Title

Debiased lasso for generalized linear models with a diverging number of covariates.

Journal

Biometrics, 79(1)

Authors

Xia, Lu

Li, Yi

Nan, Bin

Publication Date

2023-03-01

DOI

10.1111/biom.13587

Peer reviewed

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2024 March 01.

Published in final edited form as:

Biometrics. 2023 March ; 79(1): 344–357. doi:10.1111/biom.13587.

De-biased Lasso for Generalized Linear Models with A Diverging Number of Covariates

Lu Xia[1], **Bin Nan**[2], **Yi Li**[3]

[1]Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.

[2]Department of Statistics, University of California, Irvine, Irvine, CA, U.S.A.

[3]Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.

Summary:

Modeling and drawing inference on the joint associations between single nucleotide polymorphisms and a disease has sparked interest in genome-wide associations studies. In the motivating Boston Lung Cancer Survival Cohort (BLCSC) data, the presence of a large number of single nucleotide polymorphisms of interest, though smaller than the sample size, challenges inference on their joint associations with the disease outcome. In similar settings, we find that neither the de-biased lasso approach (van de Geer et al., 2014), which assumes sparsity on the inverse information matrix, nor the standard maximum likelihood method can yield confidence intervals with satisfactory coverage probabilities for generalized linear models. Under this "large n, diverging p" scenario, we propose an alternative de-biased lasso approach by directly inverting the Hessian matrix without imposing the matrix sparsity assumption, which further reduces bias compared to the original de-biased lasso and ensures valid confidence intervals with nominal coverage probabilities. We establish the asymptotic distributions of any linear combinations of the parameter estimates, which lays the theoretical ground for drawing inference. Simulations show that the proposed *refined* de-biased estimating method performs well in removing bias and yields honest confidence interval coverage. We use the proposed method to analyze the aforementioned BLCSC data, a large scale hospital-based epidemiology cohort study, that investigates the joint effects of genetic variants on lung cancer risks.

Keywords

confidence interval; de-biased lasso; diverging dimension; generalized linear models; high-dimensional regression

* nanb@uci.edu; yili@umich.edu.

Supporting Information

Web Appendices referenced in Sections 3, 4, 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. R code for implementing the two de-biased lasso methods and a simulated example are available at https://github.com/luxia-bios/DebiasedLassoGLMs/.

1. Introduction

To identify disease-related genetic markers, traditional genome-wide association studies typically analyze the marginal associations of the disease outcome with single nucleotide polymorphisms (SNPs), one at a time. As marginal associations do not account for the dependence among SNPs, false positive discoveries may occur as SNPs can be claimed as significant when they are correlated with the causal variants (Schaid et al., 2018). Alternatively, modeling the joint effects of SNPs within the target genes can reduce false positives around true causal SNPs and improve prediction accuracy (He and Lin, 2010), and also can pinpoint functionally impactful loci in the coding regions (Taylor et al., 2001; Repapi et al., 2010) so as to better understand the molecular mechanisms underlying cancer (Guan and Stephens, 2011). For example, among a subset of 1,374 patients from the Boston Lung Cancer Survival Cohort (BLCSC), an epidemiology study that investigates molecular mechanisms underlying lung cancer, our goal is to study the joint associations of lung cancer risk with over 100 SNPs residing in nine target genes that have been reported to harbor relevant genetic variants (McKay et al., 2017). The results may aid in personalized medicine by properly implicating relevant genetic variants and their joint roles in pharmacogenomics (Evans and Relling, 2004). Statistically, the analysis requires reliable estimation and inference on a fairly large number of regression parameters.

With lung cancer mechanisms differing by smoking predisposition (Bossé and Amos, 2018), analyzing BLCSC among the 1,077 smokers and 297 non-smokers, separately, is necessary. Included in our models are 103 SNPs and 4 demographic variables, which, though smaller than the number of smokers or non-smokers, are large enough to defy the conventional maximum likelihood estimation (MLE) approach. In particular, for non-smokers, Table 2 in Section 5 has shown unreasonably large MLE estimates with wide confidence intervals, e.g. a point estimate of −19.64 with a 95% confidence interval (−6,705.04, 6,665.75) for SNP AX-62479186. Failures of MLE in similar scenarios have been documented in Sur and Candès (2019), and further evidenced by our later simulation studies.

The asymptotic framework underlying these cases can be characterized as the number of parameters p increasing with the sample size n, rather than staying fixed, which is often referred to as the "large n, diverging p" scenario. Drawing inference with generalized linear models (GLMs) under this framework may facilitate a range of applications, because the setting enables us to build valid models when the collected information increases with more subjects included in the study (Wang, 2011). Several authors (Huber, 1973; Yohai and Maronna, 1979; Portnoy, 1984, 1985) investigated the relative order between p and n that ensures the validity of M-estimators in linear regression; He and Shao (2000) studied the consistency and the asymptotic normality of the M-estimators under different conditions and showed that $p^2 \log(p)/n \to 0$ would be needed for linear and logistic regression; Wang (2011) developed an asymptotic theory for the estimated regression parameters from generalized estimating equations with clustered binary outcomes, provided $p^3/n \to 0$. However, most of these methods incur substantial biases in empirical studies unless p is very small.

Penalized regression methods have been developed over the decades to accommodate a large number of covariates. These methods, including the lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005) and the Dantzig selector (Candès and Tao, 2007) among many others, are considered to be useful alternatives to the traditional variable selection methods such as forward or stepwise selection, especially in genetic studies (Schaid et al., 2018). These regularized methods yield biased estimates, and, thus, cannot be directly used for drawing inference such as constructing confidence intervals with a nominal coverage probability.

One stream of inferential methods is the post-selection inference conditional on selected models (Lee et al., 2016), which requires conditional coverage to quantify the uncertainty associated with model selection. Other super-efficient procedures, such as SCAD (Fan and Li, 2001; Fan and Peng, 2004) and adaptive lasso (Zou, 2006), share the flavor of post-selection inference that is not the focus of this article. In particular, the inference based on the oracle estimation of Fan and Peng (2004) requires $p^5/n \to 0$.

Another school of methods is to draw inference by de-biasing the lasso estimates, termed de-biased lasso or de-sparsified lasso, which relieves the restrictions of post-selection inference and possesses nice theoretical and numerical properties in linear regression models (van de Geer et al. 2014; Zhang and Zhang 2014; Javanmard and Montanari 2014).

van de Geer et al. (2014) extended de-biased lasso to GLMs and developed the asymptotic normality theory for each component of the coefficient estimates; based on this work, Zhang and Cheng (2017) proposed a multiplier bootstrap procedure to draw inference on a group of coefficients in GLMs. However, the de-biased lasso approach presented subpar performance with non-negligible biases and poor coverage of confidence intervals, as seen from Figures 1 and 2 for a logistic example in Section 4 that mimics the BLCSC setting, because a key sparsity assumption on the inverse information matrix may not hold in GLM settings.

To address the limitation and for valid inference with GLMs, we propose a *refined* de-biased lasso estimating method specifically tailored to the "large n, diverging p" scenario as in the motivating BLCSC dataset. Our proposed method estimates the inverse information matrix by directly inverting the sample Hessian matrix, which requires no structural assumptions on the inverse information matrix. We establish the asymptotic distributions for any linear combinations of the resulting estimates, laying the theoretical foundation for applications. Simulations demonstrate its better performance in reducing biases and preserving confidence interval coverage probabilities than the conventional MLE and the original de-biased lasso (van de Geer et al., 2014) for a wide range of p/n ratios, and all three methods yield almost identical results when p is rather small relative to n.

The rest of this article is organized as follows. Section 2 describes in detail the model setup and the proposed *refined* de-biased lasso estimating method. Asymptotic results for the proposed method are provided in Section 3, followed by simulation studies in Section 4. Findings on the joint associations between SNPs in target genes and lung cancer risks by applying the proposed method to the motivating BLCSC data are reported in Section 5. Not

to deviate from the main flow, we put off the discussion of the distinctions of the proposed method from the existing high-dimensional inference literature to Section 6.

2. Method

2.1 Background and set-up in generalized linear models

We start with some commonly used notation. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|_q$ denotes its ℓ_q norm, $q \geqslant 1$. Denote by $\lambda_{\max}(\boldsymbol{A})$ and $\lambda_{\min}(\boldsymbol{A})$ the largest and the smallest eigenvalues of a symmetric matrix $\boldsymbol{A}$, respectively. For a real matrix $\boldsymbol{A} = (A_{ij})$, let $\| \boldsymbol{A} \| = \sup_{\| x \|_2 = 1} \| \boldsymbol{Ax} \|_2 = \left[\lambda_{\max}(\boldsymbol{A}^T\boldsymbol{A})\right]^{1/2}$ be the spectral norm of $\boldsymbol{A}$. The induced matrix ℓ_1 norm is $\|\boldsymbol{A}\|_1 = \max_j \Sigma_i |A_{ij}|$, and when $\boldsymbol{A}$ is symmetric, $\|\boldsymbol{A}\|_1 = \max_i \Sigma_j |A_{ij}|$ also holds. The element-wise ℓ_∞ norm is $\|\boldsymbol{A}\|_\infty = \max_{i,j} |A_{ij}|$. With two positive sequences a_n and b_n, write $a_n = \mathcal{O}(b_n)$ if there exist $c > 0$ and $N > 0$ such that $a_n < cb_n$ for all $n > N$, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We write $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$.

Denote by y_i the response variable and $\boldsymbol{x}_i = \left(1, \tilde{\boldsymbol{x}}_i^T\right)^T \in \mathbb{R}^{p+1}$ for $i = 1, \ldots, n$, where "1" corresponds to the intercept term, and $\tilde{\boldsymbol{x}}_i$ represents the p covariates. Let $\boldsymbol{X}$ be the $n \times (p+1)$ covariate matrix with $\boldsymbol{x}_i^T$ being the ith row. We assume that $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ are independent and identically distributed copies of $(y, \boldsymbol{x})$. Define the *negative* log-likelihood function as the following, up to an additive constant irrelevant to the unknown parameters, when the conditional density of y given $\boldsymbol{x}$ belongs to an exponential family:

$$\rho_\xi(y, \boldsymbol{x}) = \rho\left(y, \boldsymbol{x}^T\boldsymbol{\xi}\right) = -y\boldsymbol{x}^T\boldsymbol{\xi} + b\left(\boldsymbol{x}^T\boldsymbol{\xi}\right) \tag{1}$$

where $b(\cdot)$ is a known twice continuously differentiable function, $\boldsymbol{\xi} = \left(\beta_0, \boldsymbol{\beta}^T\right)^T \in \mathbb{R}^{p+1}$ denotes the vector of coefficients and $\beta_0 \in \mathbb{R}$ is the intercept parameter. The unknown true coefficient vector is $\boldsymbol{\xi}^0 = \left(\beta_0^0, \boldsymbol{\beta}^{0T}\right)^T$.

2.2 De-biased lasso

With $\rho_\xi(y, \boldsymbol{x}) = \rho(y, \boldsymbol{x}^T\boldsymbol{\xi})$ given in (1), denote by $\dot{\rho}_\xi$ and $\ddot{\rho}_\xi$ its first and second order derivatives with respect to $\boldsymbol{\xi}$, respectively. For any function $g(y, \boldsymbol{x})$, let $P_n g = n^{-1}\sum_{i=1}^n g(y_i, \boldsymbol{x}_i)$. Then for any $\boldsymbol{\xi} \in \mathbb{R}^{p+1}$, we denote the empirical loss function based on the random sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ by $P_n\rho_\xi = n^{-1}\sum_{i=1}^n \rho_\xi(y_i, \boldsymbol{x}_i)$, and its first and second order derivatives with respect to $\boldsymbol{\xi}$ by $P_n\dot{\rho}_\xi = n^{-1}\sum_{i=1}^n \partial\rho_\xi(y_i, \boldsymbol{x}_i)/\partial\boldsymbol{\xi}$ and $\widehat{\boldsymbol{\Sigma}}_\xi = P_n\ddot{\rho}_\xi = n^{-1}\sum_{i=1}^n \partial^2\rho_\xi(y_i, \boldsymbol{x}_i)/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^T$. Two important population-level matrices are the information matrix, $\boldsymbol{\Sigma}_\xi = E\left(\widehat{\boldsymbol{\Sigma}}_\xi\right) = E(P_n\ddot{\rho}_\xi)$, and its inverse $\boldsymbol{\Theta}_\xi = \boldsymbol{\Sigma}_\xi^{-1}$. With a tuning parameter $\lambda > 0$, the lasso estimator for $\boldsymbol{\xi}^0$ is defined as

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi} = (\beta_0, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{p+1}}{\arg\min} \{P_n\rho_\xi + \lambda \| \boldsymbol{\beta} \|_1\}, \tag{2}$$

where we suppress the dependence of λ on n and p for notational ease. We clarify that we do not penalize the intercept β_0 in (2). As such, the theoretical properties for $\hat{\boldsymbol{\xi}}$, including the bounds of estimation errors and prediction errors, are still the same as those in van de

Geer (2008) and van de Geer et al. (2014), where all of the parameters are estimated via penalization (Bühlmann and van de Geer, 2011).

We briefly review the de-biased lasso estimator and its bias decomposition. The first order Taylor expansion of $P_n\dot{\rho}_{\xi^0}$ at $\hat{\xi}$ gives

$$P_n\dot{\rho}_{\xi^0} = P_n\dot{\rho}_{\hat{\xi}} + P_n\ddot{\rho}_{\hat{\xi}}\left(\xi^0 - \hat{\xi}\right) + \boldsymbol{\Delta}, \tag{3}$$

where $\boldsymbol{\Delta}$ is a $(p+1)$-dimensional vector of remainder terms with the jth element

$$\Delta_j = \frac{1}{n}\sum_{i=1}^{n}\left\{\ddot{\rho}\left(y_i, a_j^*\right) - \ddot{\rho}\left(y_i, \boldsymbol{x}_i^T\hat{\xi}\right)\right\}x_{ij}\boldsymbol{x}_i^T\left(\xi^0 - \hat{\xi}\right),$$

in which $\ddot{\rho}(y,a) = \partial^2\rho(y,a)/\partial a^2$, and a_j^* lies between $\boldsymbol{x}_i^T\hat{\xi}$ and $\boldsymbol{x}_i^T\xi^0$. In linear regression models, $\boldsymbol{\Delta} = \mathbf{0}$, which is not always the case for GLMs. Let $\boldsymbol{M}$ be a $(p+1)\times(p+1)$ matrix approximating $\boldsymbol{\Theta}_{\xi^0}$. Multiplying both sides of (3) by $\boldsymbol{M}_j$, the jth row of $\boldsymbol{M}$, we obtain the following equality for the jth component

$$\hat{\xi}_j - \xi_j^0 + \overbrace{\left(-\boldsymbol{M}_j\boldsymbol{P}_n\dot{\rho}_{\hat{\xi}}\right)}^{I_j} + \overbrace{\left(-\boldsymbol{M}_j\boldsymbol{\Delta}\right)}^{II_j} + \overbrace{\left(\boldsymbol{M}_j\boldsymbol{P}_n\ddot{\rho}_{\hat{\xi}} - \boldsymbol{e}_j^T\right)\left(\hat{\xi} - \xi^0\right)}^{III_j} = -\boldsymbol{M}_j\boldsymbol{P}_n\dot{\rho}_{\xi^0}, \tag{4}$$

where $\boldsymbol{e}_j$ is the unit vector with the jth element being 1. van de Geer et al. (2014) obtained the above decomposition by inverting the Karush–Kuhn–Tucker condition while using the node-wise lasso estimate of $\boldsymbol{\Theta}_{\xi^0}$, denoted by $\widetilde{\boldsymbol{\Theta}}$, to be the approximation matrix $\boldsymbol{M}$. Originally proposed for neighborhood selection in high-dimensional graphs (Meinshausen and Bühlmann, 2006), the node-wise lasso approach estimates a sparse matrix $\boldsymbol{\Theta}_{\xi^0}$ that consists of many zero elements. In (4), the asymptotic bias term I_j is estimable, and $\hat{\xi}_j + I_j$ corresponds to the de-biased lasso estimator in van de Geer et al. (2014) with $\boldsymbol{M} = \widetilde{\boldsymbol{\Theta}}$. In practice, the II_j and III_j terms in (4) are not computable because they involve the unknown ξ^0, and ignoring them may not help fully remove biases. Particularly, the sparse estimator $\widetilde{\boldsymbol{\Theta}}$ may result in non-negligible II_j and III_j terms compared to I_j. Consequently, the $\widetilde{\boldsymbol{\Theta}}$-based de-biased lasso estimator (van de Geer et al., 2014) incurs much bias and possesses an unsatisfactory inference performance for GLMs as evidenced by our simulations.

On the other hand, without the matrix sparsity assumption, one may obtain $\boldsymbol{M}$ by solving an optimization problem originally proposed for linear models (Javanmard and Montanari, 2014):

$$\min\left\{\zeta^T\widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}\zeta : \zeta \in \mathbb{R}^{p+1}, \left\|\widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}\zeta - \boldsymbol{e}_j\right\|_\infty \leqslant \mu_n\right\} \tag{5}$$

for $j = 1, \ldots, p+1$ and $\mu_n \geqslant 0$. Under the conditions in Theorem 1 of Section 3, the Hessian matrix $\widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}$ is invertible with probability going to one as $n \to \infty$, and the rows of $\widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}^{-1}$ are solutions to (5) when $\mu_n = 0$. As confirmed by our simulations in a variety of regimes, $\mu_n = 0$ generally performs the best in overall bias correction to $II_j + III_j$ and statistical inference as

μ_n varies from 0 to 1; see Section 4. This motivates us to replace $\boldsymbol{M}$ with $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}^{-1}$, denote by $\widehat{\boldsymbol{\Theta}}_j$ the jth row of $\widehat{\boldsymbol{\Theta}}$, and reexpress (4) as

$$\hat{\xi} - \xi^0 + \left(-\widehat{\boldsymbol{\Theta}} P_n \dot{\rho}_{\hat{\xi}}\right) + \left(-\widehat{\boldsymbol{\Theta}}\boldsymbol{\Delta}\right) = -\widehat{\boldsymbol{\Theta}} P_n \dot{\rho}_{\xi^0}. \tag{6}$$

Therefore, we propose a refined de-biased lasso estimator based on $\widehat{\boldsymbol{\Theta}}$:

$$\hat{\boldsymbol{b}} = \hat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\Theta}} P_n \dot{\rho}_{\hat{\xi}}. \tag{7}$$

We will show that our proposed method possesses desirable asymptotic properties and, in general, performs better than the original de-biased lasso approach (van de Geer et al., 2014) in finite sample settings.

3. Theoretical results

Without loss of generality, we assume that each covariate has been centered to have mean zero. Let $\boldsymbol{X}_{\boldsymbol{\xi}} = \boldsymbol{W}_{\boldsymbol{\xi}}\boldsymbol{X}$ be the weighted design matrix, where $\boldsymbol{W}_{\boldsymbol{\xi}}$ is a diagonal matrix with elements $\omega_i(\boldsymbol{\xi}) = \left\{\ddot{\rho}\left(y_i, \boldsymbol{x}_i^T\boldsymbol{\xi}\right)\right\}^{1/2}$, $i = 1, \ldots, n$. Then, for any $\boldsymbol{\xi} \in \mathbb{R}^{p+1}$, $\widehat{\boldsymbol{\Sigma}}_{\xi}$ can be rewritten as $\widehat{\boldsymbol{\Sigma}}_{\xi} = \boldsymbol{X}_{\xi}^T\boldsymbol{X}_{\xi}/n$. Recall that the population information matrix $\boldsymbol{\Sigma}_{\xi} = E\left(\widehat{\boldsymbol{\Sigma}}_{\xi}\right) = E(P_n\ddot{\rho}_{\xi})$, and its inverse matrix is $\boldsymbol{\Theta}_{\xi} = \boldsymbol{\Sigma}_{\xi}^{-1}$, which are respectively equal to $E(\boldsymbol{X}^T\boldsymbol{X}/n)$ and $\{E(\boldsymbol{X}^T\boldsymbol{X}/n)\}^{-1}$ only for linear models, but not for GLMs. The ψ_2-norm (Vershynin, 2012) is useful for characterizing the convergence rate of $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}^{-1}$. Explicitly, for a random variable Y, its ψ_2-norm is defined as $\| Y \|_{\psi_2} = \sup_{r \geq 1} r^{-1/2}(E|Y|^r)^{1/r}$, and Y is defined to be a sub-Gaussian random variable if $\| Y \|_{\psi_2} < \infty$. For a random vector $\boldsymbol{Z} \in \mathbb{R}^{p+1}$, its ψ_2-norm is defined as $\| \boldsymbol{Z} \|_{\psi_2} = \sup_{\| \boldsymbol{a} \|_2 = 1} \| \langle \boldsymbol{Z}, \boldsymbol{a} \rangle \|_{\psi_2}$, and $\boldsymbol{Z}$ is called sub-Gaussian if $\langle \boldsymbol{Z}, \boldsymbol{a} \rangle$ is a sub-Gaussian random variable for all $\boldsymbol{a} \in \mathbb{R}^{p+1}$ with $\|\boldsymbol{a}\|_2 = 1$ (Vershynin, 2012). We list the regularity conditions as follows.

Assumption 1: The elements in $\boldsymbol{X}$ are bounded almost surely. That is, $\|\boldsymbol{X}\|_{\infty} \leqslant K$ almost surely for a constant $K > 0$. In addition, the rows of $\boldsymbol{X}$ are sub-Gaussian random vectors.

Assumption 2: $\boldsymbol{\Sigma}_{\xi^0}$ is positive definite with bounded eigenvalues such that, for two positive constants $c_{\min}$ and $c_{\max}$, $c_{\min} \leqslant \lambda_{\min}(\boldsymbol{\Sigma}_{\xi^0}) \leqslant \lambda_{\max}(\boldsymbol{\Sigma}_{\xi^0}) \leqslant c_{\max} < \infty$.

Assumption 3: The derivatives $\dot{\rho}(y, a) = \partial\rho(y, a)/\partial a$ and $\ddot{\rho}(y, a) = \partial^2\rho(y, a)/\partial a^2$ exist for all (y, a). Further in some δ-neighborhood, $\delta > 0$, $\ddot{\rho}(y, a)$ is Lipschitz such that for some absolute constant $c_{Lip} > 0$,

$$\max_{a_0 \in \left\{\boldsymbol{x}_i^T\boldsymbol{\xi}^0\right\}} \max\left(|a - a_0|, |\hat{a} - a_0|\right) \leqslant \delta \sup_{y \in \mathcal{Y}} \sup \frac{|\ddot{\rho}(y, a) - \ddot{\rho}(y, \hat{a})|}{|a - \hat{a}|} \leqslant c_{Lip}.$$

And the derivatives are bounded in the sense that there exist two constants K_1, $K_2 > 0$ such that

$$\max_{a_0 \in \{x_i^T \xi^0\}} \sup_{y \in \mathcal{Y}} |\dot{\rho}(y, a_0)| \leqslant K_1,$$

$$\max_{a_0 \in \{x_i^T \xi^0\}} \sup_{|a - a_0| \leqslant \delta} \sup_{y \in \mathcal{Y}} |\ddot{\rho}(y, a)| \leqslant K_2 .$$

Assumption 4: $\|X\xi^0\|_\infty$ is bounded from above almost surely.

Assumption 5: The covariance matrix $E(X^T X/n)$ is positive definite with eigenvalues bounded away from 0 and from above.

It is common to assume bounded covariates as in Assumption 1 and bounded eigenvalues for the information matrix as in Assumption 2 in high-dimensional inference literature (van de Geer et al., 2014; Ning and Liu, 2017). Assumption 2 is needed to derive the rate of convergence for $\hat{\xi}$. Assumption 3 specifies the required smoothness and local properties of the loss function $\rho(y, x^T \xi)$ (van de Geer et al., 2014). Since each element of $X\xi^0$ is the (transformed) conditional mean of y_i, it is reasonable to assume its boundedness in Assumption 4 as in van de Geer et al. (2014) and Ning and Liu (2017) for generalized linear models, and in Kong and Nan (2014) and Fang et al. (2017) for the Cox models. Also Assumption 4 is needed to bound the variance of y_i and keep it away from 0 for generalized linear models. Assumption 5 is a mild requirement for random covariates; a similar condition on the sample covariance matrix can be found in Wang (2011). Unlike van de Geer et al. (2014), we have avoided an assumption on the boundedness of $\| \Theta_{\xi^0} x_i \|_\infty$, which is not verifiable and closely related to the sparsity requirement of Θ_{ξ^0} under Assumption 1.

Let s_0 denote the number of non-zero elements in ξ^0, and consider $\hat{b} = \hat{\xi} - \widehat{\Theta} P_n \dot{\rho}_{\hat{\xi}}$ as defined in (7). Theorem 1 establishes asymptotic normality for (multiple) linear combinations of $\hat{b}$, with a proof provided in Web Appendix A.

Theorem 1: *With*, $\lambda \asymp \{\log(p)/n\}^{1/2}$ *assume that* $p^2/n \to 0$ *and* $s_0 \log(p)(p/n)^{1/2} \to 0$ *as* $n \to \infty$. *Under Assumptions 1–5, we have that* $\widehat{\Sigma}_{\hat{\xi}}$ *is invertible with probability going to one, and that*

i. *for a constant vector* $\alpha_n \in \mathbb{R}^{p+1}$ *with* $\|\alpha_n\|_2 = 1$,

$$\frac{n^{1/2} \alpha_n^T (\hat{b} - \xi^0)}{(\alpha_n^T \widehat{\Theta} \alpha_n)^{1/2}} \xrightarrow{\mathcal{D}} N(0, 1) \text{ as } n \to \infty;$$

ii. *for a fixed integer* $m > 1$ *and a constant matrix* $A_n \in \mathbb{R}^{m \times (p+1)}$ *satisfying* $\| A_n^T \| \leqslant c_*$ *for some constant* c_* *and* $A_n \Theta_{\xi^0} A_n^T \to F$ *for some* $F \in \mathbb{R}^{m \times m}$,

$$n^{1/2} A_n (\hat{b} - \xi^0) \xrightarrow{\mathcal{D}} N_m(0, F) \text{ as } n \to \infty .$$

Remark 1: Theorem 1 enables us to construct a 100×(1−r)% confidence interval for $\boldsymbol{\alpha}_n^T\boldsymbol{\xi}^0$ as $\left[\boldsymbol{\alpha}_n^T\hat{\boldsymbol{b}} - z_{r/2}\left(\boldsymbol{\alpha}_n^T\widehat{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n/n\right)^{1/2}, \boldsymbol{\alpha}_n^T\hat{\boldsymbol{b}} + z_{r/2}\left(\boldsymbol{\alpha}_n^T\widehat{\boldsymbol{\Theta}}\boldsymbol{\alpha}_n/n\right)^{1/2}\right]$, where $0 < r < 1$ and $z_{r/2}$ is the upper (r/2)th quantile of the standard normal distribution. Here, $\boldsymbol{\alpha}_n$ can be arbitrarily dense, instead of having only a few non-zero elements such as $\boldsymbol{\alpha}_n = \boldsymbol{e}_j$ in van de Geer et al. (2014). A 100×(1−r)% confidence region for $\boldsymbol{A}_n\boldsymbol{\xi}^0$ can be constructed as $\left\{\boldsymbol{a} \in \mathbb{R}^m : n\left(\boldsymbol{A}_n\hat{\boldsymbol{b}} - \boldsymbol{a}\right)^T\left(\boldsymbol{A}_n\widehat{\boldsymbol{\Theta}}\boldsymbol{A}_n^T\right)^{-1}\left(\boldsymbol{A}_n\hat{\boldsymbol{b}} - \boldsymbol{a}\right) \leqslant \chi^2_{m,r}\right\}$, where $\chi^2_{m,r}$ is the upper rth quantile of χ^2_m.

Remark 2: In a linear regression setting with $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, some algebra shows that the proposed estimator (7) is identical to the MLE, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, regardless of the choice of the initial estimate, $\hat{\boldsymbol{\xi}}$. Therefore, as a by-product, Theorem 1 characterizes the asymptotics of the MLE for linear models with a diverging number of coefficients, which only requires $p^2/n \to 0$. This can be shown following a similar proof of Theorem 1 with $\boldsymbol{\Delta} = 0$ for linear regression models, where $\widehat{\boldsymbol{\Theta}}$ is free of regression parameters. It is obvious that regularity conditions can be simplified for linear regression models.

Remark 3: Binary covariates, particularly dummy variables for categorical covariates, satisfy the assumptions for Theorem 1. Therefore, applications of Theorem 1 encompass inference for categorical covariates, such as drawing inference on comparisons between multiple intervention groups or testing associations of multi-level categorical covariates with outcomes.

4. Numerical experiments

Under the "large n, diverging p" scenario, we compare the estimation biases and coverage probabilities of confidence intervals across the following estimators: (i) the original de-biased lasso estimator obtained by using the node-wise lasso estimator $\widetilde{\boldsymbol{\Theta}}$ in van de Geer et al. (2014) (ORIG-DS), (ii) the conventional maximum likelihood estimator (MLE), and (iii) our proposed refined de-biased lasso estimator $\hat{\boldsymbol{b}}$, based on the inverse matrix estimation $\widehat{\boldsymbol{\Theta}} = \widehat{\boldsymbol{\Sigma}}_{\hat{\xi}}^{-1}$ (REF-DS).

Simulations using the logistic and Poisson regression models yield similar observations, and we only report results from logistic regression. A total of n = 1,000 observations, each with p = 40, 100, 300, 400 covariates, are simulated. Within $\boldsymbol{x}_i = \left(1, \tilde{\boldsymbol{x}}_i^T\right)^T$, $\tilde{\boldsymbol{x}}_i$ are independently generated from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ before being truncated at ±6, and $y_i \mid \boldsymbol{x}_i \sim Bernoulli(\mu_i)$, where $\mu_i = \exp\left(\boldsymbol{x}_i^T\boldsymbol{\xi}^0\right)/\left\{1 + \exp\left(\boldsymbol{x}_i^T\boldsymbol{\xi}^0\right)\right\}$. The intercept $\beta_0^0 = 0$, and β_1^0 varies from 0 to 1.5 with 40 equally spaced increments. In addition, four arbitrarily chosen elements of $\boldsymbol{\beta}^0$ take non-zero values, two with 0.5 and the other two with 1, and are fixed throughout the simulation. In some settings, the maximum likelihood estimates do not exist due to divergence and are not shown. The covariance matrix $\boldsymbol{\Sigma}_X$ of $\tilde{\boldsymbol{x}}_i$ takes an autoregressive structure of order 1, i.e. AR(1), with correlation ρ = 0.7, or a compound symmetry structure with correlation ρ = 0.7. The tuning parameter in the ℓ_1 penalized regression is selected by 10-fold cross-validation, and the tuning parameter for the node-wise lasso estimator $\widetilde{\boldsymbol{\Theta}}$ is selected using 5-fold cross-validation. Both tuning parameter selection procedures are implemented using glmnet (Friedman et al., 2010). For every β_1^0 value, we summarize the average bias, empirical

coverage probability, empirical standard error and model-based estimated standard error over 200 replications.

Figure 1 illustrates the simulation results for estimating β_1^0 under the autoregressive covariance structure, and Figure 2 under the compound symmetry structure. The three methods in comparison behave similarly with only 40 covariates included in the model, and the MLE yields slightly larger biases. The MLE estimates display much more biases than those obtained by the other two methods with 100 covariates, and do not always exist due to divergence. When the MLE estimates do exist, they manifest more variability than the original and refined de-biased lasso estimates, and are with lower coverage probabilities. In contrast, our refined de-biased lasso approach outperforms the MLE because the former utilizes sparse lasso estimates as the initial estimates and is numerically more stable than the latter.

There are systematic biases in the original de-biased lasso estimator by van de Geer et al. (2014), which increase with the magnitude of β_1^0. When signals are non-zero, the model-based standard errors produced by van de Geer et al. (2014) slightly underestimate the true variability. These factors contribute to the poor coverage probabilities of van de Geer et al. (2014) when the signal size is not zero. In contrast, the refined de-biased lasso estimator gives the smallest biases and has an empirical coverage probability closest to the nominal level across different settings, though with slightly higher variability than van de Geer et al. (2014). This is likely because our proposed de-biased lasso approach does not utilize a penalized estimator of the inverse information matrix. We take note that as the refined de-biased lasso method needs to invert the Hessian matrix, which could become more ill-conditioned if the dimension increases, its performance may deteriorate as the dimension of covariates increases.

As we alluded to in Section 2, the refined de-biased lasso estimator is related to Javanmard and Montanari (2014), and we have conducted additional simulations to compare them, referred to as “REF-DS” and “Tuning” respectively. Figure 3, which depicts the results of a logistic regression model with $n = 500$ observations and $p = 40, 100, 200, 300, 400$ covariates, shows that $\mu_n = 0$ generally performs the best in bias correction and honest confidence interval coverage when μ_n varies from 0 to 1; see the simulation setup and additional results in Web Appendix B.

5. Boston lung cancer data analysis

Lung cancer is the top cause of cancer death in the United States. The Boston Lung Cancer Survival Cohort (BLCSC), one of the largest hospital-based cohorts in the country, investigates the molecular causes of lung cancer. Recruited to the study were the lung cancer cases and controls from the Massachusetts General Hospital and the Dana-Farber Cancer Institute since 1992 (Miller et al., 2002). We apply the proposed refined de-biased lasso approach, together with the method by van de Geer et al. (2014) and the MLE for comparison, to a subset of the BLCSC data and examine the joint effects of SNPs from nine target genes on the overall risk of lung cancer.

Genotypes from Axiom array and clinical information were originally available on 1,459 individuals. Out of those individuals, 14 (0.96%) had missing smoking status, 8 (0.55%) had missing race information, and 1,386 (95%) were Caucasian. We include a final number of $n = 1{,}374$ Caucasians with complete data, where $n_0 = 723$ were controls and $n_1 = 651$ were cases. Denote the binary disease outcome by $y_i = 1$ for cases and 0 for controls. Among the 1,077 smokers, 595 had lung cancer, whereas out of the 297 non-smokers, 56 were cases. Other demographic characteristics, such as education level, gender and age, are summarized in Web Appendix C. Using the target gene approach, we focus on the following lung cancer-related genes: *AK5* on region 1p31.1, *RNASET2* on region 6q27, *CHRNA2* and *EPHX2* on region 8p21.2, *BRCA2* on region 13q13.1, *SEMA6D* and *SECISBP2L* on region 15q21.1, *CHRNA5* on region 15q25.1, and *CYP2A6* on region 19q13.2. These genes may harbor SNPs associated with the overall lung cancer risks (McKay et al., 2017). In our dataset, each SNP is coded as 0,1,2, reflecting the number of copies of the minor allele, and minor alleles are assumed to have additive effects. After applying filters on the minor allele frequency, genotype call rate, and excluding highly correlated SNPs, 103 SNPs remain in the model. Since smoking may modify associations between lung cancer risks and SNPs, for example, those residing in region 15q25.1 (Gabrielsen et al., 2013; Amos et al., 2008), we conduct analysis stratified by smoking status. Among the smokers and non-smokers, we fit separate logistic regression models, adjusting for education, gender and age.

We apply these methods to draw inference on all of the 107 predictors, two of which are dummy variables for education originally with three levels, no high school, high school and at least 1–2 years of college. Our data analysis may shed light on the molecular mechanism underlying lung cancer. Due to limited space, Table 1 lists the estimates for 11 selected SNPs and demographic variables among smokers, and Table 2 for non-smokers. These SNPs are listed as they are significant based on at least one of the three methods among either the smokers or the non-smokers. Details of the other SNPs are omitted. Since the number of the non-smokers is only about one third of the smokers, the MLE has the largest standard errors and tends to break down among the non-smokers; see, for example, AX-62479186 in Table 2, whereas the two de-biased lasso methods give more stable estimates. The estimates by our proposed refined de-biased lasso method (REF-DS) and the method by van de Geer et al. (2014) (ORIG-DS) share more similarities in the smokers in Table 1 than in the non-smokers in Table 2. Overall, the method by van de Geer et al. (2014) has slightly narrower confidence intervals than our proposed de-biased lasso estimator due to the penalized estimation for $\boldsymbol{\Theta}_{\xi^0}$. These results generally agree with our simulation studies.

For some SNPs, our proposed method and the method by van de Geer et al. (2014) yield estimates with opposite directions; see AX-38419741 and AX-15934253 in Table 1, and AX-42391645 in Table 2. Among the non-smokers, the 95% confidence interval for AX-31620127 in *SEMA6D* by our proposed method is all positive and excludes 0, while the confidence interval by the method of van de Geer et al. (2014) includes 0; the directions for AX-88907114 in *CYP2A6* are the opposite in Table 2. *CHRNA5* is a gene known for predisposition to nicotine dependence (Halldén et al., 2016; Amos et al., 2008; Gabrielsen et al., 2013). Though AX-39952685 and AX-88891100 in *CHRNA5* are not significant at level 0.05 in marginal analysis among the smokers, their 95% confidence intervals in

Table 1 exclude 0 by all of the three methods. Indeed, AX-88891100, or rs503464, mapped to the same physical location in the dbSNP database, was found to "decrease *CHRNA5* promoter-derived luciferase activity" (Doyle et al., 2011). The same SNP was also reported to be significantly associated with nicotine dependence at baseline, as well as response to varenicline, bupropion, nicotine replacement therapy for smoking cessation (Pintarelli et al., 2017). AX-39952685 was found to be strongly correlated with SNP AX-39952697 in *CHRNA5*, which was mapped to the same physical location as rs11633585 in dbSNP. All of these markers were found to be significantly associated with nicotine dependence (Stevens et al., 2008). The stratified analysis also suggests molecular mechanisms of lung cancer differ between smokers and non-smokers, though additional confirmatory studies are needed.

6. Concluding remarks

We have proposed a refined de-biased lasso estimating method for GLMs by directly inverting Hessian matrices in the "large n, diverging p" framework. We have showed that if $p^2/n = o(1)$ and $(p/n)^{1/2} s_0 \log(p) = o(1)$, along with some other mild conditions, any linear combinations of the resulting estimates are asymptotically normal and can be used for constructing hypothesis tests and confidence intervals. By way of empirical studies, we have showed that when p is small relative to n, the proposed refined de-biased lasso yields estimates nearly identical to the MLE and the original de-biased lasso by van de Geer et al. (2014). In contrast, the proposed method outperforms the latter two in bias correction and confidence interval coverage probabilities when $p < n$ but p is relatively large, indicating a broad applicability.

Additional simulations for linear regression models (results not shown here) indicate that, however, both our proposed method (equivalent to the MLE, see Remark 2) and the original de-biased lasso method perform well with no obvious difference between these two methods for wide ranges of p/n. This is likely due to the fact that the Hessian matrix for a linear model is free of regression parameters.

Theorem 1 gives some sufficient range of p relative to n to guide practical settings, but does not necessarily exhaust all possible working scenarios in a finite sample setting. In fact, we have shown through simulations that the asymptotic approximations given in Theorem 1 work well in finite sample settings with wide ranges of p and n. Nevertheless, searching for more relaxed conditions of p and n warrants more in-depth investigations.

With a slightly stronger requirement of $s_0 \log(p)/(p/n)^{1/2} \to 0$ than $s_0 \log(p)/\sqrt{n} \to 0$ specified in van de Geer et al. (2014), Theorem 1 obtains stronger results than theirs in i) drawing inference for any linear combinations of regression coefficients, ii) releasing sparsity assumptions on $\|\boldsymbol{\Theta}_{\xi^0} \boldsymbol{x}_i\|_\infty$, and iii) dropping the boundedness assumption on $\boldsymbol{\Theta}_{\xi^0}$; see Web Appendix D for detailed discussion. Moreover, a referee pointed out a recent work on linear regression models (Bellec et al., 2018) that may help provide slightly less stringent sparsity conditions by relaxing the logarithmic factor; however, such generalization to GLMs is beyond our scope.

Lastly, we comment on the difficulties of applying some existing methods to draw inference with high-dimensional GLMs. With extensive simulations, we have discovered unsatisfactory bias correction and confidence interval coverage with the original de-biased lasso in GLM settings (van de Geer et al., 2014); for example, see the simulation results under the "large p, small n" scenario in Web Appendix B. Our further investigation pinpoints an essential assumption that hardly holds for GLMs in general, which is that the number of non-zero elements in the rows of the high-dimensional inverse information matrix $\boldsymbol{\Theta}_{\xi^0}$ is sparse and of order $o[\{n/\log(p)\}^{1/2}]$ (van de Geer et al., 2014). The theoretical developments in van de Geer et al. (2014) rely heavily on this sparse matrix assumption. The ℓ_0 sparsity conditions on high-dimensional matrices are not uncommon in the literature of high-dimensional inference. A related ℓ_0 sparsity condition on $\boldsymbol{w}^* = \boldsymbol{I}^{*-1}_{\gamma\gamma}\boldsymbol{I}^*_{\gamma\theta}$ can be found in Ning and Liu (2017), where $\boldsymbol{I}^*$ is the information matrix under the truth, but it is not well justified in a general setting for GLMs. When testing a global null hypothesis $\boldsymbol{\beta}^0 = 0$, the sparsity of $\boldsymbol{\Theta}_{\xi^0}$ reduces to the sparsity of the covariate precision matrix, which becomes less of an issue (Ma et al., 2020). Therefore, we generally do not recommend the de-biased lasso method when $p > n$ for GLMs.

Data Availability Statement

The Boston Lung Cancer Survival Cohort data are not publicly available due to access restrictions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was partially supported by the United States National Institutes of Health (R01AG056764, R01CA249096 and U01 CA209414), and the National Science Foundation (DMS-1915711). The authors thank Dr. David Christiani for providing the Boston Lung Cancer Survival Cohort data, and the associate editor and two referees for valuable comments.

References

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1. Nature Genetics 40, 616–622. [PubMed: 18385676]

Bellec PC, Lecué G, and Tsybakov AB (2018). Slope meets lasso: improved oracle bounds and optimality. Annals of Statistics 46, 3603–3642.

Bossé Y and Amos CI (2018). A decade of GWAS results in lung cancer. Cancer Epidemiology, Biomarkers & Prevention 27, 363–379.

Bühlmann P and van de Geer S (2011). Statistics for high-dimensional data: Methods, theory and applications. Springer.

Candès E and Tao T (2007). The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35, 2313–2351.

Doyle GA, Wang M-J, Chou AD, Oleynick JU, Arnold SE, Buono RJ, et al. (2011). *In Vitro* and *Ex Vivo* analysis of *CHRNA3* and *CHRNA5* haplotype expression. PloS One 6, e23373. [PubMed: 21858091]

Evans WE and Relling MV (2004). Moving towards individualized medicine with pharmacogenomics. Nature 429, 464–468. [PubMed: 15164072]

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Fan J and Peng H (2004). Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics 32, 928–961.

Fang EX, Ning Y, and Liu H (2017). Testing and confidence intervals for high dimensional proportional hazards models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79, 1415–1437.

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22. [PubMed: 20808728]

Gabrielsen ME, Romundstad P, Langhammer A, Krokan HE, and Skorpen F (2013). Association between a 15q25 gene variant, nicotine-related habits, lung cancer and COPD among 56307 individuals from the HUNT study in Norway. European Journal of Human Genetics 21, 1293–1299. [PubMed: 23443019]

Guan Y and Stephens M (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5, 1780–1815.

Halldén S, Sjögren M, Hedblad B, Engström G, Hamrefors V, Manjer J, et al. (2016). Gene variance in the nicotinic receptor cluster (CHRNA5-CHRNA3-CHRNB4) predicts death from cardiopulmonary disease and cancer in smokers. Journal of Internal Medicine 279, 388–398. [PubMed: 26689306]

He Q and Lin D-Y (2010). A variable selection method for genome-wide association studies. Bioinformatics 27, 1–8. [PubMed: 21036813]

He X and Shao Q-M (2000). On parameters of increasing dimensions. Journal of Multivariate Analysis 73, 120–135.

Huber PJ (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. The Annals of Statistics 1, 799–821.

Javanmard A and Montanari A (2014). Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15, 2869–2909.

Kong S and Nan B (2014). Non-asymptotic oracle inequalities for the high-dimensional Cox regression via Lasso. Statistica Sinica 24, 25–42. [PubMed: 24516328]

Lee JD, Sun DL, Sun Y, and Taylor JE (2016). Exact post-selection inference, with application to the lasso. The Annals of Statistics 44, 907–927.

Ma R, Cai TT, and Li H (2020). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. Journal of the American Statistical Association 00, 1–15.

McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature Genetics 49, 1126–1132. [PubMed: 28604730]

Meinshausen N and Bühlmann P (2006). High-dimensional graphs and variable selection with the lasso. The Annals of Statistics 34, 1436–1462.

Miller DP, Liu G, De Vivo I, Lynch TJ, Wain JC, Su L, et al. (2002). Combinations of the variant genotypes of GSTP1, GSTM1, and p53 are associated with an increased lung cancer risk. Cancer Research 62, 2819–2823. [PubMed: 12019159]

Ning Y and Liu H (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. The Annals of Statistics 45, 158–195.

Pintarelli G, Galvan A, Pozzi P, Noci S, Pasetti G, Sala F, et al. (2017). Pharmacogenetic study of seven polymorphisms in three nicotinic acetylcholine receptor subunits in smokingcessation therapies. Scientific Reports 7, 16730. [PubMed: 29196725]

Portnoy S (1984). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large, i. consistency. The Annals of Statistics 12, 1298–1309.

Portnoy S (1985). Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large, ii. normal approximation. The Annals of Statistics 13, 1403–1417.

Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, et al. (2010). Genome-wide association study identifies five loci associated with lung function. Nature Genetics 42, 36–44. [PubMed: 20010834]

Schaid DJ, Chen W, and Larson NB (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics 19, 491–504.

Stevens VL, Bierut LJ, Talbot JT, Wang JC, Sun J, Hinrichs AL, et al. (2008). Nicotinic receptor gene variants influence susceptibility to heavy smoking. Cancer Epidemiology, Biomarkers & Prevention 17, 3517–3525.

Sur P and Candès EJ (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences of the United States of America 116, 14516–14525. [PubMed: 31262828]

Taylor JG, Choi E-H, Foster CB, and Chanock SJ (2001). Using genetic variation to study human disease. Trends in Molecular Medicine 7, 507–512. [PubMed: 11689336]

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58, 267–288.

van de Geer S, Bühlmann P, Ritov Y, and Dezeure R (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42, 1166–1202.

van de Geer SA (2008). High-dimensional generalized linear models and the lasso. The Annals of Statistics 36, 614–45.

Vershynin R (2012). Introduction to the non-asymptotic analysis of random matrices. In Compressed Sensing, pages 210–268. Cambridge University Press, Cambridge.

Wang L (2011). GEE analysis of clustered binary data with diverging number of covariates. The Annals of Statistics 39, 389–417.

Yohai VJ and Maronna RA (1979). Asymptotic behavior of m-estimators for the linear model. The Annals of Statistics 7, 258–268.

Zhang C-H and Zhang SS (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76, 217–42.

Zhang X and Cheng G (2017). Simultaneous inference for high-dimensional linear models. Journal of the American Statistical Association 112, 757–768.

Zou H (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

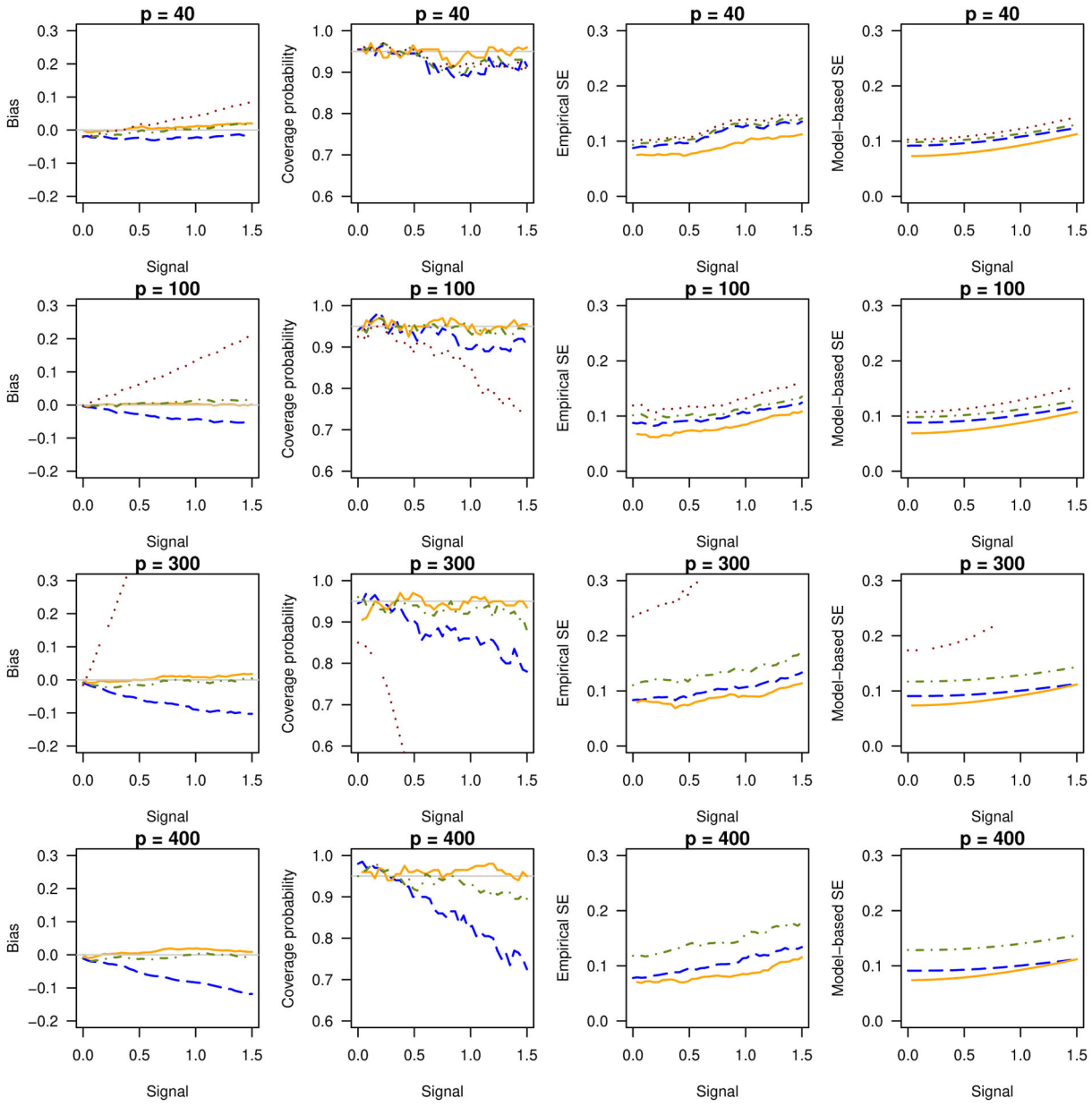

Figure 1.

Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in logistic regression. Covariates are simulated from $N_p(\mathbf{0}, \mathbf{\Sigma}_X)$ before being truncated at ±6, where $\mathbf{\Sigma}_X$ has an AR(1) with $\rho = 0.7$. The sample size is $n = 1{,}000$ and the number of covariates $p = 40, 100, 300, 400$. The oracle estimator, that is, the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparison include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

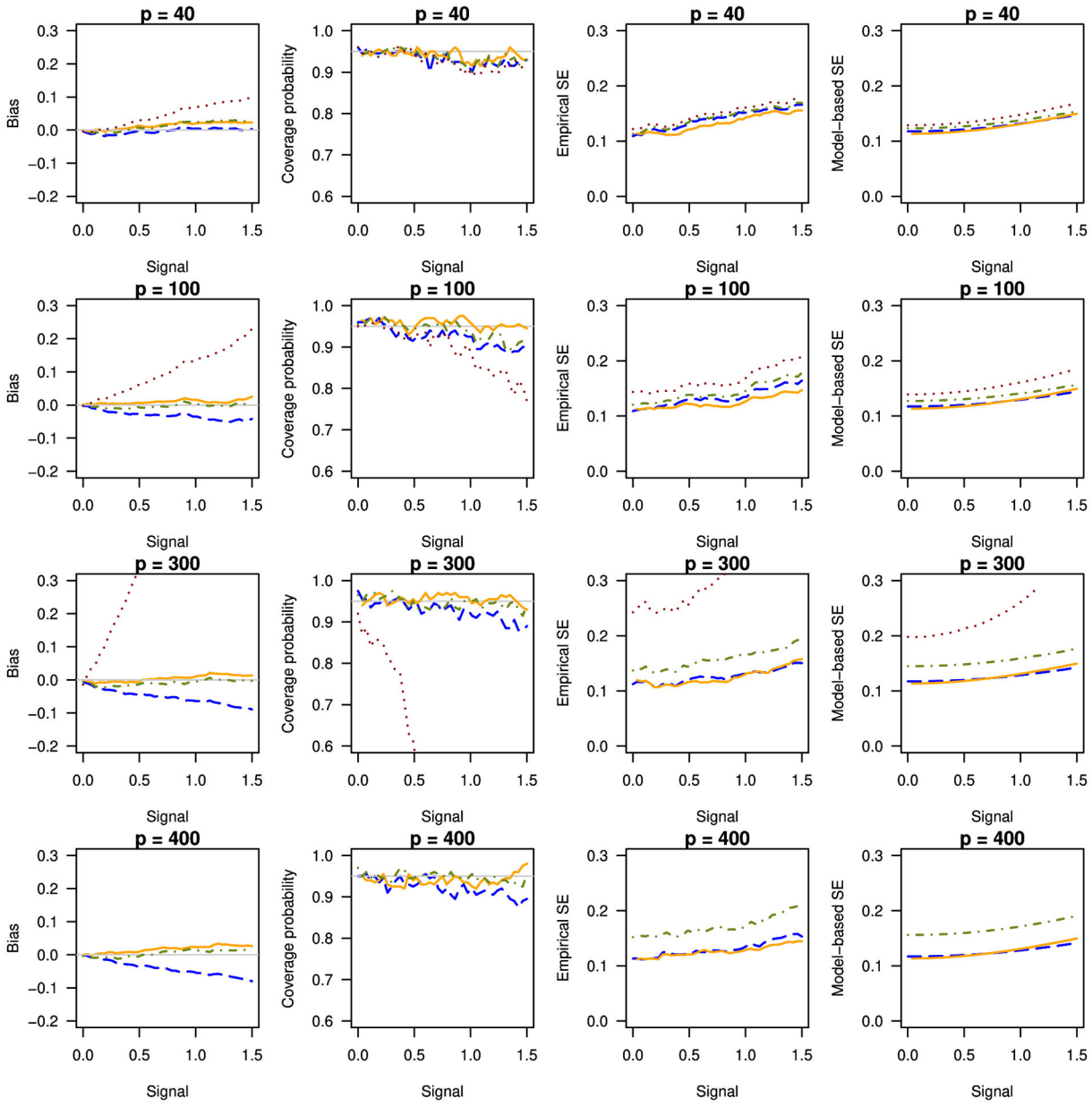

Figure 2.

Simulation results: Bias, coverage probability, empirical standard error, and model-based standard error for β_1^0 in a logistic regression. Covariates are simulated from $N_p(\mathbf{0}, \mathbf{\Sigma}_X)$ before being truncated at ±6, where $\mathbf{\Sigma}_X$ has a compound symmetry structure with $\rho = 0.7$. The sample size is $n = 1{,}000$ and the number of covariates $p = 40, 100, 300, 400$. The oracle estimator, that is the maximum likelihood estimator under the true model, is plotted as a reference in orange solid lines. The methods in comparisons include our proposed refined de-biased lasso in olive dot-dash lines, the original de-biased lasso by van de Geer et al. (2014) in blue dashed lines, and the maximum likelihood estimation in red dotted lines.

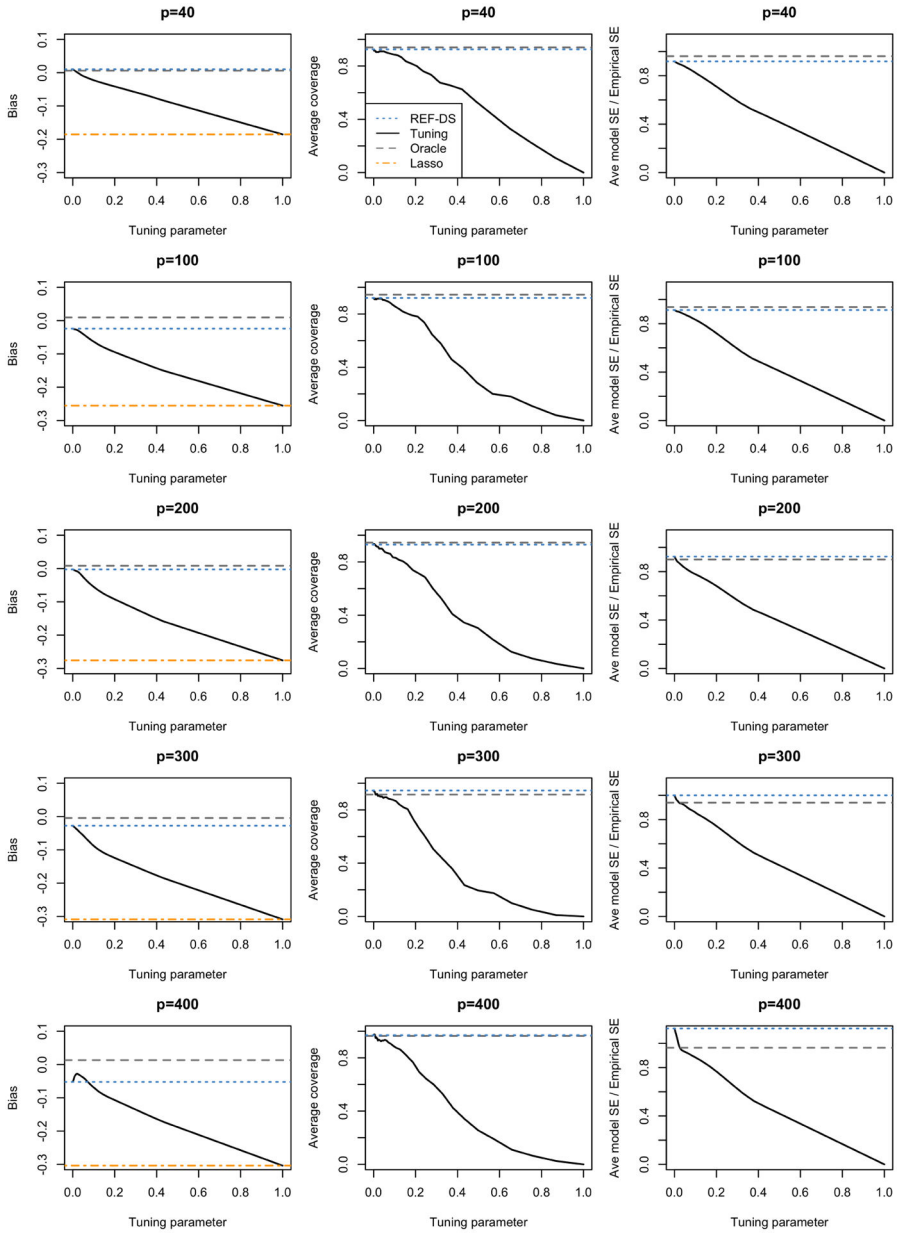

Figure 3.

Simulation results: Bias, coverage probability, ratio between average model-based standard error and empirical standard error in a logistic regression to verify the selection of the tuning parameter $\mu_n = 0$ in Eq. (5) for $\xi_j^0 = 1$.

Table 1

Estimated coefficients for demographic variables and eleven SNPs in a logistic regression model among the smokers. The other SNPs are omitted from the table

			REF-DS [a]			***ORIG-DS*** [b]			***MLE*** [c]		
	Demographic variable		**Est**[d]	**SE**[e]	**95% CI**[f]	**Est**	**SE**	**95% CI**	**Est**	**SE**	**95% CI**
	Education: No high school		0.44	0.20	(0.05, 0.83)	0.48	0.19	(0.11, 0.85)	0.52	0.22	(0.09, 0.94)
	Education: High school graduate		0.00	0.15	(−0.29, 0.28)	−0.02	0.14	(−0.29, 0.25)	−0.01	0.16	(−0.32, 0.30)
	Gender: Male		−0.14	0.13	(−0.40, 0.12)	−0.16	0.13	(−0.40, 0.09)	−0.15	0.14	(−0.43,0.13)
	Age in years		0.04	0.07	(−0.09, 0.17)	0.05	0.06	(−0.07, 0.17)	0.05	0.07	(−0.09,0.19)
SNP	Pos[g]	Gene	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
AX-15319183	6:167352075	*RNASET2*	0.01	0.19	(−0.36, 0.39)	−0.03	0.18	(−0.39, 0.33)	0.02	0.20	(−0.38, 0.42)
AX-41911849	6:167360724	*RNASET2*	0.43	0.22	(0.00, 0.86)	0.44	0.20	(0.06, 0.83)	0.49	0.24	(0.03, 0.96)
AX-42391645	8:27319769	*CHRNA2*	0.01	0.16	(−0.29, 0.32)	−0.01	0.14	(−0.28, 0.26)	0.01	0.16	(−0.31, 0.34)
AX-38419741	8:27319847	*CHRNA2*	**0.11**	0.35	(−0.59, 0.80)	**−0.14**	0.31	(−0.75, 0.48)	0.13	0.37	(−0.60, 0.86)
AX-15934253	8:27334098	*CHRNA2*	**−0.15**	0.44	(−1.02, 0.71)	**0.06**	0.39	(−0.70, 0.82)	−0.19	0.47	(−1.12, 0.74)
AX-12672764	13:32927894	*BRCA2*	−0.07	0.19	(−0.44, 0.31)	−0.10	0.16	(−0.40, 0.21)	−0.07	0.20	(−0.47, 0.33)
AX-31620127	15:48016563	*SEMA6D*	0.79	0.26	(0.28, 1.31)	0.79	0.25	(0.30, 1.28)	0.96	0.30	(0.37, 1.55)
AX-88891100	15:78857896	*CHRNA5*	0.87	0.36	(0.17, 1.57)	0.79	0.32	(0.16, 1.41)	0.98	0.39	(0.22, 1.74)
AX-39952685	15:78867042	*CHRNA5*	0.99	0.47	(0.07, 1.91)	0.82	0.38	(0.09, 1.56)	1.11	0.50	(0.13, 2.08)
AX-62479186	15:78878565	*CHRNA5*	0.41	0.41	(−0.40, 1.22)	0.46	0.39	(−0.31, 1.23)	0.46	0.45	(−0.41, 1.33)
AX-88907114	19:41353727	*CYP2A6*	0.52	0.34	(−0.16, 1.19)	0.49	0.33	(−0.15, 1.13)	0.58	0.38	(−0.15, 1.32)
⋮											

[a] The proposed refined de-biased lasso based on the inverted Hessian matrix

[b] The original de-biased lasso based on the node-wise lasso estimator for $\boldsymbol{\Theta}_{\hat{\varphi}^0}$ by van de Geer et al. (2014)

[c] The maximum likelihood estimation approach

[d] The point estimate for each coefficient

[e] The model-based standard error

[f] Confidence interval

[g] The physical position of a SNP on a chromosome based on Assembly GRCh37/hg19

Table 2

Estimated coefficients for demographic variables and eleven SNPs in a logistic regression model among the non-smokers. The other SNPs are omitted from the table

			REF-DS [a]			***ORIG-DS*** [b]			***MLE*** [c]		
	Demographic variable		**Est**[d]	**SE**[e]	**95% CI**[f]	**Est**	**SE**	**95% CI**	**Est**	**SE**	**95% CI**
	Education: No high school		−0.84	0.93	(−2.67, 0.99)	−0.58	0.78	(−2.11, 0.95)	−10.53	3.82	(−18.01, −3.04)
	Education: High school graduate		−1.68	0.52	(−2.69, −0.66)	−1.56	0.43	(−2.39, −0.72)	−11.23	3.75	(−18.58, −3.88)
	Gender: Male		−0.30	0.41	(−1.10, 0.51)	−0.16	0.32	(−0.78, 0.46)	−1.94	1.03	(−3.96, 0.09)
	Age in years		−0.52	0.20	(−0.91, −0.13)	−0.56	0.16	(−0.87, −0.26)	−2.59	0.98	(−4.51, −0.67)
SNP	Pos[g]	Gene	Est	SE	95% CI	Est	SE	95% CI	Est	SE	95% CI
AX-15319183	6:167352075	*RNASET2*	−0.71	0.55	(−1.78, 0.36)	0.01	0.40	(−0.79, 0.80)	−4.32	1.84	(−7.92, −0.71)
AX-41911849	6:167360724	*RNASET2*	0.69	0.65	(−0.59, 1.97)	0.37	0.47	(−0.55, 1.29)	4.46	2.00	(0.54, 8.39)
AX-42391645	8:27319769	*CHRNA2*	**−0.11**	0.49	(−1.07, 0.85)	**0.18**	0.30	(−0.41, 0.78)	−1.90	2.00	(−5.81, 2.02)
AX-38419741	8:27319847	*CHRNA2*	0.50	1.04	(−1.54, 2.53)	0.23	0.61	(−0.97, 1.42)	3.37	3.00	(−2.51, 9.26)
AX-15934253	8:27334098	*CHRNA2*	0.11	1.40	(−2.64, 2.86)	0.38	0.82	(−1.23, 1.98)	5.37	4.21	(−2.88, 13.62)
AX-12672764	13:32927894	*BRCA2*	−0.83	0.62	(−2.04, 0.37)	−0.57	0.38	(−1.32, 0.18)	−8.25	2.64	(−13.42, −3.08)
AX-31620127	15:48016563	*SEMA6D*	1.77	0.75	**(0.30, 3.24)**	0.43	0.46	**(−0.48, 1.34)**	9.23	3.27	(2.81, 15.64)
AX-88891100	15:78857896	*CHRNA5*	0.78	1.18	(−1.54, 3.10)	1.15	0.87	(−0.56, 2.85)	1.54	3.17	(−4.68, 7.75)
AX-39952685	15:78867042	*CHRNA5*	−0.54	1.30	(−3.09, 2.01)	−0.99	0.73	(−2.41, 0.44)	−2.85	3.98	(−10.65, 4.96)
AX-62479186	15:78878565	*CHRNA5*	−1.28	1.34	(−3.92, 1.35)	−1.33	1.10	(−3.49, 0.82)	**−19.64**	**3410.98**	(−6705.04, 6665.75)
AX-88907114	19:41353727	*CYP2A6*	0.86	0.88	**(−0.86, 2.59)**	1.40	0.68	**(0.06, 2.74)**	3.52	2.18	(−0.75, 7.78)
⋮											

[a] The proposed refined de-biased lasso based on the inverted Hessian matrix

[b] The original de-biased lasso based on the node-wise lasso estimator for $\boldsymbol{\Theta}_{\hat{\varphi}^0}$ by van de Geer et al. (2014)

[c] The maximum likelihood estimation approach

[d] The point estimate for each coefficient

[e] The model-based standard error

[f] Confidence interval

[g] The physical position of a SNP on a chromosome based on Assembly GRCh37/hg19