# Supplementary material

# SIGLE: valid Selective Inference procedure for Generalized Linear Lasso

Quentin Duchemin
Swiss Data Science Center
École polytechnique fédérale de Lausanne
1015, Lausanne, Switzerland
quentin.duchemin@epfl.ch
&
Yohann De Castro
Univ. Lyon, École Centrale de Lyon, CNRS UMR 5208
Institut Camille Jordan
36 Avenue Guy de Collongue, 69134 Écully, France
Institut Universitaire de France (IUF)
yohann.de-castro@ec-lyon.fr

April 2023

**Guidelines for the Supplementary.**

- **Section 1: Confidence region.**

  We make use of the conditional Central Limit Theorems (CLTs) presented in our main paper to show how one can get confidence region using SIGLE.

- **Section 2: Side notes about SIGLE.**

  In this section, we put in the limelight more advanced questions related to the methods proposed in this paper. We start by proposing a reinterpretation of the methods presented in this paper when we consider that the model is misspecified in the sense that the observations $y_i$'s have not been initially generated from the GLM presented in Section 1.1 of the manuscript. In a second and last part, we focus on the diffeomorphism $\Psi$ which is a key ingredient involved in SIGLE. We provide a new perspective on $\Psi$ relying on tools from convex analysis before explaining how we compute in practice quantities of the form $\Psi(\rho)$ that are involved in the algorithms presented in this paper.

- **Section 3: Inference conditional on the signs.**

  We start by a gentle introduction to the Leftover Fisher information. Introduced in Fithian et al. [2014], this concept allows to show that conditioning on both the selected support and the signs of the dual variable (i.e. $E_M^{S_M}$ with the notations of Section 1 of the manuscript) lead in general to wider (and thus worse) confidence intervals. Our goal is to use this preliminary to discuss with more details the method proposed by Taylor and Tibshirani [2018]. In particular, we explain that the former approach is doomed to work conditional to $E_M^{S_M}$ since the usual trick used in the linear model to condition only on $E_M$ does not apply for an arbitrary GLM.

# 1 Confidence region

## 1.1 Asymptotic confidence region in the selected model

### 1.1.1 Main result

In the previous section, we proved that the MLE $\widehat{\theta}$ satisfies a CLT with a centering vector that is not the parameter of interest $\theta^*$. Two questions arises at this point.

1. How can we compute a relevant estimate for $\theta^*$?
2. Can we provide theoretical guarantees regarding this estimate?

Proposition 1 answers both questions. It provides a valid confidence region with asymptotic level $1-\alpha$ for any estimate $\theta^\bigstar$ of $\theta^*$ where the width of the confidence region is asymptotically driven by $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. The proof of Proposition 1 can be found in Section 1.1.2.

**Proposition 1.** *We keep notations and assumptions of Theorem 3 (cf. the manuscript) and we assume further that there exist $p \in [1,\infty]$ and $\kappa, R > 0$ such that*

$$\theta^* \in \mathbb{B}_p(0,R) \quad \text{and} \quad \forall \theta \in \mathbb{B}_p(0,R), \quad \lambda_{\min}(\overline{\Gamma}^\theta) \geq \kappa,$$

*where $\mathbb{B}_p(0,R) := \{\theta \in \mathbb{R}^s \mid \|\theta\|_p \leq R\}$. Let us consider any estimator $\theta^\bigstar \in \mathbb{B}_p(0,R)$ of $\theta^*$. Then the probability of the event*

$$\|\theta^* - \theta^\bigstar\|_2 \leq C\,(\kappa c)^{-1} \left\{ \|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2 + \|(\sigma^{\overline{\theta}})^{-2}\|_\infty \left(Nc^2/C\right)^{-1/2} \sqrt{\chi^2_{s,1-\alpha}} \right\},$$

*tends to $1-\alpha$ as $N \to \infty$. We recall that $(\sigma^{\overline{\theta}})^2 = \sigma'(\mathbf{X}_M \overline{\theta}(\theta^*))$.*

**Remarks.** In Proposition 1, note that the constants $c$ and $C$ can be easily computed from the design matrix. Nevertheless, we point out that the confidence region from Proposition 1 involves two constants (namely $\kappa$ and $\sigma^{\overline{\theta}}$) that cannot be *a priori* easily computed in practice.
Proposition 1 proves that when $N$ is large enough, the size of our confidence region is driven by the distance $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. This remark motivates us to choose $\theta^\bigstar$ among the minimizers of the function

$$m : \theta \mapsto \|\overline{\theta}(\theta) - \widehat{\theta}\|_2^2.$$

In the sake of minimizing $m$, a large set of methods are at our disposal. In the next section, we propose a deep learning and a gradient descent approach for our numerical experiments.

### 1.1.2 Proof of Proposition 1

Let us denote $\mathcal{M} : \theta \in \mathbb{R}^s \mapsto \mathbf{X}_M^\top \overline{\pi}^\theta$. Since for any $z \in \{0,1\}^N$, $\mathbb{P}_\theta(z) = \exp(-\mathcal{L}_N(\theta,(z,\mathbf{X}_M)))$, we get $\nabla_\theta \mathbb{P}_\theta(z) = -L_N(\theta,(z,\mathbf{X}_M))\mathbb{P}_\theta(z)$. Recalling that $\overline{\pi}^\theta = \overline{\mathbb{E}}_\theta[Y]$, we have for any $k \in [s]$,

$$\begin{aligned}
\frac{\partial \overline{\pi}^\theta}{\partial \theta_k} &= \left( \sum_{w \in E_M} \mathbb{P}_\theta(w) \right)^{-2} \sum_{w,z \in E_M} \mathbb{P}_\theta(z)\mathbb{P}_\theta(w) z \left\{ L_N(\theta,(w,\mathbf{X}_M)) - L_N(\theta,(z,\mathbf{X}_M)) \right\}_k \\
&= \overline{\mathbb{E}}_\theta \left[ Z \left\{ L_N(\theta,(W,\mathbf{X}_M)) - L_N(\theta,(Z,\mathbf{X}_M)) \right\}_k \right] \\
&= \overline{\mathbb{E}}_\theta \left[ Z \left\{ \mathbf{X}_M^\top (Z-W) \right\}_k \right] \\
&= \overline{\Gamma}^\theta \mathbf{X}_{:,M[k]}, \qquad (1)
\end{aligned}$$

where $Z$ and $W$ are independent random vectors valued in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_\theta$. Note that we used that for any $W \in \{0,1\}^N$, it holds

$$L_N(\theta,(W,\mathbf{X}_M)) = \mathbf{X}_M^\top(\sigma(\mathbf{X}_M\theta) - W).$$

Hence it holds

$$\forall \theta \in \mathbb{R}^s, \quad \nabla\mathcal{M}(\theta) = \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M.$$

Suppose that we are able to compute an estimate $\theta^\bigstar \in \mathbb{B}_p(0,R)$ of $\theta^*$. Using that $\theta^* \in \mathbb{B}_p(0,R)$ and that

$$\inf_{\theta\in\mathbb{B}_p(0,R)} \lambda_{\min}\left(\nabla\mathcal{M}(\theta)\right) \geq \kappa\lambda_{\min}\left(\mathbf{X}_M^\top\mathbf{X}_M\right) \geq c\kappa N,$$

it holds

$$\begin{aligned}
\|\mathcal{M}(\theta^\bigstar) - \mathcal{M}(\theta^*)\|_2^2 &= \| \int_0^1 \nabla\mathcal{M}(t\theta^\bigstar + (1-t)\theta^*)(\theta^\bigstar - \theta^*)dt\|_2^2 \\
&= (\theta^\bigstar - \theta^*)^\top \left\{ \int_0^1 \nabla\mathcal{M}(t\theta^\bigstar + (1-t)\theta^*)dt \right\}^2 (\theta^\bigstar - \theta^*) \\
&\geq \|\theta^\bigstar - \theta^*\|_2^2 \inf_{\theta\in\mathbb{B}_p(0,R)} \lambda_{\min}(\nabla\mathcal{M}(\theta))^2 \\
&\geq (c\kappa N)^2 \|\theta^\bigstar - \theta^*\|_2^2.
\end{aligned}$$

Noticing further that

$$\sup_{\theta\in\mathbb{R}^s} \|\nabla\Psi^{-1}(\theta)\| = \sup_{\theta\in\mathbb{R}^s} \|\mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M\| \leq \frac{1}{4}CN,$$

we get

$$\begin{aligned}
\|\theta^* - \theta^\bigstar\|_2 &\leq (\kappa cN)^{-1} \|\mathbf{X}_M^\top\overline{\pi}^{\theta^\bigstar} - \mathbf{X}_M^\top\overline{\pi}^{\theta^*}\|_2 \\
&= (\kappa cN)^{-1} \|\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^\bigstar)} - \mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)}\|_2 \quad \text{(using Eq.(16) of the manuscript)} \\
&\leq (\kappa cN)^{-1} \sup_{\theta\in\mathbb{R}^s} \|\nabla\Psi^{-1}(\theta)\| \|\Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^\bigstar)}\right) - \Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)}\right)\|_2 \\
&\leq C\left(\kappa c\right)^{-1} \|\Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^\bigstar)}\right) - \Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)}\right)\|_2 \\
&= C\left(\kappa c\right)^{-1} \|\overline{\theta}(\theta^\bigstar) - \overline{\theta}(\theta^*)\|_2 \\
&\leq C\left(\kappa c\right)^{-1} \left[\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2 + \|\widehat{\theta} - \overline{\theta}(\theta^*)\|_2\right],
\end{aligned}$$

where we used that $\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)} = \mathbf{X}_M^\top\sigma\big(\mathbf{X}_M\overline{\theta}(\theta^*)\big) = \Xi\big(\overline{\theta}(\theta^*)\big) \in \mathrm{Im}(\Xi)$ and thus $\Psi(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)})$ is well-defined. Similarly, we have that $\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^\bigstar)} \in \mathrm{Im}(\Xi)$. Since Theorem 3 (see the manuscript) gives that

$$\overline{\mathbb{P}}_{\theta^*}\left(\|V_N(\theta^*)(\widehat{\theta} - \overline{\theta})\|_2^2 \leq \chi^2_{s,1-\alpha}\right) \underset{N\to+\infty}{\to} 1-\alpha,$$

with $V_N(\theta^*) := [\overline{G}_N(\theta^*)]^{-1/2} H_N(\overline{\theta}(\theta^*))$, we deduce (using the assumption of the design matrix from Section 5.1 of the manuscript) that the event

$$\|\widehat{\theta} - \overline{\theta}(\theta^*)\|_2 \leq \|[V_N(\theta^*)]^{-1}\| \|V_N(\theta^*)(\widehat{\theta} - \overline{\theta})\|_2 \leq \|(\sigma^{\overline{\theta}})^{-2}\|_\infty c^{-1} \left(N/C\right)^{-1/2} \sqrt{\chi^2_{s,1-\alpha}},$$

holds with probability tending to $1-\alpha$ as $N \to +\infty$. Note that we used that

$$\|H_N(\overline{\theta}(\theta^*))^{-1}\| \leq (cN)^{-1} \|(\sigma^{\overline{\theta}})^{-2}\|_\infty,$$

and that

$$\|[\overline{G}_N(\theta^*)]^{1/2}\| \leq (CN)^{1/2}.$$

Hence we obtain an asymptotic confidence region for $\theta^*$ of level $1-\alpha$.

### 1.1.3 Simulations

**Deep learning method** We train a feed forward neural network with ReLu activation function and three hidden layers. With this network, we aim at estimating any $\theta \in \mathbb{R}^s$ by feeding as input $\overline{\theta}(\theta)$. We generate our training dataset by first sampling $n_{train} = 500$ random vectors $\theta_i \sim \mathcal{N}(0, \mathrm{Id}_s)$, $i \in [n_{train}]$. Then, for any $i \in [n_{train}]$ we compute the estimate $\widetilde{\theta}(\theta_i)$ of $\overline{\theta}(\theta_i)$ as follows

$$\widetilde{\pi}^{\theta_i} = \frac{\sum_{t=1}^{T} \mathbb{P}_{\theta_i}(Y^{(t)}) Y^{(t)}}{\sum_{t=1}^{T} \mathbb{P}_{\theta_i}(Y^{(t)})} \quad \text{and} \quad \widetilde{\theta}(\theta_i) = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_i}),$$

where $(Y^{(t)})_{t \geq 1}$ is the sequence generated from the SEI-SLR algorithm (see Algorithm 3). We train our network using stochastic gradient descent with learning rate 0.01 and 500 epochs. At each epoch, we feed to the network the inputs $(\widetilde{\theta}(\theta_i))_{i \in [n_{train}]}$ with the corresponding target values $(\theta_i)_{i \in [n_{train}]}$. We then compute our estimate $\theta^\bigstar$ of $\theta^*$ by taking the output of our network when taking as input the unpenalized MLE $\widehat{\theta}$ using the design $\mathbf{X}_M$ (cf. Eq.(12) of the manuscript). Figure 1 illustrates the result obtained from this deep learning approach. We keep the settings of experiment section of the manuscript namely, we consider $\vartheta^* = (1\ 1\ 0\ \ldots\ 0)^\top \in \mathbb{R}^d$ and we choose the regularization parameter $\lambda$ so that the selected model corresponds to the true set of active variables, namely $M = \{1, 2\}$.

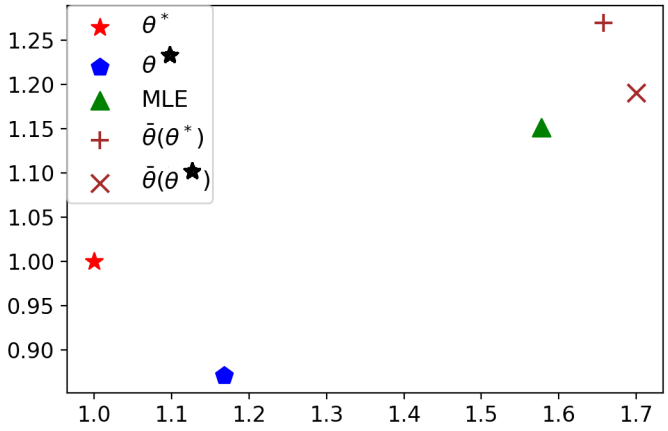


Figure 1: Visualization of the results obtained using our deep learning approach to compute an estimate $\theta^\bigstar$ (the blue hexagone) of $\theta^*$ (the red star). $\theta^\bigstar$ corresponds to the output of the neural network when feeding as input the MLE $\widehat{\theta}$ (the green triangle). We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

**Gradient descent method** As shown in the proof of the expression of Proposition 1 (cf. Eq.(1)), it holds

$$\forall \theta \in \mathbb{R}^s, \quad \nabla_\theta \overline{\pi}^\theta = \overline{\Gamma}^\theta \mathbf{X}_M.$$

Recalling additionally that $\overline{\theta}(\theta) = \Psi\left(\mathbf{X}_M^\top \overline{\pi}^\theta\right)$ (cf. Eq.(16) of the manuscript), we get that for any $\theta \in \mathbb{R}^s$,

$$\begin{aligned}
\nabla_\theta m(\theta) &= 2\nabla_\theta \overline{\theta}(\theta)(\overline{\theta}(\theta) - \widehat{\theta}) \\
&= 2\nabla\Psi(\mathbf{X}_M^\top \overline{\pi}^\theta)\mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}) \\
&= 2\nabla\Psi(\mathbf{X}_M^\top \pi^{\overline{\theta}(\theta)})\mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}) \\
&= 2\left(\mathbf{X}_M^\top \mathrm{Diag}(\pi^{\overline{\theta}(\theta)} \odot (1 - \pi^{\overline{\theta}(\theta)}))\mathbf{X}_M\right)^{-1} \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}).
\end{aligned}$$

Hence,

$$\nabla_\theta m(\theta) = 2\left[H_N(\overline{\theta}(\theta))\right]^{-1} \mathbf{X}_M^\top \overline{\Gamma}^\theta \mathbf{X}_M(\overline{\theta}(\theta) - \widehat{\theta}).$$

Given some $\theta$, $\overline{\pi}^\theta$ and $\overline{\Gamma}^\theta$ can be estimated using samples generated by the SEI-SLR algorithm (and thus the same holds for $\overline{\theta}(\theta) = \Psi(\mathbf{X}_M^\top \overline{\pi}^\theta)$ and for $H_N(\overline{\theta}(\theta))$).

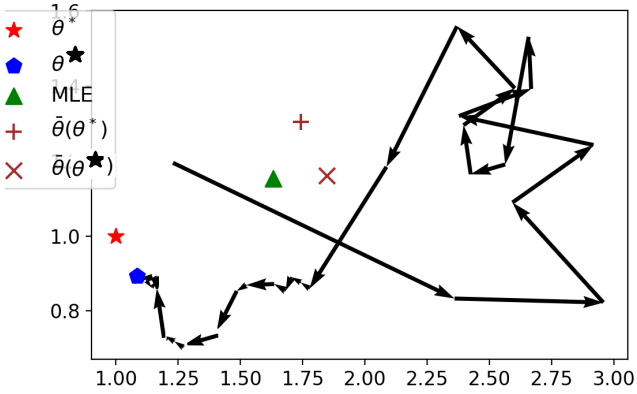


Figure 2: Visualization of our gradient descent procedure to compute an estimate $\theta^\bigstar$ (the blue hexagone) of $\theta^*$ (the red star). The MLE $\widehat{\theta}$ is the green triangle. We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

## 1.2 Asymptotic confidence region in the saturated model

With Theorem 2 of the manuscript, we proved that $\mathbf{X}_M^\top Y$ with $Y$ distributed according to $\overline{\mathbb{P}}_{\pi^*}$ satisfies a CLT with an asymptotic Gaussian distribution centered at $\mathbf{X}_M^\top \overline{\pi}^{\pi^*}$. Using an approach analogous to Section 1.1, we propose here to build an asymptotic confidence region for $\pi^*$.

**Proposition 2.** *We keep notations and assumptions of Theorem 2 of the manuscript and we consider $\alpha \in (0,1)$. We assume further that there exist $p \in [1,\infty]$ and $\kappa, R > 0$ such that*

$$\pi^* \in \mathbb{B}_p\Big(\frac{\mathbf{1}_N}{2}, R\Big) \quad \text{and} \quad \forall \pi \in \mathbb{B}_p\Big(\frac{\mathbf{1}_N}{2}, R\Big), \quad \lambda_{\min}(\overline{\Gamma}^\pi) \geq \kappa.$$

*Let us consider any estimator $\pi^\bigstar \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)$ of $\pi^*$. Then the probability of the event*

$$\|\pi^* - \pi^\bigstar\|_2 \leq (4\kappa)^{-1}\Big\{\|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\bigstar})\|_2 + Cc^{-1}\sqrt{\chi^2_{s,1-\alpha}} + \|\mathrm{Proj}^\perp_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\bigstar})\|_2\Big\},$$

*tends to $1-\alpha$ as $N \to \infty$.*

**Proof of Proposition 2.** Let us denote $\mathcal{R} : \pi \in (0,1)^N \mapsto \overline{\pi}^{\pi}$. It holds for any $i \in [N]$,

$$\begin{aligned}\frac{\partial \overline{\pi}^{\pi}}{\partial \pi_i} &= \left(\sum_{w \in E_M} \mathbb{P}_\pi(w)\right)^{-2} \sum_{w,z \in E_M} \mathbb{P}_\pi(z)\mathbb{P}_\pi(w) z \left\{z - w\right\}_i \left(\pi_i(1-\pi_i)\right)^{-1} \\ &= \overline{\mathbb{E}}_\pi \left[Z(Z-W)_i^\top\right] \left(\pi_i(1-\pi_i)\right)^{-1},\end{aligned}$$

where $Z$ and $W$ are independent random vectors valued in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_\pi$. Hence it holds

$$\forall \pi \in (0,1)^N, \quad \nabla \mathcal{R}(\pi) = \overline{\Gamma}^{\pi} \mathrm{Diag}(\pi \odot (1-\pi))^{-1}.$$

Suppose that we are able to compute an estimate $\pi^{\bigstar} \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)$ of $\pi^*$. Then since it holds for any $v \in \mathbb{R}^N$,

$$\inf_{\pi \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R)} \|\nabla \mathcal{R}(\pi) v\|_2 \geq 4\kappa \|v\|_2,$$

we get that

$$\begin{aligned}\|\mathcal{R}(\pi^{\bigstar}) - \mathcal{R}(\pi^*)\|_2 &= \| \int_0^1 \nabla \mathcal{R}(t\pi^{\bigstar} + (1-t)\pi^*)(\pi^{\bigstar} - \pi^*) dt\|_2 \\ &\geq 4\kappa \|\pi^{\bigstar} - \pi^*\|_2.\end{aligned}$$

Hence we have that

$$\begin{aligned}\|\pi^* - \pi^{\bigstar}\|_2 &\leq (4\kappa)^{-1} \|\overline{\pi}^{\pi^{\bigstar}} - \overline{\pi}^{\pi^*}\|_2 \\ &\leq (4\kappa)^{-1} \big\{ \|\mathrm{Proj}_{\mathbf{X}_M}(\overline{\pi}^{\pi^{\bigstar}} - Y)\|_2 + \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^*})\|_2 \\ &\quad + \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^{\bigstar}} - \overline{\pi}^{\pi^*})\|_2 \big\}.\end{aligned}$$

Since Theorem 2 of the manuscript gives that

$$\overline{\mathbb{P}}_{\pi^*} \left( \|[\overline{G}_N(\pi^*)]^{-1/2} (\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*})\|_2^2 \leq \chi^2_{s,1-\alpha} \right) \underset{N \to +\infty}{\to} 1 - \alpha,$$

we deduce that the event

$$\begin{aligned}\|\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*}\|_2 &\leq \|[\overline{G}_N(\pi^*)]^{1/2}\| \|[\overline{G}_N(\pi^*)]^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*})\|_2 \\ &\leq (CN)^{1/2} \sqrt{\chi^2_{s,1-\alpha}},\end{aligned}$$

holds with probability tending to $1-\alpha$ as $N \to +\infty$. Noticing further that for any vector $v \in \mathbb{R}^N$,

$$\|\mathrm{Proj}_{\mathbf{X}_M} v\|_2 \leq \|\mathbf{X}_M \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1}\| \times \|\mathbf{X}_M^\top v\|_2 \leq (CN)^{1/2} (cN)^{-1} \|\mathbf{X}_M^\top v\|_2,$$

we get that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$, it holds with at least $1 - \alpha - \epsilon$,

$$\begin{aligned}\|\pi^* - \pi^{\bigstar}\|_2 \leq (4\kappa)^{-1} \big\{ &\|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^{\bigstar}})\|_2 + Cc^{-1} \sqrt{\chi^2_{s,1-\alpha}} \\ &+ \|\mathrm{Proj}^{\perp}_{\mathbf{X}_M}(\overline{\pi}^{\pi^{\bigstar}} - \overline{\pi}^{\pi^*})\|_2 \big\}.\end{aligned}$$

Hence we obtain an asymptotic confidence region for $\pi^*$ of level $1-\alpha$.

**Remarks.**

- Analogously to Section 1.1, Proposition 2 motivates us to choose $\pi^\star$ among the minimizers of the function
$$M : \pi \mapsto \|\mathbf{X}_M^\top \overline{\pi}^\pi - \mathbf{X}_M^\top Y\|_2^2.$$
As mentioned in the Section 1.1, one can rely for example on a deep learning or a gradient descent method in order to reach a local minimum $\pi^\star$ for $M$.

- The term $\|\mathrm{Proj}^\perp_{\mathbf{X}_M}(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2$ arising in the confidence region from Proposition 2 illustrates that our conditional CLT from Theorem 2 of the manuscript holds on $\mathbf{X}_M^\top Y$ and that we do not control what occurs in the orthogonal complement of the span of the columns of $\mathbf{X}_M$. Nevertheless, let us comment informally our result in the case where $E_M = \{0,1\}^N$ (meaning that there is no conditioning) and where $\vartheta^*$ is close to 0 (meaning that $\pi^*$ is close to $\mathbf{1}_N/2$). In this framework, $\overline{\Gamma}^\pi = \mathrm{Diag}(\pi \odot (1-\pi))$ is close to $\frac{1}{4}\mathrm{Id}_N$ for $\pi$ in a small neighbourhood around $\mathbf{1}_N/2$. Hence, we get that $\kappa$ is approximately $\frac{1}{4}$. Since it also holds that $\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star} = \pi^* - \pi^\star$ (since $E_M = \{0,1\}^N$), we obtain from Proposition 2 that a CR for $\mathrm{Proj}_{\mathbf{X}_M}\pi^*$ with asymptotic coverage $1-\alpha$ is
$$\|\mathrm{Proj}_{\mathbf{X}_M}(\pi^* - \pi^\star)\|_2 \leq \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\star})\|_2 + Cc^{-1}\sqrt{\chi^2_{s,1-\alpha}}.$$

# 2 Side notes about SIGLE

## 2.1 SIGLE for a misspecified model from the start

In this paper, we have considered the case where the observed data $y_i \in \mathcal{Y}$ has indeed by generated from the GLM presented in Section 1.1 of the manuscript. Can we extend the methods presented in this paper when we remove this assumption?

In this section, we consider that the $y_i$'s are i.i.d. and distributed according to an arbitrary probability distribution $\mathbb{P}$.

### 2.1.1 SIGLE in the selected model

In the case of a misspecified model from the start, the assumption made to be in the selected model is
$$\sigma^{-1}(\mathbb{E}[Y]) \in \mathrm{Im}(\mathbf{X}_M),$$
where the expectation is taken with respect to $\mathbb{P}$. We define
$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^s} \overline{\mathbb{E}}\big[-\log \overline{\mathbb{P}}_\theta(Y)\big]. \tag{2}$$
$\mathbb{P}_{\theta^*}$ can be understood as the probability distribution belonging to the GLM family with design matrix $\mathbf{X}_M$ leading to the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ that is the closest possible to $\overline{\mathbb{P}}$. More precisely, for any GLM distribution $\mathbb{P}_\theta$, $\theta \in \mathbb{R}^s$, we have
$$\mathrm{KL}(\overline{\mathbb{P}} \mid \overline{\mathbb{P}}_\theta) \geq \mathrm{KL}(\overline{\mathbb{P}} \mid \overline{\mathbb{P}}_{\theta^*}).$$

In the following, we reinterpret the methods of this paper relaxing the assumption that the model is well-specified from the start, as it might happen that the true initial distribution of the observation $\mathbb{P}$ does not belong to the GLM family. More precisely, considering the null hypothesis:
$$\mathbb{H}_0 : \quad \{\overline{\mathbb{P}} \equiv \overline{\mathbb{P}}_{\theta_0^*}\},$$
we can fall into one of the following cases:

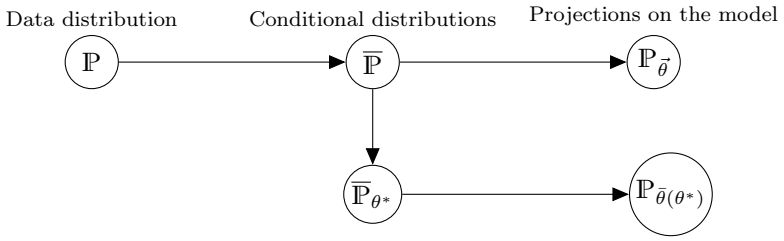


Figure 3: Visualizations of all distributions that we consider if the model is a priori not necessarily well-specified from the start.

1. If the model was well-specified initially, this means that there exists some $\vartheta^* \in \mathbb{R}^d$ such that $\mathbb{P} = \mathbb{P}_{\vartheta^*}$ and thus $\overline{\mathbb{P}} \equiv \overline{\mathbb{P}}_{\vartheta^*_M}$ (in the selected model). Namely, the null hypothesis is true for at least one parameter vector $\theta_0^* \in \mathbb{R}^s$.
2. If the model was not well-specified initially but the null is true for some $\theta_0^* \in \mathbb{R}^s$, this means that by conditioning on the selection event, we lost the information regarding the fact that the model was misspecified initially.
3. If the model was not well-specified initially and the null is false for any $\theta_0^*$, this means that $\overline{\mathbb{P}}$ still carries the information of the initial model misspecificity.

**A predictive viewpoint on SIGLE in the selective model.** To obtain the SIGLE statistic in the selected model, we need to compute $\overline{\theta}(\theta_0^*)$ which is defined by

$$\overline{\theta}(\theta_0^*) \in \arg\min_{\theta \in \mathbb{R}^s} \overline{\mathbb{E}}_{\theta_0^*}\big[-\log \mathbb{P}_\theta(Y)\big] = \arg\min_{\theta \in \mathbb{R}^s} \mathrm{KL}(\overline{\mathbb{P}}_{\theta_0^*} \mid \mathbb{P}_\theta).$$

The question that we ask is how far is the distribution $\mathbb{P}_{\overline{\theta}(\theta_0^*)}$ from $\overline{\mathbb{P}}$. This can be of interest for a prediction task where one might want to use $\overline{\theta}(\theta_0^*)$ to predict the response to new entries.

The best approximation of $\overline{\mathbb{P}}$ that we can get considering an unconditional GLM distribution of the form $\mathbb{P}_\theta$ is $\mathbb{P}_{\vec{\theta}}$ where

$$\vec{\theta} \in \arg\min_{\theta \in \mathbb{R}^s} \overline{\mathbb{E}}\big[-\log \mathbb{P}_\theta(Y)\big] = \arg\min_{\theta \in \mathbb{R}^s} \mathrm{KL}(\overline{\mathbb{P}} \mid \mathbb{P}_\theta).$$

Therefore, we want to compare the difference between the KL divergence between $\overline{\mathbb{P}}$ and $\mathbb{P}_{\vec{\theta}}$, and the KL divergence between $\overline{\mathbb{P}}$ and $\mathbb{P}_{\overline{\theta}(\theta_0^*)}$. It holds

$$\mathrm{KL}(\overline{\mathbb{P}} \mid \mathbb{P}_{\overline{\theta}(\theta_0^*)}) = \mathrm{KL}(\overline{\mathbb{P}} \mid \mathbb{P}_{\vec{\theta}}) + \overline{\mathbb{E}}\big[\log \frac{\mathbb{P}_{\vec{\theta}}}{\mathbb{P}_{\overline{\theta}(\theta_0^*)}}\big],$$

where $\overline{\mathbb{E}}\big[\log \frac{\mathbb{P}_{\vec{\theta}}}{\mathbb{P}_{\overline{\theta}(\theta_0^*)}}\big] \geq 0$ by definition of $\vec{\theta}$. Therefore, $\overline{\mathbb{E}}\big[\log \frac{\mathbb{P}_{\vec{\theta}}}{\mathbb{P}_{\overline{\theta}(\theta_0^*)}}\big]$ corresponds to the additional error we make in terms of KL divergence by working with the proxy $\overline{\mathbb{P}}_{\theta_0^*}$ instead of the true conditional distribution of the observations $\overline{\mathbb{P}}$.

## 2.2 Inverting the first order optimality condition

When characterizing the selection event $E_M^{S_M}$ (see Theorem 1 of the manuscript), we highlighted the crucial role of the diffeomorphism $\Xi = \Psi^{-1}$ arising in the first order optimality condition. In this section, we aim at presenting

- a different view on $\Psi$ using tools from convex analysis,

- the practical methods we use to compute quantities involving $\Psi$ in our simple hypothesis testing method in the selected model.

### 2.2.1 SIGLE through the lens of convex analysis

Recalling the definition of the negative log-likelihood $\mathcal{L}_N(\theta,(Y,\mathbf{X}_M))$, we will denote in this section

$$\mathcal{L}_{N,0,M}(\theta) := \mathcal{L}_N(\theta,(0,\mathbf{X}_M)) = \sum_{i=1}^N \xi(\langle \mathbf{X}_{i,M},\theta\rangle).$$

Let us recall that the Fenchel conjugate of the map $\mathcal{L}_{N,0,M}$ is defined by

$$\mathcal{L}^*_{N,0,M} : \rho \in \mathbb{R}^s \mapsto \sup_{\theta\in\mathbb{R}^s} \left\{\langle \rho,\theta\rangle - \mathcal{L}_{N,0,M}(\theta)\right\}.$$

Since $\xi$ is a convex and $C^{m+1}$ function, $\mathcal{L}_{N,0,M}$ is also a convex and a $C^{m+1}$ map which implies that $L_{N,0,M} = \nabla\mathcal{L}_{N,0,M}$ is a homeomorphism. We deduce that for any $\rho\in\mathbb{R}^s$,

$$\begin{aligned}\mathcal{L}^*_{N,0,M}(\rho) &= \langle \rho, L^{-1}_{N,0,M}(\rho)\rangle - \mathcal{L}_{N,0,M}(L^{-1}_{N,0,M}(\rho)),\\ \nabla\mathcal{L}^*_{N,0,M}(\rho) &= L^{-1}_{N,0,M}(\rho).\end{aligned}$$

For any $Y\in\{0,1\}^N$, the unpenalized MLE $\widehat{\theta}$ with the design matrix $\mathbf{X}_M$ and the observed response $Y$ is given by (using the first order optimality condition)

$$\widehat{\theta} = L^{-1}_{N,0,M}(\mathbf{X}_M^\top Y).$$

We deduce that

$$\widehat{\theta} = \nabla\mathcal{L}^*_{N,0,M}(\mathbf{X}_M^\top Y).$$

Similarly, using Eq.(16) of the manuscript we get

$$\overline{\theta}(\theta^*) = \nabla\mathcal{L}^*_{N,0,M}(\mathbf{X}_M^\top \overline{\pi}^{\theta^*}).$$

We deduce that $\Psi = \nabla\mathcal{L}^*_{N,0,M}$.

In order to provide a concrete interpretation of the function $\Psi$, let us first characterize the Fenchel conjugate $\mathcal{L}^*_{N,0,M}$:

$$\begin{aligned}\mathcal{L}^*_{N,0,M}(\rho) &= \sup_{\theta\in\mathbb{R}^s}\{\langle\rho,\theta\rangle - \mathcal{L}_{N,0,M}(\theta)\}\\ &= \sup_{\theta\in\mathbb{R}^s}\sum_{i=1}^N\{\rho_i\theta_i - \xi(\mathbf{X}_{i,M}\theta)\}\\ &= (\sum_{i=1}^N f_i)^*(\rho)\\ &= (f_1^*\square\cdots\square f_N^*)(\rho)\\ &:= \min_{\rho=\rho^{(1)}+\cdots+\rho^{(N)}}\left\{f_1^*(\rho^{(1)})+\cdots+f_N^*(\rho^{(N)})\right\},\end{aligned} \tag{3}$$

where in the last equality we used [Laurent, 1972, Theorem 6.5.8] and where for any $i\in[N]$,

$$f_i : \theta\in\mathbb{R}^s \mapsto \xi(\mathbf{X}_{i,M}\theta).$$

Using Lemma 1, we obtain that $\mathcal{L}^*_{N,0,M}(\rho)$ is the minimal entropy obtained among the vectors of probabilities $\pi\in(0,1)^N$ satisfying $\rho = \mathbf{X}_M^\top\pi$.

**Lemma 1.** *The inf-convolution in Eq.(3) is attained for* $(\rho^{(i)})_{i\in[N]} \in (\mathbb{R}^s)^N$ *such that for any* $i \in [N]$, $\rho^{(i)} = \pi_i \mathbf{X}_{i,M}$ *for some* $\pi_i \in (0,1)$. *Moreover,*

$$\mathcal{L}^*_{N,0,M}(\rho) = \min_{\pi\in(0,1)^N \ s.t. \ \rho=\mathbf{X}_M^\top \pi} H(\pi), \tag{4}$$

*where*

$$H(\pi) = \sum_{i=1}^N \{\pi_i \ln(\pi_i) + (1-\pi_i)\ln(1-\pi_i)\}.$$

*Proof of Lemma 1.*

- **The inf-convolution in Eq.(3) is attained.**
  First, we know from [Laurent, 1972, Theorem 6.5.8] that the minimum in the inf-convolution of Eq.(3) is attained.

- **$\rho^{(i)}$ in Eq.(3) can be chosen in the span of $\mathbf{X}_{i,M}$.**
  Let us consider $i \in [N]$ and some $\rho^{(i)} \in \mathbb{R}^s$. Let us assume by contradiction that $\rho^{(i)} \notin \mathrm{Span}(\mathbf{X}_{i,M})$. Then considering

  $$\theta^{(i)}(t) := t\big(\mathrm{Id} - \frac{1}{\|\mathbf{X}_{i,M}\|_2^2}\mathbf{X}_{i,M}^\top \mathbf{X}_{i,M}\big)\rho^{(i)},$$

  we have

  $$\lim_{t\to+\infty}\Big\{\langle\rho^{(i)}, \theta^{(i)}(t)\rangle - \xi(\mathbf{X}_{i,M}\theta^{(i)}(t))\Big\} = \lim_{t\to+\infty}\Big\{t[\rho^{(i)}]^\top \mathrm{Proj}^\perp_{\mathbf{X}_{i,M}}\rho^{(i)} - 0\Big\} = +\infty,$$

  which means that $f_i^*(\rho^{(i)}) = +\infty$ since for any $t > 0$ it holds

  $$\begin{aligned} f_i^*(\rho^{(i)}) &= \sup_{\theta\in\mathbb{R}^s}\Big\{\langle\rho^{(i)},\theta\rangle - \xi(\mathbf{X}_{i,M}\theta)\Big\} \\ &\geq \Big\{\langle\rho^{(i)}, \theta^{(i)}(t)\rangle - \xi(\mathbf{X}_{i,M}\theta^{(i)}(t))\Big\}. \end{aligned}$$

  We deduce that in the inf-convolution of Eq.(3), we can consider that for any $i \in [N]$, $\rho^{(i)} = \pi_i \mathbf{X}_{i,M}$ for some $\pi_i \in \mathbb{R}$.

- **$\rho^{(i)}$ in Eq.(3) can be chosen as $\pi_i\mathbf{X}_{i,M}$ with $\pi_i \in (0,1)$.**
  We have already proved that $\rho^{(i)}$ in Eq.(3) can be chosen as $\rho^{(i)} = \pi_i\mathbf{X}_{i,M}$. It holds

  $$\begin{aligned} f_i^*(\pi_i\mathbf{X}_{i,M}) &= \sup_{\theta\in\mathbb{R}^s}\{\langle\pi_i\mathbf{X}_{i,M},\theta\rangle - \xi(\mathbf{X}_{i,M}\theta)\} \\ &= \sup_{\theta\in\mathbb{R}^s}\{\langle\pi_i, \mathbf{X}_{i,M}\theta\rangle - \xi(\mathbf{X}_{i,M}\theta)\} \\ &= \sup_{r\in\mathbb{R}}\{\pi_i r - \xi(r)\} \\ &= \xi^*(\pi_i) \\ &= H(\pi_i), \end{aligned}$$

  where we used that the Fenchel conjugate of the softmax function is the entropy $H$ defined by

  $$H(p) = \begin{cases} p\ln(p) + (1-p)\ln(1-p) & \text{if } p \in (0,1), \\ +\infty & \text{otherwise.} \end{cases}$$

Since in Eq.(3) we aim a reaching a minimum, we deduce from these computations that one can restrict $\rho^{(i)}$ to be of the form $\pi_i\mathbf{X}_{i,M}$ with $\pi_i \in (0,1)$. □

**Interpretation of $\Psi(\rho)$.** Lemma 1 shows that $\mathcal{L}^*_{N,0,M}(\rho)$ is the minimum entropy of a population characterized by $N$ binary features with the constraint that we have some information on the population given by $\rho \in \mathbb{R}^s$. We assume that $\rho$ depends linearly on the proportion of the population with the different features, namely

$$\rho = \mathbf{X}_M^\top \pi,$$

where for all $i \in [N]$, $\pi_i$ represents the proportion of people with feature $i$. Hence, given the observation of $s$ aggregated properties about the population (namely $\rho$), $\mathcal{L}^*_{N,0,M}(\rho)$ is the entropy corresponding to the most uniform allocation of the $N$ binary features in the population. Hence, $\Psi(\rho) = \nabla \mathcal{L}^*_{N,0,M}(\rho)$ quantifies how much the entropy of this ideal description of the population is changed when a small shift in the observation of the $s$ properties occurs.

Taking a concrete example, one can consider that the $N$ features are the following: age between 20 and 40, age between 40 and 60, age above 60, manager, manual labourer, lives in a big city, lives in a small town, ... The vector $\rho$ represents the number of votes obtained by $s$ different candidates during an election. We assume that the number of votes obtained by each candidate is a linear function of the proportion of the population with the different features. We observe only the number of votes obtained by each candidate. Then $\mathcal{L}^*_{N,0,M}(\rho)$ represents the entropy of the population assuming that the different features are distributed as uniformly as possible in the population. $\Psi(\rho)$ measures the variation of the entropy of the population when a small change in the number of votes obtained by the different candidates is observed.

### 2.2.2 Practical implementation of SIGLE in the selected model

The PSI method in the selected model for the $\ell^1$-penalized logistic regression proposed in this paper requires the ability to compute efficiently

- $\Psi(\mathbf{X}_M^\top Y)$ for any $Y \in \{0,1\}^N$,
- $\Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*})$.

As already mentioned in Eq.(16) of the manuscript, for any $Y \in \{0,1\}^N$, $\Psi(\mathbf{X}_M^\top Y)$ corresponds to the unpenalized MLE $\widehat{\theta}$ computed using the design $\mathbf{X}_M$ (see Eq.(12) of the manuscript). As a result, we compute $\Psi(\mathbf{X}_M^\top Y)$ by simply solving the unpenalized MLE for logistic regression using standard open source libraries (such as scikit-learn in Python where we remove the $\ell^2$-regularization which is applied by default).

Solvers computing the MLE for logistic regression require - as far as we know - the response vector to have binary entries. As a consequence, a different approach is required to compute $\Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*})$ since $\overline{\pi}^{\theta_0^*} \in (0,1)^N$. We found our method to be extremely accurate in our numerical experiments and it works as follows. First, we compute

$$\theta^c \in \arg\min_{\theta \in \mathbb{R}^s} \|\mathbf{X}_M \theta - \sigma^{-1}(\overline{\pi}^{\theta_0^*})\|_2^2,$$

and we end up with two possible cases:

1. Either it holds
$$\mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta^c) = \mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}, \tag{5}$$
which is equivalent to $\overline{\theta}(\theta^*) = \theta^c$ (see Eq.(16) of the manuscript). In this case, we output $\theta^c$. Note that this situation occurs in particular when
$$\sigma^{-1}(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}) \in \mathrm{Im}(X_M),$$
which can be understood as a conditional selected model-type assumption.

2. Or Eq.(5) does not hold and we consider a gradient descent approach using as warm start the vector $\theta^c$ to minimize the map
$$G : \theta \mapsto \|\mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta) - \mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}\|_2^2.$$

Note that the gradient of $G$ at $\theta \in \mathbb{R}^s$ is given by

$$\nabla G(\theta) = 2\mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M\mathbf{X}_M^\top \left(\sigma(\mathbf{X}_M\theta) - \overline{\pi}^{\theta_0^*}\right),$$

and satisfies

$$\forall \theta \in \mathbb{R}^s, \quad \|G(\theta)\|_2 \leq \frac{1}{4}\|\mathbf{X}_M^\top \mathbf{X}_M\| \times \|\mathbf{X}_M\|_{1,2} =: L_G,$$

where $\|\mathbf{X}_M\|_{1,2} := \sqrt{\sum_{i=1}^N \|\mathbf{X}_{i,M}\|_1^2}$.

Our method is summarized with Algorithm 1.

**Algorithm 1** Computing $\overline{\theta}(\theta_0^*)$

**Input:** $t_{\max}, \epsilon, \ell_r$
$\theta^c \in \arg\min_{\theta \in \mathbb{R}^s} \|\mathbf{X}_M\theta - \sigma^{-1}(\overline{\pi}^{\theta_0^*})\|_2^2$
**if** $G(\theta^c) < \epsilon$ **then**
  **return** $\theta^c$
**else**
  $\theta^{(0)} \leftarrow \theta^c$
  $t \leftarrow 0$
  **while** $t < t_{\max}$ and $G(\theta^{(t)}) > \epsilon$ **do**
    $t \leftarrow t + 1$
    $\theta^{(t)} \leftarrow \theta^{(t-1)} - \frac{\ell_r}{L_G}\nabla G(\theta^{(t-1)})$
  **end while**
  **return** $\theta^{(t)}$
**end if**

# 3 Inference conditional on the signs

## 3.1 Leftover Fisher information

As highlighted in Fithian et al. [2014], conducting inference conditional on some random variable prevents the use of this variable as evidence against a hypothesis. Selective inference should be understood as partitioning the observed information in two sets: the one used to select the model and the one used to make inference. This communicating vessels principle is illustrated with the following inclusions borrowed from Fithian et al. [2014].

$$\mathcal{F}_0 \underbrace{\subset}_{\text{used for selection}} \mathcal{F}(\mathbb{1}_{Y\in\mathcal{M}}) \underbrace{\subset}_{\text{used for inference}} \mathcal{F}(Y).$$

Typically, let us assume that we condition on both the selected support $\widehat{M}(Y) = M$ and the observed vector of signs $\widehat{S}_M(Y) = S_M \in \{0,1\}^{|M|}$, meaning that $\mathcal{M} = E_M^{S_M}$ (cf. Eq.(5) of the manuscript). Even if the vector of signs $S_M$ is surprising under $\mathbb{H}_0$, we will not reject unless we are surprised anew by observing the response variable $Y$. Stated otherwise, when we condition on both the selected support and the vector of signs, we cannot take advantage of the possible unbalanced probability distribution of the vector of signs $\widehat{S}_M(Y)$ conditionally on $E_M$. Hence, conditioning on a finer $\sigma$-algebra results in some information loss. Fithian et al. [2014] explain that we can actually quantify this waste of information. The Hessian of the log-likelihood can be decomposed as

$$\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, E_M) = \nabla^2_\vartheta \mathcal{L}_N(\vartheta, \widehat{S}_M(Y) \,|\, E_M) + \nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \{E_M, \widehat{S}_M(Y)\}). \tag{6}$$

For any $\sigma$-algebra $\mathcal{F} \subseteq \sigma(Y)$, we consider the conditional expectation

$$\mathcal{I}_{Y\,|\,\mathcal{F}}(\vartheta) := -\mathbb{E}\left[\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \mathcal{F}) \,|\, \mathcal{F}\right].$$

The *leftover Fisher information* after selection at $\widehat{S}_M(Y)$ is defined by $\mathcal{I}_{Y \mid \{E_M, \widehat{S}_M(Y)\}}(\vartheta)$. Taking expectation in both sides of Eq.(6) leads to

$$\begin{aligned}\mathbb{E}\left[\mathcal{I}_{Y \mid \{E_M, \widehat{S}_M(Y)\}}(\vartheta)\right] &= \mathbb{E}\,\mathcal{I}_{Y \mid E_M}(\vartheta) - \mathbb{E}\,\mathcal{I}_{\widehat{S}_M(Y) \mid E_M}(\vartheta) \\ &\preceq \mathbb{E}\,\mathcal{I}_{Y \mid E_M}(\vartheta),\end{aligned}$$

which can also be written as

$$\sum_{S_M \in \{\pm 1\}^s} \mathbb{P}(\widehat{S}_M(Y) = S_M \mid E_M)\, \mathbb{E}\mathcal{I}_{Y \mid E_M^{S_M}}(\vartheta) \preceq \mathbb{E}\,\mathcal{I}_{Y \mid E_M}(\vartheta).$$

In expectation, the loss of information induced by conditioning further on the vector of signs is quantified by the information $\widehat{S}_M(Y)$ carries about $\vartheta$. Let us stress that this conclusion is only true in expectation and it may exist some vector of signs $S_M \in \{-1, +1\}^s$ such that

$$\mathcal{I}_{Y \mid E_M}(\vartheta) \preceq \mathcal{I}_{Y \mid E_M^{S_M}}(\vartheta).$$

Hence, conditioning on the signs will generally lead to wider confidence intervals. Nevertheless, let us stress that inference procedures correctly calibrated conditional on $E_M^{S_M}$ will be also valid conditional on $E_M$. More precisely, considering some transformation $T : \mathbb{R}^N \to \mathbb{R}$ and real valued random variables $L(Y, S_M) < U(Y, S_M)$ such that for any vector of signs $S_M \in \{-1, +1\}^s$ it holds

$$\mathbb{P}\left(T(\pi^*) \in [L(Y, S_M), U(Y, S_M)] \,|\, E_M^{S_M}\right) = 1 - \alpha,$$

the confidence interval has also $(1-\alpha)$ coverage conditional on the $E_M = \{\widehat{M}(Y) = M\}$ since

$$\begin{aligned}&\mathbb{P}(T(\pi^*) \in [L(Y, \widehat{S}_M(Y)), U(Y, \widehat{S}_M(Y))] \mid E_M) \\ &= \sum_{S_M \in \{\pm 1\}^s} \mathbb{P}(\widehat{S}_M(Y) = S_M \mid E_M)\, \underbrace{\mathbb{P}(T(\pi^*) \in [L(Y, S_M), U(Y, S_M)] \mid E_M^{S_M})}_{=1-\alpha} \\ &= 1 - \alpha.\end{aligned}$$

## 3.2 Discussion

Let us recall that in Taylor and Tibshirani [2018], the authors work in the selected model for logistic regression. They consider a selected model $M \subseteq [d]$ associated to a response vector $Y = (y_i)_{i \in [n]} \in \{0, 1\}^N$ where for any $i \in [N]$, $y_i$ is a Bernoulli random variable with parameter $\{\sigma(\mathbf{X}_M \theta^*)\}_i$ for some $\theta^* \in \mathbb{R}^s$ $(s = |M|)$. As presented in the first section of the Appendix of our manuscript, in Taylor and Tibshirani [2018] the authors claim the following asymptotic distribution

$$\underline{\theta} \sim \mathcal{N}(\vartheta_M^*, H_N(\vartheta_M^*)^{-1}), \tag{7}$$

where $\underline{\theta} = \hat{\vartheta}_M^\lambda + \lambda H_N(\widehat{\vartheta}_M^\lambda)^{-1} \widehat{S}_M(Y)$. Note that this approximation corresponds to the one usually made to form Wald tests and confidence intervals in generalized linear models. They claim that the selection event $\{Y \in \{0, 1\}^N \,:\, \widehat{M}(Y) = M, \widehat{S}_M(Y) = S_M\}$ can be asymptotically approximated by

$$\{Y \,:\, \mathrm{Diag}(S_M)\left(\underline{\theta} - H_N(\vartheta_M^*)^{-1} \lambda S_M\right) \geq 0\}.$$

Let us denote by $F_{\mu,\sigma^2}^{[a,b]}$ the CDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable truncated to the interval $[a, b]$. Then they use the polyhedral lemma to state that for some random variables $\mathcal{V}^-$ and $\mathcal{V}^+$ it holds

$$\left[F^{[\mathcal{V}_{S_M}^-, \mathcal{V}_{S_M}^+]}_{\vartheta^*_{M[j]}, \left[H_N(\vartheta_M^*)^{-1}\right]_{j,j}}(\underline{\theta}_j) \mid \widehat{M}(Y) = M,\ \widehat{S}_M(Y) = S_M\right] \sim \mathcal{U}([0, 1]).$$

Several problems arise at this point.

1. **Lack of theoretical guarantee due to the use of Monte-Carlo estimates.**
The first problem is that both $\underline{\theta}$ and the selection event $\{\widehat{M}(Y) = M,\ \widehat{S}_M(Y) = S_M\}$ involve the unknown parameter $\vartheta^*_M$ through $H_N(\vartheta^*_M)$. Taylor and al. propose to use a Monte-Carlo estimate for $H_N(\vartheta^*_M)$ by replacing it with $H_N(\widehat{\theta}^{\lambda})$. Using this Monte-Carlo estimate, one can compute $L$ and $U$ such that

$$F^{[\mathcal{V}^-_{S_M},\mathcal{V}^+_{S_M}]}_{L,\left[H_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = 1 - \frac{\alpha}{2} \quad \text{and} \quad F^{[\mathcal{V}^-_{S_M},\mathcal{V}^+_{S_M}]}_{U,\left[H_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = \frac{\alpha}{2}.$$

Then, $[L, U]$ is claimed to be a confidence interval with (asymptotic) $(1-\alpha)$ coverage for $\vartheta^*_{M[j]}$ conditional on $\{\widehat{M}(Y) = M, \widehat{S}_M(Y) = S_M\}$, that is,

$$\mathbb{P}(\vartheta^*_{M[j]} \in [L, U] \mid \widehat{M}(Y) = M,\ \widehat{S}_M(Y) = S_M) = 1 - \alpha.$$

2. **Their approach is not well suited to provide more powerful inference procedures by conditioning only on $E_M$.**
In the linear model, Lee et al. [2016] also start by deriving a pivotal quantity by conditioning on both the selected variables and the vector of signs. However, in the context of linear regression, the vector of signs only appears in the threshold values $\mathcal{V}^-$ and $\mathcal{V}^+$. Hence, conditioning only on the selected variables $\{\widehat{M}(Y) = M\}$ simply reduces to take the union $\cup_{S_M \in \{\pm 1\}^s}[\mathcal{V}^-_{S_M}, \mathcal{V}^+_{S_M}]$ for the truncated Gaussian. In the method proposed by Taylor and Tibshirani [2018], the vector of signs also appears in the computation of $\underline{\theta}$. The consequence is that the (asymptotic) distribution of $\underline{\theta}$ conditional on $\{\widehat{M}(Y) = M\}$ is not a truncated Gaussian anymore but a mixture of truncated Gaussians. In this situation, it seems unclear how to take advantage of this structure to provide more powerful inference procedures.

# References


W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

P. Laurent. *Approximation et optimisation*. Collection Enseignement des sciences. Hermann, 1972. URL https://books.google.fr/books?id=h8OmAAAAIAAJ.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

J. Taylor and R. Tibshirani. Post-selection inference for $\ell_1$-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018. doi: https://doi.org/10.1002/cjs.11313. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11313.