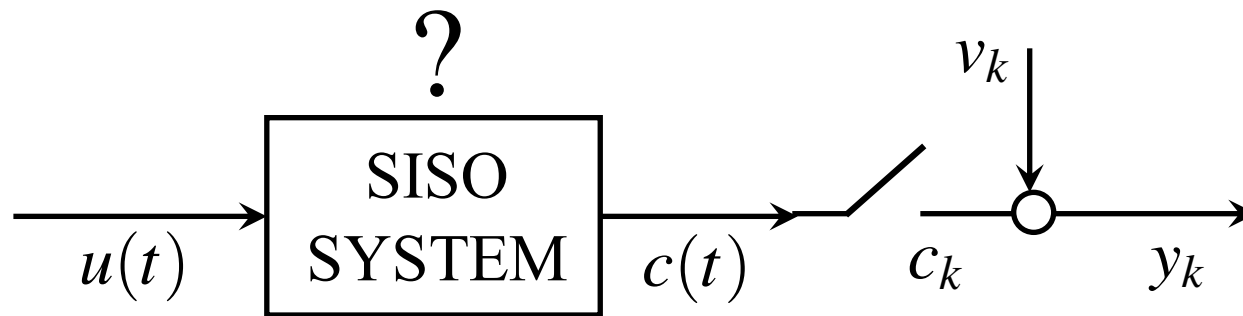
Assumptions:

- system is linear, time-invariant and BIBO-stable
- noise is white and Gaussian, independent of u

$$y_k = (g \otimes u)_k + v_k \quad \otimes = \text{convolution}$$

$$v_k \sim \mathcal{N}(0, \sigma^2) \quad k = 1, \dots, n$$

Problem statement

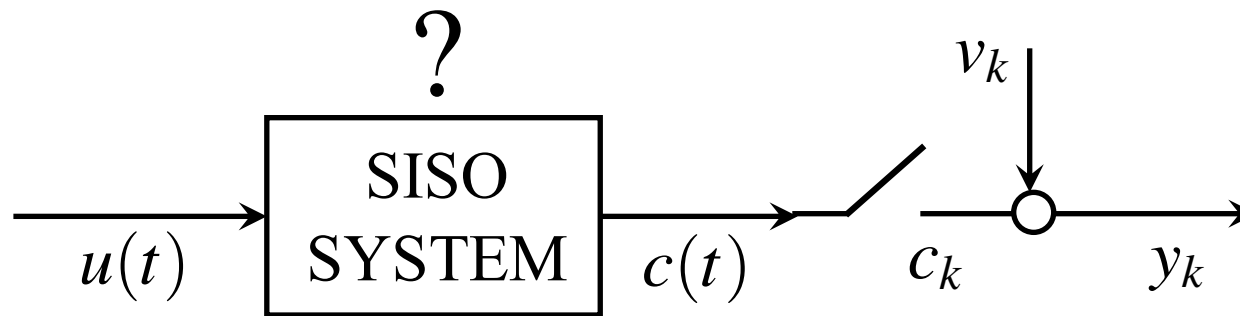


Assumptions:

- system is linear, time-invariant and BIBO-stable
- noise is white and Gaussian, independent of u

$$y_k = \overset{\text{impulse response}}{(g \otimes u)_k} + v_k \quad \otimes = \text{convolution}$$
$$v_k \sim \mathcal{N}(0, \sigma^2) \quad k = 1, \dots, n$$

Problem statement



Assumptions:

- system is linear, time-invariant and BIBO-stable
- noise is white and Gaussian, independent of u

$$y_k = \overset{\text{impulse response}}{(g \otimes u)_k} + v_k \quad \otimes = \text{convolution}$$
$$v_k \sim \mathcal{N}(0, \sigma^2) \quad k = 1, \dots, n$$

Problem: to estimate g from
observed input-output (u, y)

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

L. Ljung System identification, Prentice Hall, 1999
T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989
B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

Model $\mathcal{M}(\theta)$ and associated predictor $\hat{y}_\theta(t)$

L. Ljung System identification, Prentice Hall, 1999

T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989

B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

Model $\mathcal{M}(\theta)$ and associated predictor $\hat{y}_\theta(t)$

Impulse response models

$$\hat{y}_\theta(t) = (g_\theta \otimes u)_t$$

L. Ljung System identification, Prentice Hall, 1999

T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989

B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

Model $\mathcal{M}(\theta)$ and associated predictor $\hat{y}_\theta(t)$

Impulse response models

$$\hat{y}_\theta(t) = (\underbrace{g_\theta}_{\substack{\text{parametrized} \\ \text{impulse response}}} \otimes u)_t$$

L. Ljung System identification, Prentice Hall, 1999

T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989

B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

Model $\mathcal{M}(\theta)$ and associated predictor $\hat{y}_\theta(t)$

Impulse response models

$$\hat{y}_\theta(t) = (g_\theta \otimes u)_t$$

- $g_\theta(t) = \sum_{i=1}^d \theta_i \phi_i(t)$

Linear models using basis functions
(e.g. FIR, Laguerre polynomials)

L. Ljung System identification, Prentice Hall, 1999

T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989

B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

Classical approach: models

Let

$\mathcal{M}$: model structure

$\theta \in \mathbb{R}^d$: parameters

Model $\mathcal{M}(\theta)$ and associated predictor $\hat{y}_\theta(t)$

Impulse response models

$$\hat{y}_\theta(t) = (g_\theta \otimes u)_t$$

- $g_\theta(t) = \sum_{i=1}^d \theta_i \phi_i(t)$

Linear models using basis functions
(e.g. FIR, Laguerre polynomials)

- $G_\theta(s) = \frac{s^a + \theta_1 s^{a-1} + \dots + \theta_a}{s^a + \theta_{a+1} s^{a-1} + \dots + \theta_d}$

Rational transfer functions
(Laplace domain)

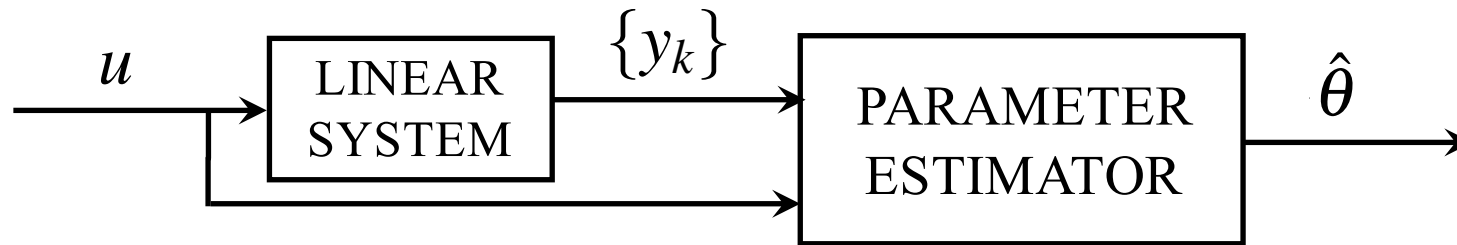
L. Ljung System identification, Prentice Hall, 1999

T. Soderstrom and P. Stoica System identification, Prentice Hall, 1989

B. Wahlberg System identification using Laguerre models, IEEE TAC, 1991

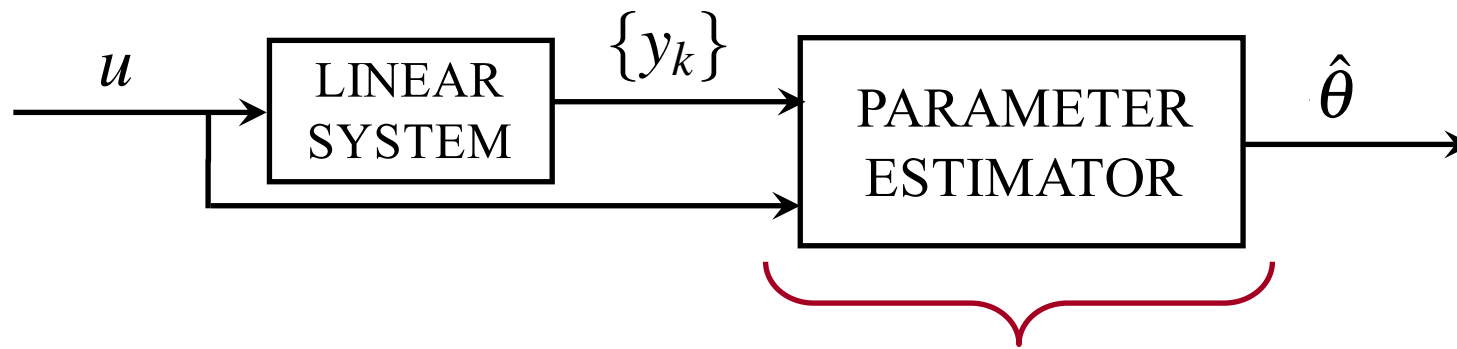
Classical approach: estimation

A simple case:
known model structure and order d



Classical approach: estimation

A simple case:
known model structure and order d

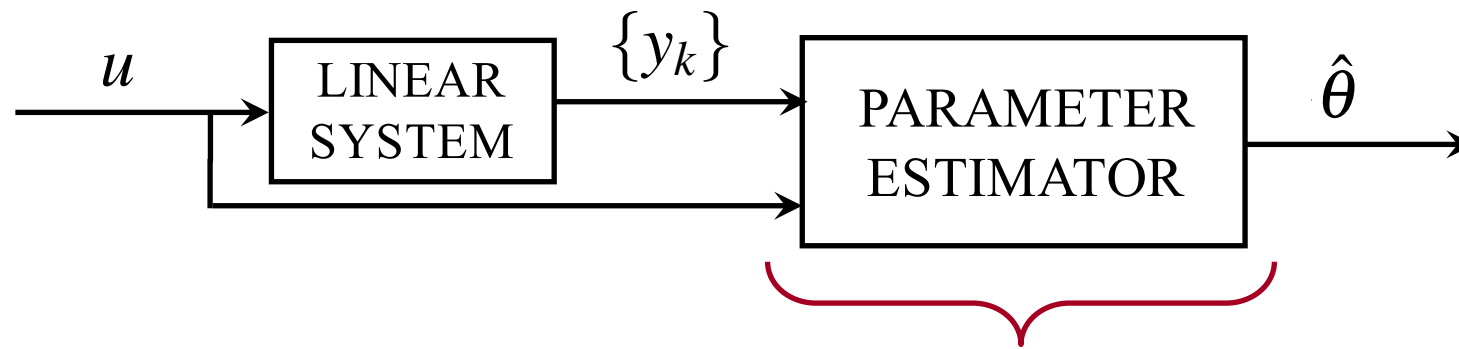


Prediction error (PEM),
maximum likelihood (ML)

$$\hat{\theta} = \arg \min_{\theta} \sum_k (y_k - \hat{y}_{\theta}(t_k))^2$$

Classical approach: estimation

A simple case:
known model structure and order d



Prediction error (PEM),
maximum likelihood (ML)

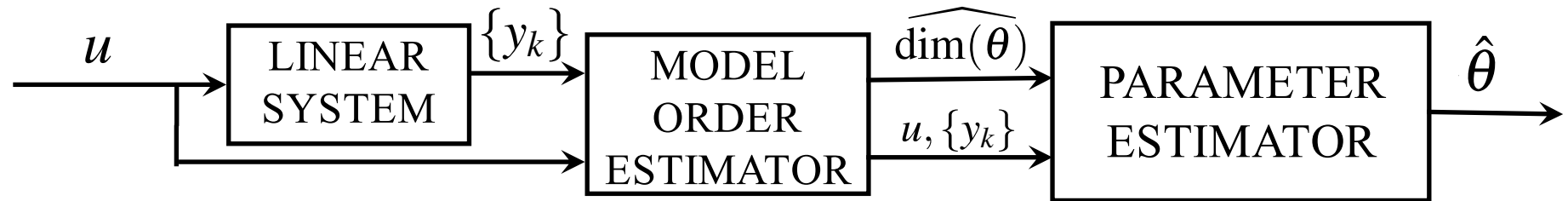
$$\hat{\theta} = \arg \min_{\theta} \sum_k (y_k - \hat{y}_{\theta}(t_k))^2$$

PEM asymptotic properties:

convergence to the true model and
smallest variance for unbiased estimators (Cramér-Rao)

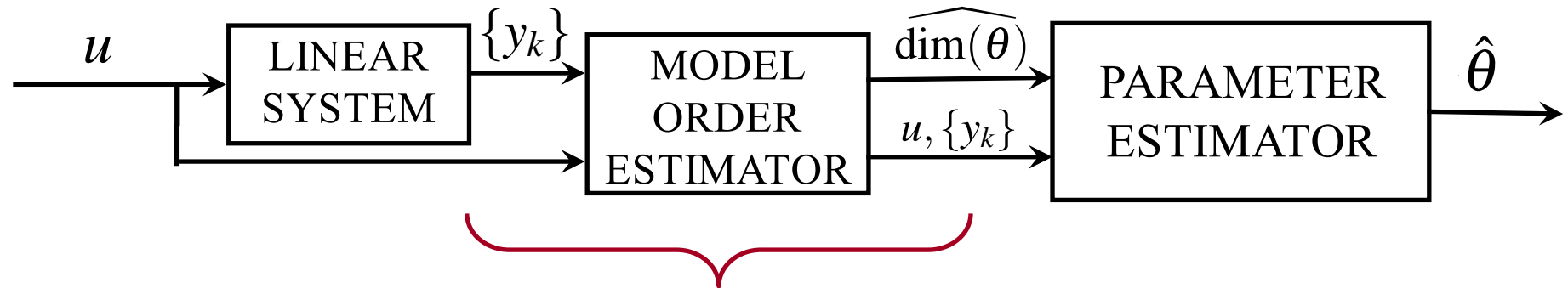
Classical approach: estimation

More difficult case:
unknown model order



Classical approach: estimation

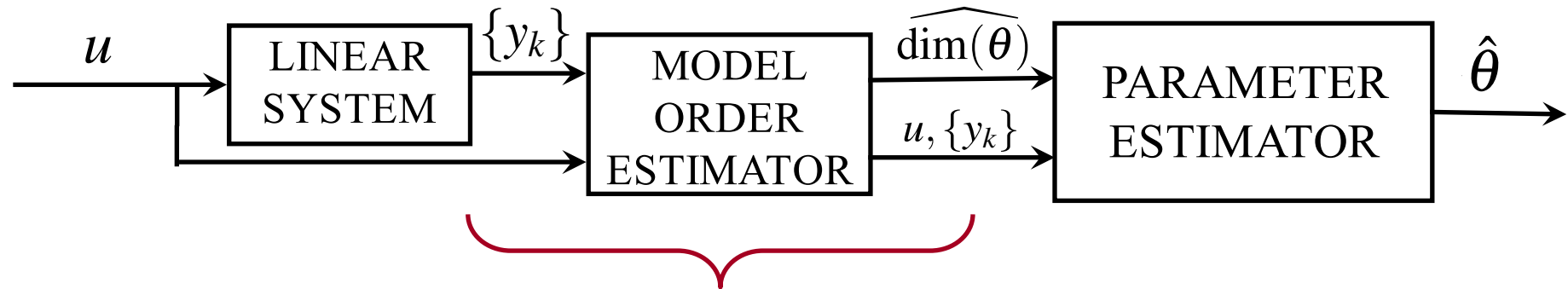
More difficult case:
unknown model order



Key issue: bias-variance trade-off
(flexible structure but avoid fitting noise to data)

Classical approach: estimation

More difficult case:
unknown model order

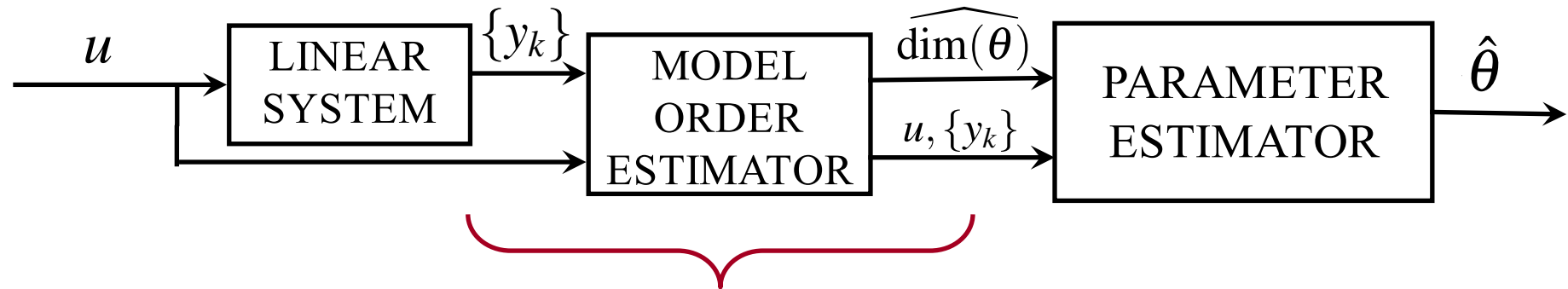


Key issue: bias-variance trade-off
(flexible structure but avoid fitting noise to data)

Leading criteria: Cross-validation techniques,
Schwarz (BIC), Akaike (AIC)

Classical approach: estimation

More difficult case:
unknown model order



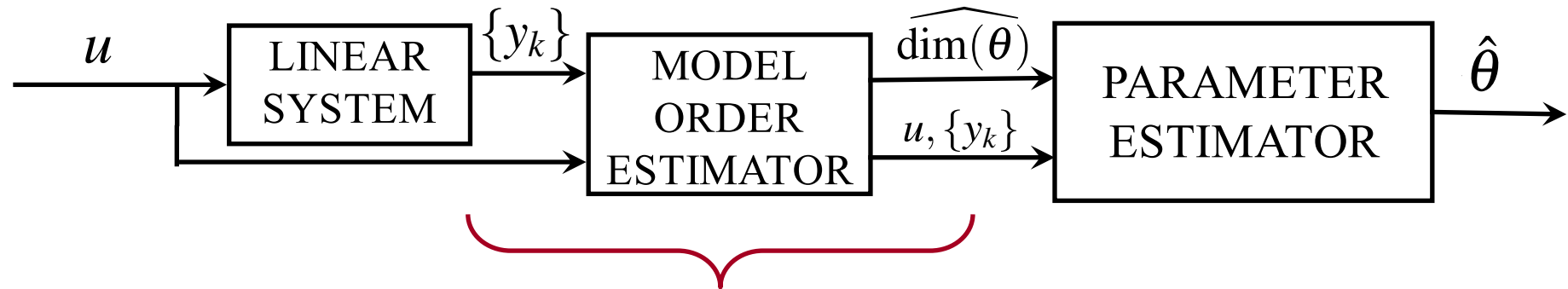
Key issue: bias-variance trade-off
(flexible structure but avoid fitting noise to data)

Leading criteria: Cross-validation techniques,
Schwarz (BIC), Akaike (AIC)

$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

Classical approach: estimation

More difficult case:
unknown model order



Key issue: bias-variance trade-off
(flexible structure but avoid fitting noise to data)

Leading criteria: Cross-validation techniques,
Schwarz (BIC), Akaike (AIC)

$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + \underbrace{J(n) \dim \theta}_{\text{Penalty term to balance overfit}}$$

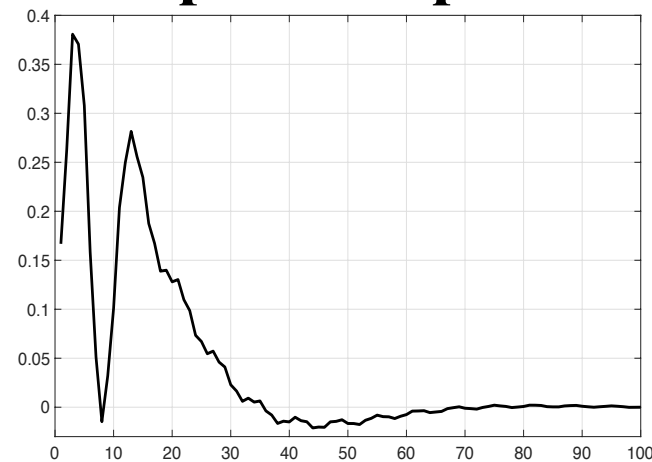
$$\text{AIC: } J(n) = 2$$

$$\text{BIC: } J(n) = \log n$$

Numerical example – discrete time system



Impulse response

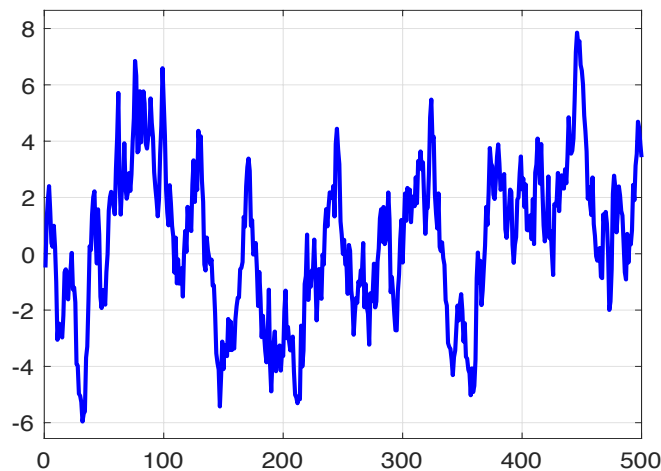


A randomly generated
6th order rational
transfer function

Numerical example – discrete time system

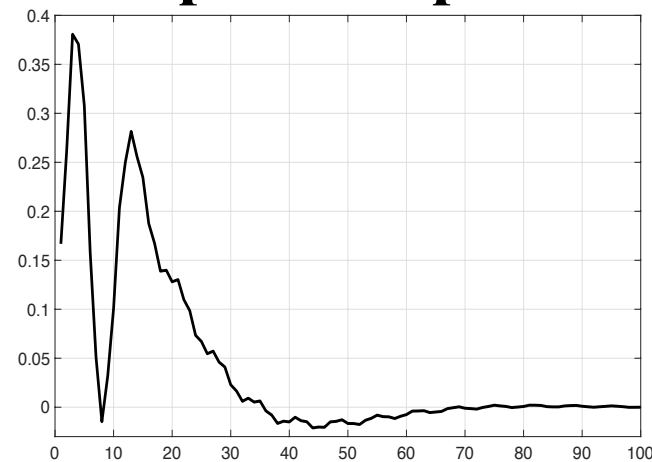


Input



Low-pass input
(quite typical in
real applications)

Impulse response

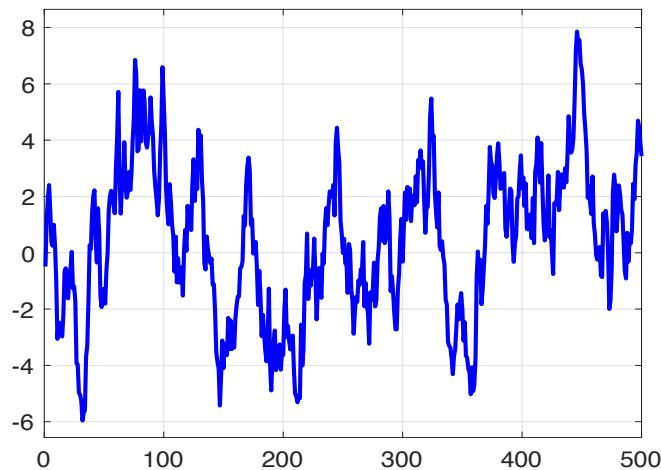


A randomly generated
6th order rational
transfer function

Numerical example – discrete time system

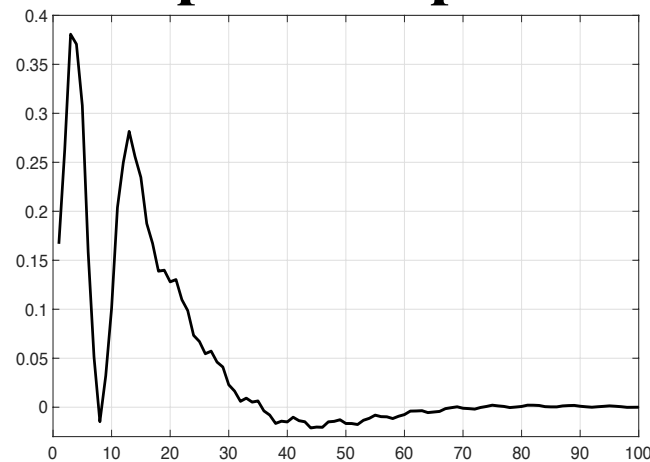


Input



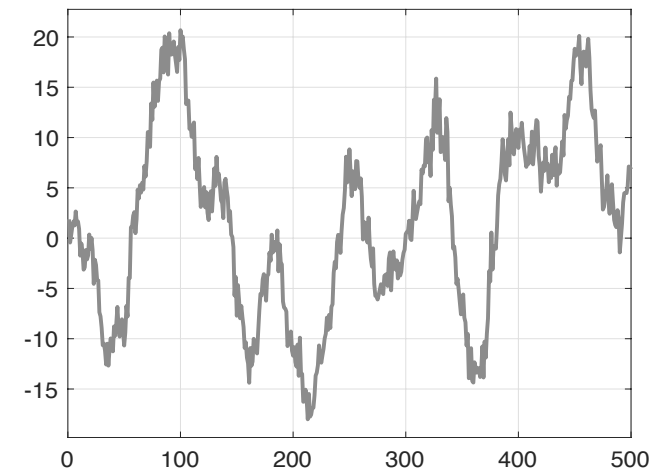
Low-pass input
(quite typical in
real applications)

Impulse response



A randomly generated
6th order rational
transfer function

Output

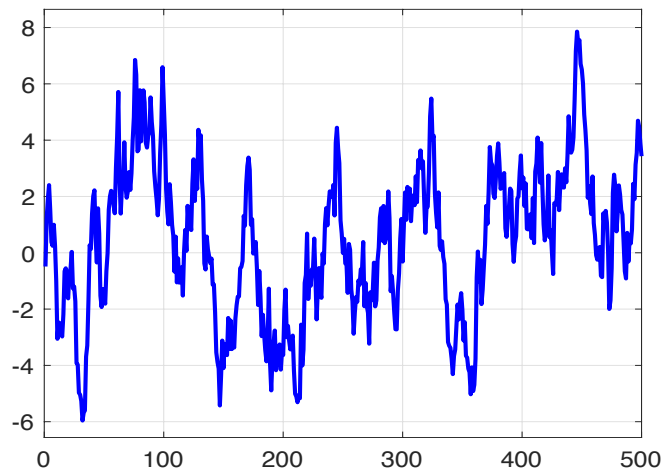


500 output
measurements
(SNR=50)

Numerical example – discrete time system

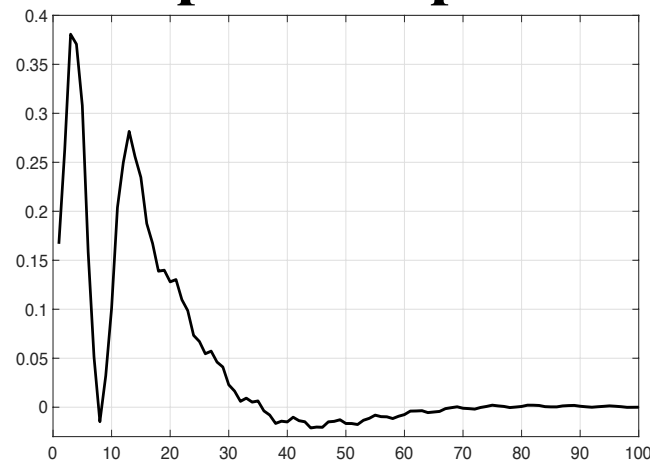


Input



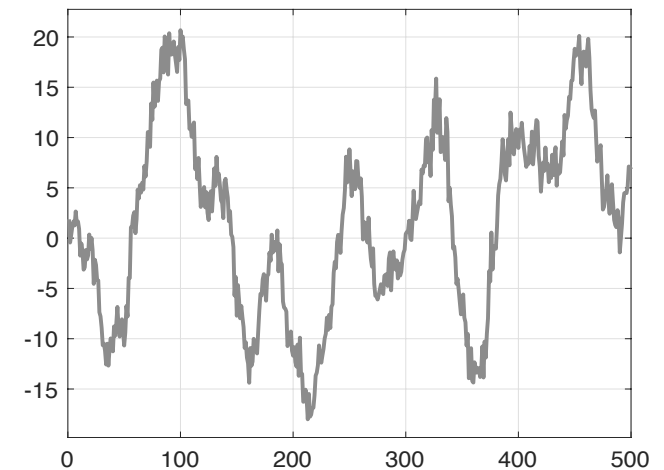
Low-pass input
(quite typical in
real applications)

Impulse response



A randomly generated
6th order rational
transfer function

Output



500 output
measurements
(SNR=50)

Let us test the classical approach using PEM
with rational transfer functions as structures

Ideal model order choice: PEM+Oracle

FIT measure

$$\text{FIT}(d) = 100\% \left(1 - \frac{\sqrt{\sum_{k=1}^{100} (g_k - \hat{g}_k(d))^2}}{\sqrt{\sum_{k=1}^{100} g_k^2}} \right)$$

Ideal model order choice: PEM+Oracle

FIT measure

$$\text{FIT}(d) = 100\% \left(1 - \frac{\sqrt{\sum_{k=1}^{100} \overset{\text{True}}{\textcircled{g_k}} - \hat{g}_k(d))^2}}{\sqrt{\sum_{k=1}^{100} g_k^2}} \right)$$

Ideal model order choice: PEM+Oracle

FIT measure

$$\text{FIT}(d) = 100\% \left(1 - \frac{\sqrt{\sum_{k=1}^{100} \overset{\text{True}}{g_k} - \overset{\text{PEM estimates}}{\hat{g}_k(d)} }^2}{\sqrt{\sum_{k=1}^{100} g_k^2}} \right)$$

Ideal model order choice: PEM+Oracle

Oracle choice

$$\text{FIT}(d) = 100\% \left(1 - \frac{\sqrt{\sum_{k=1}^{100} (g_k - \hat{g}_k(d))^2}}{\sqrt{\sum_{k=1}^{100} g_k^2}} \right)$$

$$\hat{d} = \arg \max_d \text{FIT}(d)$$

Ideal model order choice: PEM+Oracle

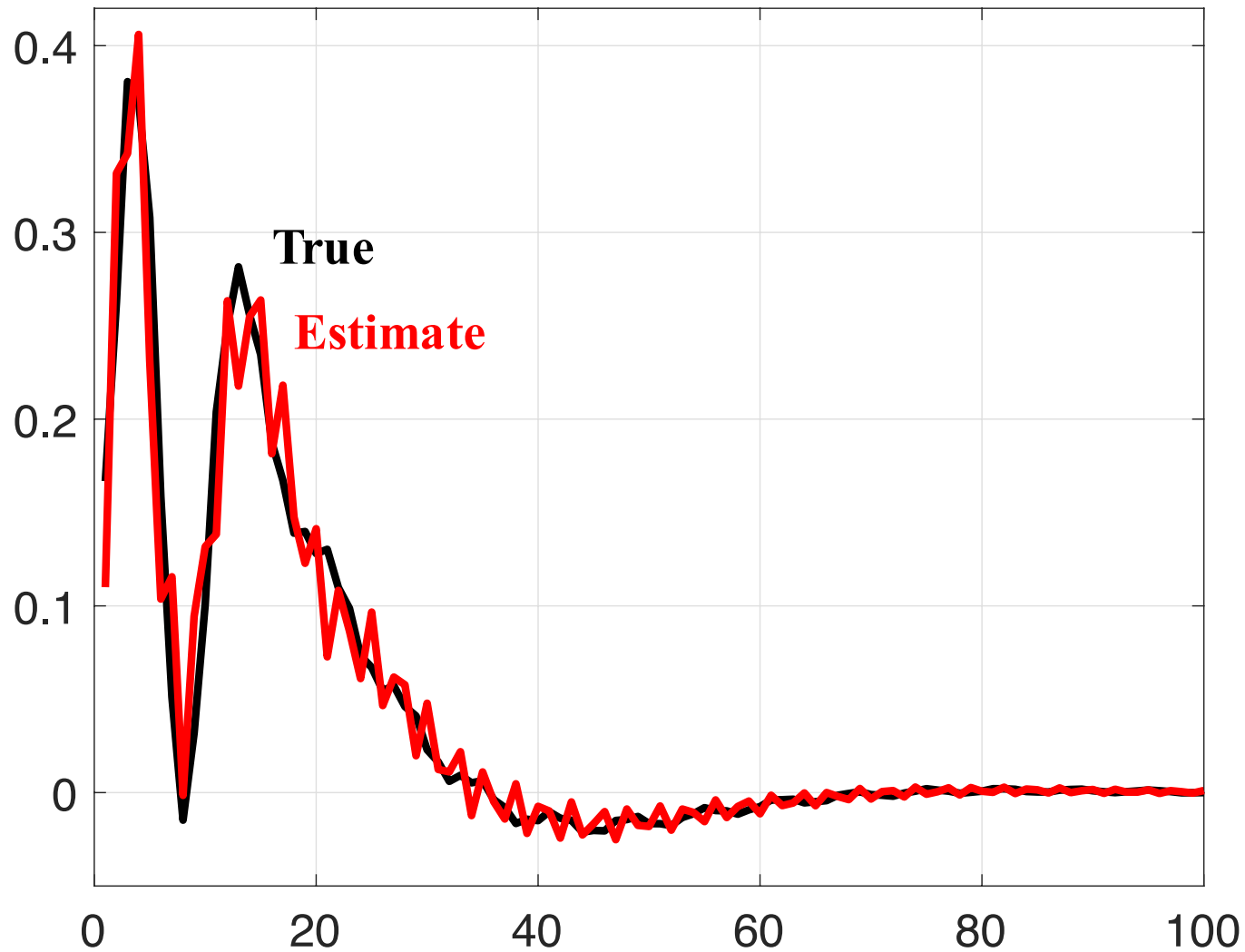
Oracle choice

$$\text{FIT}(d) = 100\% \left(1 - \frac{\sqrt{\sum_{k=1}^{100} (g_k - \hat{g}_k(d))^2}}{\sqrt{\sum_{k=1}^{100} g_k^2}} \right)$$
$$\hat{d} = \arg \max_d \text{FIT}(d)$$

In our experiment, the oracle maximizes the FIT
using the estimated rational transfer
functions up to order 15

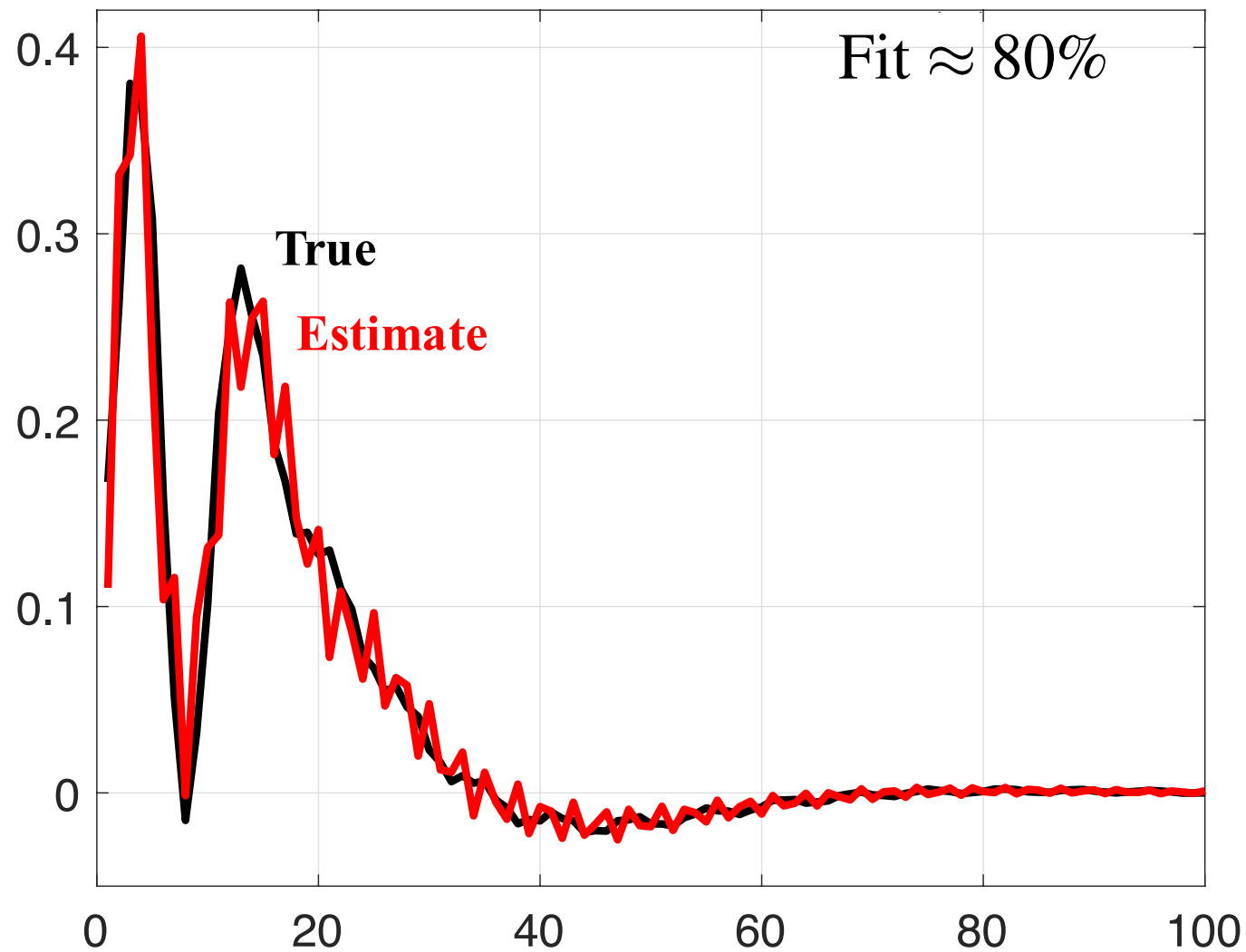
Ideal model order choice: PEM+Oracle

Oracle selection: 5th order system



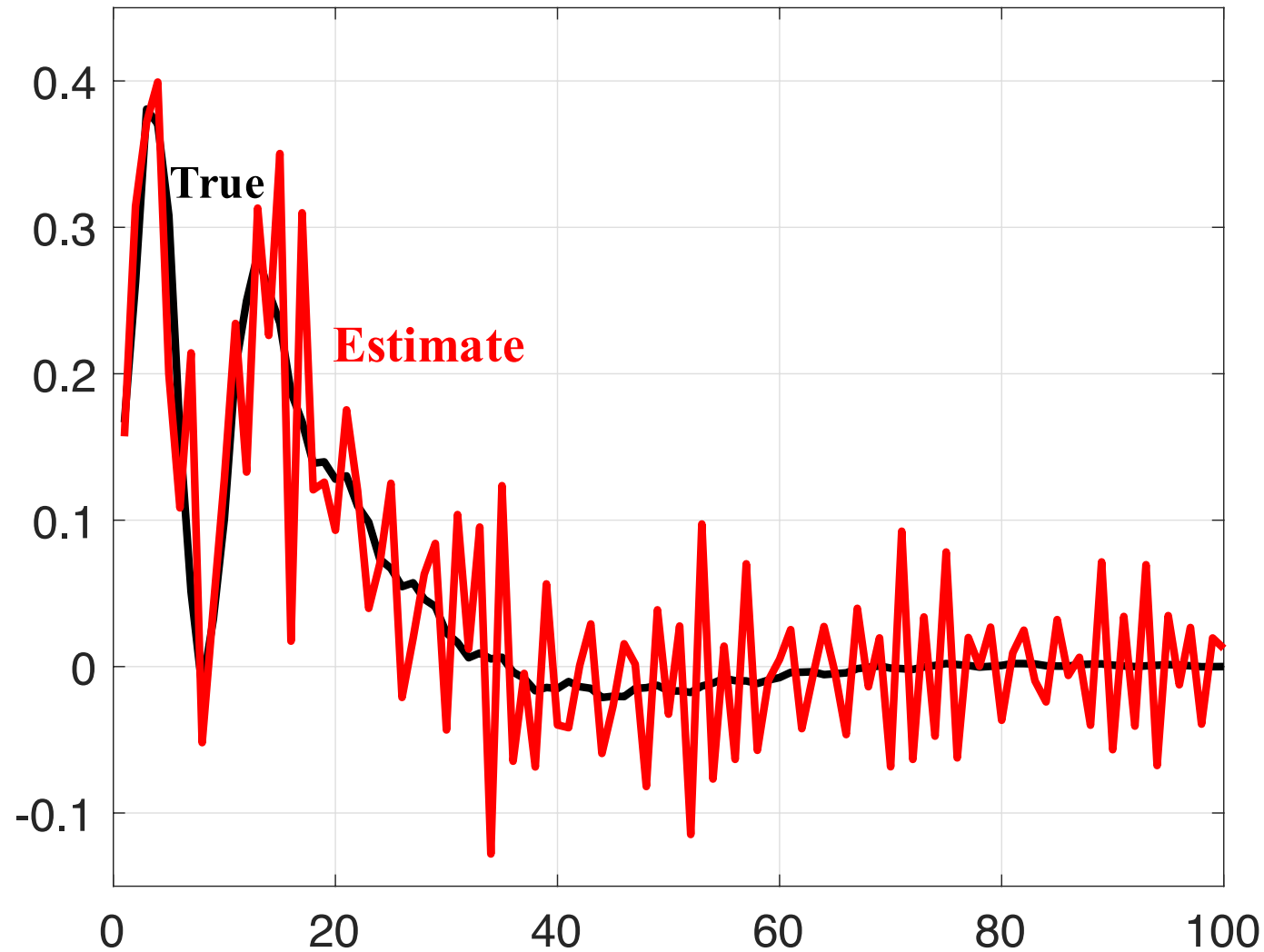
Ideal model order choice: PEM+Oracle

Oracle selection: 5th order system



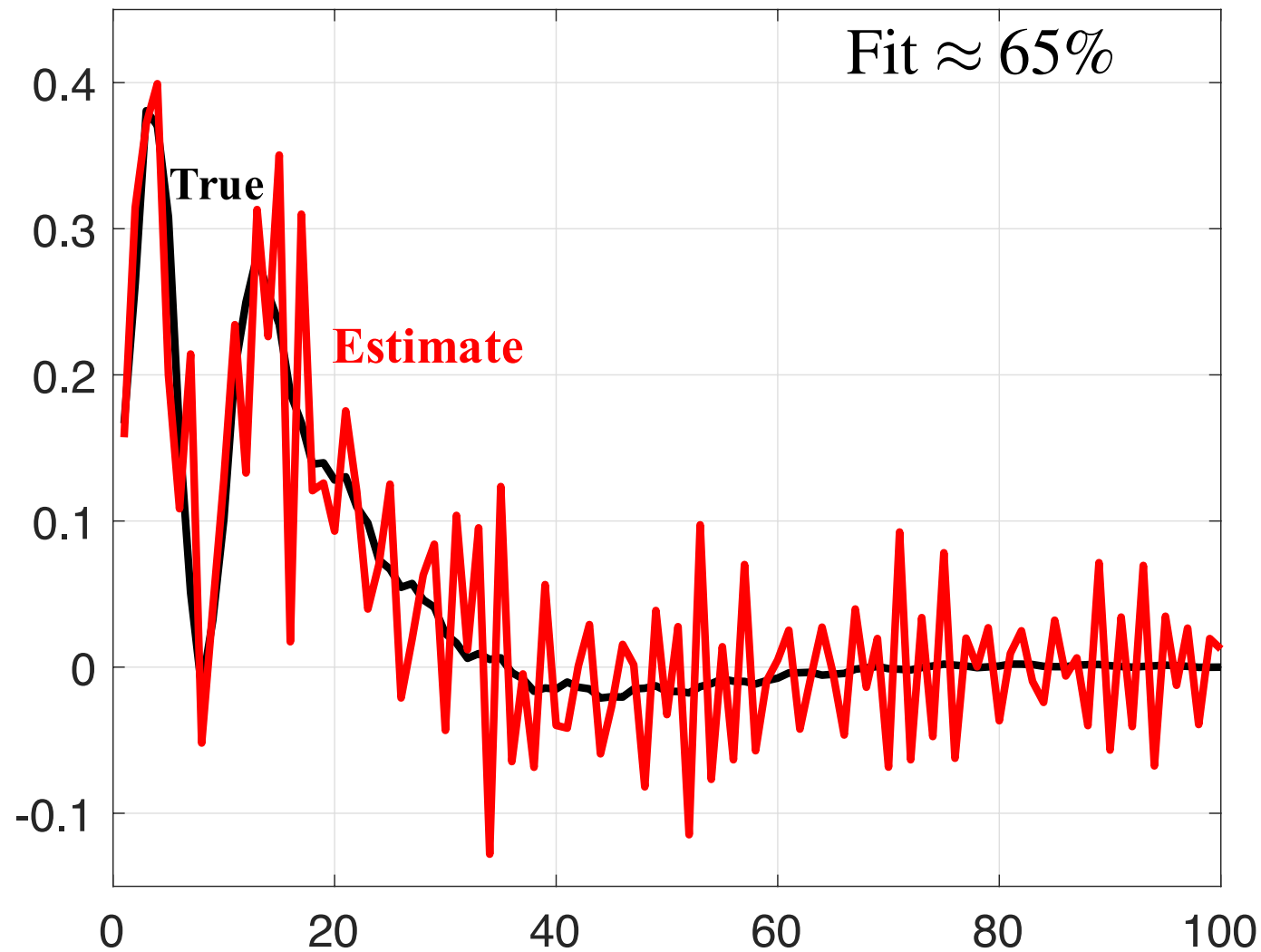
Realistic model order choice: PEM+AIC

AIC selection: 9th order system



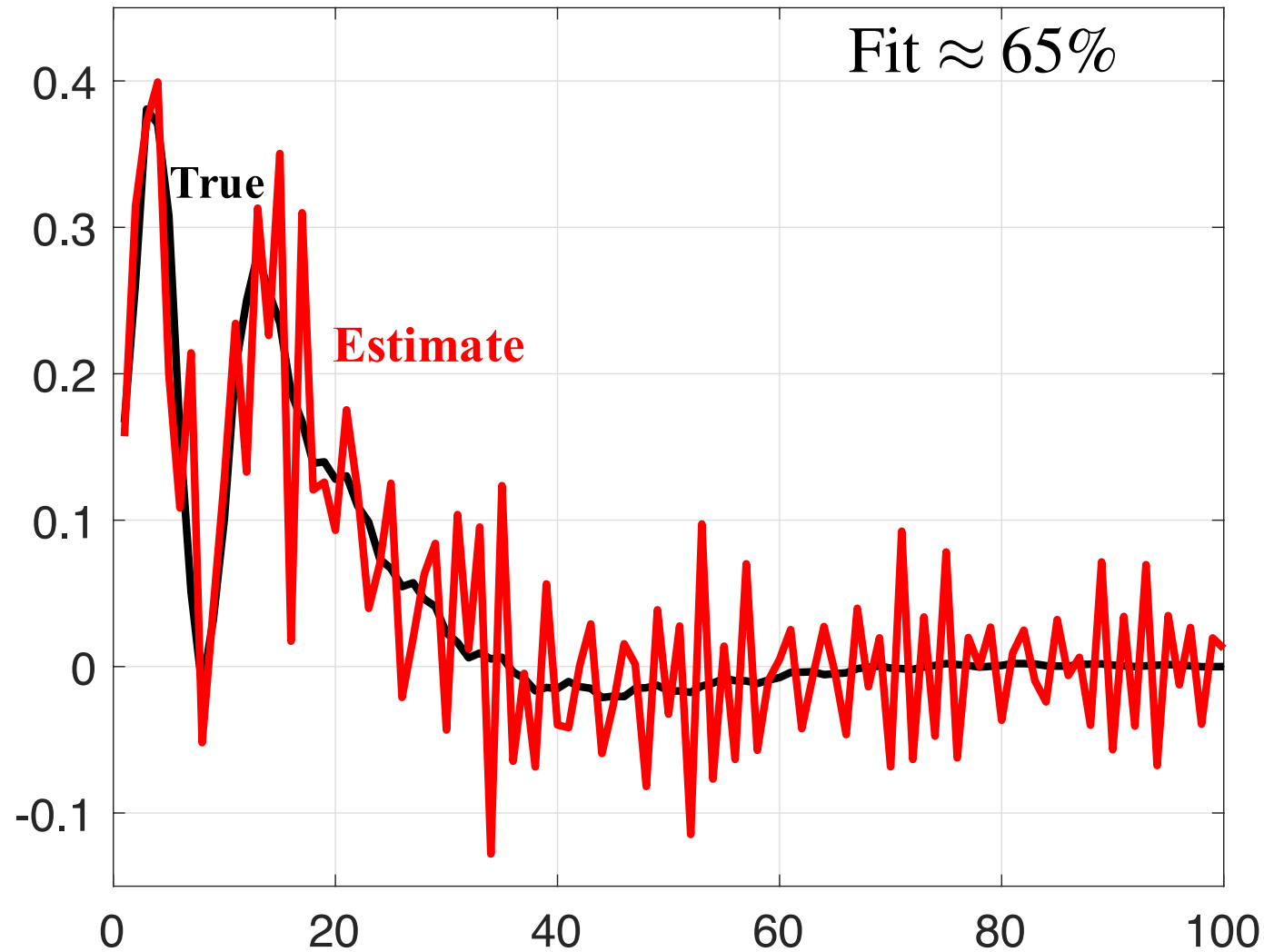
Realistic model order choice: PEM+AIC

AIC selection: 9th order system



Realistic model order choice: PEM+AIC

AIC selection: 9th order system



BIC and cross-validation (7th order selected) give similar results

Comment on numerical experiment

A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

Comment on numerical experiment

A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

- Non-convex estimator possibly subject to local minima.

Comment on numerical experiment

A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

- Non-convex estimator possibly subject to local minima.
- AIC/BIC criteria rely on asymptotic arguments irrelevant for small data sizes.

Comment on numerical experiment

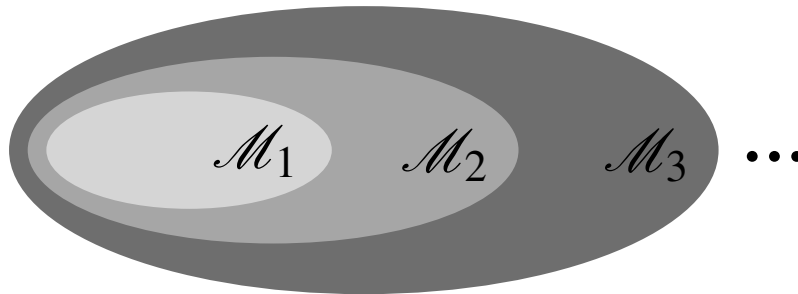
A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

- Non-convex estimator possibly subject to local minima.
- AIC/BIC criteria rely on asymptotic arguments irrelevant for small data sizes.
Typical system order choices: 5 for Oracle, 10 for AIC (small bias but large variance).

Comment on numerical experiment

A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

- Non-convex estimator possibly subject to local minima.
- AIC/BIC criteria rely on asymptotic arguments irrelevant for small data sizes.
Typical system order choices: 5 for Oracle, 10 for AIC (small bias but large variance).



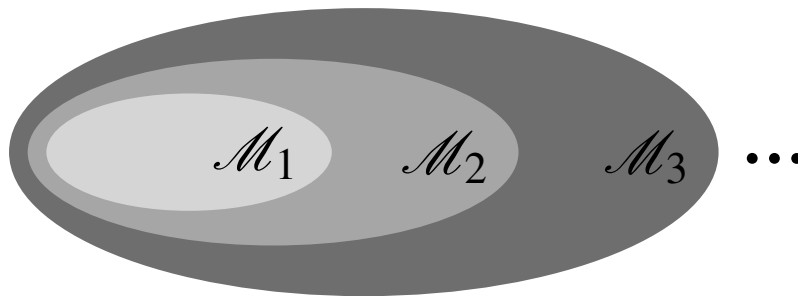
Classical approach:
complexity controlled by
discrete model order

$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

Comment on numerical experiment

A Monte Carlo confirms the result:
in this case AIC/BIC are significantly outperformed by the oracle.
What are the main problems?

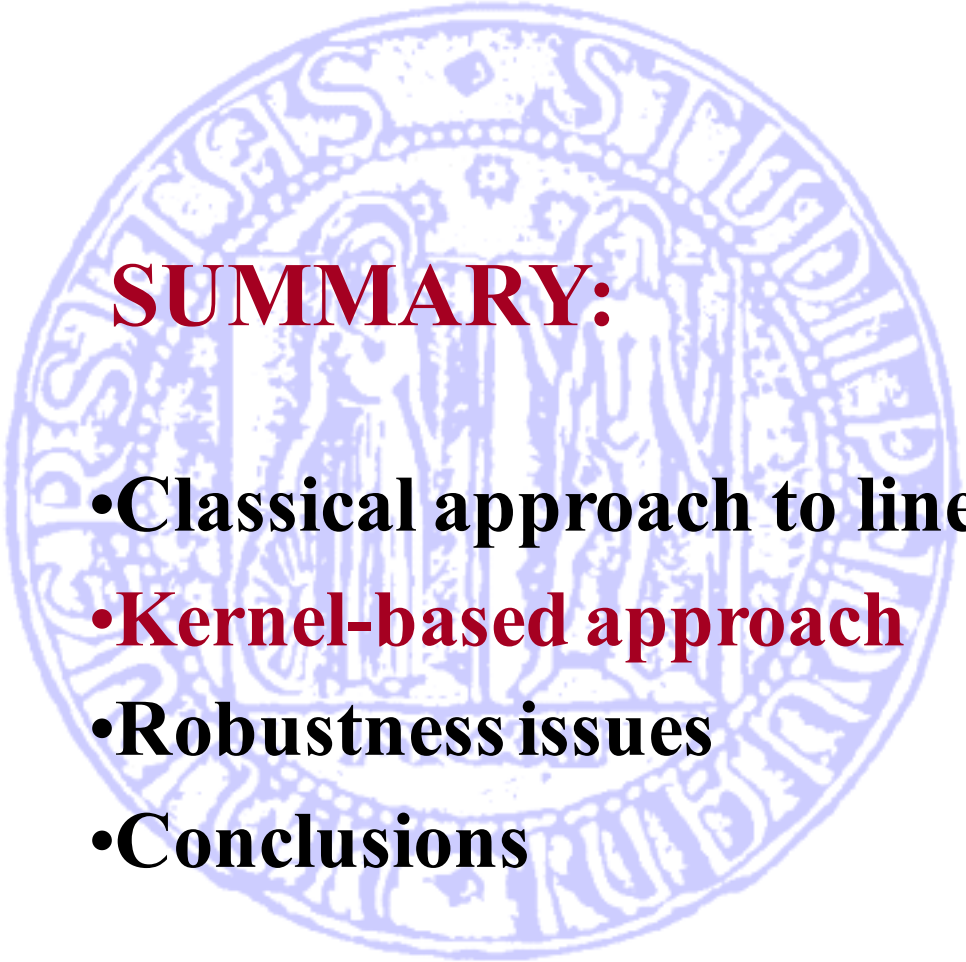
- Non-convex estimator possibly subject to local minima.
- AIC/BIC criteria rely on asymptotic arguments irrelevant for small data sizes.
Typical system order choices: 5 for Oracle, 10 for AIC (small bias but large variance).



Classical approach:
complexity controlled by
discrete model order

Bias-variance:
can we do better?

$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$



SUMMARY:

- **Classical approach to linear sysid**
- **Kernel-based approach**
- **Robustness issues**
- **Conclusions**

FIR models and least squares

Linear model

FIR models of order d

FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Linear least squares

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2$$

FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

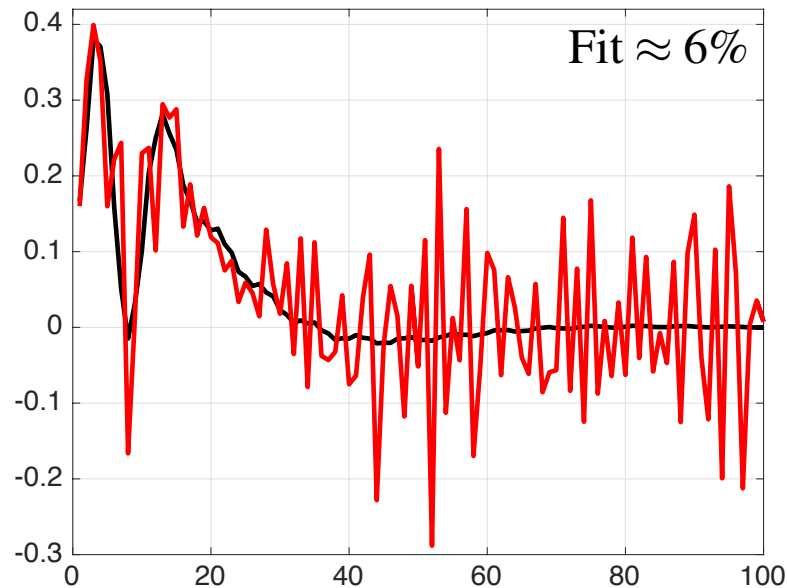
$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Linear least squares

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2$$

$d=100$



FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

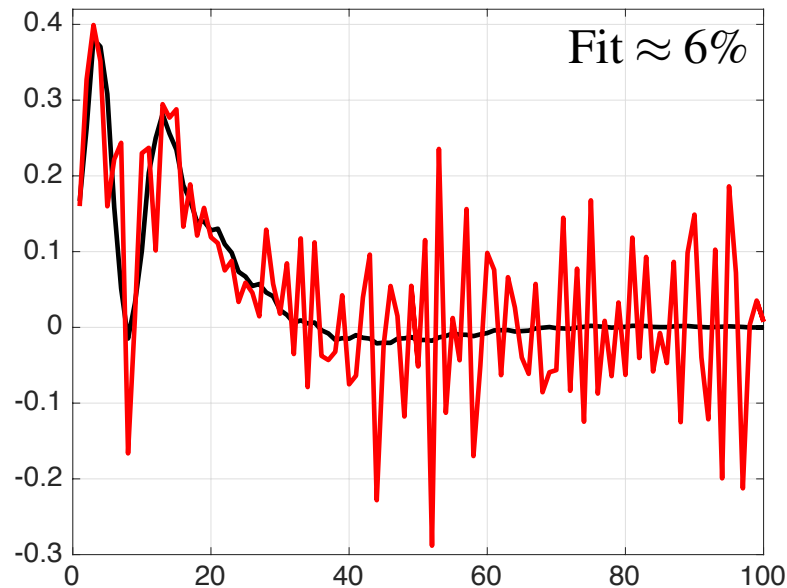
$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

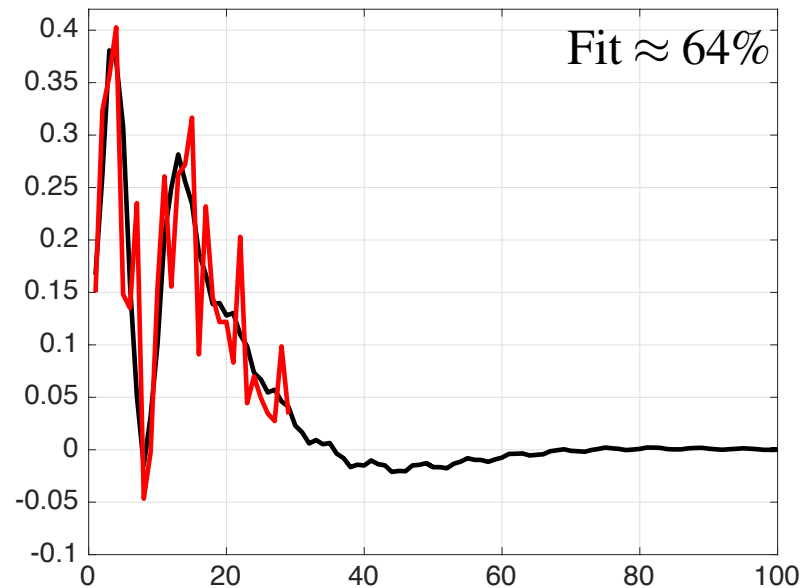
Linear least squares

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2$$

$d=100$



Oracle $d=28$



FIR models and least squares

Linear model

FIR models of order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

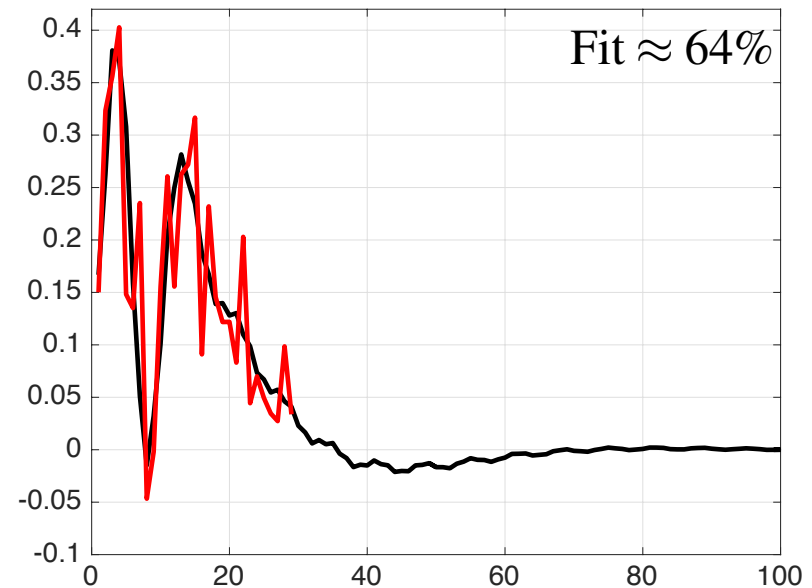
$\theta \in \mathbb{R}^d$ impulse response coefficients

Linear least squares

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2$$

Similar to
PEM+AIC

Oracle $d=28$



FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

For large d the problem is ill-posed/ill-conditioned:
how to control the variance without (discrete) tuning of d ?

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

(Hoerl and Kennard, 1970)

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

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

(Hoerl and Kennard, 1970)

Loss function

$$\hat{\theta} = \arg \min_{\theta} \left(\|Y - \Phi\theta\|^2 + \gamma \|\theta\|^2 \right)$$

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

(Hoerl and Kennard, 1970)

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

**Euclidean
norm penalty**

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

(Hoerl and Kennard, 1970)

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

**Regularization
parameter**

FIR models and ridge regression

Linear model

FIR of fixed high-order d

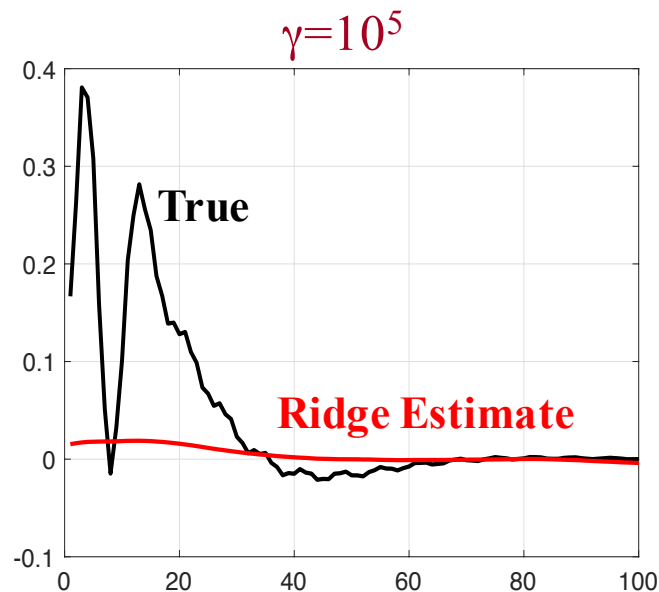
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

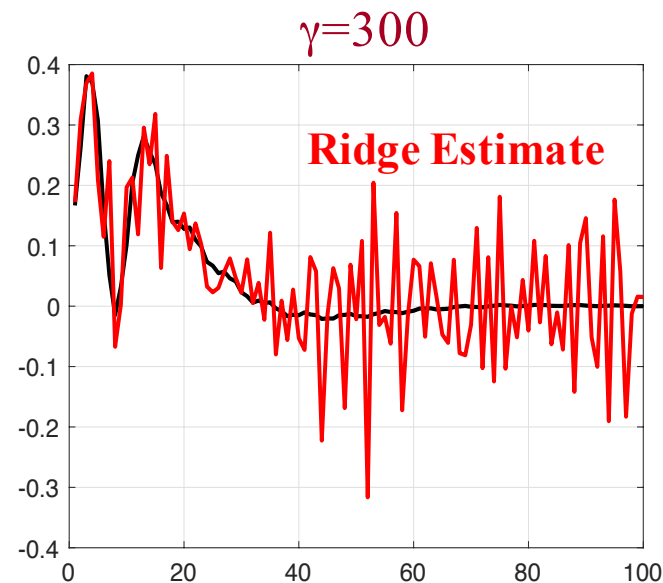
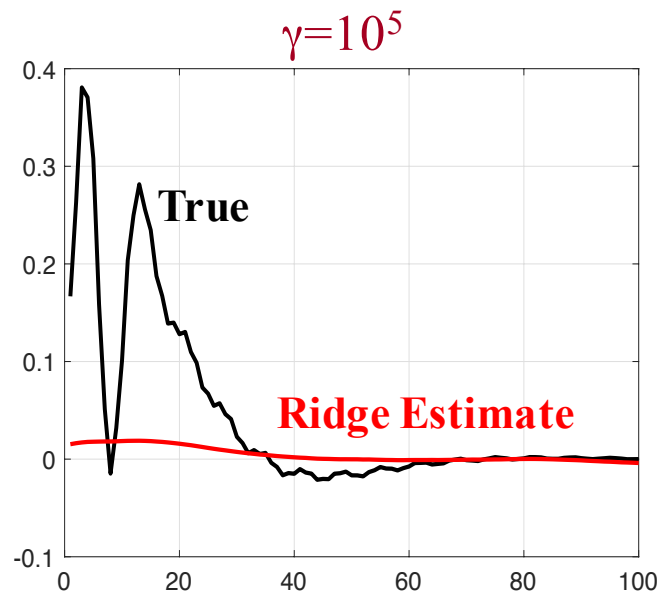
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

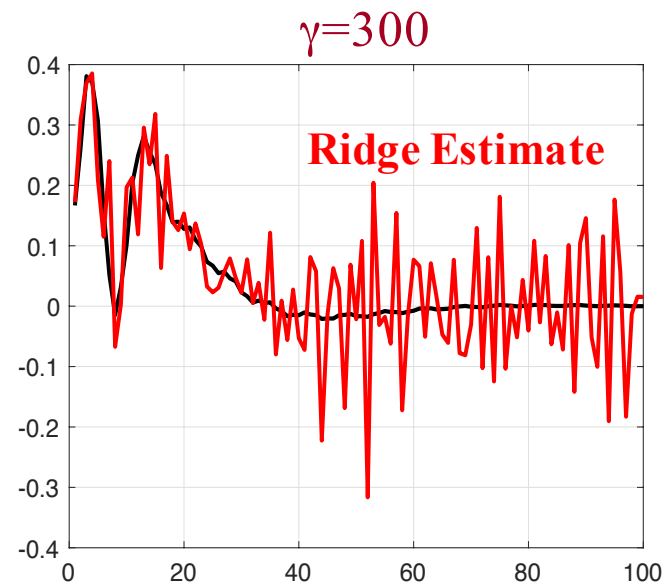
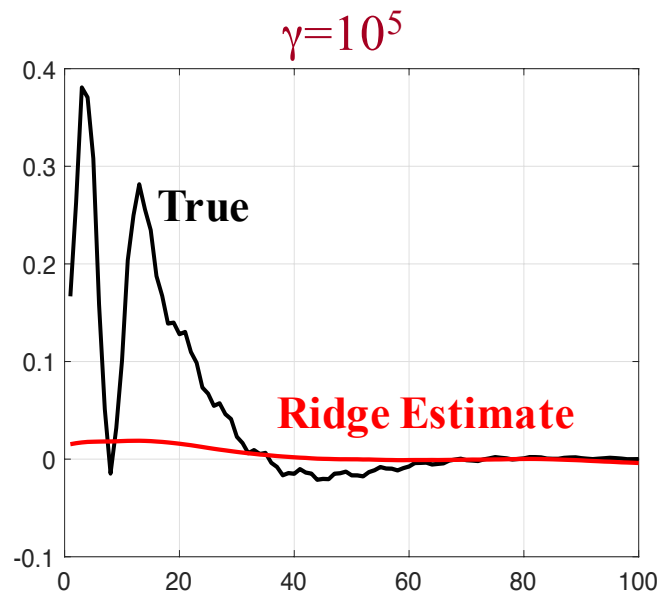
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



Difficult to understand complexity level from γ

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T\Phi + \gamma I_d)^{-1}\Phi^T)$$

$$0 \leq \text{dof}(\gamma) \leq n, \quad n = \text{data set size}$$

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T\Phi + \gamma I_d)^{-1}\Phi^T)$$

$$0 \leq \text{dof}(\gamma) \leq n, \quad n = \text{data set size}$$

γ determines a
“continuous model order”

FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

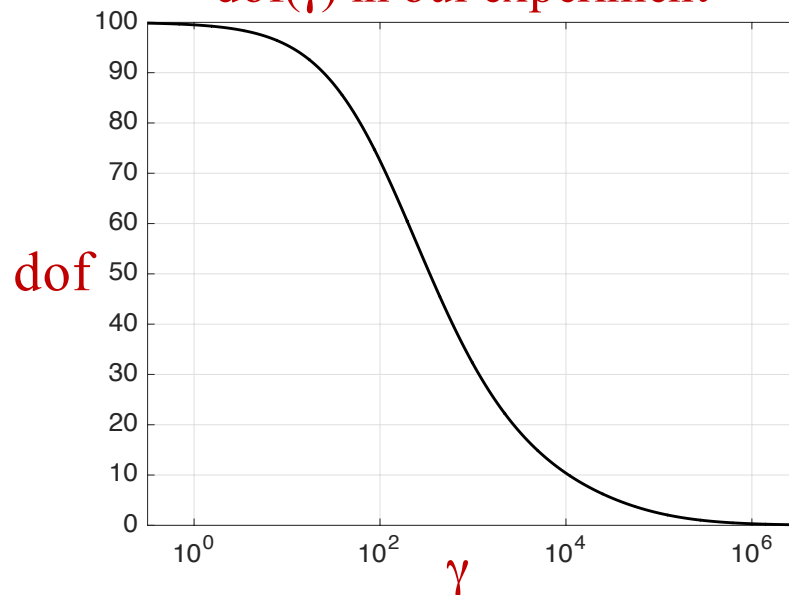
$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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

dof(γ) in our experiment



FIR models and ridge regression

Linear model

FIR of fixed high-order d

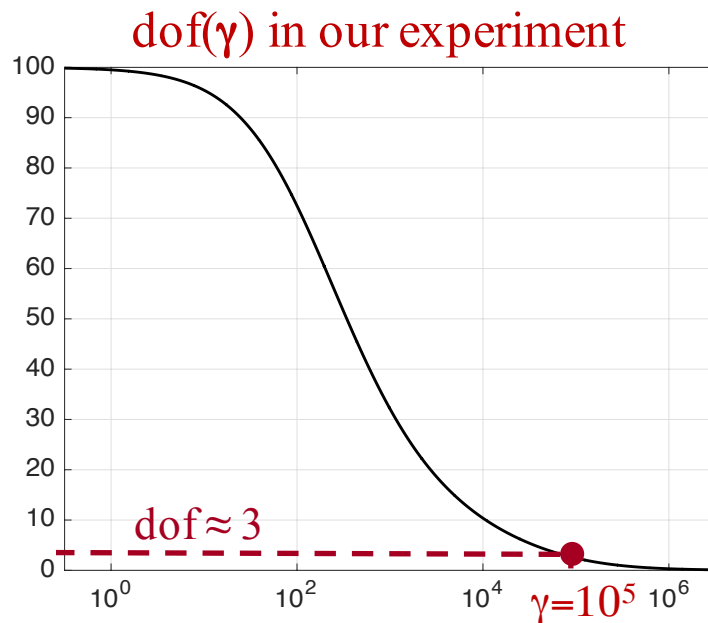
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

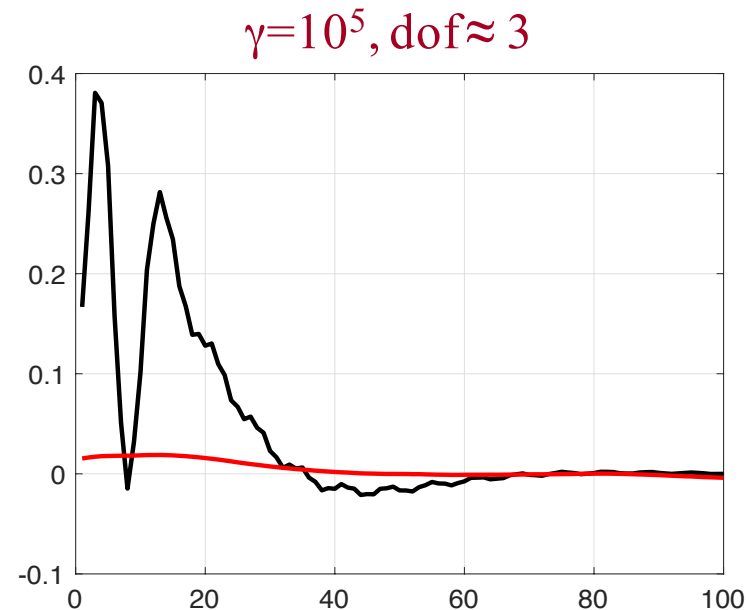
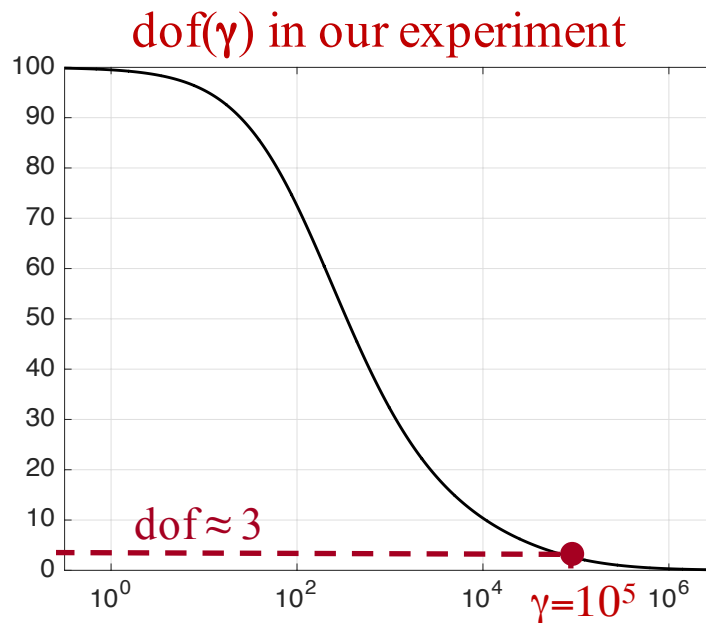
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

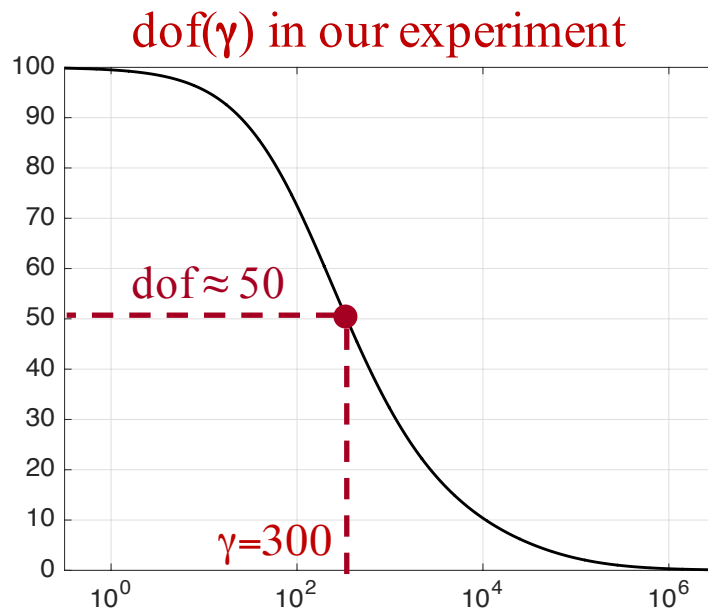
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

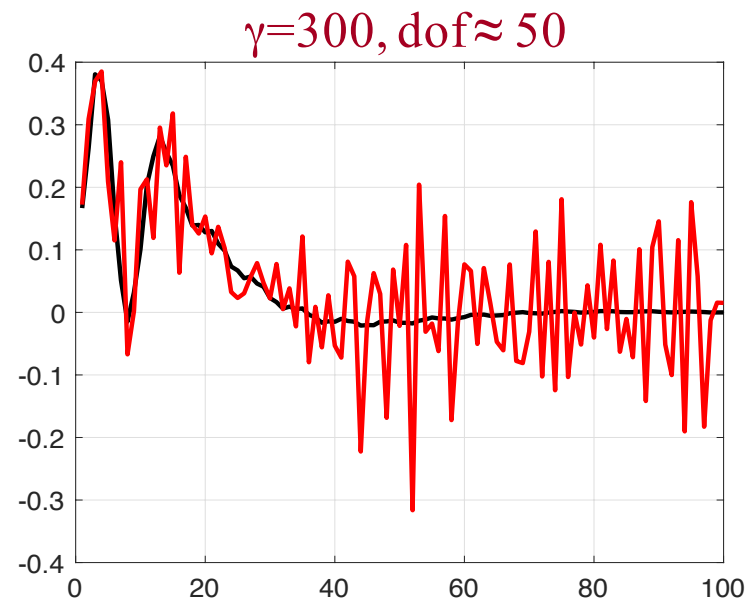
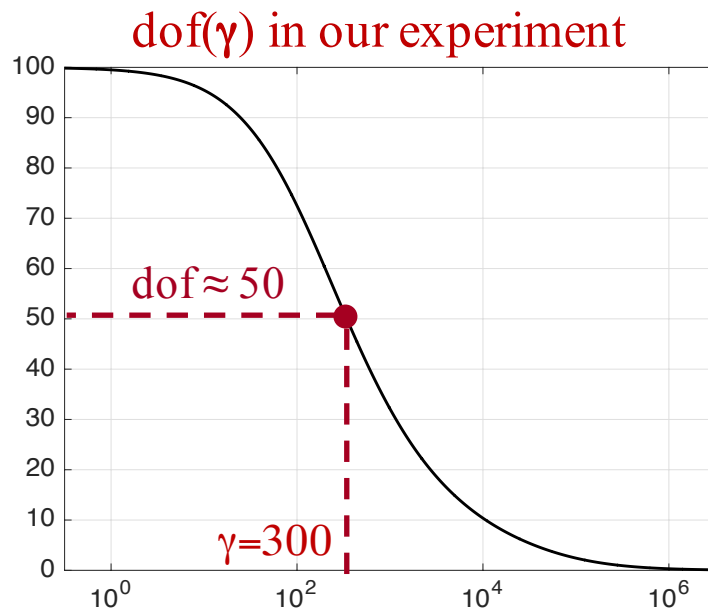
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

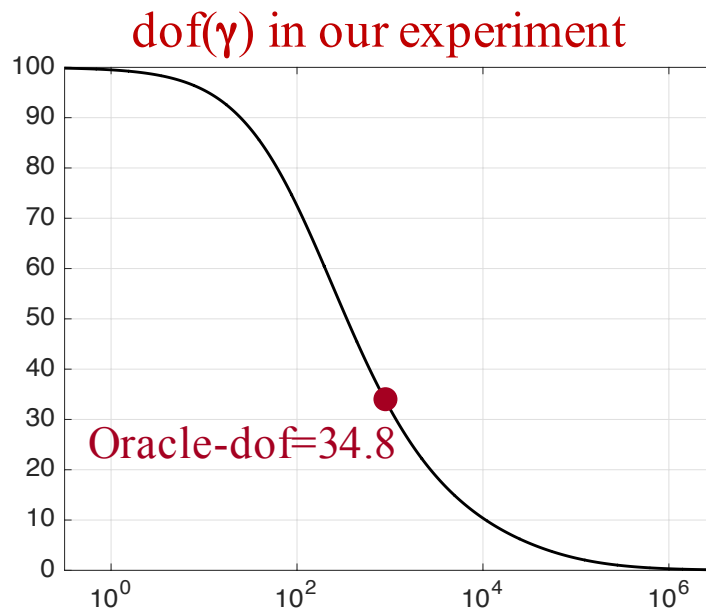
$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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



FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

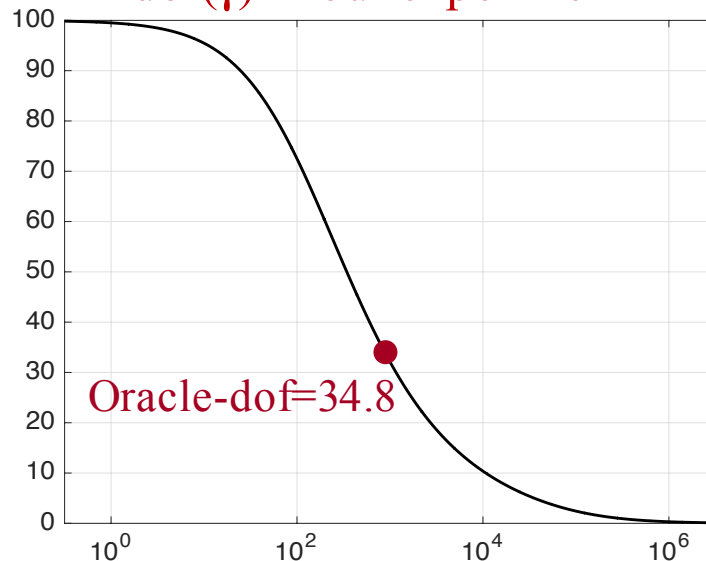
$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

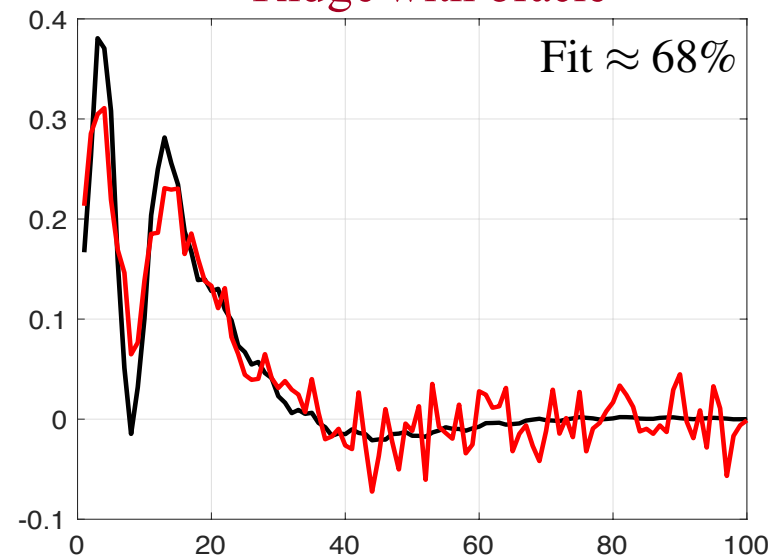
Ridge regression

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

dof(γ) in our experiment



Ridge with oracle



FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

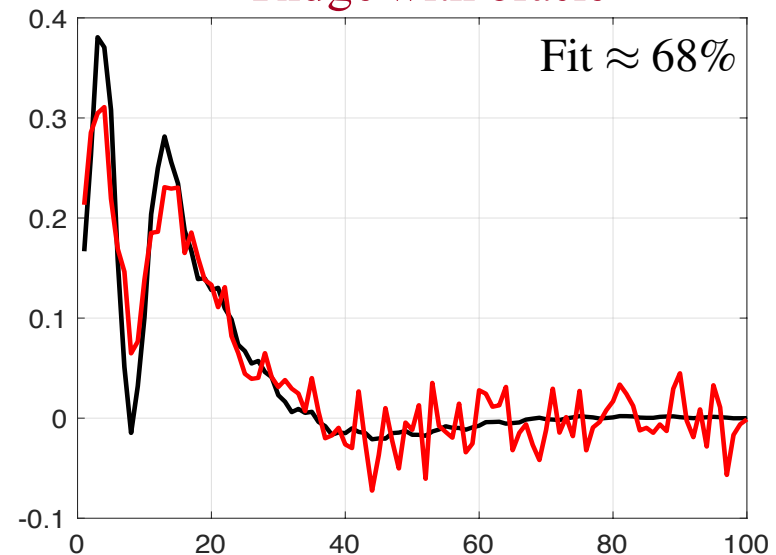
$\theta \in \mathbb{R}^d$ impulse response coefficients

Ridge regression

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

Not much better than
PEM+AIC

Ridge with oracle



FIR models and ridge regression

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

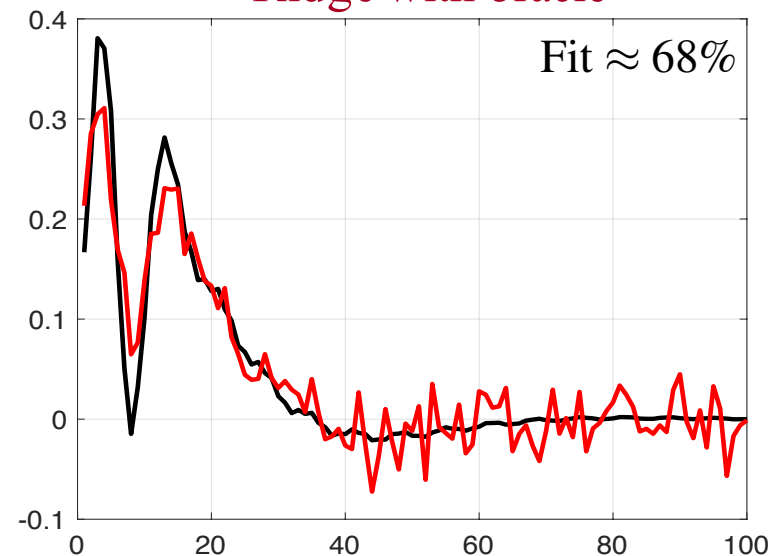
Ridge regression

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

Not much better than
PEM+AIC

**Can ridge regression
be made more effective?**

Ridge with oracle



Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov regularization theory in
reproducing kernel Hilbert spaces (RKHS)

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

RKHS and regularization theory

(Aronszajn 1950; Bergman 1950; Tikhonov 1960)

N. Aronszajn Theory of reproducing kernels, Trans. of the American Mathematical Society, 1950

S. Bergman The Kernel Function and Conformal Mapping, AMS 1950

A.N. Tikhonov and V.Y. Arsenin Solutions of Ill-Posed Problems, Winston/Wiley, 1977.

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

RKHS and regularization theory
(Aronszajn 1950; Bergman 1950; Tikhonov 1960)

**First applications in statistics/approximation theory/computer vision
(Bertero, Poggio and Torre 1988; Poggio and Girosi 1990; Wahba 1990)**

M. Bertero, T. Poggio, and V. Torre Ill-posed problems in early vision. Proceedings of the IEEE, 1988

T. Poggio and F. Girosi Networks for approximation and learning. Proceedings of the IEEE, 1990

G. Wahba Spline models for observational data, SIAM 1990

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

RKHS and regularization theory
(Aronszajn 1950; Bergman 1950; Tikhonov 1960)

First applications in statistics/approximation theory/computer vision
(Bertero, Poggio and Torre 1988; Poggio and Girosi 1990; Wahba 1990)

Machine learning:
regularization networks and support vector machines
(Girosi 1997; Vapnik 1998; Scholkopf et al. 2000; Evgeniou et al. 2000; Suykens et al. 2002)

F. Girosi An equivalence between sparse approximation and support vector machines. Technical report Cambridge (USA) 1997

V. Vapnik Statistical learning theory, Wiley 1998

T. Evgeniou, M. Pontil and T. Poggio Regularization networks and svm, Advances in Computational Mathematics 2000

B. Scholkopf and A.J. Smola Learning with kernels, support vector machines, regularization and beyond, MIT Press 2001

J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle Least squares support vector machines, World Scientific 2002

...

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

Stable RKHS for impulse response and predictor estimation

(Pillonetto and De Nicolao 2010; Pillonetto, Chiuso and De Nicolao 2011)

G. Pillonetto and G. De Nicolao A new kernel-based approach for linear system identification, Automatica 2010

G. Pillonetto, A. Chiuso and G. De Nicolao Prediction error identification of linear systems: a Gaussian regression approach, Automatica 2011

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

Stable RKHS for impulse response and predictor estimation
(Pillonetto and De Nicolao 2010; Pillonetto, Chiuso and De Nicolao 2011)

Connection with regularized IIR or high-order FIR models

Distributed lag estimators, smoothness regularizers for sysid, stochastic embedding
(Leamer, 1972; Shiller 1973; Akaike 1979; Kitagawa and Gersch 1996; Goodwin, Gevers and Ninness 1992)

E. Leamer A class of informative priors and distributed lag analysis, *Econometrics* 1972
R. Shiller A distributed lag estimator derived from smoothness priors, *Econometrics* 1973
H. Akaike Smoothness priors and the distributed lag estimator, *Tech. Stanford* 1979
G. Kitagawa and W. Gersch Smoothness priors analysis of time series, Springer 1996
G. Goodwin, M. Gevers and B. Ninness Quantifying the error in estimated transfer functions with application to model order selection, *IEEE TAC* 1992

Kernel-based regularization

Search for an unknown function in a high-dimensional space $\mathcal{H}$

Infinite dimensional space

Tikhonov + RKHS



Kernel-based methods

Stable RKHS for impulse response and predictor estimation
(Pillonetto and De Nicolao 2010; Pillonetto, Chiuso and De Nicolao 2011)

Distributed lag estimators, smoothness regularizers for sysid, stochastic embedding
(Leamer, 1972; Shiller 1973; Akaike 1979; Kitagawa and Gersch 1996; Goodwin, Gevers and Ninness 1992)

More recent work on regularized FIR and a survey

(Chen, Ohlsson and Ljung, 2012; Pillonetto, Dinuzzo, Chen, De Nicolao, Ljung 2014)

T. Chen, H. Ohlsson and L. Ljung On the estimation of transfer functions, regularization and Gaussian processes – revisited, Automatica 2012

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 2014

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Which role can kernel-based methods play
for our high-order FIR?

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Kernel-based approach

- rank candidate solutions
using a kernel K

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Kernel-based approach

- rank candidate solutions
using a kernel K
- in our setting, K is a nonnegative definite matrix
measuring “FIR complexity”

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Kernel-based approach

- rank candidate solutions
using a kernel K
- in our setting, K is a nonnegative definite matrix
measuring “FIR complexity”

Regularized least squares estimator

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Kernel-based regularization

Search for the impulse response in a high-dimensional space $\mathcal{H}$

Linear model

FIR of fixed high-order d

$$Y = \Phi\theta + v$$

$\Phi \in \mathbb{R}^{n \times d}$ with input values

$\theta \in \mathbb{R}^d$ impulse response coefficients

Kernel-based approach

- rank candidate solutions
using a kernel K
- in our setting, K is a nonnegative definite matrix
measuring “FIR complexity”

Regularized least squares estimator

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta$$

Regularizer
depends on K

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Key quantities

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Key quantities

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T \Phi + \gamma K^{-1})^{-1} \Phi^T)$$

$$0 \leq \text{dof}(\gamma) \leq n, \quad n = \text{data set size}$$

γ determines a
“continuous model order”

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Key quantities

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T \Phi + \gamma K^{-1})^{-1} \Phi^T)$$

$$0 \leq \text{dof}(\gamma) \leq n, \quad n = \text{data set size}$$

γ determines a
“continuous model order”

Classical kernel choices

Ridge regression

$$K_{ij} = \delta_{ij} \quad \begin{array}{l} \delta = \text{Kronecker} \\ \text{delta} \end{array}$$

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

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Key quantities

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T \Phi + \gamma K^{-1})^{-1} \Phi^T)$$

$$0 \leq \text{dof}(\gamma) \leq n, \quad n = \text{data set size}$$

γ determines a
“continuous model order”

Classical kernel choices

Ridge regression

$$K_{ij} = \delta_{ij}$$

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

Gaussian kernel

$$K_{ij} = e^{-(i-j)^2}$$

most used in machine learning

Degrees of freedom and kernel choice

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Key quantities

Equivalent degrees of freedom (dof)

$$\text{dof}(\gamma) = \text{Trace}(\Phi(\Phi^T \Phi + \gamma K^{-1})^{-1} \Phi^T)$$

$0 \leq \text{dof}(\gamma) \leq n$, n =data set size

γ determines a “continuous model order”

Classical kernel choices

Ridge regression

$$K_{ij} = \delta_{ij}$$

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

Gaussian kernel

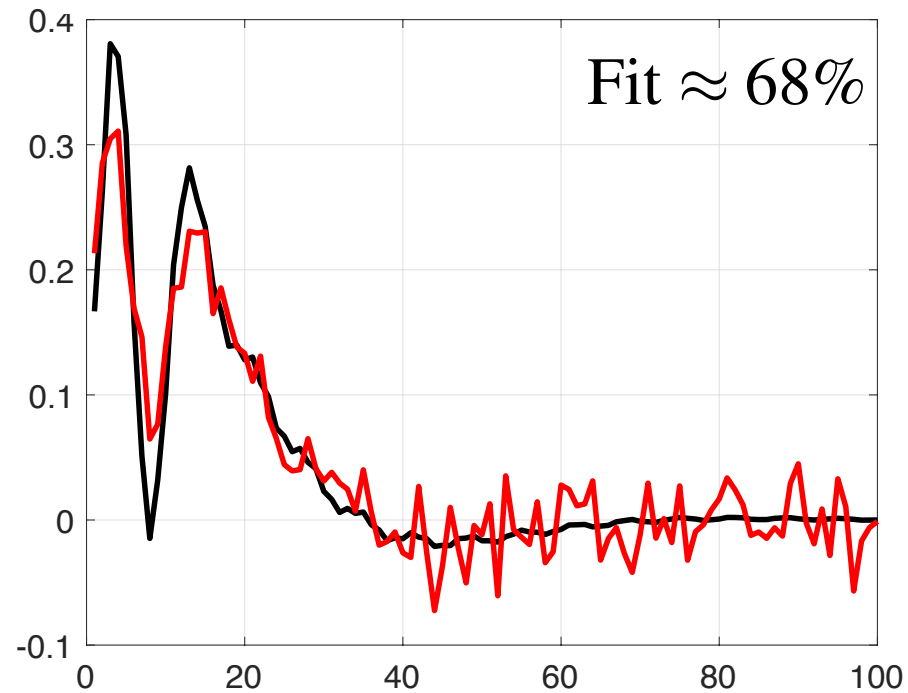
$$K_{ij} = e^{-(i-j)^2}$$

most used in machine learning

Let us now compare kernels performance
with dof chosen by an oracle

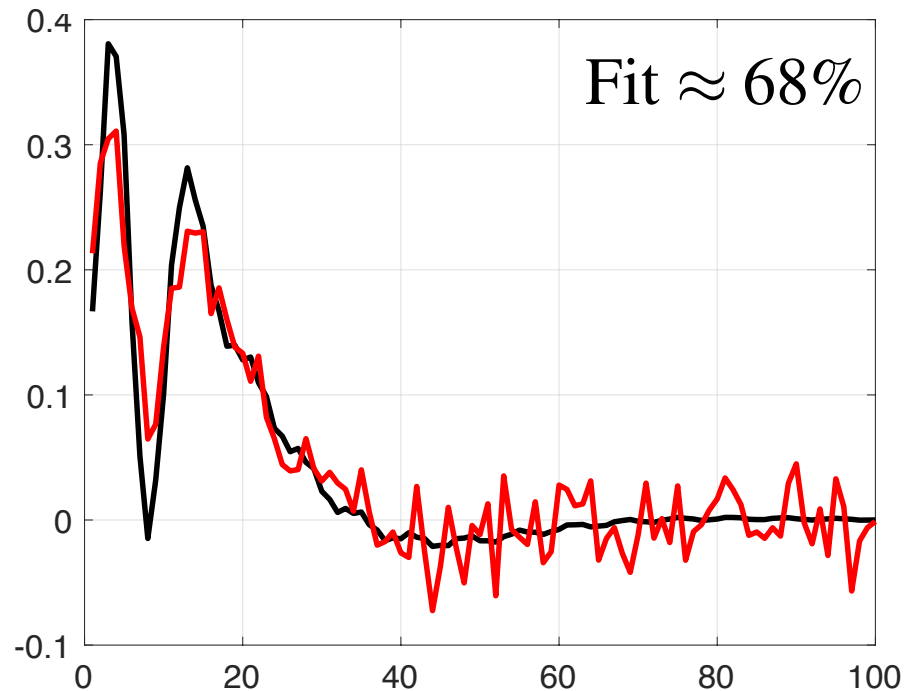
Testing Ridge regression and Gaussian kernel

Ridge regression
with oracle (dof = 34.8)

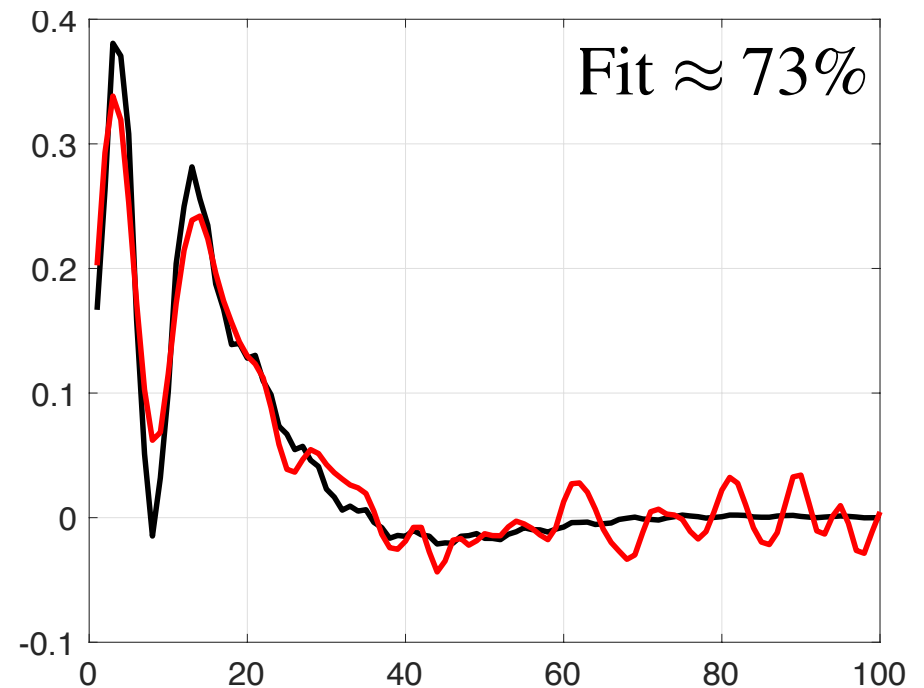


Testing Ridge regression and Gaussian kernel

Ridge regression
with oracle (dof = 34.8)

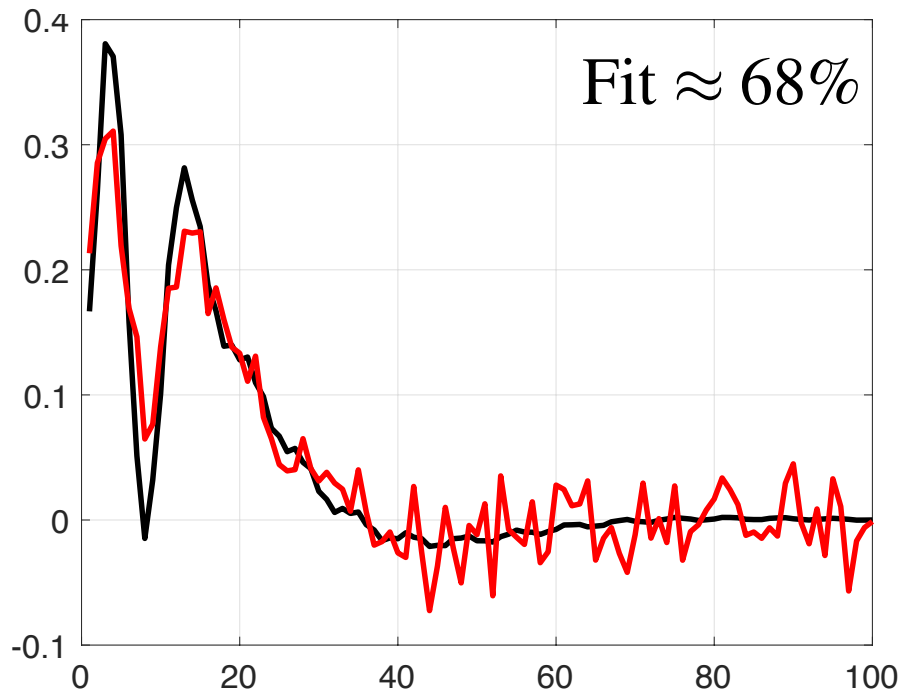


Gaussian kernel
with oracle (dof = 29.3)

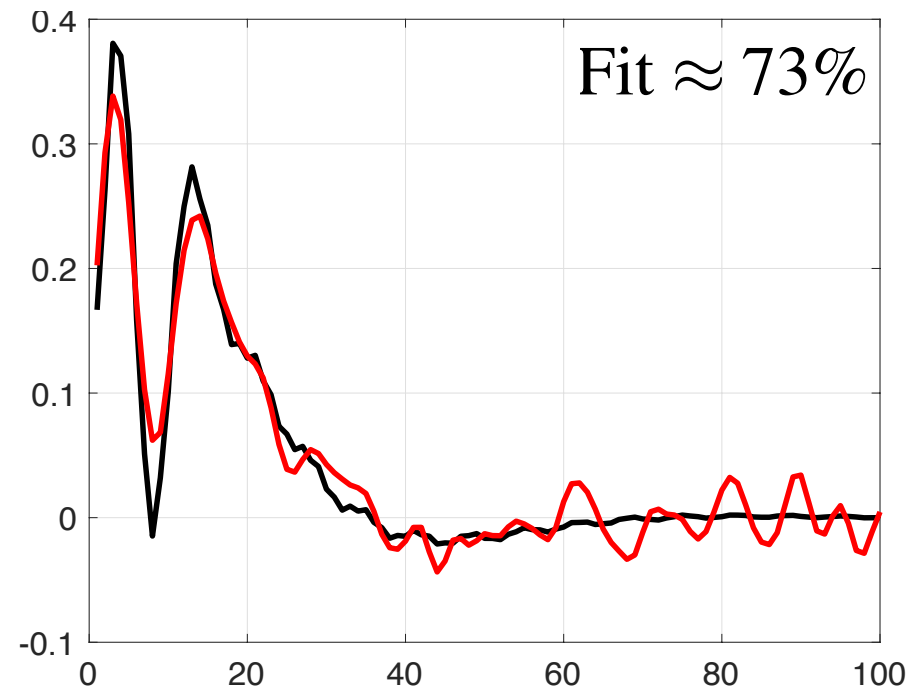


Testing Ridge regression and Gaussian kernel

Ridge regression
with oracle (dof = 34.8)



Gaussian kernel
with oracle (dof = 29.3)



Gaussian kernel only slightly better.
Not the breakthrough we were hoping for.
Why?

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

FIR of high-order d

Gaussian prior



quadratic penalty $\theta^T K^{-1} \theta$

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

FIR of high-order d

$$\theta \sim \mathcal{N}(0, \lambda K)$$

Gaussian prior



quadratic penalty $\theta^T K^{-1} \theta$

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

FIR of high-order d

Gaussian prior



quadratic penalty $\theta^T K^{-1} \theta$

$$\theta \sim \mathcal{N}(0, \lambda K)$$

Scale factor

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

FIR of high-order d

Gaussian prior



quadratic penalty $\theta^T K^{-1} \theta$

$$\theta \sim \mathcal{N}(0, \lambda K)$$

$$\theta \perp v$$

$$v \sim \mathcal{N}(0, \sigma^2 I)$$

Bayesian view

Model and estimator

$$Y = \Phi\theta + v$$

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma\theta^T K^{-1} \theta$$

Bayesian interpretation

(Kimeldorf and Wahba 1971; Williams and Rasmussen 1996;
Aravkin et al. 2015)

FIR of high-order d

Gaussian prior



quadratic penalty $\theta^T K^{-1} \theta$

$$\theta \sim \mathcal{N}(0, \lambda K)$$

$$\theta \perp v$$

$$v \sim \mathcal{N}(0, \sigma^2 I)$$

$$\hat{\theta} = \mathcal{E}[\theta|Y] \quad \text{if} \quad \gamma = \frac{\sigma^2}{\lambda}$$

Estimate = posterior mean

Bayesian view on Ridge and Gaussian kernel

Prior model

$$\theta \sim \mathcal{N}(0, \lambda K)$$

Ridge regression

$$K_{ij} = \delta_{ij}$$

white noise model
(uncorrelated coefficients)

Bayesian view on Ridge and Gaussian kernel

Prior model

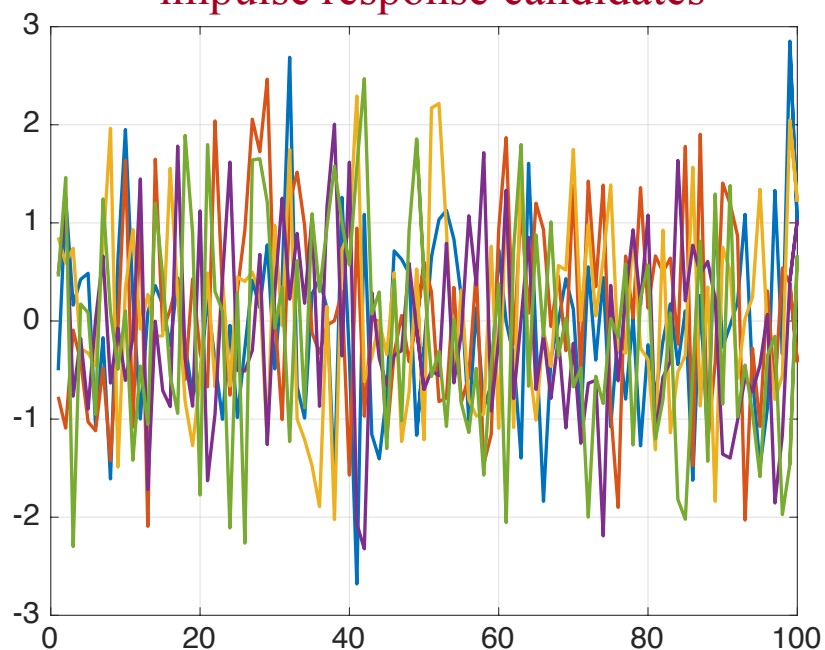
$$\theta \sim \mathcal{N}(0, \lambda K)$$

Ridge regression

$$K_{ij} = \delta_{ij}$$

white noise model
(uncorrelated coefficients)

Random sample of typical
impulse response candidates



Bayesian view on Ridge and Gaussian kernel

Prior model

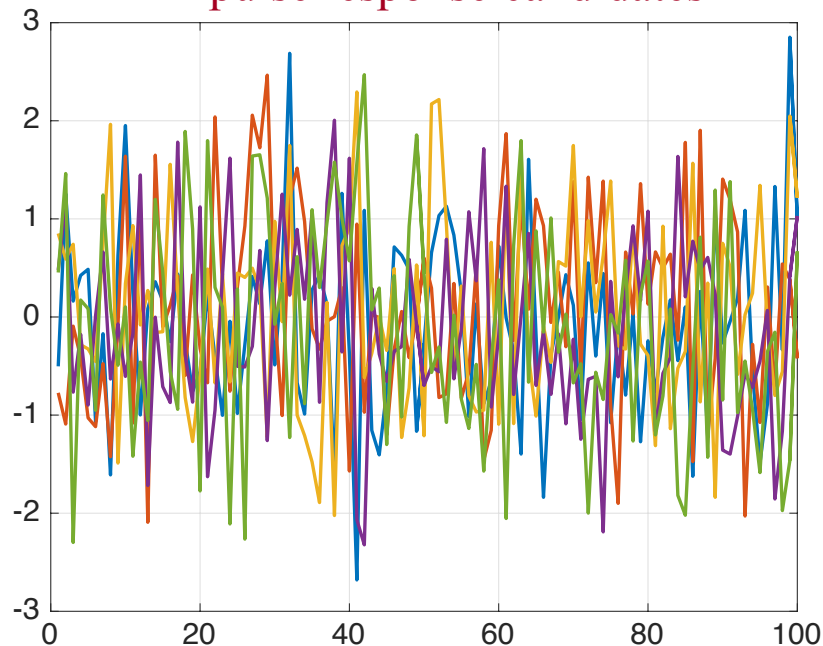
$$\theta \sim \mathcal{N}(0, \lambda K)$$

Ridge regression

$$K_{ij} = \delta_{ij}$$

white noise model
(uncorrelated coefficients)

Random sample of typical
impulse response candidates



Gaussian kernel

$$K_{ij} = e^{-(i-j)^2}$$

stationary process
(correlated but constant variance)

Bayesian view on Ridge and Gaussian kernel

Prior model

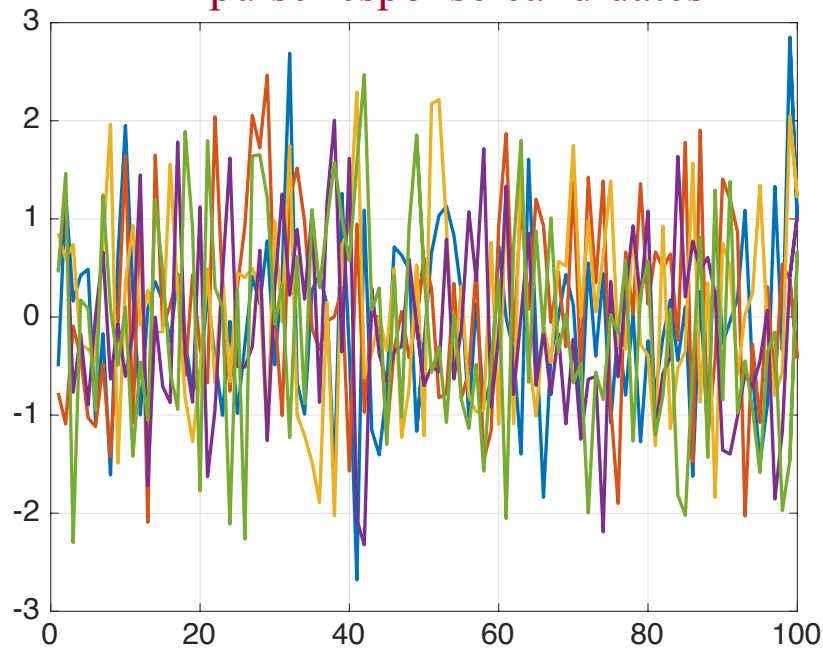
$$\theta \sim \mathcal{N}(0, \lambda K)$$

Ridge regression

$$K_{ij} = \delta_{ij}$$

white noise model
(uncorrelated coefficients)

Random sample of typical
impulse response candidates

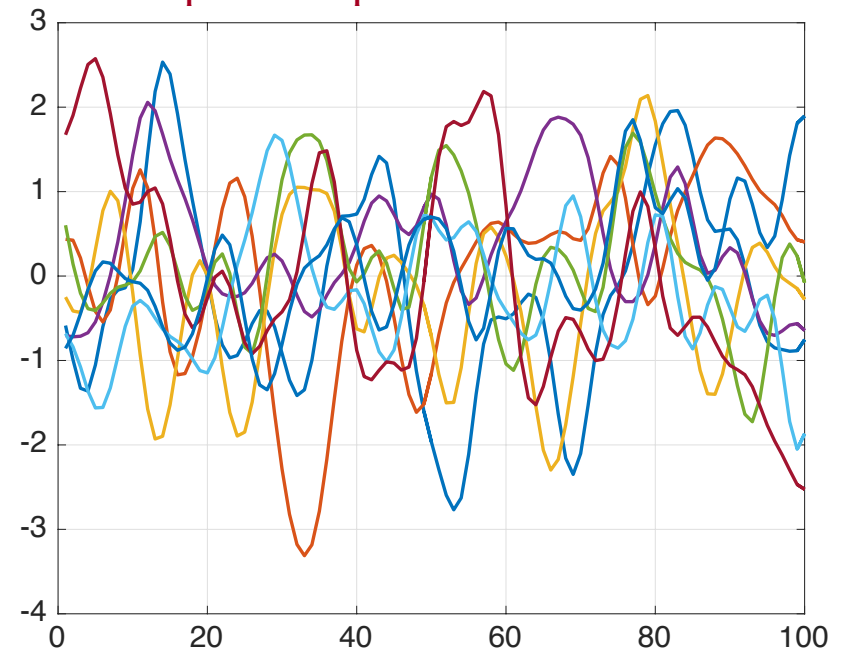


Gaussian kernel

$$K_{ij} = e^{-(i-j)^2}$$

stationary process
(correlated but constant variance)

Random sample of typical
impulse response candidates



Bayesian view on Ridge and Gaussian kernel

Prior model

$$\theta \sim \mathcal{N}(0, \lambda K)$$

Ridge regression

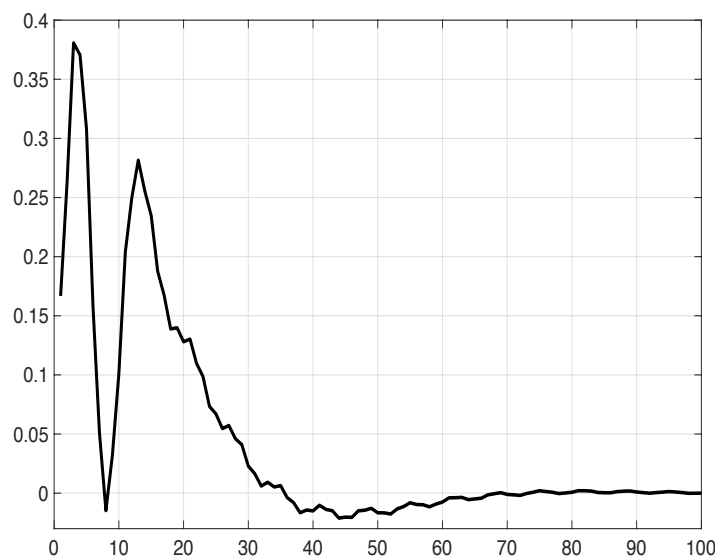
$$K_{ij} = \delta_{ij}$$

white noise model
(uncorrelated coefficients)

Gaussian kernel

$$K_{ij} = e^{-(i-j)^2}$$

stationary process
(correlated but constant variance)



Poor priors for
exponentially decaying
impulse responses

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach (Jaines, 1957)

- derives a complete a priori density function from incomplete information, e.g. expectations

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach (Jaines, 1957)

- derives a complete a priori density function from incomplete information, e.g. expectations
- the prior has to satisfy some constraints and maximize the entropy

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\begin{aligned} \max_{\mathbf{p}} \quad & - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta \\ \text{s.t.} \quad & \mathcal{E}(\theta) = 0 \end{aligned}$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta$$

$$\text{s.t. } \mathcal{E}(\theta) = 0$$

$$\text{Var } \theta_i = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta$$

$$\text{s.t. } \mathcal{E}(\theta) = 0$$

$$\text{Var } \theta_i = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

Exponential decay of
impulse response variance

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta$$

$$\text{s.t. } \mathcal{E}(\theta) = 0$$

$$\text{Var } \theta_i = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

Variance
decay rate

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\boldsymbol{\theta}) \log \mathbf{p}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$\text{s.t. } \mathcal{E}(\boldsymbol{\theta}) = 0$$

$$\text{Var } \theta_i = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

Solution: Gaussian vector with exponentially
decaying diagonal kernel

$$\boldsymbol{\theta} \sim \mathcal{N}(0, \lambda K)$$

with kernel

$$K_{ij} = \alpha^i \delta_{ij}$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta$$

$$\text{s.t. } \mathcal{E}(\theta) = 0$$

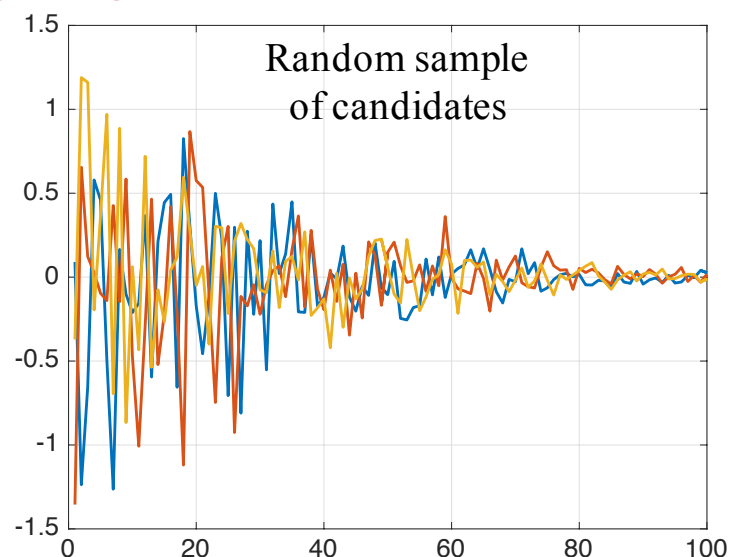
$$\text{Var } \theta_i = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

Solution: Gaussian vector with exponentially
decaying diagonal kernel

$$\theta \sim \mathcal{N}(0, \lambda K)$$

with kernel

$$K_{ij} = \alpha^i \delta_{ij}$$



Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\begin{aligned} \max_{\mathbf{p}} \quad & - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta \\ \text{s.t.} \quad & \mathcal{E}(\theta) = 0 \\ & \text{Var}(\theta_{i+1} - \theta_i) = \lambda \alpha^i, \quad 0 \leq \alpha < 1 \end{aligned}$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\begin{aligned} \max_{\mathbf{p}} \quad & - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta \\ \text{s.t.} \quad & \mathcal{E}(\theta) = 0 \\ & \text{Var}(\theta_{i+1} - \theta_i) = \lambda \alpha^i, \quad 0 \leq \alpha < 1 \end{aligned}$$

Constraints on decay of
increments variance

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\begin{aligned} \max_{\mathbf{p}} \quad & - \int \mathbf{p}(\theta) \log \mathbf{p}(\theta) d\theta \\ \text{s.t.} \quad & \mathcal{E}(\theta) = 0 \\ & \text{Var}(\theta_{i+1} - \theta_i) = \lambda \alpha^i, \quad 0 \leq \alpha < 1 \end{aligned}$$

Solution: stable spline

(Pillonetto and De Nicolao 2010; Chen et al. 2016)

$$\theta \sim \mathcal{N}(0, \lambda K)$$

with Stable
Spline kernel

$$K_{ij} = \alpha^{\max(i,j)}$$

Stable Spline kernel

Simplest prior including
smooth exponential decay?

Max-Entropy approach

$$\max_{\mathbf{p}} - \int \mathbf{p}(\boldsymbol{\theta}) \log \mathbf{p}(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

$$\text{s.t. } \mathcal{E}(\boldsymbol{\theta}) = 0$$

$$\text{Var}(\boldsymbol{\theta}_{i+1} - \boldsymbol{\theta}_i) = \lambda \alpha^i, \quad 0 \leq \alpha < 1$$

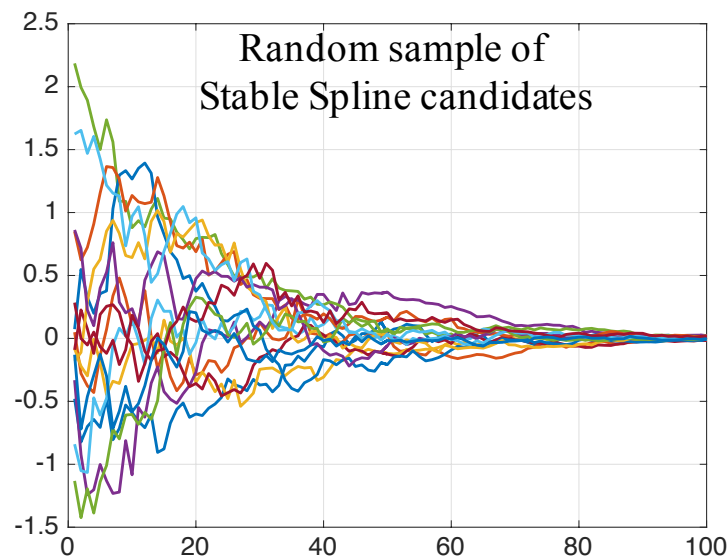
Solution: stable spline

(Pillonetto and De Nicolao 2010; Chen et al. 2016)

$$\boldsymbol{\theta} \sim \mathcal{N}(0, \lambda K)$$

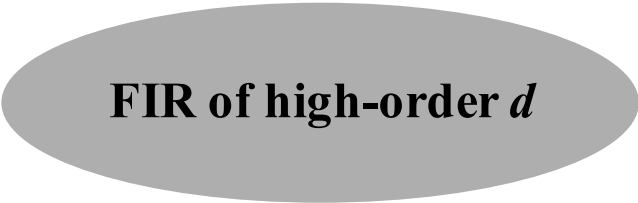
with Stable
Spline kernel

$$K_{ij} = \alpha^{\max(i,j)}$$



Ranking test

Simplified version of our sysid problem



FIR of high-order d

Ranking test

Simplified version of our sysid problem



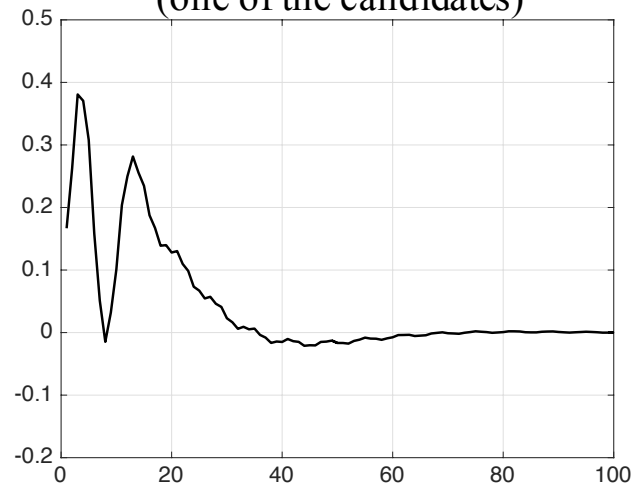
Ranking test

100 candidates (including truth)
closely fitting noiseless output

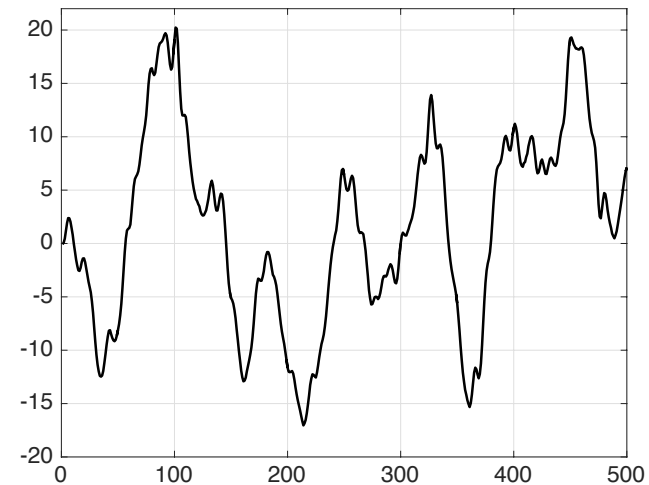
Ranking test

100 candidates (including truth)
closely fitting noiseless output

True impulse response
(one of the candidates)



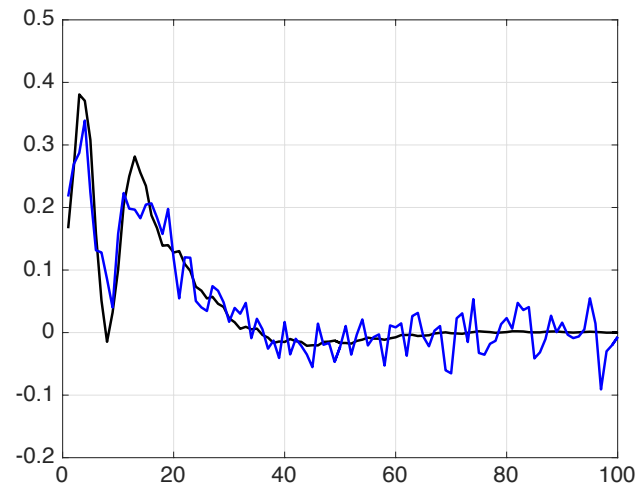
Noiseless output



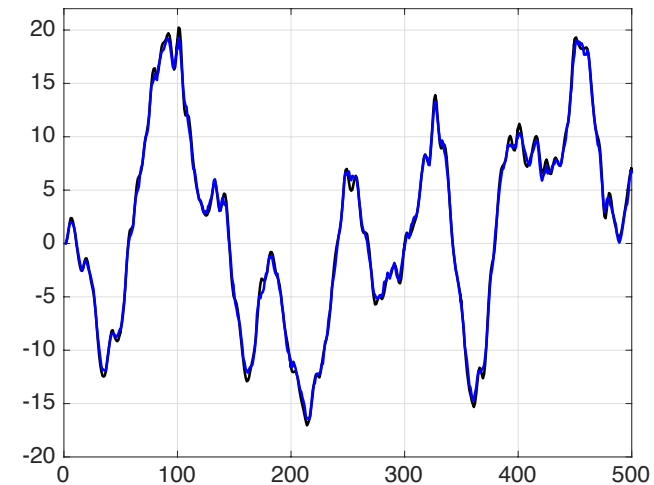
Ranking test

100 candidates (including truth)
closely fitting noiseless output

Candidates



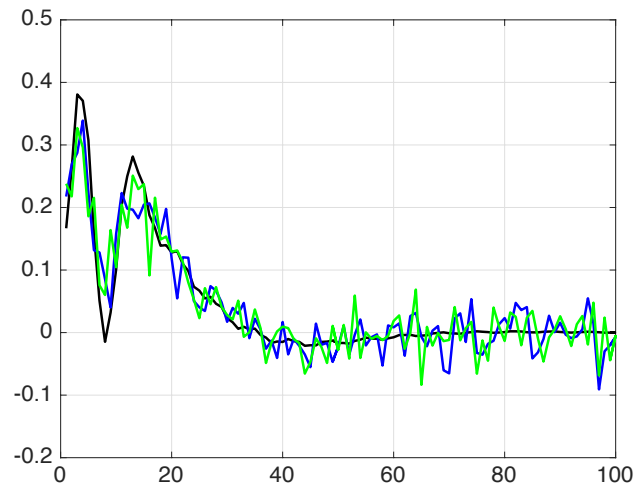
Related outputs



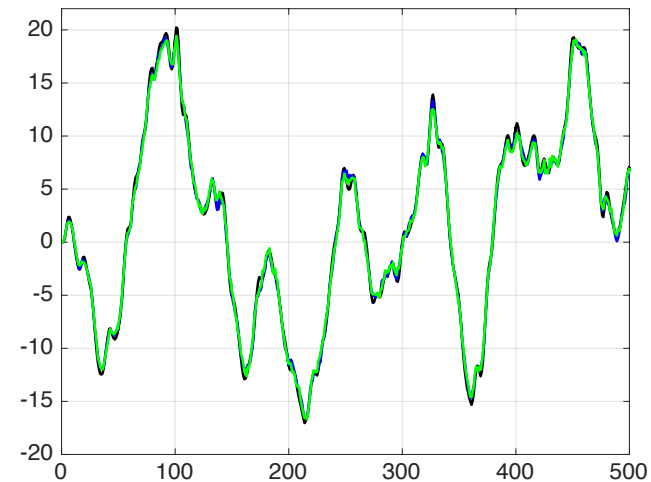
Ranking test

100 candidates (including truth)
closely fitting noiseless output

Candidates



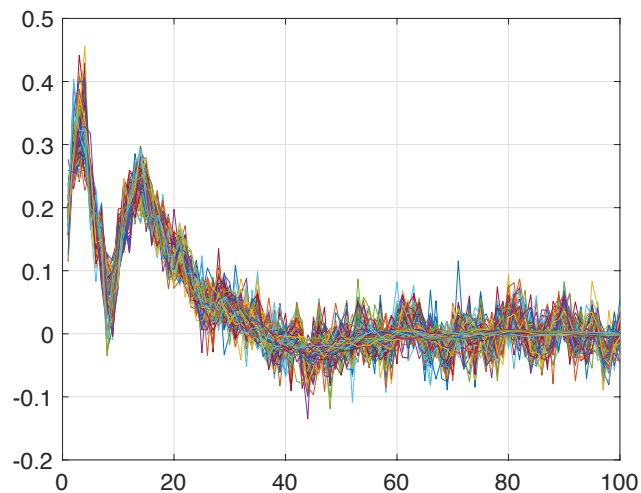
Related outputs



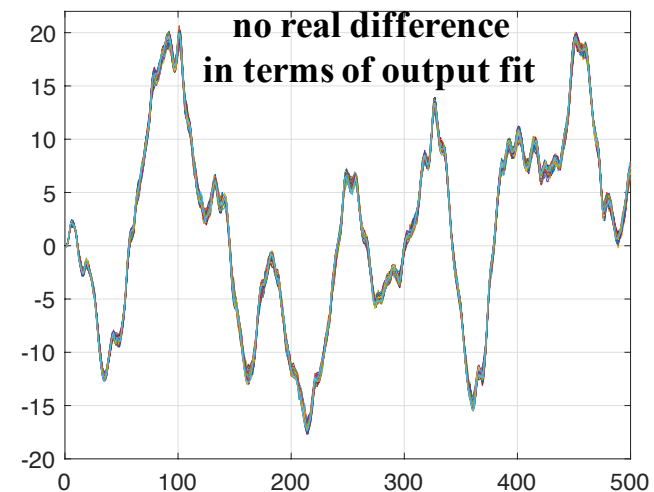
Ranking test

100 candidates (including truth)
closely fitting noiseless output

Candidates



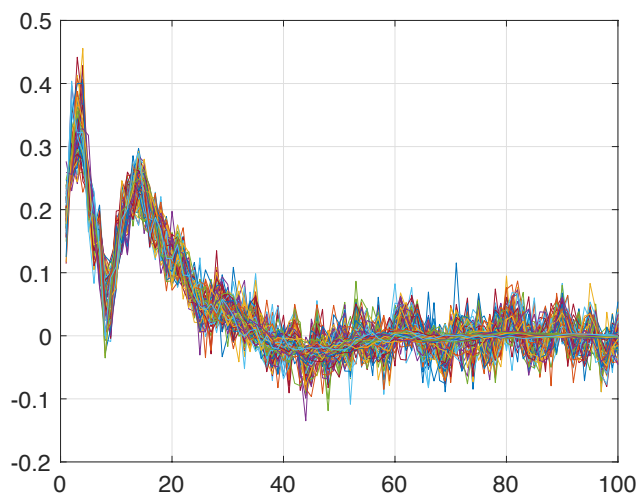
Related outputs



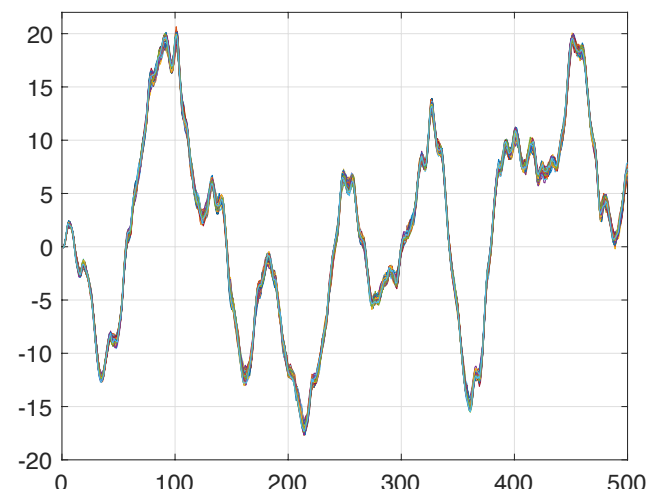
Ranking test

100 candidates (including truth)
closely fitting noiseless output

Candidates



Related outputs



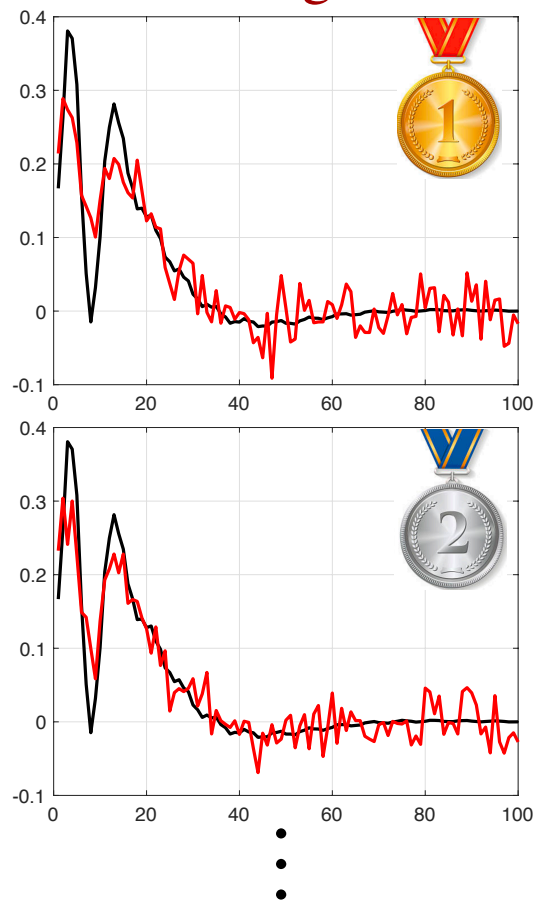
Ranking induced
by a kernel

$$\theta^{(i)} \prec \theta^{(j)} \iff \theta^{(i)\top} K^{-1} \theta^{(i)} < \theta^{(j)\top} K^{-1} \theta^{(j)}$$

Ranking test: results

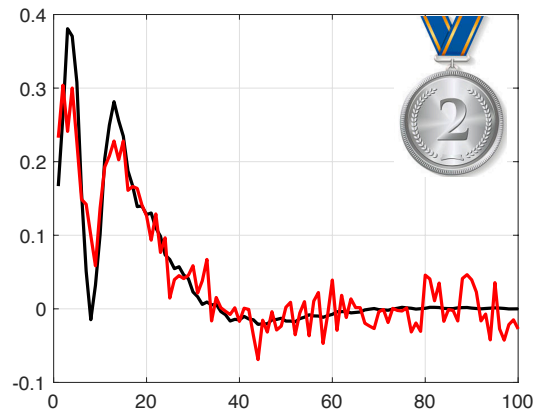
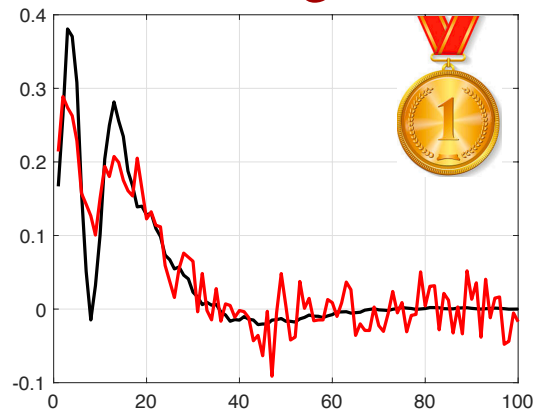
Ranking test: results

Ridge



Ranking test: results

Ridge

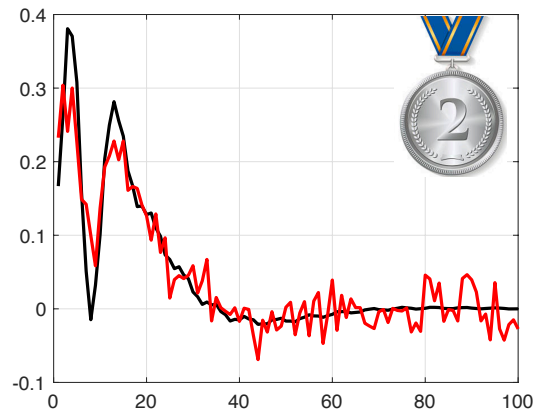
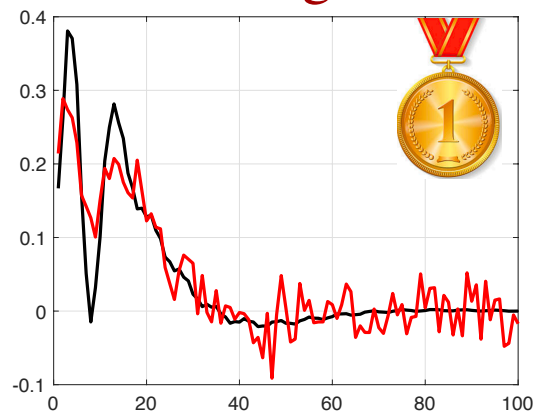


...



Ranking test: results

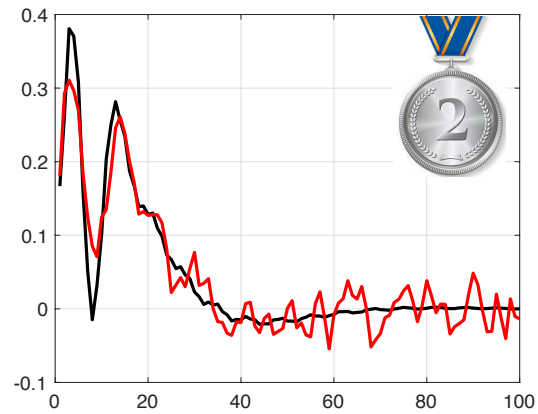
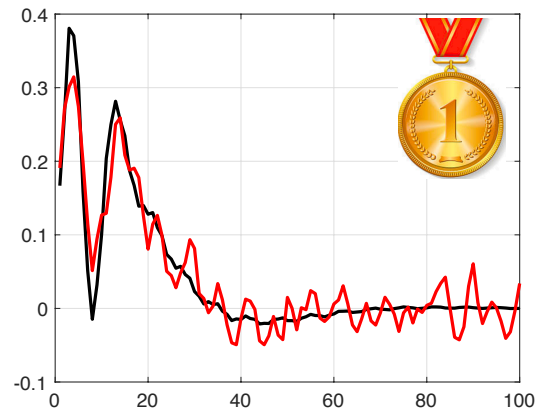
Ridge



⋮



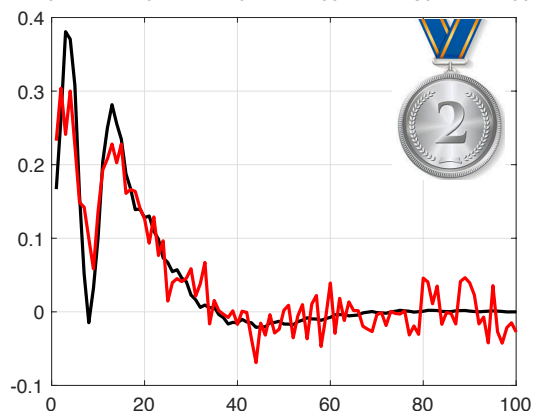
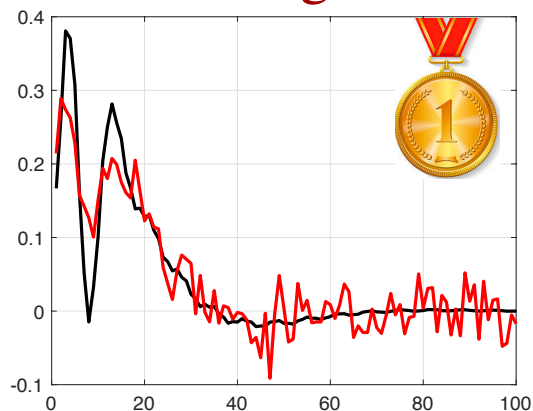
Gaussian



⋮

Ranking test: results

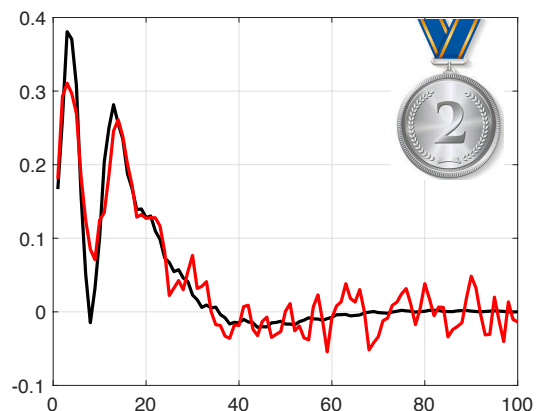
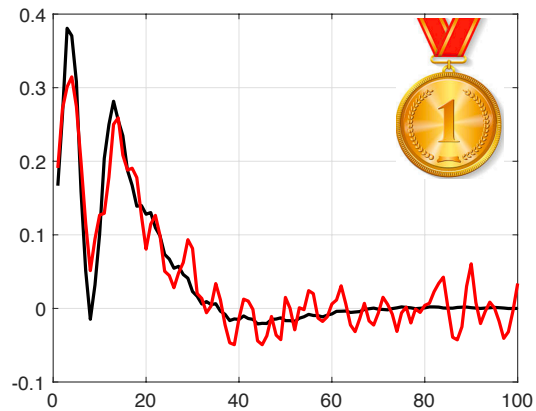
Ridge



⋮



Gaussian

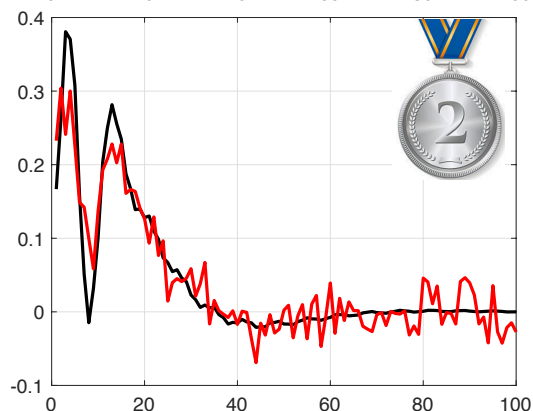
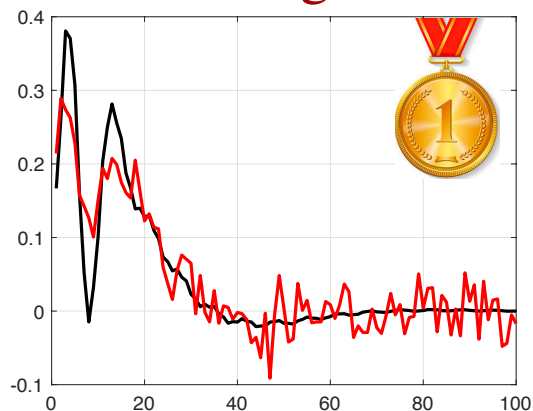


⋮



Ranking test: results

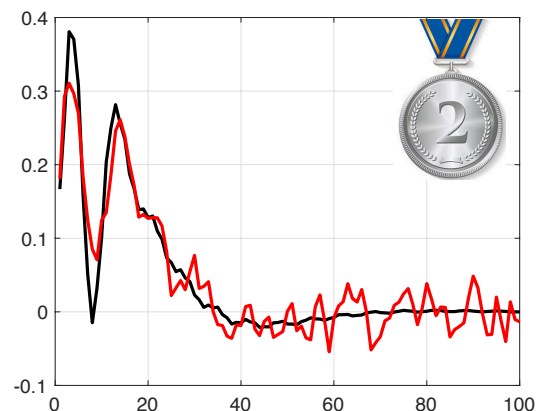
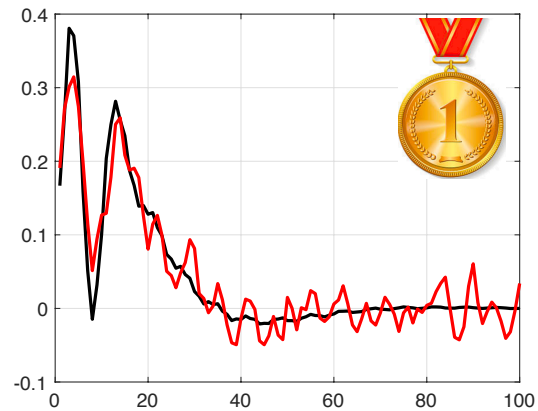
Ridge



⋮



Gaussian



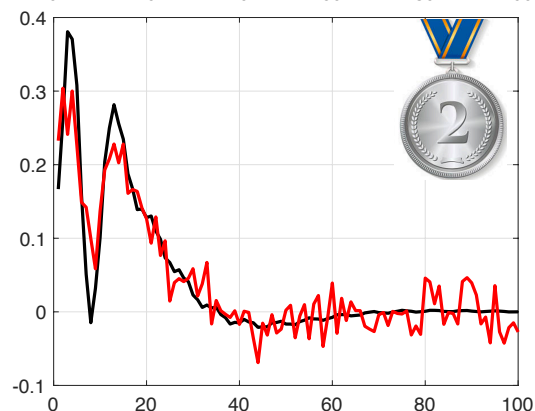
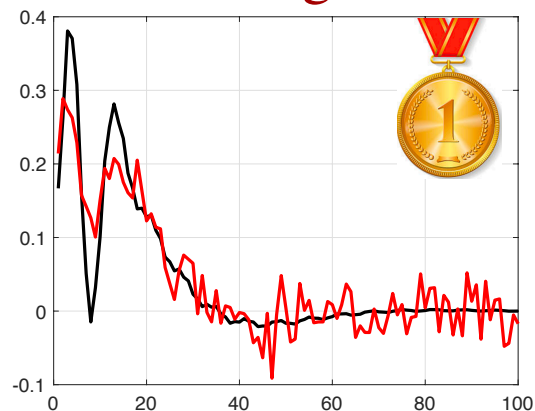
⋮



Stable Spline

Ranking test: results

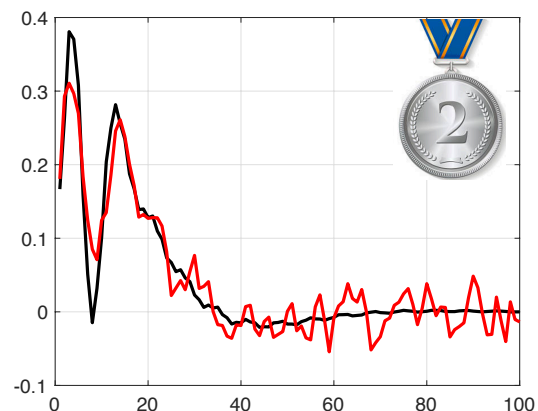
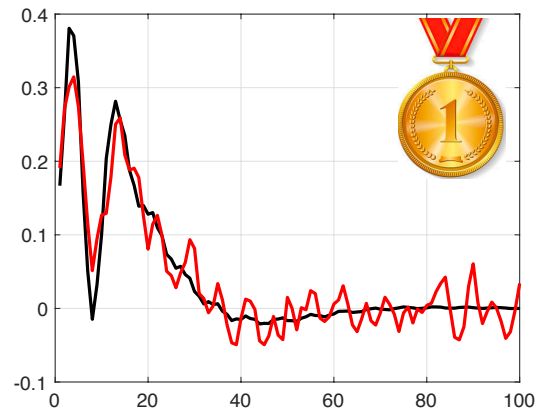
Ridge



⋮



Gaussian



⋮

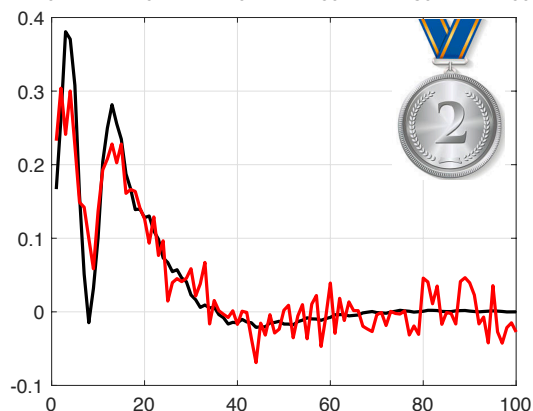
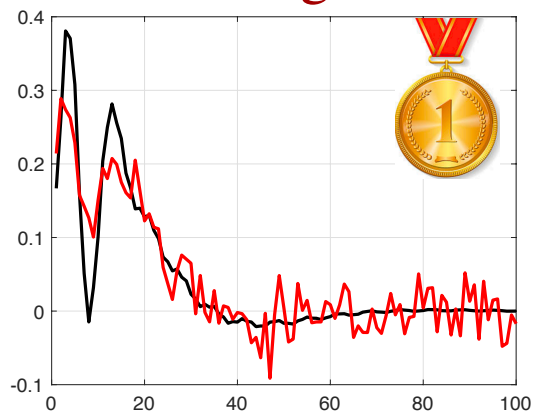


Stable Spline

Penalty on $\theta^{(i)}$:
$$\min_{\alpha} \theta^{(i)\top} K_{\alpha}^{-1} \theta^{(i)}$$

Ranking test: results

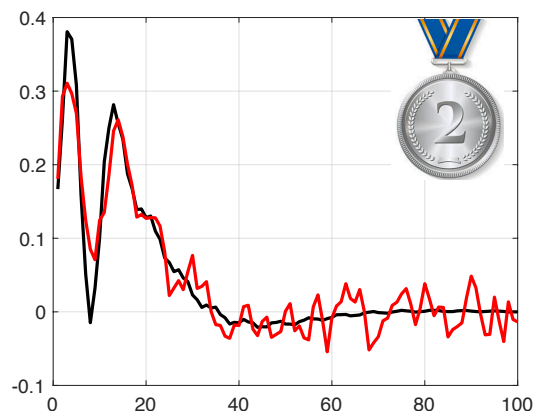
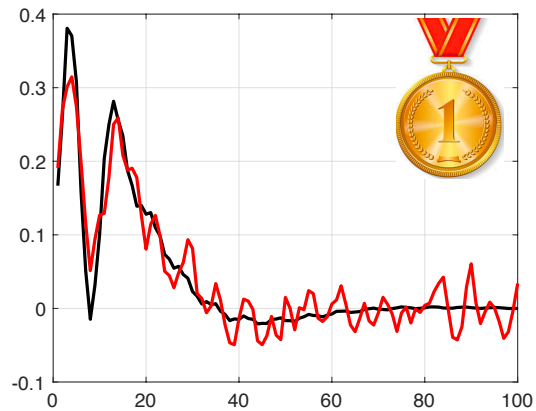
Ridge



⋮



Gaussian



⋮

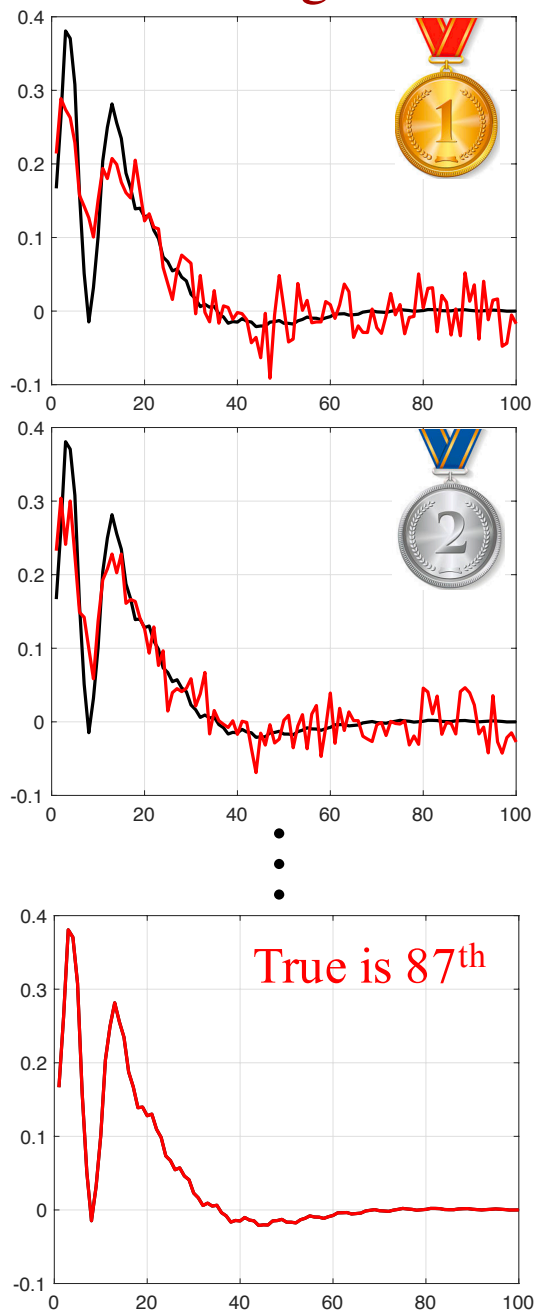


Stable Spline

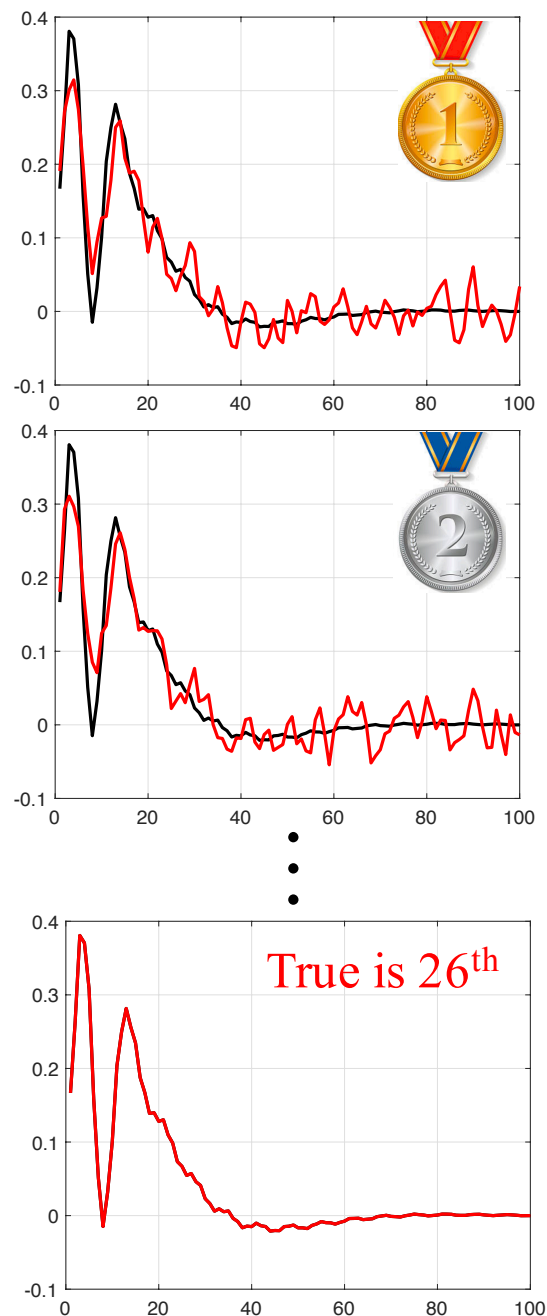


Ranking test: results

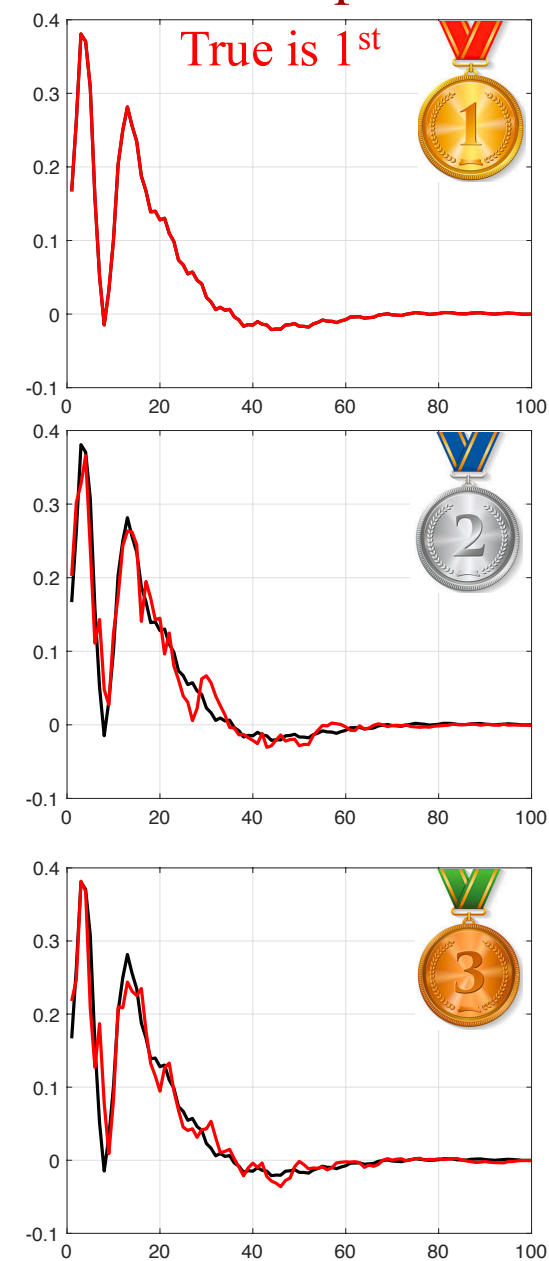
Ridge



Gaussian

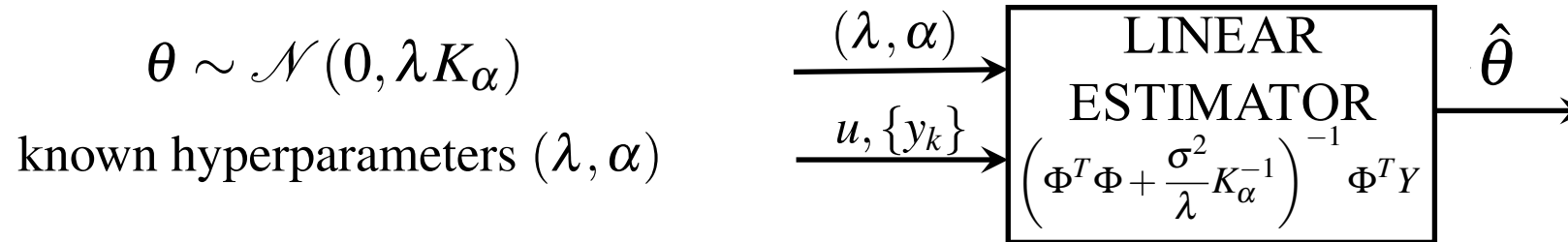


Stable Spline



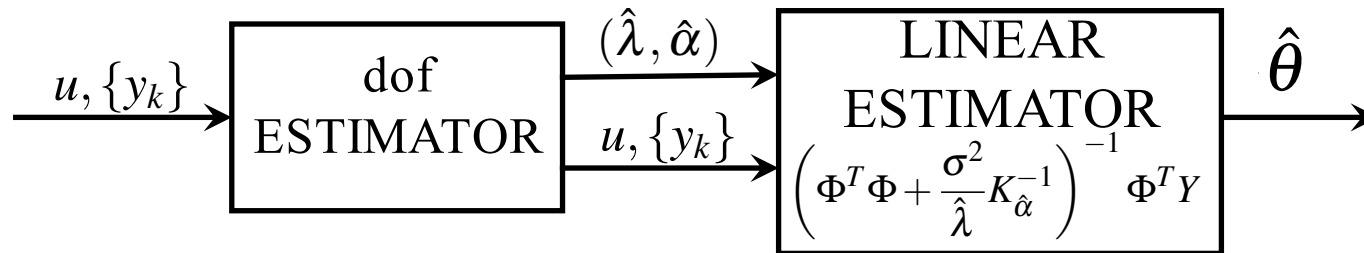
Stable spline estimator

Known complexity



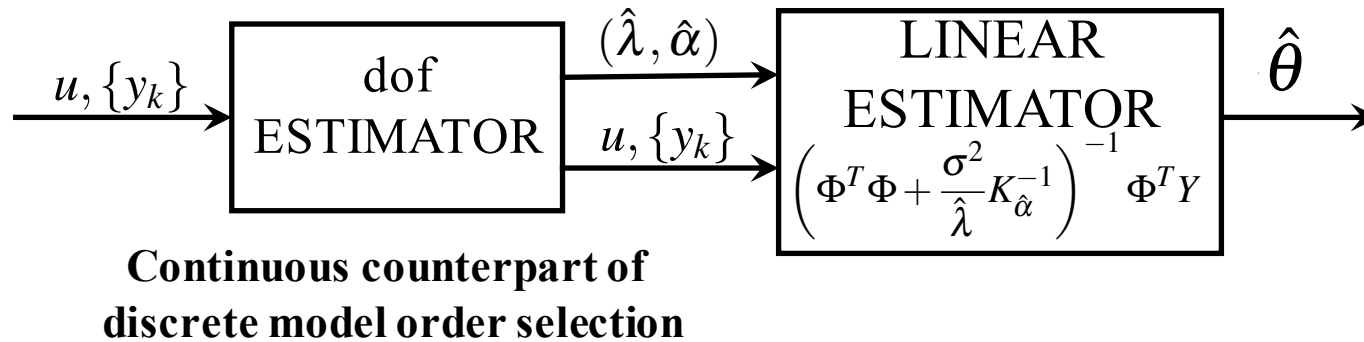
Stable spline estimator

Unknown complexity



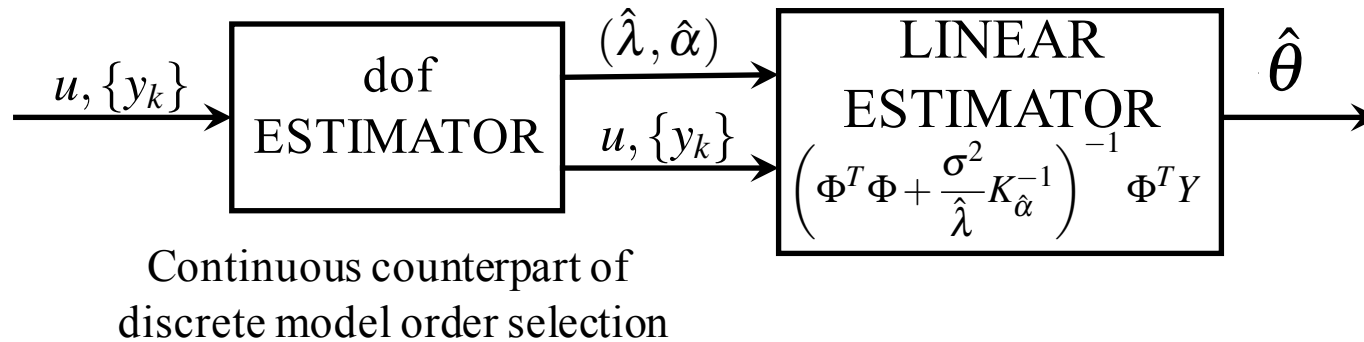
Stable spline estimator

Unknown complexity

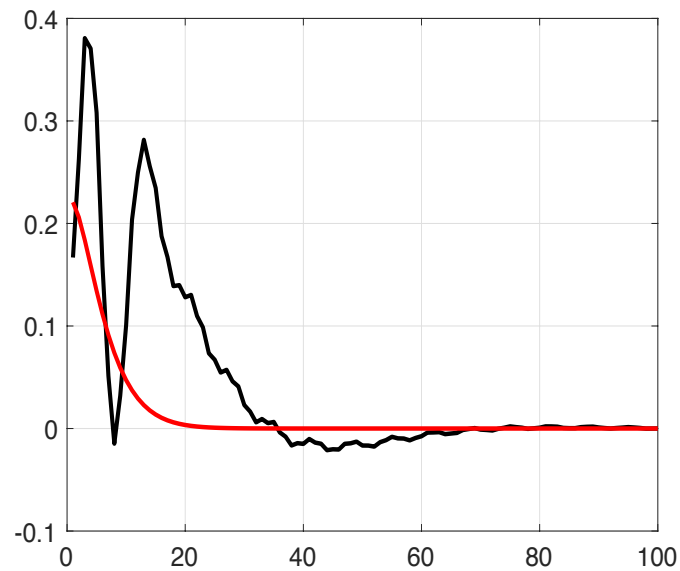


Stable spline estimator

Unknown complexity

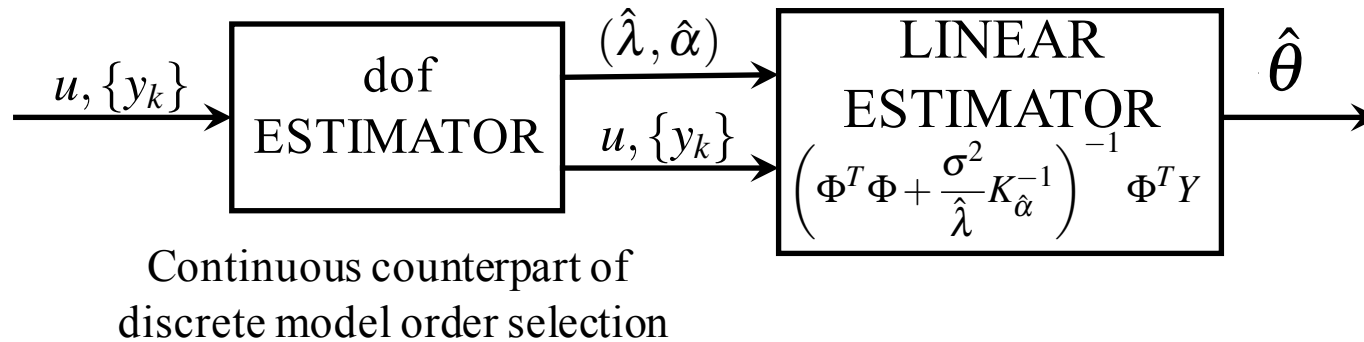


Stable spline estimate
with dof = 3

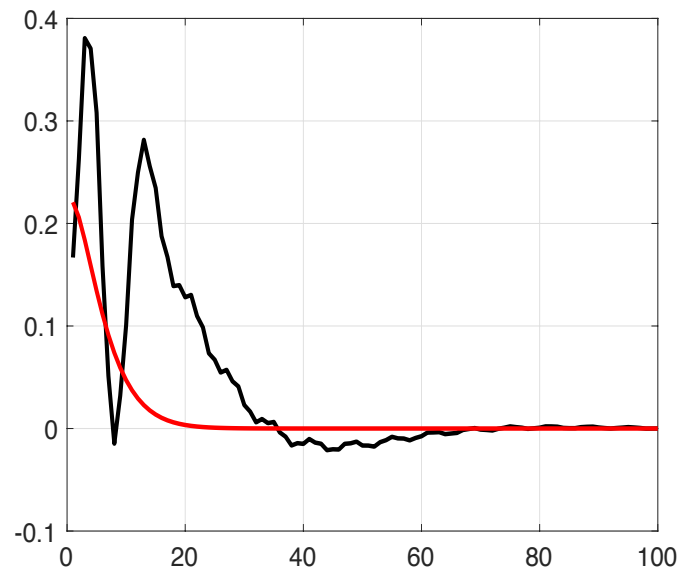


Stable spline estimator

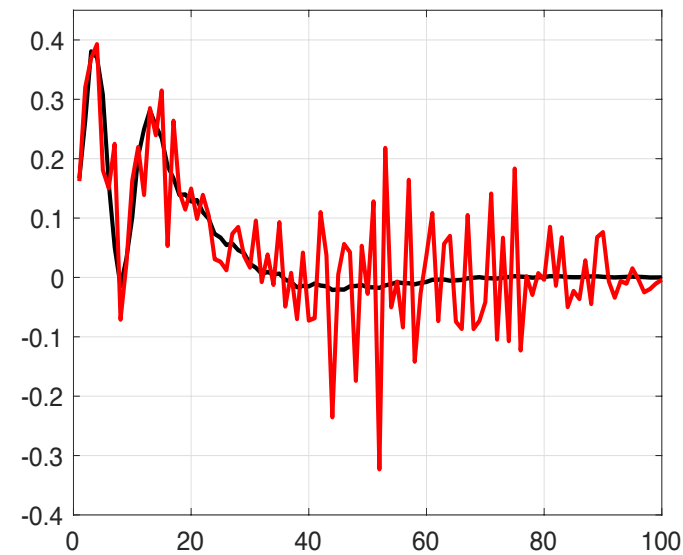
Unknown complexity



Stable spline estimate
with dof = 3

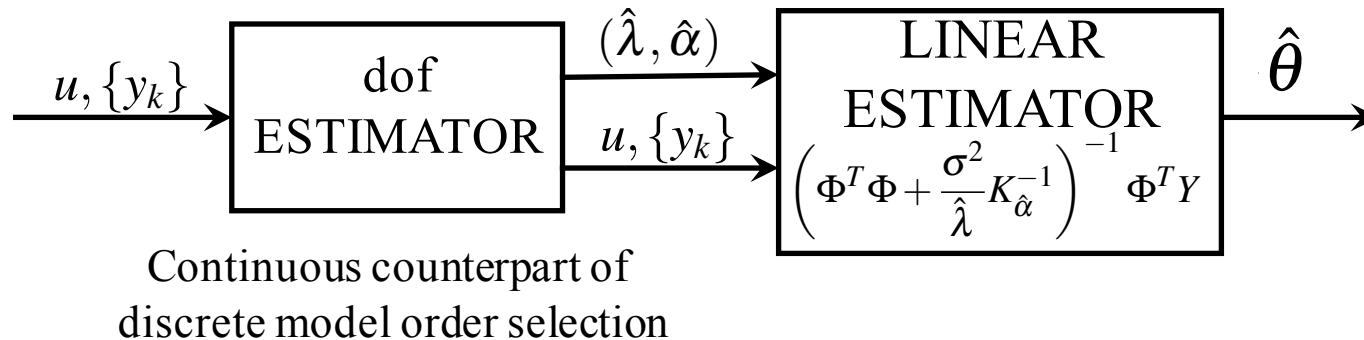


Stable spline estimate
with dof = 50

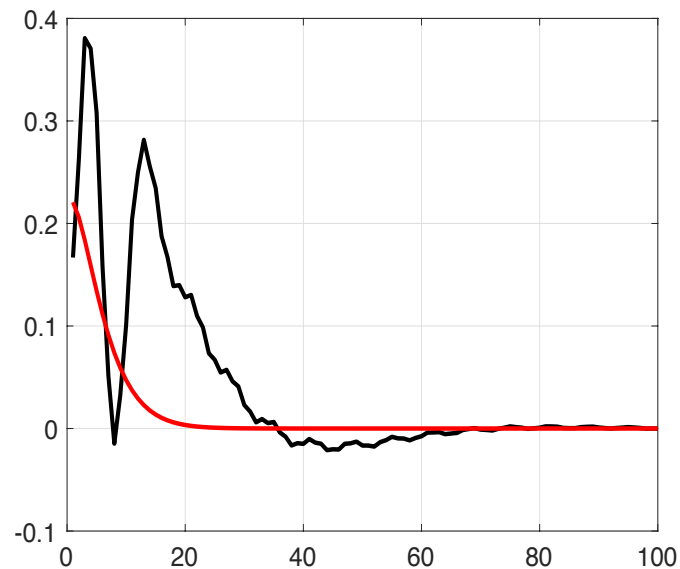


Stable spline estimator

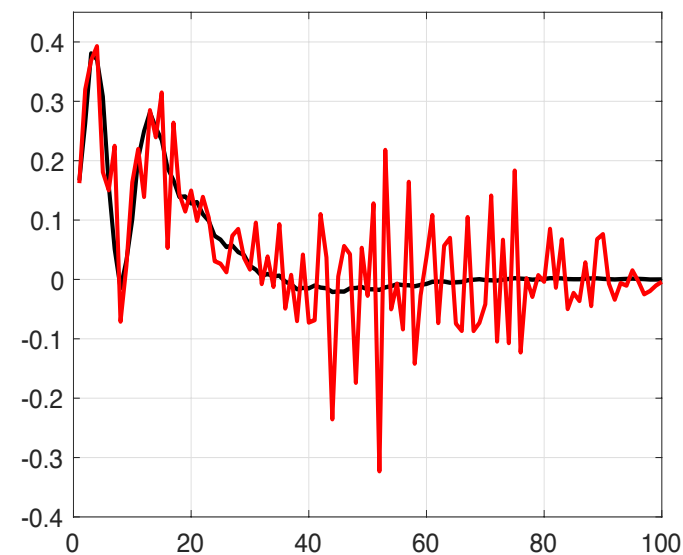
Unknown complexity



Stable spline estimate
with dof = 3



Stable spline estimate
with dof = 50



How to tune hyperparameters?

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes

(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Empirical Bayes features
(MacKay 1992; Kass and Raftery 1995)

- Connection with concepts of evidence and posterior model probability

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Empirical Bayes features
(MacKay 1992; Kass and Raftery 1995)

- Connection with concepts of evidence and posterior model probability
- Control on model complexity thanks to marginalization

$$\mathbf{p}(Y|\eta) \approx \begin{array}{c} \text{Best fit} \\ \text{likelihood} \end{array} \times \begin{array}{c} \text{Occam} \\ \text{factor} \end{array}$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Under linear assumptions

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_{\alpha}) \quad v \sim \mathcal{N}(0, \sigma^2 I) \quad \theta \perp v$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Under linear assumptions

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_{\alpha}) \quad v \sim \mathcal{N}(0, \sigma^2 I) \quad \theta \perp v$$

$$Y \sim \mathcal{N}(0, \Sigma_{\eta}), \quad \Sigma_{\eta} = \lambda \Phi K_{\alpha} \Phi^T + \sigma^2 I$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \text{Empirical Bayes estimate of } \eta$$

Under linear assumptions

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_{\alpha}) \quad v \sim \mathcal{N}(0, \sigma^2 I) \quad \theta \perp v$$

$$Y \sim \mathcal{N}(0, \Sigma_{\eta}), \quad \Sigma_{\eta} = \lambda \Phi K_{\alpha} \Phi^T + \sigma^2 I$$

$$\hat{\eta} = \arg \min_{\eta} Y^T \Sigma_{\eta}^{-1} Y + \log \det \Sigma_{\eta} \quad \text{Empirical Bayes for linear Gaussian model}$$

Tuning the regularization via marginal likelihood

η = hyperparameter vector, e.g. $\eta = [\lambda \ \alpha]$ for Stable Spline

Empirical Bayes
(Robbins 1956)

$$\mathbf{p}(Y|\eta) = \int \mathbf{p}(Y|\theta)\mathbf{p}(\theta|\eta)d\theta \quad \text{Marginal Likelihood (ML)}$$

$$\hat{\eta} = \arg \max_{\eta} \mathbf{p}(Y|\eta) \quad \begin{array}{l} \text{Empirical Bayes} \\ \text{estimate of } \eta \end{array}$$

Under linear assumptions

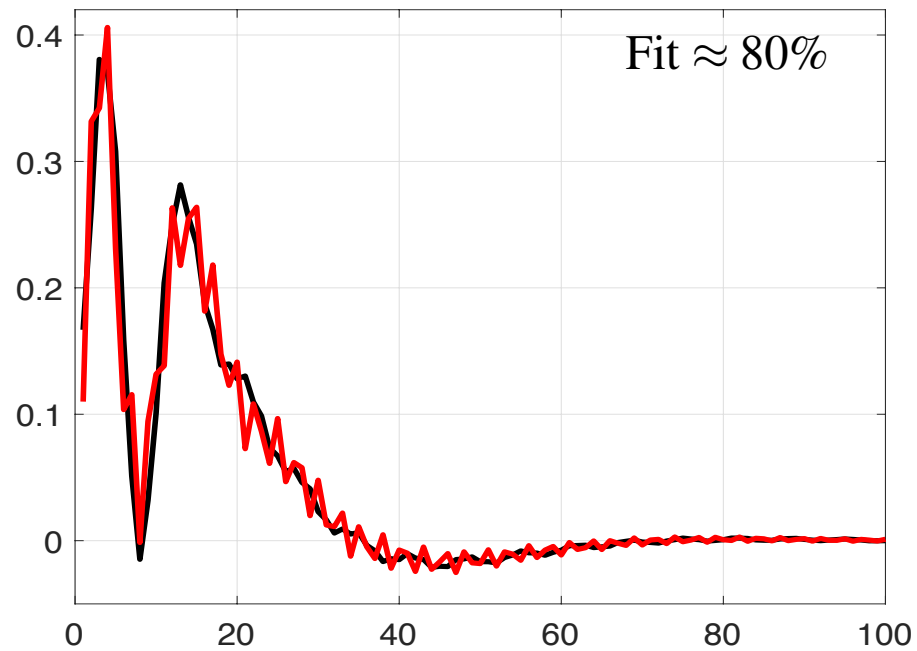
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_{\alpha}) \quad v \sim \mathcal{N}(0, \sigma^2 I) \quad \theta \perp v$$

$$Y \sim \mathcal{N}(0, \Sigma_{\eta}), \quad \Sigma_{\eta} = \lambda \Phi K_{\alpha} \Phi^T + \sigma^2 I$$

$$\hat{\eta} = \arg \min_{\eta} Y^T \Sigma_{\eta}^{-1} Y + \underbrace{\log \det \Sigma_{\eta}}_{\text{Occam factor}} \quad \begin{array}{l} \text{Empirical Bayes for} \\ \text{linear Gaussian model} \end{array}$$

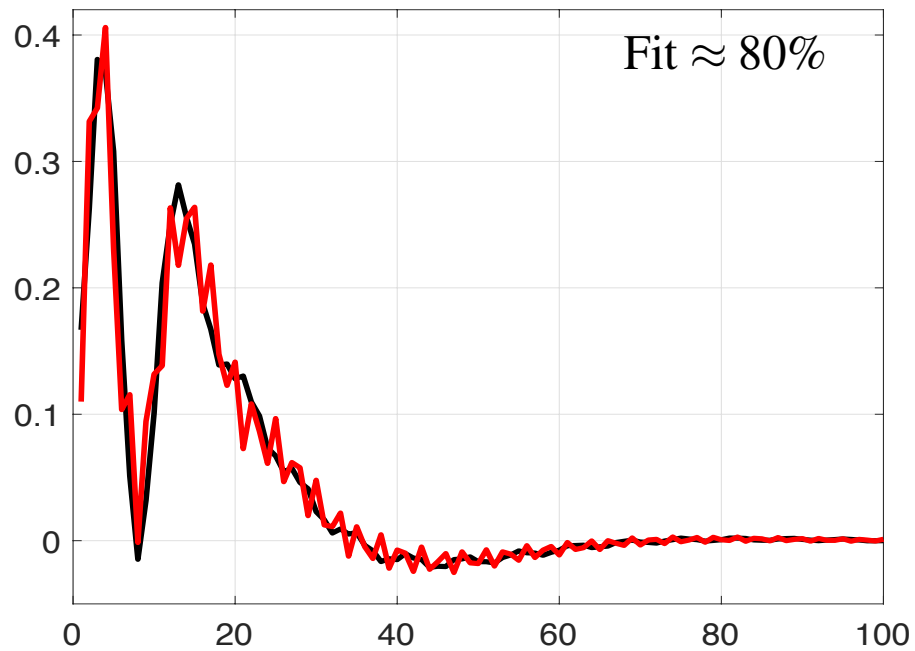
Coming back to the numerical experiment

PEM+Oracle (model order 5)

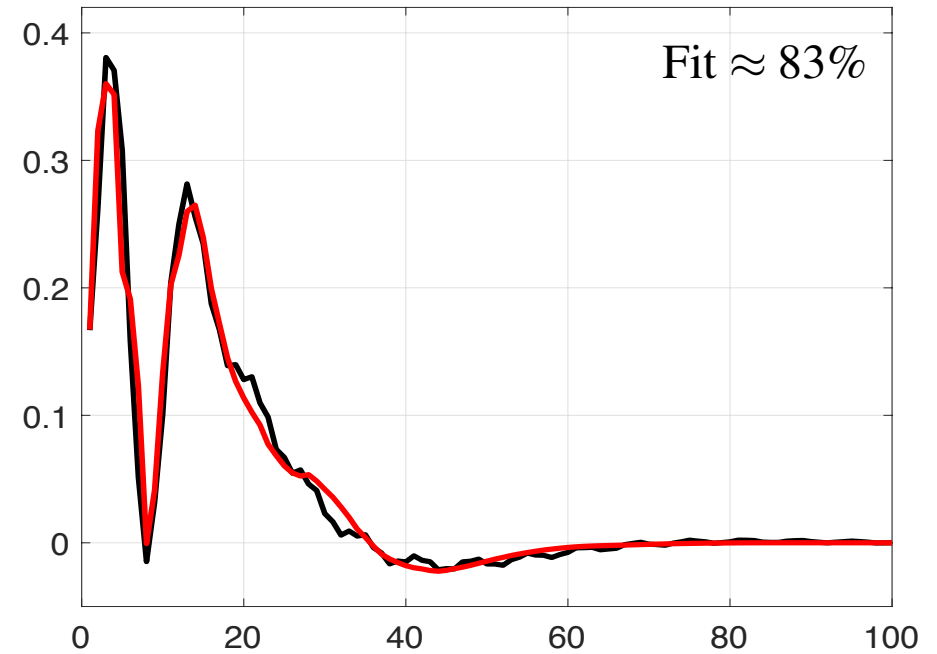


Coming back to the numerical experiment

PEM+Oracle (model order 5)

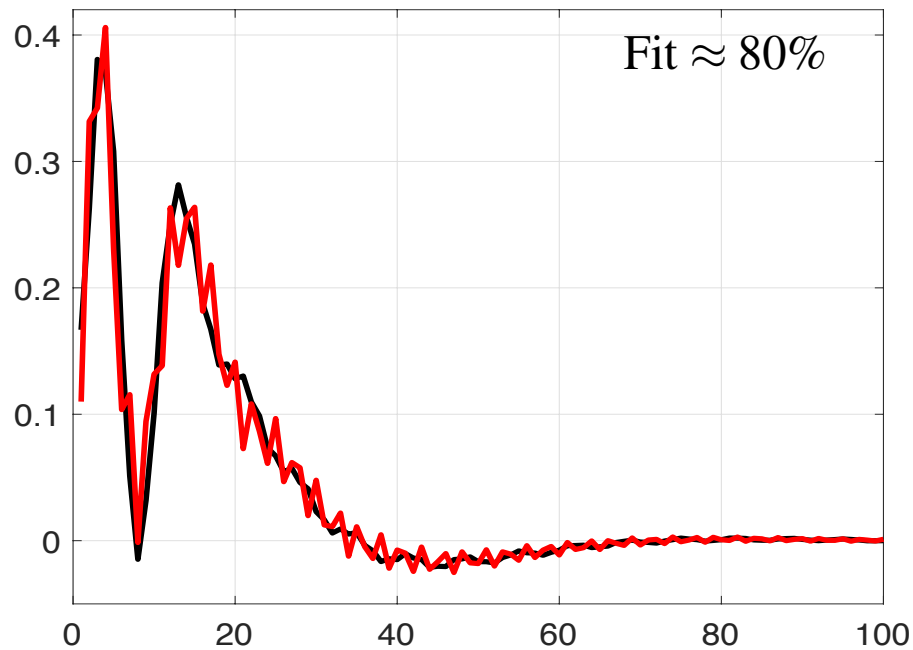


SS+ML (dof = 25.4)

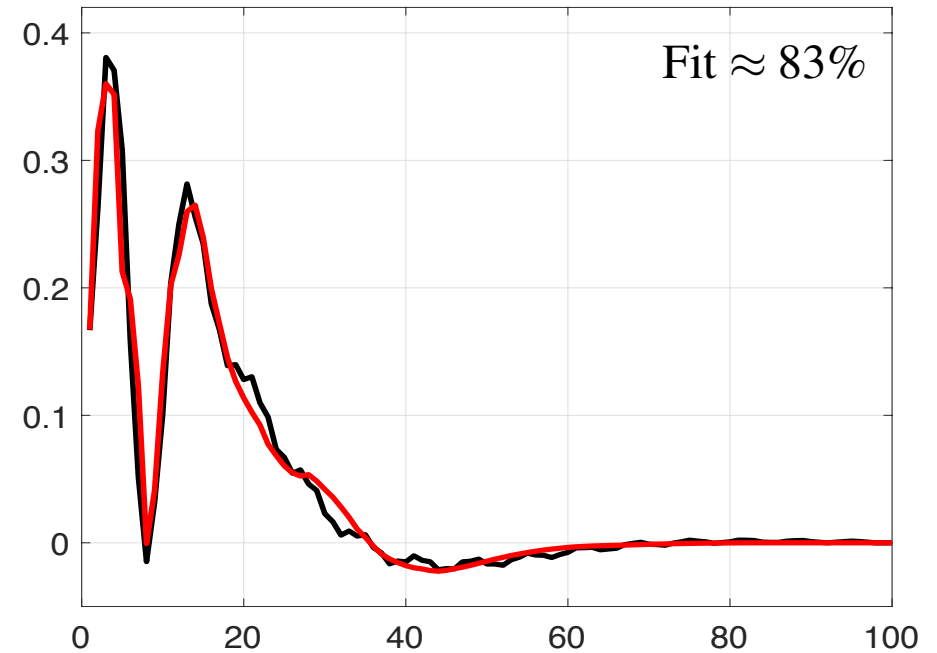


Coming back to the numerical experiment

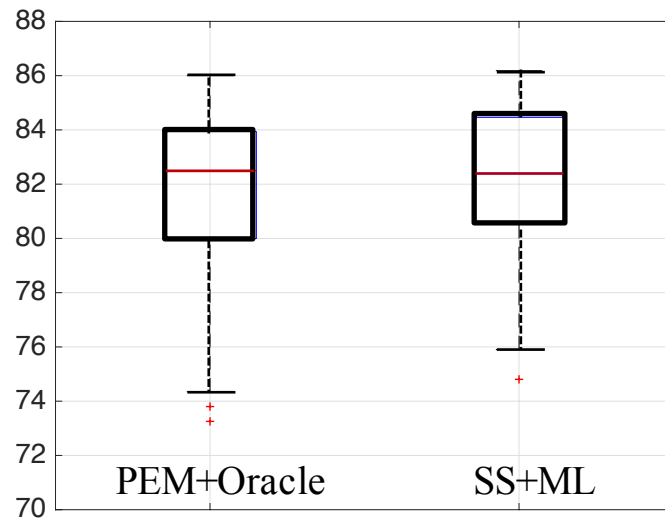
PEM+Oracle (model order 5)



SS+ML (dof = 25.4)



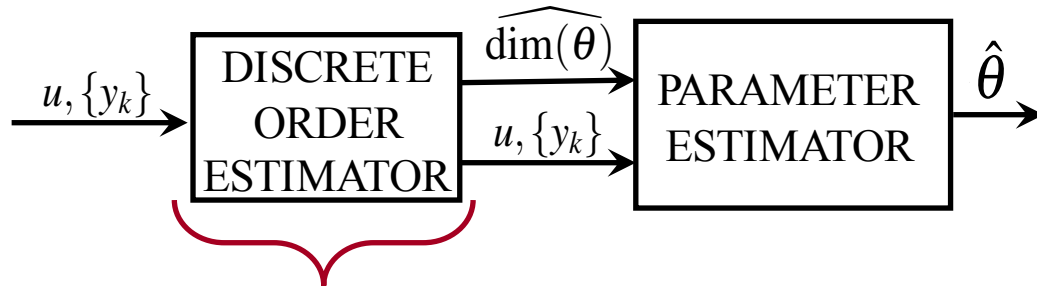
Fits after 300 Monte Carlo runs



Average fits:
SS+ML: 82.6%
PEM+Oracle: 79.4%

Why regularized may outperform classical?

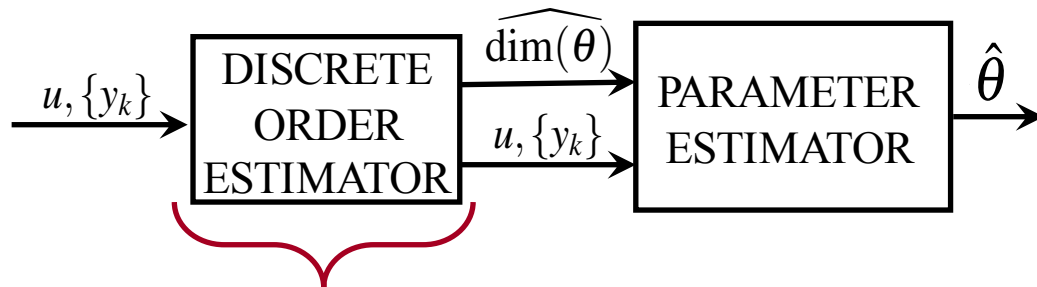
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

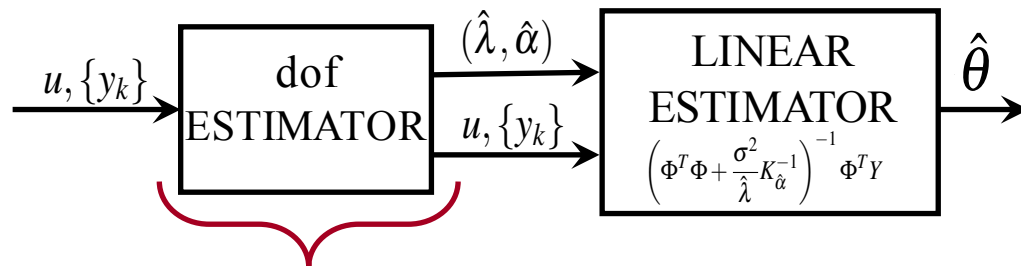
Why regularized may outperform classical?

Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

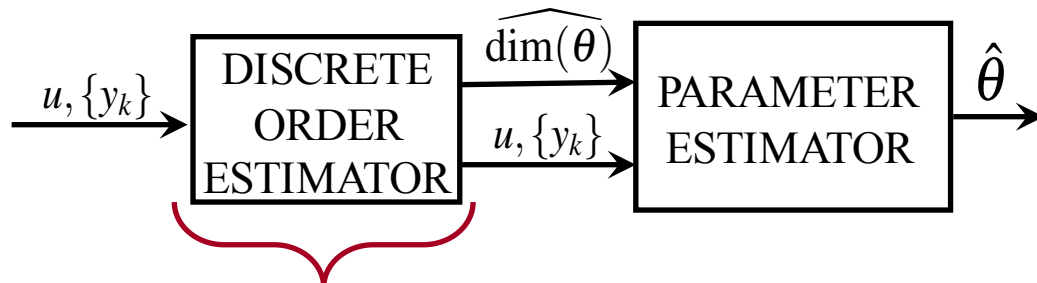
Regularized



$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

Why regularized may outperform classical?

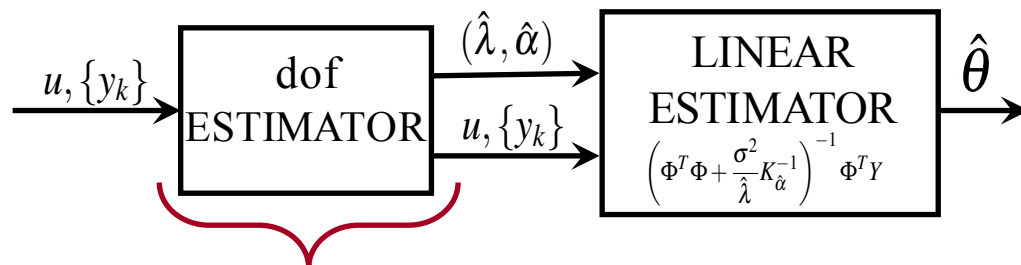
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve

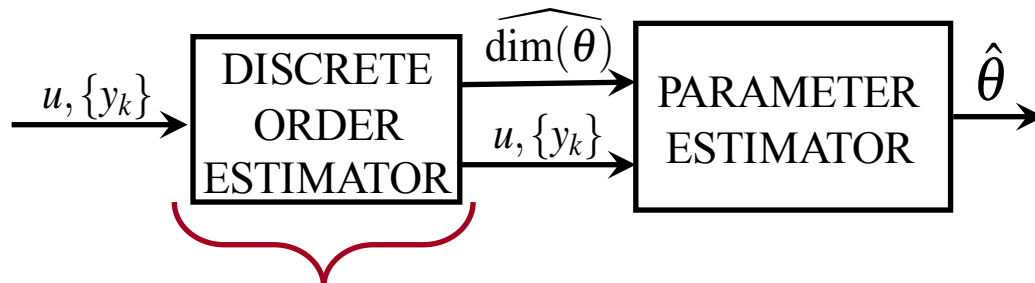
Regularized



$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

Why regularized may outperform classical?

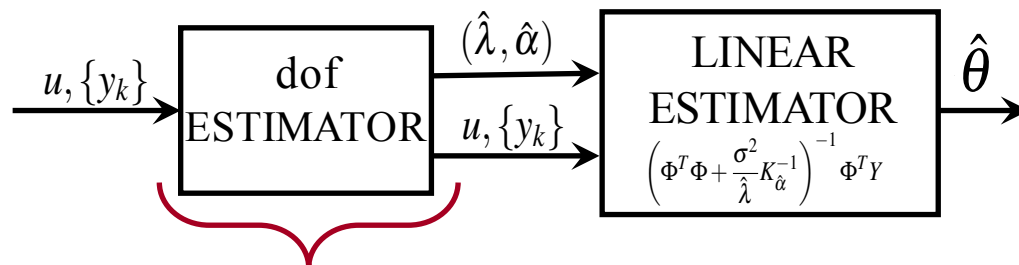
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve

Regularized

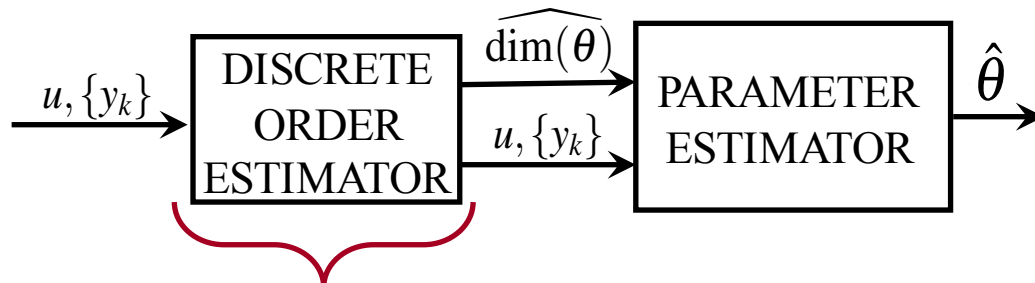


$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve

Why regularized may outperform classical?

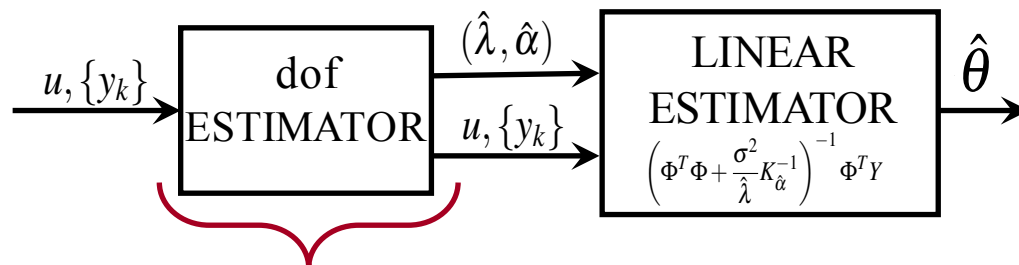
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve

Regularized



$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

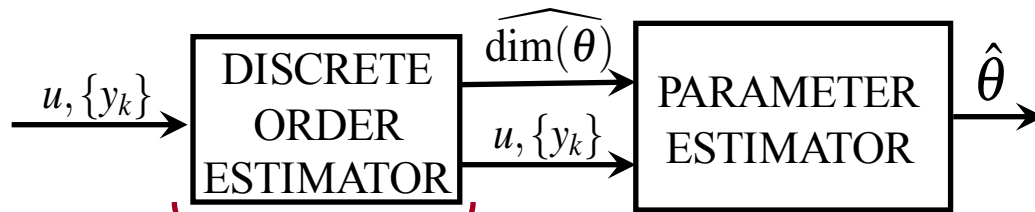
- just a two-dimensional problem to solve

Efficient/numerically stable algorithms recently included in the MATLAB SysId toolbox

T. Chen and L. Ljung Implementation of algorithms for tuning parameters in ReLS, Automatica, 2013

Why regularized may outperform classical?

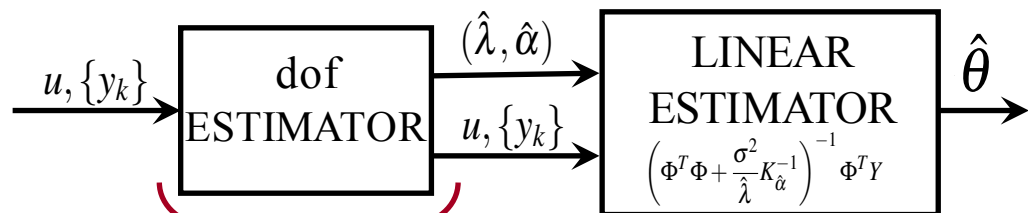
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve
- AIC/BIC using asymptotic likelihoods

Regularized

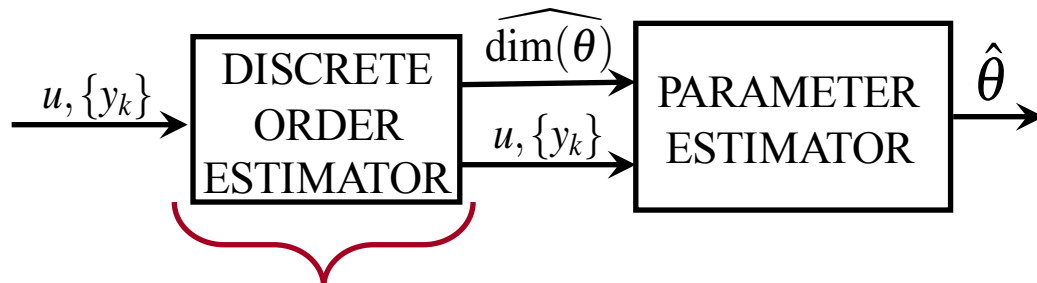


$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve

Why regularized may outperform classical?

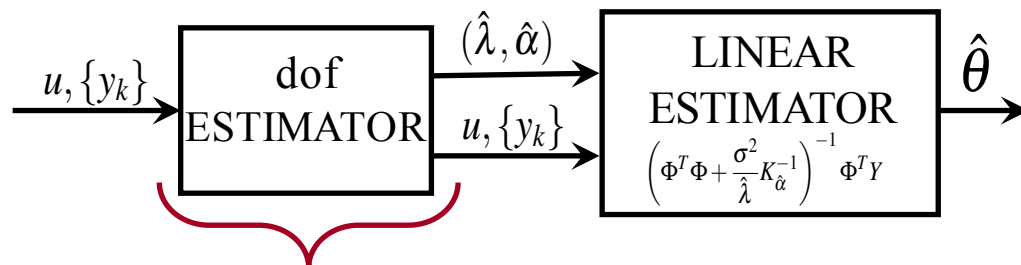
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve
- AIC/BIC using asymptotic likelihoods

Regularized

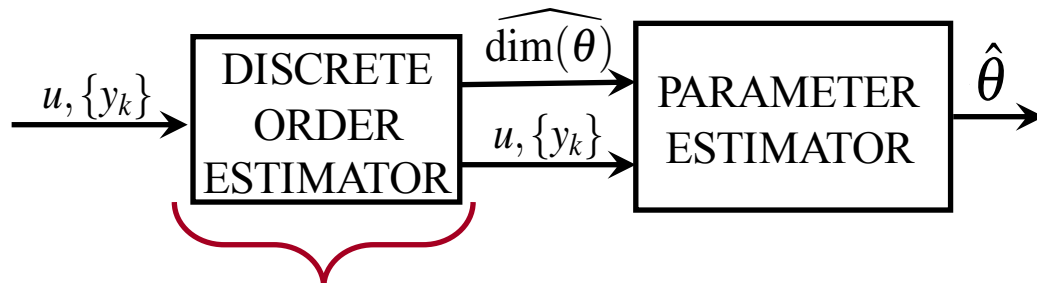


$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve
- exact (non asymptotic) likelihood

Why regularized may outperform classical?

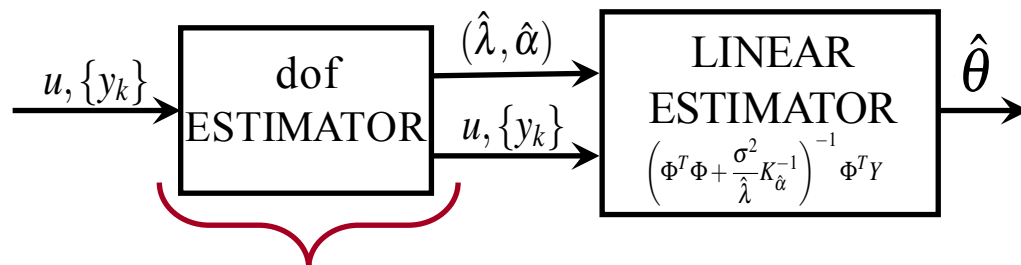
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve
- **AIC/BIC using asymptotic likelihoods**

Regularized



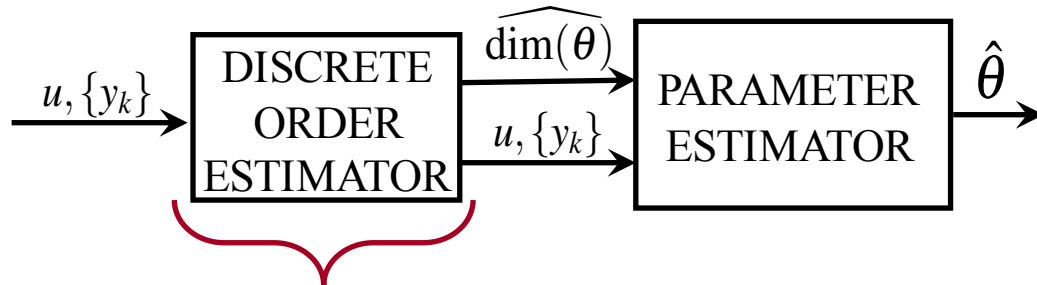
$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve
- **exact (non asymptotic) likelihood**
non-asymptotic “deterministic” ML features
(independent of prior correctness)
also recently derived:

G. Pillonetto and A. Chiuso Tuning complexity in kernel-based sysid: the robustness of ML, Automatica 2014

Why regularized may outperform classical?

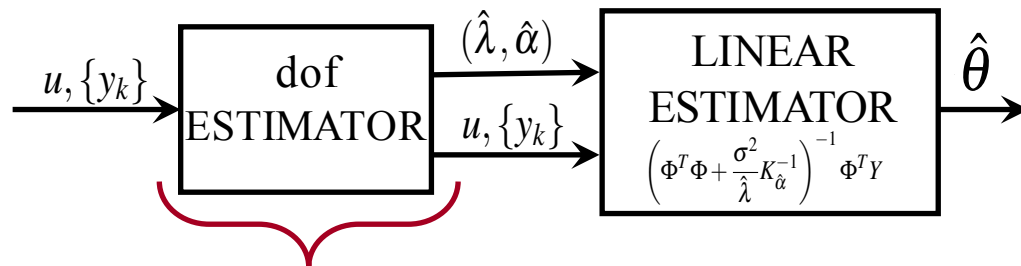
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve
- AIC/BIC using asymptotic likelihoods
- **discrete tuning**

Regularized

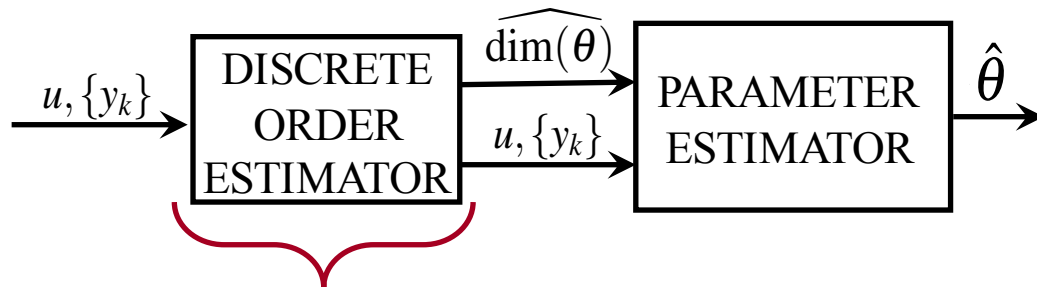


$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve
- exact (non asymptotic) likelihood

Why regularized may outperform classical?

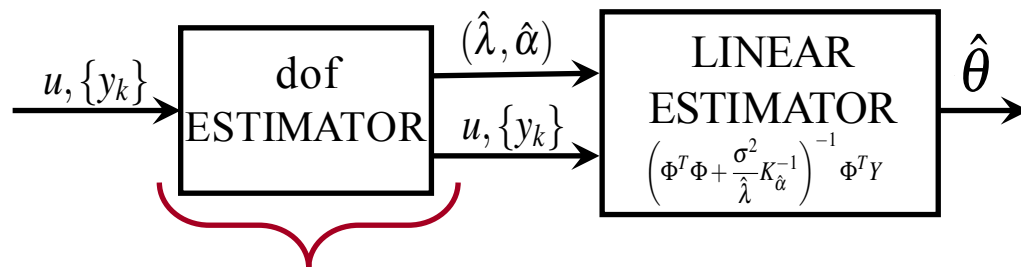
Classical



$$\mathcal{M}(\hat{\theta}) = \arg \min_{\mathcal{M}, \theta} \sum_k \frac{(y_k - \hat{y}_{\theta}(t_k))^2}{\sigma^2} + J(n) \dim \theta$$

- many (possibly) high-dimensional non-convex problems to solve
- AIC/BIC using asymptotic likelihoods
- discrete tuning

Regularized



$$(\hat{\lambda}, \hat{\alpha}) = \arg \max_{\lambda, \alpha} \mathbf{p}(Y|\lambda, \alpha)$$

- just a two-dimensional problem to solve
- exact (non asymptotic) likelihood
- continuous tuning possibly leading to better bias–variance trade off

Additional remarks

- No claim of superiority of *regularized* over *classical* here.
Main message is that stable spline can be a valuable alternative, especially for poor input excitation and small data sets

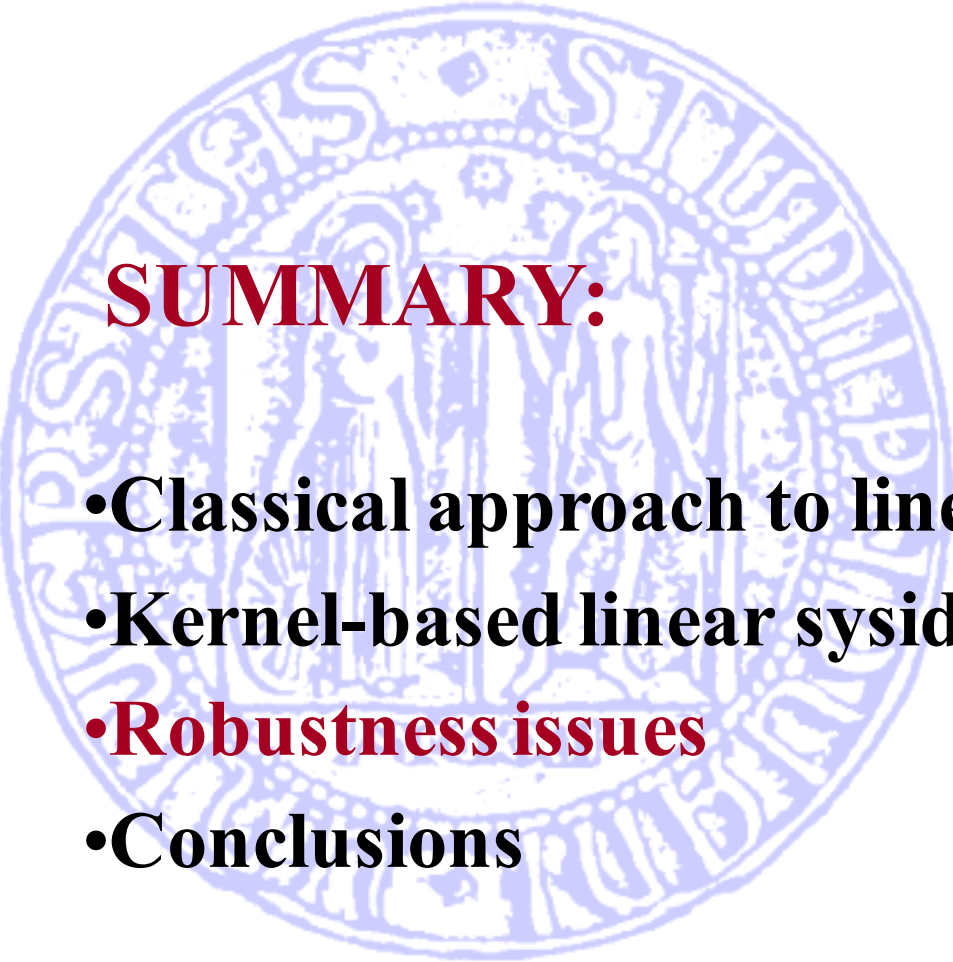
Additional remarks

- No claim of superiority of *regularized* over *classical* here.
Main message is that stable spline can be a valuable alternative, especially for poor input excitation and small data sets
- Possible synergy between *regularized* and *classical* should be also pursued, e.g. using regularization also as a tool for discrete model order selection
Pillonetto, G. , T. Chen and L. Ljung. Kernel-based model order selection for identification and prediction of linear dynamic systems, CDC 2013

Additional remarks

- No claim of superiority of *regularized* over *classical* here.
Main message is that stable spline can be a valuable alternative, especially for poor input excitation and small data sets
- Possible synergy between *regularized* and *classical* should be also pursued, e.g. using regularization also as a tool for discrete model order selection
Pillonetto, G. , T. Chen and L. Ljung. Kernel-based model order selection for identification and prediction of linear dynamic systems, CDC 2013
- Many other regularizers (including stable spline kernels evolutions) now available to further improve performance using e.g. mixtures of kernels, sparsity priors, Hankel nuclear norm

Z. Liu and L. Vandenberghe Interior-point method for nuclear norm approximation with applications to SysId, SIAM 2009
K. Mohan and M. Fazel Reweighted nuclear norm minimization with application to system identification, ACC 2010
H. Hjalmarsson, J. Welsh, C. Rojas Identification of Box-Jenkins models using structured ARX models and nuclear norm relaxation, Proc. of IFAC SysId 2012
B. Yilmaz, C. Lagoa , M. Sznajder An efficient atomic norm minimization approach to identify low order models, CDC 2013
C. Rojas, R. Toth, H. Hjalmarsson Sparse estimation of polynomials and rational dynamical models, IEEE TAC 2014
T. Chen, M. Andersen, L. Ljung, A. Chiuso, G. Pillonetto System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques, IEEE TAC 2014
G. Bottegal et al. Robust EM kernel-based methods for linear system identification, Automatica 2016
G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, L. Ljung Regularized linear SysId using atomic, nuclear and kernel-based norms: the role of the stability constraint, Automatica 2016
J. Lataire and T. Chen Transfer function and transient estimation by Gaussian regression in the frequency domain, Automatica 2016
G. Prando et al. Maximum entropy kernels for MIMO system identification, Automatica 2017
M. Darwish et al. The quest of the right kernel in Bayesian impulse response identification, Automatica 2018
...



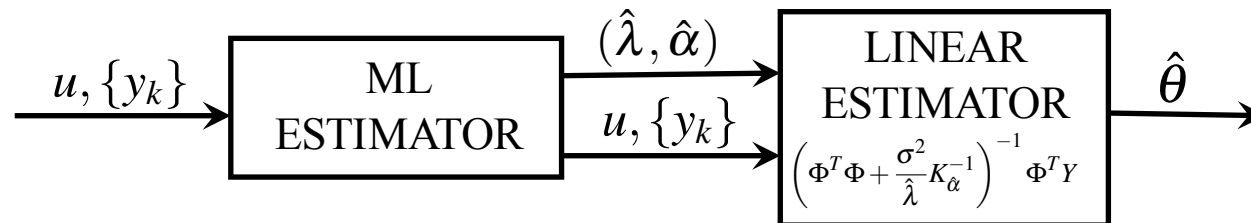
SUMMARY:

- **Classical approach to linear sysid**
- **Kernel-based linear sysid**
- **Robustness issues**
- **Conclusions**

Robustness issues

Model and estimator

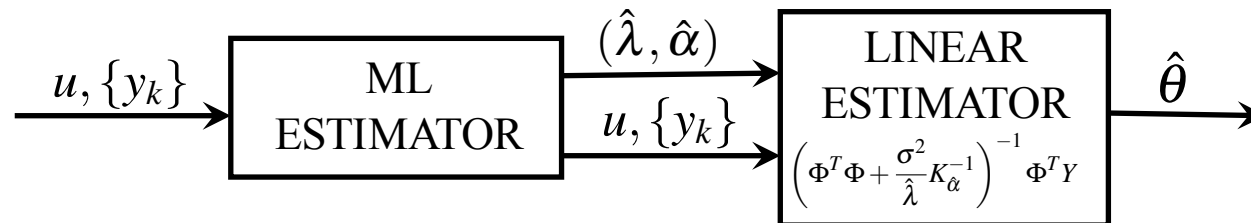
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$

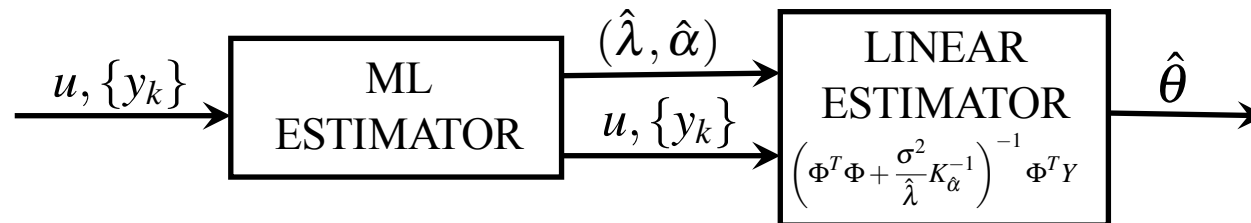


Is this model/estimator “robust”
and in which sense?

Robustness issues

Model and estimator

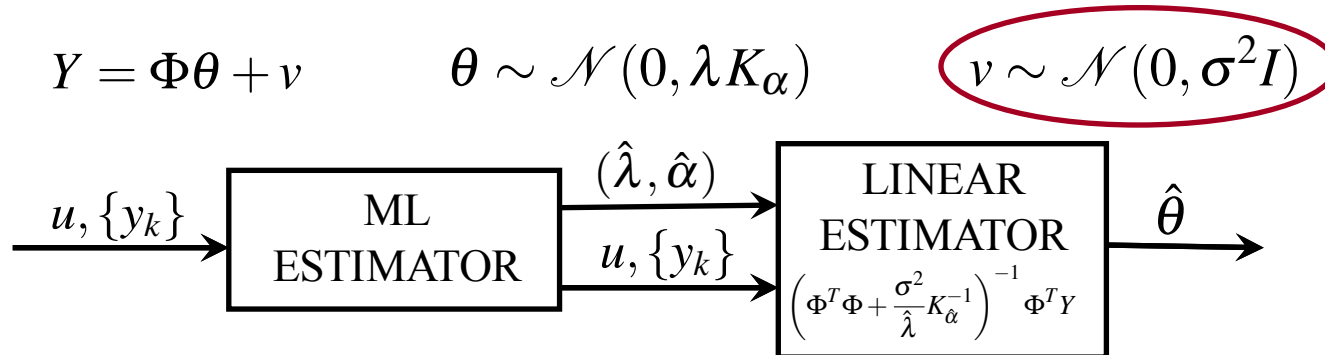
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Robustness w.r.t. uncorrect prior models

Robustness issues

Model and estimator



Robust estimation

(output outliers/non Gaussian noises)

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta$$

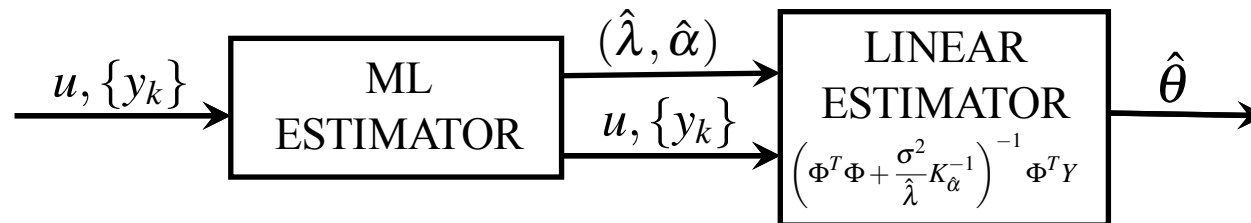
Robustness issues

Model and estimator

$$Y = \Phi\theta + v$$

$$\theta \sim \mathcal{N}(0, \lambda K_\alpha)$$

$$v \sim \mathcal{N}(0, \sigma^2 I)$$

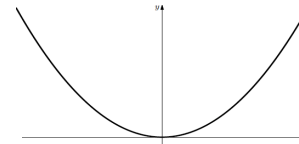


Robust estimation

(output outliers/non Gaussian noises)

$$\hat{\theta} = \arg \min_{\theta} \left(\|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta \right)$$

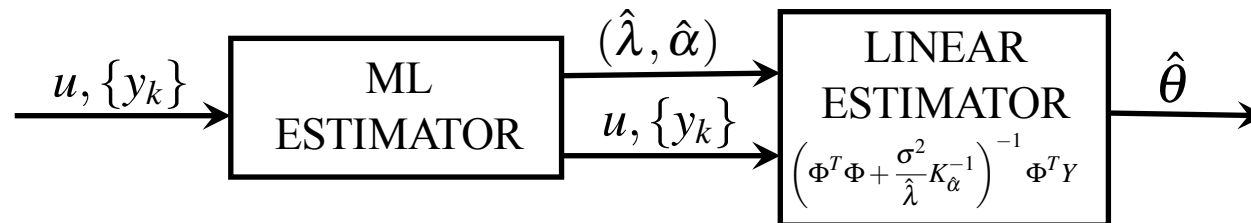
Quadratic loss
(vulnerable to outliers)



Robustness issues

Model and estimator

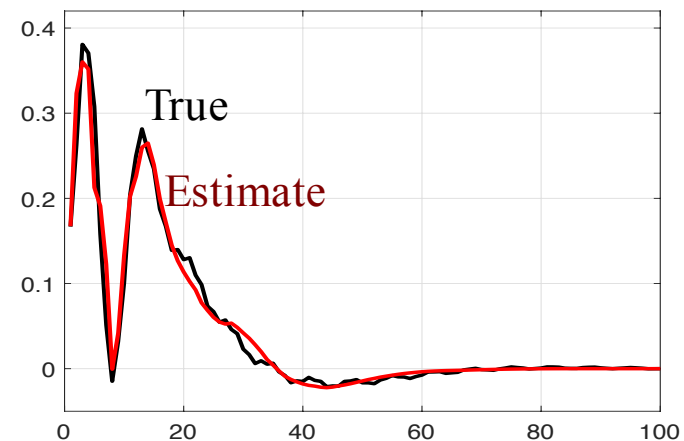
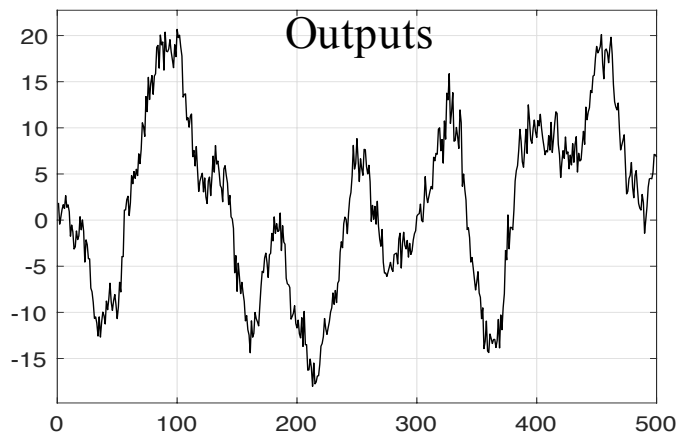
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Robust estimation (output outliers/non Gaussian noises)

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta$$

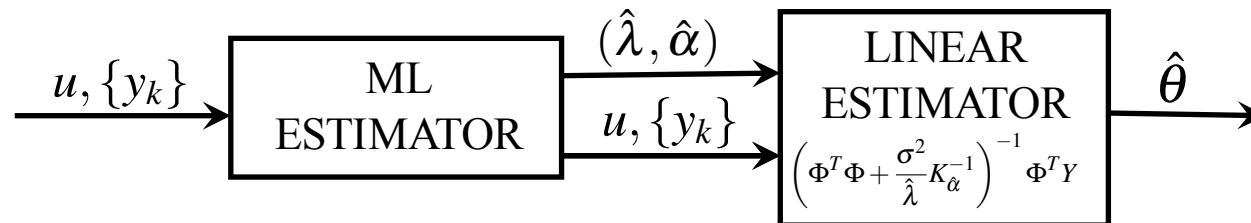
Numerical example without outliers



Robustness issues

Model and estimator

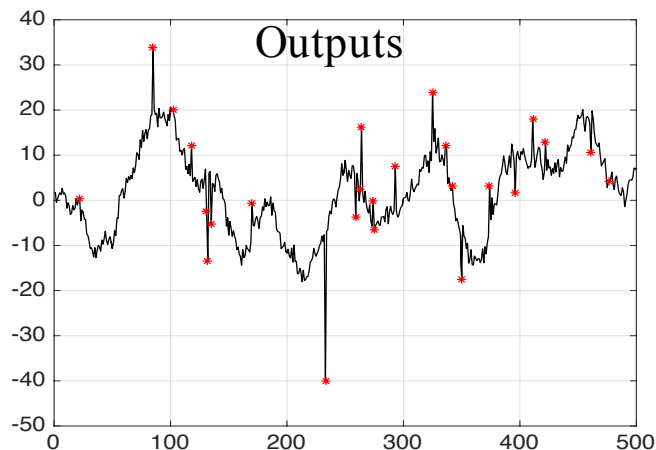
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Robust estimation (output outliers/non Gaussian noises)

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta$$

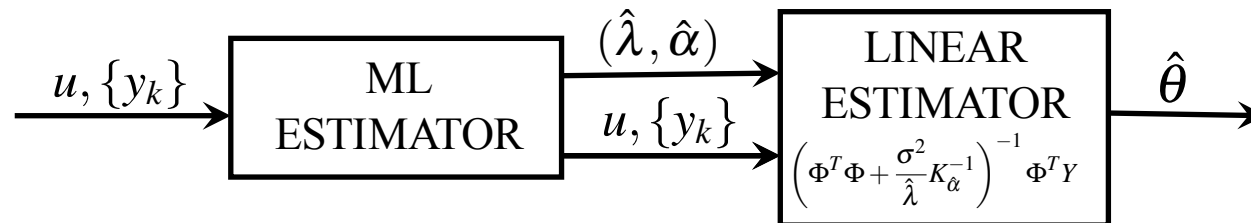
Numerical example with outliers



Robustness issues

Model and estimator

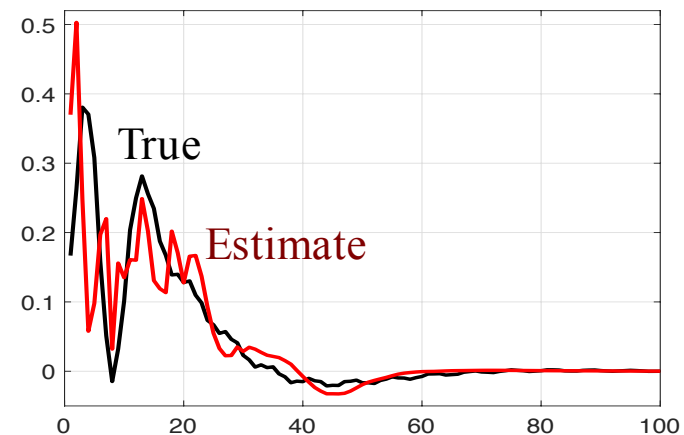
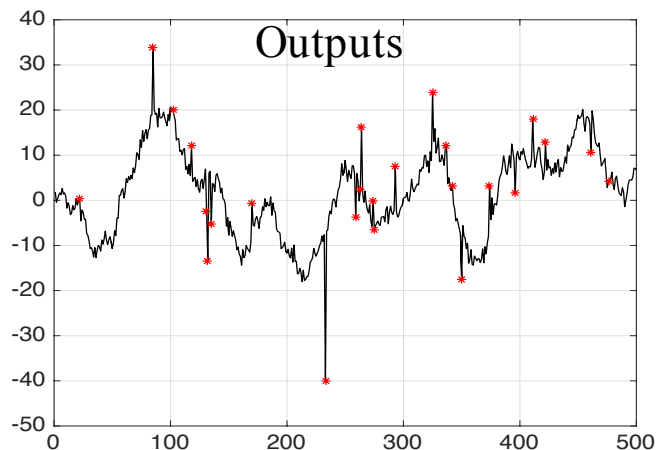
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Robust estimation (output outliers/non Gaussian noises)

$$\hat{\theta} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 + \gamma \theta^T K^{-1} \theta$$

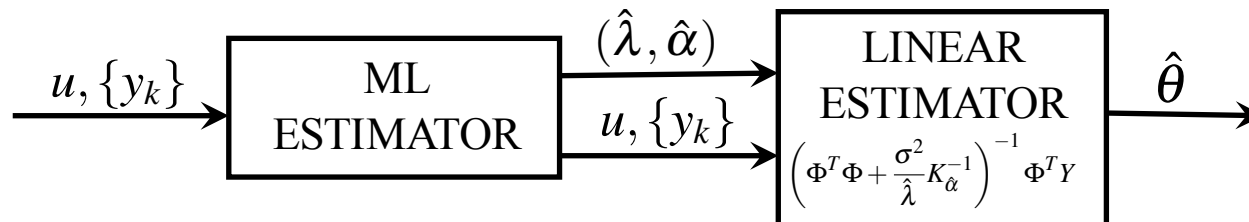
Numerical example with outliers



Robustness issues

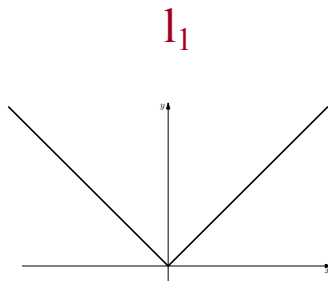
Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



More robust convex losses

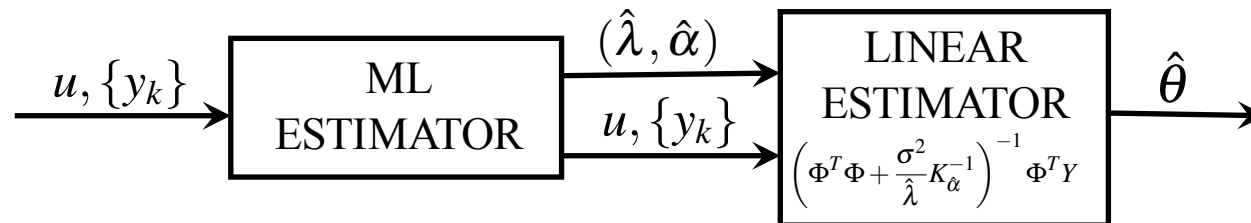
(Huber 1964; Drucker, Burgers, Kaufman, Smola, Vapnik 1997; Hastie et al 2009)



Robustness issues

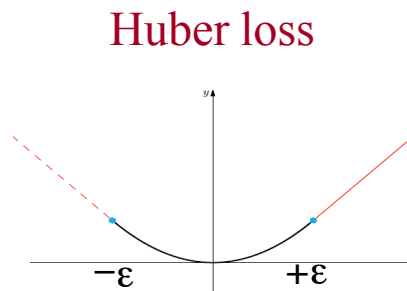
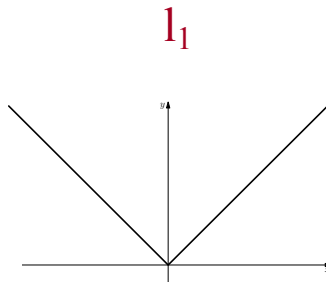
Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



More robust convex losses

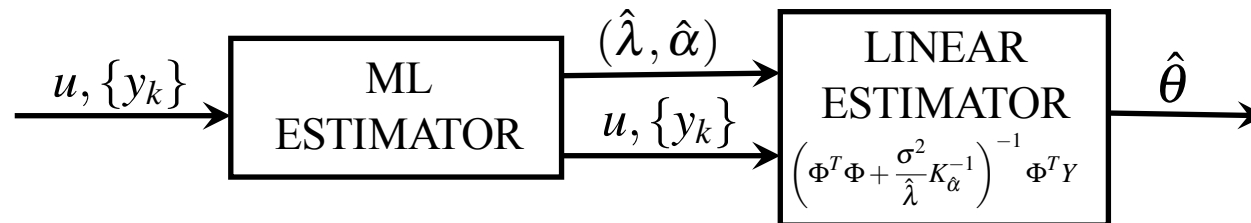
(Huber 1964; Drucker, Burgers, Kaufman, Smola, Vapnik 1997; Hastie et al 2009)



Robustness issues

Model and estimator

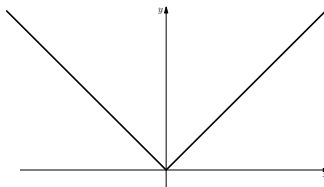
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



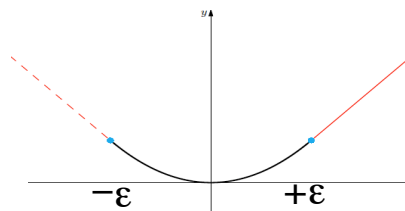
More robust convex losses

(Huber 1964; Drucker, Burgers, Kaufman, Smola, Vapnik 1997; Hastie et al 2009)

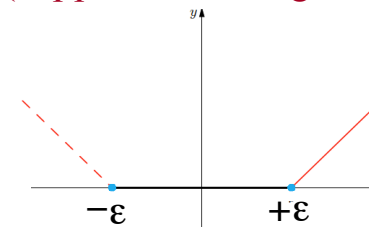
l_1



Huber loss



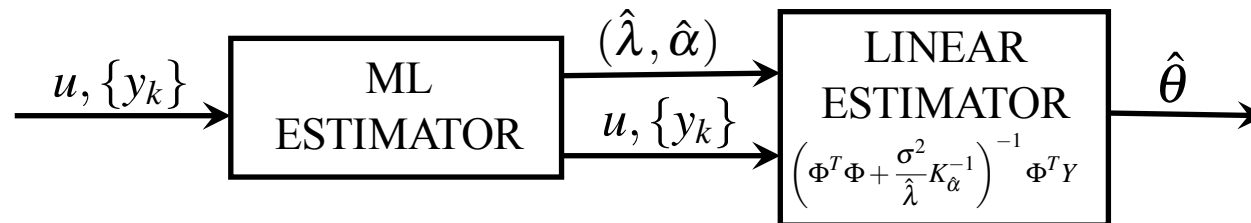
Vapnik loss
(support vector regression)



Robustness issues

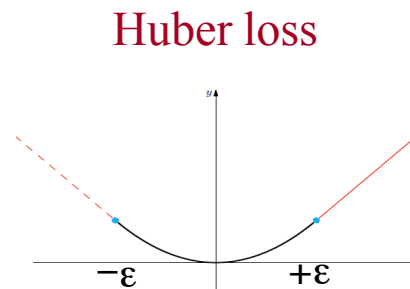
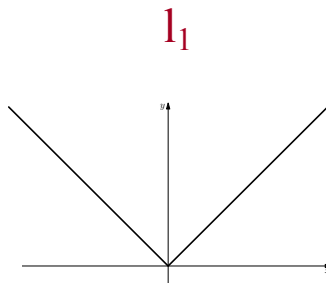
Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$

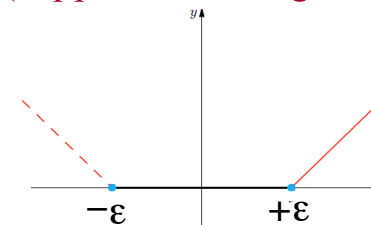


More robust convex losses

(Huber 1964; Drucker, Burgers, Kaufman, Smola, Vapnik 1997; Hastie et al 2009)



Vapnik loss
(support vector regression)

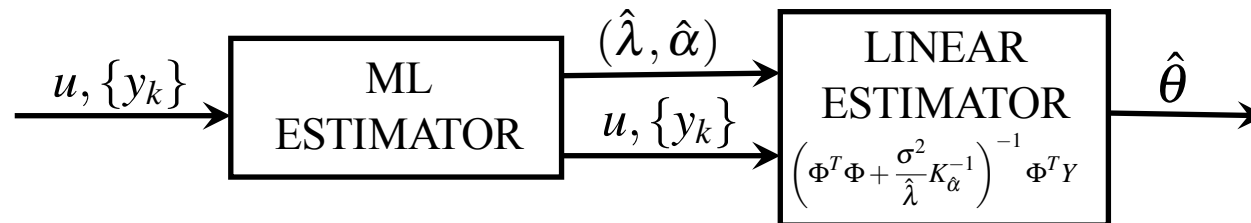


Different shapes inducing
different estimation features

Robustness issues

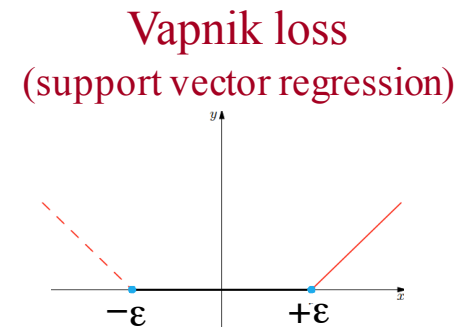
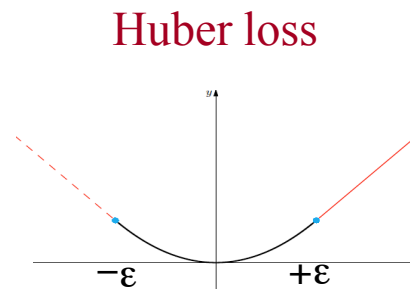
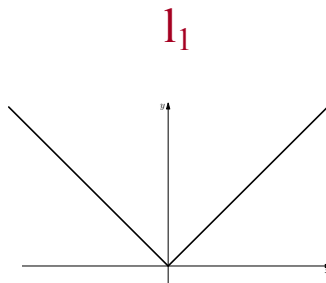
Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



More robust convex losses

(Huber 1964; Drucker, Burgers, Kaufman, Smola, Vapnik 1997; Hastie et al 2009)

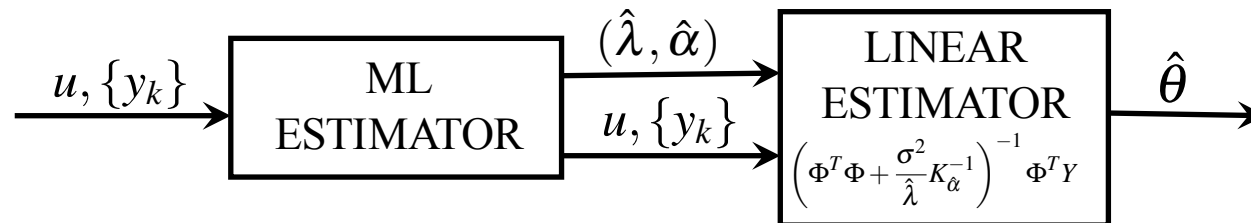


Different shapes inducing
different estimation features
but a **common linear trend against outliers**

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



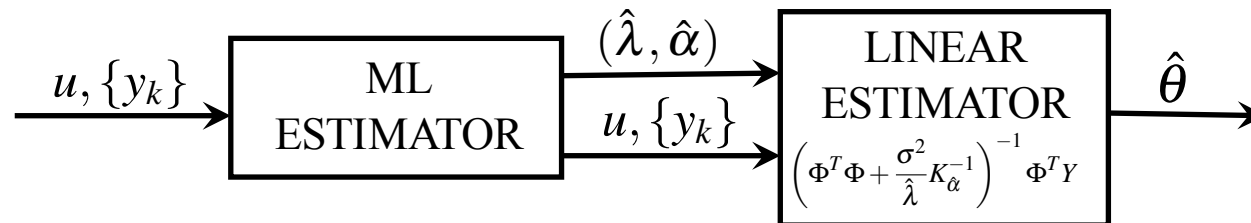
A computational framework
for robust linear system identification
(Aravkin, Burke, Pillonetto 2014)

$$\hat{\theta} = \arg \min_{\theta} \mathcal{V}(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework
for robust linear system identification
(Aravkin, Burke, Pillonetto 2014)

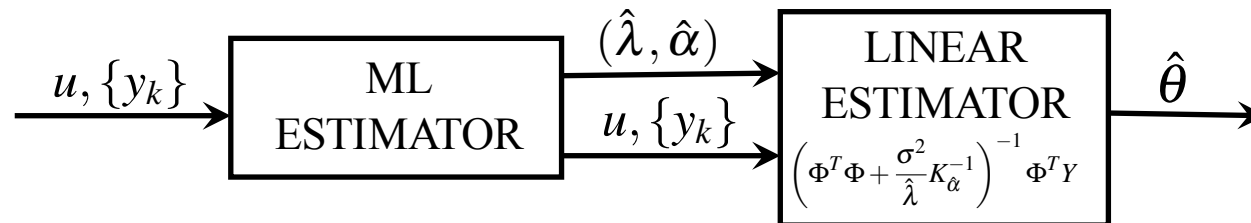
$$\hat{\theta} = \arg \min_{\theta} \mathcal{V}(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

convex piecewise linear quadratic loss,
e.g. l_1 , Huber, Vapnik,...

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework
for robust linear system identification
(Aravkin, Burke, Pillonetto 2014)

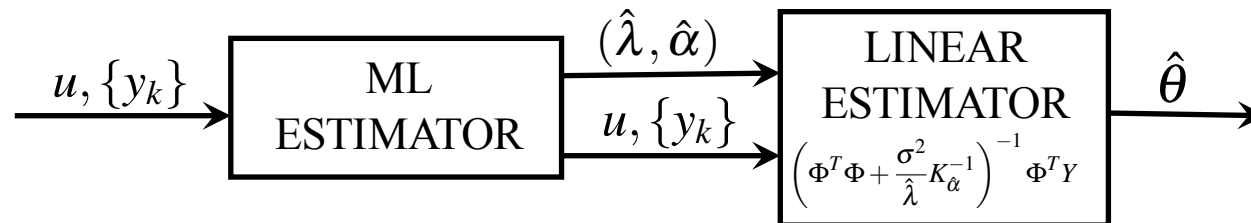
$$\hat{\theta} = \arg \min_{\theta} \mathcal{V}(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

convex piecewise linear quadratic loss:
all problems solvable with $O(d^2n)$ cost

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework
for robust linear system identification
(Aravkin, Burke, Pillonetto 2014)

$$\hat{\theta} = \arg \min_{\theta} \underbrace{\mathcal{V}(Y - \Phi\theta)}_{\text{convex piecewise linear quadratic loss}} + \gamma \theta^T K^{-1} \theta$$

convex piecewise linear quadratic loss:
all problems solvable with $O(d^2n)$ cost

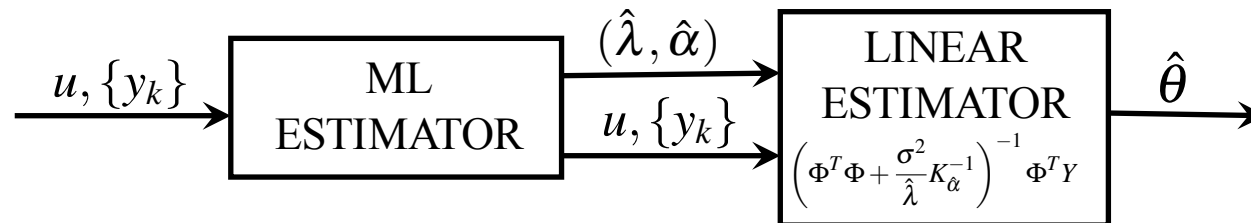
Toolbox for robust kernel-based linear
system identification available:

<https://github.com/saravkin/IPsolve>

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework for robust linear system identification (Aravkin, Burke, Pillonetto 2014)

$$\hat{\theta} = \arg \min_{\theta} \mathcal{V}(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

convex piecewise linear quadratic loss:
all problems solvable with $O(d^2n)$ cost

Toolbox for robust kernel-based linear
system identification available:

<https://github.com/saravkin/IPsolve>

For a survey linking robust estimation,
convex optimization and dynamic systems, see

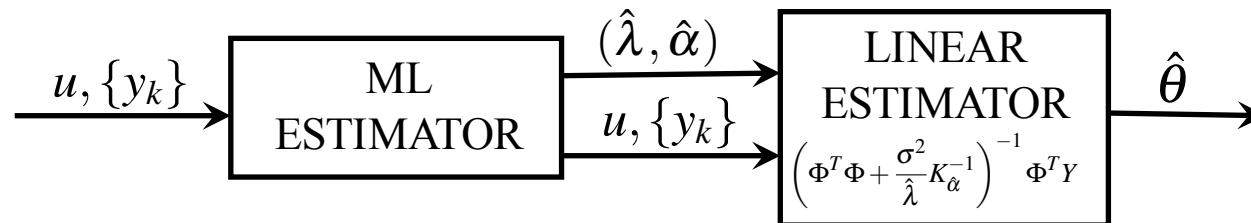
A. Aravkin, J.V. Burke, L. Ljung, A. Lozano, G. Pillonetto

Generalized Kalman smoothing: modeling and algorithms, Automatica 2017

Robustness issues

Model and estimator

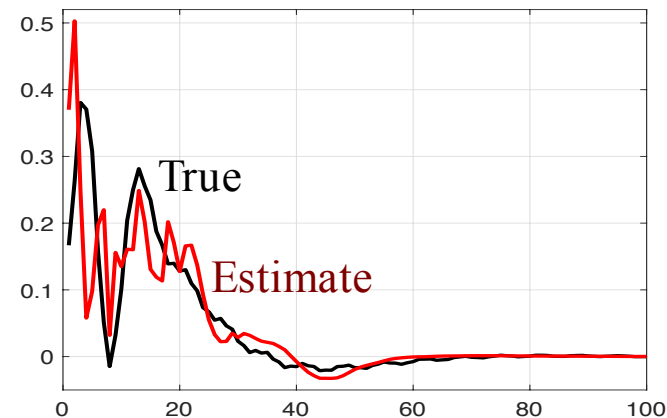
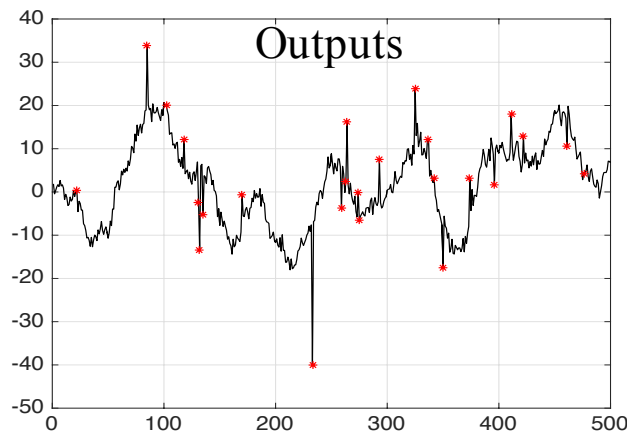
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework for robust linear system identification

$$\hat{\theta} = \arg \min_{\theta} \psi(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

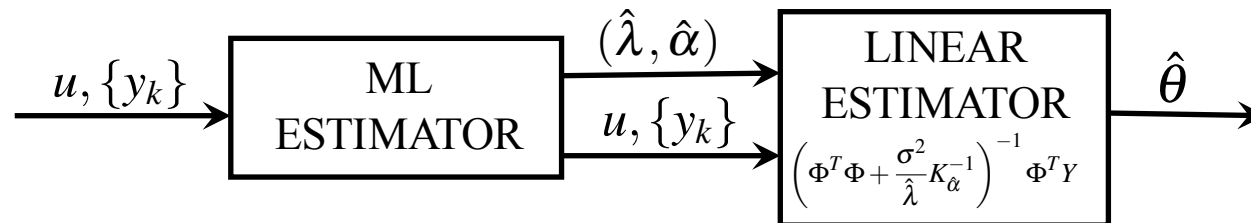
Estimation with quadratic loss 



Robustness issues

Model and estimator

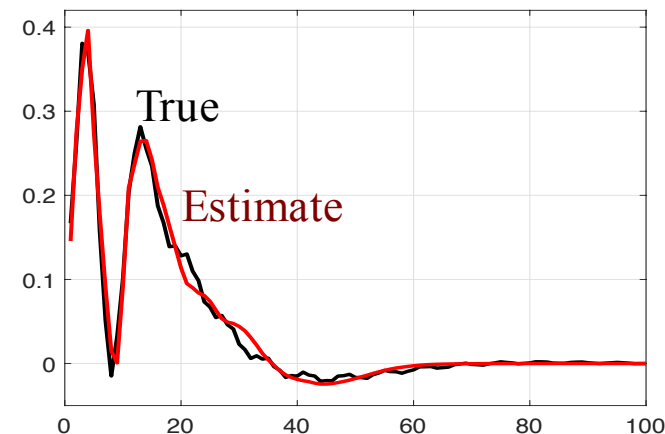
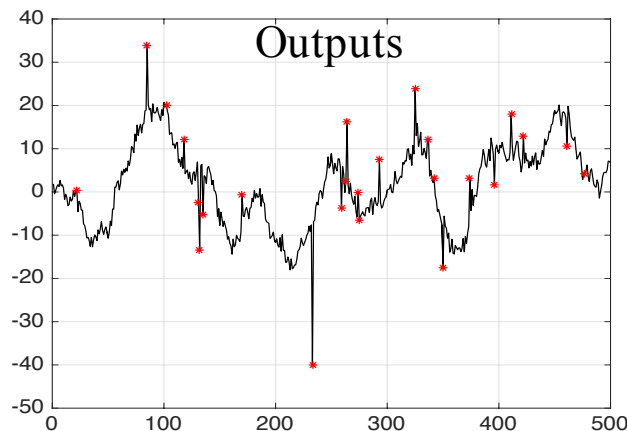
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



A computational framework for robust linear system identification

$$\hat{\theta} = \arg \min_{\theta} \psi(Y - \Phi\theta) + \gamma \theta^T K^{-1} \theta$$

Estimation with l_1 loss

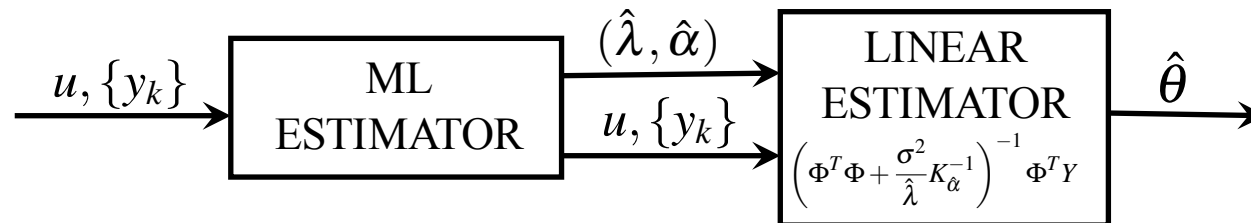


<https://github.com/saravkin/IPsolve>

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$

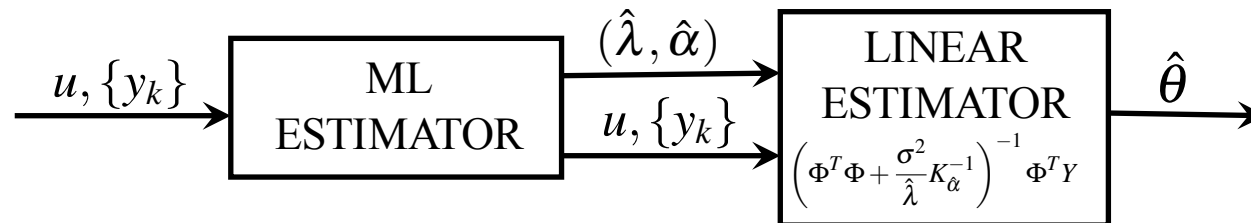


Robust system identification:
derivation of impulse response confidence intervals

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



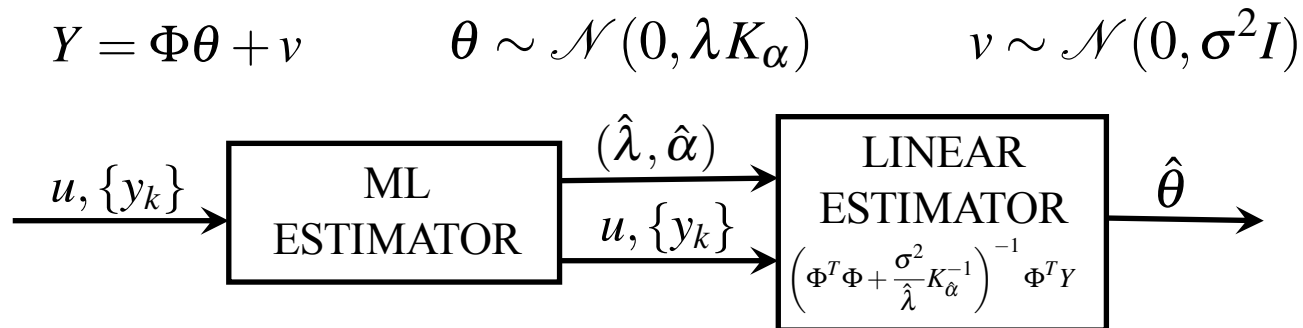
Robust system identification: derivation of impulse response confidence intervals

Model error modeling, stochastic embedding,
 H_∞ identification, worst-case set membership

(Goodwin et al., 1992; Milanese et al. 1996; Hakvoort and Van Den Hof 1997;
Garulli et al. 2000; Lecchini and Gevers 2004; Ljung et al. 2014)

Robustness issues

Model and estimator



Robust system identification: derivation of impulse response confidence intervals

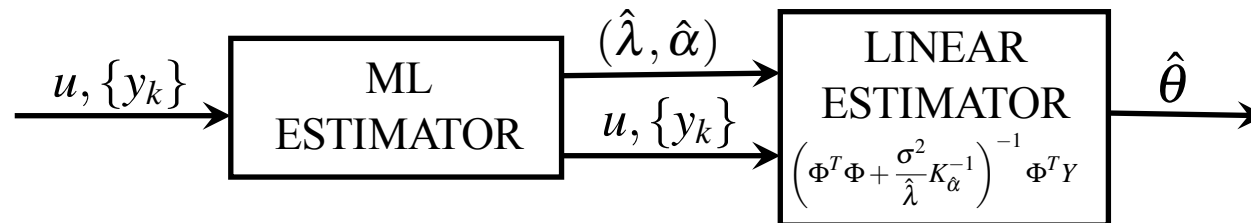
Model error modeling, stochastic embedding,
 H_∞ identification, worst-case set membership
(Goodwin et al., 1992; Milanese et al. 1996; Hakvoort and Van Den Hof 1997;
Garulli et al. 2000; Lecchini and Gevers 2004; Ljung et al. 2014)

SysId sample complexity, non asymptotic bounds,
interval predictor models, sign-perturbed sums (SPS)
(Harrison et al. 1996; Goldenshluger 1998; Smale and Zhou 2007;
Vidyasagar and Karandikar 2008; Campi, Calafiore and Garatti 2009;
Csaji, Campi and Weyer 2015; Carè et al. 2015; Hang et al. 2016)

Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Bayesian solution: posterior regions

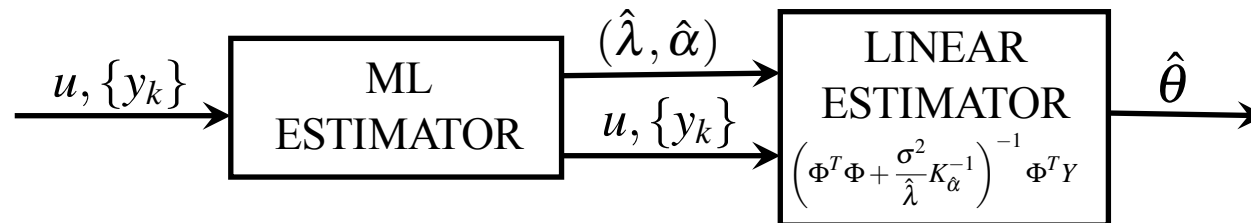
Gaussian posterior
(hyperparameters set to their estimates)

$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$

Robustness issues

Model and estimator

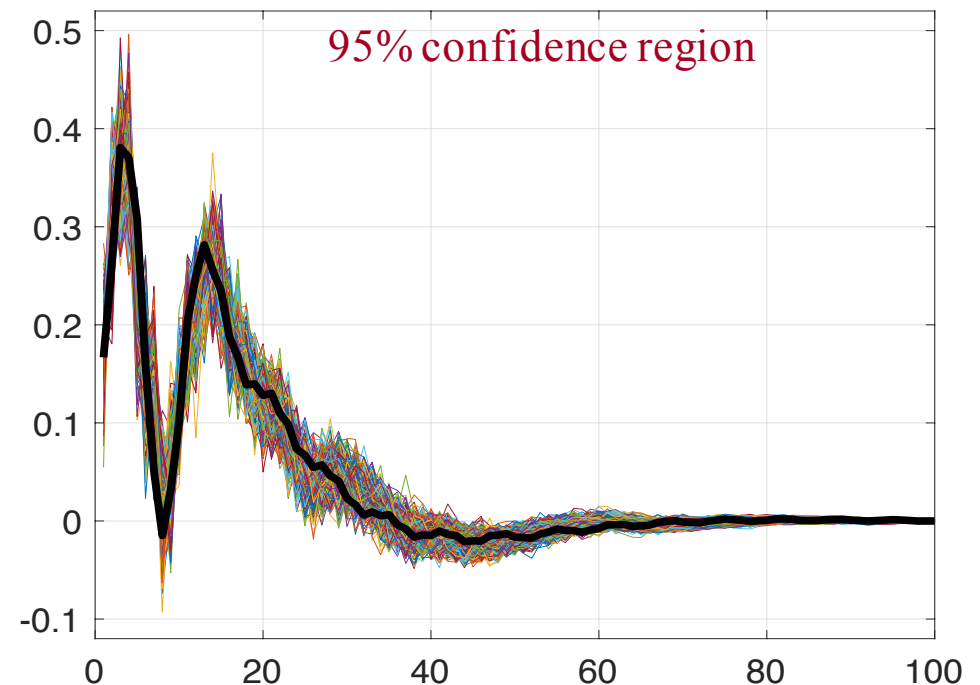
$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



Bayesian solution: posterior regions

Gaussian posterior
(hyperparameters set to their estimates)

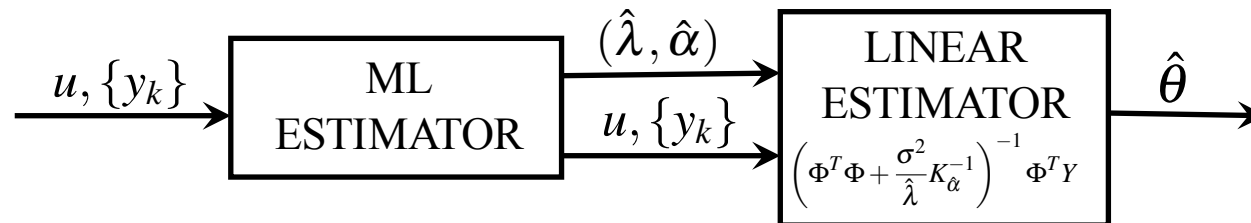
$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$



Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



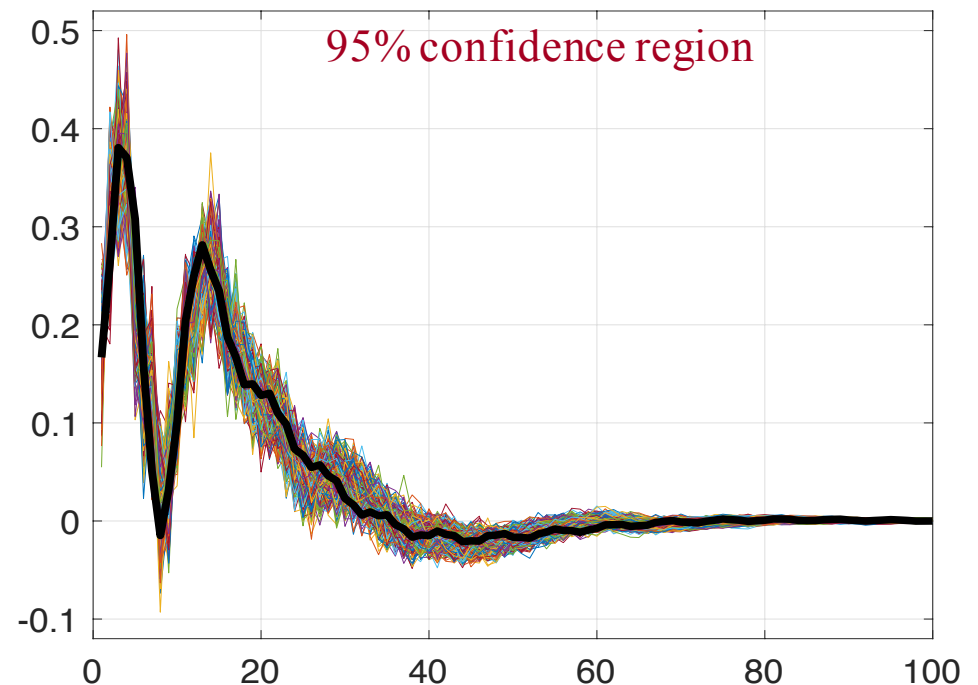
Bayesian solution: posterior regions

Gaussian posterior
(hyperparameters set to their estimates)

$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$

**Empirical studies show advantages
over classical approaches**

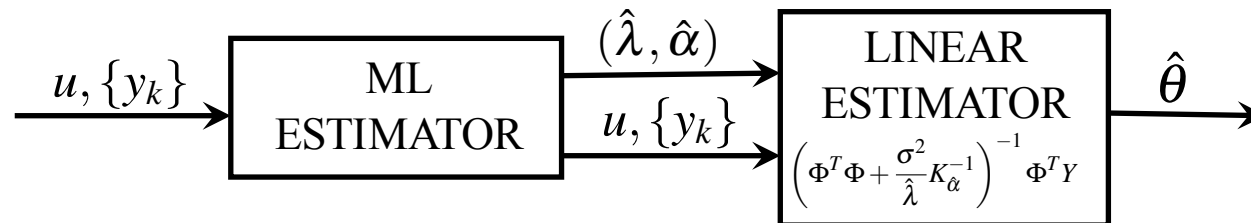
G. Prando, D. Romeres, G. Pillonetto, A. Chiuso
Classical vs. Bayesian methods for linear sysid:
point estimators and confidence sets, ECC 2016



Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



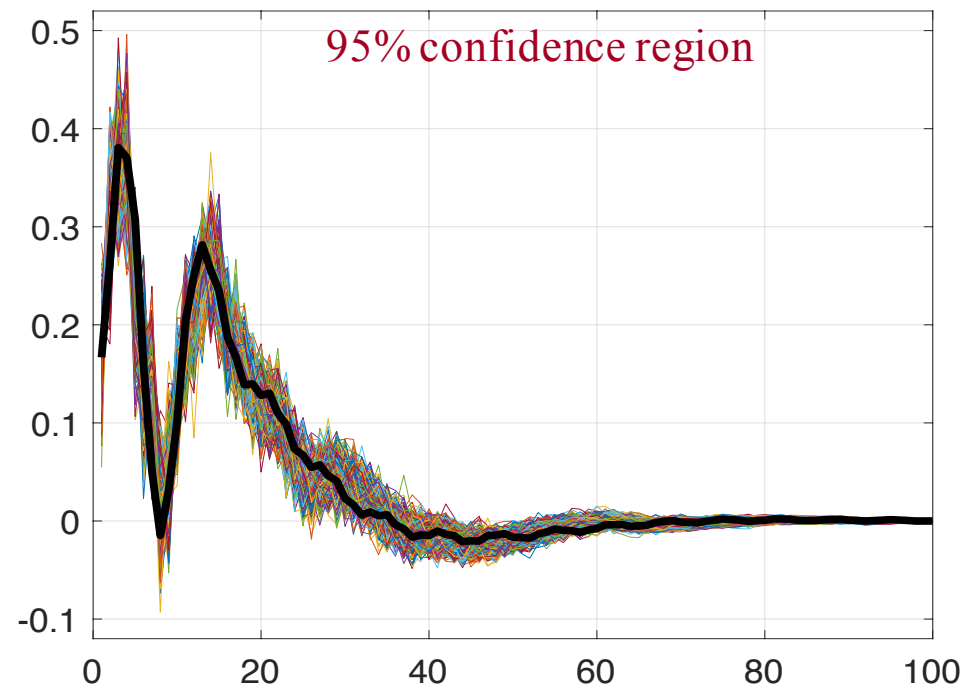
Bayesian solution: posterior regions

Gaussian posterior
(hyperparameters set to their estimates)

$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$

Theoretical issues to be investigated

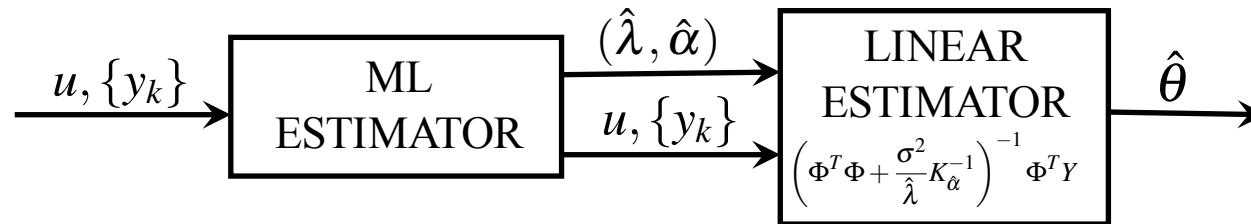
- robustness to deviations from Gaussianity



Robustness issues

Model and estimator

$$Y = \Phi\theta + v \quad \theta \sim \mathcal{N}(0, \lambda K_\alpha) \quad v \sim \mathcal{N}(0, \sigma^2 I)$$



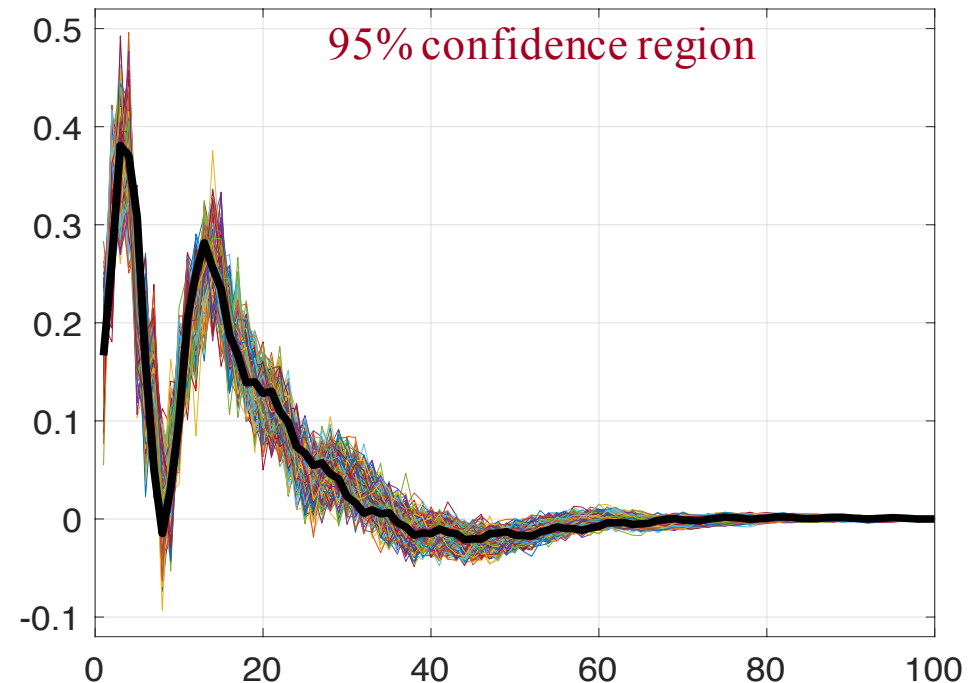
Bayesian solution: posterior regions

Gaussian posterior
(hyperparameters set to their estimates)

$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$

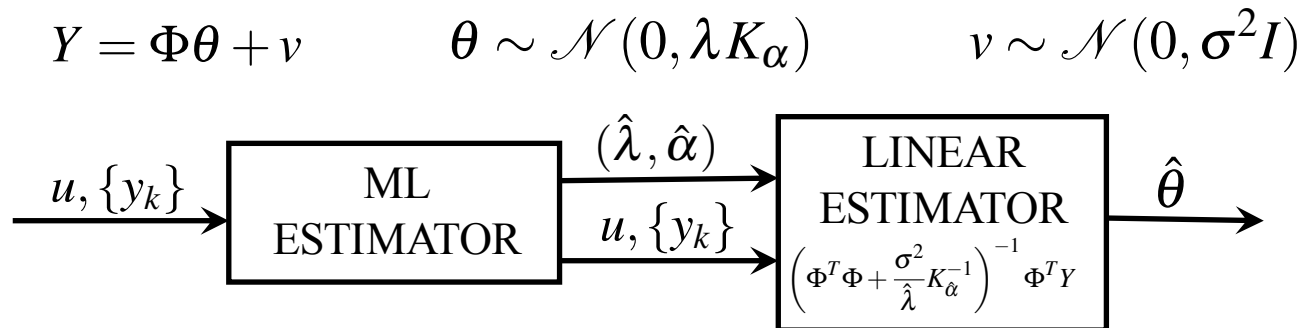
Theoretical issues to be investigated

- robustness to deviations from Gaussianity
- kernel-based exact confidence regions without prior correctness



Robustness issues

Model and estimator



Bayesian solution: posterior regions

Gaussian posterior
(hyperparameters set to their estimates)

$$\theta|Y \sim \mathcal{N} \left(\hat{\theta}, \left(\frac{\Phi^T \Phi}{\sigma^2} + \frac{K_{\hat{\alpha}}^{-1}}{\hat{\lambda}} \right)^{-1} \right)$$

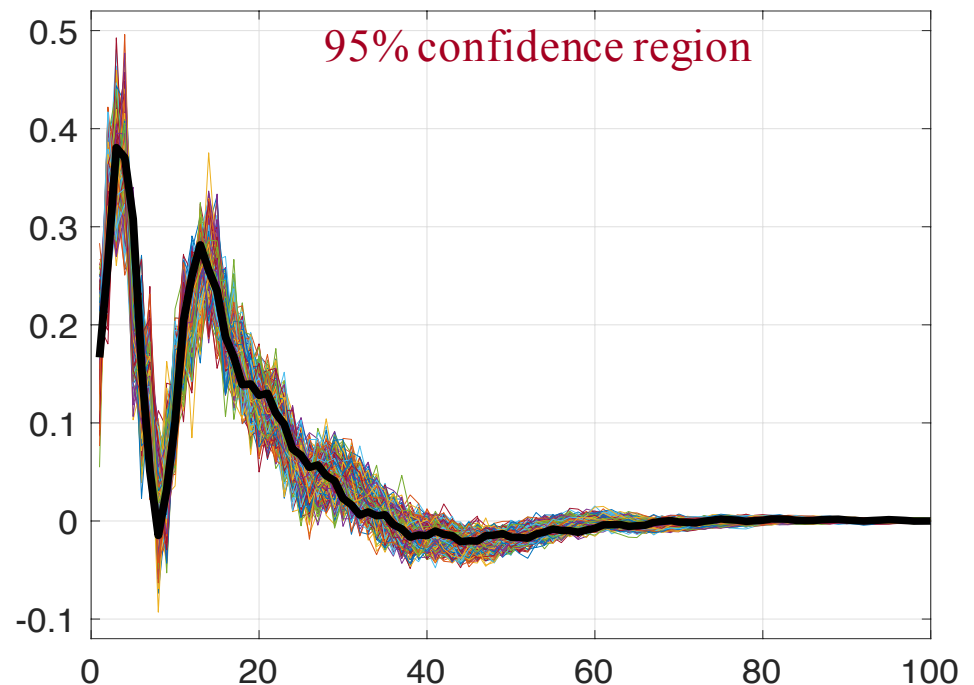
Theoretical issues to be investigated

- robustness to deviations from Gaussianity
- kernel-based exact confidence regions without prior correctness

A first step: Kernels + SPS approach

G. Pillonetto, A. Care', M.C. Campi

Kernel-based SPS, Proceedings of 18th IFAC SysId 2018



Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.

Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.
- We have reviewed kernel-based linear system identification with the impulse response estimate obtained in a high-dimensional space.

Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.
- We have reviewed kernel-based linear system identification with the impulse response estimate obtained in a high-dimensional space. Overfitting is avoided by including information on regularity and BIBO stability. Discrete model order selection is replaced by continuous tuning of hyperparameters and Marginal Likelihood seems especially effective.

Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.
- We have reviewed kernel-based linear system identification with the impulse response estimate obtained in a high-dimensional space. Overfitting is avoided by including information on regularity and BIBO stability. Discrete model order selection is replaced by continuous tuning of hyperparameters and Marginal Likelihood seems especially effective.
- Many important issues still to be investigated, e.g. study and derivation of robust uncertainty bounds.

Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.
- We have reviewed kernel-based linear system identification with the impulse response estimate obtained in a high-dimensional space. Overfitting is avoided by including information on regularity and BIBO stability. Discrete model order selection is replaced by continuous tuning of hyperparameters and Marginal Likelihood seems especially effective.
- Many important issues still to be investigated, e.g. study and derivation of robust uncertainty bounds.

The material in this lecture is the outcome of joint work with a number of colleagues and friends from whom I have learned a lot:

Giuseppe De Nicolao, Alessandro Chiuso, Bradley Bell,
Tianshi Chen, Lennart Ljung, Giorgio Picci,
Aleksandr Aravkin, Giulio Bottegal, James Burke

Conclusions

- State-of-the-art parametric identification approaches postulate a class of finite-dimensional models. An important point is model order selection.
- We have reviewed kernel-based linear system identification with the impulse response estimate obtained in a high-dimensional space. Overfitting is avoided by including information on regularity and BIBO stability. Discrete model order selection is replaced by continuous tuning of hyperparameters and Marginal Likelihood seems especially effective.
- Many important issues still to be investigated, e.g. study and derivation of robust uncertainty bounds.

The material in this lecture is the outcome of joint work with a number of colleagues and friends from whom I have learned a lot:

Giuseppe De Nicolao, Alessandro Chiuso, Bradley Bell,
Tianshi Chen, Lennart Ljung, Giorgio Picci,
Aleksandr Aravkin, Giulio Bottegal, James Burke

Thanks for the attention