Biometrika (...), **...**, ..., *pp.* 1–21
Printed in Great Britain

On inference in high-dimensional logistic regression models with separated data

BY R. M. LEWIS AND H. S. BATTEY
Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7
rebecca.lewis15@imperial.ac.uk h.battey@imperial.ac.uk

SUMMARY

Direct use of the likelihood function typically produces severely biased estimates when the dimension of the parameter vector is large relative to the effective sample size. With linearly separable data generated from a logistic regression model, the log-likelihood function asymptotes and the maximum likelihood estimator does not exist. We show that an exact analysis for each regression coefficient produces half-infinite confidence sets for some parameters when the data are separable. Such conclusions are not vacuous, but an honest portrayal of the limitations of the data. Finite confidence sets are only achievable when additional, perhaps implicit, assumptions are made. Under a notional double-asymptotic regime in which the dimension of the logistic coefficient vector increases with the sample size, the present paper considers the implications of enforcing a natural constraint on the vector of logistic-transformed probabilities. We derive a relationship between the logistic coefficients and a notional parameter obtained as a probability limit of an ordinary least squares estimator. The latter exists even when the data are separable. Consistency is ascertained under weak conditions on the design matrix.

Some key words: Binary responses; Complete and quasi-complete separation; Conditioning; Regression; Sufficiency.

1. INTRODUCTION

1.1. *Background*

The analysis of binary response data commonly assumes a logistic regression model for which the distribution of response variables $Y_1, \ldots, Y_n$ is

$$\mathrm{pr}(Y_i = 1) = \frac{e^{x_i^T \beta^*}}{1 + e^{x_i^T \beta^*}}, \qquad \mathrm{pr}(Y_i = -1) = 1 - \mathrm{pr}(Y_i = 1) \tag{1}$$

for some unknown parameter $\beta^* \in \mathbb{R}^p$ and covariates $x_1, \ldots, x_n \in \mathbb{R}^p$, treated as fixed. This was proposed by Cox (1958) and is the unique model for binary data yielding the same simple sufficient statistics for the regression coefficients as in a normal-theory linear model. His exact conditional inference based on combinatorial calculations evades maximum likelihood fitting and simultaneously achieves relevance and elimination of nuisance parameters. See Chapter 4 of Cox (1970) or Mehta & Patel (1995) for a more explicit and general exposition than that of Cox (1958).

Motivated by high-dimensional models arising in modern scientific applications, notably genomics, there has been increased interest in theoretical treatments that allow for a notional double asymptotic regime $p, n \to \infty$. Even prior to the genomics applications, this setting interested Bartlett (1936, 1937), who used it to illustrate serious difficulties with maximum-likelihood-based approaches and advantages of using marginal and conditional likelihood, when available. The approach of Cox (1958) is in this vein. It has some practical limitations, notably a possible degeneracy of the problem when covariates are continuous, and also the computational difficulty associated with the combinatorial quantities involved, see section 3.1.

Perhaps for these reasons, the exact conditional analysis is not widely used. Instead, the prevailing approach to inference in logistic regression models, and more generally, is based on maximum likelihood estimation and asymptotic calculations local to the null hypothesis, following Wald (1950). It is well-known that the resulting estimates and confidence intervals are asymptotically calibrated when p is fixed and n is large. However there exist both practical and theoretical difficulties with this approach, particularly in high-dimensional regimes. The first is that the maximum likelihood estimator does not exist if and only if the data can be separated, that is, whenever the outcome-covariate pairs $(y_1, x_1^T)^T, \ldots, (y_n, x_n^T)^T$ are such that $y_i x_i^T \beta \geq 0$ for all i and for some non-zero $\beta \in \mathbb{R}^p$ (Albert & Anderson, 1984). For centred Gaussian covariates, Candès & Sur (2020) derived the liming probability that the data can be separated in terms of the relative dimension $p/n \to \kappa$ and a function of the signal strength. This probability converges to one when κ exceeds a threshold, illustrating the difficulties encountered in high-dimensional regimes.

Issues are also encountered when the maximum likelihood estimator exists. In the same limiting setting $p, n \to \infty$ with $p/n \to \kappa > 0$, Sur & Candès (2019) showed that the maximum likelihood estimator can be severely biased when the design matrix is treated as random with independent and identically distributed entries. They further showed that standard error estimates based on fixed-p maximum likelihood theory underestimate the true variability, and that the χ_1^2 limiting approximation to the distribution of likelihood ratio test statistic is poor. Related work is due to Zhao et al. (2020) who obtained similar results for Gaussian designs with arbitrary covariance structures. Similar ideas have been explored in more general models by Coolen et al. (2020), who sought to correct the average bias in the p maximum likelihood estimates using ideas from statistical physics, and Tang & Reid (2020), who clarified the extent to which classical higher-order inference based on the so called r^* statistic continues to hold under the $p, n \to \infty$ regime.

Motivated by the issues summarised above, this work clarifies the extent to which inference is possible in logistic regression models with separated data and proposes an alternative to maximum likelihood estimation valid for these settings. We begin by studying the exact conditional inference of Cox (1958), showing that in the presence of data separation, at least one of the exact conditional confidence intervals is of infinite length in at least one direction. The results are then extended to arbitrary exact confidence sets. Such conclusions are not vacuous and are the best that could be hoped for without further assumptions or data. In high-dimensional regimes, however, it is common to make further restrictions that allow for consistent estimation of the unknown regression parameter. We introduce an approach based on least squares that is consistent in both the ℓ_∞ and ℓ_2 norms when $p, n \to \infty$ with $p < n$, under weak conditions on the design matrix. These guarantees are shown to apply to cases with separated data.

1.2. *Our approach*

Our work is concerned with settings in which the true model is thought to be logistic but that separation precludes logistic likelihood fitting. Because the sufficient statistics are the same in logistic regression as in a notional linear in probability model, the maximum likelihood estimates of the logistic coefficients, if they exist, are recoverable from the ordinary least squares estimates obtained by treating the model for the probabilities as linear, as shown in Proposition 6. This suggests an approach to inference on the logistic coefficients based on the ordinary least squares estimator.

We establish a relationship between the logistic regression coefficients and the limiting values of the ordinary least squares estimates, and use this as a basis for componentwise estimation on β^*. In particular, assuming the existence of a consistent estimator of $X\beta^*$, this typically being easier to obtain than an estimate of the entries of β^*, we manipulate the least squares estimator to obtain a corrected least square estimator whose entries converge uniformly to those of the parameter of interest. Whilst biased, we then show that the LASSO (Tibshirani, 1996) estimate of β^* produces a consistent estimator of $X\beta^*$ when the unknown parameter is suitably sparse and $\max_{i=1}^n |x_i^T\beta^*| \leq c_1\sqrt{\log n}$ for some $c_1 > 0$. These conditions bound the entries of the unknown parameter thereby avoiding the issues caused by separation. Asumptions of this form are natural in high-dimensional regimes, see for example van de Geer et al. (2014).

Least-squares fitting of a linear regression model to binary data has been explored by Cox & Wermuth (1992) and Battey et al. (2019). The latter work parameterised the linear in probability model as

$$\mathrm{pr}(Y_i = y) = (1 + yx_i^T\beta^0)/2$$

under the restriction that for all data x, $|x^T\beta^0| \leq 1$. Ordinary least squares was the recommended approach for estimating the unknown parameter β^0 as this is more robust than maximum likelihood estimation to observations that invalidate the condition $|x^T\beta^0| \leq 1$. While there are advantages, notably of interpretation and existence of estimates, there are difficulties in treating the linear in probability model as generative. Indeed, the restriction to data x satisfying $|x^T\beta^0| \leq 1$ violates McCullagh's (2002) formal definition of a statistical model. For this reason we consider the generative model as logistic and use a relationship between the logistic coefficients and the probability limit of the ordinary least squares estimator to obtain a consistent estimator of the logistic parameters.

1.3. *Related work*

To avoid the issues encountered by maximum likelihood estimation in the logistic regression model, a number of methods have been proposed for use. When the maximum likelihood estimator exists, Sur & Candès (2019) introduced the Probe-Frontier method to correct the bias and consistently estimate β^* when p is large. Yadlowsky et al. (2021) remarked on the computational difficulties involved in using the Probe-Frontier method and proposed an alternative named SLOE. Both approaches rely on the existence of the maximum likelihood estimator and so are unsuitable for settings with separated data.

When the data are separated, Firth's (1993) bias reduced estimator has been recommended for use, see for example Heinze & Schemper (2002). Kosmidis & Firth (2021) showed that Firth's (1993) estimator always exists and established an analogous result for a more general version obtained by penalising the logistic log-likelihood function using a Jeffreys-prior penalty with arbitrary tuning parameter. Additionally, when p is fixed and $n \to \infty$, the first order asymptotic distribution of Firth's estimator coincides with that of the maximum likelihood estimator (Firth, 1993). It is unclear how Firth's (1993) estimator behaves when the maximum likelihood estimator does not exist and there are currently no theoretical guarantees when $p, n \to \infty$.

Although not proposed with this situation in mind, maximum likelihood estimation with certain forms of penalty on β^* ensures existence of an estimator when data are separated. Such estimators have been shown to have low composite estimation and prediction errors with high probability under sparsity assumptions (e.g. Duffy & Santner, 1989; van de Geer, 2008; Meier et al., 2008; Fan & Peng, 2004), however their components are biased for β_j^*, $j = 1, \ldots, p$. We make use of this observation to construct a consistent estimator of logistic regression coefficients using a consistent estimator of $X\beta^*$, the latter typically being easier to obtain. Unlike the inferential procedures of van de Geer et al. (2014), Ning & Liu (2017), Ma et al. (2021), Shi et al. (2021) and Cai et al. (2021) which entail correcting the bias of penalised estimators and require a consistent estimator of β^* in either the ℓ_1 or ℓ_2 norm, our procedure only requires that $X\beta^*$ be estimated consistently, making it applicable to a broader range of settings, see for example Raskutti et al. (2011).

2. NOTATION AND LIKELIHOOD FRAMEWORK

Let n observations on p variables be represented as vectors $x_1, \ldots, x_n \in \mathbb{R}^p$, and let $X \in \mathbb{R}^{n\times p}$ be the matrix with rows x_i^T. We assume throughout that X has full-rank, a condition that can always be checked once the data have been observed and which does not affect the presence of separation. Let Col-Sp(X) denote the column-span of X and $P_X = X(X^TX)^{-1}X^T$ the projection matrix onto Col-Sp(X). Each element of the response vector $Y = (Y_1, \ldots, Y_n)^T$, taking values in $\{-1, 1\}$, is assumed to be an independent random variable with distribution given in (1). A realisation of Y is written in lower case. Define $\Gamma \in \mathbb{R}^{n\times n}$ to be the diagonal matrix with (i,i)-th entry given by $\Gamma_{ii} = \mathrm{Var}(Y_i)$. Let $\hat{\beta}^*$ be the maximum likelihood estimator or MLE, when it exists, of β^* and let $\hat{\beta}^0 = (X^TX)^{-1}X^TY$ be the ordinary least squares, or OLS, estimator. Define β^0 to be the limiting value of $\hat{\beta}^0$ as $p, n \to \infty$ with $p < n$. Unless otherwise specified, this is the notional limiting operation assumed throughout.

For a function $f : \mathbb{R} \mapsto \mathbb{R}$ and a vector v, we use $f(v)$ to denote the vector with ith entry $f(v_i)$. The vector ℓ_1, ℓ_2 and ℓ_∞ norms are given by $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$. If the argument is a matrix, these refer to the matrix norms induced by the corresponding vector norms. The Frobenius norm is written $\|\cdot\|_F$. The minimum and maximum eigenvalues of a square matrix are written $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ respectively. For a set $S \subseteq \mathbb{R}^n$, the notation $S^\perp$ refers to its orthogonal complement

$$S^\perp = \{u \in \mathbb{R}^n : u^Tv = 0 \quad \forall v \in S\}.$$

For a univariate random variable Z, the sub-Gaussian norm is given by

$$\|Z\|_{\psi_2} = \sup_{q\geq 1} q^{-1/2}(\mathbb{E}|Z|^q)^{1/q}.$$

3. EXACT INFERENCE WITH SEPARATED DATA

3.1. *The exact conditional analysis of Cox (1958)*

In the logistic regression model, the log-likelihood function at an observation $y = (y_1, \ldots, y_n)^T$ is given by

$$\ell(\beta) = \log\left\{\frac{\exp(\sum_{j=1}^p t_j\beta_j)}{\prod_{i=1}^n(1+e^{x_i^T\beta})}\right\}$$

where $t_j = \sum_{i=1}^n x_{ij}z_i$ and $z_i = (y_i+1)/2 \in \{0,1\}$. Let T_j and Z_i be the random versions of these quantities, obtained by replacing y_i by Y_i. When the data are separated, the log-likelihood function asymptotes and so inference via maximum-likelihood fitting is unavailable.

Suppose that only inference on the first component β_1^* is of interest, the other entries $\beta_2^*, \ldots, \beta_p^*$ being regarded as nuisance parameters. When all entries of β^* are of interest each entry may be treated in turn as the single interest parameter. Cox (1958) argued that inference on

β_1^* should be based on the conditional distribution of T_1 given $T_2, \ldots, T_p$ given by

$$\mathrm{pr}(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) = \frac{c(t_1, \ldots, t_p) e^{\beta_1^* t_1}}{\sum_{u \in \mathbb{T}_1} c(u, t_2, \ldots, t_p) e^{\beta_1^* u}} \tag{2}$$

where

$$c(t_1, \ldots, t_p) = \left| \left\{ \tilde{z} \in \{0,1\}^n : \sum_{i=1}^{n} x_{ik} \tilde{z}_i = t_k, \, \forall k = 1, \ldots, p \right\} \right|$$

is the number of realisations of the outcome variable that produce the same observed values of the sufficient statistics $T_1, \ldots, T_p$, and

$$\mathbb{T}_1 = \left\{ \textstyle\sum_{i=1}^{n} x_{i1} \tilde{z}_i : \tilde{z} \in \mathbb{C}_1 \right\}, \quad \mathbb{C}_1 = \left\{ \tilde{z} \in \{0,1\}^n : \textstyle\sum_{i=1}^{n} x_{ik} \tilde{z}_i = t_k, \forall k = 2, \ldots, p \right\}.$$

Let

$$f(v \mid b) = \frac{c(v, t_1, \ldots, t_p) e^{bv}}{\sum_{u \in \mathbb{T}_1} c(u, t_2, \ldots, t_p) e^{bu}}$$

be the conditional probability that $T_1 = v$ when $\beta_1^* = b$, and

$$F_1(t_1 \mid b) = \textstyle\sum_{v \geq t_1} f(v \mid b), \qquad F_2(t_1 \mid b) = \textstyle\sum_{v \leq t_1} f(v \mid b)$$

be the conditional probabilities that $T_1 \geq t_1$ or $T_1 \leq t_1$. On replacing t_1 by T_1 in the definitions above, Cox (1970) constructed a $(1 - \vartheta)$-level exact confidence set for β_1^* as $(\beta_1^-(T_1), \beta_1^+(T_1))$ where

$$\begin{cases} F_1\{T_1 \mid \beta_1^-(T_1)\} = \vartheta/2 & \text{if } t_{\min} < T_1 \leq t_{\max} \\ \beta_1^-(T_1) = -\infty & \text{if } T_1 = t_{\min} \end{cases}$$

and

$$\begin{cases} F_2\{T_1 \mid \beta_1^+(T_1)\} = \vartheta/2 & \text{if } t_{\min} \leq T_1 < t_{\max} \\ \beta_1^+(T_1) = \infty & \text{if } T_1 = t_{\max}, \end{cases}$$

with $t_{\min}$ and $t_{\max}$ the minimum and maximum values of the set $\mathbb{T}_1$.

Let $(\beta_1^-(t_1), \beta_1^+(t_1))$ be the observed confidence interval for β_1^*. We show that the observed value of t_1 coincides with either or both of $t_{\min}$ and $t_{\max}$ when the data are linearly separable. It follows that the exact conditional confidence interval defined above is of infinite length. When the data are completely separated, that is, when there exists $\beta \in \mathbb{R}^p \backslash \{0\}$ such that $y_i x_i^T \beta > 0$ for all $i = 1, \ldots, n$, the interval may equal the whole real line.

PROPOSITION 1. *Suppose the data are separated by some vector $\beta \in \mathbb{R}^p \backslash \{0\}$. If $\beta_1 > 0$ then the upper limit of the confidence interval satisfies $\beta_1^+(t_1) = \infty$ and if $\beta_1 < 0$ then the lower limit of the confidence interval satisfies $\beta_1^-(t_1) = \infty$. If additionally the data are completely separated with $\beta_1 = 0$, the conditional likelihood satisfies*

$$pr(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) = 1, \qquad \forall \beta_1^* \in \mathbb{R}$$

and the exact confidence interval is the trivial interval $(\beta_1^-(t_1), \beta_1^+(t_1)) = \mathbb{R}$.

The result above only requires existence of one such β. If the data can be separated by multiple vectors, say $\beta^{(1)}, \beta^{(2)} \in \mathbb{R}^p$ with $\beta_1^{(1)} < 0 < \beta_1^{(2)}$, then the result may be applied to each vector separately to conclude that $(\beta_1^-(t_1), \beta_1^+(t_1)) = (-\infty, \infty)$.

In section 3.2, we show that all exact confidence sets necessarily contain infinite half intervals, irrespective of how they are constructed, and thus the restrictions outlined above are limitations of the data and not the method of analysis.

3.2. *Other forms of exact analysis*

Define $\mathrm{CS}_{\vartheta}^{(1)}(T_1) \subseteq \mathbb{R}$ to be an arbitrary $(1-\vartheta)$-level exact conditional confidence set for β_1^* satisfying

$$\mathrm{pr}\{\beta_1^* \in \mathrm{CS}_{\vartheta}^{(1)}(T_1) \mid T_2 = t_2, \ldots, T_p = t_p\} \geq 1 - \vartheta, \qquad \forall \beta_1^* \in \mathbb{R}. \tag{3}$$

Let $\mathrm{CS}_{\vartheta}^{(1)}(t_1)$ be its observed value. The following result outlines the form of these sets in the presence of separation.

THEOREM 1. *Suppose the observed data are separated by $\beta \in \mathbb{R}^p \backslash \{0\}$. If $\beta_1 > 0$ then there exists $B > 0$ such that $[B, \infty) \subseteq \mathrm{CS}_{\vartheta}^{(1)}(t_1)$ and if $\beta_1 < 0$ then there exists $B > 0$ such that $(-\infty, -B] \subseteq CS_{\vartheta}^{(1)}(t_1)$. If additionally the data are completely separated with $\beta_1 = 0$, the conditional likelihood satisfies*

$$\mathrm{pr}(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) = 1, \qquad \forall \beta_1^* \in \mathbb{R}$$

and the exact confidence set is $\mathrm{CS}_{\vartheta}^{(1)}(t_1) = \mathbb{R}$.

Similar results are obtained when an unconditional analysis is performed. This makes use of the distribution of the response vector Y rather than the conditional distribution of T_1 given $T_2, \ldots, T_p$. Define $\mathrm{CS}_{\vartheta}(Y) \subseteq \mathbb{R}^p$ to be a $(1-\vartheta)$-level exact unconditional confidence set for β^* satisfying

$$\mathrm{pr}\{\beta^* \in \mathrm{CS}_{\vartheta}(Y)\} \geq 1 - \vartheta, \qquad \forall \beta^* \in \mathbb{R}^p$$

with observed value $\mathrm{CS}_{\vartheta}(y) \subseteq \mathbb{R}^p$, and let $\mathrm{CS}_{\vartheta}^{(1)}(y)$ be the projection of $\mathrm{CS}_{\vartheta}(y)$ onto its first component.

THEOREM 2. *Suppose the observed data are separated by $\beta \in \mathbb{R}^p \backslash \{0\}$ and let $m \in \{0, 1, \ldots, n\}$ be the number of observations satisfying $x_i^T \beta = 0$. Assume $\vartheta < 2^{-m}$. If $\beta_1 > 0$ then there exists $B > 0$ such that $[B, \infty) \subseteq \mathrm{CS}_{\vartheta}^{(1)}(y)$ and if $\beta_1 < 0$ then there exists $B > 0$ such that $(-\infty, -B] \subseteq \mathrm{CS}_{\vartheta}^{(1)}(y)$. If additionally the data are completely separated by β and $\beta_1 = 0$, then $\mathrm{CS}_{\vartheta}^{(1)}(y) = \mathbb{R}$.*

Theorems 1 and 2 show that all confidence sets with exact coverage guarantees, either conditional or unconditional, contain at least one unbounded interval of the form $[B, \infty)$ or $(-\infty, -B]$ where $B > 0$. In some settings, these sets are equal to the whole real line. As a result, only limited information about the unknown parameter is available from data that are linearly separable. The most severe setting occurs when the data can be completely separated by some $\beta \in \mathbb{R}^p \backslash \{0\}$ with $\beta_1 = 0$, in which case there is never enough evidence to reject a null hypothesis concerning only β_1^*, whatever this might be. Even when the data can be separated but not completely separated, one-sided hypotheses of the form $H_0 : \beta_1^* > b_0$ or $H_0 : \beta_1^* < b_0$ for $b_0 \in \mathbb{R}$ cannot be rejected, depending on the sign of the first entry of the separating parameter. Any refinement requires either additional data or further assumptions.

The non-existence of finite confidence intervals also affects estimation as, for example, it is impossible to guarantee that an estimate of an entry of β^* lies in a small region about the 220 unknown parameter with a pre-determined probability. Indeed, if there existed such an estimate $\hat{\beta}_1$ of β^*_1 satisfying

$$\text{pr}(|\hat{\beta}_1 - \beta^*_1| < \epsilon) \geq 1 - \vartheta, \qquad \forall \beta^*_1 \in \mathbb{R}$$

for some $\epsilon > 0$ and $\vartheta \in (0, 1)$, then $[\hat{\beta}_1 - \epsilon, \hat{\beta}_1 + \epsilon]$ would be an exact $(1 - \vartheta)$-level confidence interval for β^*_1 with bounded support. Markov's inequality implies that the variance of any such estimate is unbounded as a function of the unknown parameter. 225

In high-dimensional settings, restrictions on $X\beta^*$ are natural and often made. These justify our approach to estimation based on least squares, to be presented in section 5, which has statistical guarantees even when the maximum likelihood estimator does not exist or exhibits poor performance. Our results are asymptotic, allowing both the dimension p and the sample size n to diverge simultaneously. 230

4. Preliminary results

We begin by studying the limiting behaviour of the least squares estimator $\hat{\beta}^0$ in the logistic regression model. This motivates a construction that allows consistent estimation of the logistic coefficient of interest. Since the dimension p is allowed to grow under the notional operation $n \to \infty$, the limit distribution of $\hat{\beta}^0$ is not well-defined. Instead we consider the behaviour of 235 linear functions $\alpha^T \hat{\beta}^0$, where choices of particular interest are α equal to one of the canonical basis vectors for $\mathbb{R}^p$ or representing simple contrasts of the entries of $\hat{\beta}^0$. Thus assume $\alpha \in \mathcal{B}_d$ for some $d > 0$ where

$$\mathcal{B}_d = \{\alpha \in \mathbb{R}^p : \|\alpha\|_0 \leq d, \ \|\alpha\|_2 \leq 1\}$$

is the sparse ℓ_2-ball of radius one in $\mathbb{R}^p$. The following result shows that linear functions of the least squares estimator converge in probability to similar functions of 240

$$\beta^0 = (X^T X)^{-1} X^T \tanh(X\beta^*/2).$$

Proposition 2. *Define*

$$\mathbb{X}_B = \left\{ X : \sup_{\alpha \in \mathcal{B}_d} \|\alpha^T (X^T X)^{-1} X^T\|_2^2 \leq B n^{-1} \right\}$$

for some constant $B > 0$. When $d = o(n)$, $d \log(p/d)/n = o(1)$ and $t > 0$,

$$\sup_{X \in \mathbb{X}_B} \text{pr}\left\{ \sup_{\alpha \in \mathcal{B}_d} |\alpha^T (\hat{\beta}^0 - \beta^0)| \geq t \right\} = o(1). \tag{4}$$

The result in Proposition 2 is stated uniformly over all design matrices contained in the set $\mathbb{X}_B$. Justification for this will be provided at a later stage. For now, it is sufficient to consider a single design matrix satisfying $\|\alpha^T (X^T X)^{-1} X^T\|_2^2 = O(n^{-1})$, for which it follows that 245

$$\|\hat{\beta}^0 - \beta^0\|_\infty = o_P(1), \qquad p^{-1/2} \|\hat{\beta}^0 - \beta^0\|_2 = o_P(1)$$

provided the diagonal entries of $(X^TX/n)^{-1}$ are asymptotically bounded above under no restrictions on p beyond $p < n$. This may be seen by setting $d = 1$ in Proposition 2. Proposition 3 strengthens the latter result, showing that the rate of convergence in ℓ_2-norm is of order $p^{-1/2}$.

PROPOSITION 3. *Suppose* $\lambda_{\max}\{(X^TX/n)^{-1}\} = O(1)$. *Then,*

$$p^{-1/2}\|\hat{\beta}^0 - \beta^0\|_2 = O_P(p^{-1/2})$$

and when $p = o(n)$,

$$\|\hat{\beta}^0 - \beta^0\|_2 = o_P(1).$$

For inference on the entries of β^*, the limiting distribution of the least squares estimator is of interest. Proposition 4 shows that after suitable normalisation, the distribution of $\alpha^T(\hat{\beta}^0 - \beta^0)$ is asymptotically Gaussian.

PROPOSITION 4. *Let* $B_n = \|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2$ *and define* $\mathcal{R} \subseteq \mathbb{R}^p$ *to be a set satisfying*

$$\sup_{\alpha\in\mathcal{R}} B_n^{-1}\|\alpha^T(X^TX)^{-1}X^T\|_\infty = o(1). \tag{5}$$

Then,

$$\sup_{\alpha\in\mathcal{R}}\sup_{x\in\mathbb{R}}\left|\text{pr}\{B_n^{-1}\alpha^T(\hat{\beta}^0 - \beta^0) \le x\} - \Phi(x)\right| = o(1).$$

Assumption (5) arises when the quantity of interest is expressed as a sum of independent random variables and central limit type arguments are used to derive its asymptotic distribution. It is closely related to a Lindenberg condition. Similar assumptions are made by Huber (1973) and Lei et al. (2018) to establish the asymptotic normality of the least squares estimator in a different context. To understand when this assumption holds, suppose the rows of X are independently and identically distributed as centred, multivariate normal random variables with covariance matrix Σ and focus on the limiting setting where $p, n \to \infty$ with $p/n \to \kappa \in [0, 1)$ and ${\beta^*}^T\Sigma\beta^* \to \gamma^2$ for some $\gamma > 0$. This is a setting that will be considered further in section 7. Let $\mathcal{R}$ consist of the standard basis vectors of $\mathbb{R}^p$. Theorem 2.16 (Bai, 1999) shows that

$$\sup_{\alpha\in\mathcal{R}}\|\alpha^T(X^TX)^{-1}X^T\|_2^{-1} = O_P(n^{1/2}).$$

Further, $\max_{i=1}^n|x_i^T\beta^*| = O_P(\sqrt{\log n})$ and so

$$-\log(\min_{i=1}^n\Gamma_{ii}) \le \max_{i=1}^n|x_i^T\beta^*| = O_P(\sqrt{\log n}).$$

As $B_n \ge \min_{i=1}^n\Gamma_{ii}^{1/2}\|\alpha^T(X^TX)^{-1}X^T\|_2$, a sufficient condition for the left-hand side of (5) to be $o_P(1)$ is

$$\sup_{\alpha\in\mathcal{R}}\|\alpha^T(X^TX)^{-1}X^T\|_\infty = o_P\left(n^{-1/2}e^{-c\sqrt{\log n}}\right)$$

for all constants $c > 0$. Although the distribution of $(X^T X)^{-1} X^T$ is unknown, if we assume that the entries of $(X^T X/n)^{-1} X^T$ are sub-Gaussian with bounded norm, then 270

$$\begin{aligned}\|\alpha^T (X^T X)^{-1} X^T\|_\infty &= O_P(n^{-1}\sqrt{\log n}) \\ &= o_P\left(n^{-1/2} e^{-c\sqrt{\log n}}\right)\end{aligned}$$

for all $c > 0$ and so the condition is satisfied.

5. MAIN RESULTS

The previous results motivate a corrected least squares estimator, which is shown in the present section to be consistent in both the ℓ_∞ and scaled ℓ_2 norms, and to have some predictive guarantees. The considerations involved in obtaining stronger inferential guarantees are also briefly 275 discussed and assessed by simulation.

5.1. *The corrected least squares estimator*

Section 4 showed that the probability limit of a linear function of the least squares estimator is a linear function of $\beta^0 = f(\beta^*)$ where

$$f(\beta^*) = (X^T X)^{-1} X^T \tanh(X\beta^*/2).$$

If this function f were invertible, then a consistent estimator of each entry of β^* could poten- 280 tially be obtained using $f^{-1}(\hat\beta^0)$. The function is not invertible, however we show that it can be rewritten as

$$f(\beta^*) = \varsigma\beta^* + \delta$$

for some $\varsigma \in \mathbb{R}$ and $\delta \in \mathbb{R}^p$ depending only on $X\beta^*$. Whilst the entries of β^* are difficult to estimate, estimation of $X\beta^*$ is simpler and leads to an estimator of $\alpha^T\beta^*$ in the form

$$\hat\varsigma^{-1}\alpha^T(\hat\beta^0 - \hat\delta)$$

where $\hat\varsigma$ and $\hat\delta$ are estimates of ς and δ to be defined. 285

Write

$$\tanh(X\beta^*/2) = \varsigma X\beta^* + u + \Delta \tag{6}$$

where $\varsigma X\beta^*$ is the projection of $\tanh(X\beta^*/2)$ onto $X\beta^*$, $u \in \text{Col-Sp}(X)^\perp$ and Δ is the projection of $\tanh(X\beta^*/2)$ onto the subspace

$$\text{Col-Sp}(X) \cap \text{Col-Sp}(X\beta^*)^\perp = \{\varsigma X\beta^* + u : \varsigma \in \mathbb{R},\, u \in \text{Col-Sp}(X)^\perp\}^\perp.$$

Such a decomposition exists uniquely because the subspaces Col-Sp$(X\beta^*)$, Col-Sp$(X) \cap$ Col-Sp$(X\beta^*)^\perp$ and Col-Sp$(X)^\perp$ are orthogonal and span the whole of $\mathbb{R}^n$. It follows that 290

$$\varsigma = \begin{cases} 1/2 & X\beta^* = 0 \\ \frac{(X\beta^*)^T \tanh(X\beta^*/2)}{\|X\beta^*\|_2^2} & X\beta^* \neq 0, \end{cases} \qquad \Delta = (P_X - P_{X\beta^*})\tanh(X\beta^*/2)$$

where, to correspond with the definition of ς, we use the notation $P_{X\beta^*}\tanh(X\beta^*/2)$ to mean $X\beta^*/2$ when $X\beta^*$ is the zero vector. On definining $\delta = (X^T X)^{-1} X^T \Delta$,

$$\beta^0 = (X^T X)^{-1} X^T \tanh(X\beta^*/2) = \varsigma\beta^* + \delta. \tag{7}$$

Based on this observation, define the corrected least squares estimator to be

$$\tilde{\beta}^* = \hat{\varsigma}^{-1}(\hat{\beta}^0 - \hat{\delta})$$

where $\hat{\varsigma}$ and $\hat{\delta}$ are given by

$$\hat{\varsigma} = \begin{cases} 1/2 & \hat{\eta} = 0, \\ \frac{\hat{\eta}^T \tanh(\hat{\eta}/2)}{\|\hat{\eta}\|_2^2} & \hat{\eta} \neq 0 \end{cases} \qquad \hat{\delta} = (X^T X)^{-1} X^T \hat{P} \tanh(\hat{\eta}/2),$$

with $\hat{P} = P_X - P_{\hat{\eta}}$ and $\hat{\eta}$ a consistent estimator of $\eta^* = X\beta^*$ to be discussed next.

5.2. *Assumptions*

To ensure that the corrected least squares estimator may be used to estimate β^*, assumptions are required in addition to X being full rank with $p < n$. Define

$$\mathbb{X}_B^{(1)} := \{X \in \mathbb{R}^{n\times p} : \max\{\varsigma^{-1}, \lambda_{\max}\{(X^T X/n)^{-1}\}\} \leq B\},$$

for a constant $B > 0$. This set satisfies $\mathbb{X}_B^{(1)} \subseteq \mathbb{X}_B$ where $\mathbb{X}_B$ was defined in Proposition 2.

Condition 1. the unknown parameter satisfies a) $p^{-1/2}\|\beta^*\|_2 = O(1)$ or b) $\|\beta^*\|_\infty = O(1)$.

Condition 2. there exists a non-empty $\mathcal{H}_B \subseteq \mathbb{X}_B^{(1)}$ such that for all $t > 0$,

$$\sup_{X\in\mathcal{H}_B} \mathbb{P}\{h(\hat{\eta}, \eta^*) > t\} = o(1)$$

where $\eta^* = X\beta^*$ and

$$h(\hat{\eta}, \eta^*) = \begin{cases} \|\hat{\eta} - \eta^*\|_2 & \eta^* = 0, \\ \max\left\{\frac{\|\hat{\eta}-\eta^*\|_2}{\|\eta^*\|_2}, \frac{\|\hat{\eta}-\eta^*\|_2}{\sqrt{n}}\right\} & \eta^* \neq 0. \end{cases}$$

Condition 1 makes restrictions on the unknown parameter that avoids the issues outlined in section 3. Part a) follows from part b), although there will be settings where only part a) is needed and we identify these throughout. Assumptions of this form are common in high-dimensional regimes, for example, Condition 1 is implied when β^* contains at most $s = O(1)$ non-zero entries of bounded magnitude, but may also hold for dense vectors.

Condition 2 ensures the existence of a design matrix $X \in \mathcal{H}_B$ for some $B > 0$. For this matrix,

$$\max\{\varsigma^{-1}, \lambda_{\max}\{(X^T X/n)^{-1}\}\} = O(1)$$

and there exists $\hat{\eta}$ where $h(\hat{\eta}, \eta^*) = o_P(1)$. This second statement guarantees that a consistent estimator of $\eta^* = X\beta^*$ is available for estimation of ς and δ. The former allows the behaviour of the least squares estimator to be controlled, as in Propositions 2 and 3, and limits the accumulation of error when correcting the estimator using $\hat{\varsigma}$ and $\hat{\delta}$. It is sometimes sufficient to replace the eigenvalue condition by a weaker one but for notational simplicity, we do not do this here. When the rows of X are treated as independent and identically distributed observations from an appropriate distribution and $\beta^* \neq 0$,

$$\varsigma^{-1} = \frac{n^{-1}\sum_{i=1}^n (x_i^T\beta^*)^2}{n^{-1}\sum_{i=1}^n x_i^T\beta^* \tanh(x_i^T\beta^*/2)}$$

and the weak law of large numbers may be applied to both the numerator and denominator to deduce that $\varsigma^{-1} = O_P(1)$, see Lemma S8 for details. In section 6 we identify suitable choices of $\hat{\eta}$ and $\mathcal{H}_B$ that ensure Condition 2 is met.

For most of the analysis, it will be sufficient to focus on a single $X \in \mathcal{H}_B$. However, in section 7, we consider the validity of our results in settings with separated data, focusing on the accuracy of our estimator in terms of the ℓ_∞-norm. As our results are asymptotic and the limiting probability of data separation has not yet been considered for fixed designs, section 7 treats the design matrix as random. To extend our analysis to this framework, we show that consistency in terms of the ℓ_∞-norm in the fixed design setting holds uniformly over all $X \in \mathcal{H}_B$. It follows that our estimator consistently estimates the entries of β^* with respect to certain joint distributions for X and Y, even in settings where data separation occurs with probability converging to one. For clarity, we use $\mathrm{pr}(\cdot)$ to denote the probability conditional on the observed value of the design and $\mathrm{pr}_{Y,X}(\cdot)$ when considering the joint distribution. The latter only appears in section 7.

Our assumptions make no explicit restrictions on p relative to n beyond $p < n$. However there may be some implicit constraints. Section 7 considers a more refined setting where $p/n \to \kappa \in [0, 1)$ to evaluate the validity of our approach when the data are linearly separable.

5.3. *Consistent estimation*

Under the assumptions in section 5.2, the corrected least squares estimator $\tilde{\beta}^*$ is consistent, both entry-wise and in terms of the ℓ_2-norm. We refer to the latter as the composite estimation error. This is established in Theorems 3 and 4. In light of the comments in section 5.2, consistency in terms of the ℓ_∞-norm is established uniformly over design matrices $X \in \mathcal{H}_B$ in Theorem 3. All other results focus on pointwise convergence.

THEOREM 3. *Suppose conditions 1b) and 2 hold. For all $t > 0$,*

$$\sup_{X \in \mathcal{H}_B} \mathrm{pr}\Big(\sup_{\alpha \in \mathcal{B}_d} |\alpha^T(\tilde{\beta}^* - \beta^*)| \geq t\Big) = o(1)$$

as $p, n \to \infty$ with $p < n$ and $d = O(1)$. In particular,

$$\sup_{X \in \mathcal{H}_B} \mathrm{pr}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t) = o(1).$$

THEOREM 4. *Suppose conditions 1a) and 2 hold. If there exist $B, N > 0$ such that $X \in \mathcal{H}_B$ for all $n \geq N$, then,*

$$p^{-1/2}\|\tilde{\beta}^* - \beta^*\|_2 = o_P(1)$$

as $p, n \to \infty$ with $p < n$.

To derive these results, the estimation errors were decomposed into terms involving $\hat{\beta}^0 - \beta^0$, $\hat{\varsigma} - \varsigma$, and $\hat{\delta} - \delta$. The results in section 4 and Condition 2 ensure that these quantities converge to zero in probability.

For the purpose of variable selection, it is the first result that is of most interest. Provided the non-zero entries of β^* are sufficiently large, a variable selection procedure that selects the $\hat{s}$ indices corresponding to the entries of $\tilde{\beta}^*$ with the largest magnitudes will asymptotically prioritise signal variables over noise variables. For a more formal Wald-based test of $\tilde{\beta}^*_j = 0$ with appropriate calibration, the asymptotic distribution of the corrected least squares estimator is needed.

5.4. *A remark on inference*

Proposition 5 provides initial insights into the limiting distribution of the corrected least squares estimator, probed further by simulation in section 8.

PROPOSITION 5. *Suppose Condition 2 holds and there exist $B, N > 0$ such that $X \in \mathcal{H}_B$ for all $n \geq N$. Assume $\alpha \in \mathcal{B}_d$ with*

$$B_n^{-1}\|\alpha^T(X^TX)^{-1}X^T\|_\infty = o(1).$$

Then, under the null hypothesis $H_0 : \alpha^T\beta^* = 0$,

$$\frac{\alpha^T(\tilde{\beta}^* - \beta^*)}{\varsigma^{-1}\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2} = \Pi_1 + \Pi_2 + o_P(1)$$

where

$$\Pi_1 \xrightarrow{d} N(0,1), \qquad \Pi_2 = -\varsigma/\hat{\varsigma}B_n^{-1}\alpha^T(\hat{\delta} - \delta).$$

Proposition 5 shows that, up to the term Π_2, the limiting distribution of a scaled version of our estimator is standard normal. In section 8, we conduct simulations whose results suggest that Π_2 is negligible in the particular cases considered. They also suggest that a normal approximation to the distribution of the corrected least squares estimator may be accurate even when $\alpha^T\beta^* \neq 0$. Neither observation has yet been established theoretically. Part of the difficulty arises in the term B_n^{-1}, which up to the quantity Γ, is of order $n^{1/2}$. Then for Π_2 to converge to zero in probability, we would require

$$n^{1/2}\alpha^T(\hat{\delta} - \delta) = o_P(1),$$

however we have only shown that

$$n^{1/2}\alpha^T(\hat{\delta} - \delta) = O(\|\hat{\eta} - \eta^*\|_2)$$

which is not expected to be $o_P(1)$ in high-dimensional regimes.

If a closed form approximation to the distribution of the corrected least squares estimator cannot be derived, a bootstrap algorithm may serve to estimate p-values. This has not yet been considered in detail, but also presents an avenue for future work.

5.5. *Prediction error*

Although the primary aim was estimation of the unknown parameter, the prediction error of the corrected least squares estimator, defined by

$$n^{-1/2}\|X\tilde{\beta}^* - X\beta^*\|_2,$$

may be of interest. This is somewhat misleading terminology, as $X\beta^*$ is the vector of logistic-transformed probabilities. Theorem 5 shows that the above quantity converges in probability to a value that depends on the relative dimension p/n and the signal strength.

THEOREM 5. *Suppose there exist $B, N > 0$ such that $X \in \mathcal{H}_B$ for all $n \geq N$. Then,*

$$n^{-1/2}\|X(\tilde{\beta}^* - \beta^*)\|_2 = n^{-1/2}\varsigma^{-1}\|P_X\Gamma^{1/2}\|_F + o_P(1)$$

when $p, n \to \infty$ with $p < n$.

When $p = o(n)$, the error is $o_P(1)$ as $\|P_X\Gamma^{1/2}\|_F$ is bounded above by $\|P_X\|_F = \sqrt{p}$. When p/n does not converge to zero, there exist values of β^* where the prediction error does not decay to zero. As a result, whilst the corrected least squares estimator may be usefully used for estimation and variable screening, it is less suitable for inference on the logistic transforms of the individual probabilities. This is unproblematic as Condition 2 assumed the existence of an alternative estimator suitable for this purpose.

Starting with an initial estimate $\hat{\eta}^{(1)}$ of η^*, consider an iterative version of our estimator where for $i = 1, 2, \ldots$

$$\hat{\varsigma}^{(i)} = \begin{cases} 1/2 & \hat{\eta}^{(i)} = 0, \\ \frac{(\hat{\eta}^{(i)})^T \tanh(\hat{\eta}^{(i)}/2)}{\|\hat{\eta}^{(i)}\|_2^2} & \hat{\eta}^{(i)} \neq 0 \end{cases} \qquad \hat{\delta}^{(i)} = (X^T X)^{-1} X^T \{P_X - P_{\hat{\eta}^{(i)}}\} \tanh(\hat{\eta}^{(i)}/2)$$

$$\tilde{\beta}^{(i)} = (\hat{\beta}^0 - \hat{\delta}^{(i)})/\hat{\varsigma}^{(i)}, \qquad \hat{\eta}^{(i+1)} = X\tilde{\beta}^{(i)}.$$

In view of Theorem 5, the iterative version need not perform better than the once-corrected version, as $\hat{\eta} = X\tilde{\beta}^*$ violates the convergence condition in Condition 2. This conclusion was also checked by simulation but not reported.

6. Possible choices of $\hat{\eta}$

6.1. *Maximum likelihood estimation*

When the data cannot be separated, the maximum likelihood estimator exists and may be used to correct the least-squares estimator. Proposition 6 shows that our estimator obtained by setting $\hat{\eta}$ to $X\hat{\beta}^*$ recovers the original maximum likelihood estimator.

Proposition 6. *Suppose the logistic maximum likelihood estimator $\hat{\beta}^*$ exists and let $\hat{\eta} = X\hat{\beta}^*$. Define*

$$\hat{\varsigma} = \begin{cases} 1/2 & \hat{\eta} = 0, \\ \frac{\hat{\eta}^T \tanh(\hat{\eta}/2)}{\|\hat{\eta}\|_2^2} & \hat{\eta} \neq 0, \end{cases} \qquad \hat{\delta} = (X^T X)^{-1} X^T \hat{P} \tanh(\hat{\eta}/2),$$

where $\hat{P} = P_X - P_{\hat{\eta}}$. Then, $\hat{\beta}^* = \hat{\varsigma}^{-1}(\hat{\beta}^0 - \hat{\delta})$. That is, the corrected OLS estimator recovers the logistic maximum likelihood estimator.

This result only serves to supply insight and is of no practical relevance. The equivalence to maximum likelihood estimation in this setting demonstrates that the corrected OLS estimator, without exploiting further assumptions, is subject to the same considerable bias as maximum likelihood estimation. In the next section, we show that when this estimator is combined with a version of $\hat{\eta}$ that exploits constraints on $X\beta^*$, bias is substantially reduced.

6.2. *Penalised regression*

To operationalise $\tilde{\beta}^*$, a consistent estimator $\hat{\eta}$ of $\eta^* = X\beta^*$ satisfying $h(\hat{\eta}, \eta^*) = o_P(1)$ is needed. A penalised regression may be used to obtain such an estimator even when the maximum likelihood estimator does not exist. This entails setting $\hat{\eta} = X\hat{\beta}^{(\lambda)}$ where

$$\hat{\beta}^{(\lambda)} = \text{argmin}_{\beta \in \mathbb{R}^p} \{-n^{-1}\ell(\beta) + \lambda p(\beta)\}, \qquad \lambda \geq 0 \tag{8}$$

and $p : \mathbb{R}^p \to \mathbb{R}$ is a penalty function that does not depend on the data but ensures a unique maximiser exists. See Duffy & Santner (1989), Meier et al. (2008), Kosmidis & Firth (2021) and Fan & Peng (2004) for examples including the ridge, group LASSO, Jeffrey's-prior and non-concave penalty functions respectively. Unless strong conditions are imposed on the design matrix that limit the amount of correlation between covariates, $\hat{\beta}^{(\lambda)}$ rarely provides an accurate estimate of the individual entries of β^*, see section 8 for numerical examples. Nevertheless, under much weaker conditions $X\hat{\beta}^{(\lambda)}$ consistently estimates η^*. The corrected least squares estimator that makes use of $X\hat{\beta}^{(\lambda)}$ to estimate η^*, improves estimation of the entries of β^* over that of $\hat{\beta}^{(\lambda)}$.

As an example, consider the LASSO estimator that maximises (8) with $p(\beta) = \|\beta\|_1$. Lemma 1 in Meier et al. (2008) shows that $\hat{\beta}^{(\lambda)}$ exists whenever $\lambda > 0$ and $0 < \sum_{i=1}^n y_i < n$, which includes cases with separated data. Further, under a suitable sparsity assumption, the LASSO estimator produces consistent predictions whenever the design matrix is contained in

$$\mathbb{X}_B^{(2)} := \left\{ X \in \mathbb{R}^{n\times p} : \max\left\{ \max_{i=1}^{n} \frac{|x_i^T \beta^*|}{\sqrt{\log n}}, \max_{i=1}^{n} \max_{j=1}^{p} \frac{|x_{ij}|}{\sqrt{\log n}}, \frac{n^{1/2}\mathbb{1}\{\beta^* \neq 0\}}{\|X\beta^*\|_2}, \phi_0^{-2}(X) \right\} \le B \right\}$$

where $\mathcal{S} = \{j : \beta_j^* \neq 0\}$, $s = |\mathcal{S}|$, $\beta_{\mathcal{S}}$ is the vector with entries equal to those of β for all indices in $\mathcal{S}$ and 0 otherwise, and

$$\phi_0^2(X) = \inf_{\beta\in\mathbb{R}^p : \beta_{\mathcal{S}}\neq 0,\, \|\beta_{\mathcal{S}^c}\|_1 \le 3\|\beta_{\mathcal{S}}\|_1} \frac{\|X\beta\|_2^2 s}{n\|\beta_{\mathcal{S}}\|_1^2}.$$

This is established in Proposition 7.

PROPOSITION 7. *Let $\lambda = A\sqrt{(\log p \log n)/n}$ with $A > 0$ sufficiently large. For $t > 0$,*

$$\sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}\{h(X\hat{\beta}^{(\lambda)}, X\beta^*) > t\} = o(1)$$

when

$$s e^{2B\sqrt{\log n}} \sqrt{\log p \log n / n} \max\{1, e^{2B\sqrt{\log n}} \sqrt{\log p \log n / n}\} = o(1).$$

The proof of this result closely follows the arguments in Theorem 6.4 and Lemma 6.8 in Bühlmann & van de Geer (2011, p. 130-134). We make minor modifications to account for our slightly weaker assumptions, see the definition of $\mathbb{X}_B^{(2)}$ compared to the assumptions in Lemma 6.8 of Bühlmann & van de Geer (2011).

Proposition 7 establishes conditions under which Condition 2 is satisfied, with

$$\hat{\eta} = X\hat{\beta}^{(\lambda)}, \qquad \mathcal{H}_B = \mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}.$$

As a result, the least squares estimator, after correction by the LASSO, is consistent with respect to both the ℓ_2 and ℓ_∞ norms whenever there exists $B, N > 0$ such that $X \in \mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ for all $n \ge N$. We call this the OLS-LASSO estimator. In the following section, we show that when the rows of X follow a multivariate Gaussian distribution, there exist suitable constants such the probability that $X \in \mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ is arbitrarily close to one for large enough sample sizes. Thus, the set $\mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ is non-empty.

6.3. *Other approaches*

Any consistent estimator of $\eta^* = X\beta^*$ can be used for correction. Thus, if β^* is sparse, Proposition 7 shows that a LASSO penalty yields an estimator with appropriate behaviour. If instead β^* is dense but $X\beta^*$ is well-approximated by a small number of left singular vectors of X, then it may be possible to consistently estimate η^* using a sparse singular value decomposition of X. We do not explore this here, although it highlights our method's potential validity under a variety of sparsity assumptions.

7. Relevance to separated data

The applicability of our method to separated data is considered here, focusing on the OLS-LASSO estimator defined in section 6.2. As no restrictions on the observed response y are made in section 5.2, and section 6.2 only assumes $0 < \sum_{i=1}^{n} y_i < n$, the occurrence of separability does not affect the existence of our estimator. To ensure that our asymptotic guarantees are also valid, it is necessary to verify that the limiting probability of separation is non-zero. If this were not the case, then for any $t > 0$,

$$\mathrm{pr}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t) = \mathrm{pr}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t, \text{ data are not separated}) + o_P(1)$$

and so consistency may be achieved irrespective of the value of the estimator when the data are separable. The probability of data separation has not yet been studied for fixed designs and so we adopt the random design setting introduced by Candès & Sur (2020) and summarised in Condition 3.

Condition 3. The joint distribution of (Y, X) is given by

$$x_i \overset{i.i.d.}{\sim} N_p(0, \Sigma), \qquad \mathbb{P}(Y_i = 1 \mid x_i) = \frac{e^{x_i^T \beta^*}}{1 + e^{x_i^T \beta^*}}$$

and for some $\kappa \in [0, 1)$ and $\gamma^2 \geq 0$,

$$p/n \to \kappa, \qquad \beta^{*T} \Sigma \beta^* \to \gamma^2$$

as $p, n \to \infty$.

Let $\mathrm{pr}_{Y,X}$ denote the probability under the joint distribution of Y and X. Candès & Sur (2020, Theorem 2.2) show that there exists some decreasing function $h(\gamma)$ where

$$\begin{aligned} \kappa > h(\gamma) &\implies \mathrm{pr}_{Y,X}(\text{data are separated}) \to 1 \\ \kappa < h(\gamma) &\implies \mathrm{pr}_{Y,X}(\text{data are separated}) \to 0. \end{aligned}$$

Our aim is to show that the OLS-LASSO estimator is consistent with respect to the ℓ_∞-norm even in these settings where the limiting probability of separation is one. For this, we also assume Condition 4.

Condition 4. a) The unknown parameter satisfies $\|\beta^*\|_\infty = O(1)$ and $\|\beta^*\|_0 = O(n^{1/2-\xi})$ for some $\xi \in (0, 1/2)$, and b) there exist constants $\lambda_{\min}, \sigma_{\max} > 0$ such that

$$\lambda_{\min}(\Sigma) > \lambda_{\min}, \qquad \max_{j=1}^{p} \Sigma_{jj} \leq \sigma_{\max}.$$

For every $\kappa \in [0, 1)$ and $\gamma \geq 0$, there exist choices of Σ and β^* satisfying the conditions above. For example, let $\beta^* \in \mathbb{R}^p$ have exactly one non-zero entry taking the value $\gamma > 0$ and let $\Sigma = (1/4)\mathbb{1}\mathbb{1}^T + (3/4)\mathbb{I}_{p \times p}$. Thus, Condition 4 makes no further restriction on the limiting probability of separation.

Proposition 8 shows that in this random design setting, the corrected least squares estimator is consistent. If κ is sufficiently large compared to γ, then consistency is achieved alongside separation.

PROPOSITION 8. *Suppose Conditions 3 and 4 hold with $\gamma, \kappa > 0$. Let $\tilde{\beta}^*$ be the OLS-LASSO estimator. Then, for all $t > 0$*

$$\mathrm{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t) = o(1).$$

In particular, when $\kappa > h(\gamma)$,

$$\mathrm{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty < t, \textit{ data are separated}) \to 1.$$

Similar results are expected to hold for other quantities of interest, for example the composite estimation error, although we do not derive a result of this form. Instead, given our preference for conditional analyses, we highlight the importance of characterising the limiting probability of separation for fixed designs to ensure similar guarantees can also be provided in these settings.

8. NUMERICAL PERFORMANCE

Results of extensive numerical experimentation are reported in the supplementary material. In section 4.1, different versions of the corrected least squares estimators were obtained from various estimators of $\eta^* = X\beta^*$. For each estimator, error rates were examined as a function of the sample size n when $p, n \to \infty$ with p/n kept constant. Cases with separated data were included. The results show that both the component-wise and composite estimation errors converged to zero, whilst the prediction error remained stable at a non-zero value, coinciding with the analysis in section 5. The small-sample performance was analysed in section 4.2 of the supplementary material, where the average composite estimation and prediction errors of the corrected least squares estimators were recorded in multiple contexts, keeping n fixed. The results were compared to those of the maximum likelihood and Firth's (1993) estimator.

The remainder of this section fixes $n = 700$ and $p = 70$ and provides a comparison to other available methods. Based on the results in section 4.1 of the supplementary material, we focus on the SCAD (Fan & Li, 2001) correction to the least squares estimator, denoted OLS-SCAD, as it outperformed other corrections. The following four approaches were also considered: a SCAD penalised regression (Fan & Li, 2001), Firth's bias-reduced estimator (Firth, 1993), the desparsified LASSO (van de Geer et al., 2014) and the LSW estimator (Cai et al., 2021). The first comparison serves to illustrate bias removal from penalised estimators. Other penalties were examined and exhibited similar performance, therefore the results are not reported. The methods of van de Geer et al. (2014) and Cai et al. (2021) also aim to remove bias from penalised estimators in high-dimensional settings, although data separation was not explicitly considered.

The data were generated as defined in section 7, with the rows of X sampled independently from a p-dimensional multivariate Gaussian distribution with mean zero and covariance matrix Σ to be specified. The outcome Y was generated from a logistic regression model with log-odds equal to $X\beta^*$. The following three examples were considered:

Example 1. The covariance matrix and parameter vector were given by

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0.95 & \text{if } \{i,j\} = \{1,2\}, \\ 0.5 & \text{otherwise,} \end{cases} \qquad \beta_i^* = \begin{cases} 1 & \text{if } i = 1,3 \\ -1 & \text{if } i = 2 \\ 0 & \text{otherwise,} \end{cases}$$

to ensure the presence of a pair of highly correlated signal variables with small marginal correlation with the response.

Example 2. The covariance matrix and parameter vector were given by

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0.95 & \text{if } \{i,j\} = \{1,4\}, \\ 0.5 & \text{otherwise}, \end{cases} \qquad \beta_i^* = \begin{cases} 1 & \text{if } i = 1,2,3 \\ 0 & \text{otherwise}, \end{cases}$$

to ensure the presence of a signal variable that is highly correlated with a noise variable.

Example 3. The covariance matrix and parameter vector were given by

$$\Sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0.95 & \text{if } \{i,j\} = \{1,2\}, \\ 0.5 & \text{otherwise}, \end{cases} \qquad \beta_i^* = \begin{cases} 1 & \text{if } i = 1,2,3 \\ 0 & \text{otherwise}. \end{cases}$$

to ensure the presence of a pair of highly correlated signal variables with equal signal strength. 510

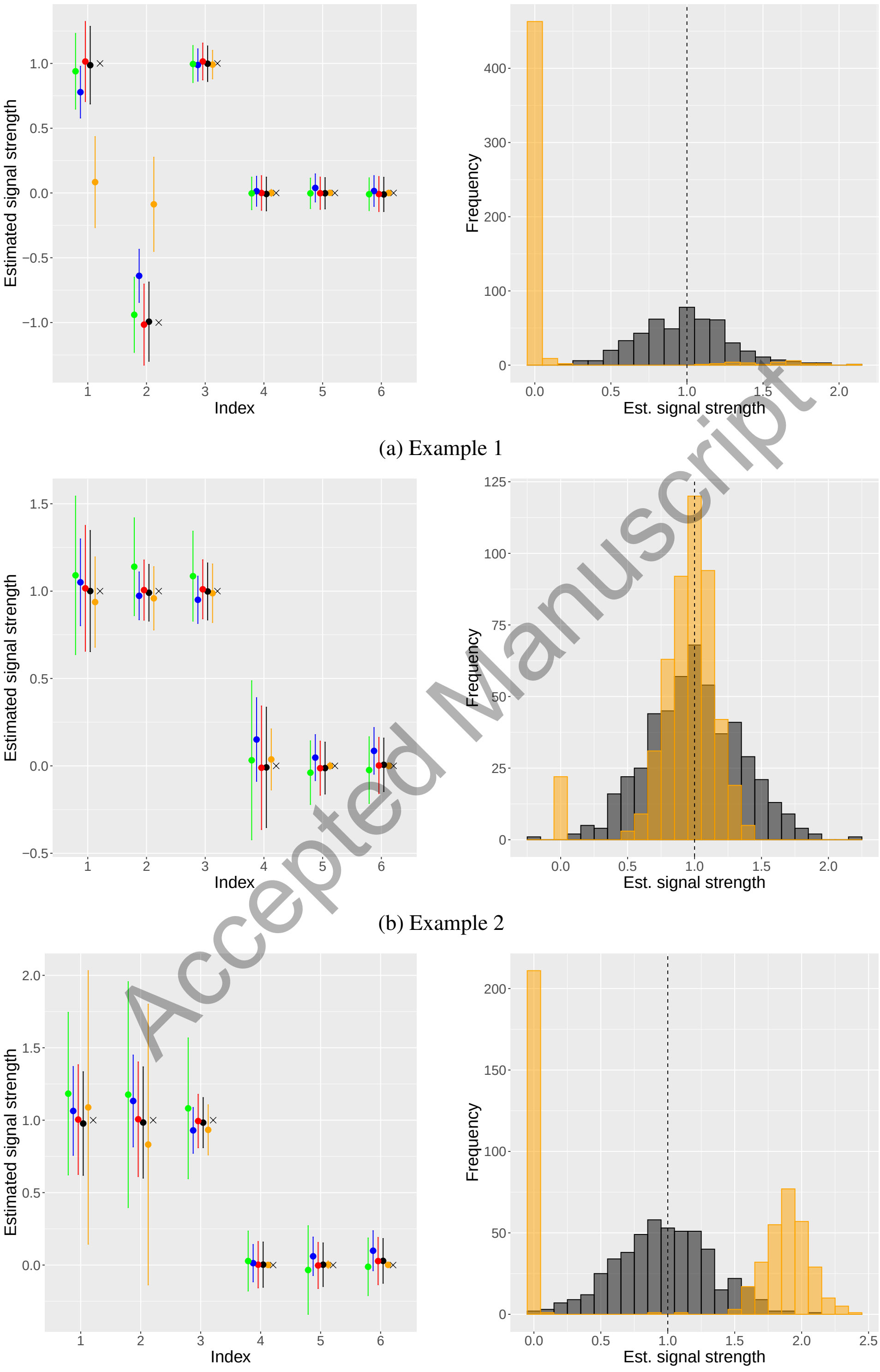
Due to the dependence structure among the covariates, these scenarios exemplify situations where the individual entries of β^* are difficult to estimate accurately. In $R = 500$ Monte Carlo replications, we obtained the aforementioned estimates of β^*. To probe the inferential capabilities beyond point estimation, for all except the SCAD penalised regression, we also computed 95% confidence intervals for β_1^* and β_4^*, the former corresponding to a signal variable and the latter to a noise variable. Motivated by the results in section 5.4, we defined a ϑ-level test for the hypothesis $H_0 : \alpha^T\beta^* = b_0$ based on the least squares estimator of the form 515

$$\Psi(Y;\vartheta,\alpha) = \begin{cases} 0 & \text{if } |T| \le z_{1-\vartheta/2}, \\ 1 & \text{if } |T| > z_{1-\vartheta/2}, \end{cases} \qquad T = \frac{\alpha^T\tilde{\beta}^* - b_0}{\hat{\varsigma}^{-1}\|\alpha^T(X^TX)^{-1}X^T\hat{\Gamma}^{1/2}\|_2} \tag{9}$$

where $z_{1-\vartheta/2}$ is the $(1-\vartheta/2)$ quantile of the $N(0,1)$ distribution and $\hat{\Gamma}$ is the diagonal matrix with entries $\hat{\Gamma}_{ii} = 1 - \tanh^2(\hat{\eta}_i/2)$ for $i = 1,\ldots,n$. Approximate confidence intervals were constructed by inverting this test. No guarantees have been provided for this construction to date, 520 the numerical results serve only to gain insights into the relevance of the term Π_2 in Proposition 5. The R functions `cv.ncvreg`, `logistf`, `lasso.proj` in the `hdi` package and `LF` in the `SIHR` package were used to compute the estimates and confidence intervals. As `ncvreg` necessarily includes an intercept term whereas `lasso.proj` does not, we fitted all models without an intercept except SCAD. This is likely to marginally favour the SCAD and OLS-SCAD results. 525

The left column of Figure 1 shows the average estimated signal strength of entries 1-6 of β^* for each example. The right column shows the distribution of estimates of β_1^* obtained via SCAD and OLS-SCAD. The results show that OLS-SCAD was able to correct the bias in the SCAD estimates. In Example 1, the SCAD estimate of β^* failed to accurately characterise the two signal variables that were only weakly marginally related to the response variable, whereas 530 OLS-SCAD was able to estimate all signal strengths accurately. In Example 2, the estimates of the highly correlated noise and signal variables were less biased for OLS-SCAD than for SCAD. This was because SCAD estimated the signal strength of the second signal variable to be zero in a non-negligible portion of cases. Finally, in Example 3, SCAD often assigned most of the signal strength corresponding to the two highly correlated signal variables to a single signal variable, 535 whereas OLS-SCAD spread the signal more evenly across the two variables. The other three estimators performed similarly to OLS-SCAD in terms of estimation.

Whilst the examples show that OLS-SCAD is able to estimate the effects of signal variables with improved accuracy over the estimator obtained directly via SCAD, it is often the case that



(a) Example 1

(b) Example 2

(c) Example 3

Fig. 1: Left column: average estimated signal strengths of entries 1-6 of β^* in each example obtained using OLS-SCAD (black), SCAD penalised regression (orange), Firth's estimator (red), desparsified LASSO (blue) and LSW (green). Error bars show one estimated standard deviation. True signal strengths are marked with black crosses. Right column: histogram of estimates of β_1^* obtained via OLS-SCAD (black) and SCAD penalised regression (orange) for the second signal variable in each example. The black dashed line represents the true signal strength.

	Method	$\lvert\mathcal{I}\rvert$	$\mathrm{pr}(\beta_i^* \in \mathcal{I})$	$\lvert\mathcal{I}\rvert$	$\mathrm{pr}(\beta_i^* \in \mathcal{I})$	$\mathrm{pr}(0 \notin \mathcal{I})$
		(null variable)		(signal variable)		
Ex. 1	OLS-SCAD	0.49 (0.01)	0.94 (0.24)	1.14 (0.02)	0.94 (0.24)	0.93 (0.26)
	FIRTH	0.50 (0.01)	0.94 (0.24)	1.19 (0.03)	0.94 (0.23)	0.93 (0.26)
	DLASSO	0.44 (0.01)	0.95 (0.21)	0.81 (0.01)	0.81 (0.39)	0.97 (0.17)
	LSW	0.54 (0.02)	0.97 (0.16)	1.28 (0.05)	0.96 (0.20)	0.84 (0.37)
Ex.2	OLS-SCAD	1.39 (0.04)	0.97 (0.18)	1.36 (0.04)	0.95 (0.21)	0.83 (0.37)
	FIRTH	1.42 (0.05)	0.96 (0.20)	1.45 (0.05)	0.96 (0.19)	0.80 (0.40)
	DLASSO	0.99 (0.03)	0.92 (0.27)	1.00 (0.03)	0.95 (0.21)	0.99 (0.11)
	LSW	2.55 (0.47)	0.98 (0.13)	2.76 (0.72)	0.99 (0.09)	0.26 (0.44)
Ex. 3	OLS-SCAD	0.63 (0.02)	0.96 (0.20)	1.45 (0.04)	0.96 (0.20)	0.76 (0.43)
	FIRTH	0.64 (0.02)	0.95 (0.21)	1.53 (0.06)	0.96 (0.19)	0.76 (0.43)
	DLASSO	0.55 (0.01)	0.98 (0.15)	1.07 (0.03)	0.92 (0.27)	0.96 (0.20)
	LSW	1.44 (0.28)	0.99 (0.10)	3.67 (0.92)	0.99 (0.09)	0.09 (0.29)

Table 1: Average length, $\lvert\mathcal{I}\rvert$, coverage probability, $\mathrm{pr}(\beta_i^* \in \mathcal{I})$, and power, $\mathrm{pr}(0 \in \mathcal{I})$, of the confidence intervals obtained via the corrected OLS estimator (OLS-SCAD), Firth's method (FIRTH), desparsified LASSO (DLASSO) and Cai et al.'s approach (LSW). Standard errors are given in brackets.

the latter has smaller cumulative estimation and prediction error, particularly when p is large and β^* is sparse. This is because the corrected versions of the OLS estimator are not sparse and so error is accumulated across all entries of the parameter vector.

Table 1 shows the average length and coverage probability of the confidence intervals. The power is also recorded. No intervals were obtained for SCAD due to the bias and highly non-Gaussian distribution of its estimates observed in Figure 1. The intervals obtained from OLS-SCAD and Firth's approach performed similarly and moderately better than the other two approaches, with OLS-SCAD producing slightly shorter intervals than Firth's approach on average for a coverage close to 95%. The LSW method produced the largest confidence intervals, resulting in uninformative intervals that contained both zero and the true non-zero signal strength in a large number of cases. The desparsified LASSO produced the shortest intervals, however this sometimes resulted in coverage probabilities substantially below the nominal level 0.95.

9. Discussion

9.1. *Extension to cases with $p \geq n$*

A limitation of our analysis is that corrections of the least squares estimator can only be used for inference in settings with $p < n$. The result is nevertheless relevant in the sparse $p > n$ setting. For instance, in Cox & Battey (2017) a large number of low-dimensional regressions are fitted, the motivation being that if a variable is causal, its explanatory power is, to a certain extent, preserved regardless of which other variables are included. Indeed, the original motivation for studying the problem of the present paper came from the difficulties in applying logistic regression in the context of Cox & Battey (2017) due to separable data. The corrected least squares estimator may be used as an alternative to logistic regression in this context.

9.2. *Extensions to other models*

Beyond its relevance to settings with separated data, there are additional benefits of the new approach that may be of interest beyond the logistic regression model. The method converts an estimation and inferential problem on the entries of β^* to a predictive one on $X\beta^*$, the latter typically being easier to solve without assuming strong conditions on the design matrix. As a result, the performance is favourable even in settings where covariates are highly correlated in sample. The numerical results of section 8 exhibit this most clearly, providing examples where a penalised regression estimator fails to accurately characterise the entries of the unknown parameter, yet when this estimator is used alongside the least squares estimator, the performance is improved.

Another favourable aspect is the method's adaptability to different forms of sparsity. In our analysis, sparsity is only assumed to ensure that a LASSO penalised regression produces a consistent estimator of $X\beta^*$. However, by making use of an alternative estimator of $X\beta^*$, there is considerable flexibility in the form that this assumption takes. This was briefly outlined in section 6.3. It would be of interest to determine whether a version of the corrected least squares estimator can be used with similar benefits in other models.

SUPPLEMENTARY MATERIAL

The supplementary file contains proofs of the theoretical results stated in the main paper and additional numerical simulations.

REFERENCES

ALBERT, A. & ANDERSON, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71**, 1–10.

BAI, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sinica* **9**.

BARTLETT, M. S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.* **32**, 560–566.

BARTLETT, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. London A* **160**, 268–282.

BATTEY, H. S., COX, D. R. & JACKSON, M. V. (2019). On the linear in probability model for binary data. *Royal Society Open Science* **6**.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

CAI, T. T., GUO, Z. & MA, R. (2021). Statistical inference for high-dimensional generalized linear models with binary outcomes. *J. Amer. Statist. Assoc.* **118**, 1319–1332.

CANDÈS, E. & SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* **48**, 27–42.

COOLEN, A. C. C., SHEIKH, M., MOZEIKA, A., AGUIRRE-LOPEZ, F. & ANTENUCCI, F. (2020). Replica analysis of overfitting in generalized linear regression models. *J. Phys. A* **53**, 365001.

COX, D. R. (1958). The regression analysis of binary sequences (with discussion). *R. Statist. Soc. B* **20**, 215–242.

COX, D. R. (1970). *Analysis of Binary Data*. Methuen.

COX, D. R. & BATTEY, H. S. (2017). Large numbers of explanatory variables, a semi-discriptive analysis. *Proc. Natl. Acad. Sci. USA* **114**, 8592–8595.

COX, D. R. & WERMUTH, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.

DUFFY, D. & SANTNER, T. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Comm. Statist. Theory Methods* **18**, 159–980.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FAN, J. & PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–961.

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

HEINZE, G. & SCHEMPER, M. (2002). A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419.

HUBER, P. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

KOSMIDIS, I. & FIRTH, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**, 71–82. 615

LEI, L., BICKEL, P. & EL KAROUI, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Probab. Theory Relat. Fields* **172**, 983–1079.

MA, R., CAI, T. T. & LI, H. (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Amer. Statist. Assoc.* **116**, 984–998.

MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30**, 1225–1267. 620

MEHTA, C. R. & PATEL, N. R. (1995). Exact logistic regression: theory and examples. *Stat. Med.* **14**, 2143–2160.

MEIER, L., VAN DE GEER, S. & BÜHLMANN, P. (2008). The group lasso for logistic regression. *J. R. Statist. Soc. B* **70**, 53–71.

NING, Y. & LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45**, 158–195. 625

RASKUTTI, G., WAINWRIGHT, M. & YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57**, 6976–6994.

SHI, C., SONG, R., LU, W. & LI, R. (2021). Statistical inference for high-dimensional models via recursive online-score estimation. *J. Amer. Statist. Assoc.* **116**, 1307–1318.

SUR, P. & CANDÈS, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 14516–14525. 630

TANG, Y. & REID, N. (2020). Modified likelihood root in high dimensions. *J. R. Stat. Soc. Ser. B.* **82**, 1349–1369.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

VAN DE GEER, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–1202. 635

WALD, A. (1950). *Statistical Decision Functions*. Wiley.

YADLOWSKY, S., YUN, T.AND MCLEAN, C. & D'AMOUR, A. (2021). SLOE: A faster method for statistical inference in high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc. 640

ZHAO, Q., SUR, P. & CANDÈS, E. (2020). The asymptotic distribution of the MLE in high-dimensional logistic models: arbitrary covariance. *arXiv:2001.09351* .

[*Received on ... ... Editorial decision on ... ...*]

Biometrika (...), ..., ..., *p.* 1
Printed in Great Britain

Supplementary material for "On inference in high-dimensional logistic regression models with separated data"

BY R. M. LEWIS AND H. S. BATTEY
Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7
rebecca.lewis15@imperial.ac.uk h.battey@imperial.ac.uk

SUMMARY

This supplementary file contains proofs of the theoretical results and additional simulations.

1. PROOFS OF MAIN RESULTS

Proof of Proposition 1. The confidence interval for β_1^* is unbounded above when t_1 is the maximum element of the set $\mathbb{T}_1$ and unbounded below when t_1 is the minimum. The result follows from Lemmas S1 and S2. □

Proof of Theorem 1. The case where the data are completely separated and $\beta_1 = 0$ follows directly from Lemma S2. Now consider the case where $\beta_1 > 0$ and suppose for a contradiction that $\forall B > 0$, there exists $b_B \geq B$ with $b_B \notin CS_\vartheta^{(1)}(t_1)$. When the data can be separated by β with $\beta_1 > 0$, Lemma S1 says the observation t_1 is the maximum value in the set $\mathbb{T}_1$. Hence, when $\beta^* = (b_B/\beta_1)\beta$ for some $B > 0$,

$$\begin{aligned}\mathrm{pr}(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) &= \frac{1}{1 + \sum_{u \in \mathbb{T}_1 \setminus \{t_1\}} \frac{c(u,t_2,\ldots,t_p)}{c(t_1,\ldots,t_p)} e^{\beta_1^*(u-t_1)}} \\ &= \frac{1}{1 + \sum_{u \in \mathbb{T}_1 \setminus \{t_1\}} \frac{c(u,t_2,\ldots,t_p)}{c(t_1,\ldots,t_p)} e^{b_B(u-t_1)}} \\ &\to 1,\end{aligned}$$

as $B \to \infty$. In particular, there exists $B > 0$ such that when $\beta^* = (b_B/\beta_1)\beta$,

$$\mathrm{pr}(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) \geq \vartheta. \tag{S1}$$

We will reach a contradiction by showing that this probability is also strictly less than ϑ. By definition, $CS_\vartheta^{(1)}(T_1)$ satisfies

$$\mathrm{pr}\{\beta_1^* \in CS_\vartheta^{(1)}(T_1) \mid T_2 = t_2, \ldots, T_p = t_p\} \geq 1 - \vartheta, \qquad \forall \beta_1^* \in \mathbb{R}.$$

In particular, this inequality should hold for $\beta_1^* = b_B$. But $\beta_1^* = b_B \notin CS_\vartheta^{(1)}(t_1)$ by assumption, and so when $\beta^* = (b_B/\beta_1)\beta$,

$$\begin{aligned}\mathrm{pr}(T_1 \neq t_1 \mid T_2 = t_2, \ldots, T_p = t_p) &\geq \mathrm{pr}\{\beta_1^* \in CS_\vartheta^{(1)}(T_1) \mid T_2 = t_2, \ldots, T_p = t_p\} \\ &\geq 1 - \vartheta\end{aligned} \tag{S2}$$

by definition of the confidence set. Inequalities (S1) and (S2) cannot both hold, and so we reach a contradiction. It follows that there exists $B > 0$ such that $[B, \infty) \subseteq CS^{(1)}_{\vartheta}(t_1)$. A similar argument establishes the case where $\beta_1 < 0$. □

Proof of Theorem 2. The proof closely follows the ideas in the proof of Theorem 1. Suppose $\beta_1 > 0$ and suppose for a contradiction that $\forall B > 0$, there exists $b_B \geq B$ with $b_B \notin CS^{(1)}_{\vartheta}(y)$. When $\beta^* = (b_B/\beta_1)\beta$ for some $B > 0$,

$$\begin{aligned} \text{pr}(Y = y) &= \prod_{i=1}^{n} \frac{1}{1 + e^{-(b_B/\beta_1) y_i x_i^T \beta}} \\ &= 2^{-m} \prod_{i : x_i^T \beta \neq 0} \frac{1}{1 + e^{-(b_B/\beta_1)|x_i^T \beta|}} \\ &\to 2^{-m}, \end{aligned}$$

as $B \to \infty$. In particular, there exists $B > 0$ such that when $\beta^* = (b_B/\beta_1)\beta$

$$\text{pr}(Y = y) \geq 2^{-m} - (2^{-m} - \vartheta) = \vartheta. \qquad \text{(S3)}$$

However, by definition, $CS_{\vartheta}(Y)$ satisfies

$$\text{pr}\{\beta^* \in CS_{\vartheta}(Y)\} \geq 1 - \vartheta, \qquad \forall \beta^* \in \mathbb{R}^p$$

and so this inequality should hold for $\beta^* = (b_B/\beta_1)\beta$. But for this choice of β^*, $\beta_1^* \notin CS^{(1)}_{\vartheta}(y)$ by construction and so $\beta^* \notin CS_{\vartheta}(y)$. It follows that,

$$\text{pr}(Y \neq y) \geq \text{pr}\{\beta^* \in CS_{\vartheta}(Y)\} \geq 1 - \vartheta \qquad \text{(S4)}$$

by definition of the confidence set. Inequalities (S3) and (S4) cannot both hold and so we reach a contradiction. Thus, there exists $B > 0$ such that $[B, \infty) \subseteq CS^{(1)}_{\vartheta}(y)$. A similar argument establishes the case where $\beta_1 < 0$.

Now assume the data are completely separated by β with $\beta_1 = 0$ and suppose for a contradiction that there exists $b \in \mathbb{R}$ with $b \notin CS^{(1)}_{\vartheta}(y)$. Define c_{ϑ} and c satisfying

$$c_{\vartheta} = \max\left\{0, \log\left(\frac{\vartheta^{1/n}}{1 - \vartheta^{1/n}}\right)\right\}, \qquad c = \frac{c_{\vartheta} + |b| \max_{i=1}^{n} |y_i x_{i1}|}{\min_{i=1}^{n} |x_i^T \beta|}.$$

Complete separation ensures c is well-defined. Consider $\beta^{*T} = c\beta^T + (b, 0, \ldots, 0)$. By definition of the confidence interval and the fact that $b \notin CS^{(1)}_{\vartheta}(y)$,

$$\text{pr}(Y \neq y) \geq \text{pr}\{\beta^* \in CS_{\vartheta}(Y)\} \geq 1 - \vartheta.$$

However,

$$\begin{aligned} \text{pr}(Y = y) &= \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i x_i^T \beta^*}} \\ &= \prod_{i=1}^{n} \frac{1}{1 + e^{-c|x_i^T \beta| - y_i x_{i1} b}} \\ &\geq \left(\frac{1}{1 + e^{-c_{\vartheta}}}\right)^n \\ &\geq \vartheta \end{aligned}$$

by definition of c and c_ϑ. Thus, we have reached a contradiction and so $CS_\vartheta^{(1)}(y) = \mathbb{R}$. □

Proof of Proposition 2. Fix $t > 0$. By Lemma S3, for all $\epsilon > 0$ there exists a set $\mathcal{N}_{\epsilon,d} \subseteq \mathcal{B}_d$ of cardinality at most

$$\left\{\frac{(2+\epsilon)ep}{\epsilon d}\right\}^d$$

such that for all possible matrices X,

$$\begin{aligned}\text{pr}\left\{\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\hat{\beta}^0-\beta^0|) \geq t\right\} &\leq \text{pr}\left\{\max_{\tilde{\alpha}\in\mathcal{N}_{\epsilon,d}} |\tilde{\alpha}^T(\hat{\beta}^0-\beta^0|) \geq t(1-\epsilon)\right\} \\ &\leq \left\{\frac{(2+\epsilon)ep}{\epsilon d}\right\}^d \max_{\tilde{\alpha}\in\mathcal{N}_{\epsilon,d}} \text{pr}\left\{|\tilde{\alpha}^T(\hat{\beta}^0-\beta^0|) \geq t(1-\epsilon)\right\}\end{aligned}$$

where we have applied a union bound to obtain the last inequality. Let $\varepsilon = Y - \mathbb{E}(Y)$ and write

$$\tilde{\alpha}^T(\hat{\beta}^0-\beta^0) = \sum_{i=1}^n v_i\varepsilon_i$$

where $v_i = \tilde{\alpha}^T(X^TX)^{-1}x_i$. The random variables ε_i lie in the range $[-2, 2]$ and hence, are sub-Gaussian with $\|\varepsilon_i\|_{\psi_2} \leq 2$. As they are also independent, $\tilde{\alpha}^T(\hat{\beta}^0-\beta^0)$ is sub-Gaussian with $\|\tilde{\alpha}^T(\hat{\beta}^0-\beta^0)\|_{\psi_2}$ bounded above by

$$\left(\sum_{i=1}^n v_i^2\right)^{1/2} = \|\tilde{\alpha}^T(X^TX)^{-1}X^T\|_2$$

up to a constant. Then, there exist constants $C, c_1, c_2 > 0$ not depending on t such that for all $X \in \mathbb{X}_B$

$$\begin{aligned}\text{pr}\left\{\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\hat{\beta}^0-\beta^0|) \geq t\right\} &\leq C\left\{\frac{(2+\epsilon)ep}{\epsilon d}\right\}^d \max_{\tilde{\alpha}\in\mathcal{N}_{\epsilon,d}} \exp\left\{\frac{-c_1t^2(1-\epsilon)^2}{\|\tilde{\alpha}^T(X^TX)^{-1}X^T\|_2^2}\right\} \\ &\leq C\exp\left\{-c_2t^2(1-\epsilon)^2n + d\log\left(C_\epsilon p/d\right)\right\}\end{aligned}$$

for some constant C_ϵ depending only on ϵ. By assumption, this bound converges to zero as $n \to \infty$ and so the result follows. □

Proof of Proposition 3. The estimation error can be written as $\|\hat{\beta}^0-\beta^0\|_2 = \|A\varepsilon\|_2$ where $A = (X^TX)^{-1}X^T$ and $\varepsilon = Y - \mathbb{E}(Y)$ consists of independent and centred sub-Gaussian random variables with $\max_{i=1}^n \|\varepsilon_i\|_{\psi_2} \leq 2$. Using the Hanson–Wright inequality (Theorem 6.2.1 in Vershynin (2018)), for any $t \geq 0$,

$$\text{pr}\left\{\left|\|A\varepsilon\|_2^2 - \mathbb{E}(\|A\varepsilon\|_2^2)\right| \geq t\right\} \leq 2\text{exp}\left(-c\min\left\{\frac{t^2}{\|A^TA\|_F^2}, \frac{t}{\|A^TA\|_2}\right\}\right)$$

for some constant $c > 0$. Note,

$$\|A^TA\|_F^2 \leq p\|A^TA\|_2^2.$$

Further, $\|A^TA\|_2^2 = \lambda_{\max}\{(X^TX)^{-2}\}$. Thus for any $t > 0$,

$$\text{pr}\left\{\left|\|A\varepsilon\|_2^2 - \mathbb{E}(\|A\varepsilon\|_2^2)\right| \geq t\right\} \leq 2\text{exp}\left[-c\min\left\{\frac{t^2n^2}{\lambda_{\max}\{(X^TX/n)^{-1}\}^2p}, \frac{tn}{\lambda_{\max}\{(X^TX/n)^{-1}\}}\right\}\right].$$

As $p < n$ and $\lambda_{\max}\{(X^TX/n)^{-1}\} = O(1)$, the right hand side is $o(1)$ and so

$$\|A\varepsilon\|_2^2 - \mathbb{E}(\|A\varepsilon\|_2^2) = o_p(1).$$

As

$$\begin{aligned}\mathbb{E}(\|A\varepsilon\|_2^2) &= \text{tr}(\Gamma A^TA)\\ &\leq \|A\|_F^2\\ &\leq p\|A\|_2^2\\ &\leq p/n\lambda_{\max}\{(X^TX/n)^{-1}\}\end{aligned}$$

the expectation is asymptotically bounded and so $\|A\varepsilon\|_2^2 = O_p(1)$. When $p = o(n)$, the expectation is $o(1)$ and so $\|A\varepsilon\|_2^2 = o_p(1)$. □

Proof of Proposition 4. Let $\varepsilon = Y - \mathbb{E}(Y)$ and write

$$B_n^{-1}\alpha^T(\hat\beta^0 - \beta^0) = \sum_{i=1}^n B_n^{-1}v_i(\alpha)\varepsilon_i$$

where $v_i(\alpha) = \alpha^T(X^TX)^{-1}x_i$. Each term $B_n^{-1}v_i(\alpha)\varepsilon_i$ is an independent random variable with zero mean and variance bounded above by one. Define

$$\begin{aligned}\Lambda_n(\alpha) &= \sum_{i=1}^n \mathbb{E}\{|B_n^{-1}v_i(\alpha)\varepsilon_i|^2\mathbb{1}\{|B_n^{-1}v_i(\alpha)\varepsilon_i| \geq 1\}\}\\ l_n(\alpha) &= \sum_{i=1}^n \mathbb{E}\{|B_n^{-1}v_i(\alpha)\varepsilon_i|^3\mathbb{1}\{|B_n^{-1}v_i(\alpha)\varepsilon_i| < 1\}\}.\end{aligned}$$

By Theorem 5.8 in Petrov (1995, p. 154), there exists some absolute constant $A > 0$ such that

$$\sup_{x\in\mathbb{R}}\left|\text{pr}\{B_n^{-1}\alpha^T(\hat\beta^0 - \beta^0) \leq x\} - \Phi(x)\right| \leq A\{\Lambda_n(\alpha) + l_n(\alpha)\}.$$

As ε_k is a bounded random variable and $B_n^{-1}\max_{i=1}^n|v_i(\alpha)| = o(1)$ uniformly over $\alpha \in \mathcal{R}$, the event $\max_{i=1,\ldots,n}|B_n^{-1}v_i(\alpha)\varepsilon_i| < 1$ has probability one when n is large enough. Thus,

$$\sup_{\alpha\in\mathcal{R}}\Lambda_n(\alpha) = o(1).$$

Further,

$$l_n(\alpha) \leq \sum_{i=1}^n B_n^{-3}|v_i(\alpha)|^3\mathbb{E}(|\varepsilon_i|^3) \leq 2B_n^{-1}\max_{i=1}^n|v_i(\alpha)|$$

as $|\varepsilon_i| \leq 2$ and $\sum_{i=1}^n B_n^{-2}|v_i(\alpha)|^2\mathbb{E}(\varepsilon_i^2) = 1$. Then

$$\sup_{\alpha\in\mathcal{R}} l_n(\alpha) = o(1)$$

and so the result follows. □

Proof of Theorem 3. Fix $t > 0$. For $\xi < 1$, define A_ξ be the event that $h(\hat\eta, \eta^*) \leq \xi$. By definition of $\mathcal{H}_B$,

$$\sup_{X\in\mathcal{H}_B}\text{pr}(A_\xi^c) = o(1). \tag{S5}$$

Our aim is to show that for appropriately chosen ξ,

$$\sup_{X\in\mathcal{H}_B} \mathrm{pr}\left\{ \sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\tilde{\beta}^* - \beta^*)| \geq t \mid A_\xi \right\} = o(1), \tag{S6}$$

in which case the result follows from

$$\sup_{X\in\mathcal{H}_B} \mathrm{pr}\left\{ \sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\tilde{\beta}^* - \beta^*)| \geq t \right\} \leq \sup_{X\in\mathcal{H}_B} \mathrm{pr}\left\{ \sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\tilde{\beta}^* - \beta^*)| \geq t \mid A_\xi \right\} + \sup_{X\in\mathcal{H}_B} \mathrm{pr}(A_\xi^c).$$

Assume the event A_ξ holds and recall that $\beta^* = \varsigma^{-1}(\beta^0 - \delta)$ and $\tilde{\beta}^* = \hat{\varsigma}^{-1}(\hat{\beta}^0 - \hat{\delta})$. Then,

$$\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\tilde{\beta}^* - \beta^*)| \leq \Pi_1 + \Pi_2 + \Pi_3$$

where

$$\begin{aligned}
\Pi_1 &= \sup_{\alpha\in\mathcal{B}_d} \hat{\varsigma}^{-1}|\alpha^T(\hat{\beta}^0 - \beta^0)| \\
\Pi_2 &= \sup_{\alpha\in\mathcal{B}_d} \hat{\varsigma}^{-1}|\alpha^T(\hat{\delta} - \delta)| \\
\Pi_3 &= \sup_{\alpha\in\mathcal{B}_d} \left|\hat{\varsigma}^{-1} - \varsigma^{-1}\right| |\alpha^T(\beta^0 - \delta)|.
\end{aligned}$$

By Lemmas S5 and S6, and our assumptions, there exist constants $C, N > 0$ such that when $X \in \mathcal{H}_B$,

$$|\hat{\varsigma} - \varsigma| \leq C\xi, \qquad \sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\hat{\delta} - \delta)| \leq C\xi$$

and when $n \geq N$,

$$\max\{d, \|\beta^*\|_\infty\} \leq C.$$

As $\varsigma^{-1} \leq B$ when $X \in \mathcal{H}_B$, it follows that when ξ is small enough,

$$\hat{\varsigma}^{-1} \leq \frac{1}{\varsigma - C\xi} \leq \frac{1}{B^{-1} - C\xi},$$

and

$$|\hat{\varsigma}^{-1} - \varsigma^{-1}| \leq |\hat{\varsigma}^{-1}\varsigma^{-1}||\hat{\varsigma} - \varsigma| \leq \frac{BC\xi}{B^{-1} - C\xi}.$$

On choosing

$$\xi < \min\left\{1, (BC)^{-1} \min\left\{\frac{t}{3+t}, \frac{t}{t + 3C^2B}\right\}\right\}$$

it follows that $\Pi_2 < t/3$ and $\Pi_3 < t/3$ when $n \geq N$, where we have used the fact that $\alpha^T(\beta^0 - \delta) = \varsigma\alpha^T\beta^* \leq d\|\beta^*\|_\infty \leq C^2$. Thus, for $n \geq N$,

$$\sup_{X\in\mathcal{H}_B} \mathrm{pr}(\Pi_2 \geq t/3 \mid A_\xi) = \sup_{X\in\mathcal{H}_B} \mathrm{pr}(\Pi_3 \geq t/3 \mid A_\xi) = 0.$$

Further, by Proposition 2 and equation (S5),

$$\sup_{X\in\mathcal{H}_B} \mathrm{pr}(\Pi_1 \geq t/3 \mid A_\xi) \leq \frac{\sup_{X\in\mathcal{H}_B} \mathrm{pr}\{\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\hat\beta^0-\beta^0)| \geq t(B^{-1}-C\xi)/3\}}{1-\sup_{X\in\mathcal{H}_B}\mathrm{pr}(A_\xi^c)} = o(1).$$

Applying a union bound,

$$\sup_{X\in\mathcal{H}_B} \mathrm{pr}\left\{\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\tilde\beta^*-\beta^*)| \geq t \mid A_\xi\right\} \leq \sum_{i=1}^{3}\sup_{X\in\mathcal{H}_B}\mathrm{pr}(\Pi_i \geq t/3 \mid A_\xi) = o(1)$$

and so the result in (S6) follows. □

Proof of Theorem 4. The estimation error may be bounded by

$$\begin{aligned}\|\tilde\beta^*-\beta^*\|_2 &= \|\hat\varsigma^{-1}(\hat\beta^0-\hat\delta)-\varsigma^{-1}(\beta^0-\delta)\|_2\\ &\leq |\hat\varsigma^{-1}-\varsigma^{-1}|\|\beta^0-\delta\|_2 + |\hat\varsigma^{-1}|(\|\hat\beta^0-\beta^0\|_2+\|\hat\delta-\delta\|_2).\end{aligned}$$

Consider each of the terms. By Lemma S5 and the assumption that $\varsigma^{-1}=O(1)$, we have $\hat\varsigma^{-1}=O_P(1)$ and $|\hat\varsigma^{-1}-\varsigma^{-1}|=o_P(1)$. Also, $\|\beta^0-\delta\|_2=\varsigma\|\beta^*\|_2=O(\sqrt{p})$ by assumption and the fact that $\varsigma\leq 1/2$. For the second term, Proposition 3 shows that

$$\|\hat\beta^0-\beta^0\|_2 = o_p(\sqrt{p}).$$

By the proof of Lemma S6,

$$\begin{aligned}\|\hat\delta-\delta\|_2 &\leq \|(X^TX)^{-1}X^T\|_2\|(P_X-P_{\hat\eta})\tanh(\hat\eta/2)-(P_X-P_{\eta^*})\tanh(\eta^*/2)\|_2\\ &= O\{h(\hat\eta,\eta^*)\}\\ &= o_P(1).\end{aligned}$$

Combining results, we conclude that $p^{-1/2}\|\tilde\beta^*-\beta^*\|_2=o_P(1)$. □

Proof of Proposition 5. Recall that $\beta^*=\varsigma^{-1}(\beta^0-\delta)$ and $\tilde\beta^*=\hat\varsigma^{-1}(\hat\beta^0-\hat\delta)$. The statistic of interest may then be decomposed as

$$\frac{\alpha^T(\tilde\beta^*-\beta^*)}{\varsigma^{-1}\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2} = \Pi_1+\Pi_2+\Pi_3+\Pi_4 \tag{S7}$$

where

$$\begin{aligned}\Pi_1 &= \frac{\alpha^T(\hat\beta^0-\beta^0)}{\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2}\\ \Pi_2 &= -\frac{\varsigma}{\hat\varsigma}\left\{\frac{\alpha^T(\hat\delta-\delta)}{\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2}\right\}\\ \Pi_3 &= \left(\frac{\varsigma}{\hat\varsigma}-1\right)\frac{\alpha^T(\hat\beta^0-\beta^0)}{\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2}\\ \Pi_4 &= \left(\frac{\varsigma}{\hat\varsigma}-1\right)\frac{\alpha^T(\beta^0-\delta)}{\|\alpha^T(X^TX)^{-1}X^T\Gamma^{1/2}\|_2}.\end{aligned}$$

By Proposition 4, Π_1 converges in distribution to a standard normal random variable. By assumption $\varsigma^{-1}=O(1)$, and so applying Lemma S5, we have $\varsigma-\hat\varsigma=o_P(1)$ and $\hat\varsigma^{-1}=O_P(1)$. Combining this with Proposition 4, it follows that $\Pi_3=o_P(1)$. Under the null hypothesis $H_0:\alpha^T\beta^*=0$, we have $\Pi_4=0$ and so the result follows. □

Proof of Theorem 5. By definition,

$$n^{-1/2}\|X(\tilde{\beta}^* - \beta^*)\|_2 = n^{-1/2}\|\hat{\varsigma}^{-1}X(\hat{\beta}^0 - \hat{\delta}) - \varsigma^{-1}X(\beta^0 - \delta)\|_2$$

and so we can write

$$\big|\|\Pi_1\|_2 - \|\Pi_2 + \Pi_3 + \Pi_4\|_2\big| \leq n^{-1/2}\|X(\tilde{\beta}^* - \beta^*)\|_2 \leq \|\Pi_1\|_2 + \|\Pi_2 + \Pi_3 + \Pi_4\|_2$$

where

$$\begin{aligned}
\Pi_1 &= n^{-1/2}\varsigma^{-1}X(\hat{\beta}^0 - \beta^0)\\
\Pi_2 &= n^{-1/2}(\hat{\varsigma}^{-1} - \varsigma^{-1})X(\hat{\beta}^0 - \beta^0)\\
\Pi_3 &= -n^{-1/2}(\varsigma^{-1} - \hat{\varsigma}^{-1})X(\beta^0 - \delta)\\
\Pi_4 &= -n^{-1/2}\hat{\varsigma}^{-1}X(\hat{\delta} - \delta).
\end{aligned}$$

By Lemma S7, $\|\Pi_1\|_2$ converges in probability to $n^{-1/2}\varsigma^{-1}\|P_X\Gamma^{1/2}\|_F$. We now show that $\|\Pi_2 + \Pi_3 + \Pi_4\|_2 = o_p(1)$ by showing that the individual terms are $o_p(1)$. By Lemma S5 and the assumption that $\varsigma^{-1} = O(1)$, it follows that $|\hat{\varsigma}^{-1} - \varsigma^{-1}| = o_P(1)$. As the projection matrix P_X has rank p,

$$n^{-1/2}\|P_X\Gamma^{1/2}\|_F \leq n^{-1/2}\|P_X\|_F = (p/n)^{1/2} \leq 1,$$

and so by Lemma S7,

$$n^{-1/2}\|X(\hat{\beta}^0 - \beta^0)\|_2 = n^{-1/2}\|P_X\Gamma^{1/2}\|_F + o_P(1) = O_P(1).$$

Then, $\|\Pi_2\|_2 = o_P(1)$. For Π_3, $X(\beta^0 - \delta) = \varsigma X\beta^* = \varsigma\eta^*$ by definition, and

$$n^{-1/2}\varsigma\|\eta^*\|_2 = \frac{\eta^{*T}\tanh(\eta^*/2)}{\sqrt{n}\|\eta^*\|_2} \leq \frac{\|\tanh(\eta^*/2)\|_2}{\sqrt{n}} \leq 1.$$

Thus, $\|\Pi_3\|_2 = o_P(1)$. Finally, $\hat{\varsigma}^{-1} = O_P(1)$ as $\varsigma^{-1} = O(1)$. By the proof of Lemma S6

$$\begin{aligned}
n^{-1/2}\|X(\hat{\delta} - \delta)\|_2 &= n^{-1/2}\|(P_X - P_{\hat{\eta}})\tanh(\hat{\eta}/2) - (P_X - P_{\eta^*})\tanh(\eta^*/2)\|_2\\
&= O\{h(\hat{\eta}, \eta^*)\},
\end{aligned}$$

and so $\|\Pi_4\|_2 = o_P(1)$. Thus the terms Π_2, Π_3 and Π_4 are all $o_P(1)$, and so the result follows.□

Proof of Proposition 6. The maximum likelihood estimator satisfies

$$\hat{\beta}^* = \operatorname*{argmax}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \Big\{ \Big(\frac{Y_i + 1}{2}\Big) x_i^T\beta - \log\big(1 + e^{x_i^T\beta}\big) \Big\}.$$

Taking derivatives,

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \left(\frac{Y_i + 1}{2} - \frac{e^{x_i^T\hat{\beta}^*}}{1 + e^{x_i^T\hat{\beta}^*}} \right) x_i\\
&= \frac{X^TY}{2} - \sum_{i=1}^{n} \frac{(e^{x_i^T\hat{\beta}^*} - 1)x_i}{2(1 + e^{x_i^T\hat{\beta}^*})}\\
&= \frac{1}{2}\Big\{X^TY - X^T\tanh(X\hat{\beta}^*/2)\Big\}
\end{aligned}$$

and so

$$\hat{\beta}^0 = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T\tanh(X\hat{\beta}^*/2). \tag{S8}$$

Write

$$\tanh(X\hat{\beta}^*/2) = \hat{\varsigma}X\hat{\beta}^* + u + \hat{\Delta}$$

where $u \in \text{Col-Sp}(X)^\perp$, $\hat{\varsigma}$ is defined in the statement of the proposition and $\hat{\Delta} = (P_X - P_{X\hat{\beta}^*})\tanh(X\hat{\beta}^*/2)$. Then, by (S8),

$$\hat{\beta}^0 = \hat{\varsigma}\hat{\beta}^* + \hat{\delta},$$

and so the result follows by inverting this relationship. □

Proof of Proposition 7. Deferred to section 3 of the supplementary material. □

Proof of Proposition 8. Throughout this proof, we will use $\text{pr}_{Y|X}(\cdot)$, $\text{pr}_{Y,X}(\cdot)$ and $\text{pr}_X(\cdot)$ to denote the probabilities under the conditional distribution of Y given X, the joint distribution of Y and X, and the marginal distribution of X. Fix $t_1, t_2 > 0$. Our aim is to show that there exists $N > 0$ such that

$$\text{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t_1) \leq t_2, \qquad \forall n \geq N.$$

Define $\mathcal{H}_B = \mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ for every $B > 0$. Lemmas S8-S11 show that there exist $B, N_1 > 0$ such that

$$\text{pr}_X(X \in \mathcal{H}_B) > 1 - t_2/2, \qquad \forall n > N_1. \tag{S9}$$

We focus on this choice of B from now on. When $s = \|\beta^*\|_0 = O(n^{1/2-\xi})$, Proposition 7 shows that for all $t_1 > 0$,

$$\sup_{X \in \mathcal{H}_B} \text{pr}_{Y|X}\{h(X\hat{\beta}^{(\lambda)}, X\beta^*) > t_1\} = o(1).$$

Further $\|\beta^*\|_\infty = O(1)$ as $\|\beta^*\|_\infty \leq \|\beta^*\|_2$ and

$$\|\beta^*\|_2^2 = \frac{\beta^{*T}\beta^*}{\beta^{*T}\Sigma\beta^*} \times (\beta^{*T}\Sigma\beta^*) \leq \frac{\beta^{*T}\Sigma\beta^*}{\lambda_{\min}} = O(1)$$

by assumption. Applying Theorem 3, there exists N_2 such that

$$\sup_{X \in \mathcal{H}_B} \text{pr}_{Y|X}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t_1) \leq t_2/2, \qquad \forall n \geq N_2. \tag{S10}$$

Let $N = \max\{N_1, N_2\}$ and assume $n \geq N$. Then, by (S9) and (S10),

$$\begin{aligned}\text{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t_1) &\leq \int_{X \in \mathcal{H}_B} \text{pr}_{Y|X}(\|\tilde{\beta}^* - \beta^*\|_\infty \geq t_1)dF(X) + \text{pr}_X(X \notin \mathcal{H}_B)\\ &\leq t_2\end{aligned}$$

where F is the distribution function of X. The first result follows. To obtain the second statement, let $\mathcal{E}$ be the event that the data (Y, X) are separated. Candès & Sur (2020) showed that $\text{pr}(\mathcal{E})$ converges to one in the setting of interest. So,

$$\text{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty < t_1, \mathcal{E}) \geq \text{pr}_{Y,X}(\|\tilde{\beta}^* - \beta^*\|_\infty < t_1) + \text{pr}_{Y,X}(\mathcal{E}) - 1$$

which converges to one. □

2. PROOFS OF ADDITIONAL RESULTS

LEMMA S1. *Suppose the observed data are separated by $\beta \in \mathbb{R}^p \backslash \{0\}$. If $\beta_1 > 0$ then t_1 is the largest element in the set $\mathbb{T}_1$. If $\beta_1 < 0$ then t_1 is the smallest element in the set $\mathbb{T}_1$.*

Proof. Fix $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_n)^T \in \mathbb{C}_1$, let $\mathcal{A} = \{i : z_i \neq \tilde{z}_i\}$ denote the set of indices where z and $\tilde{z}$ differ, and let $\tilde{y}_i = 2\tilde{z}_i - 1$. Then, for all $k \neq 1$,

$$\begin{aligned}\sum_{i=1}^n x_{ik} z_i = \sum_{i=1}^n x_{ik} \tilde{z}_i &\iff \sum_{i=1}^n x_{ik} y_i = \sum_{i=1}^n x_{ik} \tilde{y}_i \\ &\iff \sum_{i\in\mathcal{A}} x_{ik} y_i = \sum_{i\in\mathcal{A}} x_{ik} \tilde{y}_i \\ &\iff \sum_{i\in\mathcal{A}} x_{ik} y_i = -\sum_{i\in\mathcal{A}} x_{ik} y_i \\ &\iff \sum_{i\in\mathcal{A}} x_{ik} y_i = 0.\end{aligned}$$

As $y_i x_i^T \beta \geq 0$ for all $i = 1, \ldots, n$,

$$\begin{aligned}0 &\leq \sum_{i\in\mathcal{A}} y_i x_i^T \beta \\ &= \sum_{k=1}^p \beta_k \sum_{i\in\mathcal{A}} y_i x_{ik} \\ &= \beta_1 \sum_{i\in\mathcal{A}} y_i x_{i1} \qquad \text{(S11)}\end{aligned}$$

by the arguments above and the definition of $\mathbb{C}_1$. Consider possible cases for β_1. When $\beta_1 > 0$, inequality (S11) implies $\sum_{i\in\mathcal{A}} y_i x_{i1} \geq 0$. Then

$$\begin{aligned}\sum_{i=1}^n x_{i1} z_i - \sum_{i=1}^n x_{i1} \tilde{z}_i &= \sum_{i=1}^n (y_i - \tilde{y}_i) x_{i1}/2 \\ &= \sum_{i\in\mathcal{A}} y_i x_{i1} \\ &\geq 0\end{aligned}$$

and so it follows that $\sum_{i=1}^n x_{i1} z_i \geq \sum_{i=1}^n x_{i1} \tilde{z}_i$. As $\tilde{z}$ was arbitrary, the result holds for all $\tilde{z} \in \mathbb{C}_1$ and so $t_1 = \sum_{i=1}^n x_{i1} z_i$ is the largest element of the set $\mathbb{T}_1$. When $\beta_1 < 0$, we must have $\sum_{i\in\mathcal{A}} y_i x_{i1} \leq 0$. The result follows analogously, this time showing that t_1 is the smallest element of the set $\mathbb{T}_1$. □

LEMMA S2. *Suppose the observed data are completely separated by $\beta \in \mathbb{R}^p \backslash \{0\}$. If $\beta_1 = 0$ then t_1 is the unique element of the set $\mathbb{T}_1$ and*

$$pr(T_1 = t_1 \mid T_2 = t_2, \ldots, T_p = t_p) = 1, \qquad \forall \beta^* \in \mathbb{R}^p.$$

Proof. Using the notation and results in Lemma S1, when the data are separated and $\beta_1 = 0$ it must hold that

$$y_i x_i^T \beta = 0, \qquad \forall i \in \mathcal{A}$$

by (S11). This establishes the fact z and $\tilde{z} \in \mathbb{C}_1$ can only differ at indices i where $x_i^T \beta = 0$. When the data are completely separated, $x_i^T \beta > 0$ for all $i = 1, \ldots, n$. Thus, $\mathbb{C}_1$, and hence $\mathbb{T}_1$, contains a unique element. It follows that the conditional probability is equal to one for all values of the unknown parameter. □

LEMMA S3. *Let $\mathcal{B}_d = \{\alpha \in \mathbb{R}^p : \|\alpha\|_0 \leq d, \|\alpha\|_2 \leq 1\}$. For any $\epsilon > 0$, there exists $\mathcal{N}_{\epsilon,d} \subseteq \mathcal{B}_d$ such that*

$$\sup_{\alpha \in \mathcal{B}_d} |\alpha^T(\hat{\beta}^0 - \beta^*)| \leq (1-\epsilon)^{-1} \max_{\tilde{\alpha} \in \mathcal{N}_{\epsilon,d}} |\tilde{\alpha}^T(\hat{\beta}^0 - \beta^*)|$$

and

$$|\mathcal{N}_{\epsilon,d}| \leq \left\{ \frac{(2+\epsilon)ep}{\epsilon d} \right\}^d .$$

Proof. The following arguments closely resemble those in Appendix A.4 of Fan et al. (2018). For a set $\tilde{\mathcal{S}} \subseteq \{1, \ldots, p\}$ of size d, define $\mathcal{B}(\tilde{\mathcal{S}}) = \{\alpha \in \mathbb{R}^p : \text{support}(\alpha) \subseteq \tilde{\mathcal{S}}, \|\alpha\|_2 \leq 1\}$ and $\mathcal{N}_\epsilon(\tilde{\mathcal{S}}) \subseteq \mathcal{B}(\tilde{\mathcal{S}})$ to be an ϵ-net (see Definition 4.2.1 in Vershynin (2018)) of $\mathcal{B}(\tilde{\mathcal{S}})$ of minimum size. Write

$$\mathcal{B}_d = \bigcup_{\substack{\tilde{\mathcal{S}} \subseteq \{1,\ldots,p\} \\ |\tilde{\mathcal{S}}| = d}} \mathcal{B}(\tilde{\mathcal{S}}), \qquad \mathcal{N}_{\epsilon,d} = \bigcup_{\substack{\tilde{\mathcal{S}} \subseteq \{1,\ldots,p\} \\ |\tilde{\mathcal{S}}| = d}} \mathcal{N}_\epsilon(\tilde{\mathcal{S}})$$

where $\mathcal{N}_{\epsilon,d}$ is an ϵ-net of $\mathcal{B}_d$ satisfying

$$|\mathcal{N}_{\epsilon,d}| \leq \binom{p}{d} \left(1 + \frac{2}{\epsilon}\right)^d \leq \left\{ \frac{(2+\epsilon)ep}{\epsilon d} \right\}^d$$

by Lemma 4.2.13 in Vershynin (2018) and equation (A.12) in Fan et al. (2018). For any $\alpha \in \mathcal{B}(\tilde{\mathcal{S}})$, there exists $\tilde{\alpha} \in \mathcal{N}_\epsilon(\tilde{\mathcal{S}})$ such that $\|\alpha - \tilde{\alpha}\|_2 \leq \epsilon$ and the support of $\alpha - \tilde{\alpha}$ is a subset of $\tilde{\mathcal{S}}$. So,

$$\begin{aligned} |\alpha^T(\hat{\beta}^0 - \beta^*)| &\leq |(\alpha - \tilde{\alpha})^T(\hat{\beta}^0 - \beta^*)| + |\tilde{\alpha}^T(\hat{\beta}^0 - \beta^*)| \\ &\leq \epsilon \sup_{a \in \mathcal{B}(\tilde{\mathcal{S}})} |a^T(\hat{\beta}^0 - \beta^*)| + \max_{\tilde{\alpha} \in \mathcal{N}_\epsilon(\tilde{\mathcal{S}})} |\tilde{\alpha}^T(\hat{\beta}^0 - \beta^*)|. \end{aligned}$$

In particular,

$$\begin{aligned} \sup_{\alpha \in \mathcal{B}(\tilde{\mathcal{S}})} |\alpha^T(\hat{\beta}^0 - \beta^*)| &\leq (1-\epsilon)^{-1} \max_{\tilde{\alpha} \in \mathcal{N}_\epsilon(\tilde{\mathcal{S}})} |\tilde{\alpha}^T(\hat{\beta}^0 - \beta^*)| \\ &\leq (1-\epsilon)^{-1} \max_{\tilde{\alpha} \in \mathcal{N}_{\epsilon,d}} |\tilde{\alpha}^T(\hat{\beta}^0 - \beta^*)|. \end{aligned}$$

By taking the maximum over all sets $\tilde{\mathcal{S}}$ and noting that the upper bound does not depend on $\tilde{\mathcal{S}}$, the result follows. □

LEMMA S4. *Define*

$$f(x) = \begin{cases} \frac{\tanh(x)}{x} & x \neq 0 \\ 1 & x = 0. \end{cases}$$

Then for any $x \in \mathbb{R}$,

$$1 - \frac{x^2}{3 + x^2} \leq f(x) \leq 1$$

Proof. The result holds for the case $x = 0$ by inspection. When $x \neq 0$, it is sufficient to consider $x > 0$ as f is even. As the functions $-\tanh(x)$ and $(3 + x^2)\tanh(x)$ are convex over the positive real line, they may be bounded below by the linear terms in their respective Taylor expansions. It follows that

$$-\tanh(x) \geq -x, \qquad\qquad (3 + x^2)\tanh(x) \geq 3x$$

and so

$$\frac{3x}{3 + x^2} \leq \tanh(x) \leq x$$

which establishes the result. □

LEMMA S5. *Let*

$$h_1(\hat{\eta}, \eta^*) = \begin{cases} \|\hat{\eta} - \eta^*\|_2, & \eta^* = 0, \\ \frac{\|\hat{\eta} - \eta^*\|_2}{\|\eta^*\|_2}, & \eta^* \neq 0. \end{cases}$$

Then there exists a universal constant $C > 0$ not depending on X such that

$$|\hat{\varsigma} - \varsigma| \leq C \max\{h_1(\hat{\eta}, \eta^*), h_1^2(\hat{\eta}, \eta^*)\}.$$

Proof. Define $f(x)$ as in Lemma S4. When $\eta^* = 0$,

$$|\hat{\varsigma} - \varsigma| = \left| \frac{\sum_{i=1}^n \hat{\eta}_i^2 \{f(\hat{\eta}_i/2) - 1\}}{2\|\hat{\eta}\|_2^2} \right| \leq \frac{\|\hat{\eta}\|_4^4}{24\|\hat{\eta}\|_2^2} \leq \frac{\|\hat{\eta}\|_2^2}{24}$$

where we have used Lemma S4 to bound $|f(\hat{\eta}_i/2) - 1|$ by $\hat{\eta}_i^2/12$. When $\eta^* \neq 0$, we have $|\hat{\varsigma} - \varsigma| \leq \Pi_1 + \Pi_2 + \Pi_3$ where

$$\begin{aligned} 2\Pi_1 &= \left| \sum_{i=1}^n \hat{\eta}_i^2 f(\hat{\eta}_i/2) \right| \left| \frac{1}{\|\hat{\eta}\|_2^2} - \frac{1}{\|\eta^*\|_2^2} \right| \\ 2\Pi_2 &= \frac{\sum_{i=1}^n |\hat{\eta}_i| \, |\hat{\eta}_i f(\hat{\eta}_i/2) - \eta_i^* f(\eta_i^*/2)|}{\|\eta^*\|_2^2} \\ 2\Pi_3 &= \frac{\sum_{i=1}^n \eta_i^* f(\eta_i^*/2) |\hat{\eta}_i - \eta_i^*|}{\|\eta^*\|_2^2}. \end{aligned}$$

Using the Cauchy-Schwartz inequality and the fact that $f(\hat{\eta}_i/2) \leq 1$,

$$\begin{aligned} 2\Pi_1 &\leq \left| \frac{\|\eta^*\|_2^2 - \|\hat{\eta}\|_2^2}{\|\eta^*\|_2^2} \right| \\ &= \left| \frac{\|\eta^*\|_2 - \|\hat{\eta}\|_2}{\|\eta^*\|_2} \right| \left(\frac{\|\eta^*\|_2 + \|\hat{\eta}\|_2}{\|\eta^*\|_2} \right) \\ &\leq \left(\frac{\|\eta^* - \hat{\eta}\|_2}{\|\eta^*\|_2} \right) \left(2 + \frac{\|\eta^* - \hat{\eta}\|_2}{\|\eta^*\|_2} \right) \end{aligned}$$

and

$$2\Pi_3 \leq \frac{\sum_{i=1}^n |\eta_i^*| |\hat{\eta}_i - \eta_i^*|}{\|\eta^*\|_2^2} \leq \frac{\|\hat{\eta} - \eta^*\|_2}{\|\eta^*\|_2}.$$

The function $xf(x/2) = 2\tanh(x/2)$ is Lipschitz continuous with constant one and so,

$$2\Pi_2 \leq \frac{\sum_{i=1}^n |\hat{\eta}_i| |\hat{\eta}_i - \eta_i^*|}{\|\eta^*\|_2^2} \leq \left(1 + \frac{\|\eta^* - \hat{\eta}\|_2}{\|\eta^*\|_2}\right) \left(\frac{\|\hat{\eta} - \eta^*\|_2}{\|\eta^*\|_2}\right).$$

The result follows on combining the bounds. □

LEMMA S6. *Let*

$$h_2(\hat\eta,\eta^*) = \begin{cases} n^{-1/2}\|\hat\eta-\eta^*\|_2, & \eta^*=0,\\ n^{-1/2}\|\hat\eta-\eta^*\|_2\left(1+\frac{\|\hat\eta-\eta^*\|_2}{\|\eta^*\|_2}\right), & \eta^*\neq 0.\end{cases}$$

Then there exists a constant $C>0$ such that for all $X\in\mathbb{X}_B^{(1)}$,

$$\sup_{\alpha\in\mathcal{B}_d} |\alpha^T(\hat\delta-\delta)| \le CB^{1/2}h_2(\hat\eta,\eta^*).$$

Proof. For $\alpha\in\mathcal{B}_d$,

$$|\alpha^T(\hat\delta-\delta)| \le \|\alpha^T(X^TX)^{-1}X^T\|_2\|(P_X-P_{\hat\eta})\tanh(\hat\eta/2)-(P_X-P_{\eta^*})\tanh(\eta^*/2)\|_2.$$

By definition of $\mathbb{X}_B^{(1)}$,

$$\sup_{\alpha\in\mathcal{B}_d}\|\alpha^T(X^TX)^{-1}X^T\|_2^2 \le n^{-1}\lambda_{\max}\{(X^TX/n)^{-1}\}\le Bn^{-1}.$$

As the largest eigenvalue of a projection matrix is one and $\tanh(\cdot)$ is Lipschitz continuous with constant one,

$$\|P_X\{\tanh(\hat\eta/2)-\tanh(\eta^*/2)\}\|_2 \le \|\tanh(\hat\eta/2)-\tanh(\eta^*/2)\|_2\le\|\hat\eta-\eta^*\|_2/2.$$

Further, there exists some constant $C>0$ such that

$$\begin{aligned}\|P_{\hat\eta}\tanh(\hat\eta/2)-P_{\eta^*}\tanh(\eta^*/2)\|_2 &= \|\hat\varsigma\hat\eta-\varsigma\eta^*\|_2\\ &\le|\hat\varsigma|\|\hat\eta-\eta^*\|_2+|\hat\varsigma-\varsigma|\|\eta^*\|_2\\ &= Cn^{1/2}h_2(\hat\eta,\eta^*)\end{aligned}$$

by Lemma S5 and the fact that $|\hat\varsigma|\le 1/2$. The result follows on combining these inequalities. □

LEMMA S7. *For any $X\in\mathbb{R}^{n\times p}$ of rank $p<n$, the prediction error of the OLS estimator satisfies*

$$n^{-1/2}\|X(\hat\beta^0-\beta^0)\|_2 = n^{-1/2}\|P_X\Gamma^{1/2}\|_F+o_P(1)$$

as $p,n\to\infty$ with $p<n$.

Proof. The unscaled prediction error can be written as $\|X(\hat\beta^0-\beta^0)\|_2=\|P_X\varepsilon\|_2$ where $\varepsilon=Y-\mathbb{E}(Y)$ consists of independent and centred sub-Gaussian random variables with $\max_{i=1}^n\|\varepsilon_i\|_{\psi_2}\le 2$. The expected prediction error is

$$\mathbb{E}\{\|X(\hat\beta^0-\beta^0)\|_2^2\}=\mathrm{tr}(\Gamma P_X)=\|P_X\Gamma^{1/2}\|_F^2$$

and so by the Hanson–Wright inequality (Theorem 6.2.1 in Vershynin (2018)), for any $t>0$,

$$\mathrm{pr}\left\{\left|\|X(\hat\beta^0-\beta^0)\|_2^2-\|P_X\Gamma^{1/2}\|_F^2\right|\ge tn\right\}\le 2\exp\left(-C\min\left\{\frac{t^2n^2}{\|P_X\|_F^2},\frac{tn}{\|P_X\|_2}\right\}\right)$$

for some universal constant $C>0$. As P_X is a projection matrix of rank p, $\|P_X\|_F^2=p$ and $\|P_X\|_2=1$. Thus, as $p\le n$,

$$\mathrm{pr}\left\{\left|\|X(\hat\beta^0-\beta^0)\|_2^2-\|P_X\Gamma^{1/2}\|_F^2\right|\ge tn\right\}=o(1)$$

and so the result follows. □

LEMMA S8. *Suppose Conditions 3 and 4 hold with $\gamma>0$. Then, $\varsigma^{-1}=O_P(1)$.*

Proof. Recall that

$$\varsigma = \begin{cases} \frac{(X\beta^*)^T \tanh(X\beta^*/2)}{\|X\beta^*\|_2^2}, & X\beta^* \neq 0 \\ 1/2, & X\beta^* = 0. \end{cases}$$

By the weak law of large numbers and the assumption that $\mathbb{E}\{(x_1^T\beta^*)^2\} \to \gamma^2$,

$$\frac{(X\beta^*)^T \tanh(X\beta^*/2)}{n} \xrightarrow{p} \mathbb{E}(V), \qquad \frac{\|X\beta^*\|_2^2}{n} \xrightarrow{p} \gamma^2$$

where $V = (x_1^T\beta^*)\tanh(x_1^T\beta^*/2)$. As V is a non-negative random variable,

$$\mathbb{E}(V) \geq \tanh(1/2)\text{pr}(|x_1^T\beta^*| \geq 1) = 2\tanh(1/2)[1 - \Phi\{(\beta^{*T}\Sigma\beta^*)^{-1/2}\}].$$

When $\gamma > 0$, this is bounded away from zero by a constant for large enough sample sizes. Further, $\mathbb{E}(V) \leq 1/2$. Using Slutsky's Theorem, $\varsigma \xrightarrow{p} \gamma^{-2}\mathbb{E}(V)$ and so $\varsigma^{-1} = O_P(1)$. □ 210

LEMMA S9. *Suppose Conditions 3 and 4 hold with $\kappa \in (0, 1)$. Then,*

$$\lambda_{\max}\{(X^TX/n)^{-1}\} = O_P(1).$$

Proof. The matrix X has the same distribution as the matrix $Z\Sigma^{1/2}$ where each row of Z is an independent sample from the distribution $N_p(0, I)$. Then,

$$\begin{aligned} \lambda_{\min}(X^TX/n) &\overset{d}{=} n^{-1}\lambda_{\min}(\Sigma^{1/2}Z^TZ\Sigma^{1/2}) \\ &> \lambda_{\min}(Z^TZ/n)\lambda_{\min}(\Sigma) \\ &> \lambda_{\min}(Z^TZ/n)\lambda_{\min}. \end{aligned}$$

The lower bound converges to a positive constant in probability as $p, n \to \infty$ with $p/n \to \kappa \in (0, 1)$ by Theorem 2.16 (Bai, 1999). As 215

$$\lambda_{\max}\{(X^TX/n)^{-1}\} = \lambda_{\min}^{-1}(X^TX/n)$$

the result follows. □

LEMMA S10. *Suppose Conditions 3 and 4 hold with $\gamma > 0$. Then,*

$$\max_{i=1}^{n} |x_i^T\beta^*| = O_P(\sqrt{\log n}), \quad \max_{i=1}^{n}\max_{j=1}^{p} |x_{ij}| = O_P(\sqrt{\log n}),$$

and when $\beta^* \neq 0$,

$$\|X\beta^*\|_2^{-1} = O_P(n^{-1/2}).$$

Proof. By assumption, $x_i^T\beta^* \sim N(0, \beta^{*T}\Sigma\beta^*)$ and so $x_i^T\beta^*$ is sub-Gaussian with norm bounded by $(\beta^{*T}\Sigma\beta^*)^{1/2}$ up to a constant. Then, there exist constants $C, c_1, c_2 > 0$ such that, 220

$$\text{pr}(\max_{i=1}^{n} |x_i^T\beta^*| > c_1\sqrt{\log n}) \leq Cn\exp\left(-\frac{c_2\log n}{\beta^{*T}\Sigma\beta^*}\right) \to 0$$

provided c_1 is large enough as $\beta^{*T}\Sigma\beta^* \to \gamma^2$. It follows that $\max_{i=1}^{n} |x_i^T\beta^*| = O_P(\sqrt{\log n})$.

We know that $x_{ij} \sim N(0, \Sigma_{jj})$ for $j \in \{1, \ldots, p\}$. Thus, x_{ij} is sub-Gaussian with norm bounded by $\Sigma_{jj}^{1/2}$ up to a constant. Then, there exist constants $C, c_1, c_2 > 0$ such that

$$\text{pr}(\max_{i=1}^{n}\max_{j=1}^{p} |x_{ij}| > c_1\sqrt{\log n}) \leq Cnp\exp\left(-\frac{c_2\log n}{\Sigma_{jj}}\right) \to 0$$

provided c_1 is large enough as $\max_{j=2}^p \Sigma_{jj}$ is bounded. It follows that $\max_{i=1}^n \max_{j=1}^p |x_{ij}| = O_P(\sqrt{\log n})$.

Finally by the weak law of large numbers, $\|X\beta^*\|_2^2/n \xrightarrow{p} \gamma^2$ and so $\|X\beta^*\|_2^{-1} = O_P(n^{-1/2})$. □

LEMMA S11. *Suppose $\lambda_{\max}\{(X^TX/n)^{-1}\} = O(1)$. Then $\phi_0^2(X) = O(1)$.*

Proof. For a vector $\beta \in \mathbb{R}^p$, let $\beta_{\mathcal{S}}$ be the vector with entries equal to those of β for all indices in $\mathcal{S}$ and 0 otherwise. The Cauchy-Schwartz inequality implies $\|\beta_{\mathcal{S}}\|_1^2 \leq s\|\beta_{\mathcal{S}}\|_2^2$. Then,

$$\frac{\|X\beta\|_2^2 s}{n\|\beta_{\mathcal{S}}\|_1^2} \geq \frac{\|X\beta\|_2^2}{n\|\beta\|_2^2} \geq \lambda_{\min}(X^TX/n)$$

whenever $\beta \neq 0$. Thus,

$$\phi_0^{-2}(X) \leq \lambda_{\min}^{-1}(X^TX/n)$$

where the upper bound is of $O(1)$. □

3. THE LASSO ESTIMATOR OF $X\beta^*$

In this section, we prove Proposition 7, establishing that the LASSO estimator of $X\beta^*$ satisfies the requirement in Condition 2. The arguments closely follow those in Theorem 6.4 in Bühlmann & van de Geer (2011, p. 130-133). Define $Z_i = (Y_i + 1)/2$ and $\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^n [\log\{1 + \exp(x_i^T\beta)\} - Z_i x_i^T \beta]$ and $R(\beta) = \mathbb{E}\{\hat{R}(\beta)\}$. For a set $\mathcal{S}$, let $\beta_{\mathcal{S}}$ be the vector with entries equal to those of β for all indices in $\mathcal{S}$ and 0 otherwise. Let $s = \|\beta^*\|_0$ and x_j denote the j-th column of X. Recall the index i is reserved for indexing rows of X. Define E_λ to be the event that

$$\|X^T\varepsilon\|_\infty \leq 2n\lambda$$

where $\varepsilon = Y - \mathbb{E}(Y)$.

LEMMA S12. *There exist universal constants $C, c > 0$ such that,*

$$\inf_{X \in \mathbb{X}_B^{(2)}} pr(E_\lambda) > 1 - Cp\exp\{-cn\lambda^2/(B^2\log n)\}.$$

Proof. Let $S_j = \sum_{i=1}^n \varepsilon_i x_{ij}$. As $\varepsilon_i = Y_i - \mathbb{E}(Y_i) \in [-2, 2]$, these random variables are independent and sub-Gaussian with $\|\varepsilon_i\|_{\psi_2} \leq 2$. It follows that S_j is sub-Gaussian with norm bounded by $\|x_j\|_2 \leq B\sqrt{n\log n}$ up to a universal constant when $X \in \mathbb{X}_B^{(2)}$. Using a union bound, there exist universal constants $C, c > 0$ such that

$$\begin{aligned}\sup_{X \in \mathbb{X}_B^{(2)}} \text{pr}(E_\lambda^c) &= \sup_{X \in \mathbb{X}_B^{(2)}} \text{pr}\left(\cup_{j=1}^p \{|S_j| > 2n\lambda\}\right) \\ &\leq Cp\exp\{-cn\lambda^2/(B^2\log n)\}\end{aligned}$$

and so the result follows. □

LEMMA S13. *Suppose $\beta^* = 0$. For all $X \in \mathbb{X}_B^{(2)}$ and $t > 0$,*

$$pr(\|X\hat{\beta}^{(\lambda)} - X\beta^*\|_2 \geq t) \leq pr(E_\lambda^c).$$

Proof. By definition of $\hat{\beta}^{(\lambda)}$,

$$\hat{R}(\hat{\beta}^{(\lambda)}) + \lambda\|\hat{\beta}^{(\lambda)}\|_1 \leq \hat{R}(\beta^*) + \lambda\|\beta^*\|_1.$$

On the event E_λ we have $(2n)^{-1}\|X^T\varepsilon\|_\infty \leq \lambda$ and so

$$\begin{aligned} R(\hat\beta^{(\lambda)}) - R(\beta^*) &\leq (2n)^{-1}\varepsilon^T X(\hat\beta^{(\lambda)} - \beta^*) + \lambda\|\beta^*\|_1 - \lambda\|\hat\beta^{(\lambda)}\|_1 \\ &\leq \lambda\|\beta^* - \hat\beta^{(\lambda)}\|_1 + \lambda\|\beta^*\|_1 - \lambda\|\hat\beta^{(\lambda)}\|_1. \end{aligned}$$

By the triangle inequality,

$$R(\hat\beta^{(\lambda)}) - R(\beta^*) \leq 2\lambda\|\beta^*\|_1 = 0$$

and so $X\hat\beta^{(\lambda)} = X\beta^*$. It follows that

$$\mathrm{pr}(\|X\hat\beta^{(\lambda)} - X\beta^*\|_2 \geq t) \leq \mathrm{pr}(E^c_\lambda)$$

for all $t > 0$. □

LEMMA S14. *Suppose* $\max_{i=1}^n |x_i^T\beta| \leq c_1\sqrt{\log n}$ *and* $\max_{i=1}^n |x_i^T\beta^*| \leq c_2\sqrt{\log n}$. *Then,*

$$R(\beta) - R(\beta^*) \geq \frac{\|X\beta - X\beta^*\|_2^2}{8ne^{\max\{c_1,c_2\}\sqrt{\log n}}}.$$

Proof. Using a Taylor expansion, there exists $\beta_v = v\beta + (1-v)\beta^*$ with $v \in [0,1]$ such that

$$R(\beta) - R(\beta^*) = \frac{1}{2n}\sum_{i=1}^n (\beta - \beta^*)^T x_i \left\{\frac{e^{x_i^T\beta_v}}{(1+e^{x_i^T\beta_v})^2}\right\} x_i^T(\beta - \beta^*).$$

By assumption,

$$|x_i^T\beta_v| \leq v|x_i^T\beta| + (1-v)|x_i^T\beta^*| \leq \max\{c_1,c_2\}\sqrt{\log n}$$

and so

$$\frac{e^{x_i^T\beta_v}}{(1+e^{x_i^T\beta_v})^2} \geq \frac{e^{\max\{c_1,c_2\}\sqrt{\log n}}}{(1+e^{\max\{c_1,c_2\}\sqrt{\log n}})^2} \geq e^{-\max\{c_1,c_2\}\sqrt{\log n}}/4$$

where the last line follows because $(3e^x+1)(e^x-1) \geq 0$ for all $x \geq 0$. It follows that

$$R(\beta) - R(\beta^*) \geq \frac{\|X\beta - X\beta^*\|_2^2}{8ne^{\max\{c_1,c_2\}\sqrt{\log n}}}$$

and so the result is obtained. □

LEMMA S15. *Let* $\lambda = A\sqrt{(\log p \log n)/n}$ *with* $A > 0$, $\beta^* \neq 0$, $B > 0$ *and suppose*

$$se^{2B\sqrt{\log n}}\sqrt{\log p \log n/n} = o(1).$$

There exists $N > 0$, *such that when* $X \in \mathbb{X}_B^{(2)}$ *and* $n \geq N$,

$$E_{\lambda/6} \qquad \Longrightarrow \qquad n^{-1}\|X\beta^* - X\hat\beta^{(\lambda)}\|_2^2 \leq \frac{768A^2e^{4B\sqrt{\log n}}s\log p \log n}{n\phi_0^2(X)}.$$

Proof. Define

$$M^*/6 = 32\phi_0^{-2}(X)se^{2B\sqrt{\log n}}\lambda.$$

Let $\tilde\beta = t\hat\beta^{(\lambda)} + (1-t)\beta^*$ where

$$t = \frac{M^*}{M^* + \|\hat\beta^{(\lambda)} - \beta^*\|_1}.$$

By convexity of the logistic log-likelihood and the ℓ_1-norm,

$$\begin{aligned}\hat{R}(\tilde{\beta}) + \lambda\|\tilde{\beta}\|_1 &\leq t\{\hat{R}(\hat{\beta}^{(\lambda)}) + \lambda\|\hat{\beta}^{(\lambda)}\|_1\} + (1-t)\{\hat{R}(\beta^*) + \lambda\|\beta^*\|_1\} \\ &\leq \hat{R}(\beta^*) + \lambda\|\beta^*\|_1.\end{aligned}$$

Then, with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ where $\varepsilon_i = Y_i - \mathbb{E}(Y_i)$,

$$R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}\|_1 \leq (2n)^{-1}\varepsilon^T X(\tilde{\beta} - \beta^*) + \lambda\|\beta^*\|_1.$$

By definition of $\tilde{\beta}$, $\|\tilde{\beta} - \beta^*\|_1 \leq M^*$. So, on the event $E_{\lambda/6}$,

$$R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}\|_1 \leq \lambda M^*/6 + \lambda\|\beta^*\|_1.$$

Recall $\mathcal{S} = \{j : \beta_j^* \neq 0\}$. Then, as $\|\tilde{\beta}\|_1 = \|\tilde{\beta}_{\mathcal{S}}\|_1 + \|\tilde{\beta}_{\mathcal{S}^c}\|_1$,

$$\begin{aligned}R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}_{\mathcal{S}^c}\|_1 &\leq \lambda M^*/6 + \lambda\|\beta^*\|_1 - \lambda\|\tilde{\beta}_{\mathcal{S}}\|_1 \\ &\leq \lambda M^*/6 + \lambda\|\beta^* - \tilde{\beta}_{\mathcal{S}}\|_1. \end{aligned} \tag{S12}$$

First suppose $\lambda\|\beta^* - \tilde{\beta}_{\mathcal{S}}\|_1 \leq \lambda M^*/6$. Then,

$$\begin{aligned}R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta} - \beta^*\|_1 &= R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\tilde{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_1 \\ &\leq \lambda M^*/6 + 2\lambda\|\tilde{\beta}_{\mathcal{S}} - \beta^*\|_1 \\ &\leq \lambda M^*/2. \end{aligned} \tag{S13}$$

Let $\delta = \beta^* - \tilde{\beta}$. If $\lambda\|\beta^* - \tilde{\beta}_{\mathcal{S}}\|_1 > \lambda M^*/6$, then by equation (S12),

$$\begin{aligned}\lambda\|\delta_{\mathcal{S}^c}\|_1 &= \lambda\|\tilde{\beta}_{\mathcal{S}^c}\|_1 \\ &\leq \lambda M^*/6 + \lambda\|\beta^* - \tilde{\beta}_{\mathcal{S}}\|_1 \\ &< 2\lambda\|\delta_{\mathcal{S}}\|_1\end{aligned}$$

and so by definition of $\phi_0^2(X)$,

$$\|\beta^* - \tilde{\beta}_{\mathcal{S}}\|_1^2 \leq \frac{\|X(\beta^* - \tilde{\beta})\|_2^2 s}{n\phi_0^2(X)}.$$

Combining this with equation (S12), we have

$$\begin{aligned}R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta} - \beta^*\|_1 &= R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}_{\mathcal{S}^c}\|_1 + \lambda\|\tilde{\beta}_{\mathcal{S}} - \beta^*_{\mathcal{S}}\|_1 \\ &\leq \lambda M^*/6 + 2\lambda\|\tilde{\beta}_{\mathcal{S}} - \beta^*\|_1 \\ &\leq \lambda M^*/6 + \frac{2\lambda\|X(\beta^* - \tilde{\beta})\|_2 s^{1/2}}{\sqrt{n}\phi_0(X)}. \end{aligned} \tag{S14}$$

As $\|\tilde{\beta} - \beta^*\|_1 \leq M^*$ and $\max_{i,j}|x_{ij}| \leq B\sqrt{\log n}$ by assumption,

$$\|X(\beta^* - \tilde{\beta})\|_\infty \leq \max_{i=1}^{n}\sum_{j=1}^{n}|x_{ij}||\beta_j^* - \tilde{\beta}_j| \leq BM^*\sqrt{\log n}$$

and so, $\|X\tilde{\beta}\|_\infty \leq B(M^*+1)\sqrt{\log n} \leq 2B\sqrt{\log n}$ for n large enough. The last inequality follows because $M^* = o(1)$ under the assumptions of this lemma. By Lemma S14,

$$\begin{aligned}\frac{2\lambda\|X(\beta^*-\tilde{\beta})\|_2 s^{1/2}}{\sqrt{n}\phi_0(X)} &\leq \frac{4\sqrt{2}e^{B\sqrt{\log n}}\lambda s^{1/2}}{\phi_0(X)} \times \sqrt{R(\tilde{\beta})-R(\beta^*)}\\ &\leq \frac{32e^{2B\sqrt{\log n}}\lambda^2 s}{\phi_0^2(X)} + \frac{R(\tilde{\beta})-R(\beta^*)}{4}\end{aligned}$$

where the last inequality follows because $ab \leq a^2 + b^2/4$ for all $a, b \in \mathbb{R}$. Combining this with equation (S14),

$$R(\tilde{\beta}) - R(\beta^*) + \lambda\|\tilde{\beta}-\beta^*\|_1 \leq \lambda M^*/6 + \frac{32e^{2B\sqrt{\log n}}\lambda^2 s}{\phi_0^2(X)} + \frac{R(\tilde{\beta})-R(\beta^*)}{4}.$$

Rearranging, we obtain,

$$\frac{3}{4}\{R(\tilde{\beta}) - R(\beta^*)\} + \lambda\|\tilde{\beta}-\beta^*\|_1 \leq \lambda M^*/6 + \frac{32e^{2B\sqrt{\log n}}\lambda^2 s}{\phi_0^2(X)} \leq 2\lambda M^*/6. \tag{S15}$$

In particular, using this and equation (S13), we find $\|\tilde{\beta}-\beta^*\|_1 \leq M^*/2$ and so $\|\hat{\beta}^{(\lambda)}-\beta^*\|_1 \leq M^*$. The desired result can then be obtained by repeating the arguments with $\tilde{\beta}$ replaced by $\hat{\beta}^{(\lambda)}$. In particular, replacing $\tilde{\beta}$ with $\hat{\beta}^{(\lambda)}$ in equations (S13) and (S15) yields

$$R(\hat{\beta}^{(\lambda)}) - R(\beta^*) \leq \lambda M^*/2.$$

Using Lemma S14, this implies that

$$\frac{\|X\beta^* - X\hat{\beta}^{(\lambda)}\|_2^2}{n} \leq 4e^{2B\sqrt{\log n}}\lambda M^*$$

and the result is obtained by replacing M^* and λ by their defined values.

Proof of Proposition 7. By Lemmas S12 and S13, when $\beta^* = 0$,

$$\begin{aligned}\sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}\{h(X\hat{\beta}^{(\lambda)}, X\beta^*) > t\} &= \sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}(\|X(\hat{\beta}^{(\lambda)}-\beta^*)\|_2 > t)\\ &\leq \sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}(E_\lambda^c)\\ &= o(1)\end{aligned}$$

as long as A is large enough. Now suppose $\beta^* \neq 0$. Then,

$$\sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}\{h(X\hat{\beta}^{(\lambda)}, X\beta^*) > t\} \leq \sup_{X\in\mathbb{X}_B^{(2)}} \mathrm{pr}\{n^{-1/2}\|X(\hat{\beta}^{(\lambda)}-\beta^*)\|_2 \max\{B,1\} > t\}.$$

Let

$$f_n = e^{4B\sqrt{\log n}} s\log p \log n/n$$

satisfying $f_n = o(1)$ by assumption. Combining Lemmas S12 and S15, there exist positive constants $N, C_1, C_2, c_1 > 0$ such that when $n \geq N$,

$$\begin{aligned}\sup_{X \in \mathbb{X}_B^{(2)}} \operatorname{pr}\left(n^{-1}\|X\hat{\beta}^{(\lambda)} - X\beta^*\|_2^2 > t^2 \max\{B,1\}^{-2}\right) &\leq \sup_{X \in \mathbb{X}_B^{(2)}} \operatorname{pr}\left(n^{-1}\|X\hat{\beta}^{(\lambda)} - X\beta^*\|_2^2 > C_1 f_n\right) \\ &\leq \sup_{X \in \mathbb{X}_B^{(2)}} \operatorname{pr}(E^c_{\lambda/6}) \\ &\leq C_2 \exp\{-(c_1 A^2 - 1)\log p\}.\end{aligned}$$

This probability also converges to zero when A is large enough and so the result follows. □

4. FURTHER NUMERICAL PERFORMANCE

4.1. *Asymptotic performance*

In this section, we consider the asymptotic performance of the corrected least square estimator by observing how various error rates behave as $p, n \to \infty$ with $p/n = \kappa$. Data were generated based on the distribution in Condition 3 of the main paper to ensure that cases with and without data separation were included in the study. This placed focus on the pair (γ, κ) where $\gamma = \mathbb{E}\{(x_i^T\beta^*)^2\}$ and $\kappa = p/n$. The values $\gamma = 3$ and $\kappa \in \{0.1, 0.5\}$ were chosen for our analysis. Based on the work of Candès & Sur (2020), data separation occurs with probability converging to zero when $\kappa = 0.1$ and $\gamma = 3$, whereas this probability converges to one when $\kappa = 0.5$.

The design matrix X was generated as consisting of n observations on p variables with $n \in \{100, 200, \dots, 1000\}$ and $p = \kappa n$. Excluding the intercept term, the covariates corresponding to a given observation were generated from a $p-1$ dimensional multivariate normal distribution with mean zero and covariance matrix $\Sigma = \rho\mathbb{1}\mathbb{1}^T + (1-\rho)\mathbb{I}$ with $\rho = 0.5$. The response variable Y was generated from a logistic regression model with logarithmic odds given by $X\beta^*$. The parameter β^* had exactly $s = 5$ randomly chosen non-zero entries with equal signal strength. The intercept term always had a non-zero effect. To guarantee $\gamma = 3$, each non-zero entry of β^* was set to approximately 0.90.

For each sample size, various corrections to the least squares estimator were obtained by making use of a variety of estimates of η^*. The first was the oracle estimator that made use of the true value of η^*. This is unavailable in practice but serves as a useful reference point. The predictor η^* was also estimated using the ridge (Hoerl & Kennard, 1970), LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001) and MCP (Zhang, 2010) penalised regressions. Additionally, a logistic regression was fitted using only the left singular vectors of X corresponding to the largest $\hat{r}$ eigenvalues of X^TX where $\hat{r} = \operatorname{argmax}_{j \leq 20} \lambda_j/\lambda_{j+1}$ and $\lambda_1 \geq \cdots \geq \lambda_p$ are the eigenvalues of X^TX. This approach is referred to as SVD based on its relation to the singular value decomposition of X. In $R = 500$ repetitions, a new randomly generated response variable was generated and the average composite estimation and prediction errors were recorded. For a chosen null and signal variable, the estimated signal strength of each estimator was also obtained. The results are given in Figure S1.

We also considered the average behaviour of the test defined in (9) of the original paper with $\vartheta = 0.05$, $b_0 = 0$ and α equal to one of the standard basis vectors. The proportion of times the null hypothesis was rejected for the chosen null and signal value were recorded to examine the Type I error and power. The results are given in Figure S2. For the chosen null variable, Figure S3 shows the ordered p-values $2\{1 - \Phi(|T|)\}$ plotted against their theoretical expectation under a uniform distribution.

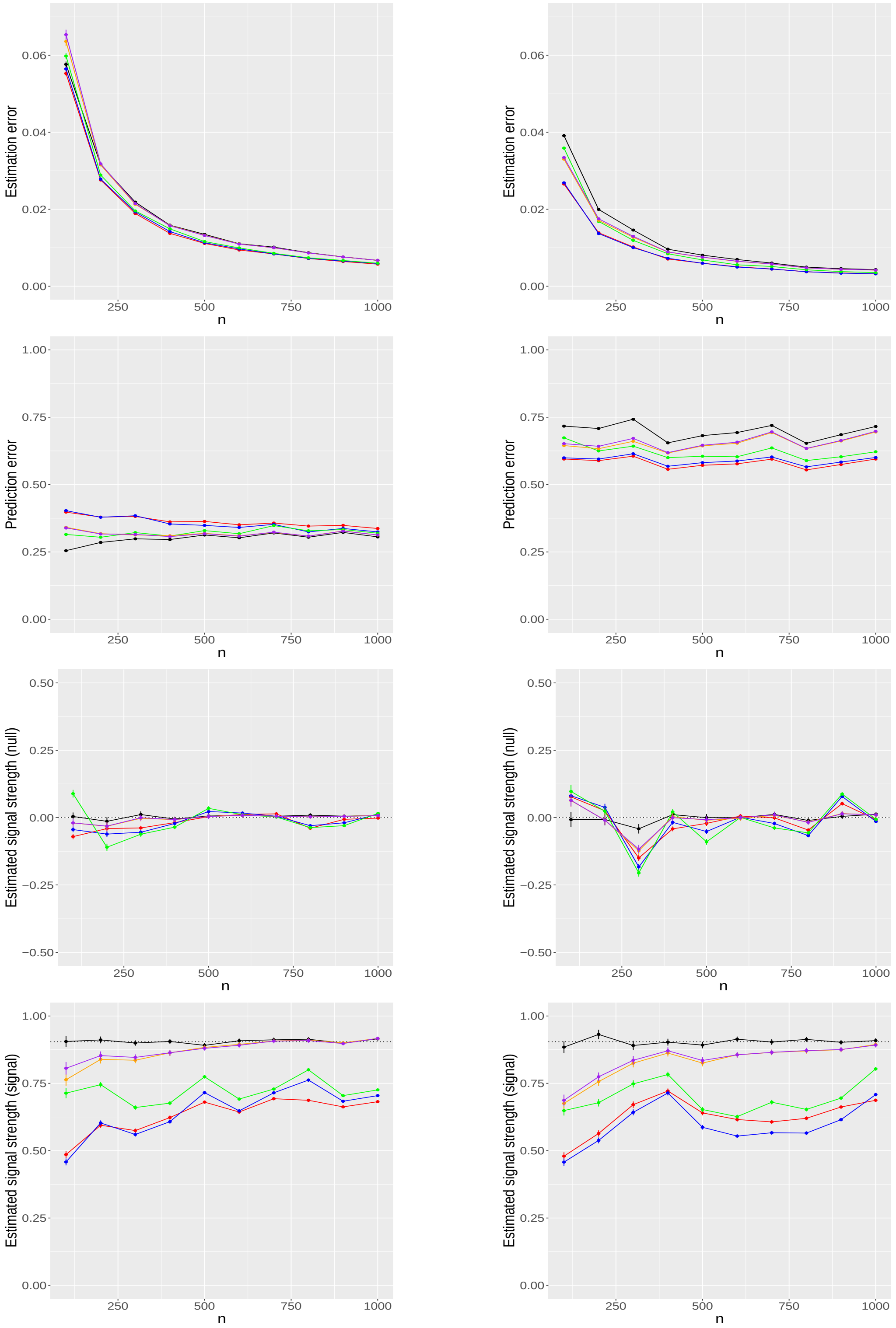


Fig. S1: Average performance of the corrected least squares estimator for various values of n and $s = 5$. The left column corresponds to $\kappa = 0.1$ and the right column to $\kappa = 0.5$. Various corrections to the OLS estimator were used: oracle (black), LASSO (red), ridge regression (blue), SVD (green), SCAD (orange), MCP (purple). The dashed lines in the bottom two rows denote the true signal strength. Error bars represent empirical standard errors.

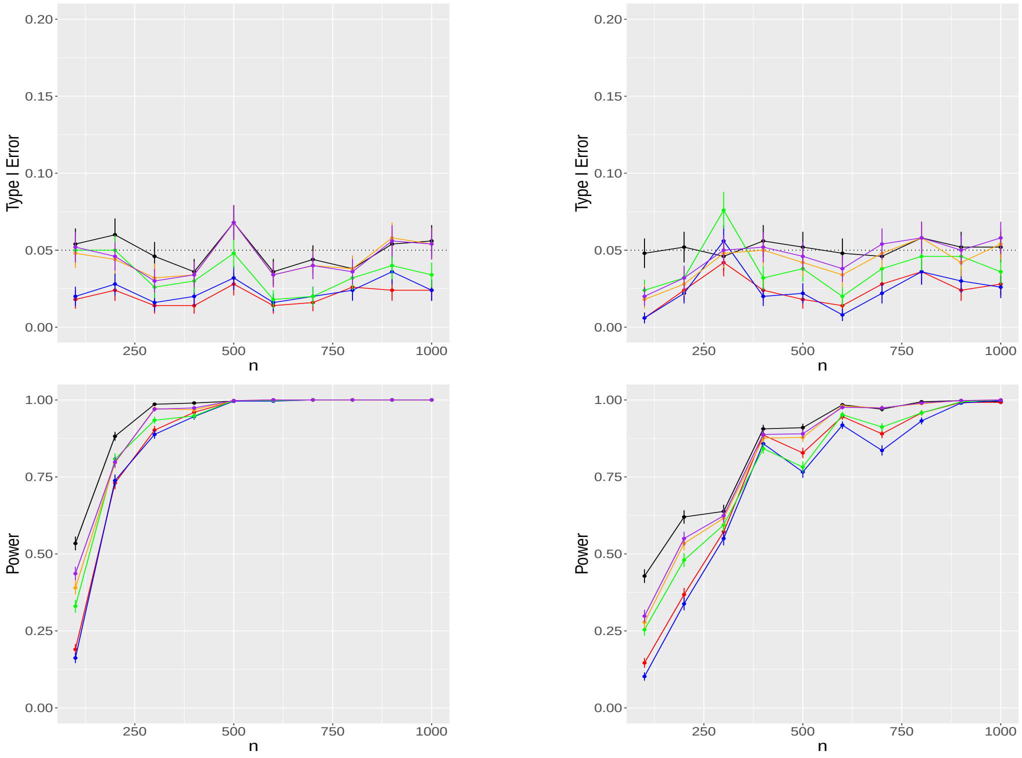
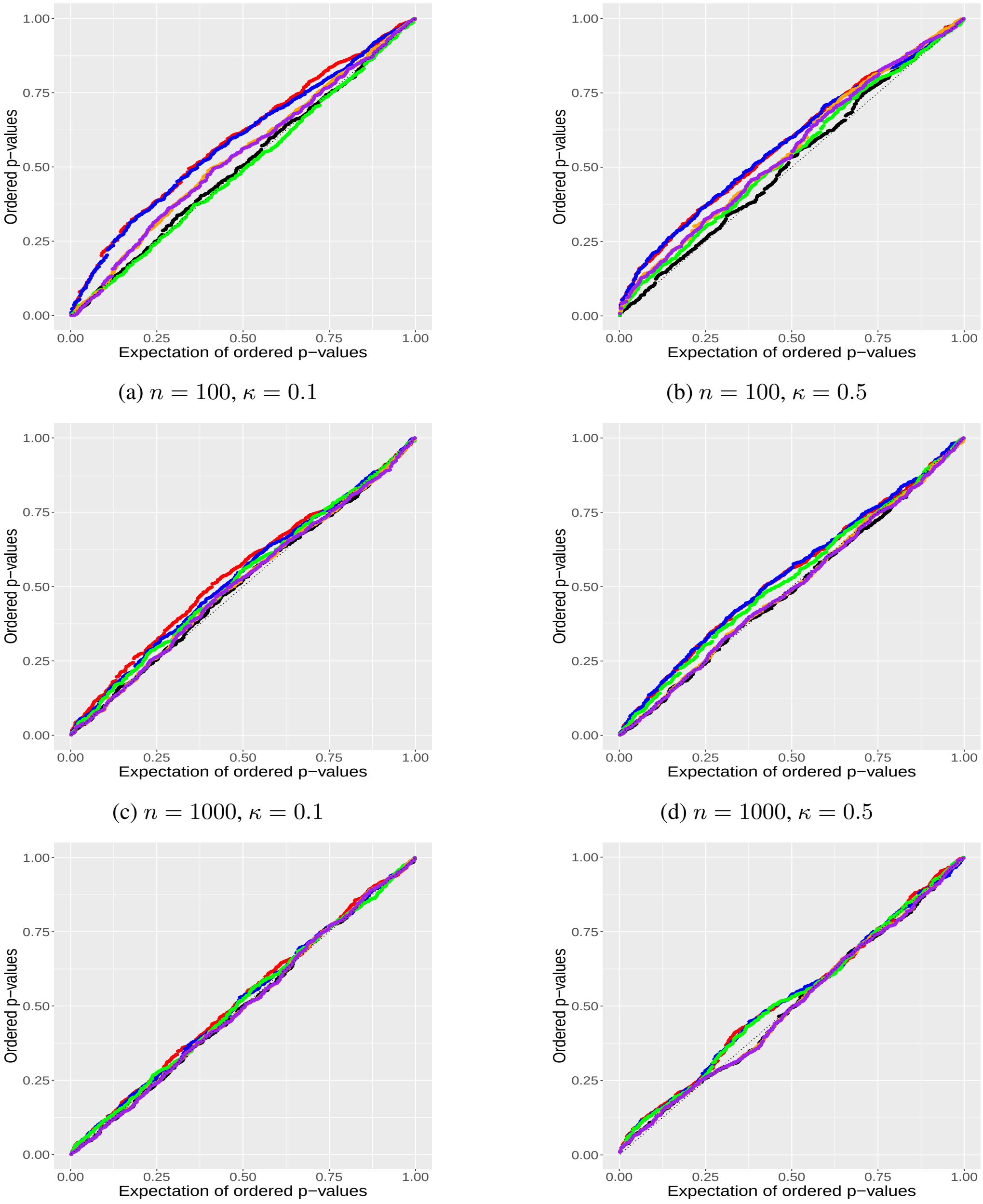


Fig. S2: Average Type I error and power of the test $\psi(Y; 0.05, \alpha)$ for various values of n, $s = 5$ and α equal to a standard basis vector. The left column corresponds to $\kappa = 0.1$ and the right column to $\kappa = 0.5$. Various estimators of η were used: oracle (black), LASSO (red), ridge regression (blue), SVD (green), SCAD (orange), MCP (purple). Error bars represent empirical standard errors.

The results show that the composite estimation error decreased as a function of n. On the other hand, the prediction error remained relatively stable at a non-zero value which coincides with the analysis in section 5.5. The average biases of the corrected least squares estimators were often close to zero for null variables. For signal variables, the bias increased slightly for the ridge, LASSO and SVD estimators, but remained small for the oracle, SCAD and MCP estimators. All estimators controlled the Type-I error close to the intended level of 0.05. The power of the test increased with the sample size and for large enough sample sizes, was very close to one. The plots displayed in Figure S3 show that the distribution of the p-values under the null hypothesis was in close agreement with a uniform distribution.

The simulations were repeated, this time allowing $s = \lfloor n^{1/3} \rfloor$ to grow with the sample size. By definition of γ and the form of β^*, this altered the signal strength as a function of n, and so the average estimated signal strength was replaced by the average estimation error $\tilde{\beta}_i^* - \beta_i^*$. The results are shown in Figures S4, S5 and S6 . They closely resemble the case with s fixed.



(a) $n = 100$, $\kappa = 0.1$

(b) $n = 100$, $\kappa = 0.5$

(c) $n = 1000$, $\kappa = 0.1$

(d) $n = 1000$, $\kappa = 0.5$

(e) $n = 3000$, $\kappa = 0.1$

(f) $n = 3000$, $\kappa = 0.5$

Fig. S3: Plots of the ordered p-values $2\{1 - \Phi(|T|)\}$ against their theoretical expectation under the null hypothesis $H_0 : \beta_j^* = 0$ when $s = 5$. Various estimators of η^* were used to compute T: oracle (black), LASSO (red), ridge regression (blue), SVD (green), SCAD (orange), MCP (purple).

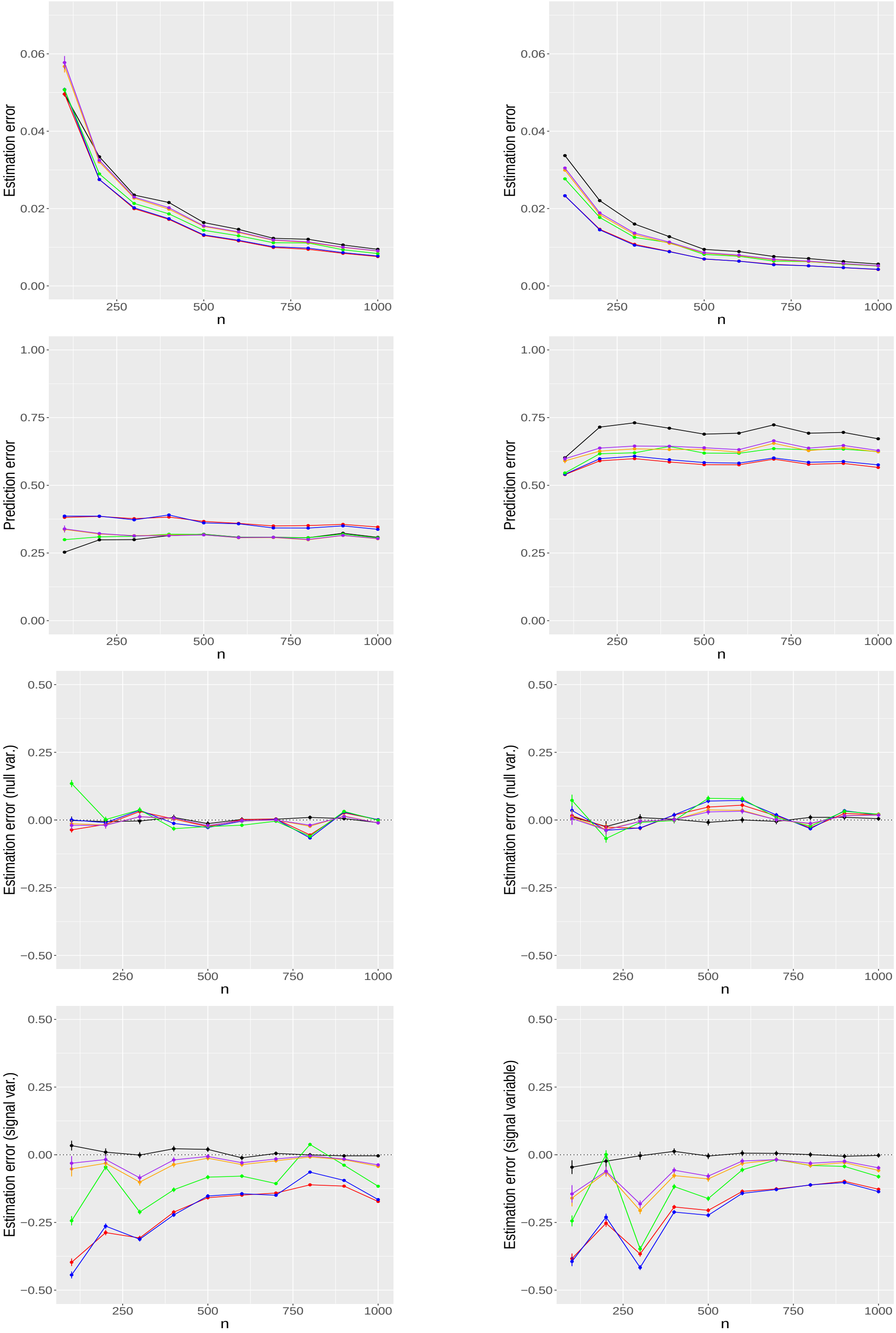


Fig. S4: As in Figure S1 with $s = \lfloor n^{1/3} \rfloor$.

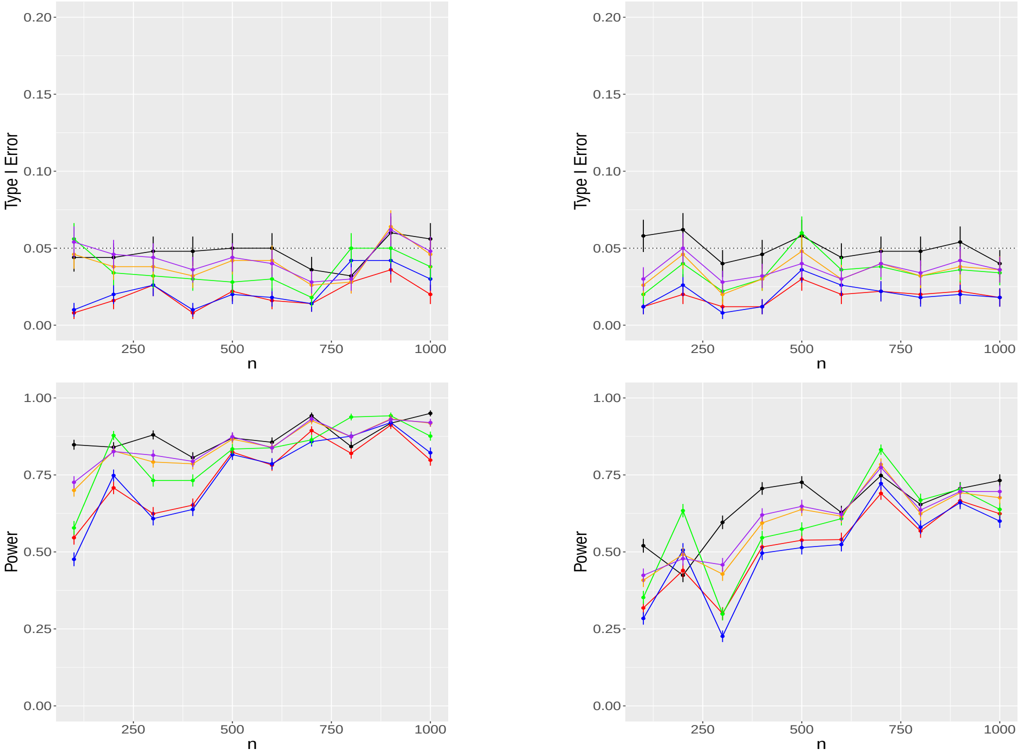
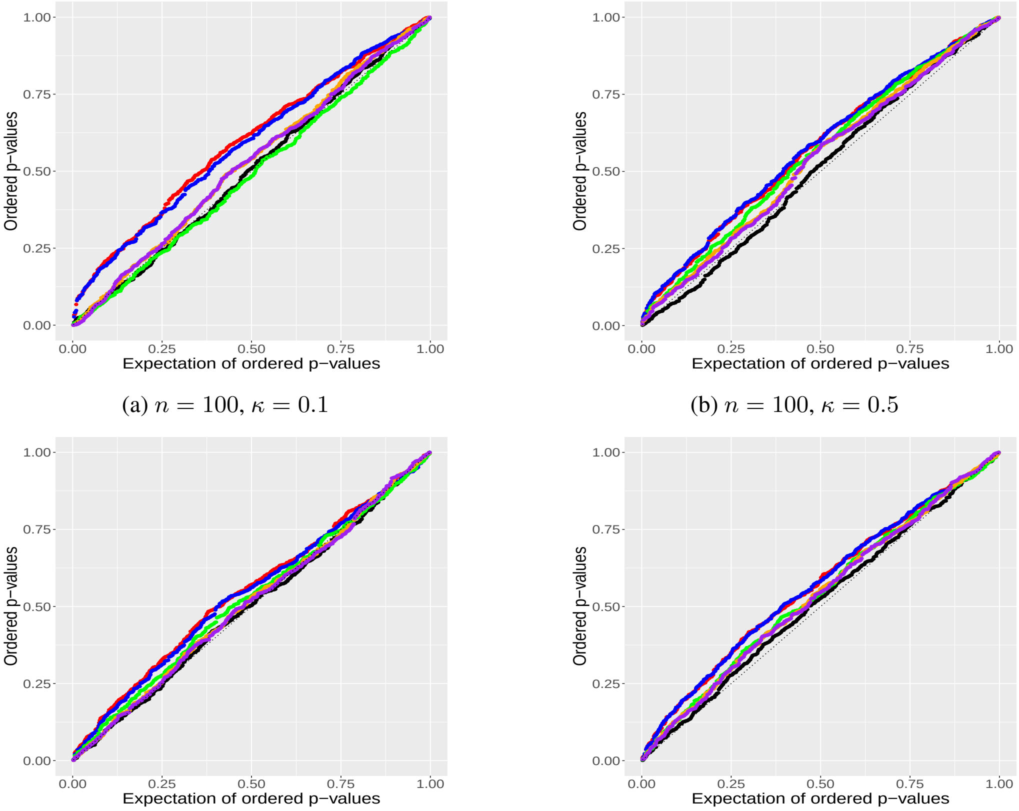


Fig. S5: As in Figure S2 with $s = \lfloor n^{1/3} \rfloor$.



(a) $n = 100$, $\kappa = 0.1$

(b) $n = 100$, $\kappa = 0.5$

(c) $n = 1000$, $\kappa = 0.1$

(d) $n = 1000$, $\kappa = 0.5$

Fig. S6: As in Figure S3 with $s = \lfloor n^{1/3} \rfloor$.

κ	γ	ρ	Statistic	ML	Firth's ML	OLS-Oracle	OLS-LASSO	OLS-Ridge	OLS-SVD	OLS-SCAD	OLS-MCP
0.1	3	0.5	P(exists)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	0.27 (0.13)	0.20 (0.07)	0.19 (0.05)	0.17 (0.03)	0.17 (0.03)	0.18 (0.05)	0.19 (0.05)	0.20 (0.08)
			Pred. Err.	0.52 (0.31)	0.37 (0.14)	0.30 (0.07)	0.41 (0.08)	0.40 (0.08)	0.33 (0.08)	0.35 (0.09)	0.37 (0.16)
0.1	3	0.9	P(exists)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	0.69 (0.39)	0.53 (0.21)	0.52 (0.13)	0.36 (0.10)	0.36 (0.09)	0.53 (0.15)	0.45 (0.17)	0.49 (0.14)
			Pred. Err.	0.55 (0.46)	0.37 (0.17)	0.29 (0.07)	0.41 (0.07)	0.40 (0.08)	0.36 (0.12)	0.34 (0.12)	0.33 (0.10)
0.1	8	0.5	P(exists)	0.78 (0.42)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	0.35 (0.32)	0.15 (0.07)	0.10 (0.03)	0.18 (0.03)	0.20 (0.01)	0.15 (0.02)	0.16 (0.18)	0.14 (0.06)
			Pred. Err.	0.92 (0.98)	0.36 (0.27)	0.17 (0.04)	0.56 (0.09)	0.61 (0.03)	0.42 (0.07)	0.43 (0.59)	0.38 (0.18)
0.1	8	0.9	P(exists)	0.62 (0.49)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	0.91 (0.84)	0.43 (0.31)	0.24 (0.07)	0.20 (0.04)	0.19 (0.03)	0.27 (0.07)	0.24 (0.07)	0.21 (0.03)
			Pred. Err.	1.25 (1.57)	0.44 (0.44)	0.16 (0.04)	0.49 (0.11)	0.54 (0.06)	0.28 (0.11)	0.37 (0.16)	0.59 (0.02)
0.5	3	0.5	P(exists)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	NA	0.37 (0.21)	0.35 (0.05)	0.22 (0.03)	0.22 (0.02)	0.27 (0.03)	0.26 (0.05)	0.27 (0.04)
			Pred. Err.	NA	0.94 (0.51)	0.84 (0.08)	0.65 (0.05)	0.65 (0.05)	0.69 (0.05)	0.69 (0.10)	0.70 (0.08)
0.5	3	0.9	P(exists)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	NA	0.70 (0.19)	0.94 (0.12)	0.60 (0.08)	0.62 (0.08)	0.98 (0.14)	0.72 (0.12)	0.76 (0.11)
			Pred. Err.	NA	0.65 (0.08)	0.64 (0.06)	0.56 (0.05)	0.56 (0.05)	0.70 (0.10)	0.56 (0.06)	0.56 (0.05)
0.5	8	0.5	P(exists)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	NA	0.13 (0.01)	0.19 (0.03)	0.12 (0.01)	0.13 (0.01)	0.15 (0.02)	0.13 (0.02)	0.13 (0.02)
			Pred. Err.	NA	0.79 (0.01)	0.39 (0.06)	0.65 (0.08)	0.69 (0.05)	0.57 (0.04)	0.50 (0.09)	0.50 (0.09)
0.5	8	0.9	P(exists)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
			Est. Err.	NA	0.21 (0.03)	0.54 (0.11)	0.27 (0.03)	0.28 (0.04)	0.49 (0.09)	0.36 (0.09)	0.26 (0.03)
			Pred. Err.	NA	0.79 (0.01)	0.39 (0.06)	0.57 (0.11)	0.58 (0.07)	0.44 (0.12)	0.45 (0.14)	0.66 (0.01)

Table S1: The proportion of times an estimator exists, P(exists), as well as the average composite estimation error, Est. Err, and average prediction error, Pred. Err., are given for various estimators of β^* when $n = 100$. The sample standard deviation is given in brackets. The columns named MLE and Firth ML correspond to the maximum likelihood estimator and Firth's bias-reduced estimator (Firth, 1993) respectively. The other columns correspond to estimators obtained from corrected versions of the least squares estimator. The Oracle version uses the true value to estimate η^*, the other five estimators use Ridge regression, LASSO, SVD, SCAD and MCP estimators of η^*.

4.2. *Small-sample performance*

The finite sample performance of our estimator was tested by computing average composite estimation and prediction errors in various settings with n fixed. The data were generated as in section 4.1 with $n = 100$, $\rho \in \{0.5, 0.9\}$, and $\gamma \in \{3, 8\}$. The parameter β^* consisted of exactly $s = 5$ randomly chosen non-zero entries with equal and positive signal strength. In this case, we allowed the intercept effect to be zero in some cases.

For each combination of parameter values, $R = 100$ Monte Carlo replications were performed where the design matrix was kept fixed but a new random response variable was sampled each time. In each repetition, multiple estimators of β^* were obtained. The first was the usual logistic maximum likelihood estimator. We also calculated Firth's bias-reduced estimator (Firth, 1993). The others were corrections to the least squares estimator obtained by using the oracle, LASSO, ridge, SVD, SCAD and MCP estimators of $\hat{\eta}$. The results are given in Table S1. For each estimate $\hat{\beta}$ of β^*, the proportion of times the estimator existed, the average relative composite estimation error $\|\hat{\beta} - \beta^*\|_2/(\sqrt{p}\|\beta^*\|_2)$ and the average relative prediction error $\|X(\hat{\beta} - \beta^*)\|_2/\|X\beta^*\|_2$ were recorded as well as the standard errors for these quantities over replications. The estimation error was divided by $\sqrt{p}\|\beta^*\|_2$ to make the entries of Table S1 comparable on account of the varying dimension p and signal strength. Note that in some cases the maximum likelihood estimator did not exist and so the errors for the maximum likelihood estimator were averaged only over the simulations that returned a solution.

The results show that the corrected least squares estimators perform favourably in comparison to the maximum likelihood estimator and Firth's (1993) estimator, both in cases with and without data separation. This was most evident for the prediction error. In view of Theorem 5, this suggests that Firth's estimator may produce inconsistent predictions when $\kappa \neq 0$. A more formal analysis of Firth's estimator in high-dimensional settings is required to establish this theoretically. Comparing the various corrections to the least-squares estimator, the LASSO version performed the best in terms of estimation error, often out-performing even the oracle estimator. This is possible as the oracle estimator obtains β^* exactly from β^0, but it may not be the best correction of the estimator $\hat{\beta}^0$. Increasing the sample size severely increased the time required to compute Firth's estimator whilst the times required to compute the corrected least squares estimators were less affected.

We also obtained the Probe-Frontier correction (Sur & Candès, 2019) although omitted the results from Table S1. This gave very similar results to those obtained using Firth's estimator when the maximum likelihood estimator existed but provided no estimate when the data was separated. It was also considerably more computationally demanding. SLOE (Yadlowsky et al., 2021) may be used to reduce the computational burden, although SLOE is also unavailable when the data is separated and so is omitted from this study.

REFERENCES

BAI, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statist. Sinica* **9**.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

CANDÈS, E. & SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* **48**, 27–42.

FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FAN, J., SHAO, Q.-M. & ZHOU, W.-X. (2018). Supplement to "Are discoveries spurious? Distributions of maximum spurious correlations and their applications." .

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

HOERL, A. E. & KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

PETROV, V. V. (1995). *Limit theorems of probability theory: sequences of independent random variables*. Oxford University Press.

SUR, P. & CANDÈS, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 14516–14525.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

VERSHYNIN, R. (2018). *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press.

YADLOWSKY, S., YUN, T.AND MCLEAN, C. & D'AMOUR, A. (2021). SLOE: A faster method for statistical inference in high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

[*Received on ... ... Editorial decision on ... ...*]