Supplementary materials for "The restricted consistency property of leave-n_v-out cross-validation for high-dimensional variable selection"

Yang Feng and Yi Yu

Columbia University and University of Bristol

The supplementary material includes all the technical details and additional simulation results.

A.1 Additional lemmas and proofs

The following Lemma is adapted from Lalley (2013). It helps us to develop the asymptotic theory where N, the size of the candidate models, is allowed to diverge with the sample size.

Lemma 1 (Gaussian concentration)**.** *Let γ be the standard Gaussian probability measure on $\mathbb{R}^n$ (that is, the distribution of a $\mathcal{N}(0, I_n)$ random vector), and let $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz in each variable separately relative to the Euclidean metric, with Lipschitz constant c. Then for every $t > 0$,*

$$\gamma\{|F - E_\gamma(F)| \geq t\} \leq 2\exp\left(-\frac{t^2}{c^2\pi^2}\right).$$

Lemma 2. *With $p < n$, let $\tilde{\beta}$ be the MLEs of a generalized linear model. Assume the penalty function $p(\cdot)$ is separable, and assume Conditions 1 - 6 hold. Furthermore, assume $n_c \to \infty$ and $n_c/n \to 0$ as $n \to \infty$, and the size of the splits K satisfies*

$$K^{-1}n_c^{-2}n^2 \to 0.$$

Then, CV(n_v) with K times subsampling is restricted model selection consistent.

Proof of Lemma 2. Due to the properties of generalized linear models with canonical parameter, we have

$$E(y_i \mid \boldsymbol{x}_i) = \dot{b}(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \quad \sigma_i^2 = a(\phi)\ddot{b}(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \quad i = 1, \cdots, n,$$

and define $\sigma^2 = (1/n)\sum_{i=1}^n \sigma_i^2$. The target is to select the model that minimizes the loss

$$\tilde{\Gamma}_\alpha = \frac{1}{Kn_v}\sum_{s\in\mathcal{S}}\Big\{-\boldsymbol{y}_s^\top(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha}) + \mathbf{1}^\top b(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha})\Big\}, \tag{1}$$

where $\mathcal{S}$ represents the collection of validation sets in different splits and $\mathbf{1}$ is an all-one vector.

Denote $E_{\mathcal{S}}$ and $\text{var}_{\mathcal{S}}$ as the expectation and variance with respect to the random selection of $\mathcal{S}$. By using the equality

$$E_{\mathcal{S}}\left(\frac{1}{r}\sum_{s\in\mathcal{S}} a_s\right) = \binom{n}{n_v}^{-1} \sum_{s\in \text{ all } s} E(a_s),$$

rewriting (1), and denoting $\ell_s(\boldsymbol{\beta}) = \boldsymbol{y}_s^\top(X_s\boldsymbol{\beta}) - \mathbf{1}^\top b(X_s\boldsymbol{\beta})$ and $\ell_n(\tilde{\boldsymbol{\beta}}_\alpha) = \boldsymbol{y}^\top(X_\alpha\tilde{\boldsymbol{\beta}}_\alpha) - \mathbf{1}^\top b(X_\alpha\tilde{\boldsymbol{\beta}}_\alpha)$, we have

$$\begin{aligned}
E_{\mathcal{S}}(\tilde{\Gamma}_\alpha) =& E_{\mathcal{S}}\left(-\frac{1}{Kn_v}\sum_{s\in\mathcal{S}}\ell_s(\boldsymbol{\beta}^o)\right) + E_{\mathcal{S}}\left(\frac{1}{Kn_v}\sum_{s\in\mathcal{S}}\left\{\ell_s(\boldsymbol{\beta}^o) - \left(\boldsymbol{y}_s^\top(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha) - \mathbf{1}^\top b(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha)\right)\right\}\right)\\
&+ E_{\mathcal{S}}\left(\frac{1}{Kn_v}\sum_{s\in\mathcal{S}}\left\{\left(\boldsymbol{y}_s^\top(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha) - \mathbf{1}^\top b(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha)\right) - \left(\boldsymbol{y}_s^\top(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha}) - \mathbf{1}^\top b(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha})\right)\right\}\right)\\
=& E\left(-\frac{1}{n}\ell_n(\boldsymbol{\beta}^o) + \frac{1}{n}\left(\ell_n(\boldsymbol{\beta}^o) - \ell_n(\tilde{\boldsymbol{\beta}}_\alpha)\right)\right.\\
&\left. + \binom{n}{n_v}^{-1}\sum_{s\in \text{ all } s}\frac{1}{n_v}\left\{\boldsymbol{y}_s^\top(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha - X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha}) - \mathbf{1}^\top\left(b(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha) - b(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha})\right)\right\}\right)\\
=& -\frac{1}{n}E(\ell_n(\boldsymbol{\beta}^o)) + E(A_{\alpha1}) + \binom{n}{n_v}^{-1}\sum_{s\in \text{ all } s}E(A_{\alpha2,s}).
\end{aligned}$$

For different α, $E(\ell_n(\boldsymbol{\beta}^o))$ stays the same, so we only need to focus on $A_{\alpha1}$ and $A_{\alpha2,s}$.

From Wilks' theorem, it is known that, if $\alpha\in\mathcal{A}\setminus\mathcal{A}_c$, as $n\to\infty$, we have $A_{\alpha1}\xrightarrow{\mathcal{D}}(1/2)\chi^2(k_\alpha)$, where $k_\alpha = d_0 - d_{\alpha0}$, $d_{\alpha0} = |\{j : \beta_j\in\alpha\cap\alpha_0\}|$, i.e., k_α is the number of false negatives. This means $E(A_{\alpha1}) = k_\alpha$; otherwise, $E(A_{\alpha1}) = O(1/n)$.

For any s,

$$\begin{aligned}
&\mathbf{1}^\top\left(b(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha) - b(X_s^\alpha\tilde{\boldsymbol{\beta}}_{s^c,\alpha})\right) = \left(\dot{b}(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha)\right)^\top X_s^\alpha(\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha})\\
&-\frac{1}{2}(\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha})^\top(X_s^\alpha)^\top\ddot{b}(X_s^\alpha\tilde{\boldsymbol{\beta}}_\alpha)X_{s,\alpha}(\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha}) + o(1).
\end{aligned}$$

Define $u_{s^c}(\boldsymbol{\gamma}) = (1/n_c)(X_{s^c}^\alpha)^\top\left(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha\boldsymbol{\gamma})\right)$, then $\tilde{\boldsymbol{\beta}}_{s^c,\alpha}$ is the solution to $u_{s^c}(\boldsymbol{\gamma}) = 0$. By Taylor expansion, we get

$$\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha} = \left(\dot{u}_{s^c}(\tilde{\boldsymbol{\beta}}_\alpha)\right)^{-1}u_{s^c}(\tilde{\boldsymbol{\beta}}_\alpha)(1+o(1)),$$

where $\dot{u}_{s^c}(\tilde{\boldsymbol{\beta}}_\alpha) = -(1/n_c)(X_{s^c}^\alpha)^\top\ddot{b}(X_{s^c}^\alpha\tilde{\boldsymbol{\beta}}_\alpha)X_{s^c}^\alpha$.

Define $D_{s,\alpha} = \ddot{b}^{1/2}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha) X_{s^c}^\alpha$, then

$$\begin{aligned}
A_{\alpha 2,s} &= \frac{1}{n_v}\big(\boldsymbol{y}_s - \dot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)^\top X_s^\alpha (\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha}) \\
&+ \frac{1}{2n_v}(\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha})^\top (X_s^\alpha)^\top \ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha) X_s^\alpha (\tilde{\boldsymbol{\beta}}_\alpha - \tilde{\boldsymbol{\beta}}_{s^c,\alpha}) + o(1/n_v) \\
&= \frac{1}{n_v}\big(\boldsymbol{y}_s - \dot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)^\top X_s^\alpha \big(\dot{u}_{s^c}(\tilde{\boldsymbol{\beta}}_\alpha)\big)^{-1} u_{s^c}(\tilde{\boldsymbol{\beta}}_\alpha) + o(1/n_v) \\
&+ \frac{1}{2n_v}\big(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)^\top \big(\ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)^{-1/2}\big) D_{s,\alpha}\big(D_{s,\alpha}^\top D_{s,\alpha}\big)^{-1} \\
&\times \big((X_s^\alpha)^\top \ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha) X_s^\alpha\big)\big((X_{s^c}^\alpha)^\top \ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha) X_{s^c}^\alpha\big)^{-1} \\
&\times D_{s,\alpha}^\top \big(\ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)^{-1/2}\big)\big(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)(1 + o(1)) \\
&= B_\alpha + C_\alpha.
\end{aligned}$$

By plugging in the expansion form of $\dot{u}_{s^c}(\cdot)$ and $u_{s^c}(\cdot)$,

$$B_\alpha = -\frac{1}{n_v}\big(\boldsymbol{y}_s - \dot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)^\top X_s^\alpha \big((X_{s^c}^\alpha)^\top \ddot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha) X_{s^c}^\alpha\big)^{-1} (X_{s^c}^\alpha)^\top \big(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)(1 + o(1)).$$

From Conditions 5 and 6, straight calculations lead to

$$E(B_\alpha) = 0, \quad \mathrm{var}(B_\alpha) = d_\alpha a(\phi)(n_c n_v)^{-1/2}(1 + o(1)).$$

For C_α we have,

$$\begin{aligned}
C_\alpha &= \frac{1}{2n_c}\big(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)^\top \big(\ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)^{-1/2}\big) D_{s,\alpha}\big(D_{s,\alpha}^\top D_{s,\alpha}\big)^{-1} D_{s,\alpha}^\top \\
&\times \big(\ddot{b}(X_s^\alpha \tilde{\boldsymbol{\beta}}_\alpha)^{-1/2}\big)\big(\boldsymbol{y}_{s^c} - \dot{b}(X_{s^c}^\alpha \tilde{\boldsymbol{\beta}}_\alpha)\big)(1 + o(1)).
\end{aligned}$$

Thus, after taking expectation we have,

$$E(A_{\alpha 2,s}) = d_\alpha a(\phi)/n_c + o(1/n_c).$$

If $\alpha \in \mathcal{A} \setminus \mathcal{A}_c$,

$$\tilde{\Gamma}_{\alpha_*} - \tilde{\Gamma}_\alpha = \frac{1}{n}\big(\ell_n(\tilde{\boldsymbol{\beta}}_{\alpha_*}) - \ell_n(\tilde{\boldsymbol{\beta}}_\alpha)\big) + O(1/n_c).$$

From Lemma 1 and Condition 3, by exploiting Gaussian concentration, $\forall \varepsilon > 0$, we have

$$R \cdot \mathrm{pr}\left\{ n_c \left| \max_{\alpha \in \mathcal{A}\setminus\mathcal{A}_c} \left| \frac{1}{n}\big(\ell_n(\tilde{\boldsymbol{\beta}}_{\alpha_*}) - \ell_n(\tilde{\boldsymbol{\beta}}_\alpha)\big) \right| - E\left(\max_{\alpha \in \mathcal{A}\setminus\mathcal{A}_c} \left| \frac{1}{n}\big(\ell_n(\tilde{\boldsymbol{\beta}}_{\alpha_*}) - \ell_n(\tilde{\boldsymbol{\beta}}_\alpha)\big) \right| \right) \right| > \varepsilon \right\} \to 0.$$

The parallel result for $\alpha \in \mathcal{A}_c$ but $\alpha \neq \alpha_*$ holds similarly. Therefore, as $n \to \infty$, $\mathrm{pr}\{\hat{\alpha} \in \alpha_*\} \to 1$.

□

A.2 Additional numerical results

We conducted an additional simulation for the setting in Example 1(i) when $\rho = -0.5$ with the results summarized in Table 1. In this case, CV(n_v) works very well compared with other methods and we skip the detailed discussion since the message is very similar to the cases of $\rho = 0$ and $\rho = 0.5$.

Table 1: Comparisons in linear regression with $\rho = -0.5$. Results are reported in the form of mean (standard error). FP, false positive; FN, false negative; PE, prediction error.

Method		$\rho = -0.5$	
Lasso	FP	FN	PE
CV(n_v)	0.03(0.02)	0.02(0.01)	1.01(0.01)
K-fold	30.53(2.84)	0.00(0.00)	1.09(0.01)
1SE	1.54(0.21)	0.00(0.00)	1.15(0.01)
AIC	469.97(1.39)	0.00(0.00)	1.38(0.01)
BIC	2.18(0.17)	0.00(0.00)	1.12(0.01)
EBIC	0.91(0.10)	0.00(0.00)	1.13(0.01)
SCAD	FP	FN	PE
CV(n_v)	0.06(0.03)	0.01(0.01)	1.01(0.01)
K-fold	24.48(2.70)	0.00(0.00)	1.03(0.01)
1SE	0.30(0.09)	0.00(0.00)	1.08(0.01)
AIC	25.20(2.02)	0.05(0.03)	1.09(0.03)
BIC	0.70(0.09)	0.05(0.03)	1.10(0.03)
EBIC	0.16(0.04)	0.05(0.03)	1.11(0.03)
MCP	FP	FN	PE
CV(n_v)	0.02(0.01)	0.00(0.00)	1.01(0.01)
K-fold	4.76(0.82)	0.00(0.00)	1.02(0.01)
1SE	0.04(0.04)	0.00(0.00)	1.07(0.01)
AIC	77.29(0.96)	0.00(0.00)	1.15(0.01)
BIC	0.52(0.11)	0.00(0.00)	1.02(0.01)
EBIC	0.06(0.03)	0.00(0.00)	1.02(0.01)

Bibliography

LALLEY, S. P. (2013). Concentration inequalities. URL `http://www.stat.uchicago.edu/$\sim$lalley/Courses/386/Concentration.pdf`.

Department of Statistics, Columbia University

E-mail: yang.feng@columbia.edu

School of Mathematics, University of Bristol

E-mail: y.yu@bristol.ac.uk