A new procedure for Selective Inference with the Generalized Linear Lasso

Quentin Duchemin*
LAMA, Univ Gustave Eiffel, Univ Paris Est Créteil, CNRS
F-77447 Marne-la-Vallée, France.
quentin.duchemin@univ-eiffel.fr
&
Yohann De Castro
Institut Camille Jordan - École Centrale de Lyon
Lyon, France
yohann.de-castro@ec-lyon.fr

Abstract

This articles investigates the distribution of the solutions of the generalized linear lasso (GLL), conditional on some selection event. In this framework of post-selection inference (PSI), we provide rigorous definitions of the selected and saturated models: two different paradigms that determine the hypothesis being tested. Based on a conditional Maximum Likelihood Estimator (MLE) approach, we give a procedure to obtain asymptotically valid PSI confidence regions and testing procedures for Generalized Linear Models (GLMs).

In a second stage, we focus on the sparse logistic regression and we exhibit conditions ensuring that our conditional MLE method is valid. We present extensive numerical simulations supporting our theoretical results.

1 Introduction

In modern statistics, the number of predictors can far exceed the number of observations available. However, in this high-dimensional context, ℓ_1 regularisation allows for a small number of predictors to be selected (referred to as the selected support) while allowing for a minimax optimal prediction error, see for instance [Van de Geer, 2016, Chapter 2]. The estimated parameters and support are not explicitly known and are obtained by solving a convex optimisation program in practice. This makes inference of the model parameters difficult if not impossible. One solution is to infer conditionally on the selected support. In this framework, it is possible to give a confidence interval and test any linear statistic.

The ubiquity of the logistic model to solve practical regression problems and the surge of high dimensional data-sets make the sparse logistic regression (SLR) more and more attractive. In this context, it becomes essential to provide certifiable guarantees on the output of the SLR, e.g. confidence intervals.

*This work was supported by grants from Région Ile-de-France.

1.1 Post-Selection Inference for high-dimensional generalized linear model

We are interested in a target parameter $\vartheta^\star \in \Theta \subseteq \mathbb{R}^d$ attached to the distribution $\mathbb{P}_{\vartheta^\star}$ of N independent response variables $Y := (y_1, \dots, y_N) \in \mathcal{Y}^N \subseteq \mathbb{R}^N$ given by the data $Z := (z_1, \dots, z_N)$ where $z_i = (\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ with $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ a covariate, namely a vector of d predictors. The family of generalized linear models, or GLMs for short, is based on modeling the conditional distribution of the responses $y_i \in \mathcal{Y}$ given the covariate $\mathbf{x}_i \in \mathcal{X}$ in an exponential family form, namely

$$\mathbb{P}_{\vartheta^\star}(y|\mathbf{x}) = h_{\vartheta^\star}(y) \exp\left\{ \frac{y\langle \mathbf{x}, \vartheta^\star \rangle - \xi(\langle \mathbf{x}, \vartheta^\star \rangle)}{v} \right\},$$

where $v > 0$ is a scale parameter, and $\xi : \mathbb{R} \to \mathbb{R}$ is the partition function which is assumed to be of class $\mathcal{C}^m$ (with m a non-negative integer). With a slight abuse of notations, we will simply denote $\mathbb{P}_{\vartheta^*}(\cdot \,|\, \mathbf{x})$ by $\mathbb{P}_{\vartheta^*}(\cdot)$. Standard examples are $\xi(t) = t^2/2$ for the Gaussian linear model with noise variance v and observation space $\mathcal{Y} = \mathbb{R}$, or $v = 1$, $\xi(t) = \log(1 + \exp(t))$ and $\mathcal{Y} = \{0, 1\}$ for the logistic regression. The negative log-likelihood takes the form

$$\forall \vartheta \in \Theta\,,\ \mathcal{L}_N(\vartheta, Z) := \sum_{i=1}^{N} \xi(\langle \mathbf{x}_i, \vartheta \rangle) - \langle y_i \mathbf{x}_i, \vartheta \rangle\,. \tag{1}$$

We assume that the partition function ξ is differentiable, then the score function $\nabla_\vartheta \mathcal{L}_N(\vartheta)$ is given by

$$\forall \vartheta \in \Theta\,,\ \nabla_\vartheta \mathcal{L}_N(\vartheta, Z) = \mathbf{X}^\top \big(\sigma(\mathbf{X}\vartheta) - Y\big)\,,$$

where $\sigma = \xi'$ is the derivative of the partition function and $\mathbf{X} \in \mathbb{R}^{N \times d}$ is referred to as the design matrix whose rows are the covariates and the columns are the predictors. In a high-dimensional context one has more predictors than observations (*i.e.,* $N \ll d$), and one would like to select a small number of predictors to explain the response. We use an ℓ_1-regularization to enforce a structure of sparsity in ϑ. Our overall estimator is based on solving the generalized linear Lasso

$$\hat{\vartheta}^\lambda \in \arg\min_{\vartheta \in \Theta} \big\{ \mathcal{L}_N(\vartheta, Z) + \lambda \|\vartheta\|_1 \big\} \tag{2}$$

where $\lambda > 0$ is a user-defined regularization hyperparameter. We assume that the negative log-likelihood is strictly convex. This assumption is satisfied for instance in the Gaussian linear model or logistic regression. In this case, it is necessary and sufficient that the solutions $\hat{\vartheta}^\lambda$ to (2) satisfy the following Karush–Kuhn–Tucker (KKT) conditions

$$\begin{cases} \mathbf{X}^\top \Big(Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)\Big) = \lambda \widehat{S} & \text{(3a)} \\ \widehat{S}_k = \operatorname{sign}(\hat{\vartheta}^\lambda_k) \qquad \text{if } \widehat{\vartheta}^\lambda_k \neq 0 & \text{(3b)} \\ \widehat{S}_k \in [-1, 1] \qquad \text{if } \hat{\vartheta}^\lambda_k = 0 & \text{(3c)} \end{cases}$$

Proposition 1 shows that there exists only one vector of signs $\widehat{S} \in \mathbb{R}^d$ such that $(\hat{\vartheta}^\lambda, \hat{S})$ satisfies the KKT conditions for some $\hat{\vartheta}^\lambda \in \Theta$. The proof of Proposition 1 can be found in Section A.1.

Proposition 1. *Let $Y \in \mathcal{Y}^N$ and let the partition function ξ be strictly convex. Then, there exists a unique $\widehat{S}(Y)$ such that for any couple $(\hat{\vartheta}^\lambda, \hat{S})$ satisfying the KKT conditions (cf. Eq.(3)) it holds that $\hat{S} = \widehat{S}(Y)$. Furthermore, one has*

$$\widehat{S}(Y) := \frac{1}{\lambda}\mathbf{X}^\top(Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)),$$

where $\hat{\vartheta}^\lambda$ is any solution of the generalized linear Lasso as defined in (2).

We define the *equicorrelation set* as

$$\widehat{M}(Y) := \{k \in [d] \mid |\widehat{S}_k(Y)| = 1\}.$$

In the following, we will identify the equicorrelation set and the set of predictors with nonzero coefficients $\{k \in [d] \mid \hat{\vartheta}^\lambda_k \neq 0\}$, also called "selected" model. Since $|\widehat{S}_k(Y)| = 1$ for any $\hat{\vartheta}^\lambda_k \neq 0$, the equicorrelation set does in fact contain all predictors with nonzero coefficients, although it may also include some predictors with zero coefficients. However, we work in this paper with Assumption 1, ensuring that the equicorrelation set is precisely the set of predictors with nonzero coefficients.

Assumption 1. *Problem (2) is non degenerate: $\widehat{S}(Y) \in \operatorname{relint} \partial\|\cdot\|_1$, where* relint *denotes the relative interior.*

Let us highlight that this assumption has already been used in the context of GLMs [cf. Massias et al., 2020, Assumption 8], and is common in works on support identification (cf. Candes and Recht [2013], Vaiter et al. [2015]).

For any set of indexes $M \subseteq [d]$ with cardinality s, we denote by Θ_M the set of target parameters induced by the support M namely,

$$\Theta_M := \{\vartheta_M \,|\, \vartheta \in \Theta\} \subseteq \mathbb{R}^s.$$

We aim at making inference conditionally on the *selection event* E_M defined as

$$E_M := \left\{Y \in \mathcal{Y}^N \mid \widehat{M}(Y) = M\right\}, \tag{4}$$

namely, the set of all observations Y that induced the same equicorrelation set M with the generalized linear lasso.

1.2 Characterization of the selection event

Following the approach of Lee et al. [2016], given some $M \subseteq [d]$ with $|M| = s$ and $S_M \in \{-1, +1\}^s$, we first characterize the event

$$E_M^{S_M} := \{Y \in E_M \mid \widehat{S}_M(Y) = S_M\}, \tag{5}$$

and we obtain E_M as a corollary by taking a union over all possible vectors of signs S_M. Proposition 2 gives a first description of $E_M^{S_M}$ and its proof is postponed to Section A.2.

Proposition 2. *Let us consider $M \subseteq [d]$ with $|M| = s$ and $S_M \in \{-1,+1\}^s$. It holds*

$$E_M^{S_M} = \Big\{ Y \in \mathcal{Y}^N \mid \exists \theta \in \Theta_M \text{ s.t. } (i)\ \mathbf{X}_M^\top \left(Y - \sigma(\mathbf{X}_M\theta)\right) = \lambda S_M \quad (ii)\ \operatorname{sign}(\theta) = S_M \quad (iii)\ \left\|\mathbf{X}_{-M}^\top \left(Y - \sigma(\mathbf{X}_M\theta)\right)\right\|_\infty < \lambda \Big\}, \tag{6}$$

were $\mathbf{X}_M \in \mathbb{R}^{s\times N}$ (resp. $\mathbf{X}_{-M} \in \mathbb{R}^{(d-s)\times N}$) is the submatrix obtained from $\mathbf{X}$ by keeping the columns indexed by M (resp. its complement).

With Proposition 1, we proved the uniqueness of the vector of signs satisfying the KKT conditions as soon as ξ is strictly convex. By considering additionally that $\mathbf{X}_M$ has full column rank, we claim that there exists a unique $\theta \in \Theta_M$ that satisfies the condition (i) in the definition of the selection event $E_M^{S_M}$ (see Eq.(6)). This statement will be a direct consequence of Proposition 3 (proved in Section A.3) which ensures that the map Ξ arising in Eq.(6) and defined by

$$\begin{aligned} \Xi : \Theta_M &\to \mathbb{R}^s \\ \theta &\mapsto \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta) \end{aligned} \tag{7}$$

is a $\mathcal{C}^m$-diffeomorphism whose inverse is denoted by Ψ.

Proposition 3. *We consider that the partition function ξ is strictly convex and we further assume that the set $M \subseteq [d]$ is such that $\mathbf{X}_M$ has full column rank. Then Ξ is a $\mathcal{C}^m$-diffeomorphism from Θ_M to $\operatorname{Im}(\Xi) = \{\mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta) \mid \theta \in \Theta_M\}$.*

Using Propositions 2 and 3, we are able to provide a new description of the selection event $E_M^{S_M}$ which can be understood as the counterpart of [Lee et al., 2016, Proposition 4.2].

Theorem 1. *Suppose that ξ is strictly convex. Given some $M \subseteq [d]$ with cardinal s such that $\mathbf{X}_M$ has full column rank and $S_M \in \{-1,1\}^s$, it holds*

$$E_M^{S_M} = \Big\{ Y \in \mathcal{Y}^N \mid \text{s.t. } \rho = -\lambda S_M + \mathbf{X}_M^\top Y \text{ satisfies } (a)\ \rho \in \operatorname{Im}(\Xi) \quad (b)\ \operatorname{Diag}(S_M)\Psi(\rho) \geq 0 \quad (c)\ \left\|\mathbf{X}_{-M}^\top \left(Y - \sigma(\mathbf{X}_M\Psi(\rho))\right)\right\|_\infty < \lambda \Big\}. \tag{8}$$

Remark. In the linear model, $\Xi : \theta \mapsto \mathbf{X}_M^\top \mathbf{X}_M \theta$ has full rank and thus condition (a) from Eq.(8) always holds.

1.3 Which parameters can be inferred?

Once a model has been selected, two different modeling assumptions are generally considered when we derive post-selection inference procedures, see for instance [cf. Fithian et al., 2014, Section 4]. This choice appears to be essential since it determines the parameters on which inference is conducted. In the following, we consider the mean value

$$\pi^* := \mathbb{E}_{\vartheta^*}[Y] = \sigma(\mathbf{X}\vartheta^*)\,. \tag{9}$$

Note that π^* allows to define the Bayes predictor in the logistic or the linear model. As presented in Fithian et al. [2014], the analyst should decide to work either under the so-called *saturated model* or the *selected model*. In the following, we discuss these concepts for arbitrary GLMs and Table 1 summarizes the key concepts.

Model	Selected	Weak selected	Saturated
Assumption	$\sigma^{-1}(\pi^*) \in \mathrm{Im}(\mathbf{X}_M)$	$\mathbf{X}_M^\top \pi^* \in \mathrm{Im}(\Xi)$	None
Statistic of interest	$\Psi(\mathbf{X}_M^\top Y)$	$\Psi(\mathbf{X}_M^\top Y)$	$\mathbf{X}_M^\top Y$
Inferred parameter	$\theta^* \in \Theta_M$ s.t. $\pi^* = \sigma(\mathbf{X}_M \theta^*)$	$\theta^* \in \Theta_M$ s.t. π^* and $\sigma(\mathbf{X}_M\theta^*)$ have the same projections on the column span of $\mathbf{X}_M$	$\mathbf{X}_M^\top \pi^*$

Table 1: Once a model has been selected, we may infer some parameters assuming one of the three modeling: selected model, weak selected model, and saturated model respectively based on the assumptions described in the first row. In this case, inference on the quantities described on the third row can be done from the statistic described in the second row.

The (weak) selected model: Parameter inference. In the *weak selected model*, we consider that the data have been sampled from the GLM (cf. Eq.(1)) and we assume that the selected model M is such that

$$\mathbf{X}_M^\top \sigma(\mathbf{X}\vartheta^*) \in \mathrm{Im}(\Xi)\,, \tag{10}$$

and recall that $\mathbf{X}_M^\top \pi^* = \mathbf{X}_M^\top \mathbb{E}_{\vartheta^*}[Y] = \mathbf{X}_M^\top \sigma(\mathbf{X}\vartheta^*)$. This is equivalent to state that there exists some vector $\theta^* \in \Theta_M$ satisfying

$$\mathbf{X}_M^\top \pi^* = \Xi(\theta^*)\,,$$

and recall that $\Xi(\theta^*) = \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta^*)$. In this framework, we have the possibility to make inference on the parameter vector $\theta^* := \Psi(\mathbf{X}_M^\top \pi^*)$ itself. Let us point out that the condition $\mathbf{X}_M^\top \pi^* \in \mathrm{Im}(\Xi)$ is equivalent to the existence of the conditional Maximum Likelihood Estimator (MLE) $\widehat{\theta}$ which is defined as the MLE working with the design $\mathbf{X}_M$ and Y distributed according to $\mathbb{P}_{\vartheta^*}(\cdot \,|\, E_M)$ (see Eq.(17)). When it exists, the conditional MLE is unique and is given by $\widehat{\theta} = \Psi(\mathbf{X}_M^\top Y)$.

In the *selected model*, we replace the condition from Eq.(10) by the stronger assumption that there exists $\theta^* \in \Theta_M$ such that

$$\mathbf{X}_M \theta^* = \mathbf{X}\vartheta^*. \tag{11}$$

This assumption is always satisfied for the global null hypothesis $\vartheta^* = 0$ for which the aforementioned condition holds with $\theta^* = 0$.

The saturated model: Mean value inference. The assumption from Eq.(10) or (11) can be understood as too restrictive since the analyst can never check in practice that this condition holds, except for the global null. This is the reason why one may prefer to consider the so-called *saturated model* where we only assume that the data have been sampled from the GLM.

In this case it remains meaningful to provide post-selection inference procedure for transformation of π^*. A typical choice is to consider linear transformation of π^* and among them, one may focus specifically on transformation of $\mathbf{X}_M^\top \pi^*$. This choice is motivated by remarking that this quantity characterizes the first order optimality condition for the unpenalized MLE $\widehat{\theta}$ for the design matrix $\mathbf{X}_M$ through $\mathbf{X}_M^\top Y = \Xi(\widehat{\theta})$, or by considering the example of linear model (as presented below).

The example of the linear model. Note that in linear regression, $\sigma = \mathrm{Id}$ and $\Psi : \rho \mapsto \left(\mathbf{X}_M^\top \mathbf{X}_M\right)^{-1} \rho$. Hence, Eq.(10) is equivalent to Eq.(11) meaning that the selected and the weak selected models coincide. Moreover, in both the saturated and the selected models, we aim at making inference on transformations of $\Psi(\mathbf{X}_M^\top \pi^*) = \mathbf{X}_M^+ \pi^*$ (where $\mathbf{X}_M^+$ is the pseudo-inverse of $\mathbf{X}_M$). While in the (weak) selected model, this quantity corresponds to the parameter vector θ^* satisfying $\pi^* = \mathbf{X}_M \theta^*$, in the saturated model, it corresponds to the best linear predictor in the population for design matrix $\mathbf{X}_M$ in the sense of the squared L^2-norm.

Notations. Let us consider some set of selected variables $M \subseteq [d]$ with $s := |M|$ and some $\vartheta^* \in \mathbb{R}^d$. By assuming that ξ is strictly convex, one can compute $\mathbf{X}\vartheta^*$ from $\pi^* = \sigma(\mathbf{X}\vartheta^*)$, allowing us to denote equivalently $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\vartheta^*}$ with an abuse of notation. In the selected model with $\theta^* \in \Theta_M$ satisfying Eq.(11), we will also denote $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\theta^*}$. Finally, we will denote by $\overline{\mathbb{P}}_{\pi^*}$ the distribution of Y conditional on E_M, namely

$$\overline{\mathbb{P}}_{\pi^*}(Y) \propto \mathbb{1}_{Y \in E_M} \mathbb{P}_{\pi^*}(Y),$$

where $\propto$ means equal up to a normalization constant.

1.4 Inference procedures

We provide a general approach to obtain asymptotically valid PSI methods, both in the saturated and the selected models. The proposed PSI methods rely on two key ingredients, namely conditional Central Limit Theorems (CLTs) and conditional sampling. Before describing our selective inference procedures, let us set the rigorous framework for which we provide our conditional CLTs.

Preliminaries. Given a non-decreasing sequence of positive integers $(d_N)_{N \in \mathbb{N}}$ converging to $d_\infty \in \mathbb{N} \cup \{+\infty\}$, we consider for any N a matrix $\mathbf{X}^{(N)} \in \mathbb{R}^{N \times d_N}$ and a vector $[\vartheta^*]^{(N)} \in \mathbb{R}^{d_N}$. Let $s \in [d_1, d_\infty] \cap \mathbb{N}$ be a fixed and finite integer and let us consider for any N a set $M^{(N)} \subseteq [d_N]$ with cardinality s. Considering further a sequence of positive reals $(\lambda^{(N)})_{N \in \mathbb{N}}$, we define $E_M^{(N)}$ as the selection event corresponding to the tuple $(\lambda^{(N)}, \mathbf{X}^{(N)}, M^{(N)})$ meaning that

$$E_M^{(N)} := \bigcup_{S_M \in \{\pm 1\}^N} \left[E_M^{S_M}\right]^{(N)},$$

where $\left[E_M^{S_M}\right]^{(N)}$ is the set given by Eq.(8) when one uses the regularization parameter $\lambda^{(N)}$, the design matrix $\mathbf{X}^{(N)}$ and the set of active variables $M^{(N)}$. Considering that the $\mathcal{Y}^N$-valued random vector Y is distributed according to

$$\overline{\mathbb{P}}_{\pi^*}^{(N)}(Y) \propto \mathbb{1}_{Y \in E_M^{(N)}} \mathbb{P}_{\pi^*}^{(N)}(Y),$$

with $\mathbb{P}_{\pi^*}^{(N)} := \mathbb{P}_{\sigma(\mathbf{X}^{(N)}[\vartheta^*]^{(N)})}$, the cornerstone of our methods consists in proving a CLT for $\left[\mathbf{X}_{M^{(N)}}^{(N)}\right]^\top Y$ in the saturated model, see Eq.(12) (resp. a CLT for the conditional MLE $\Psi(\mathbf{X}_M^\top Y)$ in the selected model, see Eq.(13)).

In Section 4, we consider the specific case of the logistic regression and we establish conditional CLTs of the form given by Eqs.(12) and (13). The proofs of our CLTs rely on triangular arrays of dependent random vectors of the form $(\xi_{i,N})_{i\in[N]}$. For any fixed N and any $i \in [N]$, $\xi_{i,N}$ is a random vector in $\mathbb{R}^s$ which can be written as a function of the deterministic quantities $\lambda^{(N)}$, $\mathbf{X}^{(N)}$, $M^{(N)}$ and of the random variable Y with probability distribution $\overline{\mathbb{P}}_{\pi^*}^{(N)}$. In Section 4, we provide conditions ensuring that the rows of the triangular system $\big((\xi_{i,N})_{i\in[N]},\, N \in \mathbb{N}\big)$ satisfy some Lindeberg's condition.

With this detailed framework now established, we will take the liberty in the remainder of this paper of adopting certain abuses of notation for the sake of readability. The notational clutter will be in particular reduced by forgetting to specify the dependence on N meaning that we will simply refer to $\mathbf{X}^{(N)}$, $M^{(N)}$, d_N, $[\vartheta^*]^{(N)}, \overline{\mathbb{P}}_{\pi^*}^{(N)}, \dots$ as $\mathbf{X}$, M, d, $\vartheta^*, \overline{\mathbb{P}}_{\pi^*}, \dots$. Nevertheless, let us stress again that the integer s is fixed and does not depend on N in our work.

Conditional CLTs. The cornerstone of our method is to establish conditional CLTs. More precisely, considering that Y is distributed according to $\overline{\mathbb{P}}_{\pi^*} = \mathbb{P}_{\pi^*}(\cdot \,|\, E_M)$, we aim at providing conditions ensuring that

- in the saturated model,

$$\overline{G}_N(\pi^*)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^*}) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s), \tag{12}$$

  for some $\overline{G}_N(\pi^*) \in \mathbb{R}^{s\times s}$ and $\overline{\pi}^{\pi^*} \in \mathbb{R}^N$ depending only on π^* and E_M,

- in the selected model where $\mathbf{X}\vartheta^* = \mathbf{X}_M\theta^*$,

$$\overline{V}_N(\theta^*)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta^*)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s), \tag{13}$$

  for some $\overline{V}_N(\theta^*) \in \mathbb{R}^{s\times s}$ and $\overline{\theta}(\theta^*) \in \mathbb{R}^s$ depending only on θ^* and E_M.

In the case of logistic regression, we give in Section 4 conditions ensuring that the CLTs from Eqs.(12) and (13) hold.

Idyllic selective inference. Using the above mentioned conditional CLTs, one can obtain confidence regions (CRs) with asymptotic level $1-\alpha$ (for some $\alpha \in (0,1)$) conditional on E_M as follows,

- in the saturated model, the CR for π^* is defined as

$$\{\pi \,|\, \|\overline{G}_N(\pi)^{-1/2}(\mathbf{X}_M Y - \mathbf{X}_M^\top \overline{\pi}^\pi)\|_2^2 \leq \chi^2_{s,1-\alpha}\}, \tag{14}$$

- in the selected model, the CR for θ^* is defined as

$$\{\theta \,|\, \|\overline{V}_N(\theta)^{1/2}(\Psi(\mathbf{X}_M Y) - \overline{\theta}(\theta))\|_2^2 \leq \chi^2_{s,1-\alpha}\}, \tag{15}$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. Obviously, covering the whole space to obtain in practice the CRs from Eqs.(14) and (15) is out of reach and one could use a discretization of a bounded domain to bypass this limitation. A more involved issue is that $\overline{G}_N(\pi)$ and $\overline{\pi}^\pi$, (resp. $\overline{V}_N(\theta)$ and $\overline{\theta}(\theta)$) can be written as expectations with respect to the conditional distribution $\overline{\mathbb{P}}_\pi$ (resp. $\overline{\mathbb{P}}_\theta$). If closed-form expressions most of time do not exist, one can estimate the latter quantities by sampling from $\overline{\mathbb{P}}_\pi$ (resp. $\overline{\mathbb{P}}_\theta$). Depending on the studied GLM, this task may be expensive and the proposed grid-based approaches to get CRs would be unusable in practice due to the curse of dimensionality. In the next subparagraph, we propose an alternative method to overcome this issue.

Confidence regions in practice. Table 2 gives the main ideas allowing us to obtain a confidence region for $\mathbf{X}_M^\top \pi^*$ (resp. θ^*) in the saturated (resp. selected) model. While the blue terms are small with high probability for N large enough thanks to the previous established conditional CLTs, the red terms motivate us to choose our estimate $\pi^\bigstar$ (resp. $\theta^\bigstar$) among the minimizers of the map $\pi \mapsto \|\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^\pi\|_2$ (resp. $\theta \mapsto \|\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta)\|_2$). This way, we circumvent the curse of dimensionality of the above mentioned idyllic approach to obtain CRs. Nevertheless, the given CRs involve quantities that are unknown in practice such as the constant κ_1 (resp. κ_2) in Table 2 that encodes the (local) Lipschitz continuity of the gradient of the inverse of the map $\pi \mapsto \mathbf{X}_M^\top \overline{\pi}^\pi$ (resp. $\theta \mapsto \overline{\theta}(\theta)$).

Saturated model	$\forall \pi^\bigstar,\ \|\mathbf{X}_M^\top \pi^* - \mathbf{X}_M^\top \pi^\bigstar\|_2 \lesssim \kappa_1 \|\mathbf{X}_M^\top \overline{\pi}^{\pi^*} - \mathbf{X}_M^\top \overline{\pi}^{\pi^\bigstar}\|_2$ $\leq \kappa_1 \left\{ \|\mathbf{X}_M^\top \overline{\pi}^{\pi^*} - \mathbf{X}_M^\top Y\|_2 + \|\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \overline{\pi}^{\pi^\bigstar}\|_2 \right\}$
Selected model	$\forall \theta^\bigstar,\ \|\theta^* - \theta^\bigstar\|_2 \lesssim \kappa_2 \|\overline{\theta}(\theta^*) - \overline{\theta}(\theta^\bigstar)\|_2$ $\leq \kappa_2 \left\{ \|\overline{\theta}(\theta^*) - \Psi(\mathbf{X}_M^\top Y)\|_2 + \|\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta^\bigstar)\|_2 \right\}$

Table 2: Confidence intervals.

> In the case of logistic regression, we present in Section 5 with full details our selective inference procedures with theoretical guarantees.

Conditional sampling: A Monte-Carlo approach for hypothesis-testing. We consider hypothesis tests with pointwise nulls as presented in Table 3. One can then compute estimate $\widetilde{G}_N(\pi_0^*), \widetilde{\pi}^{\pi_0^*}$ (resp. $\widetilde{V}_N(\theta_0^*), \widetilde{\theta}(\theta_0^*)$) of the unknown quantities $\overline{G}_N(\pi_0^*)$, $\overline{\pi}^{\pi_0^*}$ (resp. $\overline{V}_N(\theta_0^*)$, $\overline{\theta}(\theta_0^*)$) by sampling from the conditional null distribution $\overline{\mathbb{P}}_{\pi_0^*}$ (resp. $\overline{\mathbb{P}}_{\theta_0^*}$ in the selected model). Using a Monte-Carlo approach with the CLTs from Eqs.(12) and (13), one can derive testing procedures that are asymptotically correctly calibrated.

	Null and alternative	Distributed *approximately* as $\mathcal{N}(0, \mathrm{Id}_s)$ under $\mathbb{H}_0$
Saturated model	$\mathbb{H}_0 : \{\pi^* = \pi_0^*\}$, $\mathbb{H}_1 : \{\pi^* \neq \pi_0^*\}$	$\widetilde{G}_N(\pi_0^*)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \widetilde{\pi}^{\pi_0^*})$
Selected model	$\mathbb{H}_0 : \{\theta^* = \theta_0^*\}$, $\mathbb{H}_1 : \{\theta^* \neq \theta_0^*\}$	$\widetilde{V}_N(\theta_0^*)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta}(\theta_0^*))$

Table 3: Hypothesis testing.

In the case of logistic regression, we rely on a gradient alignment viewpoint of the selection event to provide in Section 3 an algorithm which allows us to sample from $\overline{\mathbb{P}}_{\pi^*}$ given any π^*. In Section 5, we present our hypothesis tests in both the saturated and the selected models with theoretical guarantees.

1.5 Related works

In the Gaussian linear model with a known variance, the distribution of the linear transformation $\eta^\top Y$ (with $\eta^\top = e_k^\top \mathbf{X}_M^+$) is a truncated Gaussian conditionally on $E_M^{S_M}$ and $\mathrm{Proj}_\eta^\perp(Y)$. This explicit formulation of the conditional distribution allows to conduct exact post-selection inference procedures [cf. Fithian et al., 2014, Section 4]. However, when the noise is assumed to be Gaussian with an unknown variance, one needs to also condition on $\|Y\|^2$ which leaves insufficient information about θ_k^* to carry out a meaningful test in the saturated model [cf. Fithian et al., 2014, Section 4.2].

Outside of the Gaussian linear model, there is little hope to obtain an easy exact characterization of the conditional distribution of some transformation of $\mathbf{X}_M^\top Y$. In the following, we sketch a brief review of this literature, see references therein for further works on this subject.

- Linear model but non-Gaussian errors.
  Let us mention for example Tian and Taylor [2017], Tibshirani et al. [2018] where the authors consider the linear model but relaxed the Gaussian distribution assumption for the error terms. They prove that the response variable is asymptotically Gaussian so that applying the well-oiled machinery from Lee et al. [2016] gives asymptotically valid post-selection inference methods.

- GLM with Gaussian errors.
  Shi et al. [2020] consider generalized linear models with Gaussian noise and can then immediately apply the polyhedral lemma to the appropriate transformation of the response.

We classify existing works with Table 4.

One important challenge that remains so far only partially answered is the case of GLMs without Gaussian noise, such as in logistic regression. In Fithian et al. [2014], the authors derive powerful unbiased selective tests and confidence intervals among all selective level-α tests for inference in exponential family models after arbitrary selection procedures. Nevertheless, their approach is not well-suited to account for discrete response variable as it is the case in logistic

Noise	Linear Model	GLM
Gaussian	Lee et al. [2016]	Shi et al. [2020]
Non-Gaussian	Tian and Taylor [2017] and Tibshirani et al. [2018]	This paper and Taylor and Tibshirani [2018]

Table 4: Positioning of this paper among some pioneering works on PSI in GLMs.

regression. In Section 6.3 of the former paper, the authors rather encourage the reader to make use of the procedure proposed by Taylor and Tibshirani [2018] in such context. Both our work and Taylor and Tibshirani [2018] are tackling the problem of post selection inference in the logisitic model. Nevertheless, the proposed methods rely on different paradigms and we explain in Section 2 this difference in perspectives.

1.6 Contributions and organization of this paper

Working with an arbitrary GLM (Sec.1 and 2).

1. We provide a new formulation of the selection event in GLMs shedding light on the $\mathcal{C}^m$-diffeomorphism Ψ that carries the geometric information of the problem (cf. Theorem 1). Ψ allows us to define rigorously the notions of selected/saturated models for arbitrary GLM (cf. Sec.1.3).

2. We provide a new perspective on post-selection inference in the selected model for GLMs through the conditional MLE approach of which Ψ is a key ingredient (cf. Sec.2).

3. We introduce the C-cube conditions that are sufficient conditions in GLMs to obtain valid post-selection inference procedures in the selected model based on the conditional MLE approach (cf. Sec.2).

Considering the Sparse Logistic Regression (SLR) (from Sec.3).

4. Under some assumptions, we prove that the C-cube conditions hold for the SLR and we conduct simulations to support our results.

5. We also derive asymptotically valid PSI methods in the saturated model for the SLR.

6. We provide an extensive comparison between our work and the heuristic from Taylor and Tibshirani [2018] which is currently considered the best to use in the context of SLR [cf. Fithian et al., 2014, Section 6.3], as far as we know.

Outline. In Section 2, we introduce the conditional MLE approach to tackle PSI in the selected model and we stress the difference with the debiasing method from Taylor and Tibshirani [2018]. From Section 3, we focus specifically on the SLR. In Section 3, we rely on a *gradient-alignment* viewpoint on the selection event to design a simulated annealing algorithm which is proved–for an appropriate

cooling scheme–to provide iterates whose distribution is asymptotically uniform on the selection event. In Section 4, we provide two conditional central limit theorems that would be key theoretical ingredients for our PSI methods presented in Section 5. In Section 5.1, we give PSI procedures in the selected model while in Section 5.2 we focus on the saturated model. In Section 6, we present the results of our simulations.

Notations. For any set of indexes $M \subseteq [d] := \{1, \dots, d\}$ and any vector v, we denote by v_M the subvector of v keeping only the coefficients indexed by M, namely $v_M = (v_k)_{k \in M}$. Analogously, v_{-M} will refer to the subvector $(v_k)_{k \notin M}$. $|M|$ denotes the cardinality of the finite set M. For any $x \in \mathbb{R}^d$, $\|x\|_\infty := \sup_{i \in [d]} |x_i|$ and for any $p \in [1, \infty)$, $\|x\|_p^p := \sum_{i \in [d]} x_i^p$. For any $A \in \mathbb{R}^{d \times p}$, we define the Frobenius norm of A as $\|A\|_F := (\sum_{i \in [d], j \in [p]} A_{i,j}^2)^{1/2}$ and the operator norm of A as $\|A\| := \sup_{x \in \mathbb{R}^p, \|x\|_2 = 1} \|Ax\|_2$. We further denote by A^+ the pseudo-inverse of A. Considering that A is a symmetric matrix, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ will refer respectively to the minimal and the maximal eigenvalue of A. $\odot$ denotes the Hadamard product namely for any $A, B \in \mathbb{R}^{d \times p}$, $A \odot B := (A_{i,j} B_{i,j})_{i \in [d], j \in [p]}$. By convention, when a function with real valued arguments is applied to a vector, one need to apply the function entrywise. $\mathrm{Id}_d \in \mathbb{R}^{d \times d}$ will refer to the identity matrix and $\mathcal{N}(\mu, \Sigma)$ will denote the multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ. For any $x \in \mathbb{R}^d$ and for $p \in [1, \infty]$, we define $\mathbb{B}_p(x, r) = \{z \in \mathbb{R}^d \,|\, \|z\|_p \leq R\}$.

2 Regularization bias and conditional MLE

In this section, we wish to emphasize the different nature of our approach and that of Taylor and Tibshirani [2018] which we consider as the more relevant point of comparison, to the best of our knowledge. While we rely on a conditional MLE viewpoint, the former paper consider a debiasing approach.

- *The debiasing approach*
  ℓ_1-penalization induced a soft-thresholding bias and one can first try to modify the solution of the penalized GLM $\hat{\vartheta}^\lambda$ to approximate the unconditional MLE of the GLM using only the features in the selected support M by some vector $\underline{\theta}$. Provided that we work with a *correctly specified model* M*–i.e.,* one that contains the true support $\{j \in [d] \,|\, \vartheta_j^* \neq 0\}$– standard results ensure that the unconditional MLE is asymptotically normal, asymptotically efficient and centered at ϑ_M^*. If one can show that the selection event only involve polyhedral constraints on a linear transformation $\eta^\top \underline{\theta}$ of the debiased vector $\underline{\theta}$, the conditional distribution of $\eta^\top \underline{\theta}$ would be a truncated Gaussian. This is the approach from Taylor and Tibshirani [2018] that we detail in Section 2.1.

- *The conditional MLE viewpoint*
  In this paper we follow a different route: one can grasp the nettle by studying directly the properties of the unpenalized conditional MLE.

2.1 Selective inference through debiasing

The idea behind the method proposed by Taylor and Tibshirani [2018] is that we need two key elements to mimic the approach from Lee et al. [2016] proposed in the linear model with Gaussian errors:

- A statistic $T(Y)$ converging in distribution to a Gaussian distribution with a mean involving the parameter of interest;
- A selection event that can be written as a union of polyhedra with respect to $\eta^\top T(Y)$ for some vector η.

In practice, a solution of the generalized linear Lasso (cf. Eq.(2)) can be approximated using the Iteratively Reweighted Least Squares (IRLS). Defining

$$W(\vartheta) = \nabla^2_\eta \mathcal{L}_N(\eta)\big|_{\eta=\mathbf{X}\vartheta} = \mathrm{Diag}(\sigma'(\mathbf{X}\vartheta)),$$
$$\text{and } z(\vartheta) = X\vartheta - [W(\vartheta)]^{-1}\nabla_\eta \mathcal{L}_N(\eta)\big|_{\eta=\mathbf{X}\vartheta} = X\vartheta + [W(\vartheta)]^{-1}(Y - \sigma(\mathbf{X}\vartheta)),$$

the IRLS algorithm works as follows.

1: Initialize $\vartheta_c = 0$.
2: Compute $W(\vartheta_c)$ and $z(\vartheta_c)$.
3: Update the current value of the parameters with
$$\vartheta_c \leftarrow \arg\min_\vartheta \frac{1}{2}(z(\vartheta_c) - \mathbf{X}\vartheta)^\top W(\vartheta_c)(z(\vartheta_c) - \mathbf{X}\vartheta) + \lambda\|\vartheta\|_1.$$
4: Repeat steps 2. and 3. until convergence.

If the IRLS has converged, we end up with a solution $\hat{\vartheta}^\lambda$ of Eq.(2) and, for $M = \{j \in [d] \,|\, \hat{\vartheta}^\lambda_j \neq 0\}$, the active block of stationary conditions (Eq. (6) (i)) can be written as

$$\mathbf{X}_M^\top W(z - \mathbf{X}_M \hat{\vartheta}^\lambda_M) = \lambda S_M,$$

where $W = W(\hat{\vartheta}^\lambda)$, $z = z(\hat{\vartheta}^\lambda)$ and $S_M = \mathrm{sign}(\hat{\theta}^\lambda_M)$. The solution $\hat{\vartheta}^\lambda_M$ should be understood as a biased version of the unpenalized MLE $\widehat{\theta}$ obtained by working on the support M, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta_M} \sum_{i=1}^N \xi(\langle \mathbf{X}_{i,M}, \theta\rangle) - \langle y_i \mathbf{X}_{i,M}, \theta\rangle.$$

If we work with a *correctly specified model* M*–i.e.,* one that contains the true support $\{j \in [d] \,|\, \vartheta^*_j \neq 0\}$–then it follows from standard results that the MLE $\widehat{\theta}$ is a consistent and asymptotically efficient estimator of ϑ^*_M (see e.g. [Van der Vaart, 2000, Theorem 5.39]). A natural idea consists in debiasing the vector of parameters ϑ^λ_M in order to get back to the parameter $\widehat{\theta}$ and to use its nice asymptotic properties for inference. We thus consider

$$\underline{\theta} = \vartheta^\lambda_M + \lambda \left(\mathbf{X}_M^\top W \mathbf{X}_M\right)^{-1} S_M,$$

so that $\underline{\theta}$ satisfies

$$\mathbf{X}_M^\top W(z - \mathbf{X}_M \underline{\theta}) = 0. \tag{16}$$

If one replaces W and z in Eq.(16) by $W(\underline{\vartheta})$ and $z(\underline{\vartheta})$ (with the obvious notation that $\underline{\vartheta}_M = \underline{\theta}$ and $\underline{\vartheta}_{-M} = 0$), Eq.(16) corresponds to the stationarity condition of the unpenalized MLE for the generalized linear regression using only the features in M.

Hence, Taylor and Tibshirani [2018] propose to treat the debiased parameters $\underline{\theta}$ has asymptotically normal centered at ϑ^*_M with covariance matrix $\left(\mathbf{X}_M^\top W(\vartheta^*)\mathbf{X}_M\right)^{-1}$. Since ϑ^* is unknown, they use a Monte-Carlo estimate and replace $W(\vartheta^*)$ by $W(\hat{\vartheta}^\lambda)$ in the Fisher information matrix. By considering that $\vartheta^* = N^{-1/2}\beta^*$ where each entry of β^* is independent of N, they claim that the selection event $E_M^{S_M}$ can be asymptotically approximated by

$$\mathrm{Diag}(S_M)\left(\underline{\theta} - \lambda\left(\mathbf{X}_M^\top W \mathbf{X}_M\right)^{-1} S_M\right) \geq 0.$$

Hence, to derive post-selection inference procedure, they apply the polyhedral lemma to the limiting distribution of $N^{1/2}\underline{\theta}$, with M and S_M fixed.

2.2 Selective inference through conditional MLE

We change of paradigm and we directly work with the conditional distribution. The conditional distribution given $Y \in E_M$ is a conditional exponential family with the same natural parameters and sufficient statistics but different support and normalizing constant:

$$\overline{\mathbb{P}}_\theta(Y) \propto \mathbb{1}_{E_M}(Y) \prod_{i=1}^N h_\theta(y_i) \exp\left\{\frac{y_i \mathbf{X}_{i,M}\theta - \xi(\mathbf{X}_{i,M}\theta)}{v}\right\},$$

where the symbol $\propto$ means "proportional to". When $E_M = \mathcal{Y}^N$ (*i.e.,* when there is no conditioning), we will simply denote $\overline{\mathbb{P}}_\theta$ by $\mathbb{P}_\theta$. In the following we will denote by $\overline{\mathbb{E}}_\theta$ (resp. $\mathbb{E}_\theta$) the expectation with respect to $\overline{\mathbb{P}}_\theta$ (resp. $\mathbb{P}_\theta$). We want to conduct inference on θ^* (from Eq.(11)) based on the conditional and unpenalized MLE computed on the selected model M, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta_M} \mathcal{L}_N(\theta, Z^M), \tag{17}$$

where $Z^M = (Y, \mathbf{X}_M)$ and where Y is distributed according to $\overline{\mathbb{P}}_{\theta^*}$. We aim at proving a Central Limit Theorem for $\widehat{\theta}$. A natural candidate for the mean of the asymptotic Gaussian distribution of $\widehat{\theta}$ is the minimizer of the conditional expected negative log-likelihood defined by

$$\overline{\theta}(\theta^*) \in \arg\min_{\theta \in \Theta_M} \overline{\mathbb{E}}_{\theta^*}\left[\mathcal{L}_N(\theta, Z^M)\right], \tag{18}$$

which is the minimizer of the conditional risk $\theta \mapsto \overline{\mathbb{E}}_{\theta^*}\left[\mathcal{L}_N(\theta, Z^M)\right]$.

In the following, when there is no ambiguity we will simply denote $\overline{\theta}(\theta^*)$ by $\overline{\theta}$. Hence, denoting

$$L_N(\theta, Z^M) = \frac{\partial \mathcal{L}_N}{\partial \theta}(\theta, Z^M) \quad \text{and} \quad \overline{L}_N(\theta, \mathbf{X}_M) = \overline{\mathbb{E}}_{\theta^*}\left[\frac{\partial \mathcal{L}_N}{\partial \theta}(\theta, Z^M)\right],$$

it holds that the conditional unpenalized MLE $\widehat{\theta}$ and the minimizer $\overline{\theta}$ of the conditional risk satisfy the first order condition

$$\begin{aligned} L_N(\widehat{\theta}, Z^M) = 0 \quad i.e. \quad \mathbf{X}_M^\top (Y - \pi^{\widehat{\theta}}) = 0 \Leftrightarrow \widehat{\theta} = \Psi(\mathbf{X}_M^\top Y), \\ \text{and } \overline{L}_N(\overline{\theta}, \mathbf{X}_M) = 0 \quad i.e. \quad \mathbf{X}_M^\top (\overline{\pi}^{\theta^*} - \pi^{\overline{\theta}}) = 0 \Leftrightarrow \overline{\theta} = \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta^*}), \end{aligned} \tag{19}$$

where $\pi^\theta = \mathbb{E}_\theta[Y] = \sigma(\mathbf{X}_M \theta)$ and $\overline{\pi}^\theta = \overline{\mathbb{E}}_\theta[Y]$.

Let us now introduce what we will call the C-cube conditions in this paper.

Conditional Sampling We are able to sample from the distribution $\overline{\mathbb{P}}_\theta$.

Computing Ψ We are able to compute efficiently $\Psi(\rho)$ for any $\rho \in \mathbb{R}^s$.

Conditional CLT Under appropriate conditions, we have the following CLT

$$u^\top [\overline{G}_N(\theta^*)]^{1/2} (\widehat{\theta} - \overline{\theta}) \xrightarrow[N \to +\infty]{(d)} \mathcal{N}(0, 1),$$

where u is a unit s-vector (with $s = |M|$), $\widehat{\theta} = \Psi(\mathbf{X}_M^\top Y)$ is the MLE, and $\overline{G}_N(\theta^*)$ is a positive semi-definite $(s \times s)$-matrix.

In any GLM where the C-cube conditions are satisfied, one can adapt the methods of this paper to design asymptotically valid PSI procedures with respect to the selected model.

2.3 Discussion

Duality between conditional MLE and debiasing approaches. Over-simplifying the situation, our approach could be understood as the dual counterpart of the one from Taylor and Tibshirani [2018] in the sense that the former paper is first focused on getting an (unconditional) CLT and deal with the selection event in a second phase. On the contrary, we are first focused on the conditional distribution (*i.e.*, we want to be able to sample from the conditional distribution) while the asymptotic (conditional) distribution considerations come thereafter.

What about the saturated model? In this section, we have presented the conditional MLE approach in the selected model. Nevertheless, we provide in this paper asymptotically valid post-selection inference procedures on $\mathbf{X}_M^\top \pi^*$ in the saturated model for logistic regression. Let us stress that this approach could also be adapted to obtain analogous methods in other GLMs.

Comprehensive comparison between our work and the one from Taylor and Tibshirani [2018]. In Taylor and Tibshirani [2018], the authors consider only the more restrictive framework of the selected model where $\mathbf{X}\vartheta^* = \mathbf{X}_M \theta^*$ for some $\theta^* \in \mathbb{R}^s$. Their method allows to conduct PSI inference on any linear transformation of θ^* (including in particular the local coordinates θ_j^* for $j \in [s]$), and can be efficiently used in practice. The authors do not provide a formal proof of their claim but rather motivate their approach with asymptotic arguments where they consider in particular that $\vartheta^* = N^{-1/2}\beta^*$ where each entry of β^* is independent of N.

On the other hand, this paper presents *global* PSI methods in both the saturated and the selected models, in the sense that statistical inference is conducted on the vector-valued parameter of interest. Our methods are computationally more expensive than the one from Taylor and Tibshirani [2018], but they are proved to be asymptotically valid under some set of assumptions that we discuss in details in Section 4.4. Table 5 sums up this comparison.

	Taylor and Tibshirani [2018]	This paper
Selected model	✓	✓
Saturated model	✗	✓
Hypotheses tested in the selected model	Local: $\theta_j^* = [\theta_0^*]_j$ for some j	Global: $\theta^* = \theta_0^*$
Formal proof	✗	✓
Assumption on $\vartheta^* = \alpha_N^{-1}\beta^*$ with entries of β^* independent of N	For the theoretical sketches supporting their result, they consider $\alpha_N = N^{1/2}$.	Require $\alpha_N = \omega(N^{1/2})$, that could be weakened (Sec.4.4).
Low computational cost	✓	✗

Table 5: Comparison between our work and the one from Taylor and Tibshirani [2018].

The logistic regression. In the remaining sections of this paper, we focus on the logistic regression case. This means in particular that $\xi(x) = \ln(1 + \exp(x))$ is the softmax function and its derivative $\sigma(x) = (1 + \exp(-x))^{-1}$ is the sigmoid function. We prove that the C-cube conditions hold in this framework under some assumptions, and we describe our methods for PSI for both the selected and the saturated models.

Note that our paper should be understood as an extension of the work from Meir and Drton [2017] to the SLR. Indeed, the authors of the former paper propose a method to compute the conditional MLE after model selection in the linear model. They show empirically that the proposed confidence intervals are close to the desired level but they are not able to provide theoretical justification of their approach.

3 Sampling from the conditional distribution

In this section, we present an algorithm based on a simulated annealing approach that is proved to sample states $Y^{(t)}$ uniformly distributed on the selection event E_M for any $M \subseteq [d]$ with cardinality s in the asymptotic regimes as $t \to \infty$. From this section, we consider the case of the logistic regression where we recall that $Y = (y_i)_{i\in[N]}$ and for all $i \in [N]$, $y_i \sim \text{Ber}(\pi_i^*)$ with $\pi^* = \sigma(\mathbf{X}\vartheta^*)$.

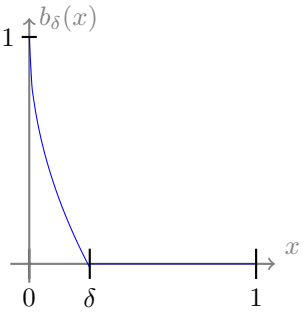
3.1 Numerical method to approximate the selection event

In this section, we present a simulated annealing algorithm to approximate the selection event E_M for some set $M \subseteq [d]$. From Proposition 1 and the KKT conditions from (3), we know that the selection event E_M can be written as

$$E_M = \left\{ Y \in \{0,1\}^N \mid \mathbb{1}_{\|\widehat{S}_{-M}(Y)\|_\infty - 1 < 0},\ \mathbb{1}_{1=\min_{k\in M}\{|\widehat{S}_k(Y)|\}} \right\}. \tag{20}$$

Based on the expression of E_M given in Eq.(20), we introduce the function

$$b_\delta(x) = 1 - \sqrt{\left(\frac{x}{\delta}\right) \wedge 1}\,,$$



for some $\delta > 0$ and we define the energy

$$\mathcal{E}(Y) := \max\left\{p_1(Y)\ ,\ p_2(Y)\right\},$$

where

$$p_1(Y) := b_\delta\left(1 - \|\widehat{S}_{-M}(Y)\|_\infty\right) \quad \text{and} \quad p_2(Y) := \frac{1}{|M|}\sum_{k\in M}(1 - |\widehat{S}_k(Y)|).$$

The energy $\mathcal{E}$ measures how close some vector $Y \in \{0,1\}^N$ is to E_M. With Lemma 1, we make this claim rigorous by proving that for $\delta > 0$ small enough, the selection event E_M corresponds to the set of vectors $Y \in \{0,1\}^N$ satisfying $\mathcal{E}(Y) = 0$.

Lemma 1. *For any $M \subseteq [d]$, there exists $\delta_c := \delta_c(M, \mathbf{X}, \lambda) > 0$ such that for all $\delta \in (0, \delta_c)$, the selection event $E_M = \{Y \in \{0,1\}^N \mid \widehat{M}(Y) = M\}$ is equal to the set*

$$\left\{Y \in \{0,1\}^N \mid p_1(Y) = 0 \quad \textit{and} \quad p_2(Y) = 0\right\}.$$

Proof. Let us consider some $\delta \in (0, \delta_c)$ where

$$\delta_c := \min_{Y\in E_M}\{1 - \|\widehat{S}_{-M}(Y)\|_\infty\}.$$

Note that Eq.(20) ensures that for any $Y \in E_M$, $\|\widehat{S}_{-M}(Y)\|_\infty < 1$. This implies that $\delta_c > 0$ since the set E_M is finite.

It is obvious that for any $Y \in \{0,1\}^N$, the fact that $p_2(Y) = 0$ is equivalent to $\min_{k\in M}|\widehat{S}_k(Y)| = 1$. Moreover, thanks to our choice for the constant δ, it also holds that $p_1(Y) = 0$ is equivalent to $\|\widehat{S}_{-M}(Y)\|_\infty < 1$. The characterization of the selection event E_M given by Eq.(20) allows to conclude the proof. □

Lemma 1 states that-provided δ is small enough–the selection event E_M corresponds to the set of global minimizers of the energy $\mathcal{E} : \{0,1\}^N \to \mathbb{R}_+$. This characterization allows us to formulate a simulating annealing (SA) procedure in order to estimate E_M. Let us briefly recall that simulated annealing algorithms are used to estimate of set of global minimizers of a given function. At each time step, the algorithm considers some neighbour of the current state and probabilistically decides between moving to the proposed neighbour or staying

at its current location. While a transition to a state inducing a lower energy compared to the current one is always performed, the probability of transition towards a neighbour that leads to increase the energy is decreasing over time. The precise expression of the probability of transition is driven by a chosen *cooling schedule* $(\mathrm{T}_t)_t$ where T_t are called *temperatures* and vanish as $t \to \infty$. Intuitively, in the first iterations of the algorithm the temperature is high and we are likely to accept most of the transitions proposed by the SA. In that way, we give our algorithm the chance to escape from local minimum. As time goes along, the temperature decreases and we expect to end up at a global minima of the function of interest.

We refer to [Brémaud, 2013, Chapter 12] for further details on SA. Our method is described in Algorithm 1 and in the next section, we provide theoretical guarantees. In Algorithm 1, $P : \{0,1\}^N \times \{0,1\}^N \to [0,1]$ is the Markov transition kernel such that for any $Y \in \{0,1\}^N$, $P(Y, \cdot)$ is the probability measure on $\{0,1\}^N$ corresponding to the uniform distribution on the vectors in $\{0,1\}^N$ that differs from Y in exactly one coordinate.

Algorithm 1 SEI-SLR: Selection Event Identification for SLR

Data: $\mathbf{X}$, Y, λ, K_0, T

Compute $\hat{\vartheta}^\lambda \in \underset{\vartheta \in \mathbb{R}^d}{\arg\min}\{\mathcal{L}_N(\vartheta, (Y, \mathbf{X})) + \lambda\|\vartheta\|_1\}$
Set $M = \{k \in [d] \mid \hat{\vartheta}^\lambda_k \neq 0\}$
$Y^{(0)} \leftarrow Y$
for $t = 1$ to T **do**
  $Y^{\mathrm{c}} \sim P(Y^{(t-1)}, \cdot)$
  $\hat{\vartheta}^{\lambda,\mathrm{c}} \in \underset{\vartheta \in \mathbb{R}^d}{\arg\min}\{\mathcal{L}_N(\vartheta, (Y^{\mathrm{c}}, \mathbf{X})) + \lambda\|\vartheta\|_1\}$
  $\widehat{S}(Y^{\mathrm{c}}) = \frac{1}{\lambda}\mathbf{X}^\top(Y^{\mathrm{c}} - \sigma(\mathbf{X}\hat{\vartheta}^{\lambda,\mathrm{c}}))$
  $\Delta\mathcal{E} = \mathcal{E}(Y^{\mathrm{c}}) - \mathcal{E}(Y^{(t-1)})$
  $U \sim \mathcal{U}([0,1])$
  $\mathrm{T}_t \leftarrow \frac{K_0}{\log(t+1)}$
  if $\exp\left(-\frac{\Delta\mathcal{E}}{\mathrm{T}_t}\right) \geq U$ **then**
    $Y \leftarrow Y^{\mathrm{c}}$
  end if
end for

3.2 Proof of convergence of the algorithm

To provide theoretical guarantees on our methods in the upcoming sections, we need to understand what is the distribution of $Y^{(t)}$ as $t \to \infty$. This is the purpose of Proposition 4 which shows that the SEI-SLR algorithm generates states uniformly distributed on E_M in the asymptotic $t \to \infty$.

Proposition 4. *[Brémaud, 2013, Example 12.2.12]*
For a cooling schedule satisfying $\mathrm{T}_t \geq 2^{N+1}/\log(t+1)$*, the limiting distribution of the random vectors* $Y^{(t)}$ *is the uniform distribution on the selection event* E_M*.*

Proposition 4 has the important consequence that we are able to compute the distribution of the binary vector $Y = (y_i)_{i \in [N]}$ where each y_i is a Bernoulli random variable with parameter $\pi^*_i \in (0,1)^N$ conditionally on the selection event.

The formal presentation of this result is given by Proposition 5 which will be the cornerstone of our inference procedures presented in Section 4.

Proposition 5. *Let us consider $M \subseteq [d]$ and some $\vartheta^* \in \mathbb{R}^d$. We consider the random vector Y with distribution $\overline{\mathbb{P}}_{\pi^*}$ where $\pi^* = \sigma(\mathbf{X}\vartheta^*)$. For a cooling schedule satisfying $\mathrm{T}_t \geq 2^{N+1}/\log(t+1)$, it holds for any function $h : \{0,1\}^N \to \mathbb{R}$,*

$$\frac{\sum_{t=1}^{T} h(Y^{(t)})\mathbb{P}_{\pi^*}(Y^{(t)})}{\sum_{t=1}^{T} \mathbb{P}_{\pi^*}(Y^{(t)})} \underset{T\to\infty}{\to} \overline{\mathbb{E}}_{\pi^*}\left[h(Y)\right] \quad \textit{almost surely.}$$

Proof. Let us consider some map $h : \{0,1\}^N \to \mathbb{R}$. Then,

$$\overline{\mathbb{E}}_{\pi^*}\left[h(Y)\right] = \frac{\sum_{y\in E_M} h(y)\mathbb{P}_{\pi^*}(y)}{\sum_{y\in E_M} \mathbb{P}_{\pi^*}(y)} = \frac{\mathbb{E}(h(U_M)\mathbb{P}_{\pi^*}(Y = U_M))}{\mathbb{E}(\mathbb{P}_{\pi^*}(Y = U_M))},$$

where U_M is a random variable taking values in $\{0,1\}^N$ which is uniformly distributed over E_M. Then the conclusion directly follows from Proposition 4. □

4 Conditional Central Limit Theorems

4.1 Preliminaries

Before presenting our conditional CLTs, let us remind the framework in which we state our asymptotic results. Let $(d_N)_{N\in\mathbb{N}}$ be a non-decreasing sequence of positive integers converging to $d_\infty \in \mathbb{N} \cup \{+\infty\}$ and let $s \in [d_1, d_\infty] \cap \mathbb{N}$. For any N, we consider $[\vartheta^*]^{(N)} \in \mathbb{R}^{d_N}$, $\lambda^{(N)} > 0$, $M^{(N)} \subseteq [d_N]$ with cardinality s and a design matrix $\mathbf{X}^{(N)} \in \mathbb{R}^{N\times d_N}$. We recall the definitions of the selection event $E_M^{(N)}$ corresponding to the tuple $(\lambda^{(N)}, M^{(N)}, \mathbf{X}^{(N)})$ and of the conditional probability distribution $\overline{\mathbb{P}}_{\pi^*}^{(N)}$ given in Section 1.4. We assume that it holds

- $K := \sup_{N\in\mathbb{N}} \max_{i\in[N], j\in M^{(N)}} |\mathbf{X}_{i,j}^{(N)}| < \infty$,
- there exist constants $C, c > 0$ (independent of N) such that for any $N \in \mathbb{N}$,

$$cN \leq \lambda_{\min}(\left[\mathbf{X}_{M^{(N)}}^{(N)}\right]^\top \mathbf{X}_{M^{(N)}}^{(N)}) \leq \lambda_{\max}(\left[\mathbf{X}_{M^{(N)}}^{(N)}\right]^\top \mathbf{X}_{M^{(N)}}^{(N)}) \leq CN.$$

Remark. Note that the latter assumption holds in particular if the matrices $\left(\frac{\mathbf{X}^{(N)}}{\sqrt{N}}\right)_{N\in\mathbb{N}}$ satisfy (uniformly) the so-called s-Restricted Isometry Property (RIP) condition [cf. Wainwright, 2019, Definition 7.10]. Let us recall that a matrix $A \in \mathbb{R}^{N\times p}$ satisfies the s-RIP condition if there exists a constant $\delta_s \in (0,1)$ such that for any $N \times s$ submatrix A_s of A, it holds

$$1 - \delta_s \leq \lambda_{\min}(A_s^\top A_s) \leq \lambda_{\max}(A_s^\top A_s) \leq 1 + \delta_s.$$

In Section 4.2, we start by presenting our first CLT for $\left[\mathbf{X}_M^{(N)}\right]^\top Y$ where Y is distributed according to $\overline{\mathbb{P}}_{\pi^*}^{(N)}$. This will be the cornerstone of our PSI procedures holding for the saturated model and presented in Section 5.2. Thereafter, we prove in Section 4.3 a CLT for the conditional unpenalized MLE $\widehat{\theta}$ working with the design $\mathbf{X}_M^{(N)}$ (see Eq.(17)). This conditional CLT will be the key theoretical

ingredient to derive the PSI methods presented in Section 5.1 when considering the selected model.

The proofs of our conditional CLTs make use of [Bardet et al., 2008, Thm.1] and rely on triangular arrays $\vec{\xi} := \big((\xi_{i,N})_{i\in[N]},\, N \in \mathbb{N}\big)$ where $\xi_{i,N}$ is a random vector in $\mathbb{R}^s$ and is a function of the deterministic quantities $\lambda^{(N)}$, $\mathbf{X}^{(N)}$, $M^{(N)}$ and of the random variable Y with probability distribution $\overline{\mathbb{P}}^{(N)}_{\pi^*}$. Most dependent CLTs have been proven for causal time series (typically satisfying some mixing condition) and are not well-suited to our case since conditioning on the selection event introduces a complex dependence structure.

$$\begin{array}{ccccccc}
\xi_{1,1} & & & & & & \\
\xi_{1,2} & \xi_{2,2} & & & & & \\
\xi_{1,3} & \xi_{2,3} & \xi_{3,3} & & & & \\
\dots & \dots & \dots & \dots & & & \\
\xi_{1,N} & \xi_{2,N} & \xi_{3,N} & \dots & \xi_{N,N} & & \\
\dots & \dots & \dots & \dots & \dots & \dots &
\end{array}$$

The dependent Lindeberg CLT from [Bardet et al., 2008, Thm.1] gives us the opportunity to find conditions involving mainly the covariance matrix of Y under which our conditional CLTs hold. More precisely, we provide conditions ensuring that the lines of the $\mathbb{R}^s$-valued process indexed by a triangular system $\vec{\xi}$ satisfy some Lindeberg's condition. Let us stress that we discuss the assumptions of the theorems presented in Sections 4.2 and 4.3 in Section 4.4.

To alleviate this notational burden, we will not specify the dependence on N in the remainder of the paper, meaning that we will simply refer to $\mathbf{X}^{(N)}$, $M^{(N)}$, d_N, $[\vartheta^*]^{(N)}, \overline{\mathbb{P}}^{(N)}_{\pi^*}, \dots$ as $\mathbf{X}$, M, d, $\vartheta^*, \overline{\mathbb{P}}_{\pi^*}, \dots$. Nevertheless, let us stress again that the integer s is fixed and does not depend on N in our work.

4.2 A conditional CLT for the saturated model

We aim at providing a global testing procedure and a confidence interval for the parameter $\mathbf{X}_M^\top \pi^*$ conditionally on the selection event E_M. To do so, we prove in this section a CLT for $\mathbf{X}_M^\top Y$ when Y is a random variable on $\{0,1\}^N$ following the multivariate Bernoulli distribution with parameter $\pi^* \in [0,1]^N$ conditionally on the event $\{Y \in E_M\}$. Let us first recall the notation for the distribution of Y conditional on E_M in the saturated model

$$\overline{\mathbb{P}}_{\pi^*}(Y) \propto \mathbb{1}_{E_M}(Y)\mathbb{P}_{\pi^*}(Y),$$

where the symbol $\propto$ means "proportional to". In the following, we will denote by $\overline{\mathbb{E}}_{\pi^*}$ the expectation with respect to $\overline{\mathbb{P}}_{\pi^*}$. With Theorem 2, we give a conditional CLT that holds under some conditions that involve in particular the covariance matrix of the response Y under the distribution $\overline{\mathbb{P}}_{\pi^*}$, namely

$$\overline{\Gamma}^{\pi^*} := \overline{\mathbb{E}}_{\pi^*}\left[(Y - \overline{\pi}^{\pi^*})(Y - \overline{\pi}^{\pi^*})^\top\right] \in [-1,1]^{N\times N},$$

where $\overline{\pi}^{\pi^*} = \overline{\mathbb{E}}_{\pi^*}[Y]$.

Theorem 2. *We keep the notations and assumptions from Section 4.1. We denote* $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ *and* Y *the random vector taking values in* $\{0,1\}^N$ *and distributed according to* $\overline{\mathbb{P}}_{\pi^*}$*. Assume further that*

1. $\displaystyle\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2} \underset{N\to+\infty}{=} o(N),$

2. *there exists* $\overline{\sigma}^2_{\min} > 0$ *such that* $\overline{\pi}_i^{\pi^*}(1 - \overline{\pi}_i^{\pi^*}) \geq \overline{\sigma}^2_{\min}$ *for all* $i \in [N]$.

Then it holds

$$u^\top [\overline{G}_N(\pi^*)]^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}) \xrightarrow[N\to+\infty]{(d)} \mathcal{N}(0,1),$$

where u *is a unit* s*-vector and where* $\overline{G}_N(\pi^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\pi^*})^2)\mathbf{X}_M$ *with* $(\overline{\sigma}^{\pi^*})^2 := \overline{\pi}^{\pi^*} \odot (1 - \overline{\pi}^{\pi^*})$.

4.3 A conditional CLT for the selected model

We now work under the condition that there exists $\theta^* \in \mathbb{R}^s$ such that $\mathbf{X}_M\theta^* = \mathbf{X}\vartheta^*$. Given some $Y \in \{0,1\}^N$ and provided that $\mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi)$, $\Psi(\mathbf{X}_M^\top Y)$ is the MLE $\widehat{\theta}$ of the unpenalized logistic model. [Sur and Candès, 2019, Theorem 1] ensures that the MLE exists asymptotically almost surely.

We aim at providing a global testing procedure and a confidence interval for the parameter θ^* conditionally on the selection event. To do so, we first prove a CLT for the MLE $\widehat{\theta}$ when Y is distributed according to $\overline{\mathbb{P}}_{\theta^*}$ (*i.e.,* Y is a random variable on $\{0,1\}^N$ following the multivariate Bernoulli distribution with parameter $\sigma(\mathbf{X}_M\theta^*)$ conditioned on the event $\{Y \in E_M\}$). The unconditional MLE $\widehat{\theta}$ (using only the features indexed by M) is known to be consistent and asymptotically efficient meaning that when Y is distributed according to $\mathbb{P}_{\theta^*}$,

$$u^\top [G_N(\theta^*)]^{1/2} (\widehat{\theta} - \theta^*) \xrightarrow[N\to+\infty]{(d)} \mathcal{N}(0,1), \tag{21}$$

where u is a unit s-vector and where

$$G_N(\theta) := \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = \mathbf{X}_M^\top \mathrm{Diag}((\sigma^\theta)^2)\mathbf{X}_M,$$

is the Fisher information matrix with $(\sigma^\theta)^2 := \pi^\theta \odot (1 - \pi^\theta)$ with $\pi^\theta = \mathbb{E}_\theta[Y]$.

In the following, we will consider the natural counterpart of the Fisher information matrix $G_N(\theta^*)$ when we work under the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$,

$$\overline{G}_N(\theta^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\theta^*})^2)\mathbf{X}_M, \quad (\overline{\sigma}^{\theta^*})^2 := \overline{\pi}^{\theta^*} \odot (1 - \overline{\pi}^{\theta^*}),\ \overline{\pi}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}[Y].$$

Theorem 3 proves that the MLE $\widehat{\theta}$ under the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ also satisfies a CLT analogous to Eq.(21) by replacing respectively θ^* and $G_N(\theta^*)^{1/2}$ by $\overline{\theta}(\theta^*)$ (cf. Eq.(18)) and $[\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))$. This conditional CLT holds under some conditions that involve in particular the covariance matrix of the response Y under the distribution $\overline{\mathbb{P}}_{\theta^*}$, namely

$$\overline{\Gamma}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}\left[(Y - \overline{\pi}^{\theta^*})(Y - \overline{\pi}^{\theta^*})^\top\right] \in [-1,1]^{N\times N}.$$

Theorem 3. *We keep the notations and assumptions from Section 4.1. Let us consider $\theta^* \in \mathbb{R}^s$ and let us denote by Y the random vector taking values in $\{0,1\}^N$ and distributed according to $\overline{\mathbb{P}}_{\theta^*}$. Assume further that*

1. $\displaystyle\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\theta^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\theta^*}\right)^2} \underset{N\to+\infty}{=} o(N),$

2. *there exists $\overline{\sigma}^2_{\min} > 0$ such that for any N and for any $i \in [N]$,*

$$\overline{\pi}_i^{\theta^*}(1 - \overline{\pi}_i^{\theta^*}) \wedge \sigma'(\mathbf{X}_{i,M}\overline{\theta}(\theta^*)) \geq \overline{\sigma}^2_{\min}.$$

3. *there exists some $\mathfrak{K} > 0$ such that for any $N \in \mathbb{N}$,*

$$\mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right] < \mathfrak{K}.$$

Then,

$$u^\top [\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))\big(\widehat{\theta} - \overline{\theta}(\theta^*)\big) \xrightarrow[N\to+\infty]{(d)} \mathcal{N}(0,1),$$

where u is a unit s-vector and where we recall that $\widehat{\theta} = \Psi(\mathbf{X}_M^\top Y)$ is the MLE.

The proof of Theorem 3 can be found with full details in Section A.5 and we only provide here the main arguments. First we use Theorem 2 that shows that the distribution of $[\overline{G}_N(\theta^*)]^{-1/2} L_N(\overline{\theta}, Z^M)$ is asymptotically Gaussian using a Lindeberg Central Limit Theorem for dependent random variables from Bardet et al. [2008]. Then, we show that the MLE $\widehat{\theta}$ exists almost surely asymptotically and is almost surely contained within an ellipsoid centered at $\overline{\theta}$ with vanishing volume. This kind of result has already been studied in Liang and Du [2012] but the proof provided by Liang and Du is wrong (Eq.(3.7) is in particular not true). As far as we know, we are the first to provide a correction of this proof in Section A.5. Let us also stress that working with the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ brings extra-technicalities that need to be handled carefully.
Using this consistency of $\widehat{\theta}$ together with the smoothness of the map $\theta \mapsto L_N(\theta, Z^M)$, one can convert the previously established result for $[\overline{G}_N(\theta^*)]^{-1/2} L_N(\overline{\theta}, Z^M) = [\overline{G}_N(\theta^*)]^{-1/2}(L_N(\overline{\theta}, Z^M) - L_N(\widehat{\theta}, Z^M))$ into a CLT for $\widehat{\theta}$.

4.4 Discussion

In this section, we discuss *informally* the assumptions of both Theorems 2 and 3. The conditions of Theorems 2 and 3 can be seen at first glance as arcane or restrictive. Without pretending that those conditions are easy to check in practice, looking at these requirements through the lens of the usual asymptotic alternative where ϑ^* itself depends on N gives a different perspective. Such assumption on ϑ^* has been considered for example in Bunea [2008] or [Taylor and Tibshirani, 2018, Section 3.1]. Following this line of work, we consider that $\vartheta^* = \alpha_N^{-1}\beta^*$ where each entry of β^* is independent of N and $(\alpha_N)_N$ is a sequence of increasing positive numbers such that $\alpha_N \underset{N\to\infty}{\to} +\infty$. We further assume β^* is s^*-sparse with support M^* (and with s^* independent of N). Let us analyze the conditions of our theorems in this framework by considering that $E_M = \{0,1\}^N$ (i.e. there is no conditioning). Then, condition 3 of Theorem 3 holds automatically since

in this case $\mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M = G_N(\theta^*)$ and $\overline{G}_N^{-1} = [G_N(\theta^*)]^{-1}$, meaning that $\mathfrak{K} = s$ works. The condition 2 of Theorems 2 and 3 holds also automatically since $\alpha_N \underset{N\to\infty}{\to} +\infty$, while the condition 1 is satisfied as soon as $\alpha_N \underset{N\to\infty}{=} \omega(N^{1/2})$.

The quantity α_N is quantifying the dependence arising from conditioning on the selection event: the weaker the dependence between the entries of the random response $Y \sim \overline{\mathbb{P}}_{\pi^*}$, the smaller α_N can be chosen while preserving the asymptotic normal distribution. Note that in the papers Bunea [2008] and [Taylor and Tibshirani, 2018, Section 3.1], the authors typically consider the case where $\alpha_N \underset{N\to\infty}{\sim} N^{1/2}$, corresponding to the regime at which the validity of our CLTs may be questioned based on the simple analysis previously conducted. Nevertheless, we stress that stronger assumptions on the design could allow to bypass this apparent limitation. A promising line of investigation is the following: taking a closer at the proofs of Theorems 2 and 3, one can notice that the condition 1 can actually be weakened by

$$\min_{\nu\in\mathfrak{S}_N} \sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{\nu([i-1]),M})^\top \overline{\Gamma}^{\pi^*}_{\nu([i-1]),\nu([i-1])} \mathbf{X}_{\nu([i-1]),M}\|_F \left(1 - 2\overline{\pi}^{\pi^*}_{\nu(i)}\right)^2} \underset{N\to+\infty}{=} o(N),$$

where $\mathfrak{S}_N$ is the set of permutations of $[N]$.

5 Selective inference

5.1 In the selected model

5.1.1 Global testing procedure

We keep the notations and the assumptions of Theorem 3. Given some $\theta_0^* \in \mathbb{R}^s$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\theta^* = \theta_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\theta^* \neq \theta_0^*\}. \tag{22}$$

The CLT from Theorem 3 naturally leads us to introduce the ellipsoid W_N given by

$$W_N := \left\{ Y \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| [\overline{G}_N(\theta_0^*)]^{-1/2} G_N(\overline{\theta}(\theta_0^*)) \left(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta_0^*)\right) \right\|_2^2 > \chi^2_{s,1-\alpha} \end{array} \right\},$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. If $\overline{\pi}^{\theta_0^*}$ was known, we could compute $\overline{\theta}(\theta_0^*)$ (using Eq.(19)) and thus $\overline{G}_N(\theta_0^*)$. Then the test with rejection region W_N would be asymptotically of level α since Theorem 3 gives that

$$\overline{\mathbb{P}}_{\theta_0^*}\left(Y \in W_N\right) \underset{N\to+\infty}{\longrightarrow} \alpha.$$

Based on this result, we construct an asymptotically valid global testing procedure for the test (22). Our method consists in finding an estimate of the parameter $\overline{\pi}^{\theta_0^*}$ in order to approximate the rejection region W_N with a Monte-Carlo approach. From Proposition 4, we know that under an appropriate cooling scheme, the asymptotic distribution of the states visited by our SEI-SLR algorithm (cf. Algorithm 1) is the uniform distribution on the selection event.

We deduce that under the null, we are able to estimate $\overline{\pi}^{\theta^*}$ and thus $\overline{\theta}$ using Eq.(19). This leads to the testing procedure presented in Proposition 6, whose proof is postponed to Section A.7.

Proposition 6. *We keep notations and assumptions of Theorem 3. We consider two independent sequences of vectors $(Y^{(t)})_{t\geq 1}$ and $(Z^{(t)})_{t\geq 1}$ generated by Algorithm 1. Let us denote*

$$\widetilde{\pi}^{\theta_0^*} = \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})}, \quad \widetilde{\theta} = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_0^*}), \quad \widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\theta_0^*}(1-\widetilde{\pi}^{\theta_0^*})\right)\mathbf{X}_M,$$

$$\text{and } W_N := \left\{ Y \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Y \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta} \right) \right\|_2^2 > \chi^2_{s,1-\alpha} \end{array} \right\}.$$

Then the procedure consisting in rejecting the null hypothesis $\mathbb{H}_0$ when

$$\zeta_{N,T} := \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)} \in \widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})} > \alpha,$$

has an asymptotic level lower than α in the sense that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$ it holds,

$$\mathbb{P}\Big(\bigcup_{T_N \in \mathbb{N}} \bigcap_{T \geq T_N} \{\zeta_{N,T} \leq \alpha + \epsilon\} \Big) = 1.$$

5.1.2 Asymptotic confidence region

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 θ^*. Two questions arises at this point.

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

Proposition 7 answers both questions. It provides a valid confidence region with asymptotic level $1-\alpha$ for any estimate $\theta^\bigstar$ of θ^* where the width of the confidence region is asymptotically driven by $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. The proof of Proposition 7 can be found in Section A.8.

Proposition 7. *We keep notations and assumptions of Theorem 3 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 θ^*. 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 7, 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 7 involves two constants (namely κ and $\sigma^{\overline{\theta}}$) that cannot be *a priori* easily computed in practice.
Proposition 7 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 Section 6, we propose a deep learning and a gradient descent approach for our numerical experiments.

5.2 In the saturated model

5.2.1 Global testing procedure

We keep the notations and the assumptions of Theorem 2. Given some $\pi_0^* \in \mathbb{R}^N$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\pi^* = \pi_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\pi^* \neq \pi_0^*\}. \tag{23}$$

The CLT from Theorem 2 naturally leads us to introduce the ellipsoid W_N given by

$$W_N = \left\{ Y \in \{0,1\}^N \mid \left\| [\overline{G}_N(\pi_0^*)]^{-1/2} \mathbf{X}_M^\top \left(Y - \overline{\pi}^{\pi_0^*} \right) \right\|_2^2 \geq \chi^2_{s,1-\alpha} \right\},$$

where $\chi^2_{s,1-\alpha}$ is the quantile of order $1-\alpha$ of the χ^2 distribution with s degrees of freedom. If $\overline{\pi}^{\pi_0^*}$ was known, we could compute $\overline{G}_N(\pi_0^*)$. Then the test with rejection region W_N would be asymptotically of level α since Theorem 2 gives that

$$\overline{\mathbb{P}}_{\pi_0^*} \left(Y \in W_N \right) \underset{N \to +\infty}{\longrightarrow} \alpha.$$

Based on this result, we construct an asymptotically valid global testing procedure for the test (23). Our method consists in finding an estimate of the parameter $\overline{\pi}^{\pi_0^*}$ in order to approximate the rejection region W_N with a Monte-Carlo approach. From Proposition 4, we know that under an appropriate cooling scheme, the asymptotic distribution of the states visited by the SEI-SLR algorithm (cf. Algorithm 1) is the uniform distribution on the selection event. We deduce that under the null, we are able to estimate $\overline{\pi}^{\pi_0^*}$ and thus $\overline{G}_N(\pi_0^*)$. This leads to the testing procedure presented in Proposition 8, whose proof is strictly analogous to the one of Proposition 6.

Proposition 8. *We keep notations and assumptions of Theorem 2. We consider two independent sequences of vectors $(Y^{(t)})_{t\geq 1}$ and $(Z^{(t)})_{t\geq 1}$ generated by Algorithm 1. Let us denote*

$$\widetilde{\pi}^{\pi_0^*} = \frac{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)})Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)})}, \qquad \widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\pi_0^*}(1-\widetilde{\pi}^{\pi_0^*})\right)\mathbf{X}_M,$$

and $\widetilde{W}_N := \left\{ Y \in \{0,1\}^N \mid \left\| \widetilde{G}_N^{-1/2}\mathbf{X}_M^\top \left(Y - \widetilde{\pi}^{\pi_0^*}\right)\right\|_2^2 > \chi^2_{s,1-\alpha} \right\}$. *Then the procedure consisting of rejecting the null hypothesis $\mathbb{H}_0$ when*

$$\zeta_{N,T} := \frac{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Z^{(t)})\mathbb{1}_{Z^{(t)}\in\widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Z^{(t)})} > \alpha,$$

has an asymptotic level lower than α in the sense that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$ it holds,

$$\mathbb{P}\Big(\bigcup_{T_N\in\mathbb{N}} \bigcap_{T\geq T_N} \{\zeta_{N,T} \leq \alpha + \epsilon\}\Big) = 1.$$

5.2.2 Asymptotic confidence region

With Theorem 2, 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 5.1.2, we propose here to build an asymptotic confidence region for $\mathbf{X}_M^\top \pi^*$. The proof of Proposition 9 is postponed to Section A.9.

Proposition 9. *We keep notations and assumptions of Theorem 2 and we consider $\alpha \in (0,1)$. We assume further that there exist $p \in [1,\infty]$, $\kappa, R > 0$ and $r \in (0, \frac{1}{2})$ such that*

$$\pi^* \in \mathbb{B}_{p,\infty}(R,r) \quad \text{and} \quad \forall \pi \in \mathbb{B}_{p,\infty}(R,r), \quad \lambda_{\min}(\overline{\Gamma}^{\pi}) \geq \kappa,$$

where $\mathbb{B}_{p,\infty}(R,r) := \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R) \cap \mathbb{B}_\infty(\frac{\mathbf{1}_N}{2}, r)$. Let us consider any estimator $\pi^\bigstar \in \mathbb{B}_{p,\infty}(R,r)$ of π^*. Then the probability of the event

$$\|\mathbf{X}_M^\top \pi^* - \mathbf{X}_M^\top \pi^\bigstar\|_2 \leq \left(\kappa^2 cN\right)^{-1/2} \|\mathbf{X}_M^\top \overline{\pi}^{\pi^\bigstar} - \mathbf{X}_M^\top Y\|_2 + \kappa^{-1}\sqrt{Cc^{-1}\chi^2_{s,1-\alpha}},$$

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

Remark. Analogously to Section 5.1.2, Proposition 9 motivates us to choose $\pi^\bigstar$ 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 5.1.2, one can rely for example on a deep learning or a gradient descent method in order to reach a local minimum $\pi^\bigstar$ for M.

6 Numerical results

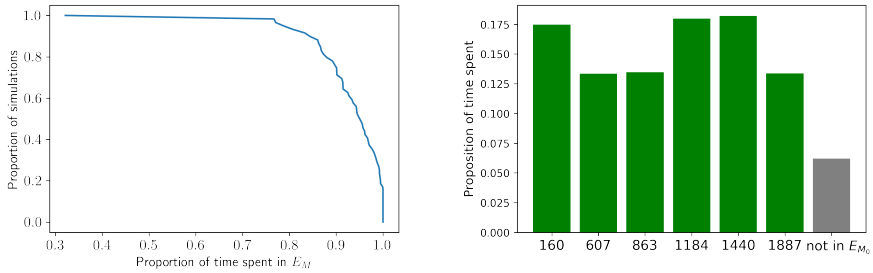
The code to reproduce our results is available at the following url: https://github.com/quentin-duchemin/LogPSI.

6.1 Sampling from the condition distribution

Description of the experiment. We test our approach under the global null, namely we consider $\vartheta^* = 0$. We work with $N = 11$, $p = 20$, $\lambda = 1.7$ and $\delta = 0.01$. The entries of the design matrix $\mathbf{X}$ are i.i.d. with a standard normal distribution. By choosing this toy example with a small value for N, we are able to compute exactly the selection event by running over the 2^N possible vector $Y \in \{0,1\}^N$. We start by sampling some vector $Y_0 \in \{0,1\}^N$ with i.i.d. entries with a Bernoulli distribution of parameter $1/2$. Note the couple $(\mathbf{X}, Y_0)$ determined the set M which corresponds to the set of indexes $i \in [d]$ such that $\hat{\vartheta}_i^\lambda \neq 0$ where $\hat{\vartheta}^\lambda$ is defined by Eq.(2). Given the equicorrelation set M, we run 40 simulated annealing paths of length $T = 150,000$ using the SEI-SLR algorithm (see Algorithm 1).

Uniform distribution on the selection event. In the following, we identify each vector $Y \in \{0,1\}^N$ with the number between 0 and $2^N - 1 = 2047$ that it represents in the base-2 numeral system. Using this identification, it holds on our example that $E_M = \{160, 607, 863, 1184, 1440, 1887\}$.

Figure 1.(a) shows the time spent in the selection event over the last $15,000$ time steps of the simulated annealing path for each of our 40 simulations. One can see that around 75% of the generated paths are spending at least 90% of their time in the selection event for the last $15,000$ time steps. Figure 1.(b) presents the proportion of time spent in the different states of the selection event and outside of E_M working again with the last $15,000$ visited states for each of the 40 simulations.



(a) Proportion of simulations (vertical axis) spending at least some x % of their time (horizontal axis) in the selection event.

(b) Time spent in each state of E_M and outside of E_M.

Figure 1: Visualization of the time spent in the selection event keeping the last $15,000$ visited states of each sequence provided by our algorithm SEI-SLR.

Figure 2 shows the last $15,000$ visited states for two specific simulated annealing paths. On the vertical axis, we have the integers encoded by all

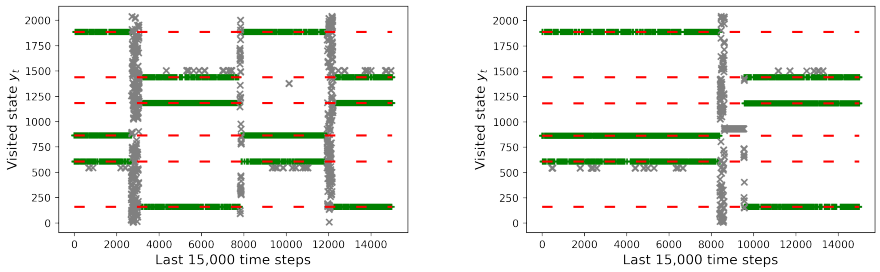


Figure 2: Last visited states of two simulated annealing paths. The dotted red lines indicate all the states belonging to the selection event. The gray (respectively green) crosses indicate visited states that do not belong (respectively that belong) to the selection event.

possible vectors $Y \in \{0,1\}^N$. The red dashed lines represent the states that belongs to the selection event E_M. While crosses are showing the visited states on the last $15,000$ time steps of the path, green crosses are emphasizing the ones that belong to the selection event. On this example, we see that the SEI-SLR algorithm covers properly the selection event without being stuck in one specific state of E_M. Each simulated annealing path is jumping from a state of E_M to another, ending up with an asymptotic distribution of the visited states that approximates the uniform distribution on E_M (see Figure 1.(b)). Let us point that two neighboring states in space $\{0,1\}^N$ will not necessarily be encoded by close integers.

Figure 2 suggests that the vectors encoded by the integers 160, 1184 and 1440 are close in the space $\{0,1\}^N$. Indeed, we see for example on the right plot of Figure 2 that in the last $5,000$ visited states, our algorithm goes from one of these three states to another passing through almost no state that does not belong to the selection event (this can be seen because there are only few gray crosses in the last $5,000$ iterations). The same remark holds for the three states encoded by the integers 607, 863 and 1887. However, we observe a large number of visited states that do not belong to E_M when we perform a transition from one of the state of the first group $\{160, 1184, 1440\}$ to one of the state of the second group $\{607, 863, 1887\}$. We can therefore legitimately think that the selection event separates into two groups of fairly distant states. This is confirmed by Figure 3 which presents the Hamming distances between the different vectors of E_M and reveals the existence of two clusters.

Comparison with the linear model. The previous theoretical and numerical results show that our approach allows to correctly identify the selection event E_M. Nevertheless, this method suffers from the curse of dimensionality since the random walks in the simulated annealings need to cover a state space of 2^N points. Let us mention that even in the linear model where the selection event E_M has the nice property to be a union of polyhedra, the method from Lee et al. [2016] to provide inference on a linear transformation of Y also copes with the curse of dimensionality. Indeed, the construction of confidence intervals conditionally on the event E_M requires the computation of 2^s intervals (while

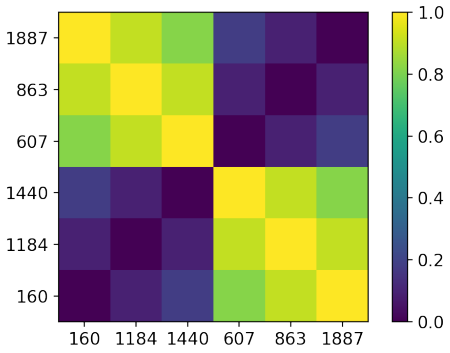


Figure 3: Normalized (by N) Hamming distances between the different states of the selection event. We observe two distinct clusters.

the computation of each of them requires at least N^3 operations) where $s = |M|$ (see [Lee et al., 2016, Section 6]). Roughly speaking, both our approach in the logistic model and the one from [Lee et al., 2016, Section 6] in the linear model are limited in large dimensions. While in the linear case, computational efficiency of the known methods mainly depends on $s = |M|$, the extra cost arising from the non-linearity of the logistic model is their dependence on N.

Let us finally mention that in the linear model, one can bypass the limitation of computing the 2^s intervals for each possible vector of dual signs on the equicorrelation set M by conditioning further on the observed vector of signs $\widehat{S}_M(Y) = \text{sign}(\widehat{\theta}^\lambda)_M$. Stated otherwise, instead of conditioning on E_M, we condition on $E_M^{S_M}$ where $S_M = \widehat{S}_M(Y)$. This method reduces the computational burden but it will lead in general to less powerful inference procedures due to some information loss which can be quantified through the so-called left-over Fisher information. In Section B, we discuss with further details PSI when we condition additionally on the observed vector of signs.

6.2 Hypothesis Testing

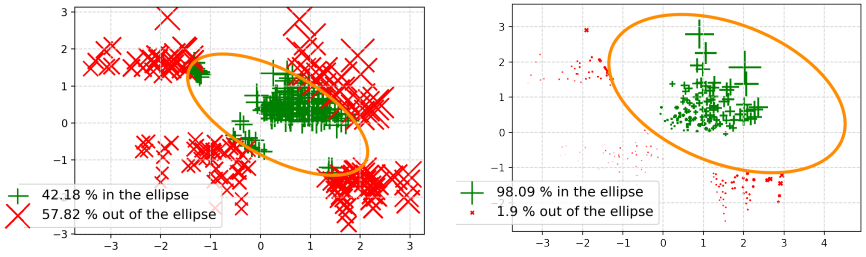
Description of the experiment. We consider $d = 40$, $N = 15$, $s = 2$, $\vartheta^* = [1, 1, 0, 0, \ldots, 0]$ and $\mathbf{X} \in \mathbb{R}^{N \times d}$ constructed as follows:

- We first sample $\widetilde{\mathbf{X}} \in \mathbb{R}^{N \times d}$ with independent standard normal entries.
- We normalize the first two columns of $\widetilde{\mathbf{X}}$ (i.e. we divide each entry of one column by the L^2-norm of this column) and the result gives the two first columns of $\mathbf{X}$.
- We project each column $j \in \{3, \ldots, d\}$ onto $\text{Span}(\mathbf{X}_1, \mathbf{X}_2)^\perp$. Thereafter, we normalize the resulting columns and we stack them to obtain the columns with index from 3 to d for the matrix $\mathbf{X}$.

Using this design matrix allows us to the It is well known that support recovery for the lasso requires such condition and by choosing a regularization parameter $\lambda =$

1.4, the selected support is equal to the true support $\widehat{M} = \{1, 2\}$. Working with $s = 2$, we can easily visualize the results of our global testing procedure. For any $\nu > 0$, we consider $\theta_0^* = [\nu, \nu]$ and we consider the hypothesis test (22).

Results. Figure 4 shows the way we compute the test statistic from Proposition 8: each visited state $Z^{(t)}$ of the SEI-SLR algorithm is plotted using a different color either $Z^{(t)} \in \widetilde{W}_N$ or $Z^{(t)} \notin \widetilde{W}_N$. Each sample is weighted proportionally to $\mathbb{P}_{\theta_0^*}(Z^{(t)})$ so that we reject the null if and only if the sum of weights of samples falling outside of the orange ellipse is larger than ν times the total mass of samples. In Figure 5, we show for ν ranging from 0 to 2 the total mass of samples falling into the ellipse $\left(\widetilde{W}_N\right)^c$. We see that our test controls that type I error at level $\alpha = 5\%$. Moreover, the test seems much more powerful when $\nu < 1$ compared to the case where $\nu > 1$.



(a) $\theta_0^* = [0.1, 0.1]$ (b) $\theta_0^* = [0.9, 0.9]$

Figure 4: The orange ellipse represents the set of parameter $\theta \in \mathbb{R}^s$ such that $\|\widetilde{G}_N^{-1/2} G_N(\widetilde{\theta})(\theta - \widetilde{\theta})\|_2^2 = \chi^2_{s,1-\alpha}$. For each t, we plot the MLE $\Psi(\mathbf{X}_M^\top Y^{(t)})$ with a green plus if the point falls into the orange ellipse and with a red cross otherwise. The size of the markers is proportional to $\overline{\mathbb{P}}_{\theta_0^*}(Y^{(t)})$.

Let us highlight that we trained a feed-forward neural network with three hidden layers to approximate $\Psi = \Xi^{-1}$ where we recall the $\Xi(\theta) = \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta)$ for any $\theta \in \mathbb{R}^s$.

6.3 Confidence region

As presented in Proposition 7 and the subsequent remark, the size of our confidence region is mainly driven by the distance $\|\overline{\theta}(\theta^\star) - \widehat{\theta}\|_2$. This encourages us to choose our estimate $\widehat{\theta}^*$ among the local minimizers of the function $m : \theta \mapsto \|\overline{\theta}(\theta) - \widehat{\theta}\|_2$. In the following, we propose a deep-learning and a gradient descent approach to achieve this goal.

6.3.1 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 $n_{train} = 500$

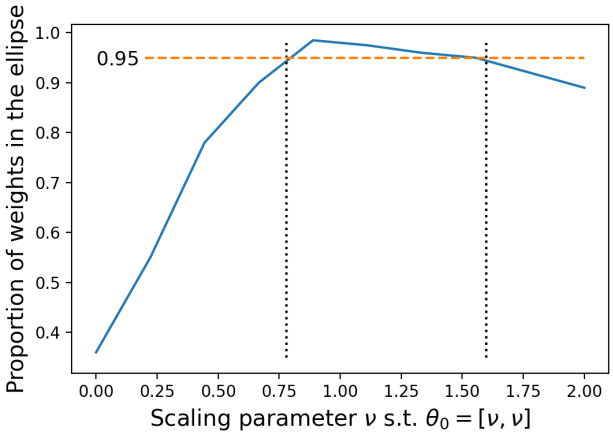


Figure 5: Value of the test statistic for $\theta_0^* = [\nu, \nu]$ with ν ranging from 0 to 2. The dashed vertical lines show the values of ν so that we reject the null at level 5%.

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 1). 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 θ^* by taking the output of our network when taking as input the unconditional MLE $\widehat{\theta}$ using the design $\mathbf{X}_M$ (cf. Eq.(17)). Figure 6 illustrates the result obtained from this deep learning approach. We keep the experiment settings of Section 6.2 namely, we consider $\vartheta^* = (1\ 1\ 0\ \ldots\ 0)^\top \in \mathbb{R}^d$ and we choose the regularization parameter λ so that the selected model corresponds to the true set of active variables, namely $M = \{1, 2\}$.

6.3.2 Gradient descent method

As shown in the proof of the expression of Proposition 7 (cf. Eq.(42)), it holds

$$\forall \theta \in \mathbb{R}^s, \quad \nabla_\theta \overline{\pi}^\theta = -\frac{1}{2} \overline{\Gamma}^\theta \mathbf{X}_M.$$

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

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

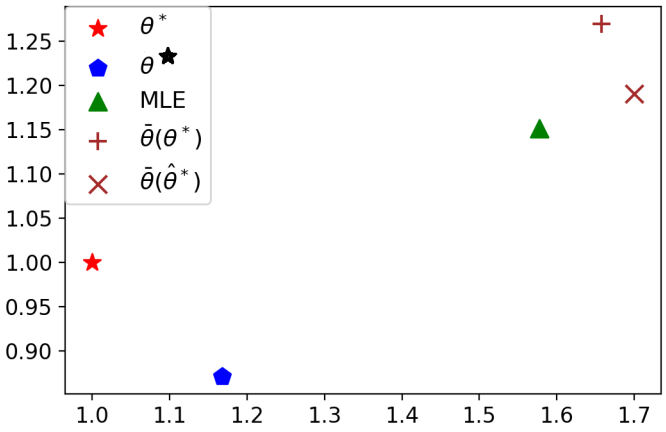


Figure 6: Visualization of the results obtained using our deep learning approach to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). $\theta^\bigstar$ corresponds to the output of the neural network when feeding as input the MLE $\widehat{\theta}$ (the green triangle). We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

Hence,

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

Given some θ, $\overline{\pi}^\theta$ 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 $G_N(\overline{\theta}(\theta))$).

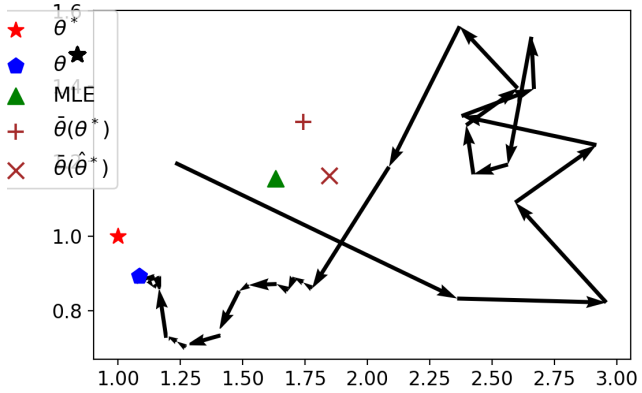


Figure 7: Visualization of our gradient descent procedure to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). We initialize the gradient descent to the MLE $\widehat{\theta}$ (the green triangle). We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

References

J.-M. Bardet, P. Doukhan, G. Lang, and N. Ragache. Dependent Lindeberg central limit theorem and some applications. *ESAIM: Probability and Statistics*, 12:154–172, 2008.

P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31. Springer Science & Business Media, 2013.

F. Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2(none):1153 – 1194, 2008. doi: 10.1214/08-EJS287. URL https://doi.org/10.1214/08-EJS287.

E. Candes and B. Recht. Simple bounds for recovering low-complexity models. *Mathematical Programming*, 141(1):577–589, 2013.

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

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.

H. Liang and P. Du. Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic Journal of Statistics*, 6: 1838–1846, 2012.

M. Massias, S. Vaiter, A. Gramfort, and J. Salmon. Dual extrapolation for sparse generalized linear models. *Journal of Machine Learning Research*, 21 (234):1–33, 2020.

A. Meir and M. Drton. Tractable Post-Selection Maximum Likelihood Inference for the Lasso. *arXiv: Methodology*, 2017.

R. T. Powers and E. Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1 – 33, 1970. doi: cmp/1103842028. URL https://doi.org/.

X. Shi, B. Liang, and Q. Zhang. Post-selection inference of generalized linear models based on the Lasso and the elastic net. *Communications in Statistics - Theory and Methods*, 0(0):1–18, 2020. doi: 10.1080/03610926.2020.1821892. URL https://doi.org/10.1080/03610926.2020.1821892.

P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

J. Taylor and R. Tibshirani. Post-selection inference for ℓ_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.

X. Tian and J. Taylor. Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499, 2017. doi: https://doi.org/10.1111/sjos.12261. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/sjos.12261.

R. J. Tibshirani, A. Rinaldo, R. Tibshirani, and L. Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287, 2018. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/26542824.

S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3): 230–287, 2015.

S. A. Van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

A Proofs

A.1 Proof of Proposition 1

Let us consider ϑ_1, ϑ_2 two vectors in Θ achieving the minimum in (2). Then, denoting $\vartheta_3 = \frac{1}{2}\vartheta_1 + \frac{1}{2}\vartheta_2$ it holds

$$\frac{\mathcal{L}_N(\vartheta_1, Z) + \mathcal{L}_N(\vartheta_2, Z)}{2} + \lambda \frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2} \leq \mathcal{L}_N(\vartheta_3, Z) + \lambda \|\vartheta_3\|_1.$$

Since the triangle inequality gives $\|\vartheta_3\|_1 \leq \frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2}$ and since the function ξ is strictly convex, it holds that $\mathbf{X}\vartheta_1 = \mathbf{X}\vartheta_2$. Indeed, otherwise we would have by strict convexity

$$\begin{aligned}
& \mathcal{L}_N(\vartheta_3, Z) + \lambda \|\vartheta_3\|_1 \\
= & \sum_{i=1}^{N} \left(\xi(\langle \mathbf{x}_i, \vartheta_3 \rangle) - \langle y_i \mathbf{x}_i, \vartheta_3 \rangle\right) + \lambda \|\vartheta_3\|_1 \\
\leq & \sum_{i=1}^{N} \left(\xi(\langle \mathbf{x}_i, \frac{\vartheta_1 + \vartheta_2}{2} \rangle) - \frac{1}{2}\langle y_i \mathbf{x}_i, \vartheta_1 \rangle - \frac{1}{2}\langle y_i \mathbf{x}_i, \vartheta_2 \rangle\right) + \frac{1}{2}\lambda \|\vartheta_1\|_1 + \frac{1}{2}\lambda \|\vartheta_2\|_1 \\
< & \frac{\mathcal{L}_N(\vartheta_1, Z) + \mathcal{L}_N(\vartheta_2, Z)}{2} + \lambda \frac{\|\vartheta_1\|_1 + \|\vartheta_2\|_1}{2}.
\end{aligned}$$

From the KKT conditions, we deduce that for a given $Y \in \mathcal{Y}^N$, all solutions $\hat{\vartheta}^\lambda$ of (2) have the same vector of signs denoted $\widehat{S}(Y)$ which is given

$$\widehat{S}(Y) = \frac{1}{\lambda} \mathbf{X}^\top \left(Y - \sigma(\mathbf{X}\hat{\vartheta}^\lambda)\right),$$

where $\hat{\vartheta}^\lambda$ is any solution to (2).

A.2 Proof of Proposition 2

Partitioning the KKT conditions of Eq.(3) according to the equicorrelation set $\widehat{M}(Y)$ leads to

$$\begin{aligned}
\mathbf{X}_{\widehat{M}(Y)}^\top \left(Y - \sigma(\mathbf{X}_{\widehat{M}(Y)} \hat{\vartheta}^\lambda_{\widehat{M}(Y)})\right) &= \lambda \widehat{S}_{\widehat{M}(Y)} \\
\mathbf{X}_{-\widehat{M}(Y)}^\top \left(Y - \sigma(\mathbf{X}_{\widehat{M}(Y)} \hat{\vartheta}^\lambda_{\widehat{M}(Y)})\right) &= \lambda \widehat{S}_{-\widehat{M}(Y)} \\
\operatorname{sign}(\hat{\vartheta}^\lambda_{\widehat{M}(Y)}) &= \widehat{S}_{\widehat{M}(Y)} \\
\|\widehat{S}_{-\widehat{M}(Y)}\|_\infty &< 1
\end{aligned}$$

Since the KKT conditions are necessary and sufficient for a solution, we obtain that Y belongs to $E_M^{S_M}$ if and only if there exist $\theta \in \Theta_M$ satisfying

$$\begin{aligned}
\mathbf{X}_M^\top \left(Y - \sigma(\mathbf{X}_M \theta)\right) &= \lambda S_M \\
\operatorname{sign}(\theta) &= S_M \\
\|\mathbf{X}_{-M}^\top \left(Y - \sigma(\mathbf{X}_M \theta)\right)\|_\infty &< \lambda
\end{aligned}$$

A.3 Proof of Proposition 3

Let us consider $\theta, \theta' \in \Theta_M$ such that $\Xi(\theta) = \Xi(\theta')$. Then we have

$$\begin{aligned}
0 &= \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta) - \mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta') \\
&= \Xi(\theta) - \Xi(\theta') \\
&= \int_0^1 \nabla \Xi(\theta t + (1-t)\theta') \cdot (\theta - \theta') dt \\
&= \int_0^1 \mathbf{X}_M^\top \mathrm{Diag}\left[\sigma'(\mathbf{X}_M \theta t + (1-t)\mathbf{X}_M \theta')\right] \mathbf{X}_M (\theta - \theta') dt \\
&= \mathbf{X}_M^\top \underbrace{\left(\int_0^1 \mathrm{Diag}\left[\sigma'(\mathbf{X}_M \theta t + (1-t)\mathbf{X}_M \theta')\right] dt \right)}_{=:D} \mathbf{X}_M (\theta - \theta'). \qquad (24)
\end{aligned}$$

Note that for any $t \in [0,1]$ and for any $i \in [N]$, $\{\sigma'(\mathbf{X}_M \theta t + (1-t)\mathbf{X}_M \theta')\}_i > 0$ since $\xi''(u) = \sigma'(u) > 0$ for any $u \in \mathbb{R}$. We deduce that $D \in \mathbb{R}^{N \times N}$ is a diagonal matrix with strictly positive coefficients on the diagonal. Eq.(24) gives that $\theta - \theta' \in \mathrm{Ker}(\mathbf{X}_M^\top D \mathbf{X}_M)$ which implies that $(\theta - \theta')^\top \mathbf{X}_M^\top D \mathbf{X}_M (\theta - \theta') = 0$. This means that

$$\sum_{i=1}^N D_{i,i} \left[\mathbf{X}_M (\theta - \theta')\right]_i^2 = 0.$$

Since $D_{i,i} > 0$ for all $i \in [N]$, we get that $\mathbf{X}_M(\theta - \theta') = 0$, i.e. $\mathbf{X}_M \theta = \mathbf{X}_M \theta'$. Since $\mathbf{X}_M$ has full column rank, this leads to $\theta = \theta'$.

Since Ξ is injective and of class $\mathcal{C}^m$ with a differential given by $\nabla_\theta \Xi(\theta) = \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta)) \mathbf{X}_M$ which is invertible at any $\theta \in \Theta_M$ under the assumptions of Proposition 3 Hence the global inversion theorem gives Proposition 3.

A.4 Proof of Theorem 2

For the sake of brevity, we will simply denote $\overline{G}_N(\pi^*)$ by $\overline{G}_N$. Let us further denote $\mathbf{X}_M^\top = [\mathbf{w}_1 \mid \mathbf{w}_2 \mid \ldots \mid \mathbf{w}_N]$, where $\mathbf{w}_i = \mathbf{x}_{i,M} \in \mathbb{R}^s$.

The proof of Theorem 2 relies on [Bardet et al., 2008, Theorem 1]. In the following, we check that all the assumptions of [Bardet et al., 2008, Theorem 1] are satisfied. Denoting for any $i \in [N]$, $\xi_{i,N} = \overline{G}_N^{-1/2} \mathbf{w}_i (y_i - \overline{\pi}_i^{\pi^*})$, it holds

$$\overline{G}_N^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}) = \sum_{i=1}^N \overline{G}_N^{-1/2} \mathbf{w}_i (y_i - \overline{\pi}_i^{\pi^*}) = \sum_{i=1}^N \xi_{i,N}.$$

Let us also point that $\overline{\mathbb{E}}_{\pi^*}[\xi_{i,N}] = 0$. In the following, we will simply refer to $\xi_{i,N}$ as ξ_i to ease the reading of the proof. Let us denote further

$$A_N = \sum_{i=1}^N \overline{\mathbb{E}}_{\pi^*} \left(\|\xi_i\|_2^3 \right).$$

One can notice that

$$\begin{aligned}
\overline{\mathbb{E}}_{\pi^*} \left(\|\xi_i\|_2^3 \right) &= \overline{\mathbb{E}}_{\pi^*} \left[(y_i - \overline{\pi}_i^{\pi^*})^3 \right] \|\overline{G}_N^{-1/2} \mathbf{w}_i\|_2^3 \\
&\leq \left(\frac{K}{\sqrt{c} \overline{\sigma}_{\min}} \right)^3 N^{-3/2} s^{3/2},
\end{aligned}$$

where we used that

$$\|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^2 \leq \|\overline{G}_N^{-1/2}\|^2 \times \|\mathbf{w}_i\|_2^2 \leq \|\overline{G}_N^{-1}\|(sK^2) \leq (c\overline{\sigma}_{\min}^2 N)^{-1}(sK^2).$$

We deduce that

$$A_N \leq \left(\frac{K}{\sqrt{c}\overline{\sigma}_{\min}}\right)^3 N^{-1/2}s^{3/2}.$$

Hence $A_N \underset{N\to\infty}{\to} 0$ which the first condition that needed to be checked to apply [Bardet et al., 2008, Theorem 1].

Let us now check the second condition from that Bardet et al. [2008] that consists in identifying the appropriate asymptotic covariance matrix.

$$\begin{aligned}
\sum_{i=1}^N \overline{Cov}_{\pi^*}(\xi_i) &= \sum_{i=1}^N \overline{\mathbb{E}}_{\pi^*}\left[\overline{G}_N^{-1/2}\mathbf{w}_i\mathbf{w}_i^\top\overline{G}_N^{-1/2}(y_i-\overline{\pi}_i^{\pi^*})^2\right] \\
&= \sum_{i=1}^N \overline{G}_N^{-1/2}\mathbf{w}_i \underbrace{\overline{\mathbb{E}}_{\pi^*}(y_i-\overline{\pi}_i^{\pi^*})^2}_{=(\overline{\sigma}_i^{\pi^*})^2}\mathbf{w}_i^\top\overline{G}_N^{-1/2} \\
&= \overline{G}_N^{-1/2}\sum_{i=1}^N \mathbf{w}_i(\overline{\sigma}_i^{\pi^*})^2\mathbf{w}_i^\top\overline{G}_N^{-1/2} \\
&= \overline{G}_N^{-1/2}\mathbf{X}_M^\top\mathrm{Diag}\big((\overline{\sigma}^{\pi^*})^2\big)\mathbf{X}_M\overline{G}_N^{-1/2} \\
&= \overline{G}_N^{-1/2}\overline{G}_N\overline{G}_N^{-1/2} \\
&= \mathrm{Id}_s.
\end{aligned}$$

To apply [Bardet et al., 2008, Theorem 1], it remains to check that the dependent Lindeberg conditions hold. For this, we consider some map $f \in \mathcal{C}_b^3(\mathbb{R}^s,\mathbb{R})$ where $\mathcal{C}_b^3(\mathbb{R}^s,\mathbb{R})$ is the set of functions from $\mathbb{R}^s$ to $\mathbb{R}$ with bounded and continuous partial derivatives up to order 3. In the following, we denote

$$W_i = \overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top(Y-\overline{\pi}^{\pi^*})_{[i-1]} = \sum_{a=1}^{i-1}\xi_a.$$

First dependent Lindeberg condition.

For any $i \in [N]$, let us consider W_i' (resp. ξ_i') an independent copy of the random vector W_i (resp. ξ_i). Let us recall the following well-known result

Lemma 2. *Let us consider two real valued random variables A, B on some probability space $(\Omega,\mathcal{F},\mathbb{P})$. Let us consider (A',B') an independent copy of the random vector (A,B). Then it holds,*

$$Cov(A,B) = \frac{1}{2}\mathbb{E}\big[(A-A')(B-B')\big].$$

Using Lemma 2, the Cauchy-Schwarz inequality and Jensen's inequalities, we get,

$$\sum_{k,l=1}^s\sum_{i=1}^N |\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l\partial x_k}(W_i),(\xi_i)_k(\xi_i)_l)|$$

$$
\begin{aligned}
&= \sum_{k,l=1}^{s}\sum_{i=1}^{N} |\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i), (\xi_i)_k(\xi_i)_l)| \\
&= \sum_{k,l=1}^{s}\sum_{i=1}^{N} \frac{1}{2}|\overline{\mathbb{E}}_{\pi^*}\left[\left(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i) - \frac{\partial^2 f}{\partial x_l \partial x_k}(W_i')\right)((\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l)\right]| \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N} \frac{1}{2}\|\nabla^3 f\|_\infty \overline{\mathbb{E}}_{\pi^*}\left(\|W_i - W_i'\|_2 \times |(\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l|\right) \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N} \frac{1}{2}\|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{E}}_{\pi^*}\left(\|W_i - W_i'\|_2^2\right)} \times \sqrt{\overline{\mathbb{E}}_{\pi^*}\left(|(\xi_i)_k(\xi_i)_l - (\xi_i')_k(\xi_i')_l|^2\right)} \\
&\leq \sum_{k,l=1}^{s}\sum_{i=1}^{N} \|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}\left(\|W_i\|_2\right)} \times \sqrt{\overline{\mathbb{V}ar}_{\pi^*}\left(|(\xi_i)_k(\xi_i)_l|\right)} \\
&\leq s\sum_{i=1}^{N} \|\nabla^3 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}\left(\|W_i\|_2\right)} \times \sqrt{\sum_{k,l=1}^{s} \overline{\mathbb{V}ar}_{\pi^*}\left(|(\xi_i)_k(\xi_i)_l|\right)},
\end{aligned}
$$

where in the last inequality we used Jensen's inequality. Let us upper-bound the terms $\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)$ and $\sum_{k,l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_k(\xi_i)_l|)$ independently. We have

$$
\begin{aligned}
&\overline{\mathbb{V}ar}_{\pi^*}\left(\|W_i\|_2\right) \\
&\leq \overline{\mathbb{E}}_{\pi^*}\left(\|W_i\|_2^2\right) \\
&= \overline{\mathbb{E}}_{\pi^*}\left[(Y - \overline{\pi}^{\pi^*})_{[i-1]}^\top \mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top (Y - \overline{\pi}^{\pi^*})_{[i-1]}\right] \\
&= \overline{\mathbb{E}}_{\pi^*}\left[\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top (Y - \overline{\pi}^{\pi^*})_{[i-1]}(Y - \overline{\pi}^{\pi^*})_{[i-1]}^\top \mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)\right] \\
&= \mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}_{[i-1],[i-1]}^{\pi^*}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
&\sum_{k,l=1}^{s} \overline{\mathbb{V}ar}_{\pi^*}\left(|(\xi_i)_k(\xi_i)_l|\right) \\
&= \sum_{k,l=1}^{s} ((\overline{G}_N^{-1/2})_{k,:}\mathbf{w}_i)^2((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 \left\{\overline{\mathbb{E}}_{\pi^*}\left[(y_i - \overline{\pi}_i^{\pi^*})^4\right] - \overline{\mathbb{E}}_{\pi^*}\left[(y_i - \overline{\pi}_i^{\pi^*})^2\right]^2\right\} \\
&= \sum_{k,l=1}^{s} ((\overline{G}_N^{-1/2})_{k,:}\mathbf{w}_i)^2((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 (\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2 \\
&= \|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^4 (\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2 \\
&\leq K^4(c\overline{\sigma}_{\min}^2)^{-2}\frac{s^2}{N^2}(\overline{\sigma}_i^{\pi^*})^2(1 - 2\overline{\pi}_i^{\pi^*})^2,
\end{aligned}
$$

where $(\overline{\sigma}_i^{\pi^*})^2 = \overline{\pi}_i^{\pi^*}(1 - \overline{\pi}_i^{\pi^*})$. Hence, coming back the first Lindeberg condition, we have (forgetting to mention the constants $K, s, c, \overline{\sigma}_{\min}^2$ that do not depend on

N, which is the sense of the symbol $\lesssim$),

$$
\begin{aligned}
&\sum_{k,l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial^2 f}{\partial x_l \partial x_k}(W_i),(\xi_i)_k(\xi_i)_l)| \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\|\overline{G}_N^{-1}\|_F\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{N}\sum_{i=1}^{N}\|\nabla^3 f\|_\infty \sqrt{\frac{1}{N}\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\theta^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\leq \frac{1}{N^{3/2}}\|\nabla^3 f\|_\infty \sum_{i=1}^{N}\sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F(1-2\overline{\pi}_i^{\pi^*})^2(\overline{\sigma}_i^{\pi^*})^2},
\end{aligned}
$$

where we used that $\|G_N^{-1}\|_F \leq \sqrt{s}\|G_N^{-1}\| \lesssim N^{-1}$ (since $\overline{G}_N^{-1}$ has rank s, see Section 4.1). Hence, the first dependent Lindeberg condition from Bardet et al. [2008] holds thanks to the assumptions made in Theorem 2.

Second dependent Lindeberg condition.
Using an approach analogous to the one conducted for the first dependent Lindeberg condition, one can obtain

$$
\begin{aligned}
&\sum_{l=1}^{s}\sum_{i=1}^{N}|\overline{Cov}_{\pi^*}(\frac{\partial f}{\partial x_l}(W_i),(\xi_i)_l)| \\
&\leq \sqrt{s}\sum_{i=1}^{N}\|\nabla^2 f\|_\infty \sqrt{\overline{\mathbb{V}ar}_{\pi^*}(\|W_i\|_2)} \times \sqrt{\sum_{l=1}^{s}\overline{\mathbb{V}ar}_{\pi^*}(|(\xi_i)_l|)} \\
&\lesssim \frac{1}{\sqrt{N}}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\mathrm{Tr}\left(\overline{G}_N^{-1/2}(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\overline{G}_N^{-1/2}\right)\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{\sqrt{N}}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\|\overline{G}_N^{-1}\|_F\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2} \\
&\lesssim \frac{1}{N}\|\nabla^2 f\|_\infty \sum_{i=1}^{N}\sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]}\mathbf{X}_{[i-1],M}\|_F\left(1-2\overline{\pi}_i^{\pi^*}\right)^2(\overline{\sigma}_i^{\pi^*})^2},
\end{aligned}
$$

where we used that

$$
\begin{aligned}
&\overline{\mathbb{V}ar}_{\pi^*}\left(|(\xi_i)_l|\right) \\
&= \overline{\mathbb{E}}_{\pi^*}\left(|(\xi_i)_l|^2\right) - \left(\overline{\mathbb{E}}_{\pi^*}|(\xi_i)_l|\right)^2 \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 \left\{ \overline{\mathbb{E}}_{\pi^*}\left((y_i - \overline{\pi}_i^{\pi^*})^2\right) - \left(\overline{\mathbb{E}}_{\pi^*}|y_i - \overline{\pi}_i^{\pi^*}|\right)^2 \right\} \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 \left\{ \overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*}) - \left(\overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*}) + (1-\overline{\pi}_i^{\pi^*})\overline{\pi}_i^{\pi^*}\right)^2 \right\} \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 \overline{\pi}_i^{\pi^*}(1-\overline{\pi}_i^{\pi^*}) \left(1 - 4(1-\overline{\pi}_i^{\pi^*})\overline{\pi}_i^{\pi^*}\right) \\
&= ((\overline{G}_N^{-1/2})_{l,:}\mathbf{w}_i)^2 (\overline{\sigma}_i^{\pi^*})^2 \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2 \\
&\lesssim \frac{1}{N}(\overline{\sigma}_i^{\pi^*})^2 \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2 .
\end{aligned}
$$

Assuming that

$$
\sum_{i=1}^{N} \sqrt{\|(\mathbf{X}_{[i-1],M})^\top \overline{\Gamma}^{\pi^*}_{[i-1],[i-1]} \mathbf{X}_{[i-1],M}\|_F \left(1 - 2\overline{\pi}_i^{\pi^*}\right)^2} \underset{N\to\infty}{=} o(N),
$$

we obtain applying [Bardet et al., 2008, Theorem 1] the following CLT

$$
\overline{G}_N^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}) \underset{N\to+\infty}{\overset{(d)}{\longrightarrow}} \mathcal{N}(0, \mathrm{Id}_s).
$$

A.5 Proof of Theorem 3

To make the notations less cluttered, we will simply denote in the following $\overline{G}_N(\theta^*)$ by $\overline{G}_N$ and $\overline{\theta}(\theta^*)$ by $\overline{\theta}$.

First step. We use Theorem 2 where we established a CLT for

$$
-L_N(\overline{\theta}, (Y, \mathbf{X}_M)) = \mathbf{X}_M^\top (Y - \pi^{\overline{\theta}}) = \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*}) = \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*}).
$$

Let us highlight that the first equality comes directly from the definition of $L_N(\overline{\theta}, (Y, \mathbf{X}_M))$ (see Section 2.2), the second equality comes from Eq.(19) and the last equality holds since we work under the selected model meaning that $\pi^* = \sigma(\mathbf{X}\vartheta^*) = \sigma(\mathbf{X}_M\theta^*)$ (and thus that $\overline{\mathbb{P}}_{\theta^*} \equiv \overline{\mathbb{P}}_{\pi^*}$). Let us recall that to prove Theorem 2, we used a variant of the Linderberg CLT for dependent random variables proved by Bardet et al. [2008]. The proof of Theorem 2 is given in Section A.4.

Second step. We now prove that for any $\epsilon > 0$ there is some $\delta > 0$ such that when N is large enough

$$
\overline{\mathbb{P}}_{\theta^*}\left(\text{there is } \widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta) \text{ such that } L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0\right) > 1 - \epsilon,
$$

with $\mathcal{N}_N(\overline{\theta}, \delta) = \{\theta \ : \ \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 \leq \delta\}$. Stated otherwise, we will prove that there exists a constant $\delta > 0$ and an integer $N_\delta \in \mathbb{N}$ such that for any $N \geq N_\delta$, the following holds with high probability,

- the conditional MLE $\widehat{\theta}$ exists,
- the conditional MLE $\widehat{\theta}$ is contained in the ellipsoid $\mathcal{N}_N(\overline{\theta}, \delta)$ centered at $\overline{\theta}$.

Let us denote

$$\begin{aligned} F : \theta \in \mathbb{R}^s \mapsto & \overline{G}_N^{-1/2}(L_N(\overline{\theta}, (Y, \mathbf{X}_M)) - L_N(\theta, (Y, \mathbf{X}_M))) \\ & = \overline{G}_N^{-1/2}\mathbf{X}_M^\top(\pi^{\overline{\theta}} - \pi^\theta). \end{aligned}$$

Note that F is a deterministic function and does not depend on the random variable Y. Moreover we choose to leave implicit the dependence on N of F. We also point out that it holds for any $\theta \in \mathbb{R}^s$,

$$\nabla_\theta F(\theta) = -\overline{G}_N^{-1/2}\mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = -\overline{G}_N^{-1/2}G_N(\theta).$$

Hence F is a $\mathcal{C}^1$ map with invertible Jacobian at any $\theta \in \mathbb{R}^s$ and is injective (thanks to Proposition 3). Applying the global inversion theorem, we deduce that F is a $\mathcal{C}^1$-diffeomorphism from $\mathbb{R}^s$ to $\mathbb{R}^s$.

Sketch of proof.
In the following, we prove that for any ϵ, we can choose $\delta > 0$ such that for some $N_\delta \in \mathbb{N}$ and for any $N \geq N_\delta$, it holds on some event E_N satisfying $\overline{\mathbb{P}}_{\theta^*}(E_N) \geq 1 - \epsilon$,

$$\begin{aligned} & \overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \in F(\mathcal{N}_N(\overline{\theta}, \delta)) \\ \Leftrightarrow \quad & \overline{G}_N^{-1/2}(\underbrace{\mathbf{X}_M^\top \overline{\pi}^{\theta^*}}_{=\mathbf{X}_M^\top \pi^{\overline{\theta}}} - \mathbf{X}_M^\top Y) \in F(\mathcal{N}_N(\overline{\theta}, \delta)). \end{aligned} \tag{25}$$

This would mean (by definition of F) that on E_N, there exists some $\widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta)$ such that $\overline{G}_N^{-1/2}L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$ or equivalently that $L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$. A sufficient condition for Eq.(25) to hold is to check that on the event E_N it holds

$$\|\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 < \inf_{\theta \in \partial\mathcal{N}_N(\overline{\theta}, \delta)} \|F(\theta)\|_2, \tag{26}$$

where $\partial\mathcal{N}_N(\overline{\theta}, \delta) := \{\theta \in \mathbb{R}^s \mid \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 = \delta\}$. This sufficient condition is a direct consequence of Lemma 3 and Figure 8 gives a visualization of our proof strategy.

Lemma 3. *Let $f : \mathbb{R}^s \to \mathbb{R}^s$ be a $\mathcal{C}^1$-diffeomorphism from $\mathbb{R}^s$ to $f(\mathbb{R}^s)$. Then for any closed space $D \subset \mathbb{R}^s$ it holds*

$$f(\partial D) = \partial f(D),$$

where for any set $U \subseteq \mathbb{R}^s$, $\partial U = \overline{U} \backslash \mathring{U}$ with $\overline{U}$ the closure of the set U and $\mathring{U}$ the interior of the set U.

Proof. As a C^1-diffeomorphism, f is in particular a homeomorphism, and as such, it preserves the topological structures. □

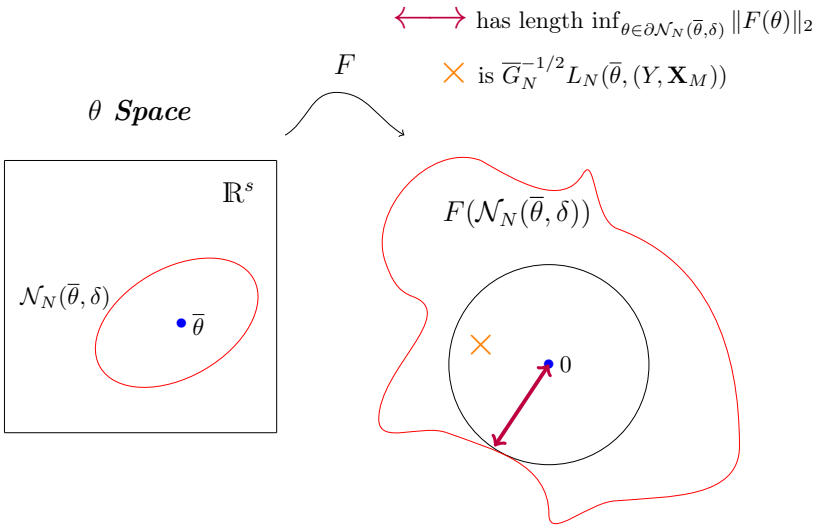


Figure 8: Visualization support for the proof of the existence of the MLE with large probability in a neighbourhood of $\overline{\theta}$. We show that with large probability, the orange cross is in the black circle (*i.e.,* Eq.(26) holds) which implies that the orange cross belongs to $F(\mathcal{N}_N(\overline{\theta},\delta))$ (*i.e.,* Eq.(25) holds). The MLE is then defined as $\widehat{\theta} = F^{-1}(\overline{G}_N^{-1/2}L_N(\overline{\theta},(Y,\mathbf{X}_M)) \in \mathcal{N}_N(\overline{\theta},\delta)$.

Let $\epsilon > 0$ and let us consider

$$\delta := \frac{\mathfrak{K}^{1/2}}{\epsilon^{1/2}2C^{-1}c\overline{\sigma}^2_{\min}}, \tag{27}$$

(the reason of this choice will become clear with Eq.(32)). Let us first notice that for any $\theta \in \mathbb{R}^s$,

$$\begin{aligned} &L_N(\overline{\theta},(Y,\mathbf{X}_M)) - L_N(\theta,(Y,\mathbf{X}_M)) && (28)\\ &= \mathbf{X}_M^\top(\pi^{\overline{\theta}} - \pi^\theta) && (29)\\ &= \underbrace{\int_0^1 G_N(t\overline{\theta}+(1-t)\theta)dt}_{=:Q_N(\theta)}\,(\overline{\theta}-\theta), && (30)\end{aligned}$$

where we used that the Jacobian of the map $\theta \mapsto \mathbf{X}_M^\top\pi^\theta = \mathbf{X}_M^\top\sigma(\mathbf{X}_M\theta)$ is $\mathbf{X}_M\mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = G_N(\theta)$. Recalling further that $\|\overline{G}_N^{-1/2}(\theta-\overline{\theta})\|_2 = \delta$

for any $\theta \in \partial\mathcal{N}_N(\overline{\theta}, \delta)$, it holds,

$$
\begin{aligned}
&\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|F(\theta)\|_2 \\
&= \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|\overline{G}_N^{-1/2} Q_N(\theta)(\theta-\overline{\theta})\|_2 \quad \text{(using Eq.(30))} \\
&= \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \|\overline{G}_N^{-1/2} Q_N(\theta)(\theta-\overline{\theta})\|_2 \times \frac{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \\
&\geq \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \frac{(\theta-\overline{\theta})^\top Q_N(\theta)(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \quad \text{(using the Cauchy Schwarz's inequality)} \\
&= \delta \inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \frac{(\theta-\overline{\theta})^\top \overline{G}_N^{1/2}}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \overline{G}_N^{-1/2} Q_N(\theta) \overline{G}_N^{-1/2} \frac{\overline{G}_N^{1/2}(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2} \\
&\geq \delta \inf_{\|e\|_2=1, \theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} Q_N(\theta) \overline{G}_N^{-1/2} e \\
&= \delta \inf_{\|e\|_2=1, \theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} \int_0^1 G_N(t\overline{\theta}+(1-t)\theta) dt \overline{G}_N^{-1/2} e \\
&= \delta \inf_{\|e\|_2=1, \theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)} \int_0^1 \left(e^\top \overline{G}_N^{-1/2} G_N(t\overline{\theta}+(1-t)\theta) \overline{G}_N^{-1/2} e \right) dt \\
&\geq \delta \inf_{\|e\|_2=1, \theta\in\mathcal{N}_N(\overline{\theta},\delta)} e^\top \overline{G}_N^{-1/2} G_N(\theta) \overline{G}_N^{-1/2} e \\
&\geq \delta \left\{ \inf_{\|e\|_2=1} e^\top \overline{G}_N^{-1/2} G_N(\overline{\theta}) \overline{G}_N^{-1/2} e - \mathcal{C} \frac{\delta}{N^{1/2}} \right\} =: \mathcal{I}_N(\delta, \overline{\theta}),
\end{aligned} \tag{31}
$$

where in the penultimate inequality we used that $\overline{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta)$ and the convexity of $\mathcal{N}_N(\overline{\theta}, \delta)$. In the last inequality, we used Lemma 4 whose proof is postponed to Section A.6.

Lemma 4. *Let us consider some $\delta > 0$. Then for any $N \in \mathbb{N}$ and for any unit vector $u \in \mathbb{R}^s$, it holds*

$$
\sup_{\theta\in\mathcal{N}_N(\overline{\theta},\delta)} |u^\top \overline{G}_N^{-1/2} (G_N(\theta) - G_N(\overline{\theta})) \overline{G}_N^{-1/2} u| \leq \mathcal{C} \frac{\delta}{N^{1/2}},
$$

where $\mathcal{N}_N(\overline{\theta}, \delta) = \{\theta \in \mathbb{R}^s \ : \ \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 \leq \delta\}$ and where $\mathcal{C}$ is a constant that only depends on the quantities $s, K, c, \overline{\sigma}^2_{\min}$ (that do not depend on N).

To lower bound uniformly in N the term $\mathcal{I}_N(\delta, \overline{\theta})$, we notice that

$$\begin{aligned}
&\inf_{\|e\|_2=1} e^\top \overline{G}_N^{-1/2} G_N(\overline{\theta}) \overline{G}_N^{-1/2} e \\
&= \inf_{\|e\|_2=1} \frac{e^\top \overline{G}_N^{-1/2}}{\|\overline{G}_N^{-1/2} e\|_2} G_N(\overline{\theta}) \frac{\overline{G}_N^{-1/2} e}{\|\overline{G}_N^{-1/2} e\|_2} \|\overline{G}_N^{-1/2} e\|_2^2 \\
&\geq \lambda_{\min}(G_N(\overline{\theta})) \inf_{\|e\|_2=1} \|\overline{G}_N^{-1/2} e\|_2^2 \\
&\geq \lambda_{\min}(G_N(\overline{\theta}))\, \lambda_{\min}(\overline{G}_N^{-1}) \\
&\geq \left(\overline{\sigma}_{\min}^2 cN\right) \times \left(4C^{-1}N^{-1}\right) \\
&\geq 4C^{-1} c \overline{\sigma}_{\min}^2,
\end{aligned}$$

where we used that for any $i \in [N]$, $\sigma'(\mathbf{x}_{i,M}\overline{\theta}) \geq \overline{\sigma}_{\min}^2$. Let us denote $N_\delta := \lceil \left(\frac{\mathcal{C}\delta}{2C^{-1}c\overline{\sigma}_{\min}^2}\right)^2 \rceil$ so that for any $N \geq N_\delta$ it holds

$$\mathcal{I}_N(\delta, \overline{\theta}) \geq \delta 2C^{-1} c \overline{\sigma}_{\min}^2.$$

Using Markov's inequality, we get that for any $N \geq N_\delta$,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 \geq \mathcal{I}_N(\delta, \overline{\theta})) \\
&\leq (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2^2) \\
&\leq (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}((Y - \overline{\pi}^{\theta^*})^\top \mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\mathrm{Tr}\left[(Y - \overline{\pi}^{\theta^*})^\top \mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})\right]) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \overline{\mathbb{E}}_{\theta^*}(\mathrm{Tr}\left[\mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top (Y - \overline{\pi}^{\theta^*})(Y - \overline{\pi}^{\theta^*})^\top\right]) \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \mathrm{Tr}\left[\mathbf{X}_M \overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*}\right] \\
&= (\mathcal{I}_N(\delta, \overline{\theta}))^{-2} \mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right].
\end{aligned}$$

Hence, it holds for any $N \geq N_\delta$,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 \geq \mathcal{I}_N(\delta, \overline{\theta})) \\
&\leq \frac{\mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right]}{\mathcal{I}_N(\delta, \overline{\theta})^2} \\
&< \frac{\mathfrak{K}}{\delta^2 (2C^{-1} c \overline{\sigma}_{\min}^2)^2} \\
&\leq \epsilon,
\end{aligned} \tag{32}$$

where the last inequality comes from the choice of δ (see Eq.(27)). From Eq.(31) and Eq.(32), we deduce that for any $N \geq N_\delta$, it holds

$$\overline{\mathbb{P}}_{\theta^*}(E_N) \geq 1 - \epsilon,$$

where

$$E_N := \left\{ \|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 < \inf_{\theta \in \partial \mathcal{N}_N(\overline{\theta}, \delta)} \|F(\theta)\|_2 \right\}.$$

Hence, on the event E_N, we define $\widehat{\theta} = F^{-1}(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)))$ which means by definition of F that $\widehat{\theta}$ is the conditional MLE, namely

$$L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0.$$

Third and final step. In the previous step, we proved that for N large enough, the MLE exists and is contained in an ellipsoid centered at $\overline{\theta}$ with vanishing volume with high probability. Now we show how using this result to turn the CLT on $L_N(\overline{\theta}, (Y, \mathbf{X}_M))$ from Theorem 2 into a CLT for $\widehat{\theta}$.

We consider $N \geq N_\delta$ and we work on the event E_N of the previous step. Since $L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) = 0$ by definition of $\widehat{\theta}$, we get that

$$\begin{aligned} L_N(\overline{\theta}, (Y, \mathbf{X}_M)) &= L_N(\overline{\theta}, (Y, \mathbf{X}_M)) - L_N(\widehat{\theta}, (Y, \mathbf{X}_M)) \\ &= \mathbf{X}_M^\top (\pi^{\overline{\theta}} - \pi^{\widehat{\theta}}) \\ &= \underbrace{\int_0^1 G_N(t\overline{\theta} + (1-t)\widehat{\theta}) dt}_{= Q_N(\widehat{\theta})} (\overline{\theta} - \widehat{\theta}), \end{aligned}$$

where we used that the Jacobian of the map $\theta \mapsto \mathbf{X}_M^\top \pi^\theta = \mathbf{X}_M \sigma(\mathbf{X}_M \theta)$ is $\mathbf{X}_M \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta)) \mathbf{X}_M = G_N(\theta)$. From the Portmanteau Theorem [cf. Van der Vaart, 2000, Lemma 2.2]), we know that a sequence of $\mathbb{R}^s$-valued random vectors $(X_n)_n$ converges weakly to a random vector X if and only if for any Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$ it holds

$$\mathbb{E} h(X_n) \underset{n \to \infty}{\to} \mathbb{E} h(X).$$

Hence, we consider a Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$. We denote by $L_h > 0$ the Lipschitz constant of h. It holds for any $N \geq N_\delta$,

$$\begin{aligned} &|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right]| \\ &= |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta}))]| \\ &\leq |\overline{\mathbb{E}}_{\theta^*}\left[\mathbb{1}_{E_N}\left\{h(\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta})) - h\big(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\big)\right\}\right]| + 2\|h\|_\infty \overline{\mathbb{P}}_{\theta^*}(E_N^c) \\ &\leq \overline{\mathbb{E}}_{\theta^*}\left[L_h \mathbb{1}_{E_N} \|\overline{G}_N^{-1/2} G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}) - \overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\|_2\right] + 2\|h\|_\infty \epsilon \\ &\leq L_h \overline{\mathbb{E}}_{\theta^*}\left[\mathbb{1}_{E_N} \|\overline{G}_N^{-1/2} (G_N(\overline{\theta}) - Q_N(\widehat{\theta})) \overline{G}_N^{-1/2}\| \|\overline{G}_N^{1/2}(\overline{\theta} - \widehat{\theta})\|_2\right] + 2\|h\|_\infty \epsilon \\ &\leq L_h \delta \sup_{\theta \in \mathcal{N}_N(\overline{\theta}, \delta)} \|\overline{G}_N^{-1/2} (G_N(\overline{\theta}) - Q_N(\theta)) \overline{G}_N^{-1/2}\| + 2\|h\|_\infty \epsilon, \end{aligned} \tag{33}$$

where we used that on the event E_N, $\widehat{\theta} \in \mathcal{N}_N(\overline{\theta}, \delta)$, i.e. $\|\overline{G}_N^{1/2}(\overline{\theta} - \widehat{\theta})\|_2 \leq \delta$. Moreover, for any $\theta' \in \mathcal{N}_N(\overline{\theta}, \delta)$ we have,

$$
\begin{aligned}
&\|\overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2}\| \\
&= \sup_{\|u\|_2=1} |u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2}u \\
&\leq \sup_{\|u\|_2=1} \int_0^1 \left|u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - G_N(t\overline{\theta} + (1-t)\theta'))\overline{G}_N^{-1/2}u\right| dt \\
&\leq \sup_{\|u\|_2=1} \sup_{\theta \in \mathcal{N}_N(\overline{\theta},\delta)} |u^\top \overline{G}_N^{-1/2}(G_N(\overline{\theta}) - G_N(\theta))\overline{G}_N^{-1/2}u| \\
&\leq \mathcal{C}\frac{\delta}{N^{1/2}}, \qquad (34)
\end{aligned}
$$

where in the penultimate inequality we used the convexity of the set $\mathcal{N}_N(\overline{\theta})$ and in the last inequality we used Lemma 4 (which is proved in Section A.6). Using Eq.(33) and Eq.(34), we deduce that for $G \sim \mathcal{N}(0, \mathrm{Id}_s)$ we have

$$
\begin{aligned}
&|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2}G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \mathbb{E}[h(G)]| \\
&\leq |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2}G_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right]| \\
&\quad + |\overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right] - \mathbb{E}[h(G)]| \\
&\leq L_h\delta\mathcal{C}\frac{\delta}{N^{1/2}} + 2\|h\|_\infty\epsilon + |\overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right] - \mathbb{E}[h(G)]|. \qquad (35)
\end{aligned}
$$

The CLT from Theorem 2 states that

$$
\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \xrightarrow[N\to\infty]{(d)} \mathcal{N}(0, \mathrm{Id}_s),
$$

which means by the Portmanteau Theorem [cf. Van der Vaart, 2000, Lemma 2.2]) that

$$
|\overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right] - \mathbb{E}[h(G)]| \underset{N\to+\infty}{\to} 0.
$$

We deduce that for any $\epsilon > 0$ and for any Lipschitz and bounded function $h : \mathbb{R}^s \to \mathbb{R}$, one can choose N large enough to ensure that the right hand side of Eq.(35) is smaller than $4\|h\|_\infty\epsilon$. Note that this is true since the constant δ does not depend on N. This concludes the proof thanks to the Portmanteau Theorem.

A.6 Proof of Lemma 4

Let us first recall that $G_N(\overline{\theta}) = \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\overline{\theta}))\mathbf{X}_M$ and that $\mathbf{X}_M^\top = [\mathbf{w}_1 \mid \mathbf{w}_2 \mid \ldots \mid \mathbf{w}_N]$, where $\mathbf{w}_i = \mathbf{x}_{i,M} \in \mathbb{R}^s$. Let us consider some $\theta \in \mathcal{N}_N(\overline{\theta}, \delta)$. We have that

$$
\begin{aligned}
G_N(\theta) - G_N(\overline{\theta}) &= \sum_{i=1}^N \mathbf{w}_i \left[\sigma'(\mathbf{w}_i^\top\theta) - \sigma'(\mathbf{w}_i^\top\overline{\theta})\right] \mathbf{w}_i^\top \\
&= \sum_{i=1}^N \mathbf{w}_i \underbrace{\int_0^1 \sigma''(t\mathbf{w}_i^\top\theta + (1-t)\mathbf{w}_i^\top\overline{\theta})dt}_{=:H_i} \mathbf{w}_i^\top(\theta - \overline{\theta})\mathbf{w}_i^\top. \qquad (36)
\end{aligned}
$$

We get using Eq.(36) that for any unit vector $u \in \mathbb{R}^s$,

$$
\begin{aligned}
&|u^\top \overline{G}_N^{-1/2}(G_N(\theta) - G_N(\overline{\theta}))\overline{G}_N^{-1/2}u| \\
&= \left|\sum_{i=1}^N u^\top \overline{G}_N^{-1/2}\mathbf{w}_i H_i \mathbf{w}_i^\top(\theta-\overline{\theta})\mathbf{w}_i^\top \overline{G}_N^{-1/2}u\right| \\
&= \left|\sum_{i=1}^N \mathbf{w}_i^\top(\theta-\overline{\theta}) \times u^\top \overline{G}_N^{-1/2}\mathbf{w}_i H_i \mathbf{w}_i^\top \overline{G}_N^{-1/2}u\right| \\
&= \left|\sum_{i=1}^N \mathbf{w}_i^\top(\theta-\overline{\theta}) \times H_i|\mathbf{w}_i^\top \overline{G}_N^{-1/2}u|^2\right| \\
&\leq \max_{1\leq j\leq N}|\mathbf{w}_j^\top(\theta-\overline{\theta})| \sum_{i=1}^N |H_i||\mathbf{w}_i^\top \overline{G}_N^{-1/2}u|^2 \\
&= \max_{1\leq j\leq N}|\mathbf{w}_j^\top(\theta-\overline{\theta})| \, \|\mathbf{H}^{1/2}\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2, \qquad (37)
\end{aligned}
$$

where $\mathbf{H}^{1/2} := \mathrm{Diag}((|H_i|^{1/2})_{i\in[N]})$. The proof is concluded by upper-bounding both terms involved in the product of the right hand side of Eq.(37). Using the assumption of the design matrix presented in Section 4.1 and recalling that $\theta \in \mathcal{N}_N(\overline{\theta}, \delta)$, we have

$$
\begin{aligned}
\max_{1\leq j\leq N}|\mathbf{w}_j^\top(\theta-\overline{\theta})| &\leq \max_{1\leq j\leq N}\|\overline{G}_N^{-1/2}\mathbf{w}_j\|_2 \underbrace{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}_{\leq \delta} \\
&= \delta K\sqrt{(\overline{\sigma}_{\min}^2 c)^{-1}s}N^{-1/2},
\end{aligned}
$$

where we used that $\|\overline{G}_N^{-1/2}\|^2 = \|\overline{G}_N^{-1}\| \leq (c\overline{\sigma}_{\min}^2 N)^{-1}$ and that for any $i \in [N]$, $\|\mathbf{w}_i\|_2^2 \leq sK^2$. Since $|H_i| \leq 1$ for any $i \in [N]$,

$$
\begin{aligned}
\|\mathbf{H}^{1/2}\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2 &\leq \|\mathbf{X}_M^\top \overline{G}_N^{-1/2}u\|_2^2 \\
&= \sum_{i=1}^N (\mathbf{w}_i^\top \overline{G}_N^{-1/2}u)^2 \\
&\leq \sum_{i=1}^N \|\overline{G}_N^{-1/2}\mathbf{w}_i\|_2^2 \leq (\overline{\sigma}_{\min}^2 c)^{-1}sK^2,
\end{aligned}
$$

where in the penultimate inequality we used Cauchy-Schwarz inequality.

A.7 Proof of Proposition 6

For any $N \in \mathbb{N}$, let us denote

$$
\mathcal{E}_N := \{Z \in \{0,1\}^N \,|\, \mathbf{X}_M^\top Z \in \mathrm{Im}(\Xi)\}. \qquad (38)
$$

In order to clarify the notations of this proof, let us stress that we denote in the following by $\overline{\mathbb{P}}_{\theta_0^*}$ the distribution of Y, $\mathbb{P}_1$ the distribution of the sequence $(Y^{(t)})_{t\geq 1}$ and $\mathbb{P}_2$ the distribution of $(Z^{(t)})_{t\geq 1}$. Let us consider some $\epsilon > 0$.

Step 1: $\mathbb{P}_1$ almost sure convergences.
From Proposition 5, we know that under the null $\mathbb{H}_0$

$$\frac{\sum_{t=1}^{T} Y^{(t)} \mathbb{P}_{\theta_0^*}(Y^{(t)})}{\sum_{t=1}^{T} \mathbb{P}_{\theta_0^*}(Y^{(t)})} \underset{T\to\infty}{\to} \overline{\mathbb{E}}_{\theta_0^*}[Y] = \overline{\pi}^{\theta_0^*} \quad \mathbb{P}_1 - \text{almost surely.} \tag{39}$$

Since $\widetilde{\pi}^{\theta_0^*} \underset{T\to\infty}{\to} \overline{\pi}^{\theta_0^*}$ $\mathbb{P}_1$-a.s., we know that $\mathbb{P}_1$-a.s, there exists some $T_1 \in \mathbb{N}$ such that for any $T \geq T_1$ it holds

$$\|\widetilde{\pi}^{\theta_0^*} \odot (1 - \widetilde{\pi}^{\theta_0^*}) - \overline{\pi}^{\theta_0^*} \odot (1 - \overline{\pi}^{\theta_0^*})\|_\infty < \epsilon,$$

and since $(\overline{\sigma}^{\theta_0^*})^2 \geq (\sigma_{\min})^2 > 0$, we get by continuity of the inverse of a matrix that $\mathbb{P}_1$-a.s, there exists some $T_2 \in \mathbb{N}$ such that for any $T \geq T_2$, it holds

$$\|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\| < \epsilon^2,$$

where we recall that

$$\widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}(\widetilde{\pi}^{\theta_0^*} \odot (1 - \widetilde{\pi}^{\theta_0^*}))\mathbf{X}_M,$$

and

$$\overline{G}_N = \mathbf{X}_M^\top \mathrm{Diag}(\overline{\pi}^{\theta_0^*} \odot (1 - \overline{\pi}^{\theta_0^*}))\mathbf{X}_M.$$

From Eq.(39) and by continuity of the map Ψ, we get that $\mathbb{P}_1$-a.s. $\widetilde{\theta} = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_0^*}) \underset{T\to\infty}{\to} \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}) = \overline{\theta}(\theta_0^*)$ (see Eq.(19)). Hence, $\mathbb{P}_1$-a.s, there exists some $T_3 \in \mathbb{N}$ such that for any $T \geq T_3$, it holds

$$\|\widetilde{\theta} - \overline{\theta}\|_2 \leq \epsilon.$$

Note that we left the dependence of $\widetilde{\pi}^{\theta_0^*}$ and $\widetilde{\theta}$ on T implicit.

Step 2: Comparing $\widetilde{W}_N$ and W_N.
It holds for any $Z \in \mathcal{E}_N$,

$$\begin{aligned}
&\left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\leq \left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\quad + \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\overline{\theta} - \widetilde{\theta} \right) \right\|_2 \\
&\leq \|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\| \|G_N(\widetilde{\theta})\| \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 \\
&\quad + \|\overline{G}_N^{-1/2}\| \|G_N(\widetilde{\theta}) - G_N(\overline{\theta})\| \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 + \left\| \mathbf{X}_M^\top \mathbf{X}_M \right\| \|\overline{\theta} - \widetilde{\theta}\|_2.
\end{aligned}$$

Using the Powers–Størmer inequality [cf. Powers and Størmer, 1970, Lemma 4.1] and denoting $\|M\|_1$ the Schatten 1-norm of any matrix M, it holds

$$\|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\|^2 \leq \|\widetilde{G}_N^{-1/2} - \overline{G}_N^{-1/2}\|_F^2 \leq \|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\|_1 \leq 2s\|\widetilde{G}_N^{-1} - \overline{G}_N^{-1}\|,$$

where in the last inequality we used that $\widetilde{G}_N$ and $\overline{G}_N$ have rank at most s. Hence, $\mathbb{P}_1$-a.s, for any $T \geq T_N(\epsilon) := \max(T_1, T_2, T_3)$ it holds

$$\begin{aligned}
&\left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right| \\
&\leq \left\| \Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right\|_2 \left\{ \epsilon 2sCN + (c(\overline{\sigma}_{\min})^2 N)^{-1/2} CN\epsilon \right\} + CN\epsilon =: \mathcal{C}_N(Z, \epsilon).
\end{aligned}$$

We get that $\mathbb{P}_1$-a.s, for any $T \geq T_N(\epsilon)$ it holds

$$\sup_{Z\in\mathcal{E}_N} \left| \left\| \widetilde{G}_N^{-1/2} G_N(\widetilde{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \widetilde{\theta} \right) \right\|_2 - \left\| \overline{G}_N^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2 \right|$$
$$\leq \sup_{Z\in\mathcal{E}_N} \mathcal{C}_N(Z,\epsilon) =: \mathcal{C}_N(\epsilon).$$

Step 3: Conclusion.
Let us consider some $\eta \in (0, 1-\alpha)$. Since $\mathcal{C}_N(\epsilon)$ goes to 0 as $\epsilon \to 0$, we deduce that we can choose ϵ small enough such that $\mathbb{P}_1$-a.s., for any $T \geq T_N(\epsilon)$ it holds

$$\forall Z \in \mathcal{E}_N, \quad \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)}, \tag{40}$$

where

$$W_N(\alpha+\eta) := \left\{ Z \in \{0,1\}^N \;\middle|\; \begin{array}{l} \diamond\ \mathbf{X}_M^\top Z \in \mathrm{Im}(\Xi) \\ \diamond\ \left\| [\overline{G}_N]^{-1/2} G_N(\overline{\theta}) \left(\Psi(\mathbf{X}_M^\top Z) - \overline{\theta} \right) \right\|_2^2 > \chi^2_{s,1-\alpha-\eta} \end{array} \right\},$$

Recalling the definition of $\mathcal{E}_N$ from Eq.(38) and using the definitions of $W_N(\alpha+\eta)$ and $\widetilde{W}_N$, it also holds trivially

$$\forall Z \in \{0,1\}^N \backslash \mathcal{E}_N, \quad 0 = \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)} = 0. \tag{41}$$

Using both Eq.(40) and Eq.(41), we deduce that

$$\forall Z \in \{0,1\}^N, \quad \mathbb{1}_{Z\in\widetilde{W}_N} \leq \mathbb{1}_{Z\in W_N(\alpha+\eta)},$$

and we then get that $\mathbb{P}_1$-a.s., for any $T \geq T_N(\epsilon)$, we have

$$\zeta_{N,T} = \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)}\in\widetilde{W}_N}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})} \leq \frac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)}) \mathbb{1}_{Z^{(t)}\in W_N(\alpha+\eta)}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Z^{(t)})}.$$

The right hand side of the previous inequality converges $\mathbb{P}_2$-a.s. to $\overline{\mathbb{P}}_{\theta_0^*}(Y \in W_N(\alpha+\eta))$ as $T \to +\infty$ thanks to Proposition 5. Since from Theorem 3 it holds,

$$\limsup_{N\to+\infty} \overline{\mathbb{P}}_{\theta_0^*}(Y \in W_N(\alpha+\eta)) \leq \alpha+\eta,$$

we get that for any $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any $N \geq N_0$ it holds,

$$\mathbb{P}\Big(\bigcup_{T_N\in\mathbb{N}} \bigcap_{T\geq T_N} \{\zeta_{N,T} \leq \alpha+\epsilon\} \Big) = 1.$$

A.8 Proof of Proposition 7

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 $k \in [s]$,

$$\begin{aligned}
\frac{\partial \overline{\pi}^\theta}{\partial \theta_k} &= \left(\sum_{w\in E_M} \mathbb{P}_\theta(w) \right)^{-2} \sum_{w,z\in E_M} \mathbb{P}_\theta(z)\mathbb{P}_\theta(w) z \left\{ L_N(\theta,(w,\mathbf{X}_M)) - L_N(\theta,(z,\mathbf{X}_M)) \right\}_k \\
&= \overline{\mathbb{E}}_\theta \left[Z \left\{ L_N(\theta,(W,\mathbf{X}_M)) - L_N(\theta,(Z,\mathbf{X}_M)) \right\}_k \right] \\
&= \overline{\mathbb{E}}_\theta \left[Z \left\{ \mathbf{X}_M^\top (Z-W) \right\}_k \right] \\
&= \overline{\Gamma}^\theta \mathbf{X}_{:,M[k]},
\end{aligned} \tag{42}$$

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 θ^*. 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

$$\begin{aligned}
\|\theta^* - \theta^\bigstar\|_2 &\leq (\kappa cN)^{-1} \|\mathbf{X}_M^\top\overline{\pi}^{\theta^\bigstar} - \mathbf{X}_M^\top\overline{\pi}^{\theta^*}\|_2 \\
&\leq (\kappa cN)^{-1} \sup_{\theta\in\mathbb{R}^s} \|\nabla\Psi^{-1}(\theta)\| \|\Psi\left(\mathbf{X}_M^\top\overline{\pi}^{\theta^\bigstar}\right) - \Psi\left(\mathbf{X}_M^\top\overline{\pi}^{\theta^*}\right)\|_2 \\
&\leq C\,(\kappa c)^{-1} \|\Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^\bigstar)}\right) - \Psi\left(\mathbf{X}_M^\top\pi^{\overline{\theta}(\theta^*)}\right)\|_2 \quad \text{(using Eq.(19))} \\
&= C\,(\kappa c)^{-1} \|\overline{\theta}(\theta^\bigstar) - \overline{\theta}(\theta^*)\|_2 \\
&\leq C\,(\kappa c)^{-1} \left[\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2 + \|\widehat{\theta} - \overline{\theta}(\theta^*)\|_2\right].
\end{aligned}$$

Since Theorem 3 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 $V_N(\theta^*) := [\overline{G}_N(\theta^*)]^{-1/2} G_N(\overline{\theta}(\theta^*))$, we deduce (using the assumption of the design matrix from Section 4.1) that the event

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

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

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

and that

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

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

A.9 Proof of Proposition 9

Let us denote $\mathcal{R} : \pi \in (0,1)^N \mapsto \mathbf{X}_M^\top \overline{\pi}^\pi$. 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

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

Suppose that we are able to compute an estimate $\pi^\bigstar \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R) \cap \mathbb{B}_\infty(\frac{\mathbf{1}_N}{2}, r)$ of π^*. Using that for some constant $c > 0$, it holds for any $v \in \mathbb{R}^N$,

$$\inf_{\pi \in \mathbb{B}_p(\frac{\mathbf{1}_N}{2}, R) \cap \mathbb{B}_\infty(\frac{\mathbf{1}_N}{2}, r)} \|\nabla \mathcal{R}(\pi) v\|_2 \geq (cN)^{1/2} \kappa \|v\|_2,$$

we get

$$\begin{aligned}\|\mathcal{R}(\pi^\bigstar) - \mathcal{R}(\pi^*)\|_2 &= \Big\| \int_0^1 \nabla \mathcal{R}(t\pi^\bigstar + (1-t)\pi^*)(\pi^\bigstar - \pi^*) dt \Big\|_2 \\ &\geq \kappa (cN)^{1/2} \|\pi^\bigstar - \pi^*\|_2.\end{aligned}$$

Hence we have that

$$\begin{aligned}\|\pi^* - \pi^\bigstar\|_2 &\leq \left(\kappa^2 cN\right)^{-1/2} \|\mathbf{X}_M^\top \overline{\pi}^{\pi^\bigstar} - \mathbf{X}_M^\top \overline{\pi}^{\pi^*}\|_2 \\ &\leq \left(\kappa^2 cN\right)^{-1/2} \left(\|\mathbf{X}_M^\top (\overline{\pi}^{\pi^\bigstar} - Y)\|_2 + \|\mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*})\|_2\right).\end{aligned}$$

Since Theorem 2 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

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

holds with probability tending to $1-\alpha$ as $N \to +\infty$. Hence we obtain an asymptotic confidence region for $\mathbf{X}_M^\top \pi^*$ of level $1-\alpha$.

B Inference conditional on the signs

B.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)). 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 σ-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

$$\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, E_M) = \nabla^2_\vartheta \mathcal{L}_N(\vartheta, \widehat{S}_M(Y) \,|\, E_M) + \nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \{E_M, \widehat{S}_M(Y)\}). \tag{43}$$

For any σ-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 $\mathcal{I}_{Y\,|\,\{E_M, \widehat{S}_M(Y)\}}(\vartheta)$. Taking expectation in both sides of Eq.(43) leads to

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

which can also be written as

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

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 ϑ. 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

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

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

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

B.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 Section 2, in Taylor and Tibshirani [2018] the authors claim the following asymptotic distribution

$$\underline{\theta} \sim \mathcal{N}(\vartheta^*_M, G_N(\vartheta^*_M)^{-1}), \tag{44}$$

where $\underline{\theta} = \hat{\vartheta}^\lambda_M + \lambda G_N(\widehat{\vartheta}^\lambda_M)^{-1} \widehat{S}_M(Y)$. 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

$$\{Y \ : \ \mathrm{Diag}(S_M) \left(\underline{\theta} - G_N(\vartheta^*_M)^{-1} \lambda S_M\right) \geq 0\}.$$

Let us denote by $F^{[a,b]}_{\mu,\sigma^2}$ 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

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

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 ϑ^*_M through $G_N(\vartheta^*_M)$. Taylor and al. propose to use a Monte-Carlo estimate for $G_N(\vartheta^*_M)$ by replacing it with $G_N(\widehat{\theta}^\lambda)$. Using this Monte-Carlo estimate, one can compute L and U such that

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

   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.