Nonparametric identification of LTIs: theory and practice

Damiano Varagnolo

KTH – School of Electrical Engineering – Automatic Control Lab

May 8^{th}, 2012

More from . . .

Gianluigi Pillonetto

Alessandro Chiuso

Question

$$\begin{cases} y_1 & = & \theta_1 + \nu_1 \\ & \vdots & \\ y_N & = & \theta_N + \nu_N \end{cases}$$

$N \geq 3$

θ_i deterministic

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

$\nu_i \perp \nu_j$

Find $\widehat{\boldsymbol{\theta}}$ minimizing $\mathbb{E}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right]$

Question

$$\begin{cases} y_1 & = & \theta_1 + \nu_1 \\ & \vdots & \\ y_N & = & \theta_N + \nu_N \end{cases}$$

$N \geq 3$

θ_i deterministic

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

$\nu_i \perp \nu_j$

Find $\widehat{\boldsymbol{\theta}}$ minimizing $\mathbb{E}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right]$

Answer?

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

ML

MVUE

efficient

The James-Stein estimator (1956)

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(M-2)\sigma^2}{\|\boldsymbol{y}\|_2^2}\right)\boldsymbol{y}$$

- its MSEs always better than LSs ones, for every $\boldsymbol{\theta} \in \mathbb{R}^M$
- its MSEs tend to LSs ones when $\|\boldsymbol{\theta}\|_2$ is large

Why?

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(M-2)\sigma^2}{\|\boldsymbol{y}\|_2^2}\right)\boldsymbol{y}$$

The Bias – Variance dilemma

$$\mathbb{E}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right] = \left\|\mathbb{E}\left[\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right]\right\|_2^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]\right\|_2^2\right]$$

Why?

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(M-2)\sigma^2}{\|\boldsymbol{y}\|_2^2}\right)\boldsymbol{y}$$

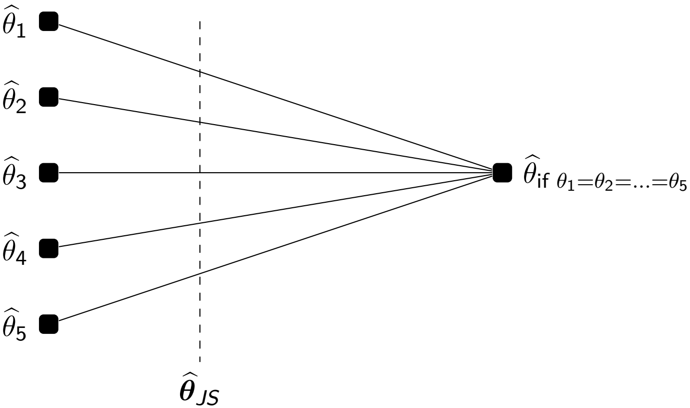
The Bias – Variance dilemma

$$\mathbb{E}\left[\left\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right\|_2^2\right] = \left\|\mathbb{E}\left[\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\right]\right\|_2^2 + \mathbb{E}\left[\left\|\widehat{\boldsymbol{\theta}} - \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]\right\|_2^2\right]$$

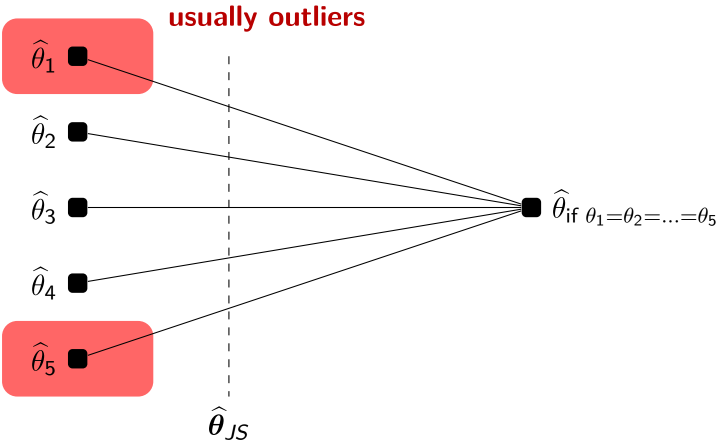
$\widehat{\boldsymbol{\theta}}_{JS}$ is a regularized estimator

Rule-of-thumb: more regularization $\Rightarrow \begin{cases} \text{more bias} \\ \text{less variance} \end{cases}$

Graphical intuition



Graphical intuition

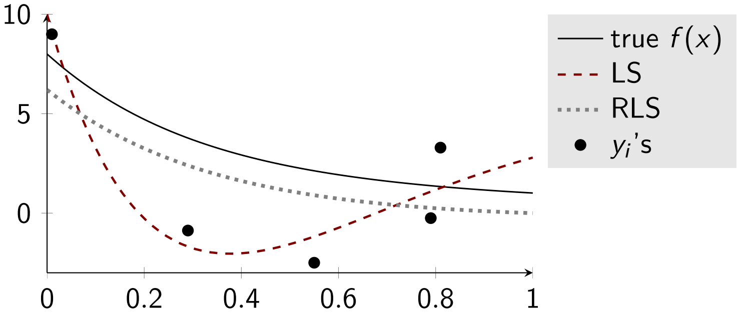


Some comments

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(M-2)\sigma^2}{\|\boldsymbol{y}\|_2^2}\right)\boldsymbol{y}$$

- $\widehat{\boldsymbol{\theta}}_{JS}$ "learns" the mechanism generating the y_m's
  (connections with Empirical Bayes)
- improves the ***total MSE*** , not the MSEs of the single components
- very close to LS if y_m's very far apart

Regularization helps: an example



$$f(x) = \theta_1 e^{-x} + \theta_2 e^{-2x} + \theta_3 e^{-3x} \qquad \boldsymbol{\theta} = \begin{bmatrix} 3 \\ -4 \\ 9 \end{bmatrix} \qquad \nu_i \sim \mathcal{N}(0, 9)$$

$$LS = \arg\min_{\bar{\boldsymbol{\theta}} \in \mathbb{R}^3} \sum_i \left(y_i - f_{\bar{\boldsymbol{\theta}}}(x_i) \right)^2 \quad RLS = \arg\min_{\bar{\boldsymbol{\theta}} \in \mathbb{R}^3} \sum_i \left(y_i - f_{\bar{\boldsymbol{\theta}}}(x_i) \right)^2 + 5 \left\| \bar{\boldsymbol{\theta}} \right\|_2^2$$

Examples of regularization

- Ridge regression (ℓ_2 norms)
- LASSO (ℓ_1 norms)
- Elastic network (combination of ℓ_1 and ℓ_2 norms)
- ...

Examples of regularization

- Ridge regression (ℓ_2 norms)
- LASSO (ℓ_1 norms)
- Elastic network (combination of ℓ_1 and ℓ_2 norms)
- ...

regularization is most useful
when problems are ill-conditioned

Examples of regularization

- Ridge regression (ℓ_2 norms)
- LASSO (ℓ_1 norms)
- Elastic network (combination of ℓ_1 and ℓ_2 norms)
- ...

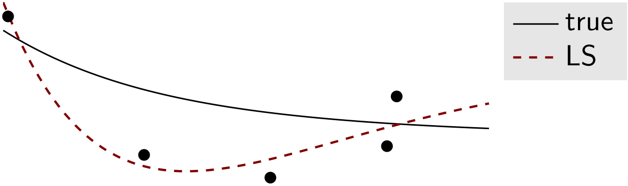
regularization is most useful
when problems are ill-conditioned

From parametric to nonparametric

previous aim: introduce regularization

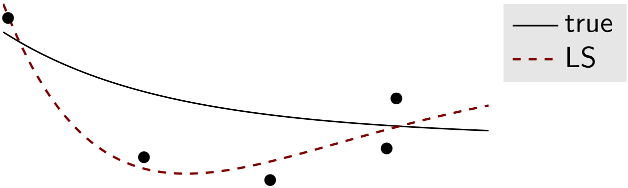
next aim: introduce nonparametric regression

Nonparametric approaches – motivations 1



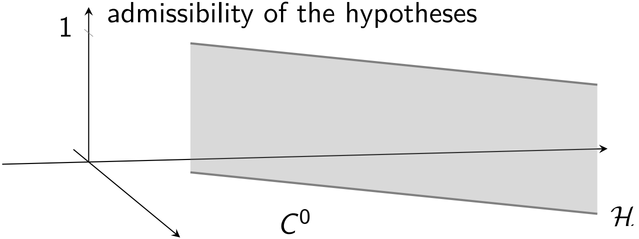
Key fact in previous example: $\mathcal{H} = span\left\{e^{-x}, e^{-2x}, e^{-3x}\right\}$

Nonparametric approaches – motivations 1

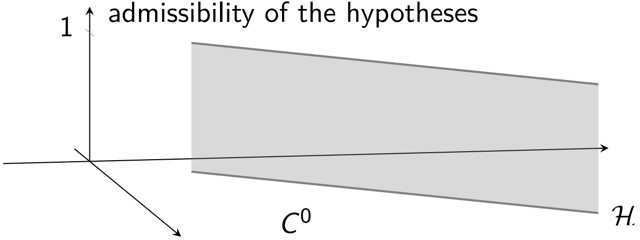


Key fact in previous example: $\mathcal{H} = span\left\{e^{-x}, e^{-2x}, e^{-3x}\right\}$

peculiarity of parametric approaches!



Nonparametric approaches – motivations 2

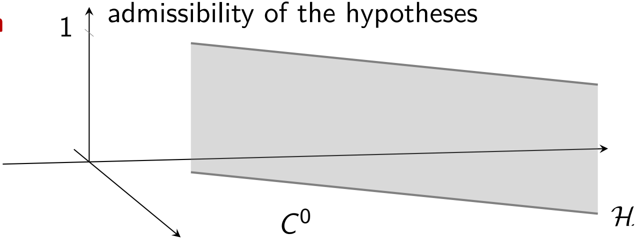


(some) drawbacks of parametric approaches

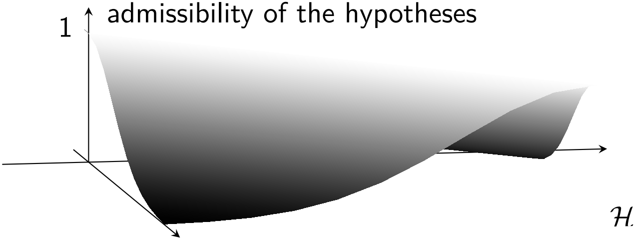
- require high levels of prior knowledge
- complex systems may lead to proliferation of parameters *(high variance!)*

Towards nonparametric approaches

From



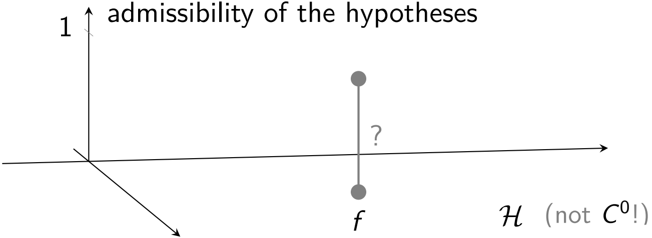
to



how should we define the novel "admissibility"?

admissibility = regularity

admissibility = regularity



concept of regularity depends on prior assumptions!

intuitive examples:

$$\|f\|_2^2 = \int_{\mathcal{X}} f(x)^2 dx \qquad \|f\|_{\text{cub.sp.}}^2 = \int_{\mathcal{X}} \ddot{f}(x)^2 dx$$

Towards RKHSs

assumption: regularity of $f := \|f\|_\star$ – hypotheses $f \in \mathcal{H}_\star$

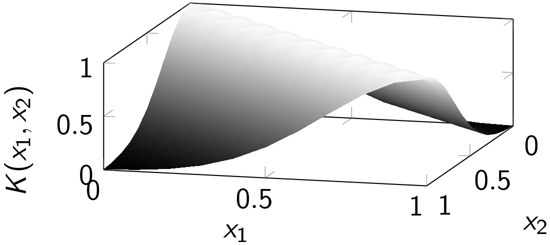
question: what can $\star$ be?

RKHSs – the key ingredient

$\star = K \qquad K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R} \qquad K : \begin{cases} \text{symmetric} \\ \text{continuous} \\ \text{positive definite} \end{cases}$

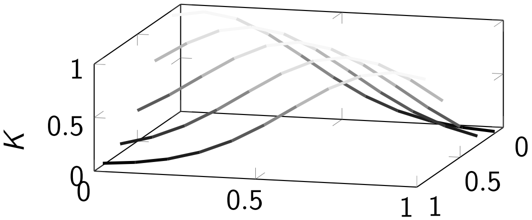
($\mathcal{X}$ = domain of f)

Example: Gaussian Kernel $K(x_1, x_2) = \exp\left(-\frac{(x_1 - x_2)^2}{2\sigma^2}\right)$



RKHSs

$$K \quad \leftrightarrow \quad \mathcal{H}_K := \overline{span\left\{K(x,\cdot) \text{ s.t. } x \in \mathcal{X}\right\}}$$



RKHSs = spaces of functions where the evaluation functional is bounded and linear

Remarks:

- $\mathcal{H}_K \subset C^0$
- $f(\cdot) = \sum_i a_i K(x_i, \cdot) \;\Rightarrow\; \|f\|_K^2 = \sum_{i,j} a_i a_j K(x_i, x_j)$

Kernels are the goggles with which one sees the world

$$f(\cdot) = \sum_i a_i K(x_i, \cdot) \;\Rightarrow\; \|f\|_K^2 = \sum_{i,j} a_i a_j K(x_i, x_j)$$

same f, different K's, different $\|f\|_K$'s

Kernels are the goggles with which one sees the world

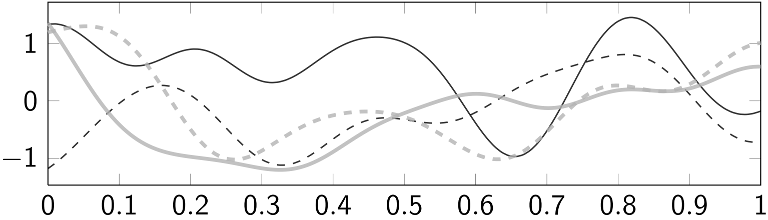
$$f(\cdot) = \sum_i a_i K(x_i, \cdot) \;\Rightarrow\; \|f\|_K^2 = \sum_{i,j} a_i a_j K(x_i, x_j)$$

same f, different K's, different $\|f\|_K$'s

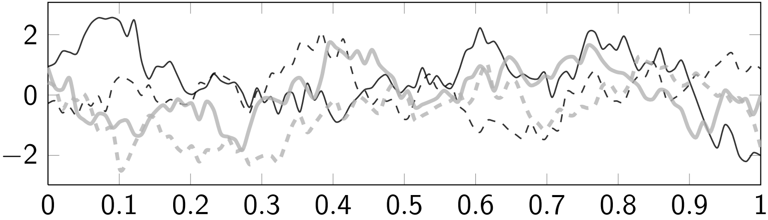
linear $x_1 x_2 + c$	polynomial $(\alpha x_1 x_2 + c)^d$	Gaussian $\exp\left(-\gamma(x_1 - x_2)^2\right)$
Laplacian $\exp\left(-\gamma(x_1 - x_2)\right)$	hyperbolic $\tanh\left((\alpha x_1 x_2 + c)^d\right)$	rational quadratic $1 - \frac{\|x_1 - x_2\|^2}{\|x_1 - x_2\|^2 + c}$
multiquadratic $\sqrt{\|x_1 - x_2\|^2 + c}$	inv. multiquadratic $\left(\sqrt{\|x_1 - x_2\|^2 + c}\right)^{-1}$	wave $\frac{\theta}{\|x_1 - x_2\|} \sin\left(\frac{\|x_1 - x_2\|}{\theta}\right)$
⋮	⋮	⋮

Examples of typical elements of $\mathcal{H}_K$

Gaussian Kernel: $K(x_1, x_2) = \exp\left(-\gamma|x_1 - x_2|^2\right)$



Laplacian Kernel: $K(x_1, x_2) = \exp\left(-\gamma|x_1 - x_2|\right)$



Example of regression with RKHSs

$$f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$$

tradeoff between fitting and regularity
– same as before!!

Example of regression with RKHSs

$$f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$$

tradeoff between fitting and regularity
– same as before!!

This case = Regularization Network

$$f^*(\cdot) = \sum_i c_i K(x_i, \cdot)$$

$$\begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} = \left(\begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_M) \\ \vdots & & \vdots \\ K(x_M, x_1) & \cdots & K(x_M, x_M) \end{bmatrix} + \gamma I \right)^{-1} \begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix}$$

important remarks

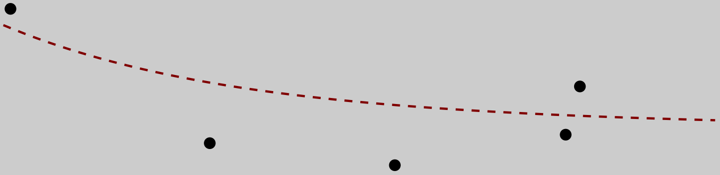
Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$

important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

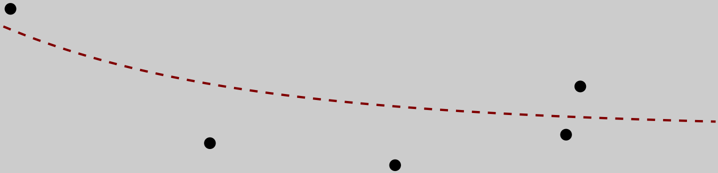
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

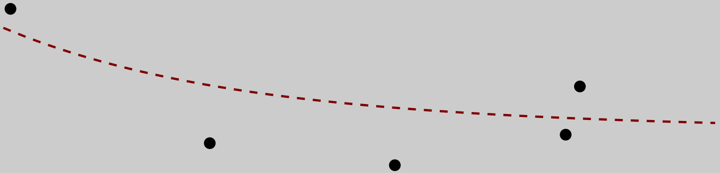
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

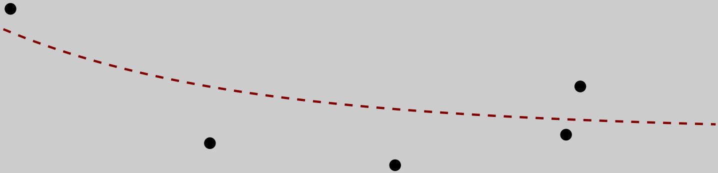
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

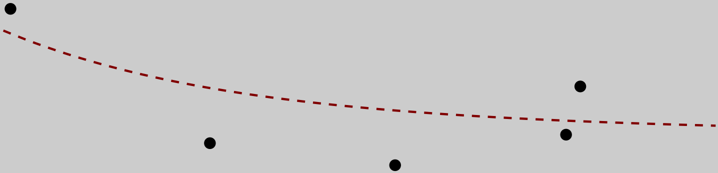
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \left\| f \right\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

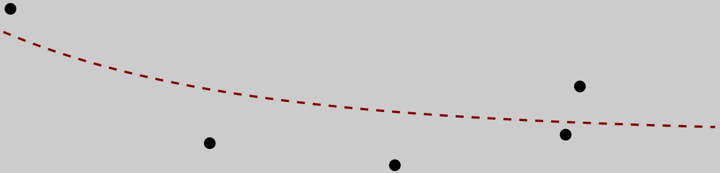
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

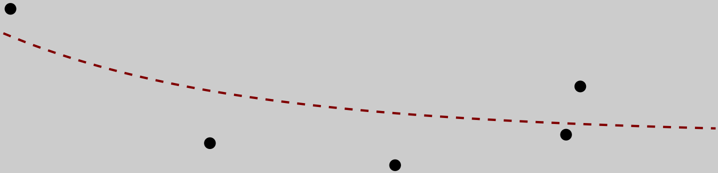
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

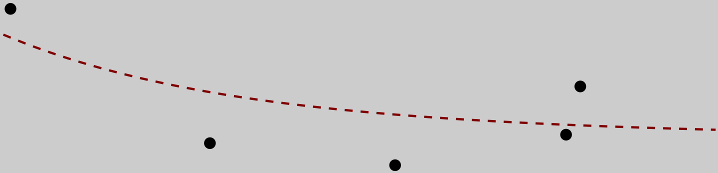
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

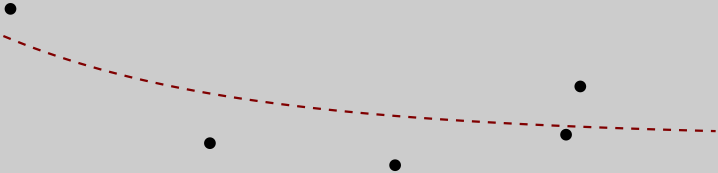
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \| f \|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

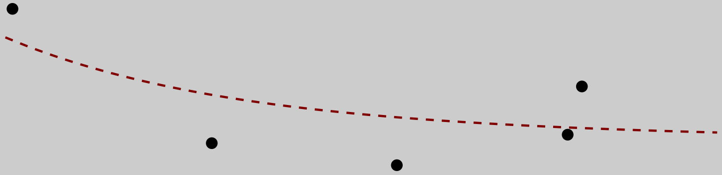
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

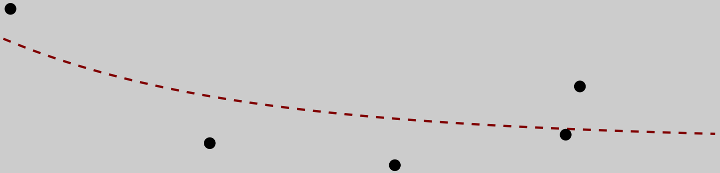
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

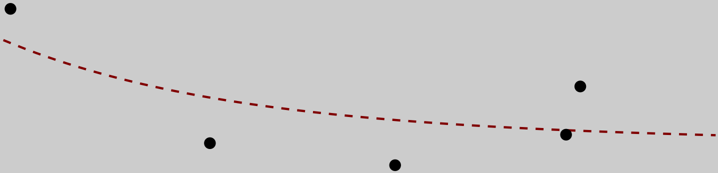
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

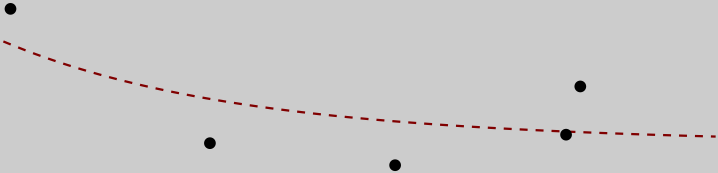
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

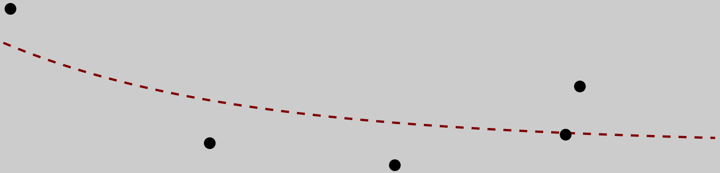
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i) \right)^2$

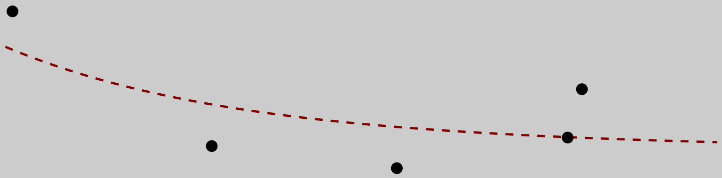
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i) \right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

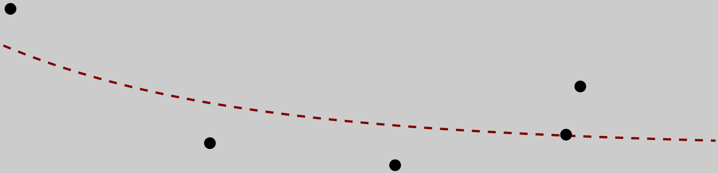
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

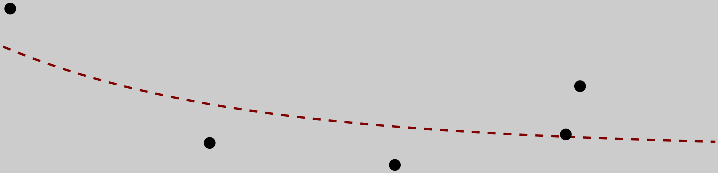
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

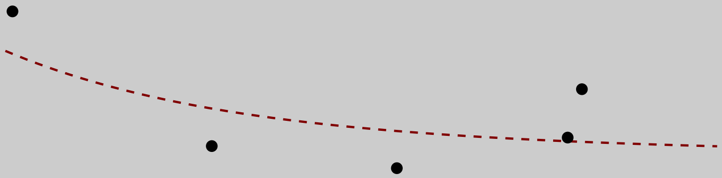
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

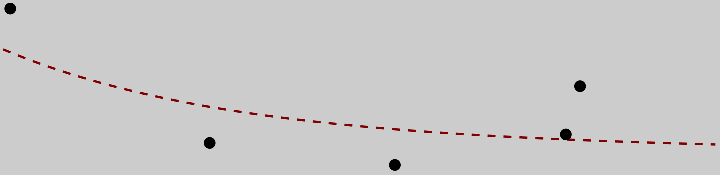
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

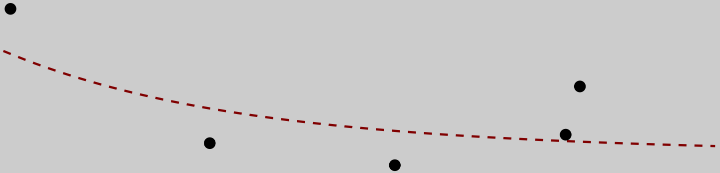
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

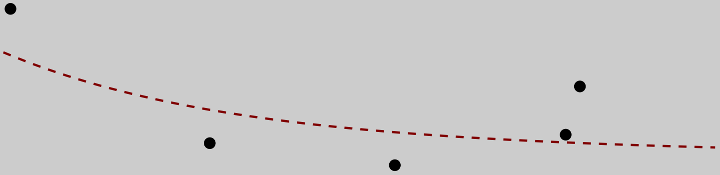
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i) \right)^2$

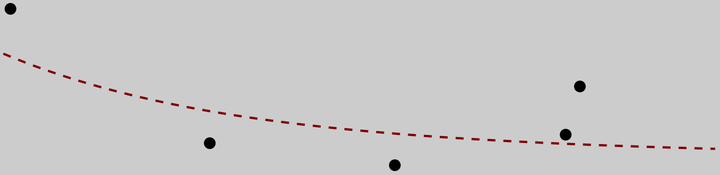
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i) \right)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

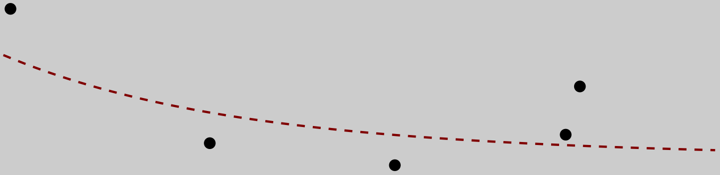
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

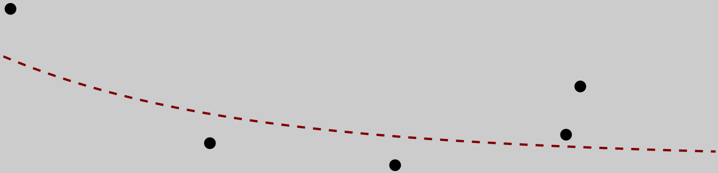
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

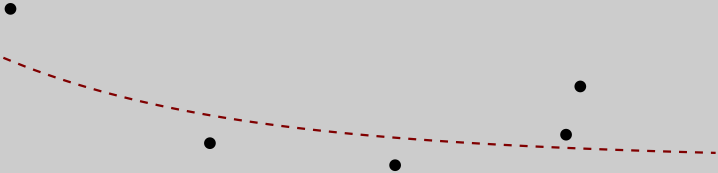
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i) \right)^2$

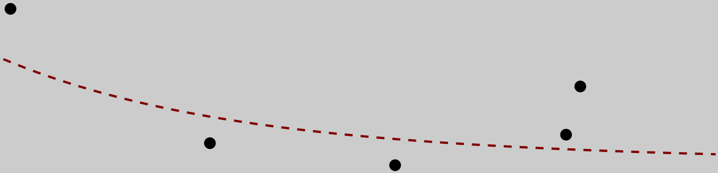
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i) \right)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

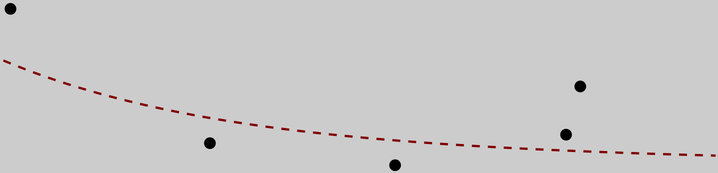
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

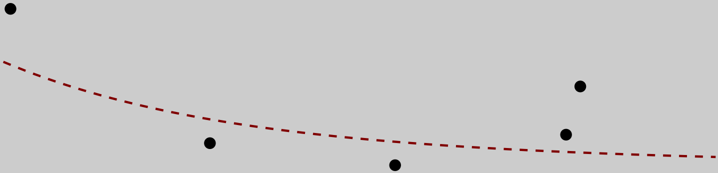
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

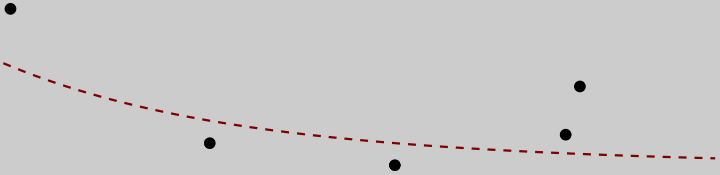
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$



important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i) \right)^2$

Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i) \right)^2 + \gamma \, \|f\|_K^2$



important remarks

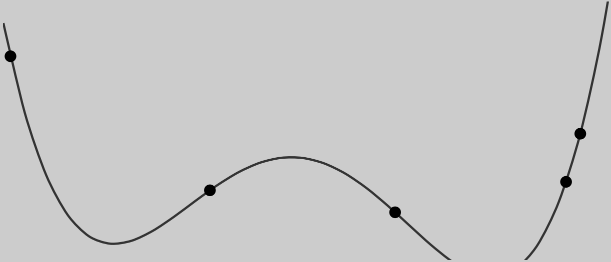
Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i)\Big)^2$

Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i)\Big)^2 + \gamma \, \|f\|_K^2$

important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \Big(y_i - f_\theta(x_i) \Big)^2$

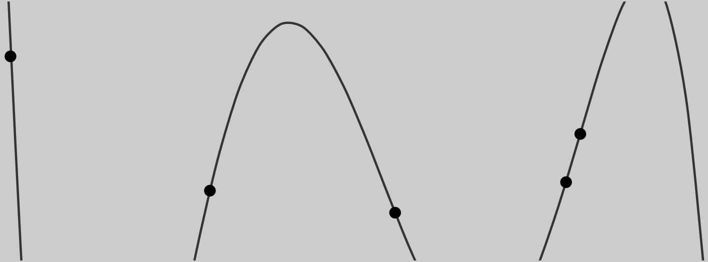
Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \Big(y_i - f(x_i) \Big)^2 + \gamma \, \|f\|_K^2$



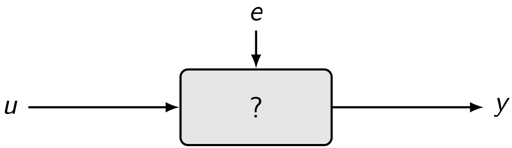
important remarks

Parametric: $\theta^* = \arg\min_{\theta \in \Theta} \sum_i \left(y_i - f_\theta(x_i)\right)^2$

Nonparametric: $f^* = \arg\min_{f \in \mathcal{H}_K} \sum_i \left(y_i - f(x_i)\right)^2 + \gamma \|f\|_K^2$



Nonparametric identification of LTIs

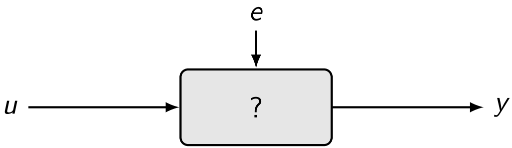


$$y_t = \sum_{i=1}^{\infty} f_i u_{t-i} + \sum_{i=1}^{\infty} g_i e_{t-i}, \qquad \text{dataset} = \{u_t\}, \{y_t\}$$

PEM approach:

$$\widehat{y}_{t|t-1} = \sum_{i=1}^{\infty} h_i^u u_{t-i} + \sum_{i=1}^{\infty} h_i^y y_{t-i}$$

Nonparametric identification of LTIs



$$y_t = \sum_{i=1}^{\infty} f_i u_{t-i} + \sum_{i=1}^{\infty} g_i e_{t-i}, \qquad \text{dataset} \; = \{u_t\}, \{y_t\}$$

PEM approach:

$$\widehat{y}_{t|t-1} \; = \; \sum_{i=1}^{\infty} h_i^u u_{t-i} \; + \; \sum_{i=1}^{\infty} h_i^y y_{t-i}$$

SysId with Regularization Networks??

$$h^* = \arg \min_{h \in \bigstar} \sum_t \left(y_t - \widehat{y}_{t|t-1} \right)^2 + \gamma \left\| h \right\|_{\bigstar}^2$$

The Stable Splines Kernel

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

$$\text{cubic splines: } W(s,t) = \begin{cases} \frac{s^2}{2}\left(t - \frac{s}{3}\right) & \text{if } s \le t \\ \frac{t^2}{2}\left(s - \frac{t}{3}\right) & \text{if } s > t \end{cases}$$

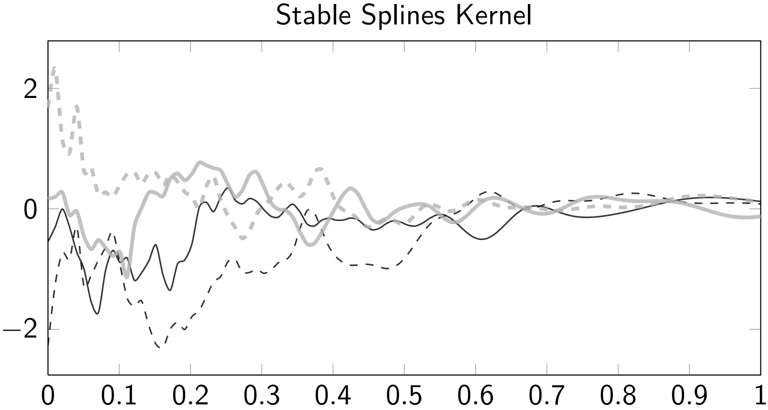
stable splines: $K(x_1, x_2) = W\left(e^{-\beta x_1}, e^{-\beta x_2}\right)$

Bayesian interpretation

Let $f \sim \mathcal{GP}(0, K)$. Then

$$\mathbb{P}\left[f = \text{imp. resp. of LTI BIBO stable system}\right] = 1$$

Examples of typical elements of $\mathcal{H}_K$



how to actually perform the identification

The model & the hyperparameters

$$A(z)y(t) = B(z)u(t) + C(z)e(t)$$

MISO $\Rightarrow$ more than one impulse response!

$i :=$ impulse response index

The model & the hyperparameters

$$A(z)y(t) = B(z)u(t) + C(z)e(t)$$

MISO $\Rightarrow$ more than one impulse response!

$i :=$ impulse response index

$$h^i \sim \mathcal{GP}\Big(0, \lambda_i^2 K(\cdot, \cdot, \; ; \; \beta)\Big)$$

The model & the hyperparameters

$$A(z)y(t) = B(z)u(t) + C(z)e(t)$$

MISO $\Rightarrow$ more than one impulse response!

$i :=$ impulse response index

$$h^i \sim \mathcal{GP}\Big(0, \lambda_i^2 K(\cdot,\cdot,\ ;\ \beta)\Big)$$

λ_i^2: "amplitude" of the i-th impulse response

β: decay ratio

Estimation of the hyperparameters

Empirical Bayes in theory

1. assume existence of prior distribution with *unknown hyperparameters*

Empirical Bayes in practice *(with some abuses of notation)*

1. $p(y|h)$, $p(h|\lambda)$ are known, λ unknown

Estimation of the hyperparameters

Empirical Bayes in theory

1. assume existence of prior distribution with *unknown hyperparameters*
2. compute the marginal likelihood

Empirical Bayes in practice *(with some abuses of notation)*

1. $p(y\,|h)$, $p(h\,|\lambda)$ are known, λ unknown
2. exploiting $p(y, h\,|\lambda) = p(y\,|h, \lambda)\,p(h\,|\lambda)$ compute

$$p(y\,|\lambda) = \int p(y, h\,|\lambda)\,p(h\,|\lambda)\,dh$$

Estimation of the hyperparameters

Empirical Bayes in theory

1. assume existence of prior distribution with *unknown hyperparameters*
2. compute the marginal likelihood
3. estimate the hyperparameters maximizing the marginal likelihood

Empirical Bayes in practice *(with some abuses of notation)*

1. $p\,(y\,|h)$, $p\,(h\,|\lambda)$ are known, λ unknown
2. exploiting $p\,(y, h\,|\lambda) = p\,(y\,|h, \lambda)\,p\,(h\,|\lambda)$ compute

$$p\,(y\,|\lambda) = \int p\,(y, h\,|\lambda)\,p\,(h\,|\lambda)\,dh$$

3. $\lambda^* = \arg\max_\lambda p\,(y\,|\lambda)$

SSpline.m: a matlab toolbox

... even if not yet publicly available

```
[M,ip,Ak]=SSpline(y,U,p,l,mv,mb,cn,r,LP,LP2,ips)
```

- `M`: estimated tf (`idpoly` object)
- `ip`: estimated hyperparameters
- `Ak`: AIC of the estimated tf
- `y`: measured outputs
- `U`: measured inputs
- `p`: max. length of the to-be estimated impulse responses
- `l`: type of tf to be estimated (ARMAX / ARX / etc.)
- `mv`: one λ in common for all the h^i's or not
- `mb`: one β in common for all the h^i's or not
- `cn`: identify high frequencies components
- `r`: number of data to be used for estimating the hyperparameters
- `LP`: obtain sparse solutions
- `LP2`: obtain approximated solutions
- `ips`: start optimization from the assigned initial point

the routine returns an `idpoly` and has a pre-fixed impulse responses max. length!

Question: at the end of the day, we obtain an object of the same kind of classical PEM approaches. So, why should this be better?

the routine returns an `idpoly` and has a pre-fixed impulse responses max. length!

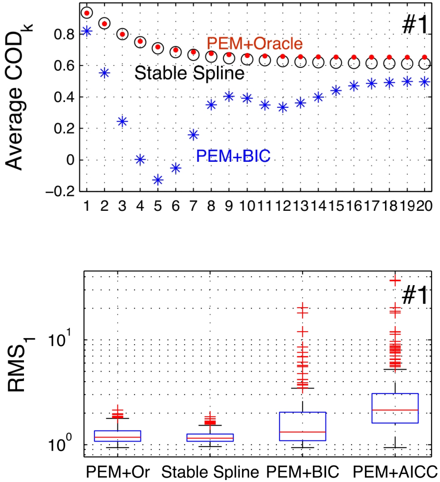
Question: at the end of the day, we obtain an object of the same kind of classical PEM approaches. So, why should this be better?

Answer: at the end of the day, ***because of the bias / variance tradeoff***

Some comparisons

1000 MC runs, for each run:

- ARMAX order $\in \{1, \ldots, 30\}$
- ARMAX model $\sim$ `drmodel`
- no delays
- inputs $\sim$ `idinput`
- training set = 200 samples
- test set = 1000 samples



And some conclusions

regularization usually has beneficial effects

nonparametric approaches are specially suited for complex situations

regularization in nonparametric approaches can be seen as using prior smoothness assumptions

regularization in nonparametric approaches leads to an extremely efficient LTIs sysid technique

Nonparametric identification of LTIs: theory and practice

Damiano Varagnolo

KTH – School of Electrical Engineering – Automatic Control Lab

May 8th, 2012

entirely written in LaTeX 2ε using Beamer and Tik Z