*Biometrika* (2024), **111**, 3, *pp.* 989–1011 https://doi.org/10.1093/biomet/asad065
Advance Access publication 2 November 2023

# 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, South Kensington Campus, London SW7 2AZ, U.K.*
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 loglikelihood 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 response; 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

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

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

© 2023 Biometrika Trust
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

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 when covariates are continuous, and also the computational difficulty associated with the combinatorial quantities involved; see § 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 $(y_1, x_1^{\mathrm{T}})^{\mathrm{T}}, \ldots, (y_n, x_n^{\mathrm{T}})^{\mathrm{T}}$ are such that $y_i x_i^{\mathrm{T}} \beta \geqslant 0$ for all $i$ and some nonzero $\beta \in \mathbb{R}^p$ (Albert & Anderson, 1984). For centred Gaussian covariates, Candès & Sur (2020) derived the limiting 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 $\kappa$ 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 $\chi_1^2$ limiting approximation to the distribution of the likelihood ratio test statistic is poor. Related work is due to Zhao et al. (2023), 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 summarized 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 $\ell_\infty$ and $\ell_2$ 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.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

### 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 $\beta^*$. 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 $\beta^*$, we manipulate the least-squares estimator to obtain a corrected least-squares estimator whose entries converge uniformly to those of the parameter of interest. Whilst biased, we then show that the lasso (Tibshirani, 1996) estimate of $\beta^*$ produces a consistent estimator of $X\beta^*$ when the unknown parameter is suitably sparse and $\max_{i=1}^{n} |x_i^{\mathrm{T}}\beta^*| \leqslant c_1(\log n)^{1/2}$ for some $c_1 > 0$. These conditions bound the entries of the unknown parameter, thereby avoiding the issues caused by separation. Assumptions 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 parameterized the linear-in-probability model as

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

under the restriction that, for all data $x$, $|x^{\mathrm{T}}\beta^0| \leqslant 1$. Ordinary least squares was the recommended approach for estimating the unknown parameter $\beta^0$ as this is more robust than maximum likelihood estimation to observations that invalidate the condition $|x^{\mathrm{T}}\beta^0| \leqslant 1$. 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 $|x^{\mathrm{T}}\beta^0| \leqslant 1$ violates the formal definition of a statistical model given by McCullagh (2002). 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 $\beta^*$ 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, the bias reduced estimator of Firth (1993) has been recommended for use; see, for example, Heinze & Schemper (2002). Kosmidis & Firth (2021)

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

showed that Firth's estimator always exists and established an analogous result for a more general version obtained by penalizing the logistic loglikelihood 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 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 $\beta^*$ 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; Fan & Peng, 2004; Meier et al., 2008; van de Geer, 2008); however, their components are biased for $\beta_j^*$, $j = 1, \ldots, p$. 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), Cai et al. (2021), Ma et al. (2021) and Shi et al. (2021) that entail correcting the bias of penalized estimators and require a consistent estimator of $\beta^*$ in either the $\ell_1$ or $\ell_2$ 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 $x_i^{\mathrm{T}}$. We assume throughout that $X$ has full rank, a condition that can always be checked once the data have been observed and that does not affect the presence of separation. Let Col-Sp($X$) denote the column span of $X$ and $P_X = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ the projection matrix onto Col-Sp($X$). Each element of the response vector $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, taking values in $\{-1, 1\}$, is assumed to be an independent random variable with distribution given in (1). A realization 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, MLE, when it exists, of $\beta^*$, and let $\hat{\beta}^0 = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ be the ordinary least-squares, OLS, estimator. Define $\beta^0$ 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 $i$th entry $f(v_i)$. The vector $\ell_1$, $\ell_2$ and $\ell_\infty$ 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 $\|\cdot\|_{\mathrm{F}}$. 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

$$S^\perp = \{u \in \mathbb{R}^n : u^{\mathrm{T}}v = 0 \text{ for all } v \in S\}.$$

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

## 3. Exact inference with separated data

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

In the logistic regression model, the loglikelihood function at an observation $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is given by

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

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 loglikelihood function asymptotes and so inference via maximum likelihood fitting is unavailable.

Suppose that only inference on the first component $\beta_1^*$ is of interest, with the other entries $\beta_2^*, \ldots, \beta_p^*$ being regarded as nuisance parameters. When all entries of $\beta^*$ are of interest, each entry may be treated in turn as the single interest parameter. Cox (1958) argued that inference on $\beta_1^*$ should be based on the conditional distribution of $T_1$ given $T_2, \ldots, T_p$ given by

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

where

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

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

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

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 let

$$F_1(t_1 \mid b) = \sum_{v\geqslant t_1} f(v \mid b), \qquad F_2(t_1 \mid b) = \sum_{v\leqslant t_1} f(v \mid b)$$

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

$(\beta_1^-(T_1), \beta_1^+(T_1))$, where

$$F_1\{T_1 \mid \beta_1^-(T_1)\} = \vartheta/2 \quad \text{if } t_{\min} < T_1 \leqslant t_{\max},$$
$$\beta_1^-(T_1) = -\infty \quad \text{if } T_1 = t_{\min},$$

and

$$F_2\{T_1 \mid \beta_1^+(T_1)\} = \vartheta/2 \quad \text{if } t_{\min} \leqslant T_1 < t_{\max},$$
$$\beta_1^+(T_1) = \infty \quad \text{if } T_1 = t_{\max},$$

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

Let $(\beta_1^-(t_1), \beta_1^+(t_1))$ be the observed confidence interval for $\beta_1^*$. 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 $y_i x_i^{\mathrm{T}} \beta > 0$ for all $i = 1, \ldots, n$, the interval may equal the whole real line.

Proposition 1. *Suppose that 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*

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

*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 $\beta$. 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 § 3.2 below, 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 $\beta_1^*$ satisfying

$$\mathrm{pr}\{\beta_1^* \in \mathrm{CS}_\vartheta^{(1)}(T_1) \mid T_2 = t_2, \ldots, T_p = t_p\} \geqslant 1 - \vartheta \quad \text{for all } \beta_1^* \in \mathbb{R}.$$

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 that 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 $(-\infty, -B] \subseteq \mathrm{CS}_\vartheta^{(1)}(t_1)$. If, additionally, the data are completely separated with $\beta_1 = 0$, the conditional likelihood satisfies*

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

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 $\beta^*$ satisfying

$$\mathrm{pr}\{\beta^* \in \mathrm{CS}_\vartheta(Y)\} \geqslant 1-\vartheta \quad \text{for all } \beta^* \in \mathbb{R}^p$$

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 that 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 $x_i^{\mathrm{T}}\beta = 0$. Assume that $\vartheta < 2^{-m}$. 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 $\beta$ 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 $\beta_1^*$, whatever this might be. Even when the data can be separated but not completely separated, one-sided hypotheses of the form $H_0 \colon \beta_1^* > b_0$ or $H_0 \colon \beta_1^* < b_0$ 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 nonexistence of finite confidence intervals also affects estimation as, for example, it is impossible to guarantee that an estimate of an entry of $\beta^*$ lies in a small region about the unknown parameter with a predetermined probability. Indeed, if there existed such an estimate $\hat{\beta}_1$ of $\beta_1^*$ satisfying

$$\mathrm{pr}(|\hat{\beta}_1 - \beta_1^*| < \epsilon) \geqslant 1-\vartheta \quad \text{for all } \beta_1^* \in \mathbb{R}$$

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 $\beta_1^*$ with bounded support. Markov's inequality implies that the variance of any such estimate is unbounded as a function of the unknown parameter.

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 § 5 below, 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.

## 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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

behaviour of linear functions $\alpha^{\mathrm{T}}\hat{\beta}^0$, where choices of particular interest are $\alpha$ 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 that $\alpha \in \mathcal{B}_d$ for some $d > 0$, where

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

is the sparse $\ell_2$ 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

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

Proposition 2. *Define*

$$\mathbb{X}_B = \Big\{X : \sup_{\alpha\in\mathcal{B}_d} \|\alpha^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\|_2^2 \leqslant Bn^{-1}\Big\}$$

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

$$\sup_{X\in\mathbb{X}_B} \mathrm{pr}\Big\{\sup_{\alpha\in\mathcal{B}_d} |\alpha^{\mathrm{T}}(\hat{\beta}^0 - \beta^0)| \geqslant t\Big\} = o(1).$$

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 $\|\alpha^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\|_2^2 = O(n^{-1})$, for which it follows that

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

provided the diagonal entries of $(X^{\mathrm{T}}X/n)^{-1}$ 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 below strengthens the latter result, showing that the rate of convergence in the $\ell_2$ norm is of order $p^{-1/2}$.

Proposition 3. *Suppose that $\lambda_{\max}\{(X^{\mathrm{T}}X/n)^{-1}\} = O(1)$. 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 $\beta^*$, the limiting distribution of the least-squares estimator is of interest. Proposition 4 below shows that, after suitable normalization, the distribution of $\alpha^{\mathrm{T}}(\hat{\beta}^0 - \beta^0)$ is asymptotically Gaussian.

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

$$\sup_{\alpha\in\mathcal{R}} B_n^{-1}\|\alpha^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\|_\infty = o(1). \tag{2}$$

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

*Then,*

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

Assumption (2) 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 that the rows of $X$ are independently and identically distributed as centred, multivariate normal random variables with covariance matrix $\Sigma$ and focus on the limiting setting where $p,n\to\infty$ with $p/n\to\kappa\in[0,1)$ and ${\beta^*}^{\mathrm{T}}\Sigma\beta^*\to\gamma^2$ for some $\gamma>0$. This is a setting that will be considered further in § 7 below. Let $\mathcal{R}$ consist of the standard basis vectors of $\mathbb{R}^p$. Theorem 2.16 of Bai (1999) shows that

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

Furthermore, $\max_{i=1}^n|x_i^{\mathrm{T}}\beta^*|=O_P\{(\log n)^{1/2}\}$ and so

$$-\log\Big(\min_{i=1}^n\Gamma_{ii}\Big)\leqslant\max_{i=1}^n|x_i^{\mathrm{T}}\beta^*|=O_P\{(\log n)^{1/2}\}.$$

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

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

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

$$\|\alpha^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\|_\infty=O_P\big(n^{-1}(\log n)^{1/2}\big)=o_P\big(n^{-1/2}e^{-c(\log n)^{1/2}}\big)$$

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

## 5. Main results

### 5.1. *The corrected least-squares estimator*

The previous results motivate a corrected least-squares estimator, which is shown in the present section to be consistent in both the $\ell_\infty$ and scaled $\ell_2$ norms, and to have some predictive guarantees. The considerations involved in obtaining stronger inferential guarantees are also briefly discussed and assessed by simulation.

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

If this function $f$ were invertible then a consistent estimator of each entry of $\beta^*$ could potentially 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 $\beta^*$ are difficult to estimate, estimation of $X\beta^*$ is simpler and leads to an estimator of $\alpha^{\mathrm{T}}\beta^*$ in the form

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

where $\hat{\varsigma}$ and $\hat{\delta}$ are estimates of $\varsigma$ and $\delta$ to be defined.

Write

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

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

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

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

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

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

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

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

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

### 5.2. *Assumptions*

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

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

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 nonempty $\mathcal{H}_B \subseteq \mathbb{X}_B^{(1)}$ 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

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

Condition 1 makes restrictions on the unknown parameter that avoids the issues outlined in § 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 $\beta^*$ contains at most $s = O(1)$ nonzero 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,

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

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 $\varsigma$ and $\delta$. 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 $\beta^* \neq 0$,

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

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 in the Supplementary Material for details. In § 6 below we identify suitable choices of $\hat{\eta}$ and $\mathcal{H}_B$ that ensure that Condition 2 is met.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

For most of the analysis, it will be sufficient to focus on a single $X \in \mathcal{H}_B$. However, in § 7 below, we consider the validity of our results in settings with separated data, focusing on the accuracy of our estimator in terms of the $\ell_\infty$ norm. As our results are asymptotic and the limiting probability of data separation has not yet been considered for fixed designs, § 7 treats the design matrix as random. To extend our analysis to this framework, we show that consistency in terms of the $\ell_\infty$ 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 $\beta^*$ 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 § 7.

Our assumptions make no explicit restrictions on $p$ relative to $n$ beyond $p < n$. However, there may be some implicit constraints. Section 7 below 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 § 5.2, the corrected least-squares estimator $\tilde{\beta}^*$ is consistent, both entrywise and in terms of the $\ell_2$ norm. We refer to the latter as the composite estimation error. This is established in Theorems 3 and 4 below. In light of the comments in § 5.2, consistency in terms of the $\ell_\infty$ 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 that Conditions 1(b) and 2 hold. For all $t > 0$,*

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

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

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

Theorem 4. *Suppose that Conditions 1(a) and 2 hold. If there exist $B, N > 0$ such that $X \in \mathcal{H}_B$ for all $n \geqslant N$ 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 § 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 that the nonzero entries of $\beta^*$ 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 prioritize 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.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

### 5.4. *A remark on inference*

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

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

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

*Then, under the null hypothesis* $H_0 \colon \alpha^{\mathrm{T}}\beta^* = 0$,

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

*where*

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

Proposition 5 shows that, up to the term $\Pi_2$, the limiting distribution of a scaled version of our estimator is standard normal. In § 8 below, we conduct simulations whose results suggest that $\Pi_2$ 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 $\alpha^{\mathrm{T}}\beta^* \neq 0$. Neither observation has yet been established theoretically. Part of the difficulty arises in the term $B_n^{-1}$, which up to the quantity $\Gamma$ is of order $n^{1/2}$. Then, for $\Pi_2$ to converge to zero in probability, we would require

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

however, we have only shown that

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

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 been considered in detail, and 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 below 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 that there exist* $B, N > 0$ *such that* $X \in \mathcal{H}_B$ *for all* $n \geqslant N$. *Then,*

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

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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

When $p = o(n)$, the error is $o_P(1)$ as $\|P_X \Gamma^{1/2}\|_{\mathrm{F}}$ is bounded above by $\|P_X\|_{\mathrm{F}} = \sqrt{p}$. When $p/n$ does not converge to zero, there exist values of $\beta^*$ where the prediction error does not decay to zero. As a result, whilst the corrected least-squares estimator may be usefully employed 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 $\eta^*$, consider an iterative version of our estimator where, for $i = 1, 2, \ldots,$

$$\hat{\varsigma}^{(i)} = \begin{cases} 1/2, & \hat{\eta}^{(i)} = 0, \\ \dfrac{(\hat{\eta}^{(i)})^{\mathrm{T}} \tanh(\hat{\eta}^{(i)}/2)}{\|\hat{\eta}^{(i)}\|_2^2}, & \hat{\eta}^{(i)} \neq 0, \end{cases}$$

$$\hat{\delta}^{(i)} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\{P_X - P_{\hat{\eta}^{(i)}}\}\tanh(\hat{\eta}^{(i)}/2),$$

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

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 below shows that our estimator obtained by setting $\hat{\eta}$ to $X\hat{\beta}^*$ recovers the original maximum likelihood estimator.

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

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

*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. *Penalized regression*

To operationalize $\tilde{\beta}^*$, a consistent estimator $\hat{\eta}$ of $\eta^* = X\beta^*$ satisfying $h(\hat{\eta}, \eta^*) = o_P(1)$ is needed. A penalized regression may be used to obtain such an estimator even when the

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

maximum likelihood estimator does not exist. This entails setting $\hat{\eta} = X\hat{\beta}^{(\lambda)}$, where

$$\hat{\beta}^{(\lambda)} = \arg\min_{\beta \in \mathbb{R}^p} \{-n^{-1}\ell(\beta) + \lambda p(\beta)\}, \qquad \lambda \geqslant 0, \tag{3}$$

and $p\colon \mathbb{R}^p \to \mathbb{R}$ is a penalty function that does not depend on the data, but ensures a unique maximiser exists. See Duffy & Santner (1989), Fan & Peng (2004), Meier et al. (2008) and Kosmidis & Firth (2021) for examples including the ridge, group lasso, Jeffreys-prior and nonconcave 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 $\beta^*$; see § 8 below for numerical examples. Nevertheless, under much weaker conditions, $X\hat{\beta}^{(\lambda)}$ consistently estimates $\eta^*$. The corrected least-squares estimator that makes use of $X\hat{\beta}^{(\lambda)}$ to estimate $\eta^*$ improves estimation of the entries of $\beta^*$ over that of $\hat{\beta}^{(\lambda)}$.

As an example, consider the lasso estimator that maximizes (3) with $p(\beta) = \|\beta\|_1$. Lemma 1 of 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. Furthermore, under a suitable sparsity assumption, the lasso estimator produces consistent predictions whenever the design matrix is contained in

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

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

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

This is established in the following proposition.

PROPOSITION 7. *Let* $\lambda = A\{(\log p \log n)/n\}^{1/2}$ *with sufficiently large* $A > 0$. *For* $t > 0$,

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

*when*

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

The proof of this result closely follows the arguments in Theorem 6.4 and Lemma 6.8 of Bühlmann & van de Geer (2011, pp. 130–4). 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).

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

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 $\ell_2$ and $\ell_\infty$ norms whenever there exists $B, N > 0$ such that $X \in \mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ for all $n \geqslant N$. We call this the OLS-LASSO estimator. In the following section we show that when the rows of $X$ follow a multivariate Gaussian distribution and $t > 0$, there exist suitable constants $B, N > 0$ such that the probability that $X \in \mathbb{X}_B^{(1)}) \cap \mathbb{X}_B^{(2)}$ is at least $1 - t$ when $n \geqslant N$. Thus, the set $\mathbb{X}_B^{(1)} \cap \mathbb{X}_B^{(2)}$ is nonempty.

### 6.3. *Other approaches*

Any consistent estimator of $\eta^* = X\beta^*$ can be used for correction. Thus, if $\beta^*$ is sparse, Proposition 7 shows that a lasso penalty yields an estimator with appropriate behaviour. If, instead, $\beta^*$ 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 $\eta^*$ 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 § 6.2. As no restrictions on the observed response $y$ are made in § 5.2, and § 6.2 only assumes that $0 < \sum_{i=1}^n y_i < n$, 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 nonzero. If this were not the case then, for any $t > 0$,

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

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 summarized in the following condition.

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

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

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

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

as $p, n \to \infty$.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

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

$$\kappa > h(\gamma) \quad \Longrightarrow \quad \mathrm{pr}_{Y,X}(\text{data are separated}) \to 1,$$
$$\kappa < h(\gamma) \quad \Longrightarrow \quad \mathrm{pr}_{Y,X}(\text{data are separated}) \to 0.$$

Our aim is to show that the OLS-LASSO estimator is consistent with respect to the $\ell_\infty$ norm even in these settings where the limiting probability of separation is one. For this, we also assume that the following condition holds.

*Condition* 4. (a) The unknown parameter satisfies $\|\beta^*\|_\infty = O(1)$ and $\|\beta^*\|_0 = O(n^{1/2-\xi})$ for some $\xi \in (0, 1/2)$.

(b) There exist constants $\lambda_{\min}, \sigma_{\max} > 0$ such that

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

For every $\kappa \in [0, 1)$ and $\gamma \geqslant 0$, there exist choices of $\Sigma$ and $\beta^*$ satisfying the conditions above. For example, let $\beta^* \in \mathbb{R}^p$ have exactly one nonzero entry taking the value $\gamma > 0$ and let $\Sigma = (1/4)\mathbb{1}\mathbb{1}^{\mathrm{T}} + (3/4)\mathbb{I}_{p\times p}$. 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 $\kappa$ is sufficiently large compared to $\gamma$ then consistency is achieved alongside separation.

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

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

*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 characterizing the limiting probability of separation for fixed designs to ensure that similar guarantees can also be provided in these settings.

## 8. Numerical performance

Results of extensive numerical experimentation are reported in the Supplementary Material. In § 4.1 in the Supplementary Material, 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 componentwise and composite estimation errors converged to zero, whilst the prediction error

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

remained stable at a nonzero value, coinciding with the analysis in § 5. The small-sample performance was analysed in § 4.2 within 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 the estimator of Firth (1993).

The remainder of this section fixes $n = 700$ and $p = 70$ and provides a comparison to other available methods. Based on the results in the Supplementary Material, we focus on the smoothly clipped absolute deviation, 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-penalized 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 penalized 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 penalized estimators in high-dimensional settings, although data separation was not explicitly considered.

The data were generated as defined in § 7, with the rows of $X$ sampled independently from a $p$-dimensional multivariate Gaussian distribution with mean zero and covariance matrix $\Sigma$ 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.

Because of the dependence structure among the covariates, these scenarios exemplify situations where the individual entries of $\beta^*$ are difficult to estimate accurately. In $R = 500$

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

Monte Carlo replications, we obtained the aforementioned estimates of $\beta^*$. To probe the inferential capabilities beyond point estimation, for all except the SCAD-penalized regression, we also computed 95% confidence intervals for $\beta_1^*$ and $\beta_4^*$, the former corresponding to a signal variable and the latter to a noise variable. Motivated by the results in § 5.4, we defined a $\vartheta$-level test for the hypothesis $H_0 : \alpha^{\mathrm{T}}\beta^* = b_0$ based on the least-squares estimator of the form

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

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; the numerical results serve only to gain insights into the relevance of the term $\Pi_2$ in Proposition 5. The R functions `cv.ncvreg`, `logistf`, `lasso.proj` in the `hdi` package and `LF` in the `SIHR` package (R Development Core Team, 2024) 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.

The left column of Fig. 1 shows the average estimated signal strength of entries 1–6 of $\beta^*$ for each example. The right column shows the distribution of estimates of $\beta_1^*$ 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 $\beta^*$ failed to accurately characterize the two signal variables that were only weakly marginally related to the response variable, whereas 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 nonnegligible 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, 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 the latter has smaller cumulative estimation and prediction error, particularly when $p$ is large and $\beta^*$ 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 Fig. 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 nonzero signal strengths 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.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

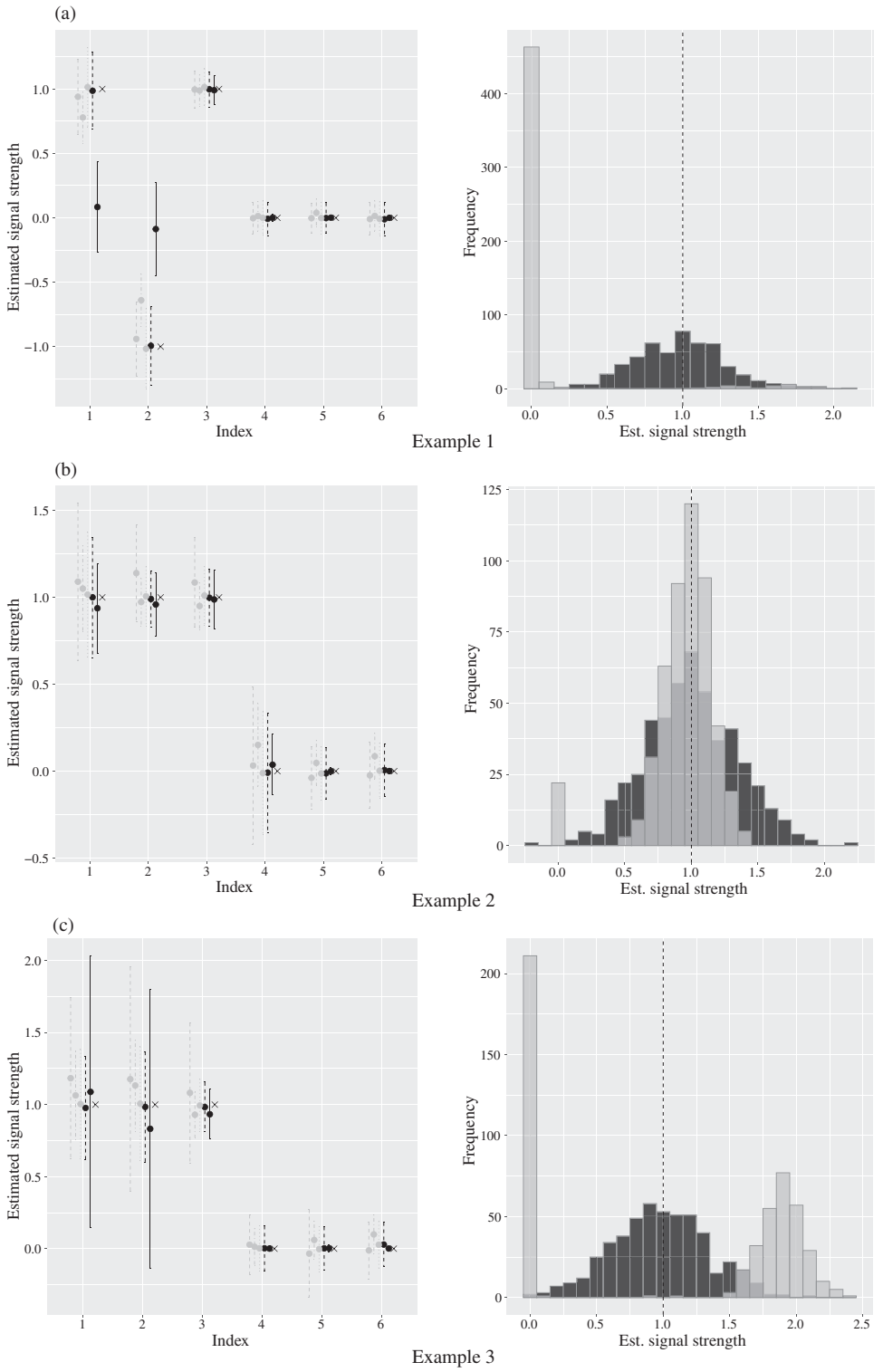



Fig. 1. Left column: average estimated signal strengths of entries 1-6 of $\beta^*$ in each example obtained using OLS-SCAD (black-dashed), SCAD-penalized regression (black-solid), Firth's estimator (grey-dotted), desparsified lasso (grey-dotdashed) and LSW (grey-dashed). Error bars show one estimated standard deviation. True signal strengths are marked with black crosses. Right column: histogram of estimates of $\beta^*$ obtained via OLS-SCAD (black) and SCAD penalized regression (grey) for the second signal variable in each example. The black dashed line represents the true signal strength.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

Table 1. *Average length* $|\mathcal{I}|$*, coverage probability* $\mathrm{pr}(\beta_i^* \in \mathcal{I})$ *and power* $\mathrm{pr}(0 \in \mathcal{I})$ *of the confidence intervals obtained via the different methods. Standard errors are given in brackets.*

| | Method | $\lvert\mathcal{I}\rvert$ (Null variable) | $\mathrm{pr}(\beta_i^* \in \mathcal{I})$ | $\lvert\mathcal{I}\rvert$ (Signal variable) | $\mathrm{pr}(\beta_i^* \in \mathcal{I})$ | $\mathrm{pr}(0 \notin \mathcal{I})$ |
|---|---|---|---|---|---|---|
| Example 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) |
| Example 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) |
| Example 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) |

OLS-SCAD, the corrected OLS estimator; FIRTH, the method of Firth (1993); DLASSO, desparsified lasso; LSW, the approach of Cai et al. (2021).

## 9. Discussion

### 9.1. *Extension to cases with* $p \geqslant n$

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, Cox & Battey (2017) fitted a large number of low-dimensional regressions, 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 $\beta^*$ 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 § 8 exhibit this most clearly, providing examples where a penalized regression estimator fails to accurately characterize 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-penalized 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

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

outlined in § 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.

## Acknowledgement

This work was partially supported by an Engineering and Physical Sciences Research Fellowship (EP/T01864X/1).

## Supplementary material

The Supplementary Material 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**, 611–62.
Bartlett, M. S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.* **32**, 560–6.
Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond.* A **160**, 268–82.
Battey, H. S., Cox, D. R. & Jackson, M. V. (2019). On the linear in probability model for binary data. *R. Soc. Open Sci.* **6**, doi: 10.1098/rsos.190067.
Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.
Cai, T. T., Guo, Z. & Ma, R. (2021). Statistical inference for high-dimensional generalized linear models with binary outcomes. *J. Am. Statist. Assoc.* **118**, 1319–32.
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). *J. R. Statist. Soc.* B **20**, 215–42.
Cox, D. R. (1970). *Analysis of Binary Data*. London: Methuen.
Cox, D. R. & Battey, H. S. (2017). Large numbers of explanatory variables, a semi-discriptive analysis. *Proc. Nat. Acad. Sci. USA* **114**, 8592–5.
Cox, D. R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–61.
Duffy, D. & Santner, T. (1989). On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Commun. Statist. A* **18**, 159–980.
Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–60.
Fan, J. & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928–61.
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. *Statist. Med.* **21**, 2409–19.
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.
Lei, L., Bickel, P. & El Karoui, N. (2018). Asymptotics for high dimensional regression M-estimates: fixed design results. *Prob. Theory Rel. Fields* **172**, 983–1079.
Ma, R., Cai, T. T. & Li, H. (2021). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. *J. Am. Statist. Assoc.* **116**, 984–98.
McCullagh, P. (2002). What is a statistical model? *Ann. Statist.* **30**, 1225–67.
Mehta, C. R. & Patel, N. R. (1995). Exact logistic regression: theory and examples. *Statist. Med.* **14**, 2143–60.

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

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**, 158–95.

R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Raskutti, G., Wainwright, M. & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE Trans. Inf. Theory* **57**, 6976–94.

Shi, C., Song, R., Lu, W. & Li, R. (2021). Statistical inference for high-dimensional models via recursive online-score estimation. *J. Am. Statist. Assoc.* **116**, 1307–18.

Sur, P. & Candès, E. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Nat. Acad. Sci. USA* **116**, 14516–25.

Tang, Y. & Reid, N. (2020). Modified likelihood root in high dimensions. *J. R. Statist. Soc.* B **82**, 1349–69.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–45.

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**, 1166–202.

Wald, A. (1950). *Statistical Decision Functions*. London: John Wiley and Sons.

Yadlowsky, S., Yun, T., 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, Ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang and J. Wortman Vaughan, pp. 29517–28. Red Hook, NY: Curran Associates.

Zhao, Q., Sur, P. & Candès, E. (2023). The asymptotic distribution of the MLE in high-dimensional logistic models: arbitrary covariance. *arXiv:* 2001.09351v2.

[*Received on* 17 *March* 2022. *Editorial decision on* 12 *October* 2023]

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024

Downloaded from https://academic.oup.com/biomet/article/111/3/989/7338235 by guest on 18 September 2024