Fast cross validation and its theoretical validation

Felix Bartel

MMCS10

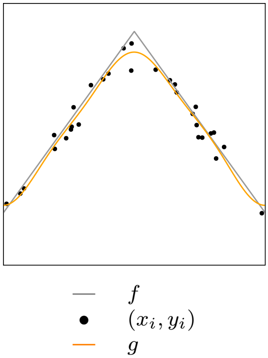
27th of June 2024

Mathematik! TU Chemnitz

Statistical learning

Given:

- $f\colon \mathbb{T} \to \mathbb{C}$
- $\boldsymbol{X} = \{x_1, \dots, x_n\} \subset \mathbb{T}$ uniformly random
- $\boldsymbol{y} = (f(x_1) + \varepsilon_1, \dots, f(x_n) + \varepsilon_n)^\top$ with
- $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(|\varepsilon_i|^2) = \sigma^2$



Goal: reconstruction $g\colon \mathbb{T} \to \mathbb{C}$ with small error

$$\|f - g\|_{L_2}^2 = \int_{\mathbb{T}} |f(x) - g(x)|^2 \,\mathrm{d}x$$

Least Squares approximation

Least squares approximation

The **least squares approximation** is given by

$$S_m^{\boldsymbol{X}} \boldsymbol{y} = \operatorname*{arg\,min}_{g \in V_m} \sum_{i=1}^{n} |y_i - g(x_i)|^2$$

for

$$V_m = \operatorname{span}\left\{ \exp(2\pi \mathrm{i} k \,\cdot),\, k = -\frac{m}{2}, \dots, \frac{m}{2} - 1 \right\}.$$

Implementation

We use the series expansion

$$S_m^{\boldsymbol{X}} \boldsymbol{y} = \sum_{k=-m/2}^{m/2-1} \hat{g}_k \exp(2\pi\mathrm{i}k\,\cdot) \quad \text{with} \quad \hat{\boldsymbol{g}} = (\boldsymbol{L}^*\boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{y}$$

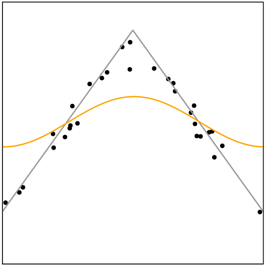
where

$$\boldsymbol{L} = \begin{pmatrix} \exp(2\pi\mathrm{i}(-\frac{m}{2})x_1) & \dots & \exp(2\pi\mathrm{i}(\frac{m}{2}-1)x_1) \\ \vdots & \ddots & \vdots \\ \exp(2\pi\mathrm{i}(-\frac{m}{2})x_n) & \dots & \exp(2\pi\mathrm{i}(\frac{m}{2}-1)x_n) \end{pmatrix} \in \mathbb{C}^{n\times m}.$$

Numerical example

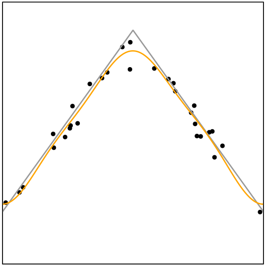
- approximation of B_2 from $n = 30$ points and noise level $\sigma^2 = 0.01$

$m = 2$



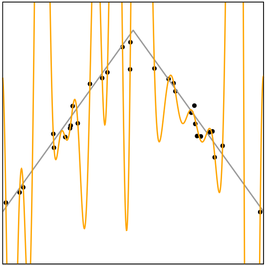
$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.13943$

$m = 7$



$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.00343$

$m = 29$



$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 10.3526$

Least Squares approximation

Theorem [B. '23]

Let
- $f\colon \mathbb{T} \to \mathbb{C}$,
- $\boldsymbol{X} = \{x_1, \dots, x_n\} \subset \mathbb{T}$ uniformly random,
- $\boldsymbol{y} = (f(x_1) + \varepsilon_1, \dots, f(x_n) + \varepsilon_n)^\top$ with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(|\varepsilon_i|^2) = \sigma^2$,
- $m \le \frac{n}{20 \log n}$, and
- $P_m f = \arg\min_{g \in V_m} \|f - g\|_{L_2}^2$.

Then we have with probability exceeding $1 - 3/n$

$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 \lesssim \|f - P_m f\|_{L_2}^2 + \sigma^2 \frac{m}{n} \,.$$

Optimal rates

- for $\|f - P_m f\|_{L_2} \lesssim m^{-s}$, we have the error behavior

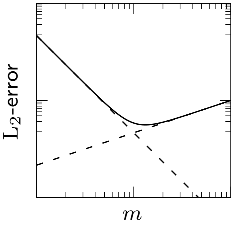
$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 \lesssim m^{-2s} + \sigma^2 \frac{m}{n}$$

- this yields the optimal number of frequencies

$$m^\star \sim n^{\frac{1}{2s+1}}$$

and the optimal error

$$\|f - S_{m^\star}^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 \lesssim n^{-\frac{2s}{2s+1}}$$



error behavior for fixed n

Cross validation

Purely data-driven method for approximating the L_2-error:

- validation set

validation	training

Cross validation

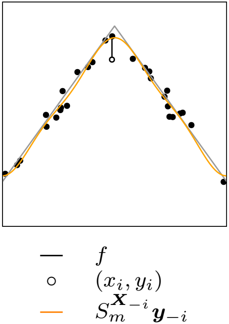
Purely data-driven method for approximating the L_2-error:

- validation set
- ”Probably the simplest and most widely used method for estimation prediction error is cross-validation.“ [Hastie '01]

<table>
<tr><td>validation</td><td colspan="4">training</td></tr>
<tr><td>training</td><td>validation</td><td colspan="3">training</td></tr>
<tr><td colspan="2">training</td><td>validation</td><td colspan="2">training</td></tr>
<tr><td colspan="3">training</td><td>validation</td><td>training</td></tr>
<tr><td colspan="4">training</td><td>validation</td></tr>
</table>

Cross validation [Stone '74]

1. compute reconstructions $S_m^{\boldsymbol{X}_{-i}} \boldsymbol{y}_{-i}$ omitting i-th sample



2. evaluate residual of $S_m^{\boldsymbol{X}_{-i}} \boldsymbol{y}_{-i}$ in i-th point x_i

$$\left| (S_m^{\boldsymbol{X}_{-i}} \boldsymbol{y}_{-i})(x_i) - y_i \right|^2$$

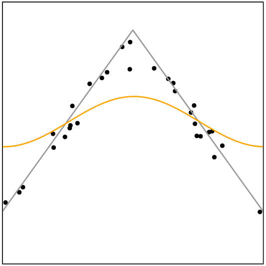
3. calculate mean value with respect to all points

$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) := \frac{1}{n} \sum_{i=1}^{n} \left| (S_m^{\boldsymbol{X}_{-i}} \boldsymbol{y}_{-i})(x_i) - y_i \right|^2$$

Numerical example

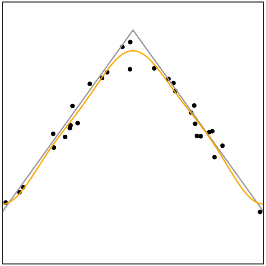
- approximation of B_2 from $n = 30$ points and noise level $\sigma^2 = 0.01$

$m = 2$



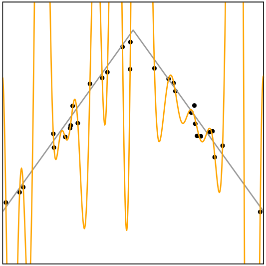
$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.13943$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.12856$$

$m = 7$



$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.00343$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.01160$$

$m = 29$



$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 10.3526$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 1.6295 \cdot 10^9$$

Cross validation – theoretical foundation

Lemma

Let
- $f\colon \mathbb{T} \to \mathbb{C}$,
- $\boldsymbol{X} = \{x_1, \dots, x_n\} \subset \mathbb{T}$ uniformly random,
- $\boldsymbol{y} = (f(x_1) + \varepsilon_1, \dots, f(x_n) + \varepsilon_n)^\top$ with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(|\varepsilon_i|^2) = \sigma^2$.

Then

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\Big(\operatorname{CV}(S_m^{\boldsymbol{X}}\boldsymbol{y})\Big) = \mathbb{E}_{\boldsymbol{X}_{-1},\boldsymbol{y}_{-1}}\Big(\|f - S_m^{\boldsymbol{X}_{-1}}\boldsymbol{y}_{-1}\|_{L_2}^2\Big) + \sigma^2 .$$

Cross validation – theoretical foundation

Theorem [B. '23]

Let $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$ be drawn uniformly random with $m \le \frac{n}{20 \log n}$.
Then we have with probability exceeding $1 - 8/\sqrt{n}$

$$\Big| \|f - S_m^{\boldsymbol{X}^{-1}} \boldsymbol{y}\|_{L_2}^2 - \mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) + \sigma^2 \Big| \lesssim \sqrt{\frac{m^3 \log n}{n}}\,.$$

In particular, when $\|f - P_I f\|_{L_2} \lesssim m^{-s}$ this gives $m^\star \sim n^{\frac{1}{2s+1}}$ and

$$\Big| \|f - S_{m^\star}^{\boldsymbol{X}^{-1}} \boldsymbol{y}\|_{L_2}^2 - \mathrm{CV}(S_{m^\star}^{\boldsymbol{X}} \boldsymbol{y}) + \sigma^2 \Big| \lesssim n^{-\frac{s-1}{2s+1}} \sqrt{\log n}\,.$$

Cross validation – fast computation

- **[Golub, Heath, Whaba '79]:** for $h_{i,i} = (\boldsymbol{L}(\boldsymbol{L}^*\boldsymbol{L})^{-1}\boldsymbol{L}^*)_{i,i}$,

$$\mathrm{CV}(S_m^{\boldsymbol{X}}\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(S_m^{\boldsymbol{X}}\boldsymbol{y})(\boldsymbol{x}^i) - y_i}{1-h_{i,i}}\right|^2$$

- **[Tasche, Weyrich '96]:** for the full tensor grid of frequencies and points

$$h_{i,i} = \frac{m}{n}$$

Cross validation – fast computation

- **[Golub, Heath, Whaba ’79]:** for $h_{i,i} = (\boldsymbol{L}(\boldsymbol{L}^*\boldsymbol{L})^{-1}\boldsymbol{L}^*)_{i,i}$,

$$\mathrm{CV}(S_m^{\boldsymbol{X}}\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(S_m^{\boldsymbol{X}}\boldsymbol{y})(\boldsymbol{x}^i) - y_i}{1-h_{i,i}}\right|^2$$

- **[Tasche, Weyrich ’96]:** for the full tensor grid of frequencies and points

$$h_{i,i} = \frac{m}{n}$$

Theorem [B., Hielscher, Potts ’19]

The above holds for all exact quadrature points, i.e., when

$$\boldsymbol{L}^*\boldsymbol{L} = n\boldsymbol{I}\,.$$

Cross validation – fast computation

- we define the **approximated cross-validation score** $\mathrm{FCV}(S_m^{\boldsymbol{X}}\boldsymbol{y})$

$$\mathrm{FCV}(S_m^{\boldsymbol{X}}\boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(S_m^{\boldsymbol{X}}\boldsymbol{y})(\boldsymbol{x}^i) - y_i}{1-\tilde{h}_{i,i}}\right|^2 \quad \text{for} \quad \tilde{h}_{i,i} = \frac{m}{n} \approx h_{i,i}$$

Theorem

[B. '23]

Let $\boldsymbol{x}^1, \dots, \boldsymbol{x}^n$ be drawn uniformly random with $m \le n/(20\log n)$.
Then we have with probability exceeding $1-2/n$

$$\frac{n}{2}\boldsymbol{I} \preccurlyeq \boldsymbol{L}^*\boldsymbol{L} \preccurlyeq \frac{3n}{2}\boldsymbol{I}$$

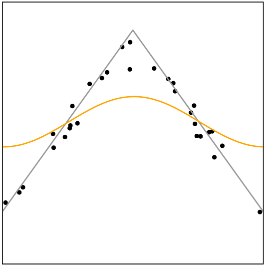
and

$$\left|\frac{\mathrm{CV}(S_m^{\boldsymbol{X}}\boldsymbol{y}) - \mathrm{FCV}(S_m^{\boldsymbol{X}}\boldsymbol{y})}{\mathrm{CV}(S_m^{\boldsymbol{X}}\boldsymbol{y})}\right| \le \frac{m/n}{(1-m/n)^2}\,.$$

Numerical example

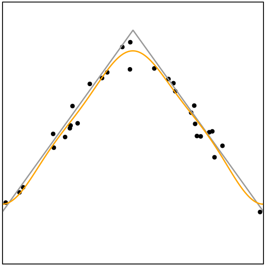
- approximation of B_2 from $n = 30$ points and noise level $\sigma^2 = 0.01$

$m = 2$



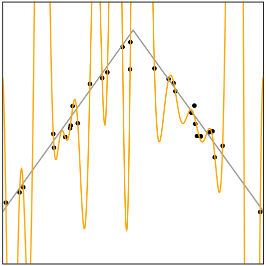
$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.13943$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.12856$$
$$\mathrm{FCV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.12831$$

$m = 7$



$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 0.00343$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.01160$$
$$\mathrm{FCV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.01160$$

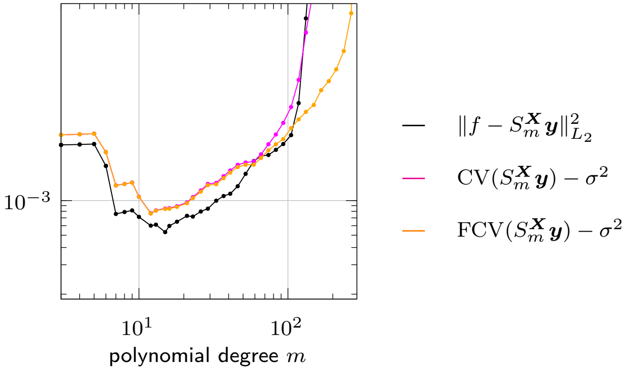
$m = 29$



$$\|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2 = 10.3526$$
$$\mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 1.6295 \cdot 10^9$$
$$\mathrm{FCV}(S_m^{\boldsymbol{X}} \boldsymbol{y}) = 0.79503$$

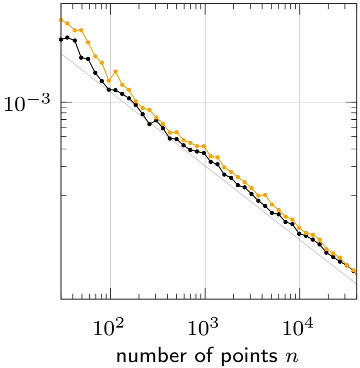
Numerical example

- approximation of B_2 from $n = 300$ points and noise level $\sigma^2 = 0.01$



Numerical example

- approximation of B_2 from different number of points and noise level $\sigma^2 = 0.01$



— $\|f - S_{m^\star}^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2$ with $m^\star = \arg\min_m \|f - S_m^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2$

— $\|f - S_{m^\blacktriangle}^{\boldsymbol{X}} \boldsymbol{y}\|_{L_2}^2$ with $m^\blacktriangle = \arg\min_m \mathrm{CV}(S_m^{\boldsymbol{X}} \boldsymbol{y})$

— $n^{-\frac{2s}{2s+1}}$ for $s = \frac{3}{2}$