SIGLE: valid Selective Inference procedure for Generalized Linear Lasso

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

April 2023

Abstract

This article investigates uncertainty quantification of the generalized linear lasso (GLL), a popular variable selection method in high-dimensional regression settings. In many fields of study, researchers use data-driven methods to select a subset of variables that are most likely to be associated with a response variable. However, such variable selection methods can introduce bias and increase the likelihood of false positives, leading to incorrect conclusions.

In this paper, we propose a post-selection inference framework that addresses these issues and allows for valid statistical inference after variable selection using GLL. We show that our method provides accurate p-values and confidence intervals, while maintaining high statistical power.

In a second stage, we focus on the sparse logistic regression, a popular classifier in high-dimensional statistics. We show with extensive numerical simulations that SIGLE is more powerful than state-of-the-art PSI methods. SIGLE relies on a new method to sample states from the distribution of observations conditional on the selection event. This method is based on a simulated annealing strategy whose energy is given by the first order conditions of the logistic lasso.

1 Introduction

In modern statistics, the number of predictors can far exceed the number of observations available. In this high-dimensional context, ℓ_1 regularisation leads to 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.

In this context, the application of standard inference methods without taking into account the use of data to select the model usually leads to undesirable statistical properties. Post-selection inference (PSI) is designed to address this issue. It consists of constructing inference procedures considering that the vector of observations Y is distributed according to the distribution conditional on the so-called selection event.

In the literature, the problem of post-selection inference has been studied mainly for linear regression with Gaussian noise and assuming that the model has been selected using LASSO. Leaving this specific framework is an essential step for applications and more challenges can be expected in the study of PSI procedures for a generalized linear model (GLM). Moreover, 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 frame, it becomes crucial to provide certifiable guarantees on the output of the SLR, e.g. confidence intervals.

Inference procedures with statistical guarantees in the Generalized Linear Model (GLM) are few, if any. The practitioner is often left with no valid option to quantifies the uncertainty of predictions in high-dimensional GLMs. To the best of our knowledge, she might use the recent work of Taylor and Tibshirani [2018] for inference with the Generalized Linear LASSO (GLL). Based on a heuristic argument, Taylor and Tibshirani [2018] quantifies the uncertainty of the solutions of GLL.

The main contribution of this article is three fold. First, we introduce **SIGLE** (Selective Inference for Generalized Linear Estimation), a new conditional MLE approach to provide testing procedures and confidence regions for the solutions of GLL. SIGLE relies one a new sampling scheme from the distribution of observations conditional on the selection event.

Second, we focus on the SLR and we introduce a new method to sample states according to the conditional distribution, allowing the use of SIGLE in this context. We empirically witness that SIGLE is more powerful than current state-of-the-art methods. On Figure 1, we observe that our testing procedure (SIGLE) is correctly calibrated and we compare its power with the method from Taylor and Tibshirani [2018] and with a *weak learner*. This weak learner is a two-sided test based on the statistic $\sum_{i=1}^{n} |\overline{\pi}_i^{\theta_0} - y_i|$ where $\overline{\pi}^{\theta_0}$ is the expectation of the vector of observations under the null conditional on the selection event (cf. Eq.(16)). More experiments can be found in Section 4.2.

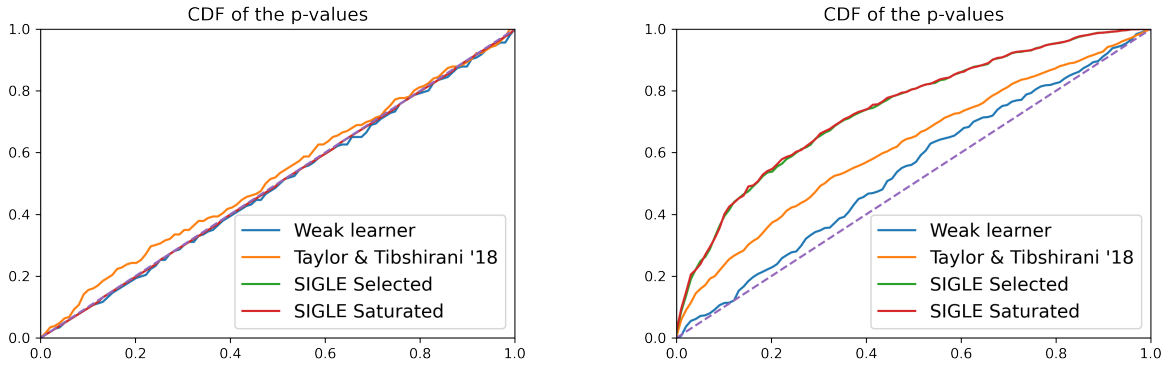
Last but not least, we prove a new conditional Central Limit Theorem (CLT) that exhibits conditions under which the SIGLE statistic is asymptotically normal. These assumptions hold under considerations similar to those commonly used in the study of asymptotic properties of subset selection via the Lasso in linear models (cf. Taylor and Tibshirani [2018], Bunea [2008]) and are not of particular interest for practical applications. This conditional CLT is a significant contribution and can be read at three levels of granularity. First it motivates the choice of the SIGLE statistic in this work. Second, it opens new perspective regarding the theoretical analysis of PSI methods in GLL. Indeed, while Taylor and Tibshirani [2018] focus first on getting unconditional asymptotic result before considering the distribution of the limit distribution conditional on the selection event, we directly consider the conditional distribution of the SIGLE statistic before analyzing its asymptotic limit. Let us stress out that the asymptotic result stated in Taylor and Tibshirani [2018] relies on non rigorous computations. Third, we believe that the proof of our conditional CLT might be of independent interest. In particular, we are–as far as we know–the first to correct the proof from Liang and Du [2012] which has been reported as false (cf. Zhang [2018]).

1.1 Post-Selection Inference for high-dimensional GLM

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

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

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 $\mathcal{C}^{m+1}$ (with m a non-negative integer). For sake of readability, the dependence on $\mathbf{X}$ will be omitted when it is clear from the context, and 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 $v = 1$,



(a) $\vartheta^* = [0, \dots, 0]$. (b) $\vartheta^* = [0.2, 0.2, 0, \dots, 0]$.

Figure 1: In the logistic model, we consider a design matrix $\mathbf{X} \in \mathbb{R}^{200\times 10}$ where we sample independently each entry with respect to a standard normal distribution. On Figures (a) and (b), we show the cumulative distribution function (CDF) of the p-values under obtained from i) a *weak learner*, ii) the procedure TT-1 (cf. Section 4.2) adapted from Taylor and Tibshirani [2018] and iii) SIGLE. On Figure (a) we work under the global null showing that SIGLE and the weak leaner are correctly calibrated. The method from Taylor and Tibshirani [2018] does not show p-values systematically larger than uniform. On Figure (b) we work under the alternative $\vartheta^* = [0.2, 0.2, 0, \dots] \in \mathbb{R}^{10}$.

$\xi(t) = \exp(t)$ and $\mathcal{Y} = \{0, 1, 2, \dots\}$ for the Poisson regression. Last but not least, we will consider in this paper the logistic regression where $v = 1$, $\xi(t) = \log(1 + \exp(t))$ and $\mathcal{Y} = \{0, 1\}$.

The negative log-likelihood takes the form

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

Assume that the partition function ξ is differentiable, then the score function is

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

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. Note that $\sigma(\mathbf{X}\vartheta)$ should be understood as applying entrywise the function σ to the vector $\mathbf{X}\vartheta$. 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 (GLL)

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

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

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

Given any $Y \in \mathcal{Y}^N$ and $\lambda > 0$, Proposition 1 shows that there exists one and 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 E.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) with Y in Eq.(3a)), 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 on 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 A useful 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 E.2.

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

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

where $\mathbf{X}_M \in \mathbb{R}^{N \times s}$ (resp. $\mathbf{X}_{-M} \in \mathbb{R}^{N \times (d-s)}$) 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 E.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* $\mathrm{Im}(\Xi) = \{\mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta) \mid \theta \in \Theta_M\}$.

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*

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

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 M has been selected, two different modeling assumptions are generally considered when we derive post-selection inference procedures, see for instance [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

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

as the parameter of interest. To support this choice, note that the Bayes predictor in the logistic or the linear model is defined from $\mathbb{E}_{\vartheta^*}[Y]$.

As presented in Fithian et al. [2014], the analyst should decide whether the model M belongs to the so-called class of *saturated models* or *selected models*. In the following, we discuss these concepts for arbitrary GLMs and Table 1 summarizes the key concepts.

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

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

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.

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.

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.

1.4 Inference procedures with SIGLE

In this section, we show how the SIGLE test statistics naturally emerge by establishing a parallel between post selection inference and M-estimation with a misspecified model.

SIGLE statistic in the selected model. In the post-selection paradigm, we work conditional to the selection event $\{Y \in E_M\}$. The conditional distribution of the observations is a conditional exponential

family with the same parameters and sufficient statistics but different support and normalizing constant:

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

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. $\mathbb{P}_\theta$). In the selected model, we want to conduct inference on θ^* (from Eq.(11)) based on the conditional and unpenalized MLE computed on the selected model M, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta_M} \mathcal{L}_N(\theta, Z^M), \qquad \mathcal{L}_N(\theta, Z^M) = \sum_{i=1}^{n} \left\{ \xi(\mathbf{X}_{i,M}\theta) - y_i \mathbf{X}_{i,M}\theta \right\}, \tag{12}$$

where $Z^M = (Y, \mathbf{X}_M)$ and where Y is distributed according to $\overline{\mathbb{P}}_{\theta^*}$. Eq.(12) can be understood as a mean-field approximation of the true likelihood where we make the assumption that the Y_i's are independent conditional to the selection event $\{Y \in E_M\}$. This simplification might make our model misspecified in that $\overline{\mathbb{P}}_{\theta^*}$ might fall out of the framework of independent Bernoulli trials. The asymptotic properties of the MLE under a misspecified model are well known. First we expect $\widehat{\theta}$ to be asymptotically consistent for a parameter vector $\overline{\theta}(\theta^*)$ which minimizes the conditional expected negative log-likelihood defined by

$$\overline{\theta}(\theta^*) \in \arg\min_{\theta \in \Theta_M} \overline{\mathbb{E}}_{\theta^*} \left[\mathcal{L}_N(\theta, Z^M) \right] . \tag{13}$$

In the following, when there is no ambiguity we will simply denote $\overline{\theta}(\theta^*)$ by $\overline{\theta}$. The density $\mathbb{P}_{\overline{\theta}}$ can be understood as the projection of the true underlying distribution $\overline{\mathbb{P}}_{\theta^*}$ on the model using the Kullback-Leibler divergence. Second, we expect $\sqrt{N}(\widehat{\theta} - \overline{\theta})$ to be asymptotically normal with zero mean and covariance matrix $\overline{V} := \lim_{N\to\infty} N\overline{V}_N(\theta^*)$ (provided that the limit exists) where

$$\overline{V}_N(\theta^*) := H_N(\overline{\theta})^{-1} \left[L_N(\overline{\theta}, Z^M) L_N(\overline{\theta}, Z^M)^\top \right] H_N(\overline{\theta})^{-1}, \tag{14}$$

where

$$L_N(\theta, Z^M) := \frac{\partial \mathcal{L}_N}{\partial \theta}(\theta, Z^M) = \mathbf{X}_M^\top \big(\sigma(\mathbf{X}_M\theta) - Y \big),$$

is the score function and where $H_N(\theta) := \frac{\partial^2 \mathcal{L}_N}{\partial \theta^2}(\theta, Z^M)$ is the Hessian of the log-likelihood. This result (provided in [White, 1982, Theorem 3.2]) holds under some regularity conditions such as the continuous differentiability of the score function and a domination assumption on the Hessian of the log-likelihood. Denoting

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

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

$$L_N(\widehat{\theta}, Z^M) = 0 \quad \text{i.e.} \quad \mathbf{X}_M^\top (Y - \pi^{\widehat{\theta}}) = 0 \quad \Leftrightarrow \quad \Xi(\widehat{\theta}) = \mathbf{X}_M^\top Y \quad \Leftrightarrow \quad \widehat{\theta} = \Psi(\mathbf{X}_M^\top Y), \tag{15}$$

$$\text{and } \overline{L}_N(\overline{\theta}, \mathbf{X}_M) = 0 \quad \text{i.e.} \quad \mathbf{X}_M^\top (\overline{\pi}^{\theta^*} - \pi^{\overline{\theta}}) = 0 \quad \Leftrightarrow \quad \Xi(\overline{\theta}) = \mathbf{X}_M^\top \overline{\pi}^{\theta^*} \quad \Leftrightarrow \quad \overline{\theta} = \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta^*}), \tag{16}$$

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

$$\overline{V}_N(\theta^*) = H_N(\overline{\theta})^{-1} \overline{G}_N^c(\theta^*) H_N(\overline{\theta})^{-1},$$

where

$$H_N(\overline{\theta}) = \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\overline{\theta}))\mathbf{X}_M \quad \text{and} \quad \overline{G}_N^c(\theta^*) = \mathbf{X}_M^\top \overline{\mathbb{E}}_{\theta^*} \left[(Y - \pi^{\overline{\theta}})(Y - \pi^{\overline{\theta}})^\top \right] \mathbf{X}_M.$$

Let us state explicitly that the previous asymptotic considerations hold under specific assumptions that are not satisfied in our setting. Nevertheless, building a bridge between the standard theory of the MLE

under model misspecification and our framework of PSI can help us choose a relevant covariance structure to design the SIGLE test statistic. In the rest of this paper, we will consider the following proxy for the covariance matrix of the score $\overline{G}_N^c(\theta^*)$:

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

$\overline{G}_N(\theta^*)$ is obtained from $\overline{G}_N^c(\theta^*)$ by using $\mathbf{X}_M^\top \overline{\pi}^{\theta^*} = \mathbf{X}_M^\top \pi^{\overline{\theta}}$ (cf. Eq.(16)) and by keeping only the diagonal elements of the covariance matrix $\overline{\mathbb{E}}_{\theta^*}\big[(Y - \overline{\pi}^{\theta^*})(Y - \overline{\pi}^{\theta^*})^\top\big]$ while setting to zero the off-diagonal entries. Therefore, in the selected model SIGLE relies on the following test statistic

$$\|[\overline{G}_N(\theta^*)]^{-1/2} H_N(\overline{\theta})(\widehat{\theta} - \overline{\theta})\|_2^2. \tag{17}$$

The choice to work with $\overline{G}_N(\theta^*)$ rather than $\overline{G}_N^c(\theta^*)$ is motivated by several reasons:

1. Working with $\overline{G}_N(\theta^*)$ makes the theoretical analysis simpler although the post-selection inference methods proposed in this paper remain valid in the case one uses $\overline{G}_N^c(\theta^*)$.
2. Extensive numerical experiments have shown that the power of hypothesis tests using SIGLE with either $\overline{G}_N(\theta^*)$ or $\overline{G}_N^c(\theta^*)$ is very similar (and some of them are presented in Section 2).
3. Only N coefficients need to be estimated to approximate $\overline{G}_N(\theta^*)$ as opposed to the N^2 coefficients required to estimate $\overline{G}_N^c(\theta^*)$. As a consequence, working with $\overline{G}_N(\theta^*)$ might allow to reduce the variance of our estimate of the SIGLE statistic and thus to get closer to the power that would give SIGLE using the unknown quantities $\overline{\pi}^{\theta^*}$, $\overline{\theta}(\theta^*)$.

SIGLE statistic in the saturated model. Let us start by introducing some notations. By assuming that ξ is strictly convex, one can compute $\mathbf{X}\vartheta^*$ from π^*, allowing us to denote equivalently $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\vartheta^*}$ with an abuse of notation. Given some set of selected variables $M \subseteq [d]$ with $s := |M|$ and some $\vartheta^* \in \mathbb{R}^d$, we denote by $\overline{\mathbb{P}}_{\pi^*}$ the distribution of Y given E_M, namely

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

$\pi^* = \sigma(\mathbf{X}\vartheta^*)$ and where $\propto$ means equal up to a normalization constant. In the selected model with $\theta^* \in \Theta_M$ satisfying Eq.(11), we will also denote $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\theta^*}$.

In the saturated, we have already explained that we focus on the statistic $\mathbf{X}_M^\top Y$. Recalling the definition of Ξ (cf. Eq.(7)) and using Eq.(15), we get that $\mathbf{X}_M^\top Y = \Xi(\widehat{\theta})$. Therefore, one can apply the delta method to convert the heuristic CLT obtained for $\widehat{\theta}$ (cf. Eq.(14)) into a similar asymptotic result for $\mathbf{X}_M^\top Y$. The delta method suggests that $\Xi(\widehat{\theta}) = \mathbf{X}_M^\top Y$ should be asymptotically normal with mean $\lim_{N\to\infty} \Xi(\overline{\theta}) = \mathbf{X}_M^\top \overline{\pi}^{\theta^*}$ (using Eq. (16)) and covariance matrix

$$\lim_{N\to\infty} \nabla\Xi(\overline{\theta})^\top \overline{V}_N(\theta^*) \nabla\Xi(\overline{\theta}) = \lim_{N\to\infty} \overline{G}_N(\theta^*),$$

where we used that $\nabla\Xi(\overline{\theta}) = H_N(\overline{\theta})$. A careful reader would note that it makes no sense to refer to θ^* in the saturated model. To overcome this issue, one can realize that θ^* only appears in the asymptotic description of $\mathbf{X}_M^\top Y$ through $\overline{\pi}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}[Y] = \overline{\mathbb{E}}_{\pi^*}[Y]$. Therefore, denoting $\overline{\pi}^{\theta^*}$ by

$$\overline{\pi}^{\pi^*} := \overline{\mathbb{E}}_{\pi^*}[Y],$$

the previous discussion suggests that $\mathbf{X}_M^\top(Y - \overline{\pi}^{\pi^*})$ should be asymptotically normal with mean 0 and covariance matrix $\lim_{N\to\infty} \overline{G}_N(\pi^*)$ where

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

Therefore, SIGLE in the saturated model relies on the following test statistic

$$\|[\overline{G}_N(\pi^*)]^{-1/2} \mathbf{X}_M^\top (Y - \overline{\pi}^{\pi^*})\|_2^2. \tag{18}$$

Discussion. The presentation of the SIGLE statistics in this section naturally gives rise to the following questions $(\mathcal{Q}_k)_{k\in[4]}$:

- **$\mathcal{Q}_1$: How to use the SIGLE statistics (17) and (18) in practice?**

  In most cases, the distribution of the observations conditional to the selection event is unknown and computing (17) or (18) requires to use sampling methods.

  We consider hypothesis tests with pointwise nulls as presented in Table 2. Assuming that we are able to sample state according the $\overline{\mathbb{P}}_{\pi_0^*}$ (resp. $\overline{\mathbb{P}}_{\theta_0^*}$), we can compute estimates $\widetilde{G}_N(\pi_0^*), \widetilde{\pi}^{\pi_0^*}$ (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).

	Null and alternative	Test statistic
Saturated model	$\mathbb{H}_0 : \{\pi^* = \pi_0^*\}$, $\mathbb{H}_1 : \{\pi^* \neq \pi_0^*\}$	$\|\widetilde{G}_N(\pi_0^*)^{-1/2}(\mathbf{X}_M^\top Y - \mathbf{X}_M^\top \widetilde{\pi}^{\pi_0^*})\|_2^2$
Selected model	$\mathbb{H}_0 : \{\theta^* = \theta_0^*\}$, $\mathbb{H}_1 : \{\theta^* \neq \theta_0^*\}$	$\|\widetilde{V}_N(\theta_0^*)^{1/2}(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta}(\theta_0^*))\|_2^2$

Table 2: Test statistics of SIGLE.

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 2, we present our hypothesis tests in both the saturated and the selected models.

- **$\mathcal{Q}_2$: What are the asymptotic properties of the SIGLE statistics (17) and (18)?**

  The way the SIGLE statistics have been motivated in this section naturally opens the question of their asymptotic properties. More precisely, can we find conditions ensuring that $[\overline{G}_N(\pi^*)]^{-1/2}\mathbf{X}_M^\top(Y - \overline{\pi}^{\pi^*})$ (resp.$[\overline{G}_N(\theta^*)]^{-1/2}H_N(\overline{\theta})(\widehat{\theta} - \overline{\theta})$ in the selected model) is asymptotically normal? Asymptotic considerations have already been used in the literature to design post-selection inference methods in GLMs such as in Taylor and Tibshirani [2018]. Such approaches often rely on non-rigorous computations conducted under (very) restrictive assumptions.

  In the case of logistic regression, we prove conditional central limit theorems (CLTs) for the SIGLE statistics in both the selected and the saturated model. As far as we know, we are the first to provide such results in the field of PSI. Our conditional CLTs hold under conditions that are similar to the ones usually considered in the literature when studying the asymptotic properties of the MLE in high dimensions (cf.Bunea [2008]).

  Furthermore, we provide an extensive comparison between our methods and the one from Taylor and Tibshirani [2018] on both the numerical side (cf. Section 4) and the theoretical side (cf. Section A).

- **$\mathcal{Q}_3$: Other statistics might have been considered. How the SIGLE statistics from (17) and (18) perform compared to other approaches?**

  At the end of Section 2, we show with numerical experiments that SIGLE statistics lead to more powerful testing procedures compared to methods based on other reasonable choices for the test statistics. In Section 4, we compare our method with state of the art approaches for PSI in logistic regression.

- **$\mathcal{Q}_4$: How the methods of this paper can be interpreted when the model is misspecified from the start?**

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

this assumption? In Section D.1, we consider that the y_i's are i.i.d. and distributed according to an arbitrary probability distribution $\mathbb{P}$.

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 a useful 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 3.

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

Table 3: Positioning of SIGLE (this paper) among some pioneering works on PSI in GLMs.

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 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 this paper and Taylor and Tibshirani [2018] are tackling the problem of post selection inference in the logistic model.

1.6 Contributions and organization of the paper

SIGLE for an arbitrary GLM (Sec.1).

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.1.4).

3. We introduce the SIGLE statistics in both the saturated and the selected model. Computing these statistics and calibrating the SIGLE hypothesis testing require to be able to sample from the distribution of the observations conditional to the selection event (cf. Sec.1.4).

SIGLE for the Sparse Logistic Regression (SLR) (from Sec.2).

4. We describe in details the way to use SIGLE in practice in both the selected in the saturated model (cf. Sec.2).

5. In the context of the SLR, we introduce a new sampling method allowing to compute the SIGLE statistics and to calibrate our methods (cf. Sec.3).

6. We provide an extensive comparison between this paper 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. The methods are compared both on the numerical side (cf. Sec.4) and the theoretical side (cf. Sec.A).

7. Going back to the motivation behind the choice of the SIGLE statistics, we study the asymptotic properties of the conditional MLE. We provide conditions under which conditional CLTs hold (cf. Sec.5).

Outline. In this paper, we focus specifically on the SLR. We start by describing the SIGLE hypothesis testing methods in this context in Section 2. 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 present the results of our simulations. We conclude in Section 5 by providing two conditional central limit theorems.

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 $x \in \mathbb{R}^d$, $R > 0$ and for $p \in [1,\infty]$, we define $\mathbb{B}_p(x,R) = \{z \in \mathbb{R}^d \,|\, \|z\|_p \le R\}$. Let us finally recall that given some set of selected variables $M \subseteq [d]$ with $s := |M|$ and some $\vartheta^* \in \mathbb{R}^d$, we denote by $\overline{\mathbb{P}}_{\pi^*}$ the distribution of Y conditional on E_M, namely

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

$\pi^* = \sigma(\mathbf{X}\vartheta^*)$ and where $\propto$ means equal up to a normalization constant. By assuming that ξ is strictly convex, one can compute $\mathbf{X}\vartheta^*$ from π^*, 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 $\mathbb{P}_{\pi^*} \equiv \mathbb{P}_{\theta^*}$.

2 Comprehensive description of SIGLE for SLR

From this section, we consider the case of the logistic regression where we recall that $Y = (y_i)_{i\in[N]}$ and for all $i \in [N]$, $y_i \sim \mathrm{Ber}(\pi_i^*)$ with $\pi^* = \sigma(\mathbf{X}\vartheta^*)$. As already explained in the introduction, the SIGLE statistics are motivated by the conditional CLTs provided in details in Section 5. In this section, we describe our methods.

SIGLE in the saturated model. Given some $\pi_0^* \in \mathbb{R}^N$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\pi^* = \pi_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\pi^* \neq \pi_0^*\}. \tag{19}$$

The statistics given by the CLT from Theorem 2 (cf. Section 5) naturally leads us to introduce the ellipsoid W_N given by

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

where

- $w_{N,1-\alpha}$ is the quantile of order $1-\alpha$ of the SIGLE statistic

$$\left\| [\overline{G}_N(\pi_0^*)]^{-1/2} \mathbf{X}_M^\top \left(Y - \overline{\pi}^{\pi_0^*} \right) \right\|_2^2,$$

- $\overline{G}_N(\pi^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\pi^*})^2) \mathbf{X}_M$ with $(\overline{\sigma}^{\pi^*})^2 := \overline{\pi}^{\pi^*} \odot (1 - \overline{\pi}^{\pi^*})$.

If $\overline{\mathbb{P}}_{\pi_0^*}$ was nice enough, we could hope to easily compute *i)* $\overline{\pi}^{\pi_0^*}$ and then $\overline{G}_N(\pi_0^*)$ and *ii)* $w_{N,1-\alpha}$. Contrary to the linear model where the conditional distribution is known to be a truncated Gaussian, we do not have such a characterization of $\overline{\mathbb{P}}_{\pi_0^*}$ in SLR. As a consequence, we propose in the paper two different ways to sample state in the selection event and to estimate the parameters $\overline{\pi}^{\pi_0^*}$ and $w_{N,1-\alpha}$ in order to approximate the rejection region W_N. Both methods are presented in Section 3. The first sampling approach is a simple rejection sampling method. This method is particularly appropriate when the number of features d is small. When d is getting large, another sampling method is needed and we introduce in this paper the SEI-SLR algorithm. From Proposition 4 (cf. Section 3), we know that under an appropriate cooling scheme, the asymptotic distribution of the states visited by the SEI-SLR algorithm (cf. Algorithm 3) 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 $\overline{G}_N(\pi_0^*)$. Algorithm 1 describes the testing procedure when we sample states in the selection event using the SEI-SLR algorithm.

When the states $(Y^{(t)})_{t\geq 1}$ in step 2. of Algorithm 1 are sampled using the rejection method instead of the SEI-SLR algorithm, one only needs to change the way $\widetilde{\pi}^{\pi_0^*}$ and $\zeta_{N,T}$ are computed by using

$$\widetilde{\pi}^{\pi_0^*} = \frac{1}{T} \sum_{t=1}^{T} Y^{(t)} \quad \text{and} \quad \zeta_{N,T} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{Y^{(t)} \in \widetilde{W}_N}.$$

SIGLE in the selected model. Given some $\theta_0^* \in \mathbb{R}^s$, we consider the hypothesis test with null and alternative hypotheses defined by

$$\mathbb{H}_0 \ : \ \{\theta^* = \theta_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\theta^* \neq \theta_0^*\}. \tag{20}$$

The statistic given by the CLT from Theorem 3 (cf. Section 5) naturally leads us to introduce the ellipsoid W_N given by

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

where

Algorithm 1 SIGLE in the saturated model.

1: **Input:** $\mathbf{X} \in \mathbb{R}^{N\times d}$, $Y \in \mathbb{R}^N$, $\lambda > 0$, $\alpha \in (0,1)$.
$(\mathbf{X}, Y, \lambda)$ characterizes the selection event E_M (cf. Eq.(4)).
2: Sample states $(Y^{(t)})_{t\geq 1}$ uniformly distribution on E_M using the SEI-SLR algorithm (cf. Algorithm 3).
3: Compute:
- $\widetilde{\pi}^{\pi_0^*} = \dfrac{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)})Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\pi_0^*}(Y^{(t)})}$,
- $\widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\pi_0^*} \odot (1-\widetilde{\pi}^{\pi_0^*})\right)\mathbf{X}_M$,
- $\widetilde{w}_{N,1-\alpha}$ which is the quantile of order $1-\alpha$ of the sequence $\left(\left\|\widetilde{G}_N^{-1/2}\mathbf{X}_M^\top\left(Y^{(t)} - \widetilde{\pi}^{\pi_0^*}\right)\right\|_2^2\right)_{t\geq 1}$.

4: Define $\widetilde{W}_N := \left\{Y \in \{0,1\}^N \mid \left\|\widetilde{G}_N^{-1/2}\mathbf{X}_M^\top\left(Y - \widetilde{\pi}^{\pi_0^*}\right)\right\|_2^2 > \widetilde{w}_{N,1-\alpha}\right\}$.
5: Reject the null hypothesis $\mathbb{H}_0$ when

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

- $w_{N,1-\alpha}$ is the quantile of order $1-\alpha$ of the SIGLE statistic

$$\left\|[\overline{G}_N(\theta_0^*)]^{-1/2} H_N(\overline{\theta}(\theta_0^*))\left(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta_0^*)\right)\right\|_2^2,$$

- $H_N(\theta) := \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = \mathbf{X}_M^\top \mathrm{Diag}((\sigma^\theta)^2)\mathbf{X}_M$ is the Fisher information matrix with $(\sigma^\theta)^2 := \pi^\theta \odot (1-\pi^\theta)$ and $\pi^\theta = \mathbb{E}_\theta[Y]$,
- $\overline{G}_N(\theta^*) := \mathbf{X}_M^\top \mathrm{Diag}((\overline{\sigma}^{\theta^*})^2)\mathbf{X}_M$ is the natural counterpart of the Fisher information matrix $H_N(\theta^*)$ when we work under the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ with $(\overline{\sigma}^{\theta^*})^2 := \overline{\pi}^{\theta^*} \odot (1-\overline{\pi}^{\theta^*})$, $\overline{\pi}^{\theta^*} = \overline{\mathbb{E}}_{\theta^*}[Y]$.

We rely - as in the saturated model - on the SEI-SLR algorithm or the rejection sampling method (cf. Section 3) to estimate the parameters $\overline{\pi}^{\theta_0^*}$ and $w_{N,1-\alpha}$ in order to approximate the rejection region W_N. The SIGLE procedure in the selected model in presented in Algorithm 1 when the SEI-SLR algorithm is used.

When the states $(Y^{(t)})_{t\geq 1}$ in step 2. of Algorithm 2 are sampled using the rejection method instead of the SEI-SLR algorithm, one only needs to change the way $\widetilde{\pi}^{\theta_0^*}$ and $\zeta_{N,T}$ are computed by using

$$\widetilde{\pi}^{\theta_0^*} = \frac{1}{T}\sum_{t=1}^T Y^{(t)} \quad \text{and} \quad \zeta_{N,T} = \frac{1}{T}\sum_{t=1}^T \mathbb{1}_{Y^{(t)}\in\widetilde{W}_N}.$$

The careful reader can notice that Algorithm 2 requires to compute efficiently $\Psi(\mathbf{X}_M^\top\pi)$ for any $\pi \in [0,1]^N$. In the specific case where $\pi = Y \in \{0,1\}^N$, we know that $\Psi(\mathbf{X}_M^\top Y$ is the conditional MLE (cf. Eq. (15)) and thus can be computed using the usual Iterative Reweighted Least Squares algorithm. For an arbitrary $\pi \in [0,1]^N$ (such as in step 3. of Algorithm 2 to compute $\widetilde{\theta}$), we need to use another approach. In Section D.2 of the Appendix, we describe in details our gradient descent-based method to compute $\Psi(\mathbf{X}_M^\top\pi)$ which proved to be extremely accurate in our numerical experiments.

Discussion regarding the choice of the SIGLE statistic. As explained in Section 1.4, the SIGLE statistics can be motivated by making a connection between PSI and asymptotic properties of the MLE with model misspecification. Let us present a numerical experiment providing an additional support for the choice of the SIGLE statistics. We consider the hypothesis test in the saturated model presented in Table 2 with $\pi_0^* = \frac{1}{2}\mathbb{1}_N$.

Algorithm 2 SIGLE in the selected model.

Input: $\mathbf{X} \in \mathbb{R}^{N\times d}$, $Y \in \mathbb{R}^N$, $\lambda > 0$, $\alpha \in (0,1)$.
$(\mathbf{X}, Y, \lambda)$ characterizes the selection event E_M (cf. Eq.(4)).
Sample states $(Y^{(t)})_{t\geq 1}$ uniformly distribution on E_M using the SEI-SLR algorithm (cf. Algorithm 3).
Compute:

- $\widetilde{\pi}^{\theta_0^*} = \dfrac{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})Y^{(t)}}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})}$,
- $\widetilde{\theta} = \Psi(\mathbf{X}_M^\top \widetilde{\pi}^{\theta_0^*})$,
- $\widetilde{G}_N = \mathbf{X}_M^\top \mathrm{Diag}\left(\widetilde{\pi}^{\theta_0^*} \odot (1-\widetilde{\pi}^{\theta_0^*})\right)\mathbf{X}_M$,
- $\widetilde{w}_{N,1-\alpha}$ which is the quantile of order $1-\alpha$ of the sequence $\left(\left\|\widetilde{G}_N^{-1/2} H_N(\widetilde{\theta})\left(\Psi(\mathbf{X}_M^\top Y^{(t)}) - \widetilde{\theta}\right)\right\|_2^2\right)_{t\geq 1}$.

Define $\widetilde{W}_N := \left\{Y \in \{0,1\}^N \mid \left\|\widetilde{G}_N^{-1/2} H_N(\widetilde{\theta})\left(\Psi(\mathbf{X}_M^\top Y) - \widetilde{\theta}\right)\right\|_2^2 > \widetilde{w}_{N,1-\alpha}\right\}$.
Reject the null hypothesis $\mathbb{H}_0$ when

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

We consider a design matrix $\mathbf{X} \in \mathbb{R}^{100\times 10}$ where the entries are i.i.d. and sampled from a standard normal distribution. We use a regularization parameter $\lambda = 5$. We work with the following three test statistics:

- the SIGLE statistic: $\left\|[\overline{G}_N(\pi_0^*)]^{-1/2}\mathbf{X}_M^\top\left(Y - \overline{\pi}^{\pi_0^*}\right)\right\|_2^2$,
- the SIGLE correlated statistic: $\left\|[\overline{G}_N^c(\pi_0^*)]^{-1/2}\mathbf{X}_M^\top\left(Y - \overline{\pi}^{\pi_0^*}\right)\right\|_2^2$ where

$$\overline{G}_N^c(\pi_0^*) = \mathbf{X}_M^\top \overline{\mathbb{E}}_{\pi_0^*}\left[(Y - \overline{\pi}^{\pi_0^*})(Y - \overline{\pi}^{\pi_0^*})^\top\right]\mathbf{X}_M,$$

- the logistic unconditional Fisher statistic: $\left\|[H_N(\pi_0^*)]^{-1/2}\mathbf{X}_M^\top(Y - \pi_0^*)\right\|_2^2$.

We calibrate each testing procedure by sampling under the null distribution. Figure 2.(a) presents the cumulative distribution function of the p-values obtained considering the alternative $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ with $\vartheta^* = 0.2\times\mathbb{1}_d$ for the different tests. We see that the SIGLE statistic leads to more powerful tests compared to the logistic unconditional Fisher statistic. Moreover, the SIGLE statistic and the SIGLE correlated statistic give similar result as already explained in Section 1.4.

Figure 2.(b) shows the pdf of the SIGLE statistic under the null and the pdf of the closer χ^2 distribution, in the sense that we chose the degree of freedom for the χ^2 distribution that gives the smallest L^2 error between the χ^2 quantiles and the SIGLE's quantiles. It appears that this optimal degree of freedom is 14. Figure 2.(b) makes clear that the SIGLE statistic is not distributed as a χ^2 random variable contrary to what our conditional CLT from Section 5 is suggesting. The obvious reason is that our conditional CLT from Section 5 holds only under restrictive conditions that are nonetheless standard in the literature (cf. Bunea [2008]). This is one reason motivating the calibration of the SIGLE procedures by sampling under the null. Let us highlight that this is not restrictive in the sense that the computation of the SIGLE statistics require anyway to sample under the null in order to estimate both $\overline{\pi}^{\pi_0^*}$ and $\overline{G}_N(\pi_0^*)$.

We conducted the same analysis in the selected model working with the following three test statistics:

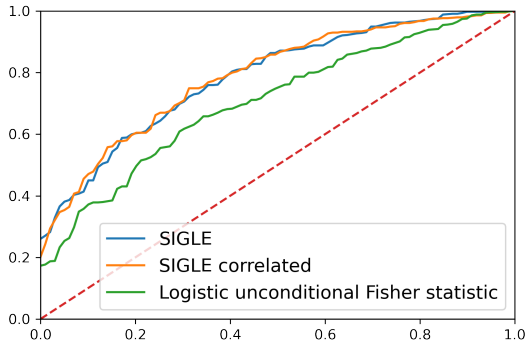
- the SIGLE statistic: $\left\|[\overline{G}_N(\theta_0^*)]^{-1/2}H_N(\theta_0^*)\left(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta_0^*)\right)\right\|_2^2$,

- the SIGLE correlated statistic: $\left\| [\overline{G}_N^c(\theta_0^*)]^{-1/2} H_N(\theta_0^*) \left(\Psi(\mathbf{X}_M^\top Y) - \overline{\theta}(\theta_0^*) \right) \right\|_2^2$, where

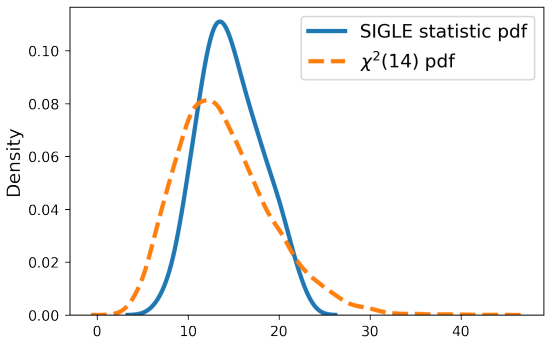
$$\overline{G}_N^c(\theta_0^*) = \mathbf{X}_M^\top \overline{\mathbb{E}}_{\theta_0^*} \left[(Y - \overline{\pi}^{\theta_0^*})(Y - \overline{\pi}^{\theta_0^*})^\top \right] \mathbf{X}_M,$$

- the logistic unconditional Fisher statistic: $\left\| [H_N(\theta_0^*)]^{-1/2} \left(\Psi(\mathbf{X}_M^\top Y) - \theta_0^* \right) \right\|_2^2$.

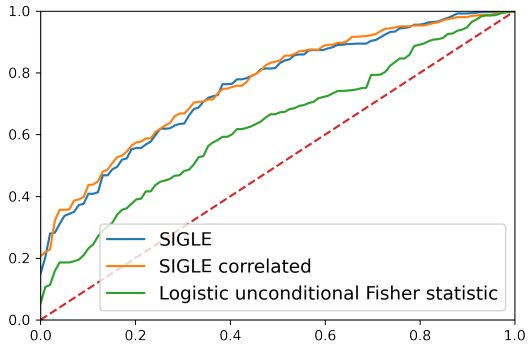
The results are presented in Figures 2.(c) and (d) with similar conclusions.



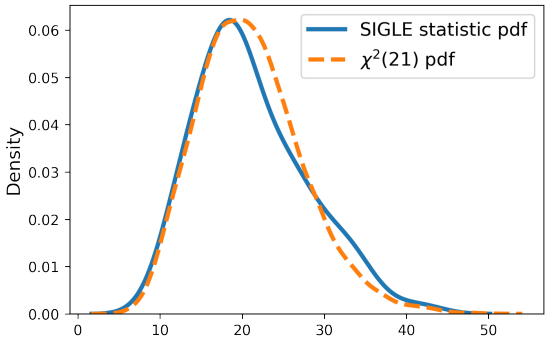
(a) CDF of p-values for the alternative $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ with $\vartheta^* = 0.2 \times \mathbb{1}_d$.



(b) Pdfs of the SIGLE statistic under the null and of the $\chi^2(14)$ distribution.



(c) CDF of p-values for the alternative $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ with $\vartheta^* = 0.2 \times \mathbb{1}_d$.



(d) Pdfs of the SIGLE statistic under the null and of the $\chi^2(21)$ distribution.

Figure 2: Figure (a) (resp. (c)) shows the CDF of p-values for the alternative $\pi^* = \sigma(\mathbf{X}\vartheta^*)$ with $\vartheta^* = 0.2 \times \mathbb{1}_d$ for the test using the SIGLE statistic, the SIGLE correlated statistic and the logistic unconditional Fisher statistic in the saturated (resp. selected) model. Figure (b) (resp. (d)) presents the probability density function (pdf) of the SIGLE statistic under the null and the one of a χ^2 distribution with 14 (resp. 21) degrees of freedom in the saturated (resp. selected) model.

3 Sampling from the conditional distribution

Let us recall that we focus on the case of the logistic regression. We propose two different approaches to compute quantities of the form $\overline{\mathbb{E}}_{\vartheta^*}[h(Y)]$ for some map $h : \{0,1\}^N \to \mathbb{R}$.

The first one is a simple rejection sampling method that can be used when carrying simple hypothesis testing as presented in Table 2. In this situation, one can sample states from $\mathbb{P}_{\pi_0^*}$ in the saturated model (resp. $\mathbb{P}_{\sigma(\mathbf{X}_M\theta_0^*)}$ in the selected model) while keeping only the ones leading to the selected support M. By construction, the distribution of the saved states is precisely $\overline{\mathbb{P}}_{\pi_0^*}$ (resp. $\mathbb{P}_{\sigma(\mathbf{X}_M\theta_0^*)}$). This method is particularly appropriate when the number of features d is small since the number of possible selected support for the lasso solution is exponential in d. When d is getting large or when we want to derive a confidence region, another sampling method is needed.
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 regime as $t \to \infty$. Contrary to the rejection sampling method, this simulated-annealing based algorithm can be used to compute expectations of the form $\overline{\mathbb{E}}_{\vartheta^*}[h(Y)]$ regardless of the inference procedure conducted or when d is large. Nevertheless, let us point out that this approach requires an appropriate tuning of some parameters, and the convergence guarantees are only asymptotic. An extensive discussion of our sampling strategies is provided in Section 4.1.

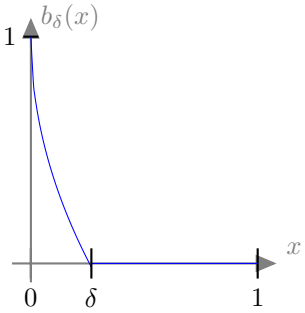
3.1 SEI-SLR: sampling the selection event

From Proposition 1 and the KKT conditions from (3), we know that the selection event E_M can be written as

$$E_M = \left\{ Y \in \{0,1\}^N \mid \mathbb{1}_{\|\widehat{S}_{-M}(Y)\|_\infty - 1 < 0},\ \mathbb{1}_{1=\min_{k\in M}\{|\widehat{S}_k(Y)|\}} \right\}. \tag{21}$$

Based on the expression of E_M given in Eq.(21), 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

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

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.(21) 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.(21) 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 the 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 3 and in the next section, we provide theoretical guarantees. In Algorithm 3, $P : \{0,1\}^N \times \{0,1\}^N \to [0,1]$ 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 3 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**
    $Y^{(t)} \leftarrow Y^{\mathrm{c}}$
  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 $\pi_i^* \in (0,1)$ conditional 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 5.

Proposition 5. *Let us consider $M \subseteq [d]$ and some $\vartheta^* \in \mathbb{R}^d$. Consider a 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 Numerical results

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

4.1 Sampling the conditional distribution with SEI-SLR

As already discussed in the beginning of Section 3, we propose two different ways to sample points on the hypercube $\{0,1\}^N$ allowing us to compute conditional expectations of the form $\overline{\mathbb{E}}_{\theta^*}[h(Y)]$ or $\overline{\mathbb{E}}_{\pi^*}[h(Y)]$ where $h : \{0,1\}^N \to \mathbb{R}$.

The first method is a simple rejection sampling approach and is described in Algorithm 4. The rejection

Algorithm 4 Rejection sampling.

Input: T, π^*, $\mathbf{X}$, M, λ
$t \leftarrow 0$
while $t < T$ **do**
  $Y \sim \mathbb{P}_{\pi^*}$
  if $Y \in E_M$ **then**
    $t \leftarrow t + 1$
    $Y^{(t)} \leftarrow Y$
  end if
end while
return $(Y^{(t)})_{t\in[T]}$

sampling algorithm does not require any parameter tuning and allow to estimate any expectation $\overline{\mathbb{E}}_{\pi^*}[h(Y)]$ by taking a simple average over the list of returned states namely $\sum_{t\in[T]} h(Y^{(t)})$. Nevertheless, a major drawback of the rejection sampling method is its large computing time when the number of features d is getting "large" (typically when d exceeds ten). Indeed, the number of possible supports for a lasso solution is equal to 2^d and increases exponentially fast with d.

In order to bypass this curse of dimensionality, we proposed in Section 3.1 the SEI-SLR algorithm: a simulated annealing-based method that is proved to generate states that are asymptotically uniformly distributed on the selection event. The SEI-SLR algorithm solves the computational issue faced by the rejection sampling for large p values. Nevertheless, the convergence of SEI-SLR algorithm requires the use of well-chosen parameters namely:

- the parameter δ involved in the energy (cf. Section 3.1),

	Rejection Sampling	SEI-SLR
Conditions for application	Simple hypothesis (cf. Table 2)	No condition
Need for hyperparameters tuning	No	Yes
Computational time	Efficient only for a small d but N can be chosen (very) large	Easier to use for relatively small N but d can be large
Distribution of the sequence of states generated $(Y^{(t)})_{t\in[T]}$	$\overline{\mathbb{P}}_{\theta_0^*}$ or $\overline{\mathbb{P}}_{\pi_0^*}$ (cf. Table 2)	Uniform distribution on E_M
	$\Downarrow$	$\Downarrow$
In simple hypothesis testing with $\mathbb{H}_0$: "$\theta^* = \theta_0^*$", $\overline{\mathbb{E}}_{\theta_0^*}[h(Y)] \approx \dots$	$\frac{1}{T}\sum_{t\in[T]} h(Y^{(t)})$	$\frac{\sum_{t\in[T]} \overline{\mathbb{P}}_{\theta_0^*}(Y^{(t)}) h(Y^{(t)})}{\sum_{r\in[T]} \overline{\mathbb{P}}_{\theta_0^*}(Y^{(r)})}$
	i.e. the estimate is obtained with a simple average on the sequence of generated states.	i.e. we need to weight properly each visited state.

Table 4: Comparison between the rejection sampling method and the SEI-SLR algorithm.

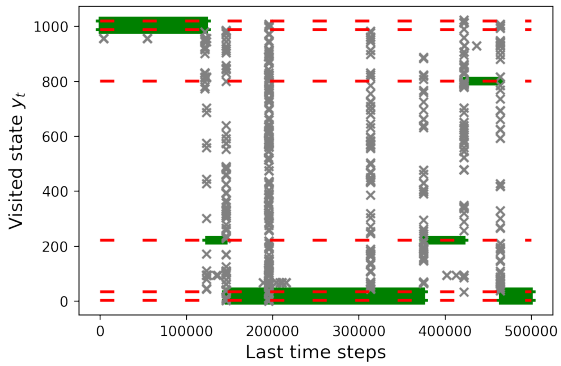
- the temperatures $(\mathrm{T}_t)_t$,
- the time horizon of the algorithm.

Let us finally mention that estimating expectations of the form $\overline{\mathbb{E}}_{\theta^*}[h(Y)]$ from the samples $(Y^{(t)})_t$ obtained with the SEI-SLR algorithm requires the computation of weighting factors that allow to go from the uniform distribution on the selection event E_M to the target conditional distribution $\overline{\mathbb{E}}_{\theta^*}$. In Table 4, we sum-up the previous discussion in order to give a comprehensive comparison between the two methods. In the rest of this section, we illustrate the performance of the SEI-SLR algorithm

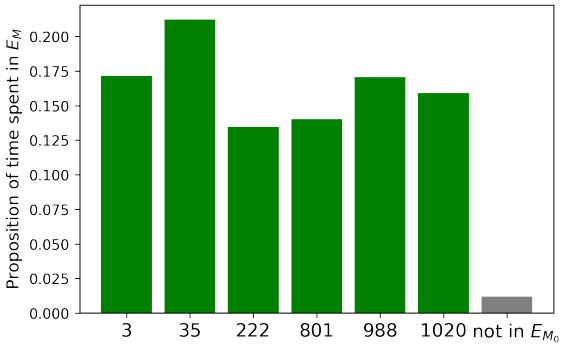
We consider a design matrix $\mathbf{X} \in \mathbb{R}^{10\times 20}$ where all entries are i.i.d. and sampled from a standard normal distribution. We consider $\delta = 0.01$, $\lambda = 1.5$ and we sample some vector $Y_0 \in \{0,1\}^N$ with i.i.d. entries with a Bernoulli distribution of parameter $1/2$. Note that the tuple $(\mathbf{X}, Y_0, \lambda)$ determined the set of active variables M (cf. Eq.(2)). We run the SEI-SLR algorithm for $3\,000\,000$ time steps. By choosing this toy example with a small value for N, we are able to compute exactly the selection event E_M by running over the 2^{10} possible vectors $Y \in \{0,1\}^N$. In the following, we identify each vector $Y \in \{0,1\}^N$ with the number between 0 and $2^N - 1 = 1024$ that it represents in the base-2 numeral system. Using this identification, it holds on our example that $E_M = \{3, 35, 222, 801, 988, 1020\}$.

Figure 3.(a) shows the last $500,000$ visited states for our simulated annealing path. On the vertical axis, we have the integers encoded by all possible vectors $Y \in \{0,1\}^N$. The red dashed lines represent the states that belong to the selection event E_M. While crosses are showing the visited states on the last $500,000$ 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. The simulated annealing path is jumping from one 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 3.(b)). Let us point that two neighboring states in space $\{0,1\}^N$ will not necessarily be encoded by close integers.

Figure 3.(a) suggests that the vectors encoded by the integers 3 and 35 are close in the space $\{0,1\}^N$. Indeed, we see on Figure 3.(a) that between indexes $180\,000$ and $350\,000$, our algorithm goes from one of



(a) Last 500 000 visited states of the SEI-SLR algorithm. The dotted red lines represent the states in E_M.



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

Figure 3: Visualization of the time spent in the selection event from the sequence of states provided by the SEI-SLR algorithm.

these 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 on this time window of the simulated annealing path). The same remark holds for the two states encoded by the integers 988 and 1020. However, we observe a large number of visited states that do not belong to E_M when we perform a transition between any other pair of states belonging to the selection event. We can therefore legitimately think that the selection event separates into four groups of fairly distant states. This is confirmed by Figure 4 which presents the Hamming distances between the different vectors of E_M and reveals the existence of two clusters.

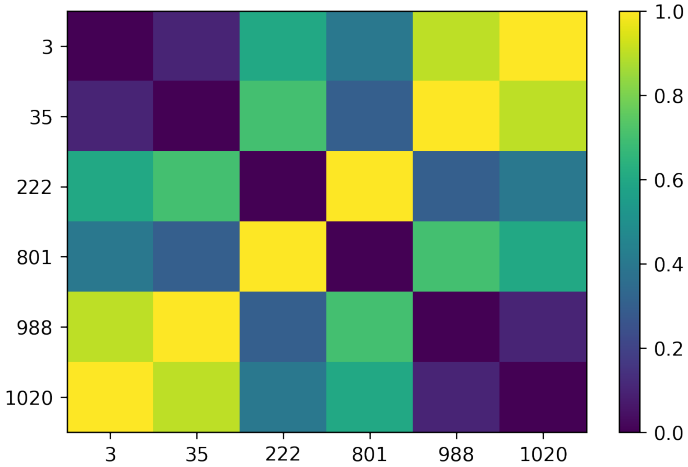
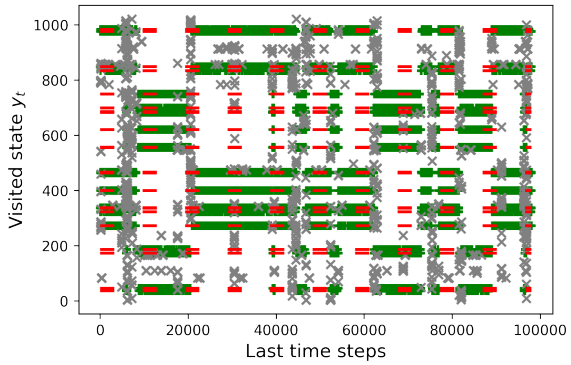
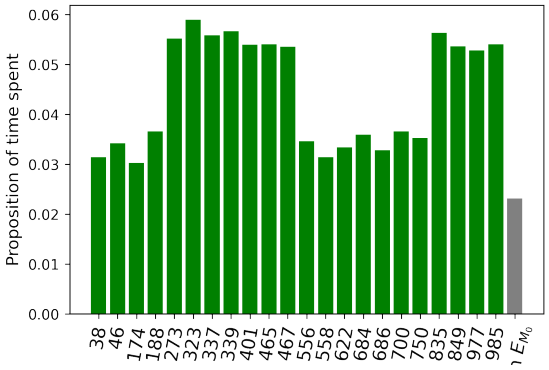


Figure 4: Normalized (by N) Hamming distances between the different states of the selection event.

With Figure 5, we show the results obtained from the SEI-SLR algorithm considering a similar experiment but taking $d = 15$ (instead of 20) and $\lambda = 2$ (instead of 1.5), which leads to a larger selection event.



(a) Last 100 000 visited states of the SEI-SLR algorithm. The dotted red lines represent the states in E_M.



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

Figure 5: Visualization of the time spent in the selection event from the sequence of states provided by the SEI-SLR algorithm.

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 can also cope with some computational issues. Indeed, the construction of confidence intervals conditionally on the event E_M requires the computation of 2^s intervals (while 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 Gaussian 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 leftover Fisher information. In Section F, we discuss with further details PSI when we condition additionally on the observed vector of signs.

4.2 Hypothesis Testing

In this section, we propose to analyze the level and the power of the SIGLE procedure considering the following simple hypothesis testing problem

$$\mathbb{H}_0 : \{\theta^* = \theta_0^*\}, \qquad \mathbb{H}_1 : \{\theta^* \neq \theta_0^*\}.$$

We compare the SIGLE method with the results obtained from a weak learner and from the heuristic method proposed by Taylor and Tibshirani [2018].

Description of the settings of our experiments. We consider a design matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ where the entries are i.i.d. and sampled from a standard normal distribution. We consider two different experiments (cf. Table 5). For the Setting 1 under the null, the set of active variables M is of size 4. We sample states from

E_M using the rejection sampling method and approximately 8% of the states sampled from $\mathbb{P}_{\theta_0^*}$ fall in the selection event with this algorithm. For the Setting 2, we use the SEI-SLR algorithm to sample states in E_M.

	N	d	$\mathbf{X}$	λ	θ_0^*	Sampling method
Setting 1	100	10	$\mathbf{X}_{i,j} \sim \mathcal{N}(0,1)$	5	$[0,\dots,0]$	Rejection sampling
Setting 2	20	15	$\mathbf{X}_{i,j} \sim \mathcal{N}(0,1)$	3	$[0,\dots,0]$	SEI-SLR

Table 5: Description of the experiments.

4.2.1 Description of the benchmark methods

A weak learner. Our weak learner is a two-sided test based on the statistic $\sum_{i=1}^n |\overline{\pi}_i^{\theta_0} - y_i|$ where $\overline{\pi}^{\theta_0}$ is the expectation of the vector of observations under the null conditional on the selection event (cf. Eq.(16)). Let us highlight that $\overline{\pi}^{\theta_0^*}$ is estimated by $\widetilde{\pi}^{\theta_0^*}$ where

- $\widetilde{\pi}^{\theta_0^*} = \sum_{t=1}^T Y^{(t)}$ if the sequence $(Y^{(t)})_{t\in[T]}$ is generated from the rejection sampling method,
- $\widetilde{\pi}^{\theta_0^*} = \frac{\sum_{t=1}^T Y^{(t)} \mathbb{P}_{\theta_0^*}(Y^{(t)})}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})}$ if we use the SEI-SLR algorithm to generate the states $(Y^{(t)})_{t\in[T]}$.

The method is calibrated empirically using the sequence $(Y^{(t)})_{t\in[T]}$.

The PSI method from Taylor and Tibshirani [2018]. The PSI method in the logistic model proposed by Taylor and Tibshirani [2018] is described in details in Section A.1. Based on heuristic justifications, this approach has the advantage to provide an hypothesis testing method for any linear transformation of the debiased lasso solution $\underline{\theta}$ (i.e. of the form $\eta^\top \underline{\theta}$) that does not require a cumbersome sampling step. We propose to compare the SIGLE methods with the PSI procedure from Taylor and Tibshirani [2018] by considering different approaches:

TT-1 We use the p-value obtained from a two-sided test based on the statistic $\underline{\theta}_1$.

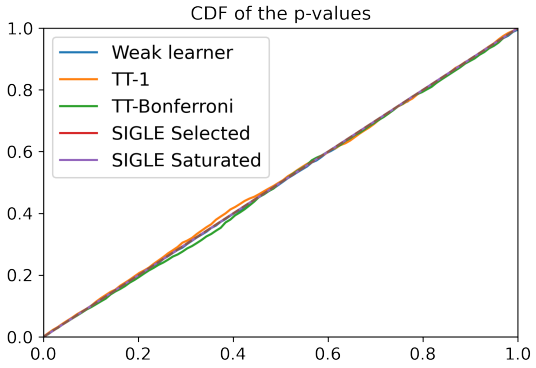
TT-Bonferroni We use a Bonferroni method from the p-values computed from the set of two-sided composite tests with null hypotheses $\mathbb{H}_0$: "$\theta_j^* = \left[\theta_0^*\right]_j$" for $j \in [s]$ where $s = |M|$.

4.2.2 Calibration

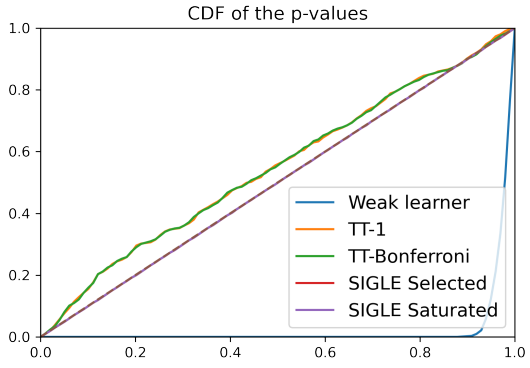
SIGLE procedures. To compute the SIGLE statistics, we need to estimate $\overline{G}_N(\pi_0^*)$ (and $\overline{\theta}(\theta_0^*)$ in the selected model). Since the conditional distribution $\overline{\mathbb{P}}_{\pi_0^*}$ (resp. $\overline{\mathbb{P}}_{\theta_0^*}$) is not known, we sample states from these distributions to estimate these quantities. We use these states sampled in the selection event E_M in order to calibrate empirically the SIGLE procedures. In the literature, one often says that we *calibrate by sampling under the null*.

PSI methods from Taylor and Tibshirani [2018]. In Taylor and Tibshirani [2018], the authors justify their approach with asymptotic considerations. Figure 6.(a) shows that for a large value for N, the methods TT-1 and TT-Bonferroni are correctly calibrated since the CDF of the p-values are uniform under the null. On the contrary, for small value of N, the calibration of these procedures may be lost as shown with Figure 6.(b).

The weak learner. By construction, the p-values of the weak learner are stochastically larger than uniform under the null. The CDF of p-values are uniformly distributed in the Setting 1 with Figure 6.(a). Note that in the Setting 2, the weak learner is irrelevant since the selection event E_M is such that $\overline{\pi}^{\theta_0^*} = \frac{1}{2}\mathbf{1}_N$. This means that the test-statistic of the weak learner is constant.



(a) Setting 1.



(b) Setting 2.

Figure 6: CDF of the p-values of the different testing procedures under the global null for the Settings given in Table 5. We calibrate empirically the SIGLE methods.

4.2.3 Power

We consider two different types of alternatives:

- *Localized alternatives.*
  A *localized signal* is of the form $\vartheta^* = [\nu\,,\, 0\,,\, ,\, \ldots\,,\, 0]$ for some $\nu > 0$.

- *Disseminated alternatives.*
  A *disseminated signal* is of the form $\vartheta^* = \nu \mathbf{1}_d$ for some $\nu > 0$.

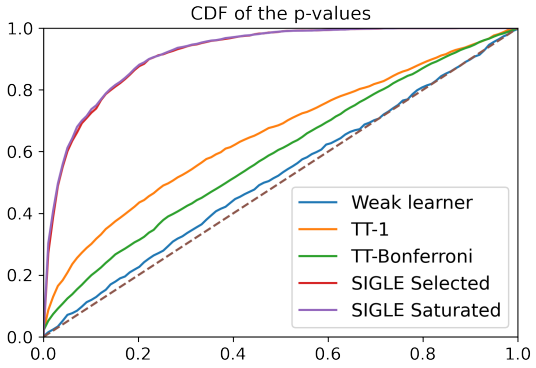
As explained in the previous sections, we calibrate the SIGLE methods empirically. Figure 6.(a) shows that the p-values for the SIGLE methods are distributed uniformly under the null. Figure 6.(a) also shows that the benchmark methods are correctly calibrated.
Figures 7 and 8 show that SIGLE is more powerful compared to the benchmark methods for the localized or disseminated alternative. In Figure 8.(b), the power of the SIGLE methods is so high that we barely identify the curve of the CDF in the top-left corner. Figure 9 gives a complete visualization of the power of the different testing methods when tests have level 5%. We see that the methods of this paper are always improving upon the benchmark methods. The superiority of the SIGLE methods regarding power becomes even more significant when we consider disseminated alternatives. This is not surprising since the methods of this paper are intrinsically designed to tackle simple hypothesis testing problem.

Another interesting remark is that the procedure TT-1 is more powerful than the procedure TT-Bonferroni when considering localized alternatives as showed by Figure 7. Again this result is not surprising: the TT-Bonferroni loses power by testing each coordinate of the parameter vector while TT-1 is focused on a single coordinate which is better suited to identify a localized signal. On the contrary, the TT-Bonferroni is more powerful when considering disseminated alternatives as observed with Figure 8 and Figure 9.(b).

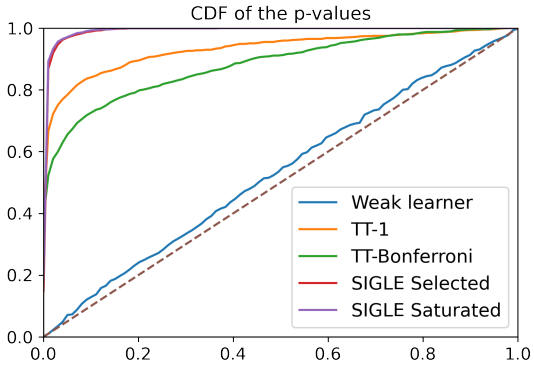
We conduct similar experiments in the Setting 2 given in Table 5. Figure 10 shows that the TT-1 and TT-Bonferroni are still less powerful than the SIGLE methods. Moreover, Figure 10.(b) illustrates that in the high dimensional setting (i.e. when d is larger than N), the size of the selection event can be small which leads to a non-smooth staircase function for the CDF of p-values. In the example of the Figure 10.(b), the selection event contains only 22 states.

4.2.4 Computational time and implementation details

Implementation of the SIGLE procedures.

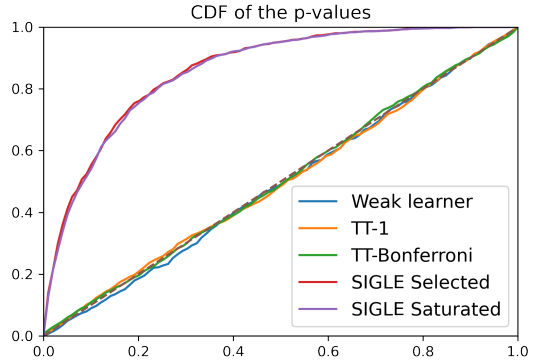


(a) CDF of p-values for the alternative $\vartheta^* = [0.4\,,\,0\,,\,\ldots\,,\,0]$.

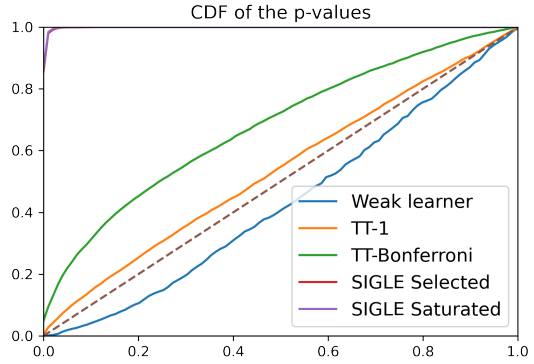


(b) CDF of p-values for the alternative $\vartheta^* = [0.9\,,\,0\,,\,\ldots\,,\,0]$.

Figure 7: CDF of the p-values for the SIGLE procedures and the benchmark methods using the **Setting 1** (cf. Table 5) for **localized alternatives**.



(a) CDF of p-values for the alternative $\vartheta^* = 0.04 \times \mathbf{1}_d$.

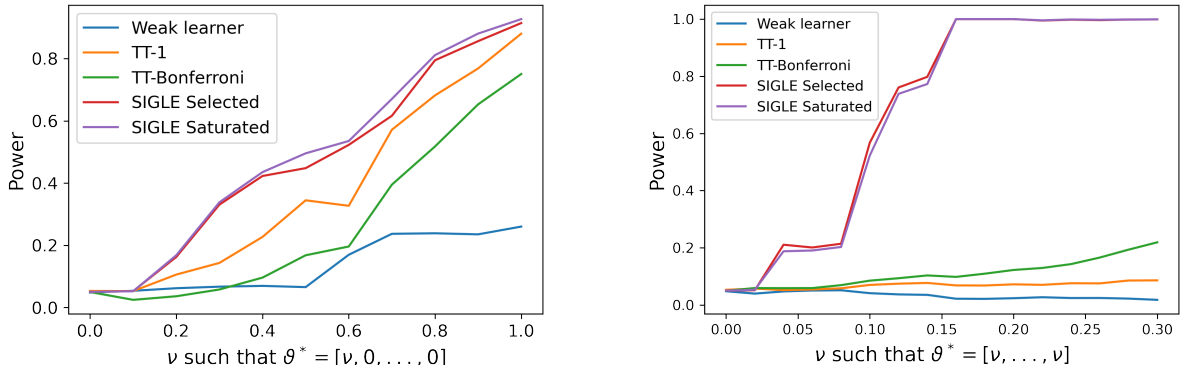


(b) CDF of p-values for the alternative $\vartheta^* = 0.3 \times \mathbf{1}_d$.

Figure 8: CDF of the p-values for the SIGLE procedures and the benchmark methods using the **Setting 1** (cf. Table 5) for **disseminated alternatives**.

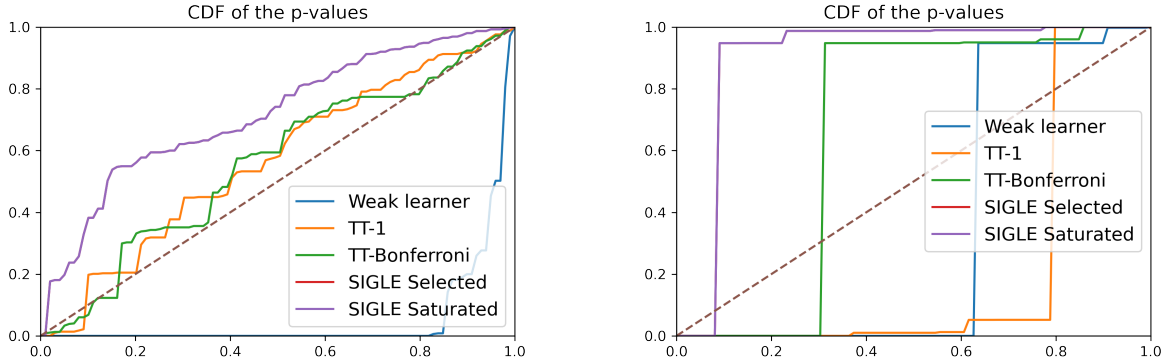
- SIGLE in the selected model.
  In the selected model, the SIGLE testing method requires to compute $\overline{\theta}(\theta_0^*) = \Psi(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*})$. Since we do not have a closed-form expression for $\Psi = \Xi^{-1}$, we first tried to learn this function by using a feed-forward neural network. We were not able to reach sufficient accuracy with this method and we proposed a gradient descent based approach to approximate $\overline{\theta}(\theta_0^*)$ from the estimate $\widetilde{\pi}^{\theta_0^*}$ of $\overline{\pi}^{\theta_0^*}$ (cf. Section 4.2.1). This algorithm is fully described in Section D.2.2. Making use of a proper warm start, we found this method highly robust and accurate to compute $\overline{\theta}(\theta_0^*)$.

- SEI-SLR algorithm and speed of convergence.
  In the previous sections, we proved the correctness of the SEI-SLR algorithm: the states visited by the algorithm are asymptotically distributed according to the uniform measure on E_M. This is an



(a) **Localized alternatives.** (b) **Disseminated alternatives.**

Figure 9: Comparison of the power of the SIGLE procedures and the benchmark methods using the **Setting 1** (cf. Table 5) for tests with level 0.05.



(a) CDF of p-values for the **localized alternative** $\vartheta^* = [3\,,\,0\,,\,\ldots\,,\,0]$.

(b) CDF of p-values for the **disseminated alternative** $\vartheta^* = 1.5 \times \mathbf{1}_d$.

Figure 10: CDF of the p-values for the SIGLE procedures and the benchmark methods using the **Setting 2** (cf. Table 5).

asymptotic result and MCMC methods are known to converge slowly. In order to increase the speed of convergence of the SEI-SLR algorithm, we found very useful in practice to introduce a *repulsing force* in the markovian transition kernel. Denoting $Y^{(t)}$ the visited state at time t, we sample a candidate $Y^c \sim P(Y^{(t)}, \cdot)$ where we recall that $P(Y^{(t)}, \cdot)$ is the uniform distribution over the neighbours of $Y^{(t)}$, i.e. the states of the hypercube $\{0,1\}^N$ that differ from $Y^{(t)}$ in exactly one coordinate. Instead of accepting the transition towards the candidate state Y^c if

$$1 - \exp(-\frac{\Delta\mathcal{E}}{\mathrm{T}_t}) \leq U_t,$$

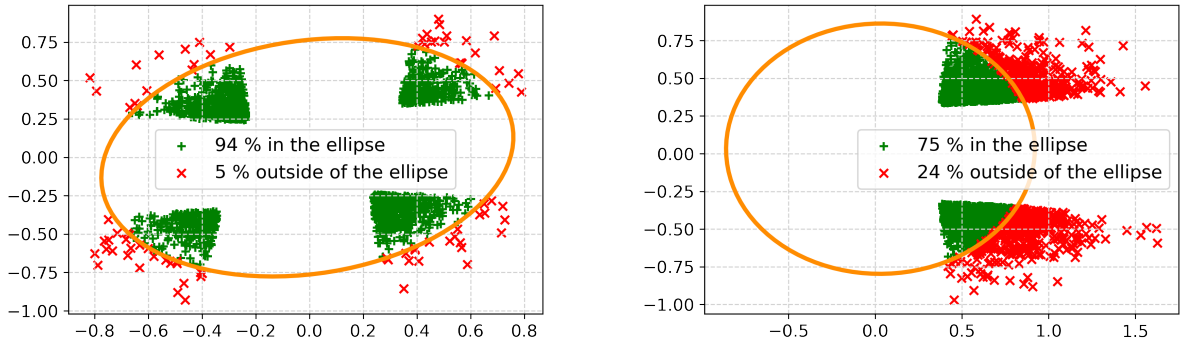
where $U_t \sim \mathcal{U}([0,1])$ and $\Delta\mathcal{E} := \mathcal{E}(Y^c) - \mathcal{E}(Y^{(t)})$, we decide to set $Y^{(t+1)} \leftarrow Y^c$ if and only if

$$\min\left(1 - \exp(-\frac{\Delta\mathcal{E}}{\mathrm{T}_t})\,,\, 1 - \mathcal{E}(Y^{(t)})\right) \leq U_t.$$

The extra term $1 - \mathcal{E}(Y^{(t)})$ in the acceptance rate acts like a *repulsion force*. If the current state $Y^{(t)}$ does not belong to the selection event, the energy $\mathcal{E}(Y^{(t)})$ is strictly positive. Nevertheless, if the neighbours of $Y^{(t)}$ have an energy which is larger than $\mathcal{E}(Y^{(t)})$, the algorithm may get stuck at $Y^{(t)}$ for some time before exploring other regions of the hypercube. Thanks to the extra term $1 - \mathcal{E}(Y^{(t)})$, the acceptance rate is boosted whenever the current state is known to be outside of the selection event.

4.2.5 Visualization of the SIGLE procedure in the selected model

Figure 11 provides a visualization of the SIGLE procedure in the selected model. We consider a design matrix $\mathbf{X} \in \mathbb{R}^{100\times 5}$ with i.i.d. entries sampled according to a standard normal distribution. We consider the null hypothesis $\mathbb{H}_0$: "$\theta^* = \mathbf{0}$". In Figure 11.(a), we work under $\mathbb{H}_0$ and we choose a regularization parameter $\lambda = 7$ in order to have a selected support of size 2 to be able to visualize in the plane the SIGLE method in the selected model. We calibrate our testing procedure empirically and we see on Figure 11.(a) that 95% of the states sampled using the rejection sampling method fall into the orange ellipse, meaning that our test has level 5%. On Figure 11.(b), we consider a localized alternative by considering $\vartheta^* = [0.5\,,\,0\,,\,\ldots\,,\,0]$ and we choose $\lambda = 8$ in order to have $|M| = 2$. In this case, the number of states falling into the orange ellipse is less than 95% which means that we reject the null hypothesis.



(a) $\vartheta^* = \mathbf{0}$, $\lambda = 7$

(b) $\vartheta^* = [0.5, 0, \ldots, 0]$, $\lambda = 8$

Figure 11: The orange ellipse represents the set of parameter $\theta \in \mathbb{R}^s$ such that $\|\widetilde{G}_N^{-1/2} H_N(\widetilde{\theta})(\theta - \widetilde{\theta})\|_2^2 = q_{1-\alpha}$ where $q_{1-\alpha}$ is the empirical quantile of order $1-\alpha$ of the test statistic of the SIGLE procedure in the selected model under the null "$\theta^* = 0$". 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.

4.3 Discussion and final remarks

Calibration. Despite the method proposed by Taylor and Tibshirani [2018] lacks theoretical guarantees, our experiments have shown that it is most of the time correctly calibrated. The calibration of SIGLE requires to sample under the null, which makes the method computationally more heavy.

Power. Our experiments have shown that the *empirically calibrated* SIGLE procedures seem to be systematically more powerful compared to the approach from Taylor and Tibshirani [2018]. We would like to point out two main possible reasons explaining the lack of power of the PSI method from Taylor and Tibshirani [2018].

(*i*) We are tackling a simple hypothesis testing problem while the method proposed in Taylor and Tibshirani [2018] is more naturally suited to address composite testing problems (typically testing the nullity of a specific coordinate of θ^*). Note that the SIGLE methods cannot easily tackle single testing problems since the whole parameter π_0^* (resp. θ_0^* in the selected model) is need to estimate $\overline{G}_N(\pi_0^*)$ (resp. $\overline{G}_N(\pi_0^*)$ and $\overline{\theta}(\theta_0^*)$). When deriving their PSI method, Taylor and Tibshirani [2018] face a similar issue and propose to use a plug-in approach by remplacing the unknown parameter θ^* by the lasso solution. A similar plug-in approximation for SIGLE could be investigated and this research direction is left for future work.

(*ii*) The method proposed by Taylor and Tibshirani [2018] is motivated by non-rigorous computations that aim at characterizing the distribution of the debiased lasso solution $\underline{\theta}$ conditional on the selection event $E_M^{S_M}$ (we refer to Section A.1 for details). It is well-known that conditioning on both the active variables and the vector of dual signs can lead to less powerful testing procedures. This statement can be made rigorous through the concept of *leftover Fisher information* (see Section F or Fithian et al. [2014] for details). As summarized in Fithian et al. [2014], "*on average, the price of conditioning on the [signs] S_M – the price of selection – is the information S_M carries about θ^**". Roughly speaking, even if the observed vector of dual signs is very surprising under the null, the method from Taylor and Tibshirani [2018] will not reject the null hypothesis unless we are surprised anew by looking at $\underline{\theta}$. On the contrary, the SIGLE methods rely on the characterization of some test statistic conditional on E_M (without conditioning on the signs).
In the Linear LASSO, the same situation arises and in Lee et al. [2016], the authors proved that one can rely on the work done conditional on $E_M^{S_M}$ in order to derive a more powerful testing method (at least on average) at the price of an additional computational cost. In this case, the linear transformation of the response vector is not distributed as a Gaussian truncated to an interval (as conditional on $E_M^{S_M}$) but is now a truncated Gaussian with a truncation set being a union of intervals. One important remark is that contrary to the Linear LASSO, the method from Taylor and Tibshirani [2018] cannot be easily adapted to get more power by working directly on E_M. The reason is that the test statistic itself depends on the vector of dual signs S_M (and not only the bounds of the truncation interval).

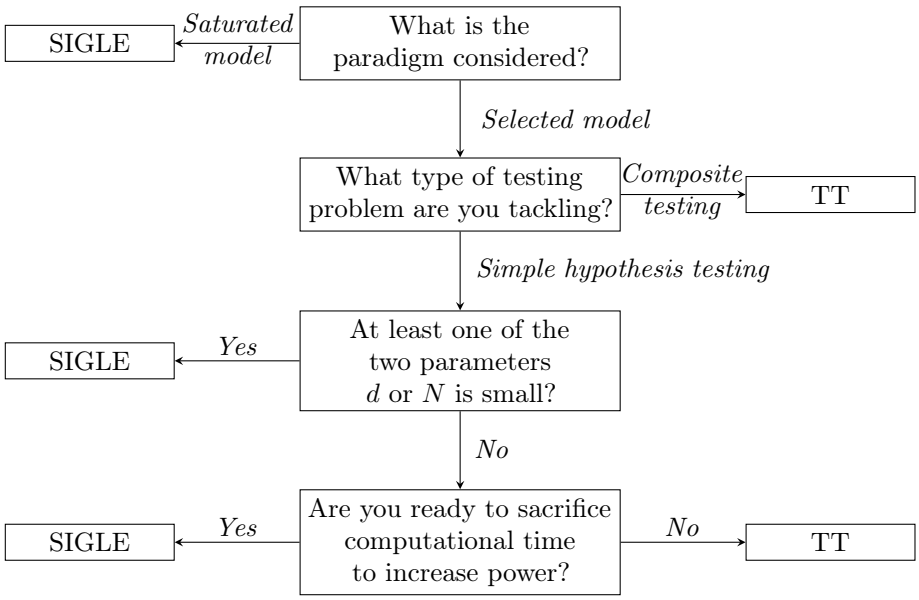


Figure 12: Choosing the PSI testing method that matches your setting: SIGLE (this paper) or TT (from Taylor and Tibshirani [2018]).

Conclusion. The SIGLE procedures require to use either the rejection sampling method or the SEI-SLR algorithm to estimate the matrix $\overline{G}_N(\pi_0^*)$ (and the parameter $\overline{\theta}(\theta_0^*)$ in the selected model) and to estimate the parameter $w_{N,1-\alpha}$ needed to define the rejection region. This sampling step is the main computational burden of the SIGLE procedures. On the contrary, the approach of Taylor and Tibshirani [2018] does not require such sampling stage and only requires to compute the bounds of the truncation interval of the distribution the $\eta^\top \underline{\theta}$ for some fixed vector $\eta \in \mathbb{R}^s$ under the null.
Figure 12 summarizes the main differences between the methods proposed in this paper and the one from Taylor and Tibshirani [2018] and provides an easy way to select the best method for a given setting. This organizational chart stresses that when d is small, the rejection sampling method allows to efficiently sample states from the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ while when N is small, the SEI-SLR algorithm allows to efficiently sample states uniformly distributed on E_M. In both cases, the SIGLE methods can be used with a small computational time and they should be preferred to get more powerful methods.

5 Conditional Central Limit Theorems for SLR

5.1 Preliminaries

Before presenting our conditional CLTs, let us present 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 $N \in \mathbb{N}$,

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

Remark. Note that the latter assumption holds in particular if the matrices $\big(\mathbf{X}^{(N)}/\sqrt{N}\big)_{N\geq 1}$ 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 5.2, we start by presenting our first CLT for $[\mathbf{X}_M^{(N)}]^\top Y$ where Y is distributed according to $\overline{\mathbb{P}}_{\pi^*}^{(N)}$. Thereafter, we prove in Section 5.3 a CLT for the conditional unpenalized MLE $\widehat{\theta}$ working with the design $\mathbf{X}_M^{(N)}$ (see Eq.(12)).

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 $\overline{\mathbb{P}}_{\pi^*}^{(N)}$. 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.

$$\begin{array}{cccccc}
\xi_{1,1} & & & & & \\
\xi_{1,2} & \xi_{2,2} & & & & \\
\xi_{1,3} & \xi_{2,3} & \xi_{3,3} & & & \\
\cdots & \cdots & \cdots & \cdots & & \\
\xi_{1,N} & \xi_{2,N} & \xi_{3,N} & \cdots & \xi_{N,N} & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array}$$

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 5.2 and 5.3 in Section 5.4.

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, ϑ^*, $\overline{\mathbb{P}}_{\pi^*}, \dots$. Nevertheless, let us stress again that the integer s is fixed and does not depend on N in this paper.

5.2 A conditional CLT for the saturated model

We aim at providing a simple hypothesis 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 5.1. 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

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

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 $(\overline{\sigma}^{\pi^*})^2 := \overline{\pi}^{\pi^*} \odot (1 - \overline{\pi}^{\pi^*})$.

5.3 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 when Y is distributed as $\mathbb{P}_{\theta^*}$. When the distribution of Y is $\overline{\mathbb{P}}_{\theta^*}$, we prove in Section E.5 a weaker counterpart of this result showing that for N large enough, the MLE exists with high probability.

We aim at providing a simple hypothesis 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 $\mathbb{P}_{\theta^*}$,

$$u^\top [H_N(\theta^*)]^{1/2}(\widehat{\theta} - \theta^*) \xrightarrow[N\to+\infty]{(d)} \mathcal{N}(0,1), \tag{22}$$

where u is a unit s-vector and where

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

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

In the following, we will consider the natural counterpart of the Fisher information matrix $H_N(\theta^*)$ 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.(22) by replacing respectively θ^* and $H_N(\theta^*)^{1/2}$ by $\overline{\theta}(\theta^*)$ (cf. Eq.(13)) and $[\overline{G}_N(\theta^*)]^{-1/2}H_N(\overline{\theta}(\theta^*))$. 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 5.1. 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,

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

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 E.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 for N large enough, the following holds with high probability: the MLE $\widehat{\theta}$ exists and is 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 E.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 $\widehat{\theta}$.

5.4 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 $\mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M = H_N(\theta^*)$ and $\overline{G}_N^{-1} = [H_N(\theta^*)]^{-1}$, 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 $[N]$.

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.

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

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

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.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

H. Zhang. A note on "mle in logistic regression with a diverging dimension", 2018. URL https://arxiv.org/abs/1801.08898.

Guidelines for the Appendix.

- **Section A: Regularization bias and conditional MLE.**

  In this first section of the Appendix, we shed light on the difference between SIGLE and the work of Taylor and Tibshirani [2018]. Both methods have already been compared on the practical side in Section 4. In Section A, we take a step back to understand the different paradigms considered in these two approaches. We describe the strengths and drawbacks of both methods, highlighting the fact that the method of Taylor and Tibshirani [2018] rely on non rigorous computations while SIGLE can be proved (see Section B) to be asymptotically valid under the set of assumptions presented in Section 5.4.

- **Section B: Theoretical guarantees for SIGLE in SLR.**

  In this section, we show how the conditional CLTs of Section 5 can be used to prove that the SIGLE methods are asymptotically correctly calibrated when the restrictive conditions of Section 5.4 are satisfied.

- **Section C: Confidence region.**

  Following the spirit of the previous section, we make use of the conditional CLTs presented in Section 5 to show how one can get confidence region using SIGLE.

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

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

- **Section E: Proofs.**

  We provide all the proofs of the theoretical results presented in this paper.

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

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

A 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 A.1.

- *SIGLE : 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.

A.1 Selective inference through debiasing

The idea behind the method proposed by Taylor and Tibshirani [2018] is that we need two key elements to deploy 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

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

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

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

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

$$\mathbf{X}_M^\top W \{z - \mathbf{X}_M \underline{\theta}\} = 0. \tag{23}$$

If one replaces W and z in Eq.(23) by $W(\underline{\vartheta})$ and $z(\underline{\vartheta})$ (with the obvious notation that $\underline{\vartheta}_M = \underline{\theta}$ and $\underline{\vartheta}_{-M} = 0$), Eq.(23) 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 plug-in 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.

A.2 Discussion

Duality between SIGLE and debiasing approaches. Oversimplifying 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. Figure 13 provides a visualization of these two different perspectives that can be used for PSI.

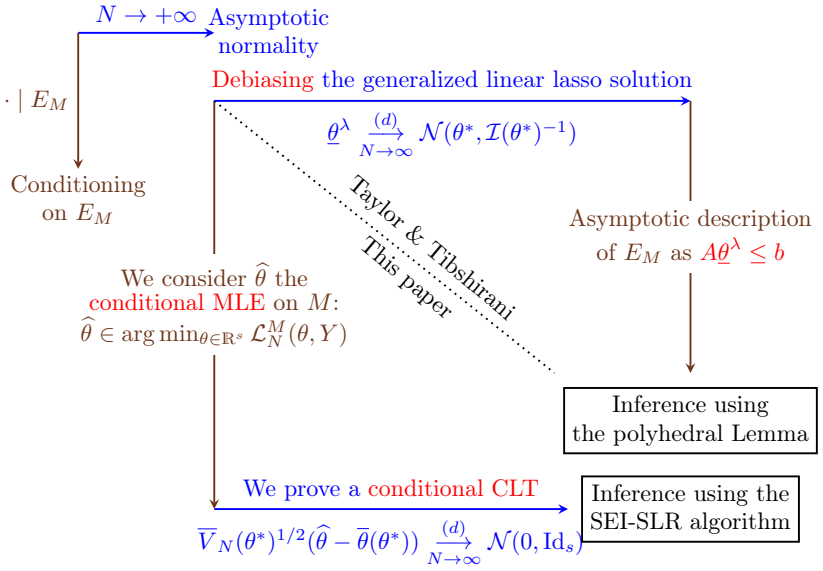


Figure 13: Duality between SIGLE and debiasing approaches.

Comprehensive comparison between SIGLE and 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 *simple hypothesis* 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 (see Section B) to be asymptotically valid under some set of assumptions that we discuss in details in Section 5.4. Table 6 sums up this comparison.

	Taylor and Tibshirani [2018]	SIGLE (this paper)
Selected model	✓	✓
Saturated model	✗	✓
Hypotheses tested in the selected model	Composite: $\theta_j^* = [\theta_0^*]_j$ for some j	Simple: $\theta^* = \theta_0^*$
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 $\alpha_N = \omega(N^{1/2})$.
Low computational cost	✓	✗

Table 6: Comparison between SIGLE and Taylor and Tibshirani [2018].

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.

In Section B, we show our the conditional CLTs provided in Section 5 can be used to derive theoretical guarantees for the SIGLE procedures under the restrictive assumptions given in Theorems 2 and 3.

B Theoretical guarantees for SIGLE in SLR

In this section, we make use of the conditional CLTs presented in Section 5 to prove that the SIGLE methods are asymptotically correctly calibrated when the assumptions of Section 5 are satisfied. Let us stress that this section is not of practical interest for two main reasons. First, the condition under which the theoretical guarantees presented in this section hold are restrictive and correspond to the ones usually considered in the literature when analyzing the asymptotic properties of the MLE in high dimensions. Second, making use of the conditional CLTs of Section 5 does not allow us to bypass the computational burden of sampling from the conditional distribution. Indeed, to get the SIGLE statistics, one still need to compute quantities such that $\overline{G}_N(\theta_0^*)$ (resp. $\overline{G}_N(\pi_0^*)$) or $\overline{\theta}(\theta_0^*)$ (resp. $\overline{\pi}^{\pi_0^*}$). Since the distribution of the observations conditional to the selection event has no closed form expression, we still need to use sampling methods such as the SEI-SLR algorithm presented in Section 3 to estimate the SIGLE statistics.

In this section, we consider the notations and assumptions described at the beginning of Section 5 and that rely on a system of triangular arrays.

B.1 In the selected model

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

$$\mathbb{H}_0 \ : \ \{\theta^* = \theta_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\theta^* \neq \theta_0^*\}\,. \tag{24}$$

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

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

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.(16)) 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 simple hypothesis testing procedure for the test (20). 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 3) 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.(16). This leads to the testing procedure presented in Proposition 6, whose proof is postponed to Section E.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 3. Let us denote*

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

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

Then the SIGLE 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,

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

B.2 In the saturated model

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

$$\mathbb{H}_0 \ : \ \{\pi^* = \pi_0^*\} \quad \text{and} \quad \mathbb{H}_1 \ : \ \{\pi^* \neq \pi_0^*\}. \tag{25}$$

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 simple hypothesis testing procedure for the test (19). 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 3) 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 7, whose proof is strictly analogous to the one of Proposition 6.

Proposition 7. *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 3. Let us denote*

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

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 SIGLE 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,

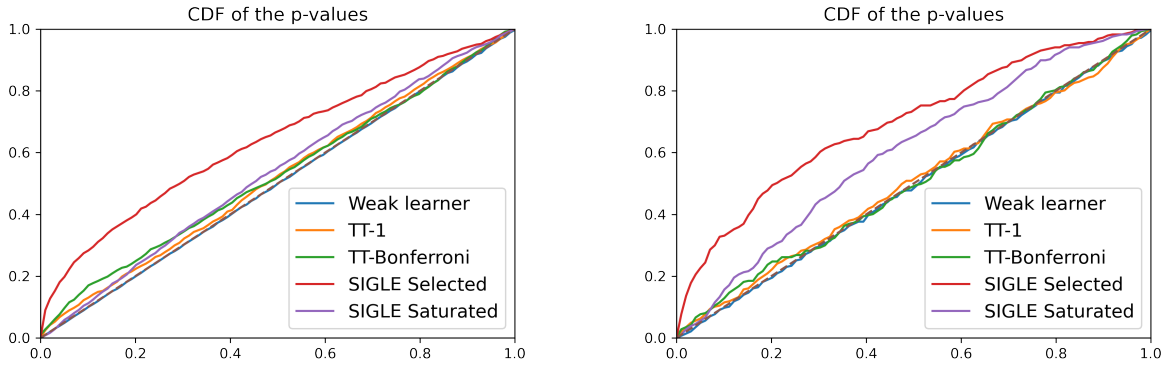
$$\mathbb{P}\Big(\bigcup_{T_N\in\mathbb{N}} \bigcap_{T\geq T_N} \{\zeta_{N,T} \leq \alpha + \epsilon\}\Big) = 1.$$

B.3 Calibration of SIGLE

In the main paper, we have clearly stated that SIGLE procedures are calibrated by sampling under the null using the SEI-SLR algorithm or the rejection sampling method. In this section, we conduct some experiments to study the distribution under the null of the p-values of the SIGLE methods when we calibrate the tests by using the theoretical quantile given by Proposition 6 and Proposition 7.

A correct calibration under weak dependence. Our experiments have shown that calibrating the SIGLE procedures using our conditional CLTs from Section 5 can lead to anti-conservative tests. This undesirable property was still observed when we conducted experiments with large values for N (typically $N = 30,000$). Based on our extensive simulations, we strongly believe that our conditional CLTs hold when the entries of the response vector $Y \sim \overline{\mathbb{P}}_{\theta^*}$ are weakly dependent. To illustrate our conclusions, we conducted simulations with different regularization parameters λ using the Setting 1 (cf. Table 5). In the first situation, a small regularization parameter is chosen (namely $\lambda = 0.1$), leading to select 9 out of the 10 features. In the second situation, we choose $\lambda = 1$ leading to a set of active variables of size 8. Figure 14 shows that for $\lambda = 0.1$, SIGLE in the saturated model is correctly calibrated while SIGLE in the selected model is anti-conservative. When λ is increased to 0.5, we see that the SIGLE procedure in both the selected and the saturated model is anti-conservative.

Despite the use the conditional CLTs for calibration, one still needs to sample under the null. Let us point out that calibrating the SIGLE procedure using the conditional CLTs from Section 5 does not exempt us from sampling states using the rejection method or the SEI-SLR algorithm since we need to estimate $\overline{G}_N(\pi_0^*)$ (and $\overline{\theta}(\theta_0^*)$ in the selected model).



(a) $\lambda = 0.1$.

(b) $\lambda = 0.5$.

Figure 14: CDF of the p-values of the different testing procedures under the global null with the **Setting 1** (cf. Table 5) for different regularization parameters λ.

C Confidence region

C.1 Asymptotic confidence region in the selected model

C.1.1 Main result

In the previous section, we proved that the MLE $\widehat{\theta}$ satisfies a CLT with a centering vector that is not the parameter of interest 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 8 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 8 can be found in Section E.8.

Proposition 8. *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 8, 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 8 involves two constants (namely κ and $\sigma^{\overline{\theta}}$) that cannot be *a priori* easily computed in practice.
Proposition 8 proves that when N is large enough, the size of our confidence region is driven by the distance $\|\overline{\theta}(\theta^\bigstar) - \widehat{\theta}\|_2$. This remark motivates us to choose $\theta^\bigstar$ among the minimizers of the function

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

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

C.1.2 Simulations

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

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

where $(Y^{(t)})_{t \geq 1}$ is the sequence generated from the SEI-SLR algorithm (see Algorithm 3). We train our network using stochastic gradient descent with learning rate 0.01 and 500 epochs. At each epoch, we feed to the network the inputs $(\widetilde{\theta}(\theta_i))_{i \in [n_{train}]}$ with the corresponding target values $(\theta_i)_{i \in [n_{train}]}$. We then compute our estimate $\theta^\bigstar$ of θ^* by taking the output of our network when taking as input the unpenalized MLE $\widehat{\theta}$ using the design $\mathbf{X}_M$ (cf. Eq.(12)). Figure 15 illustrates the result obtained from this deep learning approach. We keep the experiment settings of Section 4.2.5 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 $M = \{1, 2\}$.

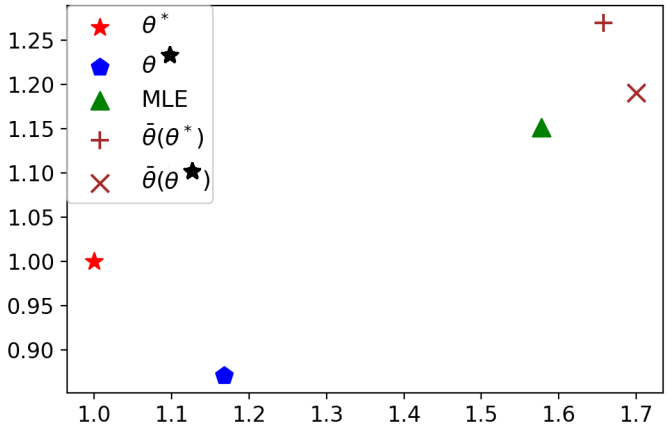


Figure 15: Visualization of the results obtained using our deep learning approach to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). $\theta^\bigstar$ corresponds to the output of the neural network when feeding as input the MLE $\widehat{\theta}$ (the green triangle). We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

Gradient descent method As shown in the proof of the expression of Proposition 8 (cf. Eq.(48)), it holds

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

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

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

Hence,

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

Given some 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 $H_N(\overline{\theta}(\theta))$).

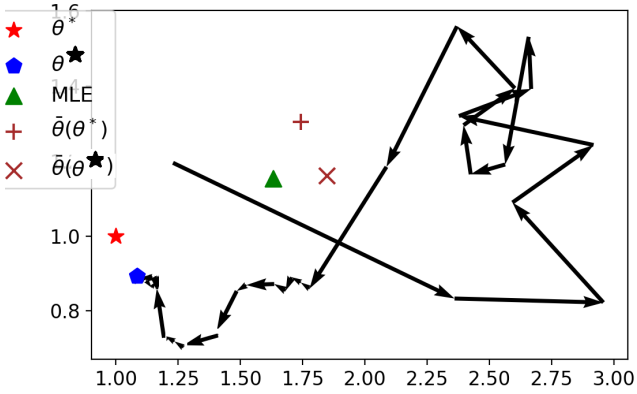


Figure 16: Visualization of our gradient descent procedure to compute an estimate $\theta^\bigstar$ (the blue hexagone) of θ^* (the red star). The MLE $\widehat{\theta}$ is the green triangle. We also plot the parameter $\overline{\theta}(\theta^*)$ (the brown plus) and $\overline{\theta}(\theta^\bigstar)$ (the brown cross).

C.2 Asymptotic confidence region in the saturated model

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 C.1, we propose here to build an asymptotic confidence region for π^*. The proof of Proposition 9 is postponed to Section E.9.

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

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

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

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

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

Remarks.

- Analogously to Section C.1, Proposition 9 motivates us to choose $\pi^\star$ among the minimizers of the function
$$M : \pi \mapsto \|\mathbf{X}_M^\top \overline{\pi}^\pi - \mathbf{X}_M^\top Y\|_2^2.$$
As mentioned in the Section C.1, one can rely for example on a deep learning or a gradient descent method in order to reach a local minimum $\pi^\star$ for M.

- The term $\|\mathrm{Proj}_{\mathbf{X}_M}^\perp(\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star})\|_2$ arising in the confidence region from Proposition 9 illustrates that our conditional CLT from Theorem 2 holds on $\mathbf{X}_M^\top Y$ and that we do not control what occurs in the orthogonal complement of the span of the columns of $\mathbf{X}_M$. Nevertheless, let us comment informally our result in the case where $E_M = \{0,1\}^N$ (meaning that there is no conditioning) and where ϑ^* is close to 0 (meaning that π^* is close to $\mathbf{1}_N/2$). In this framework, $\overline{\Gamma}^\pi = \mathrm{Diag}(\pi \odot (1-\pi))$ is close to $\frac{1}{4}\mathrm{Id}_N$ for π in a small neighbourhood around $\mathbf{1}_N/2$. Hence, we get that κ is approximately $\frac{1}{4}$. Since it also holds that $\overline{\pi}^{\pi^*} - \overline{\pi}^{\pi^\star} = \pi^* - \pi^\star$ (since $E_M = \{0,1\}^N$), we obtain from Proposition 9 that a CR for $\mathrm{Proj}_{\mathbf{X}_M}\pi^*$ with asymptotic coverage $1-\alpha$ is
$$\|\mathrm{Proj}_{\mathbf{X}_M}(\pi^* - \pi^\star)\|_2 \leq \|\mathrm{Proj}_{\mathbf{X}_M}(Y - \overline{\pi}^{\pi^\star})\|_2 + Cc^{-1}\sqrt{\chi^2_{s,1-\alpha}}.$$

D Side notes about SIGLE

D.1 SIGLE for a misspecified model from the start

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

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

D.1.1 SIGLE in the selected model

In the case of a misspecified model from the start, the assumption made to be in the selected model is
$$\sigma^{-1}(\mathbb{E}[Y]) \in \mathrm{Im}(\mathbf{X}_M),$$
where the expectation is taken with respect to $\mathbb{P}$. We define
$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^s} \overline{\mathbb{E}}\big[-\log \overline{\mathbb{P}}_\theta(Y)\big]. \tag{26}$$
$\mathbb{P}_{\theta^*}$ can be understood as the probability distribution belonging to the GLM family with design matrix $\mathbf{X}_M$ leading to the conditional distribution $\overline{\mathbb{P}}_{\theta^*}$ that is the closest possible to $\overline{\mathbb{P}}$. More precisely, for any GLM distribution $\mathbb{P}_\theta$, $\theta \in \mathbb{R}^s$, we have
$$\mathrm{KL}(\overline{\mathbb{P}} \mid \overline{\mathbb{P}}_\theta) \geq \mathrm{KL}(\overline{\mathbb{P}} \mid \overline{\mathbb{P}}_{\theta^*}).$$

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

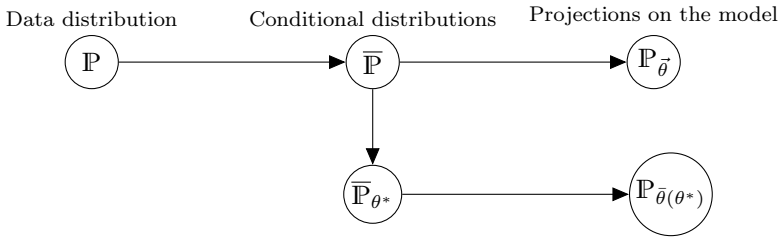


Figure 17: Visualizations of all distributions that we consider if the model is a priori not necessarily well-specified from the start.

1. If the model was well-specified initially, this means that there exists some $\vartheta^* \in \mathbb{R}^d$ such that $\mathbb{P} = \mathbb{P}_{\vartheta^*}$ and thus $\overline{\mathbb{P}} \equiv \overline{\mathbb{P}}_{\vartheta^*_M}$ (in the selected model). Namely, the null hypothesis is true for at least one parameter vector $\theta_0^* \in \mathbb{R}^s$.
2. If the model was not well-specified initially but the null is true for some $\theta_0^* \in \mathbb{R}^s$, this means that by conditioning on the selection event, we lost the information regarding the fact that the model was misspecified initially.
3. If the model was not well-specified initially and the null is false for any θ_0^*, this means that $\overline{\mathbb{P}}$ still carries the information of the initial model misspecificity.

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

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

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

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

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

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

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

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

D.2 Inverting the first order optimality condition

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

- a different view on Ψ using tools from convex analysis,
- the practical methods we use to compute quantities involving Ψ in our simple hypothesis testing method in the selected model.

D.2.1 SIGLE through the lens of convex analysis

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

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

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

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

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

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

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

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

We deduce that

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

Similarly, using Eq.(16) we get

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

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

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

$$\begin{aligned}\mathcal{L}^*_{N,0,M}(\rho) &= \sup_{\theta\in\mathbb{R}^s}\{\langle\rho,\theta\rangle - \mathcal{L}_{N,0,M}(\theta)\}\\ &= \sup_{\theta\in\mathbb{R}^s}\sum_{i=1}^N\{\rho_i\theta_i - \xi(\mathbf{X}_{i,M}\theta)\}\\ &= (\sum_{i=1}^N f_i)^*(\rho)\\ &= (f_1^*\square\cdots\square f_N^*)(\rho)\\ &:= \min_{\rho=\rho^{(1)}+\cdots+\rho^{(N)}}\left\{f_1^*(\rho^{(1)})+\cdots+f_N^*(\rho^{(N)})\right\},\end{aligned} \tag{27}$$

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

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

Using Lemma 2, we obtain that $\mathcal{L}^*_{N,0,M}(\rho)$ is the minimal entropy obtained among the vectors of probabilities $\pi\in(0,1)^N$ satisfying $\rho = \mathbf{X}_M^\top\pi$.

Lemma 2. *The inf-convolution in Eq.(27) is attained for $(\rho^{(i)})_{i\in[N]}\in(\mathbb{R}^s)^N$ such that for any $i\in[N]$, $\rho^{(i)} = \pi_i\mathbf{X}_{i,M}$ for some $\pi_i\in(0,1)$. Moreover,*

$$\mathcal{L}^*_{N,0,M}(\rho) = \min_{\pi\in(0,1)^N \ s.t.\ \rho=\mathbf{X}_M^\top\pi} H(\pi), \tag{28}$$

where

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

Proof of Lemma 2.

- **The inf-convolution in Eq.(27) is attained.**
  First, we know from [Laurent, 1972, Theorem 6.5.8] that the minimum in the inf-convolution of Eq.(27) is attained.

- **$\rho^{(i)}$ in Eq.(27) can be chosen in the span of $\mathbf{X}_{i,M}$.**
  Let us consider $i \in [N]$ and some $\rho^{(i)} \in \mathbb{R}^s$. Let us assume by contradiction that $\rho^{(i)} \notin \mathrm{Span}(\mathbf{X}_{i,M})$. Then considering

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

  we have

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

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

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

  We deduce that in the inf-convolution of Eq.(27), we can consider that for any $i \in [N]$, $\rho^{(i)} = \pi_i \mathbf{X}_{i,M}$ for some $\pi_i \in \mathbb{R}$.

- **$\rho^{(i)}$ in Eq.(27) can be chosen as $\pi_i\mathbf{X}_{i,M}$ with $\pi_i \in (0,1)$.**
  We have already proved that $\rho^{(i)}$ in Eq.(27) can be chosen as $\rho^{(i)} = \pi_i\mathbf{X}_{i,M}$. It holds

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

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

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

Since in Eq.(27) we aim a reaching a minimum, we deduce from these computations that one can restrict $\rho^{(i)}$ to be of the form $\pi_i\mathbf{X}_{i,M}$ with $\pi_i \in (0,1)$. □

Interpretation of $\Psi(\rho)$. Lemma 2 shows that $\mathcal{L}^*_{N,0,M}(\rho)$ is the minimum entropy of a population characterized by N binary features with the constraint that we have some information on the population given

by $\rho \in \mathbb{R}^s$. We assume that ρ depends linearly on the proportion of the population with the different features, namely

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

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

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

D.2.2 Practical implementation of SIGLE in the selected model

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

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

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

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

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

and we end up with two possible cases:

1. Either it holds
$$\mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta^c) = \mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}, \tag{29}$$
which is equivalent to $\overline{\theta}(\theta^*) = \theta^c$ (see Eq.(16)). In this case, we output θ^c. Note that this situation occurs in particular when
$$\sigma^{-1}(\mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}) \in \mathrm{Im}(X_M),$$
which can be understood as a conditional selected model-type assumption.

2. Or Eq.(29) does not hold and we consider a gradient descent approach using as warm start the vector θ^c to minimize the map
$$G : \theta \mapsto \|\mathbf{X}_M^\top \sigma(\mathbf{X}_M \theta) - \mathbf{X}_M^\top \overline{\pi}^{\theta_0^*}\|_2^2.$$
Note that the gradient of G at $\theta \in \mathbb{R}^s$ is given by
$$\nabla G(\theta) = 2\mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M \theta)) \mathbf{X}_M \mathbf{X}_M^\top \left(\sigma(\mathbf{X}_M \theta) - \overline{\pi}^{\theta_0^*}\right),$$

and satisfies

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

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

Our method is summarized with Algorithm 5.

Algorithm 5 Computing $\overline{\theta}(\theta_0^*)$

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

E Proofs

E.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

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

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).

E.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 exists $\theta \in \Theta_M$ satisfying

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

E.3 Proof of Proposition 3

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

$$\begin{aligned}
0 &= \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta) - \mathbf{X}_M^\top \sigma(\mathbf{X}_M\theta') \\
&= \Xi(\theta) - \Xi(\theta') \\
&= \int_0^1 \nabla\Xi(\theta t + (1-t)\theta') \cdot (\theta - \theta') dt \\
&= \int_0^1 \mathbf{X}_M^\top \operatorname{Diag}\left[\sigma'(\mathbf{X}_M\theta t + (1-t)\mathbf{X}_M\theta')\right] \mathbf{X}_M(\theta - \theta') dt \\
&= \mathbf{X}_M^\top \underbrace{\left(\int_0^1 \operatorname{Diag}\left[\sigma'(\mathbf{X}_M\theta t + (1-t)\mathbf{X}_M\theta')\right] dt\right)}_{=:D} \mathbf{X}_M(\theta - \theta'). \qquad (30)
\end{aligned}$$

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.(30) gives that $\theta - \theta' \in \operatorname{Ker}(\mathbf{X}_M^\top D \mathbf{X}_M)$ 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 $\nabla_\theta\Xi(\theta) = \mathbf{X}_M^\top \operatorname{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M$ which is invertible at any $\theta \in \Theta_M$ under the assumptions of Proposition 3. Hence the global inversion theorem gives Proposition 3.

E.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

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

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 3. *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 3, the Cauchy-Schwarz inequality and Jensen's inequalities, we get,

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

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

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

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 $K, s, c, \overline{\sigma}^2_{\min}$ 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 $\|\overline{G}_N^{-1}\|_F \leq \sqrt{s}\|\overline{G}_N^{-1}\| \lesssim N^{-1}$ (since $\overline{G}_N^{-1}$ has rank s, see Section 5.1). 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

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

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

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

E.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 1.4), the second equality comes from Eq.(16) 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 E.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 exist 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 $\theta \in \mathbb{R}^s$,

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

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

$$\begin{aligned} & \overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \in F(\mathcal{N}_N(\overline{\theta}, \delta)) \\ \Leftrightarrow \quad & \overline{G}_N^{-1/2}(\underbrace{\mathbf{X}_M^\top \overline{\pi}^{\theta^*}}_{=\mathbf{X}_M^\top \pi^{\overline{\theta}}} - \mathbf{X}_M^\top Y) \in F(\mathcal{N}_N(\overline{\theta}, \delta)). \end{aligned} \tag{31}$$

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.(31) to hold is to check that on the event E_N it holds

$$\|\overline{G}_N^{-1/2}L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 < \inf_{\theta \in \partial\mathcal{N}_N(\overline{\theta}, \delta)} \|F(\theta)\|_2, \tag{32}$$

where $\partial\mathcal{N}_N(\overline{\theta}, \delta) := \{\theta \in \mathbb{R}^s \,|\, \|\overline{G}_N^{1/2}(\theta - \overline{\theta})\|_2 = \delta\}$. This sufficient condition is a direct consequence of Lemma 4 and Figure 18 gives a visualization of our proof strategy.

Lemma 4. *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 $\mathcal{C}^1$-diffeomorphism, f is in particular a homeomorphism, and as such, it preserves the topological structures. □

Let $\epsilon > 0$ and let us consider

$$\delta := \frac{\mathfrak{K}^{1/2}}{\epsilon^{1/2} 2C^{-1}c\overline{\sigma}^2_{\min}}, \tag{33}$$

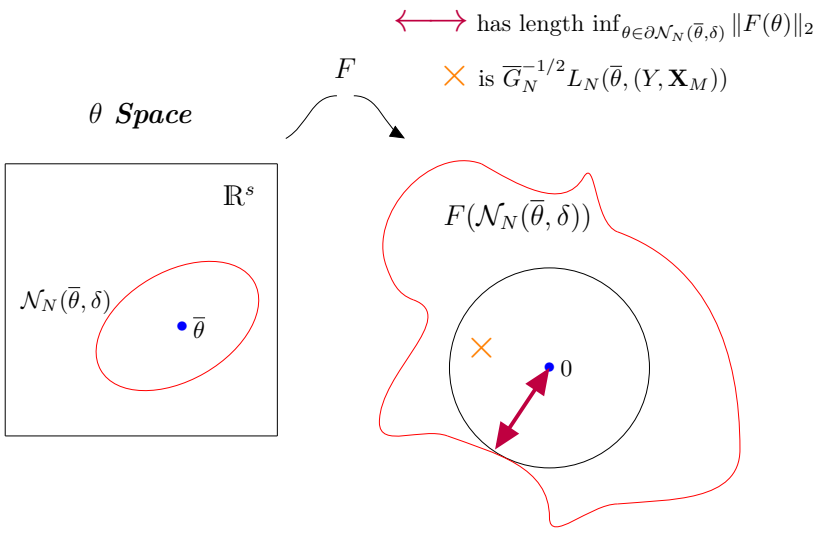


Figure 18: Visualization support for the proof of the existence of the MLE with large probability in a neighbourhood of $\overline{\theta}$. We show that with large probability, the orange cross is in the black circle (*i.e.*, Eq.(32) holds) which implies that the orange cross belongs to $F(\mathcal{N}_N(\overline{\theta},\delta))$ (*i.e.*, Eq.(31) holds). The MLE is then defined as $\widehat{\theta} = F^{-1}(\overline{G}_N^{-1/2}L_N(\overline{\theta},(Y,\mathbf{X}_M)) \in \mathcal{N}_N(\overline{\theta},\delta)$.

(the reason of this choice will become clear with Eq.(38)). Let us first notice that for any $\theta \in \mathbb{R}^s$,

$$L_N(\overline{\theta},(Y,\mathbf{X}_M)) - L_N(\theta,(Y,\mathbf{X}_M)) \tag{34}$$
$$= \mathbf{X}_M^\top(\pi^{\overline{\theta}} - \pi^\theta) \tag{35}$$
$$= \underbrace{\int_0^1 H_N(t\overline{\theta}+(1-t)\theta)dt}_{=:Q_N(\theta)}\,(\overline{\theta}-\theta), \tag{36}$$

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 $\mathbf{X}_M\mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = H_N(\theta)$.

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,

$$
\begin{aligned}
&\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\|F(\theta)\|_2\\
&=\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\|\overline{G}_N^{-1/2}Q_N(\theta)(\theta-\overline{\theta})\|_2\quad\text{(using Eq.(36))}\\
&=\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\|\overline{G}_N^{-1/2}Q_N(\theta)(\theta-\overline{\theta})\|_2\times\frac{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}\\
&\geq\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\frac{(\theta-\overline{\theta})^\top Q_N(\theta)(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}\quad\text{(using the Cauchy Schwarz's inequality)}\\
&=\delta\inf_{\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\frac{(\theta-\overline{\theta})^\top\overline{G}_N^{1/2}}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}\overline{G}_N^{-1/2}Q_N(\theta)\overline{G}_N^{-1/2}\frac{\overline{G}_N^{1/2}(\theta-\overline{\theta})}{\|\overline{G}_N^{1/2}(\theta-\overline{\theta})\|_2}\\
&\geq\delta\inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}e^\top\overline{G}_N^{-1/2}Q_N(\theta)\overline{G}_N^{-1/2}e\\
&=\delta\inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}e^\top\overline{G}_N^{-1/2}\int_0^1H_N(t\overline{\theta}+(1-t)\theta)dt\overline{G}_N^{-1/2}e\\
&=\delta\inf_{\|e\|_2=1,\theta\in\partial\mathcal{N}_N(\overline{\theta},\delta)}\int_0^1\left(e^\top\overline{G}_N^{-1/2}H_N(t\overline{\theta}+(1-t)\theta)\overline{G}_N^{-1/2}e\right)dt\\
&\geq\delta\inf_{\|e\|_2=1,\theta\in\mathcal{N}_N(\overline{\theta},\delta)}e^\top\overline{G}_N^{-1/2}H_N(\theta)\overline{G}_N^{-1/2}e\\
&\geq\delta\left\{\inf_{\|e\|_2=1}e^\top\overline{G}_N^{-1/2}H_N(\overline{\theta})\overline{G}_N^{-1/2}e-\mathcal{C}\frac{\delta}{N^{1/2}}\right\}=:\mathcal{I}_N(\delta,\overline{\theta}),
\end{aligned}
\tag{37}
$$

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 5 whose proof is postponed to Section E.6.

Lemma 5. *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*

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

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

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

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

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

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 $N \geq N_\delta$,

$$\begin{aligned}
&\overline{\mathbb{P}}_{\theta^*}(\|\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\|_2 \geq \mathcal{I}_N(\delta, \overline{\theta})) \\
&\leq \frac{\mathrm{Tr}\left[\overline{G}_N^{-1} \mathbf{X}_M^\top \overline{\Gamma}^{\theta^*} \mathbf{X}_M\right]}{\mathcal{I}_N(\delta, \overline{\theta})^2} \\
&< \frac{\mathfrak{K}}{\delta^2 (2C^{-1}c\overline{\sigma}^2_{\min})^2} \\
&\leq \epsilon,
\end{aligned} \tag{38}$$

where the last inequality comes from the choice of δ (see Eq.(33)). From Eq.(37) and Eq.(38), 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

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

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 $\mathbf{X}_M \mathrm{Diag}(\sigma'(\mathbf{X}_M\theta))\mathbf{X}_M = H_N(\theta)$. 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 $N \geq N_\delta$,

$$\begin{aligned}
&|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right]| \\
&= |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta}))]| \\
&\leq |\overline{\mathbb{E}}_{\theta^*}\left[\mathbb{1}_{E_N}\left\{h(\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta})) - h\big(\overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\big)\right\}\right]| + 2\|h\|_\infty \overline{\mathbb{P}}_{\theta^*}(E_N^c) \\
&\leq \overline{\mathbb{E}}_{\theta^*}\left[L_h \mathbb{1}_{E_N} \|\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}) - \overline{G}_N^{-1/2} Q_N(\widehat{\theta})(\overline{\theta} - \widehat{\theta})\|_2\right] + 2\|h\|_\infty \epsilon \\
&\leq L_h \overline{\mathbb{E}}_{\theta^*}\left[\mathbb{1}_{E_N} \|\overline{G}_N^{-1/2}(H_N(\overline{\theta}) - Q_N(\widehat{\theta}))\overline{G}_N^{-1/2}\| \|\overline{G}_N^{1/2}(\overline{\theta} - \widehat{\theta})\|_2\right] + 2\|h\|_\infty \epsilon \\
&\leq L_h \delta \sup_{\theta \in \mathcal{N}_N(\overline{\theta}, \delta)} \|\overline{G}_N^{-1/2}(H_N(\overline{\theta}) - Q_N(\theta))\overline{G}_N^{-1/2}\| + 2\|h\|_\infty \epsilon,
\end{aligned} \tag{39}$$

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,

$$\begin{aligned}
&\|\overline{G}_N^{-1/2}(H_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2}\| \\
&= \sup_{\|u\|_2 = 1} |u^\top \overline{G}_N^{-1/2}(H_N(\overline{\theta}) - Q_N(\theta'))\overline{G}_N^{-1/2} u| \\
&\leq \sup_{\|u\|_2 = 1} \int_0^1 \left|u^\top \overline{G}_N^{-1/2}(H_N(\overline{\theta}) - H_N(t\overline{\theta} + (1-t)\theta'))\overline{G}_N^{-1/2} u\right| dt \\
&\leq \sup_{\|u\|_2 = 1} \sup_{\theta \in \mathcal{N}_N(\overline{\theta}, \delta)} |u^\top \overline{G}_N^{-1/2}(H_N(\overline{\theta}) - H_N(\theta))\overline{G}_N^{-1/2} u| \\
&\leq \mathcal{C}\frac{\delta}{N^{1/2}},
\end{aligned} \tag{40}$$

where in the penultimate inequality we used the convexity of the set $\mathcal{N}_N(\overline{\theta}, \delta))$ and in the last inequality we used Lemma 5 (which is proved in Section E.6). Using Eq.(39) and Eq.(40), we deduce that for $G \sim \mathcal{N}(0, \mathrm{Id}_s)$ we have

$$\begin{aligned}
&|\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \mathbb{E}[h(G)]| \\
&\leq |\overline{\mathbb{E}}_{\theta^*}[h(\overline{G}_N^{-1/2} H_N(\overline{\theta})(\overline{\theta} - \widehat{\theta}))] - \overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right]| \\
&\quad + |\overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right] - \mathbb{E}[h(G)]| \\
&\leq L_h \delta \mathcal{C}\frac{\delta}{N^{1/2}} + 2\|h\|_\infty \epsilon + |\overline{\mathbb{E}}_{\theta^*}\left[h\big(\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M))\big)\right] - \mathbb{E}[h(G)]|.
\end{aligned} \tag{41}$$

The CLT from Theorem 2 states that

$$\overline{G}_N^{-1/2} L_N(\overline{\theta}, (Y, \mathbf{X}_M)) \underset{N\to\infty}{\overset{(d)}{\to}} \mathcal{N}(0, \mathrm{Id_s}),$$

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.(41) 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.

E.6 Proof of Lemma 5

Let us first recall that $H_N(\overline{\theta}) = \mathbf{X}_M^\top \mathrm{Diag}(\sigma'(\mathbf{X}_M \overline{\theta}))\mathbf{X}_M$ 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

$$\begin{aligned} H_N(\theta) - H_N(\overline{\theta}) &= \sum_{i=1}^N \mathbf{w}_i \left[\sigma'(\mathbf{w}_i^\top \theta) - \sigma'(\mathbf{w}_i^\top \overline{\theta})\right] \mathbf{w}_i^\top \\ &= \sum_{i=1}^N \mathbf{w}_i \underbrace{\int_0^1 \sigma''(t\mathbf{w}_i^\top \theta + (1-t)\mathbf{w}_i^\top \overline{\theta}) dt}_{=:H_i} \mathbf{w}_i^\top (\theta - \overline{\theta}) \mathbf{w}_i^\top . \end{aligned} \tag{42}$$

We get using Eq.(42) that for any unit vector $u \in \mathbb{R}^s$,

$$\begin{aligned} &|u^\top \overline{G}_N^{-1/2} (H_N(\theta) - H_N(\overline{\theta})) \overline{G}_N^{-1/2} u| \\ &= \left| \sum_{i=1}^N u^\top \overline{G}_N^{-1/2} \mathbf{w}_i H_i \mathbf{w}_i^\top (\theta - \overline{\theta}) \mathbf{w}_i^\top \overline{G}_N^{-1/2} u \right| \\ &= \left| \sum_{i=1}^N \mathbf{w}_i^\top (\theta - \overline{\theta}) \times u^\top \overline{G}_N^{-1/2} \mathbf{w}_i H_i \mathbf{w}_i^\top \overline{G}_N^{-1/2} u \right| \\ &= \left| \sum_{i=1}^N \mathbf{w}_i^\top (\theta - \overline{\theta}) \times H_i |\mathbf{w}_i^\top \overline{G}_N^{-1/2} u|^2 \right| \\ &\leq \max_{1 \leq j \leq N} |\mathbf{w}_j^\top (\theta - \overline{\theta})| \sum_{i=1}^N |H_i| |\mathbf{w}_i^\top \overline{G}_N^{-1/2} u|^2 \\ &= \max_{1 \leq j \leq N} |\mathbf{w}_j^\top (\theta - \overline{\theta})| \, \|\mathbf{H}^{1/2} \mathbf{X}_M^\top \overline{G}_N^{-1/2} u\|_2^2, \end{aligned} \tag{43}$$

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.(43). Using the assumption of the design matrix presented in Section 5.1 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 $i \in [N]$,

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

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

E.7 Proof of Proposition 6

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

$$\mathcal{E}_N := \{Z \in \{0,1\}^N \,|\, \mathbf{X}_M^\top Z \in \mathrm{Im}(\Xi)\}. \tag{44}$$

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 $\mathbb{H}_0$

$$\frac{\sum_{t=1}^T Y^{(t)} \mathbb{P}_{\theta_0^*}(Y^{(t)})}{\sum_{t=1}^T \mathbb{P}_{\theta_0^*}(Y^{(t)})} \underset{T\to\infty}{\to} \overline{\mathbb{E}}_{\theta_0^*}[Y] = \overline{\pi}^{\theta_0^*} \quad \mathbb{P}_1 - \text{almost surely.} \tag{45}$$

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.(45) 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.(16)). 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 $Z \in \mathcal{E}_N$,

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

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

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

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

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

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

$$\forall Z \in \mathcal{E}_N, \quad \mathbb{1}_{Z \in \widetilde{W}_N} \leq \mathbb{1}_{Z \in W_N(\alpha+\eta)}, \tag{46}$$

where

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

Recalling the definition of $\mathcal{E}_N$ from Eq.(44) and using the definitions of $W_N(\alpha+\eta)$ and $\widetilde{W}_N$, it also holds trivially

$$\forall Z \in \{0,1\}^N \backslash \mathcal{E}_N, \quad 0 = \mathbb{1}_{Z \in \widetilde{W}_N} \leq \mathbb{1}_{Z \in W_N(\alpha+\eta)} = 0. \tag{47}$$

Using both Eq.(46) and Eq.(47), 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,

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

E.8 Proof of Proposition 8

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 $k \in [s]$,

$$\begin{aligned}
\frac{\partial \overline{\pi}^\theta}{\partial \theta_k} &= \left(\sum_{w \in E_M} \mathbb{P}_\theta(w) \right)^{-2} \sum_{w,z \in E_M} \mathbb{P}_\theta(z) \mathbb{P}_\theta(w) z \left\{ L_N(\theta, (w, \mathbf{X}_M)) - L_N(\theta, (z, \mathbf{X}_M)) \right\}_k \\
&= \overline{\mathbb{E}}_\theta \left[Z \left\{ L_N(\theta, (W, \mathbf{X}_M)) - L_N(\theta, (Z, \mathbf{X}_M)) \right\}_k \right] \\
&= \overline{\mathbb{E}}_\theta \left[Z \left\{ \mathbf{X}_M^\top (Z - W) \right\}_k \right] \\
&= \overline{\Gamma}^\theta \mathbf{X}_{:, M[k]},
\end{aligned} \tag{48}$$

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

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

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

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

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

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

and that

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

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

E.9 Proof of Proposition 9

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

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

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

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

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

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

we get that

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

Hence we have that

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

Since Theorem 2 gives that

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

we deduce that the event

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

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

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

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

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

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

F Inference conditional on the signs

F.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

$$\nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, E_M) = \nabla^2_\vartheta \mathcal{L}_N(\vartheta, \widehat{S}_M(Y) \,|\, E_M) + \nabla^2_\vartheta \mathcal{L}_N(\vartheta, Y \,|\, \{E_M, \widehat{S}_M(Y)\}). \tag{49}$$

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.(49) 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

$$\begin{aligned}&\mathbb{P}(T(\pi^*) \in [L(Y,\widehat{S}_M(Y)), U(Y,\widehat{S}_M(Y))] \mid E_M) \\ &= \sum_{S_M\in\{\pm1\}^s} \mathbb{P}(\widehat{S}_M(Y) = S_M \,|\, E_M)\, \underbrace{\mathbb{P}(T(\pi^*) \in [L(Y,S_M), U(Y,S_M)] \mid E_M^{S_M})}_{=1-\alpha} \\ &= 1-\alpha.\end{aligned}$$

F.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 A, in Taylor and Tibshirani [2018] the authors claim the following asymptotic distribution

$$\underline{\theta} \sim \mathcal{N}(\vartheta^*_M, H_N(\vartheta^*_M)^{-1}), \tag{50}$$

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

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

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

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

Several problems arise at this point.

1. **Lack of theoretical guarantee due to the use of Monte-Carlo estimates.**
   The first problem is that both $\underline{\theta}$ and the selection event $\{\widehat{M}(Y) = M, \ \widehat{S}_M(Y) = S_M\}$ involve the unknown parameter ϑ^*_M through $H_N(\vartheta^*_M)$. Taylor and al. propose to use a Monte-Carlo estimate for $H_N(\vartheta^*_M)$ by replacing it with $H_N(\widehat{\theta^\lambda})$. Using this Monte-Carlo estimate, one can compute L and U such that
   $$F^{[\mathcal{V}^-_{S_M},\mathcal{V}^+_{S_M}]}_{L,\left[H_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = 1 - \frac{\alpha}{2} \quad \text{and} \quad F^{[\mathcal{V}^-_{S_M},\mathcal{V}^+_{S_M}]}_{U,\left[H_N(\vartheta^*_M)^{-1}\right]_{j,j}}(\underline{\theta}_j) = \frac{\alpha}{2}.$$
   Then, $[L,U]$ is claimed to be a confidence interval with (asymptotic) $(1-\alpha)$ coverage for $\vartheta^*_{M[j]}$ conditional on $\{\widehat{M}(Y) = M, \widehat{S}_M(Y) = S_M\}$, that is,
   $$\mathbb{P}(\vartheta^*_{M[j]} \in [L,U] \mid \widehat{M}(Y) = M, \ \widehat{S}_M(Y) = S_M) = 1 - \alpha.$$

2. **Their approach is not well suited to provide more powerful inference procedures by conditioning only on E_M.**
   In the linear model, Lee et al. [2016] also start by deriving a pivotal quantity by conditioning on both the selected variables and the vector of signs. However, in the context of linear regression, the vector of signs only appears in the threshold values $\mathcal{V}^-$ and $\mathcal{V}^+$. Hence, conditioning only on the selected variables $\{\widehat{M}(Y) = M\}$ simply reduces to take the union $\cup_{S_M\in\{\pm1\}^s}[\mathcal{V}^-_{S_M},\mathcal{V}^+_{S_M}]$ 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.