Optimality of Cross-validation in Scattered Data Approximation

Felix Bartel Ralf Hielscher

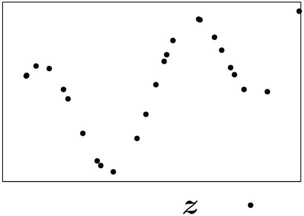
ESI Workshop on Tomographic Reconstructions and their Startling Applications

March 17, 2021

Mathematik!
TU Chemnitz

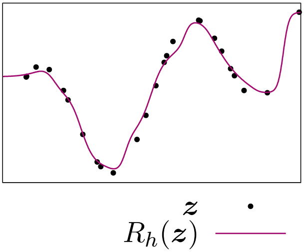
Setting

- data $\boldsymbol{z} = ((x_i, y_i))_{i=1}^n \in (\Omega \times Y)^n$
- (x_i, y_i) distributed according to ρ



Setting

- data $\boldsymbol{z} = ((x_i, y_i))_{i=1}^n \in (\Omega \times Y)^n$
- (x_i, y_i) distributed according to ρ
- reconstruction algorithm
  $R_h \colon (\Omega \times Y)^n \to Y^\Omega$



Which reconstruction algorithm R_h to choose?

- risk functional

$$\mathcal{E}(R_h(\boldsymbol{z})) = \int_{\Omega \times Y} |(R_h(\boldsymbol{z}))(x) - y|^2 \, \mathrm{d}\rho(x, y)$$

Which reconstruction algorithm R_h to choose?

- risk functional

$$\mathcal{E}(R_h(\boldsymbol{z})) = \int_{\Omega \times Y} |(R_h(\boldsymbol{z}))(x) - y|^2 \, \mathrm{d}\rho(x, y)$$

- cross-validation score (Golub, Heath, and Whaba '79)

$$\mathrm{CV}(\boldsymbol{z}, h) = \frac{1}{n} \sum_{i=1}^{n} |(R_h(\boldsymbol{z}_{-i}))(x_i) - y_i|^2$$

Goal: bound their difference w.h.p.

Concentration inequalities

Theorem (Hoeffding '63)

Let

- $Z_1, \dots, Z_n$ *independent rv's with values in* $[0,1]$ *and*
- $m = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^n Z_i\right\}$.

Then for $\varepsilon > 0$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i - m\right| > \varepsilon\right\} \leq 2\exp\left(-2n\varepsilon^2\right).$$

Concentration inequalities

A function $f\colon \Omega^n \to \mathbb{R}$ is $\boldsymbol{c}$-bounded on $\Xi \subset \Omega^n$ for $\boldsymbol{c} = (c_1, \ldots, c_n)$, iff

$$|f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i$$

for all $(z_1, \ldots, z_n), (z_1', \ldots, z_n') \in \Xi$, and $1 \leq i \leq n$.

Theorem (McDiarmid '89)

Let

- $Z_1, \dots, Z_n$ *be independent rv's with values in* Ω,
- $f : \Omega^n \to \mathbb{R}$ *be* $\boldsymbol{c}$*-bounded on* Ω^n, *and*
- $m = \mathbb{E}\{f(Z_1, \dots, Z_n)\}$.

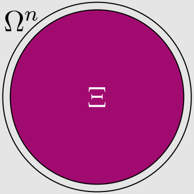
Then for $\varepsilon > 0$

$$\mathbb{P}\{|f(Z_1, \dots, Z_n) - m| > \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\|\boldsymbol{c}\|_2^2}\right).$$

Theorem (Extension of McDiarmid by Richard Combes '15)

Let

- $Z_1, \dots, Z_n$ *be independent rv's with values in* Ω,
- $f : \Omega^n \to \mathbb{R}$ *be* $\boldsymbol{c}$*-bounded on* $\Xi \subset \Omega^n$, *and*
- $m = \mathbb{E}\{f(Z_1, \dots, Z_n) | (Z_1, \dots, Z_n) \in \Xi\}$, *and*
- $\gamma = 1 - \mathbb{P}\{(Z_1, \dots, Z_n) \in \Xi\}$.



Then for $\varepsilon > \gamma \|\boldsymbol{c}\|_1$

$$\mathbb{P}\left\{|f(Z_1, \dots, Z_n) - m| > \varepsilon\right\} \leq 2\gamma + 2\exp\left(-\frac{2(\varepsilon - \gamma\|\boldsymbol{c}\|_1)^2}{\|\boldsymbol{c}\|_2^2}\right).$$

Choice of Ξ

Define $\Xi = \Xi(C_1)$ as the set of $\boldsymbol{z} = (z_1, \ldots, z_n)$ fulfilling

$$|R_h(\boldsymbol{z}_{-i})(x) - y| \leq C_1$$

for all $x \in \Omega, y \in Y$, and $1 \leq i \leq n$.

General framework

Risk functional $\boldsymbol{z} \mapsto \mathcal{E}(R_h(\boldsymbol{z}_{-i}))$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 4C_1^2\mathbb{1}$.

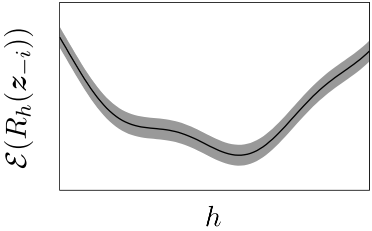
Cross-validation score $\boldsymbol{z} \mapsto \mathrm{CV}(\boldsymbol{z}, h)$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 5C_1^2\mathbb{1}$.

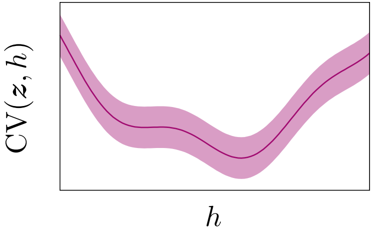
General framework

Risk functional $\boldsymbol{z} \mapsto \mathcal{E}(R_h(\boldsymbol{z}_{-i}))$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 4C_1^2\mathbb{1}$.

Cross-validation score $\boldsymbol{z} \mapsto \mathrm{CV}(\boldsymbol{z}, h)$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 5C_1^2\mathbb{1}$.

Now we apply Combes extension of McDiarmid.



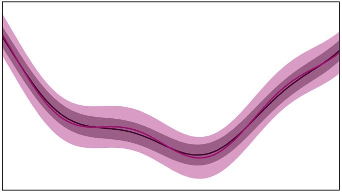


General framework

Lemma

For $\boldsymbol{Z}'$ representing $n-1$ samples and $\boldsymbol{Z}$ representing n samples element-wise distributed according to ρ, we have

$$\mathbb{E}_{\boldsymbol{Z}'}\left\{\mathcal{E}(R_h(\boldsymbol{Z}'))\right\} = \mathbb{E}_{\boldsymbol{Z}}\left\{\mathrm{CV}(\boldsymbol{Z}, h)\right\}.$$



General framework

Theorem (B., Hielscher '21)

Let

- $\boldsymbol{Z}$ *be element-wise distributed according to* ρ *with values in* $(\Omega \times Y)^n$ *and*
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Then for $\varepsilon > 2\gamma \max\{9nC_1^2, \|\operatorname{CV}(\cdot, h)\|_\infty + \|\mathcal{E}(R_h(\cdot))\|_\infty\}$ *and* $n \geq 3$

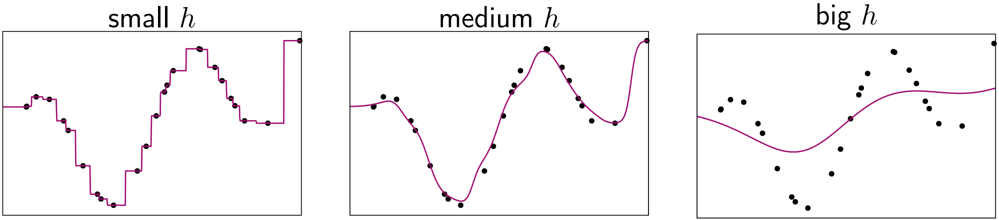
$$\mathbb{P}\left\{|\operatorname{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| > \varepsilon\right\}$$
$$\leq 2\gamma + 2\exp\left(-\left(\frac{\varepsilon}{12\sqrt{n}C_1^2} - \sqrt{2n}\gamma\right)^2\right).$$

Shepard's model for $\Omega = \mathbb{T}$ and $Y = \mathbb{R}$ without noise

$$R_h(\boldsymbol{z}) = \frac{\sum_{i=1}^{m} K_h(\cdot, x_i) f(x_i)}{\sum_{i=1}^{m} K_h(\cdot, x_i)}$$

with

$$K_h(x_1, x_2) = \max\{0, 1 - h|x_1 - x_2|\}$$



Shepard's model

We need estimates for

- $\| \mathrm{CV}(\cdot, h)\|_\infty$,
- $\|\mathcal{E}(R_h(\cdot))\|_\infty$,
- $C_1 = \max_{\boldsymbol{z}\in\Xi} \max_{x\in\Omega} |R_h(\boldsymbol{z})(x) - f(x)|$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Shepard's model

We need estimates for

- $\| \mathrm{CV}(\cdot, h)\|_\infty \leq 2\|f\|_\infty$,
- $\|\mathcal{E}(R_h(\cdot))\|_\infty \leq 2\|f\|_\infty$,
- $C_1 = \max_{\boldsymbol{z}\in\Xi} \max_{x\in\Omega} |R_h(\boldsymbol{z})(x) - f(x)|$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Shepard's model

We need estimates for

- $\| \operatorname{CV}(\cdot, h)\|_\infty \leq 2\|f\|_\infty$,
- $\|\mathcal{E}(R_h(\cdot))\|_\infty \leq 2\|f\|_\infty$,
- $C_1 = \max_{\boldsymbol{z}\in\Xi} \max_{x\in\Omega} |R_h(\boldsymbol{z})(x) - f(x)|$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Theorem (B., Hielscher '21)

Let $K_h(x, \cdot)$ be supported on $[x - 1/h, x + 1/h]$ and f be Lipschitz with constant L. Then $C_1 > L/h$ with probability

$$\gamma = 1 - \left(1 - \left(1 - \frac{2}{h}\right)^{n-2} \frac{2n + h - 4}{h}\right)^n.$$

Shepard's model

Theorem (B., Hielscher '21)

Let $\boldsymbol{Z}$ *represent* n *samples distributed according to* ρ*,* $K_h(x,\cdot)$ *be supported on* $[x-1/h, x+1/h]$*, and*

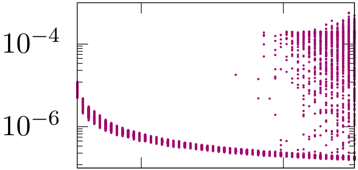
$$\gamma = 1 - \left(1 - \left(1 - \frac{2}{h}\right)^{n-2} \frac{2n+h-4}{h}\right)^{n}.$$

We then have for $\varepsilon > \gamma \max\{10nL^2/h^2, 8\|f\|_\infty\}$

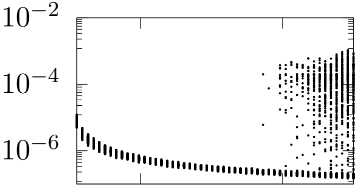
$$\mathbb{P}\left\{|\mathrm{CV}(\boldsymbol{Z},h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| > \varepsilon\right\} \leq 2\gamma + 2\exp\left(-\left(\frac{h^2\varepsilon}{12\sqrt{n}L^2} - \sqrt{2n}\gamma\right)^2\right).$$

Numerical experiment for Shepard's model

$\mathrm{CV}(\boldsymbol{z}, h)$

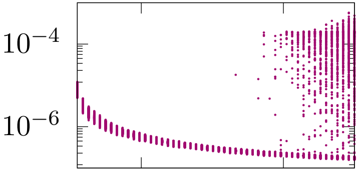


$\mathcal{E}(R_h(\boldsymbol{z}_{-1}))$

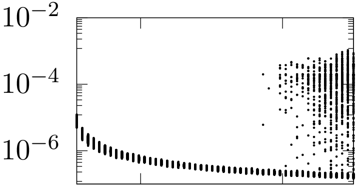


Numerical experiment for Shepard's model

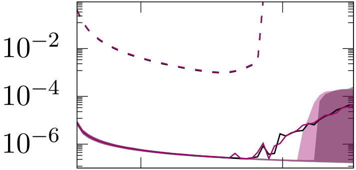
$\mathrm{CV}(\boldsymbol{z}, h)$



$\mathcal{E}(R_h(\boldsymbol{z}_{-1}))$



$\mathrm{CV}(\boldsymbol{z}, h)$ and $\mathcal{E}(R_h(\boldsymbol{z}_{-1}))$



$|\,\mathrm{CV}(\boldsymbol{z}, h) - \mathcal{E}(R_h(\boldsymbol{z}_{-1}))|$