Supplemantary Material for

Conditional Selective Inference for Robust Regression and Outlier Detection using Piecewise-Linear Homotopy Continuation

by Toshiaki Tsukurimichi, Yu Inatsu, Vo Nguyen Le Duy and Ichiro Takeuchi.

Appendix B: Breakpoints in Huber regression

In this section, we give the lemma for calculating breakpoints. Let $\mathcal{A}_z$, $S_{\mathcal{A}_z^c}$, $\hat{\boldsymbol{r}}(z)$, $\hat{\boldsymbol{u}}_{\mathcal{A}_z}(z)$, $\boldsymbol{h}_{\mathcal{A}_z^c}(z)$, $\psi(z)$ and $\gamma(z)$ be the same definitions as in Appendix Appendix A:. Then, the following lemma holds:

Lemma 5 *Consider a real value z. Then, $\mathcal{A}_z = \mathcal{A}_{z'}$ for any real value z' in the interval $[z, z + t_z]$, where $z + t_z$ is the value of transition point,*

$$t_z = \min\{t_z^1, t_z^2\},$$
$$t_z^1 = \min_{j \in \mathcal{A}_z^c} \left(-\frac{(S_{\mathcal{A}_z^c}\hat{\boldsymbol{r}}(z) - \boldsymbol{h}_{\mathcal{A}_z^c}(z))_j}{(S_{\mathcal{A}_z^c}\psi(z))_j} \right)_{++} \quad \textit{and} \quad t_z^2 = \min_{j \in \mathcal{A}_z} \left(-\frac{(\hat{u}_{\mathcal{A}_z}(z))_j}{\gamma_j(z)} \right)_{++}.$$

Here, $(a)_{++} = a$ if $a \geq 0$, and otherwise $(a)_{++} = +\infty$.

Proof We first show how to derive t_z^1. From (29), we have

$$\hat{\boldsymbol{r}}(z') = \hat{\boldsymbol{r}}(z) + \psi(z) \times (z' - z).$$

Then, we need to guarantee

$$\begin{aligned} S_{\mathcal{A}_z^c}\hat{\boldsymbol{r}}(z) - \boldsymbol{h}_{\mathcal{A}_z^c}(z) &\leq 0, \\ S_{\mathcal{A}_z^c}(\hat{\boldsymbol{r}}(z') + \psi(z) \times (z' - z)) - \boldsymbol{h}_{\mathcal{A}_z^c}(z) &\leq 0, \\ S_{\mathcal{A}_z^c}\psi(z) \times (z' - z) &\leq -(S_{\mathcal{A}_z^c}\hat{\boldsymbol{r}}(z) - \boldsymbol{h}_{\mathcal{A}_z^c}(z)). \end{aligned} \tag{36}$$

The right hand side of (36) is positive since $S_{\mathcal{A}_z^c}\hat{\boldsymbol{r}}(z) - \boldsymbol{h}_{\mathcal{A}_z^c}(z) \leq 0$. Therefore, satisfying equation (36) implies that

$$z' - z \leq \min_{j \in \mathcal{A}_z^c} \left(-\frac{(S_{\mathcal{A}_z^c}\hat{\boldsymbol{r}}(z) - \boldsymbol{h}_{\mathcal{A}_z^c}(z))_j}{(S_{\mathcal{A}_z^c}\psi(z))_j} \right)_{++} = t_z^1.$$

Next, we show how to derive t_z^2. From (30), we have

$$\hat{\boldsymbol{u}}_{\mathcal{A}_z}(z') = \hat{\boldsymbol{u}}_{\mathcal{A}_z}(z) + \gamma(z) \times (z' - z).$$

Thus, noting that $\hat{\boldsymbol{u}}_{\mathcal{A}_z}(z') \geq 0$ we need guarantee

$$\hat{\boldsymbol{u}}_{\mathcal{A}_z}(z') = \hat{\boldsymbol{u}}_{\mathcal{A}_z}(z) + \gamma(z) \times (z' - z) \geq 0. \tag{37}$$

Hence, satisfying equation (37) means that

$$z' - z \leq \min_{j \in \mathcal{A}_z} \left(-\frac{(\hat{u}_{\mathcal{A}_z}(z))_j}{\gamma_j(z)} \right)_{++} = t_z^2.$$

Appendix C: Huberized Lasso

According to Equation (8) in Chen and Bien (2020), we consider the following optimization problem

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{u}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta} - \boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1, \tag{38}$$

given $X = (x_1, ..., x_n)^\top$ and $\boldsymbol{y} = (y_1, ..., y_n)^\top$. Following Sec 3.2 in the supplementary of Chen and Bien (2020), the optimization in (38) can be transformed to

$$\hat{\boldsymbol{u}} = \underset{\boldsymbol{u}\in\mathbb{R}^n}{\arg\min} \frac{1}{2}\|\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1,$$

where $\tilde{\boldsymbol{y}} = P_X^{\perp}\boldsymbol{y}$ and $\tilde{X} = P_X^{\perp}$. We can obtain $\hat{\boldsymbol{u}}$ in (38) by using Lasso algorithm $\mathcal{A}$. Then, the set of the observed outliers is defined as

$$\mathcal{A}(\boldsymbol{y}) = \{j : \hat{u}_j \neq 0\}.$$

Finally, the inference for a selected outlier is defined as follows

$$\boldsymbol{\eta}^\top \boldsymbol{Y} \mid \{\mathcal{A}(\boldsymbol{Y}) = \mathcal{A}(\boldsymbol{y}), \boldsymbol{q}(\boldsymbol{Y}) = \boldsymbol{q}(\boldsymbol{y})\}.$$

Unfortunately, as pointed out in Lee et al. (2016), characterizing $\mathcal{A}(\boldsymbol{Y}) = \mathcal{A}(\boldsymbol{y})$ in (39) is computationally intractable because we have to consider $2^{|\mathcal{A}(\boldsymbol{y})|}$ possible sign vectors. As suggested in Lee et al. (2016), we need to consider inference conditional not only on the selected features but also on their signs to overcome the aforementioned issue. Specifically, let $\boldsymbol{s}(\boldsymbol{y})$ denote the sign vector of the selected features when applying Lasso on $\boldsymbol{y}$, the conditional inference we need to focus is

$$\boldsymbol{\eta}^\top \boldsymbol{Y} \mid \{\mathcal{A}(\boldsymbol{Y}) = \mathcal{A}(\boldsymbol{y}), \boldsymbol{s}(\boldsymbol{Y}) = \boldsymbol{s}(\boldsymbol{y}), \boldsymbol{q}(\boldsymbol{Y}) = \boldsymbol{q}(\boldsymbol{y})\}. \tag{39}$$

However, additionally considering the signs leads to low statistical power because of *over-conditioning*. This is well-known as the major drawback in SI literature.