Smoothing

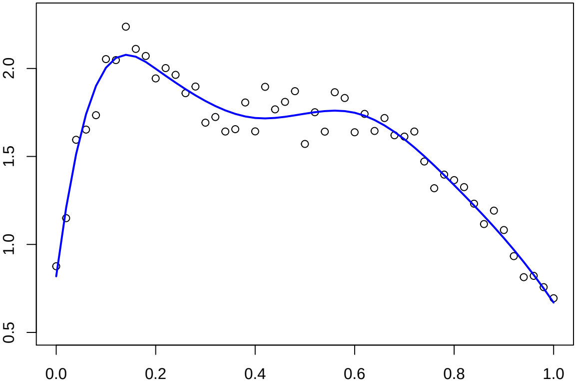
Lessons for the master course in Applied Statistics

May 2023

Laura Sangalli

MOX laboratory – Department of Mathematics - Politecnico di Milano
Email: laura.sangalli@polimi.it
Web: https://sangalli.faculty.polimi.it
Office: Building 14th, La Nave, 7th floor

POLITECNICO
MILANO 1863


2.0
1.5
1.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0

SMOOTHING

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y = f(x) + \epsilon$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

Smoothing

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

- Data: $\{(Y_1, x_1), \dots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2 < 0$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2 < 0$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

Smoothing problem:

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \text{RSS}(f) = \underset{f \in \mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 \right\}$$

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2 < 0$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

Smoothing problem:

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\ \text{RSS}(f) = \underset{f \in \mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 \right\}$$

ill posed!

Smoothing

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2 < 0$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

Smoothing problem:

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\text{argmin}}\, \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 \right\}$$

1st approach: restrict search to $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

Smoothing

- Data: $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \epsilon_i \sim \text{indep} \quad \mathbb{E}[\epsilon_i] = 0, Var[\epsilon_i] = \sigma^2 < 0$$

$f \in \mathcal{F}$ where $\mathcal{F}$ is the appropriate functional space

Smoothing problem:

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\, \text{RSS}(f) = \underset{f \in \mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \mathcal{P}(f) \right\}$$

2nd approach: do not restrict $\mathcal{F}$ but add a roughness penalty

1st approach: $\mathcal{F}$ approximated by $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\text{argmin}} \ \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

1st approach: $\mathcal{F}$ approximated by $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \ \mathrm{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

- $\psi_1, \ldots, \psi_K$: K basis functions s.t. $\mathcal{F}_K = \mathrm{span}(\psi_1, \ldots, \psi_K)$

1st approach: $\mathcal{F}$ approximated by $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\text{argmin}} \ \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

- $\psi_1, \ldots, \psi_K$: K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \ldots, \psi_K)$

Model

$$Y_i = f(x_i) + \epsilon_i$$

1st approach: $\mathcal{F}$ approximated by $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\text{argmin}}\ \text{RSS}(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

- $\psi_1, \ldots, \psi_K$: K basis functions s.t. $\mathcal{F}_K = \text{span}(\psi_1, \ldots, \psi_K)$

Model

$$Y_i = f(x_i) + \epsilon_i \qquad \rightsquigarrow \qquad Y_i = \sum_{j=1}^{K} c_j \psi_j(x_i) + \epsilon_i$$

1st approach: $\mathcal{F}$ approximated by $\mathcal{F}_K$, with $dim(\mathcal{F}_K) = K << n$

$$\hat{f} = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \ \mathrm{RSS}(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - f(x_i)\right)^2$$

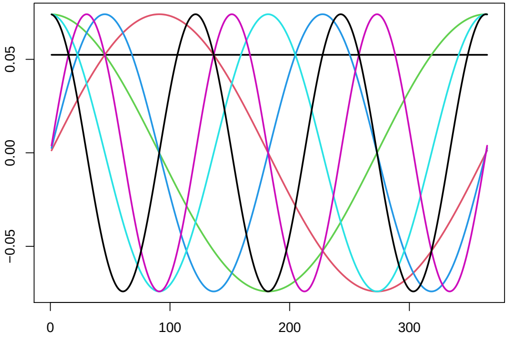
$$= \underset{\boldsymbol{c} \in \mathbb{R}^K}{\operatorname{argmin}} \ \mathrm{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c} \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{K} c_j \psi_j(x_i)\right)^2$$

- $\psi_1, \ldots, \psi_K$: K basis functions s.t. $\mathcal{F}_K = \mathrm{span}(\psi_1, \ldots, \psi_K)$

Model

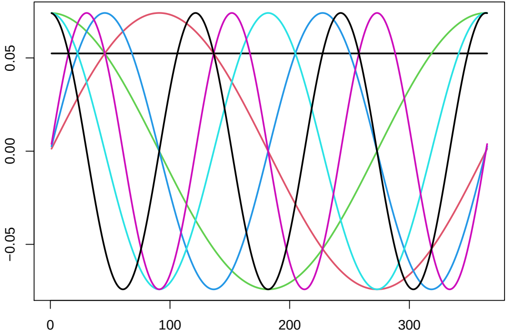
$$Y_i = f(x_i) + \epsilon_i \qquad \rightsquigarrow \qquad Y_i = \sum_{j=1}^{K} c_j \psi_j(x_i) + \epsilon_i$$

Fourier

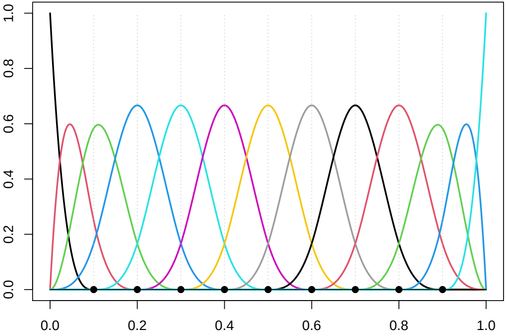


Fourier basis

Fourier

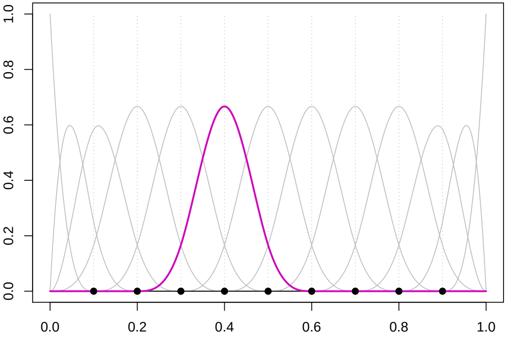


Periodic, localized in frequency, Convenient for differentiation and integration, Computationally efficient (orthogonal)



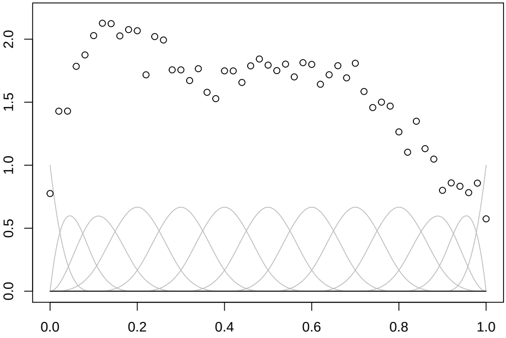
Cubic B-spline basis

B-SPLINE BASIS



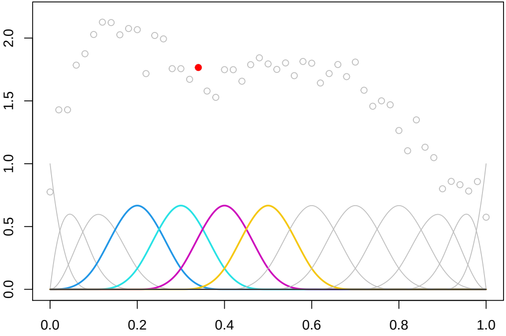
Flexible, localized in space, convenient for differentiation and integration

B-SPLINE BASIS



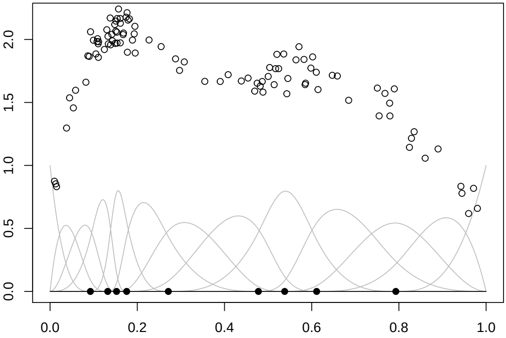
Computationally efficient (local support, band limited structure of key matrices involved, etc)

B-SPLINE BASIS



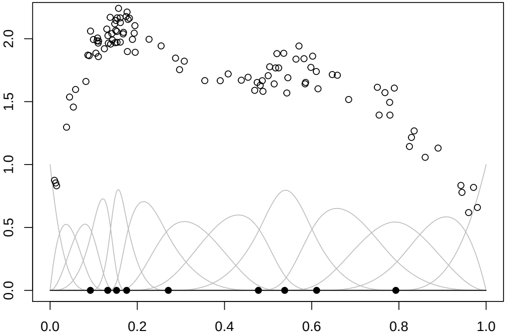
Computationally efficient (local support, band limited structure of key matrices involved, etc)

B-spline basis



Knots can be placed along the percentiles of X

B-SPLINE BASIS



Knots replication permits discontinuity in derivatives or function itself

$$\Psi = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix}$$

$$\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)^\top$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{c} = (c_1, \ldots, c_K)^\top$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{c} = (c_1, \ldots, c_K)^\top \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$$

$$\Psi = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{c} = (c_1, \ldots, c_K)^\top \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$$

Model

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon} = \Psi \boldsymbol{c} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{c} = (c_1, \ldots, c_K)^\top \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$$

Model

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon} = \boldsymbol{\Psi}\boldsymbol{c} + \boldsymbol{\epsilon} \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0} \quad Var[\boldsymbol{\epsilon}] = \sigma^2 I_n$$

$$\Psi = \begin{bmatrix} \psi_1(x_1) & \psi_2(x_1) & \cdots & \psi_K(x_1) \\ \psi_1(x_2) & \psi_2(x_2) & \cdots & \psi_K(x_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_n) & \psi_2(x_n) & \cdots & \psi_K(x_n) \end{bmatrix} \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$$

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \qquad \boldsymbol{f} = \big(f(x_1), \ldots, f(x_n)\big)^\top$$

$$\boldsymbol{c} = (c_1, \ldots, c_K)^\top \qquad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$$

Model

$$\boldsymbol{Y} = \boldsymbol{f} + \boldsymbol{\epsilon} = \Psi\boldsymbol{c} + \boldsymbol{\epsilon} \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad Var[\boldsymbol{\epsilon}] = \sigma^2 I_n$$

$$\hat{f} = \boldsymbol{\psi}^\top \hat{\boldsymbol{c}}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \, \mathrm{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{Y} - \Psi\boldsymbol{c})^\top (\boldsymbol{Y} - \Psi\boldsymbol{c}) \right\}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}}\ \mathrm{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{(\boldsymbol{Y} - \Psi\boldsymbol{c})^\top(\boldsymbol{Y} - \Psi\boldsymbol{c})\right\}$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \operatorname{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{Y} - \Psi\boldsymbol{c})^\top (\boldsymbol{Y} - \Psi\boldsymbol{c}) \right\}$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi\hat{\boldsymbol{c}} = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y} = S\boldsymbol{Y}$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\text{argmin}}\ \text{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\text{argmin}} \left\{ (\boldsymbol{Y} - \Psi\boldsymbol{c})^\top (\boldsymbol{Y} - \Psi\boldsymbol{c}) \right\}$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top\Psi\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi\hat{\boldsymbol{c}} = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y} = S\boldsymbol{Y}$$

$$S = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top : \text{projection matrix (properties: } S^\top S = S)$$

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \operatorname{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{Y} - \Psi\boldsymbol{c})^\top (\boldsymbol{Y} - \Psi\boldsymbol{c}) \right\}$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi\hat{\boldsymbol{c}} = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y} = S\boldsymbol{Y}$$

$$S = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top : \text{projection matrix (properties: } S^\top S = S)$$

$$df = K = tr(S) = tr(S^\top S) = tr(2S - S^\top S)$$

$$
\begin{aligned}
\hat{\boldsymbol{c}} &= \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \operatorname{RSS}(\boldsymbol{c}) = \underset{\boldsymbol{c}\in\mathbb{R}^K}{\operatorname{argmin}} \left\{(\boldsymbol{Y}-\Psi\boldsymbol{c})^\top(\boldsymbol{Y}-\Psi\boldsymbol{c})\right\} \\
\hat{\boldsymbol{c}} &= \left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y} \\
\hat{\boldsymbol{Y}} &= \hat{\boldsymbol{f}} = \Psi\hat{\boldsymbol{c}} = \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top\boldsymbol{Y} = S\boldsymbol{Y} \\
S &= \Psi\left(\Psi^\top\Psi\right)^{-1}\Psi^\top : \text{projection matrix (properties: } S^\top S = S) \\
df &= K = tr(S) = tr(S^\top S) = tr(2S - S^\top S) \\
\hat{\sigma}^2 &= \frac{1}{n-K}(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^\top(\boldsymbol{Y}-\hat{\boldsymbol{Y}}) = \frac{1}{n-tr(S)}(\boldsymbol{Y}-\hat{\boldsymbol{Y}})^\top(\boldsymbol{Y}-\hat{\boldsymbol{Y}})
\end{aligned}
$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{f}'(x) = \psi'(x)^\top \hat{\boldsymbol{c}} \qquad \text{where } \psi' = (\psi'_1, \dots, \psi'_k)^\top$$

$$\hat{f}''(x) = \psi''(x)^\top \hat{\boldsymbol{c}} \qquad \text{where } \psi'' = (\psi''_1, \dots, \psi''_k)^\top$$

Smoothing requires special care when the curve estimate is asked, not only to provide a good smoothing of the data, but also to reflect the differential features

Curve derivatives (or their functions) are very often:
- objects of analysis
- helpful for further processing and data analysis (curve alignment/clustering)

$$\hat{f}(x) = \boldsymbol{\psi}(x)^\top \hat{\boldsymbol{c}} = \boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\big[\hat{f}(x)\big] = \mathbb{E}\big[\psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}\big] = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \boldsymbol{c} = \psi(x)^\top \boldsymbol{c}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \boldsymbol{c} = \psi(x)^\top \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \boldsymbol{c}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\big[\hat{f}(x)\big] = \mathbb{E}\big[\psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}\big] = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \boldsymbol{c} = \psi(x)^\top \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \boldsymbol{c}$$

Source of bias: discretization

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\big[\hat{f}(x)\big] = \mathbb{E}\big[\psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \boldsymbol{Y}\big] = \psi(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \boldsymbol{c} = \psi(x)^\top \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \boldsymbol{c}$$

$$\text{Source of bias: discretization}$$

$$\mathrm{Var}[\hat{f}(x)] = \mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] = \sigma^2 \psi(x)^\top (\Psi^\top \Psi)^{-1} \psi(x)$$

$$\hat{f}(x) = \boldsymbol{\psi}(x)^\top \hat{\mathbf{c}} = \boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \mathbf{Y}\right] = \boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \Psi^\top \Psi \mathbf{c} = \boldsymbol{\psi}(x)^\top \mathbf{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \boldsymbol{\psi}(x)^\top \mathbf{c}$$

Source of bias: discretization

$$\mathrm{Var}[\hat{f}(x)] = \mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] = \sigma^2 \boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \boldsymbol{\psi}(x)$$

$$\widehat{\mathrm{Var}[\hat{f}(x)]} = \mathbb{E}[\{\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\}^2] = \hat{\sigma}^2 \boldsymbol{\psi}(x)^\top (\Psi^\top \Psi)^{-1} \boldsymbol{\psi}(x)$$

Number of bases K controls Bias-Variance trade-off.

Number of bases K controls Bias-Variance trade-off.

K can be selected by AiC, C_p Mallows, cross-validation, Generalized Cross Validation:

Number of bases K controls Bias-Variance trade-off.

K can be selected by AiC, C_p Mallows, cross-validation, Generalized Cross Validation:

$$
\begin{aligned}
GCV(K) &= \frac{n}{(n-K)} \frac{1}{(n-K)} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \\
&= \frac{n}{(n - tr(S))^2} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}})
\end{aligned}
$$

Number of bases K controls Bias-Variance trade-off.

K can be selected by AiC, C_p Mallows, cross-validation, Generalized Cross Validation:

$$\begin{aligned} GCV(K) &= \frac{n}{(n-K)} \frac{1}{(n-K)} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \\ &= \frac{n}{(n - tr(S))^2} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) \end{aligned}$$

Hopefully the chosen value of K is close to that minimizing

$$\mathrm{MSE}[\hat{f}(x)] = \mathbb{E}[\{\hat{f}(x) - f(x)\}^2] = \mathrm{Bias}^2[\hat{f}(x)] + \mathrm{Var}[\hat{f}(x)]$$

Under regularity conditions, as $n \to \infty$ and $K(n) \to \infty$ with appropriate rates

$$\mathrm{Bias}[\hat{f}(x)] \to 0 \qquad \text{and} \qquad \mathrm{Var}[\hat{f}(x)] \to 0$$

$$\mathrm{MSE}[\hat{f}(x)] \to 0$$

$$\hat{f}(x) \approx \textit{Gaussian}$$

Under regularity conditions, as $n \to \infty$
and $K(n) \to \infty$ with appropriate rates

$$\mathrm{Bias}[\hat{f}(x)] \to 0 \qquad \text{and} \qquad \mathrm{Var}[\hat{f}(x)] \to 0$$

$$\mathrm{MSE}[\hat{f}(x)] \to 0$$

$$\hat{f}(x) \approx \textit{Gaussian}$$

Limiting Gaussian distrib justifies Wald type inference on $f(x)$:

Under regularity conditions, as $n \to \infty$
and $K(n) \to \infty$ with appropriate rates

$$\mathrm{Bias}[\hat{f}(x)] \to 0 \qquad \text{and} \qquad \mathrm{Var}[\hat{f}(x)] \to 0$$

$$\mathrm{MSE}[\hat{f}(x)] \to 0$$

$$\hat{f}(x) \approx \textit{Gaussian}$$

Limiting Gaussian distrib justifies Wald type inference on $f(x)$:

- C.I. of approx level $(1-\alpha)$ on $f(x)$: $\hat{f}(x) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{f}(x)]}$

Under regularity conditions, as $n \to \infty$
and $K(n) \to \infty$ with appropriate rates

$$\mathrm{Bias}[\hat{f}(x)] \to 0 \qquad \text{and} \qquad \mathrm{Var}[\hat{f}(x)] \to 0$$

$$\mathrm{MSE}[\hat{f}(x)] \to 0$$

$$\hat{f}(x) \approx \textit{Gaussian}$$

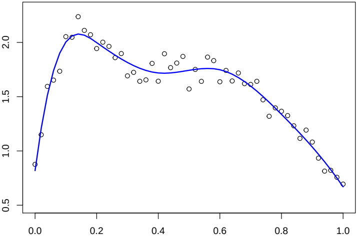
Limiting Gaussian distrib justifies Wald type inference on $f(x)$:

- C.I. of approx level $(1-\alpha)$ on $f(x)$: $\hat{f}(x) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}[\hat{f}(x)]}}$
- Test on $H_0 : f(x) = f_0(x)$ vs $H_1 : f(x) \neq f_0(x)$ of approx level α

  Reject H_0 if $|\hat{f}(x) - f_0(x)| > z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}[\hat{f}(x)]}}$

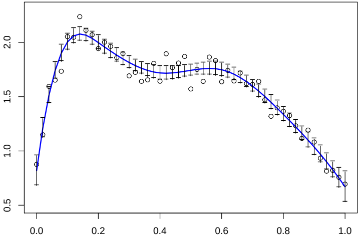
Caveats:

- This inference is only pointwise!



Caveats:

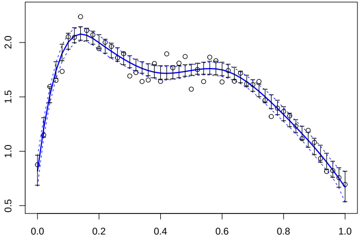
- This inference is only pointwise!



C.I.s are computed one at a time, not simultaneous!

Caveats:

- This inference is only pointwise!

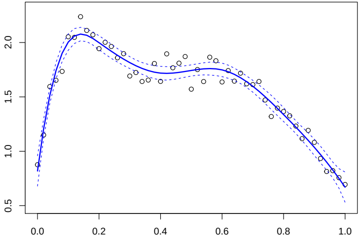


C.I.s are computed one at a time, not simultaneous!

Even if you join the various C.I.s by bands, do not forget these are C.I.s for $f(x)$ in a specific x.

Caveats:

- This inference is only pointwise!



C.I.s are computed one at a time, not simultaneous!

Even if you join the various C.I.s by bands, do not forget these are C.I.s for $f(x)$ in a specific x.

You cannot interpret the smooth dashed bands as delimiters of a region that includes the true overall f with a given confidence.

Caveats:

- This inference is only pointwise!

Caveats:

- This inference is only pointwise!
- This inference does not account for the uncertainty in the selection of K

Caveats:

- This inference is only pointwise!
- This inference does not account for the uncertainty in the selection of K
- Wald type inference is underconservative in smoothing:

Caveats:

- This inference is only pointwise!
- This inference does not account for the uncertainty in the selection of K
- Wald type inference is underconservative in smoothing:

 - IC coverage is < $(1-\alpha)$, i.e., true confidence < nominal

Caveats:

- This inference is only pointwise!
- This inference does not account for the uncertainty in the selection of K
- Wald type inference is underconservative in smoothing:

  - IC coverage is < $(1-\alpha)$, i.e., true confidence < nominal

  - $\mathbb{P}$(type I error) > α

Caveats:

- This inference is only pointwise!
- This inference does not account for the uncertainty in the selection of K
- Wald type inference is underconservative in smoothing:
  - IC coverage is < $(1-\alpha)$, i.e., true confidence < nominal
  - $\mathbb{P}$(type I error) > α

Various alternative approaches, including bootstrap, as well as undersmoothing or oversmoothing approaches (see, e.g., review in Hall and Horowitz 2013)

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \mathrm{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \mathrm{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \ \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

THM: If (a, b) s.t. $a < x_{[1]} < x_{[2]} < \ldots < x_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(x))^2 dx$ then $\hat{f}$ is a natural cubic splines over (a, b) with knots at $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$.

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \ \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

$$f = \sum_{j=1}^{K} c_j \psi_j$$

THM: If (a, b) s.t. $a < x_{[1]} < x_{[2]} < \ldots < x_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(x))^2 dx$ then $\hat{f}$ is a natural cubic splines over (a, b) with knots at $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$.

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}}\ \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

$$= \underset{\boldsymbol{c} \in \mathbb{R}^K}{\text{argmin}} \left\{ (\boldsymbol{Y} - \Psi \boldsymbol{c})^\top (\boldsymbol{Y} - \Psi \boldsymbol{c}) \quad + \lambda \boldsymbol{c}^\top P \boldsymbol{c} \right\} \qquad P \in \mathbb{R}^K \times \mathbb{R}^K$$

THM: If (a, b) s.t. $a < x_{[1]} < x_{[2]} < \ldots < x_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(x))^2 dx$ then $\hat{f}$ is a natural cubic splines over (a, b) with knots at $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$.

2nd approach: Do not restrict $\mathcal{F}$ but
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \mathrm{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

$$= \underset{\boldsymbol{c} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ (\boldsymbol{Y} - \Psi \boldsymbol{c})^\top (\boldsymbol{Y} - \Psi \boldsymbol{c}) \ + \lambda \boldsymbol{c}^\top P \boldsymbol{c} \right\} \qquad P \in \mathbb{R}^K \times \mathbb{R}^K$$

THM: If (a, b) s.t. $a < x_{[1]} < x_{[2]} < \ldots < x_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(x))^2 dx$ then $\hat{f}$ is a natural cubic splines over (a, b) with knots at $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$.

Ex: For $\mathcal{P}(f) = \int (f'')^2$ then (j, ℓ)-entry of P is $\int_a^b \psi_j''(x) \psi_\ell''(x)\, dx$

2nd approach: $\hat{f} \in \mathcal{F}_K$, with $dim(\mathcal{F}_K) = K \approx n$
minimize RSS + roughness penalty

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \ \text{RSS}_\lambda(f) = \underset{f \in \mathcal{F}_K}{\text{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \mathcal{P}(f) \right\} \qquad \lambda > 0$$

$$= \underset{\boldsymbol{c} \in \mathbb{R}^K}{\text{argmin}} \left\{ (\boldsymbol{Y} - \Psi \boldsymbol{c})^\top (\boldsymbol{Y} - \Psi \boldsymbol{c}) \ + \lambda \boldsymbol{c}^\top P \boldsymbol{c} \right\} \qquad P \in \mathbb{R}^K \times \mathbb{R}^K$$

THM: If (a, b) s.t. $a < x_{[1]} < x_{[2]} < \ldots < x_{[n]} < b$, $\mathcal{F} = H^2(a, b)$, and $\mathcal{P}(f) = \int_a^b (f''(x))^2 dx$ then $\hat{f}$ is a natural cubic splines over (a, b) with knots at $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$.

Ex: For $\mathcal{P}(f) = \int (f'')^2$ then (j, ℓ)-entry of P is $\int_a^b \psi_j''(x) \psi_\ell''(x) \, dx$

For large datasets, it may be convenient to employ a mixed approach, where one considers a roughness penalty, but also reduces the dimensionality of the data estimation problem by considering a basis of dimension K, where K is still large but somehow smaller than n.

$$\hat{\boldsymbol{c}} \quad = \quad \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

Penalized smoothing
Expression of the estimator

$$\hat{\boldsymbol{c}} = \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y} = S\boldsymbol{Y}$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y} = S\boldsymbol{Y}$$

$$S = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top : \text{sub-projection operator } (S^\top S \neq S)$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y} = S\boldsymbol{Y}$$

$$S = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top : \text{sub-projection operator } (S^\top S \neq S)$$

$$df = tr(S) < K \qquad (\text{or } df = tr(S^\top S) \text{ or } df = tr(2S - S^\top S))$$

$$\hat{\boldsymbol{c}} = \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}} = \Psi \hat{\boldsymbol{c}} = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y} = S\boldsymbol{Y}$$

$$S = \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top : \text{sub-projection operator } (S^\top S \neq S)$$

$$df = tr(S) < K \qquad (\text{or } df = tr(S^\top S) \text{ or } df = tr(2S - S^\top S))$$

$$\hat{\sigma}^2 = \frac{1}{n - tr(S)} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

$$\mathrm{Var}[\hat{f}(x)] = \sigma^2 \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \psi(x)$$

$$\hat{f}(x) = \psi(x)^\top \hat{\boldsymbol{c}} = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}$$

$$\mathbb{E}\left[\hat{f}(x)\right] = \mathbb{E}\left[\psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \boldsymbol{Y}\right] = \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

$$\mathrm{Bias}[\hat{f}(x)] = f(x) - \mathbb{E}[\hat{f}(x)] = f(x) - \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \boldsymbol{c}$$

Sources of bias: discretization and penalization

(unless true f is s.t. $\mathcal{P}(f) = 0$, the penalty induces a bias)

$$\mathrm{Var}[\hat{f}(x)] = \sigma^2 \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \psi(x)$$

$$\widehat{\mathrm{Var}[\hat{f}(x)]} = \hat{\sigma}^2 \psi(x)^\top \left(\Psi^\top \Psi + \lambda P\right)^{-1} \Psi^\top \Psi \left(\Psi^\top \Psi + \lambda P\right)^{-1} \psi(x)$$

Smoothness parameter λ controls Bias-Variance trade-off.

Smoothness parameter λ controls Bias-Variance trade-off.

Selection of smoothness parameter λ: AiC, C_p Mallows, cross-validation, Generalized Cross Validation:

$$GCV(\lambda) = \frac{n}{(n - tr(S))^2} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$

Smoothness parameter λ controls Bias-Variance trade-off.

Selection of smoothness parameter λ: AiC, C_p Mallows, cross-validation, Generalized Cross Validation:

$$GCV(\lambda) = \frac{n}{(n - tr(S))^2} (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^\top (\boldsymbol{Y} - \hat{\boldsymbol{Y}})$$

Hopefully the chosen λ is close to that minimizing $\mathrm{MSE}[\hat{f}]$.

PENALIZED SMOOTHING

Under regularity conditions, as $n \to \infty$,
and $K(n) \to \infty$ and $\lambda(n) \to 0$ with appropriate rates

$$\mathrm{Bias}[\hat{f}(x)] \to 0 \qquad \text{and} \qquad \mathrm{Var}[\hat{f}(x)] \to 0$$

$$\mathrm{MSE}[\hat{f}(x)] \to 0$$

$$\hat{f}(x) \approx \textit{Gaussian}$$

Limiting Gaussian distrib justifies Wald type inference on $f(x)$.

Physics-informed penantly: $\mathcal{P}(f) = \int (Lf - u)^2$
$Lf = u$ encodes available problem-specific information