Corrected generalized cross-validation for finite ensembles of penalized estimators

Takuya Koriyama

University of Chicago

September 4, 2024

- To appear in Journal of the Royal Statistical Society: Series B (2024)
- Joint work with Pierre C. Bellec (Rutgers), Jin-Hong Du (CMU), Kai Tan (Rutgers), and Pratik Patil (UC Berkeley).

Problem set up

- The response and feature $(y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^p$ $(i = 1, \ldots, n)$ are i.i.d. distributed.
- Consider the high-dimensional regime

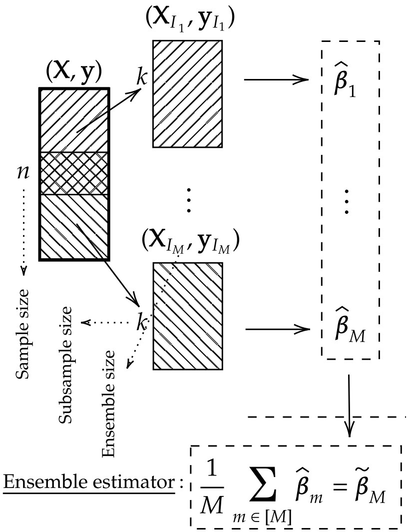
$$p/n \to \text{constant} \quad \text{for sample size } n \text{ and dimension } p.$$

- We are interested in an estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X})$ such that the prediction risk

$$\mathbb{E}\Big[\big(y_0 - \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}\big)^2 | \boldsymbol{y}, \boldsymbol{X}\Big] \quad \text{where} \quad (y_0, \boldsymbol{x}_0) =^d (y_i, \boldsymbol{x}_i)$$

is small.
- We consider ensemble estimators $\tilde{\boldsymbol{\beta}}$ (next slide).

Ensemble estimator $\tilde{\beta}$



We define ensemble estimator $\tilde{\boldsymbol{\beta}}$ as follows:

1. Subsampling

   $$(I_m)_{m=1}^M \overset{iid}{\sim} \text{Uniform}\{I \subset [n] : |I| = k$$

   for some integers $k \leq n$ and M.

2. Fit the penalized least square

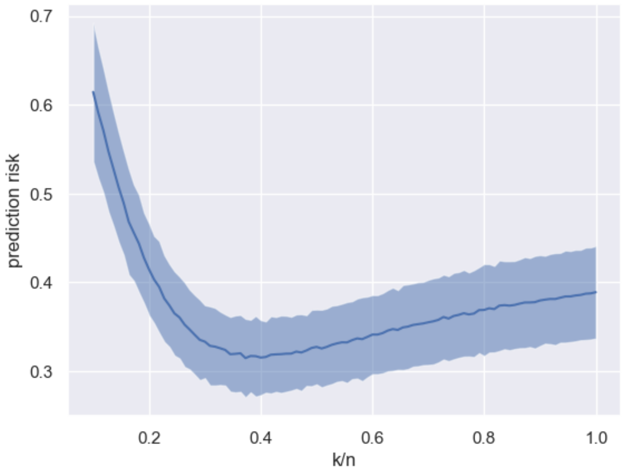
   $$\hat{\boldsymbol{\beta}}_m \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{k}\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\boldsymbol{\beta}\|^2 + g(\boldsymbol{\beta}$$

   for some convex function $g : \mathbb{R}^p \to \mathbb{R}$.

3. Ensemble $(\hat{\boldsymbol{\beta}}_m)_{m=1}^M$ together

   $$\tilde{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}_m.$$

Prediction risk is U-shape in sub-sample size k



Ensemble of Ridge estimators.

Equivalence between subsampling and regularization

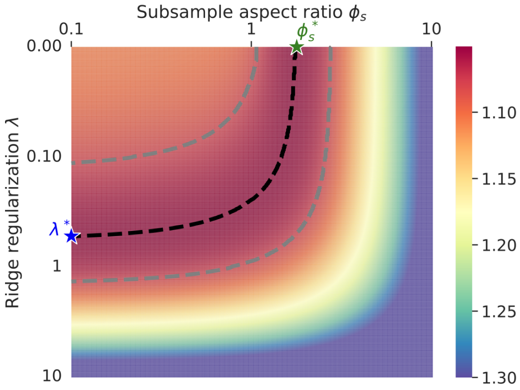


Figure 1 in Du et al. [2023].

Adaptive tuning of sub-sample size and penalty

(Recall) Ensemble estimator is $\tilde{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\beta}}_m$ where

$$\hat{\boldsymbol{\beta}}_m \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{k}\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\boldsymbol{\beta}\|^2 + g(\boldsymbol{\beta}), \quad I_m \sim \text{Uniform}\{I \subset [n] : |I| = k\}$$

for each $m \in [M]$.

- (Goal) Select sub-sample size k and penalty g in a data-driven manner so that the ensemble estimator $\tilde{\boldsymbol{\beta}}$ achieves a small prediction risk

$$\mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}] \quad \text{where} \quad (y_0, \boldsymbol{x}_0) =^d (y_i, \boldsymbol{x}_i)$$

- Since the prediction risk is not observable, we need some proxy;
  - L-fold cross-validation is biased.
  - Leave one out cross-validation is computationally hard due to high-dimension.
  - **Generalized cross-validation (GCV)**.

Generalized cross validation

For the penalized least square estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{y}, \boldsymbol{X}) \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + g(\boldsymbol{\beta}) \Big\},$$

Generalized cross-validation (GCV) of $\hat{\boldsymbol{\beta}}$ is defined by

$$(\text{GCV of } \hat{\boldsymbol{\beta}}) := \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n(1 - \hat{\mathrm{df}}/n)^2} \quad \text{where} \quad \hat{\mathrm{df}} := \mathrm{tr}\Big[\boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{y}}\Big].$$

Estimator $\hat{\boldsymbol{\beta}}$	Penalty $g(\boldsymbol{\beta})$	Degrees of freedom $\hat{\mathrm{df}}$
Lasso	$\lambda\|\boldsymbol{\beta}\|_1$	$\|\hat{S}\|$
Ridge	$\frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2$	$\mathrm{tr}\left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + n\mu\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^\top\right]$
Elastic net	$\lambda\|\boldsymbol{\beta}\|_1 + \frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2$	$\mathrm{tr}\left[\boldsymbol{X}_{\hat{S}}\left(\boldsymbol{X}_{\hat{S}}^\top\boldsymbol{X}_{\hat{S}} + n\mu\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}_{\hat{S}}^\top\right]$

Example of $\hat{\mathrm{df}}$ for specific penalties. Here, $\hat{S} = \{j \in [p] : e_j^\top \hat{\boldsymbol{\beta}} \neq 0\}$ and $\boldsymbol{X}_{\hat{S}}$ is the sub-matrix of $\boldsymbol{X}$ made of columns indexed in $\hat{S}$.

Consistency of Generalized cross-validation

Theorem (Prediction risk ≈ GCV)

$$\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big] \approx \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n(1 - \hat{\mathrm{df}}/n)^2}$$

	Penalty	Proof
Patil et al. [2021]	Ridge	Random Matrix Theory
Celentano et al. [2023]	Lasso	Convex Gaussian Min-Max Theo
Bellec and Shen [2022]	strongly convex	Second order Stein's formula

Naive GCV for ensemble estimator

For ensemble estimator $\tilde{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m$, we can think of the naive-GCV:

$$\text{naive-GCV} := \frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{n(1 - \tilde{\mathrm{df}}/n)^2} \quad \text{where} \quad \tilde{\mathrm{df}} = \mathrm{tr}\Big[\boldsymbol{X}\frac{\partial \tilde{\boldsymbol{\beta}}}{\partial \boldsymbol{y}}\Big]$$

Q. Does the naive-GCV consistently estimate the prediction risk?

$$\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big] \overset{?}{\approx} \text{naive-GCV}$$

A. No. The naive-GCV is inconsisntent.

Theorem

Under some regularity condition, there exists some positive constant $C \in (0,1)$ *such that*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\Big|\frac{\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big]}{\textit{naive-GCV}} - 1\Big| \geq C\Big) \geq C.$$

Overview of main result: corrected-GCV (CGCV)

$$\text{CGCV} := \underbrace{\frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{n(1 - \tilde{\mathrm{df}}/n)^2}}_{=\text{naive-GCV}} - \underbrace{\left(\frac{\tilde{\mathrm{df}}}{n - \tilde{\mathrm{df}}}\right)^2 \left(\frac{n}{k} - 1\right) \frac{1}{M^2} \sum_{m=1}^{M} \frac{\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\hat{\boldsymbol{\beta}}_m\|^2}{k(1 - \hat{\mathrm{df}}_m/k)^2}}_{=:\text{correction}}.$$

Theorem (Informal)

Either assumption (a) or (b) below is satisfied.

Assumption	Distribution	Response $y = f(\boldsymbol{x}, \epsilon)$	Penalty g
(a)	*Gaussian*	*Linear*	*strongly convex*
(b)	*Non-Gaussian*	*Nonlinear*	*Ridge*

Then, we have (*Prediction error*) $\approx$ *CGCV. More precisely,*

$$\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big] = \begin{cases} \mathit{CGCV} \cdot \big(1 + O_p(n^{-1/2})\big) & \text{under (a)} \\ \mathit{CGCV} + o_p(1) & \text{under (b)} \end{cases}$$

When correction term is small

The theorem implies

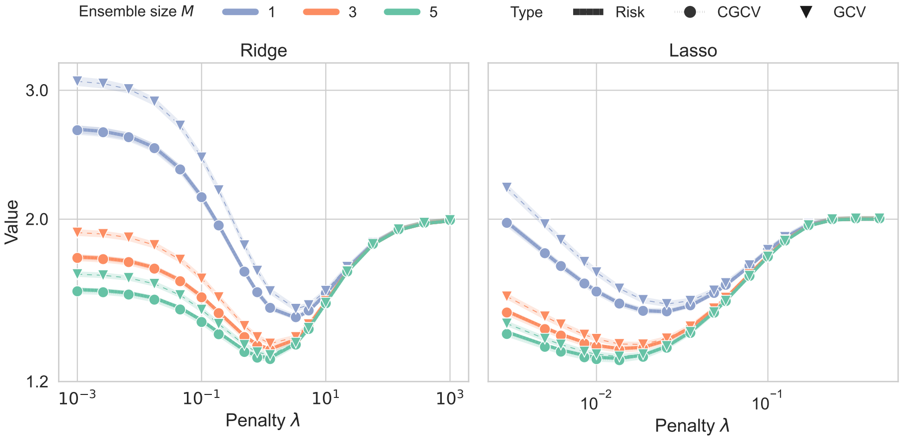
$$\text{(Prediction risk)} \approx \text{CGCV} = \underbrace{\frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{n(1 - \tilde{\mathrm{df}}/n)^2}}_{=\text{naive-GCV}} - \text{correction}$$

where

$$\text{correction} = \left(\frac{\tilde{\mathrm{df}}}{n - \tilde{\mathrm{df}}}\right)^2 \left(\frac{n}{k} - 1\right) \frac{1}{M^2} \sum_{m=1}^{M} \frac{\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\hat{\boldsymbol{\beta}}_m\|^2}{k(1 - \hat{\mathrm{df}}_m/k)^2}.$$

- Naive-GCV overestimates prediction risk.
- Correction term is exactly 0 when sub-sample size k is n.
- Correction term is $O(M^{-1})$.
  $\Rightarrow$ For infinite-ensemble $(M = \infty)$, the naive-GCV is consistent.

Comparison of CGCV and naive-GCV



Proof: Second order Stein's fomrula

Theorem (Bellec and Zhang [2021])

For almost surely differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ *and* $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$*, we have*

$$\mathbb{E}\Big[\big\{\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - \nabla \cdot \boldsymbol{f}(\boldsymbol{z})\big\}^2\Big] = \mathbb{E}\Big[\|\boldsymbol{f}(\boldsymbol{z})\|^2 + \operatorname{tr}\big\{\big(\nabla \boldsymbol{f}(\boldsymbol{z})\big)^2\big\}\Big].$$

- Many applications in single index model (Bellec, 2022), multinomial regression (Tan and Bellec, 2023), robust regression (Bellec and Koriyama, 2023).

Summary

- The naive-GCV is inconsistent to the prediction error of ensemble estimators.
- We proposed the corrected GCV and showed its consistency under Gaussian setting and non-Gaussian setting.
- arXiv:2310.01374

Reference I

P. C. Bellec and Y. Shen. Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*, 2022.

P. C. Bellec and C.-H. Zhang. Second-order stein: Sure for sure and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903, 2021.

M. Celentano, A. Montanari, and Y. Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

J.-H. Du, P. Patil, and A. K. Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*, 2023.

P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Appendix

Consistency of CGCV under assumption (a)

Assumption (a)

- $(y_i, \boldsymbol{x}_i)_{i=1}^n \in \mathbb{R} \times \mathbb{R}^p$ *are iid distributed according to*

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^* + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

  for some $\boldsymbol{\beta}^* \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \succ 0$ *and* $\sigma > 0$.
- g *is strongly convex with respect to* $\boldsymbol{\Sigma}$ [a] *(e.g., Ridge, Elastic net).*
- $p = O(k)$ *for sub-sample size* k.

[a] the map $\boldsymbol{\beta} \mapsto g(\boldsymbol{\beta}) - \mu \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta}$ is convex for some $\mu > 0$

Theorem (Prediction risk ≈ GCCV)

If the assumption (a) is satisfied, we have

$$\mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big] = \big[1 + O_P(n^{-1/2})\big] \cdot \textit{CGCV} \quad \textit{as } n \to \infty$$

Consistency of CGCV under Assumption (b)

Assumption (b)

- $g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|^2$ *for some* $\lambda > 0$.
- $\mathbb{E}[y_i] = 0$ *and* $\mathbb{E}[y_i^{4+\delta}] < +\infty$ *for some* $\delta > 0$.
- $\boldsymbol{x}_i =^d \Sigma^{1/2}\boldsymbol{z}_i$ *for some* $\boldsymbol{\Sigma} \succ 0$ *and* $\boldsymbol{z_i} \in \mathbb{R}^p$ *has iid entries such that* $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, *and* $\mathbb{E}[z_{ij}^{4+\delta}] < +\infty$.
- $p/n \to \phi \in (0, \infty)$, $p/k \to \psi \in [\phi, \infty]$.

Theorem

$$\mathbb{E}\left[(y_0 - \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\right] = \mathit{CGCV} + o_P(1) \quad \text{as } n \to +\infty$$

Proof outline

Prediction risk of $\hat{\boldsymbol{\beta}}$, denoted by $R(\hat{\boldsymbol{\beta}})$, can be written as

$$
\begin{aligned}
R(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\big[(y_0 - \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}})^2 | \boldsymbol{y}, \boldsymbol{X}\big] \\
&= \mathbb{E}\Big[\big\{\epsilon_0 - \boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\big\}^2 | \boldsymbol{y}, \boldsymbol{X}\Big] \quad \text{by } y_0 = \boldsymbol{x}_0^\top \boldsymbol{\beta}^* + \epsilon_0 \\
&= \sigma^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \qquad \text{by } \boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}),\ \epsilon_0 \sim \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

Thus, the prediction risk of the ensemble $\tilde{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^M \hat{\boldsymbol{\beta}}_m$ is given by

$$
\begin{aligned}
R(\tilde{\boldsymbol{\beta}}) &= \sigma^2 + \Big\{\big(\frac{1}{M}\sum_{m=1}^M \hat{\boldsymbol{\beta}}_m\big) - \boldsymbol{\beta}^*\Big\}\boldsymbol{\Sigma}\Big\{\big(\frac{1}{M}\sum_{m=1}^M \hat{\boldsymbol{\beta}}_m\big) - \boldsymbol{\beta}^*\Big\} \\
&= \frac{1}{M^2}\sum_{m=1}^M \sum_{\ell=1}^M \Big[\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*)\Big].
\end{aligned}
$$

Proof outline

The naive-GCV for $\tilde{\boldsymbol{\beta}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\beta}}_m$ is given by

$$\text{naive-GCV} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2}{n(1-\tilde{\mathrm{df}}/n)^2} = \frac{\frac{1}{M^2}\sum_{m=1}^{M}\sum_{\ell=1}^{M}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\ell)}{n(1-\tilde{\mathrm{df}}/n)^2}$$

Lemma

For all $m, \ell \in [M]$, *we have*

$$(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\ell) \approx \left[\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)^\top\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*)\right] \cdot D_{m\ell},$$

where $\quad D_{m\ell} = n - \mathrm{df}_m - \mathrm{df}_\ell + \dfrac{\hat{\mathrm{df}}_m\hat{\mathrm{df}}_\ell}{|I_m||I_\ell|}|I_m \cap I_\ell|.$

Using this lemma,

$$\text{naive-GCV} \approx \frac{1}{M^2}\sum_{m=1}^{M}\sum_{\ell=1}^{M}\left[\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)^\top\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*)\right] \cdot \frac{D_{m\ell}}{n(1-\tilde{\mathrm{df}}/n)^2}$$

Proof outline

Lemma (Concentration of $D_{m,\ell}$)

$$\frac{D_{m,\ell}}{n(1-\tilde{\mathrm{df}}/n)^2} \approx 1 + \mathbf{1}\{m=\ell\} \cdot (\frac{n}{k}-1)\frac{(\tilde{\mathrm{df}}/n)^2}{(1-\tilde{\mathrm{df}}/n)^2}.$$

$$\begin{aligned}
\text{naive-GCV} &\approx \frac{1}{M^2}\sum_{m=1}^{M}\sum_{\ell=1}^{M}\Big[\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*)\Big] \\
&\quad + \frac{1}{M^2}\sum_{m=1}^{M}\Big[\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)\Big] \cdot (\frac{n}{k}-1)\frac{(\tilde{\mathrm{df}}/n)^2}{(1-\tilde{\mathrm{df}}/n)^2} \\
&= R(\tilde{\boldsymbol{\beta}}) + \frac{1}{M^2}\sum_{m=1}^{M} R(\hat{\boldsymbol{\beta}}_m) \cdot (\frac{n}{k}-1)\frac{(\tilde{\mathrm{df}}/n)^2}{(1-\tilde{\mathrm{df}}/n)^2}
\end{aligned}$$

Obtain CGCV

We have shown that

$$R(\tilde{\boldsymbol{\beta}}) \approx \text{naive-GCV} - \frac{1}{M^2}(\frac{n}{k} - 1)\frac{(\tilde{\mathrm{df}}/n)^2}{(1-\tilde{\mathrm{df}}/n)^2}\sum_{m=1}^{M} R(\hat{\boldsymbol{\beta}}_m).$$

Using (prediction risk of $\hat{\boldsymbol{\beta}}_m$) $\approx$ (GCV of $\hat{\boldsymbol{\beta}}_m$ fitted on $(y_i, \boldsymbol{x}_i)_{i \in I_m}$)

$$R(\hat{\boldsymbol{\beta}}_m) \approx \frac{\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\hat{\boldsymbol{\beta}}_m\|^2}{k(1-\hat{\mathrm{df}}_m/k)},$$

we are left with

$$R(\tilde{\boldsymbol{\beta}}) \approx \underbrace{(\text{naive-GCV}) - \frac{1}{M^2}(\frac{n}{k} - 1)\frac{(\tilde{\mathrm{df}}/n)^2}{(1-\tilde{\mathrm{df}}/n)^2}\sum_{m=1}^{M} \frac{\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\hat{\boldsymbol{\beta}}_m\|^2}{k(1-\mathrm{df}_m/k)^2}}_{=\text{CGCV}}$$

Proof of Lemma 1: Second order Stein's formula

Recall that Lemma 1 claims

$$\Big(\sigma^2 + (\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}^*)^\top \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}^*)\Big) \cdot D_{m\ell} \approx (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_m)^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\ell),$$

where $D_{m\ell} = n - \mathrm{df}_m - \mathrm{df}_\ell + \frac{\hat{\mathrm{df}}_m \hat{\mathrm{df}}_\ell}{|I_m||I_\ell|} |I_m \cap I_\ell|$.

Theorem (Bellec and Zhang [2021])

For almost surely differentiable function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}_n, \boldsymbol{I}_n)$, we have

$$\mathbb{E}\Big[\big\{\boldsymbol{z}^\top \boldsymbol{f}(\boldsymbol{z}) - \nabla \cdot \boldsymbol{f}(\boldsymbol{z})\big\}^2\Big] = \mathbb{E}\Big[\|\boldsymbol{f}(\boldsymbol{z})\|^2 + \mathrm{tr}\big\{\big(\nabla \boldsymbol{f}(\boldsymbol{z})\big)^2\big\}\Big].$$