Confidence regions for high-dimensional generalized linear models under sparsity

Jana Janková    Sara van de Geer

October 6, 2016

Abstract

We study asymptotically normal estimation and confidence regions for low-dimensional parameters in high-dimensional sparse models. Our approach is based on the \( \ell_1 \)-penalized M-estimator which is used for construction of a bias corrected estimator. We show that the proposed estimator is asymptotically normal, under a sparsity assumption on the high-dimensional parameter, smoothness conditions on the expected loss and an entropy condition. This leads to uniformly valid confidence regions and hypothesis testing for low-dimensional parameters. The present approach is different in that it allows for treatment of loss functions that are not sufficiently differentiable, such as quantile loss, Huber loss or hinge loss functions. We also provide new results for estimation of the inverse Fisher information matrix, which is necessary for the construction of the proposed estimator. We formulate our results for general models under high-level conditions, but investigate these conditions in detail for generalized linear models and provide mild sufficient conditions. As particular examples, we investigate the case of quantile loss and Huber loss in linear regression and demonstrate the performance of the estimators in a simulation study and on real datasets from genome-wide association studies. We further investigate the case of logistic regression and illustrate the performance of the estimator on simulated and real data.

Keywords: Sparsity; High-dimensional; Lasso; Entropy; Generalized linear model; Inverse covariance matrix

1 Introduction

The need to develop efficient methodology for handling high-dimensional data arises in a variety of applications including genome-wide studies, image processing and pattern recognition. Penalized M-estimators have become a popular tool for point estimation in such high-dimensional settings. Our goal in this paper however goes beyond point estimation: we aim to construct and study methodology for quantifying the uncertainty of estimation.

Suppose that we observe a sample \( X_1, \ldots, X_n \in \mathcal{X} \) of independent observations from an unknown distribution \( P \) which is known to belong to a class \( \mathcal{P} = \{ P_\beta \} \) where \( \beta \) ranges over a subset of \( \mathbb{R}^p \). We assume that the setting is high-dimensional: the number of unknown parameters \( p \) may be greater than the sample size. We denote by \( P \) the mean with respect to \( P \) (assuming it exists) and by \( P_n \) the empirical mean given the sample \( X_1, \ldots, X_n \). Consider a given loss function \( \rho_\beta : \mathcal{X} \to \mathbb{R} \) and let the true unknown parameter \( \beta_0 \) be defined as

\[
\beta_0 := \arg \min_{\beta \in \mathbb{R}^p} P \rho_\beta.
\]

To estimate \( \beta_0 \), we consider \( \ell_1 \)-penalized M-estimators defined by

\[
\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} P_n \rho_\beta + \lambda \| \beta \|_1.
\]

Under restrictions on the number of non-zero coefficients in \( \beta_0 \) and several technical assumptions (compatibility condition, margin condition), consistency of \( \ell_1 \)-penalized M-estimators can
be achieved. For an overview of the theoretical results, we refer the reader to Bühlmann and van de Geer (2011) and the references therein. The so-called “oracle inequalities” show consistency in $\ell_1$-norm and consistency of the excess risk at a near-oracle rate $s\lambda$ and $s\lambda^2$, respectively, where $s$ is the number of non-zero elements of $\beta_0$. More precisely, if certain regularity conditions are satisfied (see Bühlmann and van de Geer (2011)), it holds

$$
\|\hat{\beta} - \beta_0\|_1 = O_P(s\lambda^2), \quad P(\rho_{\hat{\beta}} - \rho_{\beta_0}) = O_P(s\lambda^2),
$$

where $O_P(1)$ means boundedness in probability. Under somewhat stronger regularity conditions, consistency in $\ell_2$-norm at the near-oracle rate may be obtained.

The advantage of $\ell_1$-penalized estimators is that due to the geometry of the $\ell_1$-norm, the estimator $\hat{\beta}$ may have many coefficients set exactly to zero. In this sense, Lasso methods yield “variable selection”; however, this holds only if certain restrictive conditions are satisfied. We make the statement about variable selection more precise for the case of linear regression. To this end, we denote the true non-zero set of $\beta_0$ by $S := \{i : \beta_0^i \neq 0\}$ and its estimated analogue by $\hat{S} := \{i : \hat{\beta}_i \neq 0\}$. Then under a “beta-min condition”, which, loosely speaking, requires the non-zero coefficients in $\beta_0$ to be sufficiently large (i.e. above the noise level), it holds that $P(S \subseteq \hat{S}) \rightarrow 1$. If, in addition, the “irrepresentability condition” (which a restrictive assumption on the design matrix) is satisfied, it holds that $P(S = \hat{S}) \rightarrow 1$.

A disadvantage of Lasso-penalized estimators is that the variable selection properties are only guaranteed under restrictive conditions and the asymptotic distribution of penalized $M$-estimators is in general not tractable. The asymptotic behaviour of $\ell_1$-penalized $M$-estimators has been studied in several papers, see e.g. Knight and Fu (2000). They suggest, as one might expect, that the classical theory on asymptotic normality of $M$-estimators cannot be immediately regenerated. Nevertheless, penalized estimators might be used as initial estimators to construct estimators that are asymptotically normal and regular under mild conditions, thus moving towards asymptotically efficient estimation. These asymptotically normal estimators may then be used for variable selection in the spirit of the more classical framework of hypothesis testing.

To construct asymptotically normal estimators for (sparse) high-dimensional models, several different methods have been studied. We mention the idea of bias correction of an initial Lasso estimator, which was studied in the papers Zhang and Zhang (2011), van de Geer et al. (2014), Javanmard and Montanari (2014) for linear regression and the paper van de Geer et al. (2014) considers in addition the generalized linear models. The message of these papers is that for inference about low-dimensional parameters of interest, one needs a good initial estimator of the high-dimensional parameter and an estimator of the score of the nuisance parameter. The approach has also been applied in particular examples of nonlinear models, see e.g. Janková and van de Geer (2014) and Janková and van de Geer (2016) for Gaussian graphical models. An alternative approach, based on Neyman’s orthogonalizing conditions, was studied in Chernozhukov et al. (2015). This yields an asymptotically normal estimator for low-dimensional parameters, by solving the orthogonalizing conditions, using an initial Lasso-regularized estimator. The paper Chernozhukov et al. (2015) provides high-level conditions under which asymptotic normality can be obtained. The estimator they propose is related to the bias correction idea, although the two approaches are not identical. This approach was also studied for the least absolute deviations estimator in Belloni et al. (2015), where asymptotic normality of the estimator based on Neyman’s orthogonalizing conditions was shown. Further works on inference in high-dimensional settings include Chatterjee and Lahiri (2013), van de Geer (2014), van de Geer and Stucky (2016), Nickl and van de Geer (2012) and other.

1.1 Contributions

The paper van de Geer et al. (2014) is closely related to our work. This paper extends the analysis therein to non-differentiable loss functions and relaxes certain conditions therein. The paper van de Geer et al. (2014) constructs an asymptotically normal estimator for low-dimensional parameter in high-dimensional generalized linear models. This is done by bias correction of an initial Lasso
estimator. To calculate the bias correction, nodewise Lasso regressions are used to approximately invert a high-dimensional precision matrix, which corresponds to the inverse Fisher information. The results of van de Geer et al. (2014) assume that the loss function is twice differentiable and the second derivative is Lipschitz. We relax this assumption by considering entropy of the classes of functions instead of Lipschitz properties. This moves the differentiability and Lipschitz conditions from the loss function onto the expected loss function. We also derive alternative theoretical results for estimating the precision matrix with nodewise Lasso regressions. These results hold for generalized linear models with bounded design under mild conditions, which are alternative to the conditions in the paper van de Geer et al. (2014). For general high-dimensional models, we provide high-level conditions, which can be checked in particular situations. The theoretical results are supported by a simulation study and applications to real data from genome studies in linear and logistic regression.

1.2 Organization of the paper

In Section 2 we consider the high-dimensional generalized linear model. We describe the de-sparsifying methodology and estimation of the inverse covariance matrix in Section 2.1. Main theoretical results for generalized linear models are contained in Section 2.2. Section 3 contains main theoretical results for nodewise regression for estimation of inverse covariance matrices. In Section 4 we consider general high-dimensional models. Examples including the Lasso, least absolute deviations estimator and the Huber estimator are contained in Section 5. Section 6 summarizes the findings. Sections 7 and 8 contain simulation studies and applications to real data sets in linear and logistic regression. The proofs are deferred to Appendix A.

1.3 Notation

Let $X_1, \ldots, X_n$ be independent random variables with values in some space $\mathcal{X}$ and let $\mathcal{F}$ be a class of real valued functions on $\mathcal{X}$. For a function $f : \mathcal{X} \to \mathbb{R}$ we denote by $\mathbb{P}_n f = \sum_{i=1}^n f(X_i)/n$ its empirical measure and by $\mathbb{F} f = \sum_{i=1}^n \mathbb{E} f(X_i)/n$ its theoretical measure (assuming the integrals exist). Let $\|f\|^2_n := \mathbb{E}_n f^2$ denote the empirical norm of $f$ and let $\|f\| = \mathbb{P} f^2$ (assuming it exists) denote the theoretical norm of $f$. Let $G_n f := \sqrt{n}(\mathbb{P}_n - \mathbb{P}) f$. By $N(\epsilon, \mathcal{F}, \|\cdot\|_n)$ we denote the covering number of the set $\mathcal{F}$, which is the minimum number of $\|\cdot\|_n$-balls with radius $\epsilon$ needed to cover the set $\mathcal{F}$. For a vector $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ we denote its $\ell_r$ norm by $\|x\|_r := (\sum_{i=1}^p x_i^r)^{1/r}$ for $r \geq 1$. We further let $\|x\|_{\infty} := \max_{i=1, \ldots, p} |x_i|$ and $\|x\|_0 = |\{i : x_i \neq 0\}|$. For a matrix $A$ we denote its $j$-th column by $A_j$ and its $(i, j)$-th element by $A_{ij}$.

2 High-dimensional generalized linear models

We are given independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. $Y_i \in \mathbb{R}^p$ has the interpretation of the dependent variable and $X_i \in \mathbb{R}^p$ represents covariates. Let $\rho_{\beta} : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ for $\beta \in \mathbb{R}^p$ be a given loss function. We assume that the loss function depends on the parameter only through the linear combination $x^T \beta$, i.e.,

$$\rho_{\beta}(x, y) := \rho(x^T \beta, y).$$

The loss function is not necessarily related to the probability distribution of the instances.

**Example 2.1. (Generalized linear models)** A special case of the above setting is the generalized linear model (McCullagh and Nelder (1989)), where $\mathbb{E}(Y_i | X_i) = g(X_i^T \beta_0)$. Then the probability density function has the form $p_{\beta_0}(y|x) = f(y, g(x^T \beta))$, for some function $f$. If the loss function equals the negative log-likelihood; this corresponds to a maximum likelihood approach.

Some examples of loss functions covered in this paper include

(i) quadratic loss $\rho(u, y) = (y - u)^2$,

(ii) quantile loss $\rho(u, y) = q|y - u|1_{y-u>0} + (1 - q)|y - u|1_{y-u\leq0}, y - u \in \mathbb{R}$, for some $0 \leq q \leq 1$,
(iii) Huber loss \( \rho(u, y) = \left( (y - u)^2 1_{|y-u| \leq K} + K(2|y-u| - K)1_{|y-u| > K} \right) / (2K) \).

(iv) hinge loss \( \rho(u, y) = (1 - yu)_+ \).

(v) mixture models: \( \rho(u, y) = \log \left( \pi_1 \sigma_1 \phi \left( \frac{y-u_1}{\sigma_1} \right) + (1-\pi) \frac{1}{\sigma_2} \phi \left( \frac{y-u_2}{\sigma_2} \right) \right) \), where \( u_1 = x^T \beta_1, u_2 = x^T \beta_2 \).

(vi) logistic loss: \( \rho(u, y) = -yu + \log(1 + e^u) \).

A given loss function defines an \( \ell_1 \)-penalized M-estimator via

\[
\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(X_i^T \beta, Y_i) + \lambda \| \beta \|_1.
\]

The above optimization problem implies the first order necessary conditions, so called “estimating equations”,

\[
\frac{1}{n} \sum_{i=1}^{n} \psi_{\hat{\beta}}(X_i, Y_i) + \lambda \hat{Z} = 0,
\]

where \( \psi_{\beta} : \mathcal{X} \to \mathbb{R}^p \) is the sub-differential of \( \rho_{\beta} \) evaluated at \( \hat{\beta} \) and \( \hat{Z} \) is the sub-differential of the \( \ell_1 \)-norm evaluated at \( \hat{\beta} \):

\[
\hat{Z}_i = \begin{cases} 
\text{sign}(\hat{\beta}_i) & \text{if } \hat{\beta}_i \neq 0, \\
\hat{q}_i \in [0, 1] & \text{otherwise}.
\end{cases}
\]

When the loss function is differentiable in \( \beta \), equation (3) simply applies with \( \psi_{\beta} := \dot{\rho}_{\beta} \). When the loss function is not differentiable in \( \beta \), but it is sub-differentiable, one may still replace the derivative by sub-differential. Examples of loss functions that are not differentiable (in every point) but the sub-derivative exists at every point include e.g. quantile loss function (used in quantile regression) or hinge loss function (used in support vector machines).

\( \ell_1 \)-penalized M-estimators have been studied extensively and under certain conditions, they copy the behaviour of an “oracle”, which knows the true position of zero entries of \( \beta_0 \). The technical conditions for oracle inequalities were briefly outlined in the introduction and we do not treat them in detail in the present paper, as they are well established in literature, see e.g. the book Bühlmann and van de Geer (2011). We remark that differentiability of the loss function is not necessary for the oracle inequalities.

2.1 Methodology: De-sparsifying the Lasso

We follow the methodology from van de Geer et al. (2014), which implements a bias correction step on the initial Lasso estimator. The methodology of bias correction removes the bias associated with the \( \ell_1 \)-penalty and leads to a non-sparse estimator which recovers the desired asymptotic properties that e.g. the maximum likelihood estimator possesses in low-dimensional settings.

2.1.1 Establishing an asymptotic pivot

The estimating equations as in (3) read

\[ P_n \psi_{\beta} + \lambda \hat{Z} = 0. \]

The idea is to find a root \( \hat{b} \) which (approximately) satisfies the estimating equations without the bias term and thus asymptotically behaves as the oracle estimator, which knows the true positions of non-zero entries of \( \beta_0 \) and applies a maximum likelihood estimator. This can be done by arguments relying on second order approximations via Taylor expansions. We will proceed in an equivalent way by “inverting the estimating equations” with the Hessian matrix of the loss function. To avoid the need to assume differentiability of the loss function, we do the inversion...
with a matrix that represents the Hessian of the expected loss function. To this end, we denote \( \Theta := ([P\psi_{j\beta}]_{j=1}^{p})^{-1} \). When the loss function is twice differentiable and equals the negative log-likelihood, \( \Theta \) is the Fisher information matrix. Multiplying the estimating equations with \( \Theta \) and adding \( \hat{\beta} - \beta_0 \) to both sides yields

\[
\hat{\beta} - \beta_0 + \Theta \lambda \hat{Z} = -\Theta P_n \psi_{\beta} + \hat{\beta} - \beta_0.
\]

This leads (by rearranging) to the following (classical) decomposition (see \cite{van der Vaart}):

\[
\hat{\beta}_j - \beta_0 - \Theta_j^T P_n \psi_{\beta} = -\Theta_j^T P_n \psi_{\beta_0} - \Theta_j^T (P_n - P)(\psi_{\beta} - \psi_{\beta_0}) + \hat{\beta}_j - \beta_0 - \Theta_j^T P(\psi_{\beta} - \psi_{\beta_0})
\]

(4)

where \( \Theta_j \) is the \( j \)-th column of \( \Theta \). Contrary to the classical setting as studied in \cite{van der Vaart}, there is an extra term \( \Theta_j^T P_n \psi_{\beta} \) which corresponds to the bias of the Lasso. Thus the decomposition (4) suggests to take a new estimator corrected by the extra term as follows

\[
\tilde{b}_j := \hat{\beta}_j - \Theta_j^T P_n \psi_{\beta}.
\]

(5)

Clearly the matrix \( \Theta \) is typically not known and hence (5) is not a proper estimator. In Section 2.1.2 below, we propose an estimator of \( \Theta \).

The decomposition (4) is the main tool in our analysis and it illustrates the challenges underlying this problem. Provided that the empirical process part and the smoothness part are small enough, of order \( O(\sqrt{n}) \), asymptotic normality of \( \sqrt{n}(\hat{\beta}_j - \beta_0) \) can be established by classical arguments, under certain conditions on the pivot term.

The empirical process part in (4) is related to the complexity of the considered class of functions, which are indexed by a sparse parameter \( \beta \). Our aim in this part is to show entropy bounds for the class

\[
\mathcal{G} := \{ x \mapsto \Theta_j^T (\psi_{\beta} - \psi_{\beta_0}) : \beta \in \mathcal{B} \},
\]

where \( \mathcal{B} \) is some sparse subset of \( \mathbb{R}^p \) that will be specified later. The sparseness of the index set \( \mathcal{B} \) is crucial, in view of our results relying on entropy numbers. For an overview of results on entropy numbers, we refer the reader to e.g. \cite{van der Vaart and Wellner}.

The smoothness part in (4) is related to the smoothness of the (derivative of the) expected loss function. The smoothness part poses a problem in high-dimensional settings since Taylor expansions have to be carried out with caution because norms in \( \mathbb{R}^p \) are not equivalent asymptotically when \( p \to \infty \).

2.1.2 Nodewise Lasso regression

The next challenge is that the high-dimensional vector \( \Theta_j \in \mathbb{R}^p \) is unknown and has to be replaced by a well-behaved estimator. In view of the decomposition (4), consider now the following second decomposition, given that \( \hat{\Theta}_j \) is an estimate of \( \Theta_j \),

\[
\hat{\beta} - \beta_0 - \Theta_j^T P_n \psi_{\beta} = \hat{\beta} - \beta_0 - \Theta_j^T P_n \psi_{\beta} - (\hat{\Theta}_j - \Theta_j)^T P_n \psi_{\beta}.
\]

This implies that in order for the remainder term to be negligible, \( \hat{\Theta}_j \) must satisfy the condition

\[
||| \Theta_j - \hat{\Theta}_j |||_{\infty} = O_P(1/\sqrt{n}).
\]

By Hölder’s inequality, we have the bound \( ||| \hat{\Theta}_j - \Theta_j |||_{\infty} \leq ||| \hat{\Theta}_j - \Theta_j |||_1 ||| P_n \psi_{\beta} |||_\infty \). Hence an appropriately fast rate for \( \hat{\Theta}_j - \Theta_j \) in \( \ell_1 \)-norm and approximately satisfied estimating equations are sufficient. To this end, note that the estimating equations (3) imply \( ||| P_n \psi_{\beta} |||_\infty = O_P(\lambda) \).
In view of the above arguments, our goal is to construct an $\ell_1$-oracle estimator of the inverse of \( \Theta \). Estimation of \( \Theta \) was well explored in literature for the case of quadratic loss; this is the same problem as estimation of the edge weights in undirected graphical models (see e.g. Friedman et al. (2008), Meinshausen and Bühlmann (2006), Yuan and Lin (2007)). The challenge arises because, in general, \( \Theta \) depends on \( \beta_0 \) (an exception is the quadratic loss). We employ the fact that for generalized linear models, there is a special structure in the matrix \( \Sigma := \Theta^{-1} \equiv (P_{\psi \beta})_{\beta=\beta_0} \). We can then use as an estimator of \( \Sigma \) the empirical version

\[
\hat{\Sigma}_\beta = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}(y_i, x_i^T \hat{\beta}) x_i x_i^T.
\]

Denoting the weights by \( \hat{w}_i := \hat{\rho}(y_i, x_i^T \hat{\beta}) \) and \( \hat{W} \equiv W_{\beta} := \text{diag}(\hat{w}_i) \); then we may rewrite \( \hat{\Sigma}_\beta = X^T \hat{W}^2 X/n \). This matrix is not invertible because of the high-dimensional setting, but we can approximately invert it using nodewise regression of each column of \( \hat{W} X \) on all the other columns. To construct in this way a nodewise regression estimator of \( \Theta \), we define

\[
\gamma_{\beta,j} := \arg \min_{\gamma_j \in \mathbb{R}^{p-1}} \| \hat{W}(X_j - X_{-j} \gamma_j) \|_2^2/n + 2\lambda\|\gamma_j\|_1,
\]

\[
\hat{\gamma}_{\beta,j} := \| \hat{W}(X_j - X_{-j} \hat{\gamma}_j) \|_2^2/n + \lambda\|\hat{\gamma}_j\|_1,
\]

and

\[
\hat{\Theta}_j := (-\hat{\gamma}_{\beta,j,1}, \ldots, -\hat{\gamma}_{\beta,j,j-1,1}, -\hat{\gamma}_{\beta,j,j+1}, \ldots, -\hat{\gamma}_{\beta,j,p})/\hat{\gamma}_{\beta,j}^2.
\]

The estimator \( \hat{\Theta}_j \) was studied in the paper van de Geer et al. (2014) for generalized linear models. In the next sections, we provide alternative conditions under which the methodology yields good estimators. For generalized linear models with non-differentiable loss functions, other methods have to be used (this is discussed in Section 3.2).

Remark 2.1. Instead of the Lasso, one could use the square-root Lasso (Belloni et al. (2011)) to estimate the partial correlations in (6):

\[
\hat{\gamma}_{\beta,j} := \arg \min_{\gamma_j \in \mathbb{R}^{p-1}} \| \hat{W}(X_j - X_{-j} \gamma_j) \|_2^2/n + 2\lambda\|\gamma_j\|_1,
\]

The advantage is that the square-root Lasso automatically estimates the noise variance as well and thus uses a universal choice of the tuning parameter which is particularly useful from a practical point of view. To avoid digressions, we do not elaborate on the theoretical results for this alternative method in the present paper.

Finally, using the nodewise regression estimator \( \hat{\Theta} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_p) \), we may define the new corrected estimator

\[
\hat{b} := \hat{\beta} - \hat{\Theta}^T \rho_{\hat{\beta}}.
\]

This estimator will be referred to as the de-sparsified Lasso, in line with van de Geer (2014). In some literature, it is called the “de-biased Lasso”.

2.2 Main theoretical results

2.2.1 Model assumptions

(A1) (Observations) Assume that \( (X_i, Y_i) \) are independent for \( i = 1, \ldots, n \) and identically distributed for each fixed \( n \). Suppose that \( \|X_i\|_{\infty} \leq K_X, i = 1, \ldots, n \), \( E[\Theta_j^T X_i]^4 = O(1) \), \( 1/\Lambda_{\min}(\Sigma) = O(1) \). Define \( \beta_0 \) by \( P_{\psi \beta_0} = 0 \) and assume that \( 1/(\Theta_j^T P_{\psi \beta_0} \rho_{\beta_0} \Theta_j) = O(1) \).

(A2) (Initial estimates) Suppose that \( \|\hat{\beta} - \beta_0\| \leq C_s \lambda \) \( E\|X(\hat{\beta} - \beta_0)\|_2^2/n \leq C_s \lambda^2 \) and \( \|\hat{\beta}\| \leq C \lambda \) with high probability for some \( C > 0 \).
We remark that condition (A1) assumes that the observations are identically distributed for every fixed \( n \). This is not important for the analysis, we only assume this to keep the presentation cleaner. The moment condition \( \text{E}[\Theta_j^T X_i]^4 = \mathcal{O}(1) \) holds e.g. for a sub-Gaussian random vector \( X_i \) (see Bühlmann and van de Geer (2011) when \( \Lambda_{\max}(\Theta) = \mathcal{O}(1) \). Clearly, it is a much weaker condition than requiring sub-Gaussianity of \( X_i \). Furthermore, as already noted, the rates of convergence \( \|\hat{\beta} - \beta_0\|_1 \leq C s \lambda_n \) were derived under mild conditions in the book by Bühlmann and van de Geer (2011). As for the condition \( \|\hat{\beta}\|_0 \leq Cs \), we show in Theorem 2.2 below that this is satisfied under mild conditions.

We introduce some further notation. If \( u \mapsto \rho(y, u) \) is differentiable at \( u \) then we denote \( w(y, u) := \frac{\partial \rho(y, u)}{\partial u} \). We call \( w \) the weight function or weight. For illustration, we give a few examples of weight functions:

(i) quadratic loss: \( w_{\text{quadratic}}(y, u) = 2(y - u) \),

(ii) absolute loss: \( w_{\text{absolute}}(y, u) = \text{sign}(y - u) \), for all \( u \neq y \).

(iii) Huber loss:

\[
    w_{\text{Huber}}(y, u) = \begin{cases} 
        (y - u)/K & \text{if } |y - u| \leq K, \\
        \text{sign}(y - u) & \text{otherwise}.
    \end{cases}
\]

### 2.2.2 Main results for differentiable loss functions

In this section we consider differentiable loss functions as summarized in the following conditions.

\((B1)\) (Sparsity) Let \( \|\Theta_j\|_0 \leq s \), where it holds \( s^3 (\log p)^2 (\log n)^2 / n \to 0 \).

\((B2)\) (First order differentiability) Assume that \( w(y, u) \) is Lipschitz in \( u \) with a constant \( L = \mathcal{O}(1) \) for all \( \beta \) such that \( \|\beta - \beta_0\|_1 \leq \delta \) for some \( \delta > 0 \).

\((B3)\) (Second order differentiability) Suppose that \( u \mapsto w(y, u) \) is differentiable and its derivative \( \dot{w} \) is Lipschitz in \( u \) with \( L = \mathcal{O}(1) \) for all \( \beta \) such that \( \|\beta - \beta_0\|_1 \leq \delta \) for some \( \delta > 0 \).

Note that assumption \((B2)\) rules out the absolute loss and assumption \((B3)\) rules out the Huber loss. Relaxations of \((B2)\) were treated in Section 2.2.3 below. We remark that the analysis in van de Geer et al. (2014) requires both \((B2)\) and \((B3)\) i.e. that the loss function is twice differentiable and the second derivative is Lipschitz. The conditions \((B2)\) and \((B3)\) are stated separately only in view of Theorem 2.1 below. We further need conditions \((E1)\), \((E2)\), which are needed for the estimation of the score for the nuisance parameter as given in Section 3. The following lemma, which underlies the theoretical result of Theorem 2.1 below, gives an bound on entropy of a certain class of functions.

**Lemma 2.1.** Suppose that condition \((B2)\) is satisfied. For some constant \( C_1 > 0 \) let

\[
    \mathcal{F} := \{ x \mapsto \Theta_j^T x(\beta - \beta_0) : \|\beta - \beta_0\|_1 \leq \delta, \|\beta\|_0 \leq C_1 s \},
\]

Then for all \( \epsilon > 0 \)

\[
    N(\mathcal{F}, \| \cdot \|_n, \epsilon) \leq C_2 \left( \frac{B s}{\epsilon} \right)^s,
\]

and

\[
    \log N(\mathcal{F}, \| \cdot \|_n, \epsilon) \leq C_3 s \log p + C_3 s \log \left( \frac{1}{\epsilon} \right),
\]

for some constants \( C_2, C_3 > 0 \).

**Theorem 2.1.** Let \( \rho_\beta \) be a given loss function and assume that Conditions \((A1)\), \((A2)\), \((B1)\) and \((B2)\) are satisfied with \( \psi_\beta = \rho_\beta \). Let \( \hat{\beta} \) be defined by \((9)\). Define

\[
    \tilde{b}_j := \hat{\beta}_j - \Theta_j^T \mathbb{P}_n \hat{\rho}_\beta.
\]
Then
\[ \sqrt{n}(\hat{b}_j - \beta_0)/\sqrt{\Theta_j^T P \hat{\beta}_n \hat{\beta}_0^T \Theta_j} \to \mathcal{N}(0,1). \]

Let \( \hat{\Theta}_j \) be defined in (7). In addition, if conditions (B3), (E1), (E2) hold, then
\[ \|\hat{\Theta}_j - \Theta_j\|_1 = O_P(s^{3/2}\sqrt{\log p/n}), \]
and the estimator
\[ \hat{b}_j := \hat{\beta}_j - \hat{\Theta}_j^T P_n \hat{\beta}_0 \]
satisfies
\[ \sqrt{n}(\hat{b}_j - \beta_0)/\sqrt{\hat{\Theta}_j^T P_n \hat{\beta}_0 \hat{\beta}_0^T \hat{\Theta}_j} \to \mathcal{N}(0,1). \]

Compared to Theorem 3.1 in \cite{van2014}, Theorem 2.1 does not relax the differentiability conditions. This differentiability result is however only needed for estimation of the score for the nuisance parameter \( \Theta_j \), as can be seen from the first part of Theorem 2.1. Thus the differentiability of the loss function can be relaxed to first derivative being Lipschitz, provided that a good estimate of \( \Theta_j \) (in \( \ell_1 \)-norm) is available.

The second part of the theorem illustrates conditions for nodewise Lasso which are alternative to the conditions in \cite{van2014}. We do not need the condition \( \|X\Theta_j\|_\infty \leq K \) (where \( K > 0 \) is a constant). The price we pay for this relaxation is a stronger sparsity assumption \( s^{3/2}/\sqrt{n} \).

An application of the result of Theorem 2.1 can be considered for the case of Huber loss. Huber loss is once differentiable and the derivative is Lipschitz continuous. The second derivative exists everywhere except \( |u - y| = K \), however, it is not Lipschitz. Hence the results of \cite{van2014} do not apply to Huber loss. The first part of the Theorem 2.1 does apply, and one then needs to estimate \( \Theta \). This is treated in Section 5.3.

### 2.2.3 Main results for non-differentiable loss functions

Theorem 2.1 however still does not cover an important example such as the quantile regression due to assumption (B2). The absolute loss is differentiable everywhere except \( u = y \), however, the first derivative is not Lipschitz. In the conditions below, we do not require that \( w \) is Lipschitz, but we require that its expectation is Lipschitz, which is a much weaker assumption. We formulate a relaxation of Theorem 2.1 to non-differentiable functions below. This requires an entropy condition on the class of functions which are related to the empirical process part of the problem.

\( \text{(C1)} \) Suppose that the function \( \psi_{\beta}(y, x) \) has the form \( \psi_{\beta}(y, x) = w(y, x^T \beta)x \) for some function \( w \). Assume that the function \( u \mapsto G(u) := \int w(u, y) dP_{Y_i|x} \) is differentiable and \( u \mapsto G'(u) \) is Lipschitz. Suppose that for some \( \delta > 0 \), the function \( w_\beta \) is bounded from above and stays away from zero uniformly in \( n \) for all \( \beta \) that satisfy \( \|\beta - \beta_0\|_1 \leq \delta \).

\( \text{(C2)} \) Suppose that \( \mathbb{E}(w_\beta - w_{\beta_0})^2 = O_s^{-2}(\log p)^{-2}(\log n)^{-4} \) for all \( \|\beta - \beta_0\|_1 \leq \delta \) for some \( \delta > 0 \).

**Theorem 2.2.** Assume conditions (A1), (A2), (C1) and (C2) with some function \( \psi_\beta \). Suppose that the function \( \psi_{\beta} \) and \( \hat{\beta} \in \mathbb{R}^p \) satisfy
\[ \|P_n \psi_\beta\|_\infty = O_P(\lambda). \]

Suppose that \( \|\Theta_j\|_0 \leq s \) and \( \hat{\Theta}_j \) satisfies \( \|\hat{\Theta}_j - \Theta_j\|_1 = O_P(s\lambda) \). Consider the class of functions \( \mathcal{F} := \{ \Theta_j^T (\psi_{\beta} - \psi_{\beta_0}) : \|\beta - \beta_0\|_1 \leq s\lambda, \mathbb{E}\|X(\beta - \beta_0)\|_2^2/n \leq s^2\lambda^2, \|\beta\|_0 \leq s \} \). Suppose that
\[ \log N(\mathcal{F}, \|\cdot\|_n, \epsilon) \leq s \log \left( \frac{p\|F\|_n}{\epsilon} \right), \tag{9} \]
where \( F(x) = \sup_{f \in \mathcal{F}} |f(x)| \) is the envelope function of \( \mathcal{F} \). Then
\[ \sqrt{n}(\hat{\beta}_j + \hat{\Theta}_j^T P_n \psi_\beta - \beta_0)/\sqrt{\hat{\Theta}_j^T P_n \psi_{\beta_0} \hat{\Theta}_j^T} \to \mathcal{N}(0,1). \]
The theorem replaces the differentiability assumption by an entropy condition (see van der Vaart and Wellner (1996) for similar arguments). The theorem assumes that we can estimate the score for the nuisance parameter; this is discussed in Section 3.

2.2.4 Sparsity of the Lasso

The following lemma shows that under mild conditions, the $\ell_1$-penalized M-estimator $\hat{\beta}$ has sparsity of the same order as $\beta_0$ with high probability. It is worth out point out that we do not require differentiability of the loss function.

**Lemma 2.2.** Suppose that condition (A1) holds. Let $\hat{\beta}$ be defined by

$$P_n \psi \hat{\beta} + \lambda \hat{Z} = 0,$$

where $\hat{Z}$ is the sub-differential of the $\ell_1$-norm evaluated at $\hat{\beta}$ and where the function $\beta \mapsto \psi \beta$ satisfies condition (C1) with a function $G$ such that $|G'| \leq K$ for some constant $K > 0$. Assume that $\|X(\hat{\beta} - \beta_0)\|_2 = O_P(s\lambda^2)$ and $\|\hat{\beta} - \beta_0\|_1 = O_P(s\lambda)$. Further assume that (9) is satisfied. Then

$$\|\hat{\beta}\|_0 = O_P(s).$$

2.2.5 Estimation of asymptotic variance

To construct confidence intervals, one needs to estimate the asymptotic variance of the de-sparsified estimator. The following lemma shows that we can use $\hat{\Theta}^T_j P_n \psi \beta \hat{\beta} \hat{\Theta}^T_j$ as an estimator of the asymptotic variance of the de-sparsified estimator, where $\hat{\Theta}$ is the nodewise regression estimator.

**Lemma 2.3.** Assume conditions (A1), (A2), (B1), (B2) and suppose that $\|\hat{\Theta}_j - \Theta_j\|_1 = O_P(s\lambda^2)$. Then

$$|\hat{\Theta}^T_j P_n \psi \beta \hat{\beta} \hat{\Theta}^T_j - \Theta^T_j P \psi \beta_0 \psi \beta_0^T \Theta_j| = o_P(1).$$

3 Nodewise regression for estimation of precision matrices

Our goal in this section is to provide estimators for $\Theta := \Sigma^{-1}$, where

$$\Sigma := (\mathbb{E} \psi \beta(x,y))_{\beta = \beta_0}.$$

If the parameter of interest is $\beta_0^j$ for some $j$, we only need to estimate $\Theta_j$. In the next sections, we suggest procedures for estimation of $\Theta$ in generalized linear models. We first consider first the case when the loss function is differentiable and we also discuss the case when it is not differentiable.

3.1 Generalized linear models with differentiable loss functions

If the loss function is twice differentiable, then

$$\Sigma = \mathbb{E} \hat{\beta}(y, x^T \beta_0) xx^T = \mathbb{E} \hat{w}(y, x^T \beta_0) xx^T.$$

Hence we can approximate $\Sigma$ by the empirical version

$$\hat{\Sigma}_\beta := \frac{1}{n} \sum_{i=1}^n \hat{w}_\beta(y_i, x_i) x_i x_i^T.$$

This matrix is not invertible in high-dimensional settings, but we can use e.g. nodewise regression to approximately invert it, as outlined in Section 2. The main difficulty here is that the estimator depends on the estimator $\hat{\beta}$.

We formulate the results for general weights satisfying the conditions below. Let the weight matrix $W_\beta$ be given by

$$W_\beta := \text{diag}(v(y_i, x_i^T \hat{\beta}))_{i=1,\ldots,n},$$

for some weight function $v$. Let $\Sigma := \mathbb{E} v(y, x^T \beta) xx^T$ and $\Theta = \Sigma^{-1}$. Below we provide theoretical guarantees for the estimator $\hat{\Theta}_j$ defined in equation (7), with the weight matrix $W_\beta$. We make the following assumptions.
(E1) The matrix inverse $\Theta := \Sigma^{-1}$ exists and $\max_{j=1,\ldots,p} \|\Theta_j\|_0 \leq s$. Moreover, $1/\Lambda_{\min}(\Sigma) = O(1)$, $\|\Sigma\|_{\infty} = O(1)$.

(E2) There exists some $\delta > 0$ such that for all $\|\beta - \beta_0\|_1 \leq \delta$ it holds that $1/v_\beta = O(1)$ and $v_\beta = O(1)$. Furthermore, $v_\beta$ is Lipschitz with a universal constant.

(E3) It holds that $s^{3/2}/\sqrt{\log p/n} = o(1)$.

**Theorem 3.1.** Suppose that Conditions [(A1)] [(A2)] and [(E1)] [(E3)] are satisfied. Then using $\lambda_j \asymp \sqrt{\log p/n}$ for the nodewise Lasso estimator $\hat{\Theta}_j$ defined in (1) it holds that

$$
\|\hat{\Theta}_j - \Theta_j\|_1 = O_p(s^{3/2}/\sqrt{\log p/n}).
$$

Theorem 3.1 relaxes the condition $\|X\Theta_j\|_{\infty} = O(1)$ from van de Geer et al. [2014]. Furthermore, we remark that it is a more general result than in van de Geer et al. [2014] in that the latter only considers that $\Sigma = \mathbb{E}wxx^T$, but we allow for any $\Sigma = \mathbb{E}wxx^T$ for arbitrary weights $v$ satisfying the conditions. From the point of view of the proof, this makes no actual difference, however, the application of the result is then somewhat broader, as will be illustrated in Section 3.2.2 below.

### 3.2 Generalized linear models with non-differentiable loss functions

If the loss function is not differentiable, the above strategy clearly cannot be used to estimate $\Theta$. We discuss some alternative options that could be used.

#### 3.2.1 Special cases

In some settings, one can make use of the structure in $(P\psi)^T_{\beta_0}$. In particular, we have

$$
(P\psi)^T_{\beta} = \mathbb{E}_X (\mathbb{E}_Y w(u,Y))^T_{u=x^T\beta} x_i x_i^T.
$$

The above can be approximated by

$$
\hat{\Sigma}_{\beta} := \frac{1}{n} \sum_{i=1}^n (\mathbb{E}_Y w(u,Y))^T_{u=x^T\beta} x_i x_i^T.
$$

In some situations, it is possible to calculate $(\mathbb{E}_Y w(u,Y))^T_{u=x^T\beta_0}$ provided that we assume the distribution of $Y$, and then we can plug in an estimate $\hat{\beta}$ of $\beta_0$. Then we can estimate $\Sigma$ by

$$
\hat{\Sigma}_{\beta} := \frac{1}{n} \sum_{i=1}^n (\mathbb{E}_Y w(u,Y))^T_{u=x^T\hat{\beta}} x_i x_i^T.
$$

We can then use nodewise regression with weights, and under some conditions on the weights, the nodewise regression yields good estimators of $\Theta$, see Theorem 3.1 in Section 3. For instance, for absolute loss in the linear model, we have (if $\epsilon$ and $X$ are independent)

$$
(\mathbb{E}w_{\text{absolute}}(y, x^T\beta))'_{\beta=\beta_0} = 2\epsilon_f(0),
$$

and for Huber loss

$$
(\mathbb{E}w_{\text{Huber}}(y, x^T\beta))'_{\beta=\beta_0} = F_{\epsilon}(K) - F_{\epsilon}(-K).
$$

#### 3.2.2 Maximum likelihood

If the loss function equals the negative log-likelihood, we can consider the following approach. Denote the score function by $s_{\beta} := \frac{\partial \log p}{\partial \beta}$. Then (for differentiable loss) the following identity holds

$$
\mathbb{E}s_{\beta} s_{\beta}^T = -\mathbb{E}s_{\beta} = -\mathbb{E}\hat{\beta}.
$$

This implies that $\hat{\Sigma}$ has the form of a Gram matrix with inner products given by score functions corresponding to individual parameters. Hence we could use as an alternative estimator of $\Sigma$

$$
\frac{1}{n} \sum_{i=1}^n s_{\beta}(X_i) s_{\beta}(X_i)^T = \frac{1}{n} \sum_{i=1}^n w^2(y_i, x_i^T \hat{\beta}) x_i x_i^T.
$$
This again has the form $X^T\hat{W}X$, so we can do nodewise regression, if conditions [E1] [E3] are satisfied with weight function $v := w^2$. For instance, for absolute loss, $w^2(y, u) = 1$ if $|u| \leq K$ and $w^2(y, u) = u^2$ otherwise. This function is Lipschitz, and hence Theorem [3.1] can be applied.

4 General high-dimensional models

In this section we provide results for general models under high-level conditions. These are useful for insight into the underlying machinery and its limits. Furthermore, they can be used to obtain results for more general models than the generalized linear models. Assume we have independent data $X_1, \ldots, X_n \in \mathcal{X}$. We make the following assumptions.

(D1) The function $\beta \mapsto (P\psi_\beta)'|_{\beta = \beta_0}$ is differentiable with a matrix of first derivatives $\Sigma := (P\psi_\beta)'|_{\beta = \beta_0}$, which satisfies the eigenvalue condition $\Lambda_{\min}(\Sigma) \geq c$, for some $c > 0$. Denote $\Theta := \Sigma^{-1}$.

(D2) Suppose that the following expansion holds

$$\| (P\psi_\beta - P\psi_{\beta_0}) - \Sigma(\beta - \beta_0) \|_\infty = O(d^2(\beta, \beta_0)),$$

where $d$ is some metric.

(D3) Suppose that $\hat{\Theta}_j$ satisfies $\| \hat{\Theta}_j - \Theta_j \|_1 = O_P(s\lambda)$.

(D4) It holds that

$$G_n\Theta_j^T(P\psi_{\hat{\beta}} - P\psi_{\beta_0}) = o_P(1).$$

(D5) It holds that

$$\sqrt{n}\Theta_j^TP_n\psi_{\beta_0}/\sqrt{\Theta_j^TP\psi_{\beta_0}\psi_{\beta_0}^T\Theta_j} \sim \mathcal{N}(0, 1).$$

Theorem 4.1. Let $\psi_\beta : \mathcal{X} \to \mathbb{R}^p$ satisfy $\|P_n\psi_{\beta}\|_\infty = O_P(\lambda)$ in a given $\hat{\beta} \in \mathbb{R}^p$ and define $\beta_0$ by $P\psi_{\beta_0} = 0$. Assume that conditions [(D1) (D2) (D3) (D4) (D5)] are satisfied with some function $d(\beta, \beta_0)$. Denote $s := \|\Theta_j\|_0$ and assume the sparsity condition

$$s = o_P(\max\{n^2d^2(\hat{\beta}, \beta_0), 1/(\sqrt{n}\lambda^2)\}).$$

Then

$$\sqrt{n}(\hat{\beta}_j + \hat{\Theta}_j^TP_n\psi_{\beta} - \beta_0^j)/\sqrt{\hat{\Theta}_j^TP\psi_{\beta_0}\psi_{\beta_0}^T\Theta_j} \sim \mathcal{N}(0, 1).$$

The conditions on the initial estimator $\hat{\beta}$ are in the assumption $s = o_P(n^2d(\hat{\beta}, \beta_0)^2)$ and in the empirical process condition [(D4)]. The sparsity condition [(10)] has two parts: the first part $s = o_P(n^2d(\hat{\beta}, \beta_0)^2)$ ensures that there is enough continuity in the problem and the second part $s = o_P(1/(\sqrt{n}\lambda^2))$ ensures that the estimator of the score for the nuisance parameter is good enough.

Theorem 4.1 assumes asymptotic equicontinuity [(D4)]. The following Theorem shows sufficient conditions for the asymptotic equicontinuity to hold.

Theorem 4.2. Suppose that for the class of functions $\mathcal{F}$ it holds that

$$\log N(\mathcal{F}, \| \cdot \|_n, \epsilon) \leq s \log \left( \frac{p\|F\|_n}{\epsilon} \right),$$

for some
where \( F = \sup_{f \in \mathcal{F}} |f| \) is the envelope function of \( \mathcal{F} \). Let \( R := \sup_{f \in \mathcal{F}} \| f \| \) and suppose that

\[
R \log n \sqrt{s \log p/n} = o(1/\sqrt{n}),
\]

\[
(E \sup_{f \in \mathcal{F}} f^4(X_1))^{1/4} (\log n \sqrt{s \log p/n})^{3/2} = o(1/\sqrt{n}).
\]

Then

\[
\sup_{f \in \mathcal{F}} G_n f = o_P(1).
\]

We aim to apply Theorem 4.2 with the class of functions

\[
\mathcal{F} := \{ \Theta_f^T (\psi_\beta - \psi_{\beta_0}) : \beta \in \mathcal{B} \},
\]

for some set \( \mathcal{B} \subset \mathbb{R}^p \) which can be specified depending on the problem at hand (for generalized models, see e.g. Theorem 2.1), but we must ensure that \( \hat{\beta} \in \mathcal{B} \) and at the same time, the set \( \mathcal{B} \) must be in some sense sparse. Conditions (D1) are discussed in Section 4.1 below. They mean that the higher order remainders from the Dudley’s integral are small enough. The idea is it should be possible to get a rate for these remainders over the set \( \mathcal{F} \), since the set is shrinking with \( n \). They can be shown for generalized linear models under some sufficient conditions. Combining Theorem 4.1 with Theorem 4.2 gives explicit sufficient conditions under which asymptotic normality can be achieved.

### 4.1 Discussion of the conditions

**Condition (D1)** Condition (D1) avoids the need to assume differentiability of \( \rho_\beta \) or \( \psi_\beta \) directly. Instead we assume differentiability of the expected loss. Note that in some situations, the matrix \( \Theta \) may not exist, for instance for the linear model with fixed design. To be able to describe the asymptotics, we may then assume that there exists a non-singular matrix \( \Theta \) (with eigenvalues bounded from above and away from zero) such that

\[
\| \Theta_f^T (P\psi_\beta)'|_{\beta = \beta} - c_j \|_\infty = o(1/\sqrt{n}).
\]

Consequently, replacing assumption (D1) in Theorem 4.1 with (12), the result of Theorem 4.1 applies.

**Condition (D2)** In condition (D2) \( d(\beta, \beta_0) \) represents a metric suitable for the problem at hand. For generalized linear models, one may choose \( d^2(\beta, \beta_0) = E|x^T (\beta - \beta_0)|^2 = (\beta - \beta_0)^T \Sigma(\beta - \beta_0) \). For general models, if the function \( \beta \rightarrow \frac{\partial^2 (P\psi_{\beta_0})}{\partial \beta^2} \) is bounded for all \( i, j, k = 1, \ldots, p \), one may choose \( d^2(\beta, \beta_0) = \| \beta - \beta_0 \|^2 \) (see Lemma A.4).

**Sparse conditions (11)**. To have sufficient continuity in the model as described in condition (D2) some sparsity assumptions must be made. Naturally, considering more general models costs more. In general, we require \( d^2(\hat{\beta}, \beta_0) = o \left( \frac{1}{\sqrt{n}} \right) \). For generalized linear models this conditions amounts to \( s^3 (\log p)^2 / n \to 0 \), even for non-differentiable loss, provided that the expected loss is differentiable.

**Condition (11)**. Theorem 4.2 suggests that we need some rate on \( R = \sup \| f \| \), in particular it must be shown that \( R = o \left( \frac{1}{\sqrt{s \log p \log n}} \right) \). This can indeed be shown for e.g. generalized linear models, under sufficient sparsity conditions (see Section 2).

Condition (P sup\( f \in \mathcal{F} \) \( f^4 \)) \( (\log n)^b (s \log p)^3 / n = o(1) \) is satisfied e.g. for Lipschitz \( \psi_\beta \). An envelope function for the class \( \mathcal{F} := \{ \Theta_f^T (\psi_\beta - \psi_{\beta_0}) : \| \beta - \beta_0 \|_1 \leq s \lambda \} \) is then obtained using the following upper bounds

\[
\| \Theta_f^T (\psi_\beta - \psi_{\beta_0}) \| \leq \| \Theta_f^T \|_1 \| \psi_\beta - \psi_{\beta_0} \|_\infty \\
\leq \| \Theta_f^T \|_1 L \| \beta - \beta_0 \|_1 \\
\leq \Lambda_{\max}(\Theta) L \sqrt{s} \| \beta - \beta_0 \|_1 \leq \Lambda_{\max}(\Theta) L s^{3/2} \lambda.
\]
Then clearly, \( P \sup_{f \in \mathcal{F}} f^4 \leq (Cs^{3/2} \lambda)^4 \). Then the condition [11] is satisfied under the sparsity \( s^3(\log p)^{5/3} / (\log n)^2 / n = o(1) \).

Or, for instance, for generalized linear models, if \( w_\beta - w_{\beta_0} \) is uniformly bounded, then
\[
\left( P \sup_{f \in \mathcal{F}} f^4 \right)^{1/4} = \left( P \sup_{\beta : \beta \in \mathcal{F}} |\Theta_j^T x (w_\beta - w_{\beta_0})|^4 \right)^{1/4} \leq (P |\Theta_j^T x|^4 K)^{1/4}.
\]

Then under a moment condition \( P |\Theta_j^T x| = O(1) \), the expression [13] is bounded. Hence we would require \( s^3(\log p)^3(\log n)^6 / n = o(1) \).

5 Examples

5.1 Quadratic loss

Consider the linear model
\[
Y = X\beta^0 + \epsilon,
\]
(14)
where \( \epsilon := (\epsilon_1, \ldots, \epsilon_n) \) with \( \epsilon_i \)’s independent and \( \mathbb{E}\epsilon = 0 \). \( X \) is a \( n \times p \) matrix independent of \( \epsilon \) with i.i.d. rows with mean zero and covariance matrix \( \Sigma := \mathbb{E}X^T X \). We assume the inverse \( \Theta = \Sigma^{-1} \) exists and suppose that \( \Lambda_{\max}(\Sigma) = O(1) \) and \( 1/\Lambda_{\min}(\Sigma) = O(1) \). Moreover, we assume that \( \|X_i\|_{\infty} \leq K, \mathbb{E}[\Theta_j^T x]^4 = O(1), \|\beta_0\|_0 \leq s, \|\Theta_j\|_0 \leq s \) and \( \Lambda_{\max}(\Sigma) = O(1) \). Finally, assume the sparsity condition \( s^3(\log p)^2(\log n)^2 / n = o(1) \).

Consider the Lasso estimator
\[
\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \beta)^2 + \lambda \|\beta\|_1.
\]
and its de-sparsified version \( \hat{b} = \hat{\beta} + \hat{\Theta}(Y - X\hat{\beta}) / n \). We apply Theorem 2.1. The loss function is differentiable, so one may take
\[
\psi_\beta(x, y) := \rho_\beta(x, y) = 2(y - x^T \beta)x.
\]
By the above assumptions, the weight function is \( w(u, y) = 2(y - u), \) and hence \( \hat{\psi}(u, y) = -2 \) is Lipschitz in \( u \). Hence conditions [A1] [A2] [B1] [B2] [B3] [E1] [E2] are satisfied.

5.2 Absolute loss

Consider the linear model [14]. The \( \ell_1 \)-penalized least absolute deviations (LAD) estimator is defined by
\[
\hat{\beta}_{\text{LAD}} := \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n |Y_i - X_i^T \beta| + \lambda \|\beta\|_1.
\]
(15)
One may take
\[
\psi_\beta(x, y) := \text{sign}(y - x^T \beta)x.
\]
Lemma 5.1 below shows that the estimating equations are approximately satisfied with \( \psi_\beta \) at the point \( \beta_{\text{LAD}} \). We also need to construct an estimate of \( \Theta = \frac{1}{2f_c(0)} (EX_iX_i^T)^{-1} \). A near-oracle estimate of \( (EX_iX_i^T)^{-1} \) can be obtained using nodewise regression with input matrix \( \Sigma := X^T X / n \) under conditions [F1] [F3] below (see van de Geer et al. [2014]).

(F1) Assume that there exists a universal constant \( K > 0 \) such that \( \|X_i\|_{\infty} \leq K \). Moreover, \( X_i \) are independent of \( \epsilon_i \), for \( i = 1, \ldots, n \).

(F2) Let the distribution function \( F_c \) of \( \epsilon_i \) satisfy \( F_c(0) = 1/2 \) and let it have a density \( f_c \). Furthermore, \( f_c(0) \geq c > 0 \), where \( c \) is a fixed constant and \( |f_c(0)| \leq K \) for all \( x \in \mathbb{R} \) and a universal constant \( K \). Suppose that \( f_c \) is Lipschitz.
Theorem 5.1. Assume the model is satisfied. Let \( \hat{K} > K \) for some constant \( \sqrt{n} \). It holds
\[
\mathbb{E}[(\hat{\Theta}_j^T X_i)^4] = O(1).
\]
Define the function
\[
\ell(x, y) = \text{sign}(y - x^T \beta) x.
\]
Hence we may apply the first part of Theorem 2.1. Define the
\[
\psi_{\beta}(x, y) := \text{sign}(y - x^T \beta) x.
\]
We again consider the linear model. The loss function is given by
\[
\rho_{\beta}(x, y) := \psi_{\beta}(x, y) := \text{sign}(y - x^T \beta) x.
\]
For Huber loss, we have
\[
\mathbb{E}[\psi_{\beta}(y, x^T \beta)]_{\beta=\beta_0} = (F_\epsilon(K) - F_\epsilon(-K))/K,
\]
and hence
\[
\Sigma = (\mathbb{E}[\psi(y, x^T \beta)]_{\beta=\beta_0}) = (F_\epsilon(K) - F_\epsilon(-K))/K \mathbb{E}xx^T.
\]

Lemma 5.1. Assume the linear model is satisfied and that \( s/n = O(\lambda) \). Let \( \psi_{\beta}(x, y) := \text{sign}(y - x^T \beta) x \). Then
\[
\|\mathbb{P}_n \hat{\psi}_{\beta LAD}\|_\infty = O_{p}(\lambda).
\]

Lemma 5.2. Let
\[
\mathcal{F} := \{ \Theta_j^T (\psi_{\beta}(x, y) - \psi_{\beta_0}(x, y)) : \beta \in \mathbb{R}^p, \|\beta\|_0 \leq s \}.
\]
Then for all \( \epsilon > 0 \) it holds
\[
\log N(\epsilon\|F\|_n, \mathcal{F}, \|\cdot\|_n) \leq Cs\log p + 2Cs\log(16\epsilon/\varepsilon),
\]
for some constant \( C > 0 \).

Define \( \sigma_j^2 := \frac{1}{4\mathcal{J}(0)^2} \Theta_{j,j} \), then we have the following result for the de-sparsified LAD estimator.

Theorem 5.1. Assume the model is satisfied and that conditions \( (F1), (F2), (F3) \) are satisfied. Let \( \hat{\beta}_{LAD} \) be defined in (16). Let an estimate \( \hat{\Theta}' \) of \( \Theta' \) be constructed using nodewise regression with the input matrix \( \hat{X} := X/n, \) and let \( \hat{\Theta}_j := \hat{\Theta}'_j/(2f_\epsilon(0)) \). Let \( \|\Theta_j\|_0 \leq s \) and \( s^5(\log p)^3(\log n)^4/n = o(1) \). Then for
\[
\hat{b}_{LAD,j} := \hat{\beta}_{LAD,j} + \frac{1}{n} \sum_{i=1}^{n} \text{sign}(Y_i - X_i^T \hat{\beta}_{LAD}) \hat{\Theta}_j^T X_i
\]
\[
\sqrt{n}(\hat{b}_{LAD,j} - \beta_{j,0}^0)/\sigma_j \sim N(0, 1).
\]

5.3 Huber loss

We again consider the linear model. The loss function is given by
\[
\rho_{\beta}(x, y) := \rho(y - x^T \beta),
\]
where
\[
\rho(z) = \begin{cases} z^2/2 & |z| \leq K, \\ K(2|z| - K) & |z| > K, \end{cases}
\]
for some constant \( K > 0 \). We note that the first derivative satisfies the Lipschitz condition
\[
|\rho(u, y) - \rho(u', y)| \leq |u' - u|, \text{ for all } u, u', y.
\]

Hence we may apply the first part of Theorem 2.1. Define the \( \ell_1 \)-penalized Huber estimator
\[
\hat{\beta}_{\text{Huber}} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_{\text{Huber}}(Y_i - X_i^T \beta) + \lambda \|\beta\|_1.
\]
Define the function
\[
\psi_{\beta}(x_i, y_i) = \psi(y_i - x_i \beta) := \psi(z) = \begin{cases} z/K & \text{if } |z| \leq K, \\ \text{sign}(z) & \text{if } |z| > K. \end{cases}
\]
For Huber loss, we have
\[
(\mathbb{E}_n \psi_{\beta, j}(y, x^T \beta))_{\beta=\beta_0} = (F_\epsilon(K) - F_\epsilon(-K))/K,
\]
and hence
\[
\Sigma = (\mathbb{E}_n \psi(y, x^T \beta))_{\beta=\beta_0} = (F_\epsilon(K) - F_\epsilon(-K))/K \mathbb{E}xx^T.
\]
Furthermore, 
\[
\mathbb{E}\psi_{\beta_0}^T \psi_{\beta_0}^T = \mathbb{E}w_{\text{Huber},j}^2(y,a^T\beta_0)xx^T = \left\{ \frac{1}{K} \mathbb{E}_{\epsilon}1_{|\epsilon|<K}\epsilon^2 + [F_\epsilon(-K) + 1 - F_\epsilon(K)] \right\} xx^T.
\]
Hence the asymptotic variance per entry of the de-sparsified estimator is
\[
\sigma^2_{\text{Huber},j} := \Theta_j^T\mathbb{E}\psi_{\beta}^T\psi_{\beta}^T\Theta_j = K^2\mathbb{E}_{\epsilon}1_{|\epsilon|<K}\epsilon^2/K + [F_\epsilon(-K) + 1 - F_\epsilon(K)]/\left(F_\epsilon(K) - F_\epsilon(-K)\right)^2 \Theta_{jj}^0.
\]
One could then define an estimator of \(\sigma^2_{\text{Huber},j}\) as follows
\[
\hat{\sigma}^2_{\text{Huber},j} := \frac{1}{K^2} \frac{1}{n} \sum_{i=1}^n \frac{1}{\frac{1}{n} \sum_{i=1}^n 1_{|\epsilon_i|<K}\epsilon_i^2 + \frac{1}{n} \sum_{i=1}^n 1_{|\epsilon_i|>-K} \epsilon_i^2}{\Theta_{jj}^0},
\]
where \(\epsilon_i := Y_i - X_i^T\hat{\beta}_{\text{Huber}}\) and \(\hat{\Theta}_{jj}\) was obtained using nodewise regression with matrix \(X\).

**Theorem 5.2.** Assume the model (14) and suppose that conditions (F1), (F3) are satisfied. Let \(\hat{\beta}_{\text{Huber}}\) be defined in (17). Let an estimate \(\hat{\Theta}'\) of \(\Theta'\) be constructed using nodewise regression with the input matrix \(\hat{\Sigma} := X^TX/n\), and let \(\hat{\Theta}_j := \hat{\Theta}_j/(F(K) - F(-K))\). Let \(||\Theta_j||_0 \leq s\) and \(s^3(\log p)^2(\log n)^2/n = o(1)\). Then for
\[
\hat{b}_{\text{Huber},j} := \hat{\beta}_{\text{Huber},j} + \frac{1}{n} \sum_{i=1}^n \hat{\Theta}_j^T\psi_\beta(X_i,Y_i)
\]
it holds
\[
\sqrt{n}(\hat{b}_{\text{Huber},j} - \beta_{j0})/\sigma_{\text{Huber},j} \rightsquigarrow \mathcal{N}(0,1).
\]

6 Conclusions

A first message of our analysis is that to obtain asymptotically normal estimator in high-dimensional settings, we do not require the loss to be twice differentiable. Instead however, one must assume that the expected loss is sufficiently smooth and that the class of the “score” functions indexed by the unknown parameter satisfies a certain entropy bound. To this end, our analysis needed sparsity in the Lasso estimator, which was shown in Section 2. A second message is that we need to estimate the score of the nuisance parameter for the methodology to work. There is a certain price we pay compared to the results in [van de Geer et al. (2014)]; our analysis leads to somewhat stronger sparsity assumptions. The analysis in [van de Geer et al. (2014)] requires a sparsity condition \(s = o(\sqrt{n}/\log p)\) for generalized linear models, where \(s = \max(||\beta_0||_0, ||\Theta_j||_0\) \}. We need the somewhat stronger condition \(s^{3/2} = o(\sqrt{n}/\log p)\). This results from considering non-differentiable loss functions on one hand, and from estimation of the score for the nuisance parameter on the other.

7 Simulation study

We confirm our theoretical results with numerical experiments on synthetic data and compare the performance of our approach with other plausible procedures, such as the maximum likelihood estimator.

7.1 Models

We consider the linear model with a continuous random variable and logistic regression with a binary response variable. In both settings, the design matrix \(X \in \mathbb{R}^{n \times p}\) has independent normally distributed rows with \(\mathbb{E}X = 0\) and with covariance matrix \(\Sigma_0 := \Theta_0^{-1}\), where the precision matrix is given by
\[
\Theta_{ij}^0 = 0.3 \text{ if } |i - j| = 1, \Theta_{ij}^0 = 1 \text{ and } \Theta_{ij}^0 = 0 \text{ otherwise}.
\]
The vector of regression coefficients is sparse and is given by \(\beta_0 = (1,1,1,0,\ldots,0) \in \mathbb{R}^p\).
7.2 Confidence intervals and hypothesis testing

Our proposed methodology gives us tools to construct confidence intervals and test hypothesis about the regression parameters. We can also apply multiple testing procedures to test hypotheses about sets of regression coefficients.

We construct confidence intervals using the asymptotic normality of the de-sparsified Lasso $\hat{b}$. In particular, an asymptotic $(1 - \alpha)%$ confidence interval for $\beta_j^0$ ($j = 1, \ldots, p$) can be constructed by

$$\hat{b}_j \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}_j / \sqrt{n},$$

where $\hat{\sigma}_j$ is an estimate of the asymptotic variance of the de-sparsified estimator (see Lemma 2.3 and Section 5). To calculate the de-sparsified estimator, we first need to compute the initial Lasso estimator which is done using the function `glmnet()` (or `cv.glmnet()`) from the R package `glmnet`. The matrix $\Theta$ is estimated using the nodewise Lasso regression. The de-sparsified estimator is calculated as in (5). Asymptotic variance of the de-sparsified estimator is estimated as in Lemma 2.3.

For the confidence intervals, we report average coverages and averages lengths from $N$ independently generated samples. We give the average coverage over the “active” set $S_0 = \{j : \beta_j^0 \neq 0\}$ and average coverage over the “non-active” set $S_0^c$. Similarly, we report average lengths of the confidence intervals over $S_0$ and $S_0^c$.

For testing multiple hypothesis such as $H_0 : \beta_j^0 = 0$ among all $j = 1, \ldots, p$, we will use Bonferroni-Holm multiple testing adjustment to control the family-wise error rate (FWER) or the Benjamini-Hochberg correction for controlling the false discovery rate (FDR) (see Benjamini and Hochberg (1995)).

7.3 Linear regression

In this section, we investigate the performance of the de-sparsified Lasso estimator with different loss functions (square loss, absolute loss and Huber loss) on simulated data. We consider the linear regression setting

$$Y = X\beta_0 + \epsilon,$$

with independent errors, which are independent of the design matrix and have

(1) a Gaussian distribution with zero mean and variance one, or
(2) Student $t_5$-distribution (scaled to have variance equal to one), or
(3) Student $t_3$-distribution (scaled to have variance equal to one).

The construction of the estimators is the same as in Section 5. The tuning parameters were selected by cross-validation for square loss and Huber loss (using R packages `glmnet` and `hqreg`) and for absolute loss, the tuning parameter was selected by the method `qr.fit.lasso` from the R package `quantreg`. We assume that the weights needed to calculate the de-sparsified LAD estimator and de-sparsified Huber estimator (and their asymptotic variances) are known. Their estimation would involve e.g. density estimation and deeper analysis is omitted in this paper. We report the results on confidence intervals in Table 1 and the histograms for the three methods in Figure 1 (for the case of Gaussian error).

7.4 Logistic regression

We analyze the performance of the de-sparsified Lasso for the case of logistic regression,

$$\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = x^T\beta_0, \quad \pi(x) = P(Y = 1 | X = x).$$

The $\ell_1$-penalized estimator is defined via the logistic loss function

$$\rho_\beta(x, y) = -yx^T\beta + \log(1 + e^{x^T\beta}).$$
Asymptotic normality of regression parameters


de-sparsified Lasso


de-sparsified LAD


de-sparsified Huber estimator

Figure 1: Histograms of $\sqrt{n}(\hat{b}_j - \beta_0^j) / \hat{\sigma}_j$, $j = 1, \ldots, 4$ for the de-sparsified LAD, de-sparsified Lasso, and de-sparsified Huber estimator ($K = 0.5$). Here, $n = 500, p = 100, \beta_0 = (1, 1, 1, 0, \ldots, 0)$. The error distribution is $\mathcal{N}(0, 1)$. Superimposed is the density of $\mathcal{N}(0, 1)$ (red curve).
Table 1: A table showing a comparison of the de-sparsified Lasso (D-S Lasso), the de-sparsified LAD estimator (D-S $\ell_1$-LAD) and the de-sparsified Huber estimator with $K = 0.5$ (D-S Huber ($K = 0.5$)). Here, $\beta_0 = (1, 1, 1, 0, \ldots, 0)$. The number of generated random samples was $N = 100$. The nominal coverage is 0.95.

| Gaussian-distributed errors | Coverage | Length |
|----------------------------|----------|--------|
|                            | $p$      | $n$    | $S_0$ | $S_0^c$ | $S_0$ | $S_0^c$ |
| D-S Lasso                  | 100      | 500    | 92.67 | 95.88   | 0.17  | 0.17   |
| D-S $\ell_1$LAD            |          |        | 90.00 | 91.37   | 0.22  | 0.22   |
| D-S Huber ($K = 0.5$)      |          |        | 94.33 | 95.62   | 0.19  | 0.19   |

| Student $t_3$-distributed errors | Coverage | Length |
|----------------------------------|----------|--------|
|                                  | $p$      | $n$    | $S_0$ | $S_0^c$ | $S_0$ | $S_0^c$ |
| D-S Lasso                        | 100      | 500    | 95.00 | 95.94   | 0.17  | 0.17   |
| D-S $\ell_1$LAD                 |          |        | 91.33 | 88.19   | 0.13  | 0.13   |
| D-S Huber ($K = 0.5$)           |          |        | 94.67 | 95.51   | 0.11  | 0.11   |

| Student $t_5$-distributed errors | Coverage | Length |
|----------------------------------|----------|--------|
|                                  | $p$      | $n$    | $S_0$ | $S_0^c$ | $S_0$ | $S_0^c$ |
| D-S Lasso                        | 100      | 500    | 92.00 | 95.79   | 0.17  | 0.17   |
| D-S $\ell_1$LAD                 |          |        | 89.33 | 89.04   | 0.18  | 0.17   |
| D-S Huber ($K = 0.5$)           |          |        | 91.00 | 95.62   | 0.15  | 0.15   |

In the first part of the simulation experiment, we construct confidence intervals using the de-sparsified logistic Lasso as defined in the general formula $\hat{\beta}$. We consider a setting with $p = 100$ and $n = 400$. This also allows us to compare our approach to confidence intervals based on a maximum likelihood estimator. The maximum likelihood estimator is fitted with the function `glm()` in R. The confidence intervals are then calculated using the function `confint.default()`, which bases the confidence intervals on the standard error. The initial logistic Lasso estimator is fitted with `cv.glmnet()` with tuning parameter chosen by cross-validation. The matrix $\Theta$ is estimated by nodewise regression and the tuning parameters chosen by cross-validation.

We plot histograms for the individual entries of the de-sparsified Lasso estimator in Figure 2. For comparison, we also display histograms for the initial Lasso estimator and the maximum likelihood estimator. This demonstrates that the “de-sparsifying” is useful even in the setting where $p < n$. For the confidence intervals, we report average coverages and lengths over the active and non-active set in Table 2. The de-sparsified estimator performs significantly better than the maximum likelihood estimator.

In the second part of the experiment, we look at multiple testing. We consider testing the hypothesis $H_0 : \beta_j = 0$ among all $j = 1, \ldots, p$. We use the Bonferroni-Holm procedure for controlling FWER. From 200 generated samples, the testing procedure had 100% true positive rate and FWER value 0.015.
Figure 2: Histograms of the de-sparsified logistic Lasso (left panel) for $\beta_1, \ldots, \beta_9$. For comparison, histograms of the logistic Lasso (right panel) are also displayed. Here, $n = 800, p = 100$. Even for low-dimensional settings, the de-sparsifying step turns out to be useful.

Table 2: A table showing the average coverages and lengths over the active and non-active set for the de-sparsified logistic Lasso (D-S Logistic Lasso) and the maximum likelihood estimator (MLE). Here, $\beta_0 = (1, 1, 1, 0, \ldots, 0), p = 100$.

| Logistic regression | Coverage | Length |
|---------------------|----------|--------|
|                     | $S_0$    | $S_0^c$ | $S_0$   | $S_0^c$ |
| $n = 400$           | D-S Logistic Lasso 0.817 0.919 0.423 0.402 |
|                     | MLE 0.320 0.891 0.730 0.638 |
| $n = 800$           | D-S Logistic Lasso 0.872 0.932 0.464 0.374 |
|                     | MLE 0.657 0.929 0.433 0.348 |
Bonferroni-Holm adjustment

| Gene | D-S LAD | D-S LASSO |
|------|---------|-----------|
| RPSB_at | 0.05 | RPLX_at | 0.12 |
| YCEI_at | 0.14 | YCEI_at | 0.41 |
| YHCL_at | 0.51 | YNEF_at | 0.57 |
| YFKN_at | 0.61 | |

Benjamini-Hochberg adjustment

| Gene | D-S LAD | D-S LASSO |
|------|---------|-----------|
| RPSB_at | 0.05 | IOLL_at | 0.10 |
| YCEI_at | 0.07 | RPLX_at | 0.10 |
| RPLX_at | 0.36 | YCEI_at | 0.10 |
| YHCL_at | 0.36 | YFKN_at | 0.10 |
| NARI_at | 0.49 | YHCL_at | 0.10 |

Table 3: Variables with smallest p-values among all genes selected by initial screening are reported. D-S LAD corresponds to the de-sparsified Lasso and D-S LASSO corresponds to the de-sparsified LAD estimator. Corresponding p-values are reported next to the genes.

8 Real data experiments

In this section we investigate the practical usefulness of our methodology for gene expression studies which involve high-dimensional data.

8.1 Linear regression: Riboflavin (vitamin B2) production

The dataset riboflavin from the R package hdi contains gene expression levels of 4088 genes and the response variable represents riboflavin (vitamin B2) production. Our goal is to identify genes that significantly effect the production of riboflavin. This dataset is ultra-high-dimensional given that it contains 4088 variables and only 72 observations, but we will reduce it to a moderate high-dimensional data set as for testing such a large number of hypotheses simultaneously turns out to be very conservative. This was also demonstrated in the papers [van de Geer et al., 2014] and [Javanmard and Montanari, 2014], which previously studied this data set. The paper [van de Geer et al., 2014] did not select any gene using the de-sparsified Lasso and the procedure suggested in [Javanmard and Montanari, 2014] selected only two genes: genes YXLD_at and YXLE_at. The works apply (a version) of the de-sparsified Lasso with square loss to select significant variables using a multiple testing adjustment. We also aim to apply the de-sparsified Lasso estimator but in addition we apply the de-sparsified LAD estimator which is expected to be more robust to outliers and to the violation of the normality assumption.

To do initial variable screening, we calculate \( \omega_i := |Y^T X_i| \), \( i = 1, \ldots, 4088 \), where \( X_i \) is the \( i \)-th row of the design matrix and \( Y \) is the response. We select the first 300 variables which have the highest \( \omega_i \)'s. To do the de-sparsified estimator, we fit the initial Lasso estimators with square loss and absolute loss to the data using cross-validation to choose the tuning parameters. To calculate an estimate of \( \Theta \), we use nodewise square-root Lasso from Remark 2.1, which avoids the need to do cross-validation to choose the tuning parameters. We then test the hypotheses: \( H_0 : \beta_j^0 = 0 \), among all \( j = 1, \ldots, 300 \). For multiple testing adjustment, we use two different procedures: the Bonferroni-Holm procedure and the Benjamini-Hochberg procedure. The results are reported in Table 3.
8.2 Logistic regression: Genome-wide studies in cancer

We apply our methodology to a real data set on genome-wide association studies in cancer. The response variable indicating presence or absence of the illness (prostate cancer) is binary, therefore we model the relationship using logistic regression. The dataset contains 102 observations (52 positive, 50 control) on 6033 genes and is available from the R package spls.

We do variable screening as in Section 8.1 to reduce the ultra-high-dimensional data to a more feasible size of 200 genes. The initial logistic Lasso estimator is computed using \texttt{cv.glmnet()} with cross-validation to determine the tuning parameter. The nodewise regression estimator of $\Theta$ is computed using the square-root Lasso as in Remark 2.1.

Using the de-sparsified logistic Lasso and multiple testing adjustment (both Bonferroni-Holm and Benjamini-Hochberg yield the same result), we identify gene number 515 as significant, with a coefficient estimate $\hat{b}_{515} = -2.4677139$. By thresholding the de-sparsified logistic Lasso at the level $2\hat{\sigma}_{j}\sqrt{\log p/n}$, $j = 1, \ldots, p$ we find genes 515, 4639, 5503 significant with coefficients -2.4677139, -1.3019043, -0.7832844, respectively. For a comparison, logistic Lasso identifies 32 genes with non-zero coefficients (including genes 515, 4639, 5503).

For an illustration of the confidence intervals for individual coefficients (without adjustment), see Figure 3.

8.3 Discussion

The simulation study demonstrated that the de-sparsified estimator performs well in a variety of settings for the linear regression and logistic regression, in the setting when $p$ is moderately large. In these settings, the de-sparsified estimator proves to be useful as it clearly outperforms the maximum likelihood estimator. We further observed that multiple testing with the de-sparsified estimator turned out to be conservative and lead to only a few variables selected. However, this is to be expected in the presence of many variables.
References

Belloni, A., Chernozhukov, V., and Kato, K. (2015). Uniform Post Selection Inference for LAD Regression and Other Z-estimation problems. *Biometrika*, 102(1):77–94.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming. *Biometrika*, 98(4):791–806.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data. *Springer*.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics*, 41(3).

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). *Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach*. Annual Review of Economics.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

Janková, J. and van de Geer, S. (2014). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9(1):1205–1229.

Janková, J. and van de Geer, S. (2016). Honest confidence regions and optimality for high-dimensional precision matrix estimation. *TEST*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

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

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462.

Nickl, R. and van de Geer, S. (2012). Confidence sets in sparse regression. *Annals of Statistics*, 41(6):2852–2876.

van de Geer, S. (2014). Worst possible sub-directions in high-dimensional models. *ArXiv: 1403.7023*.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

van de Geer, S. and Stucky, B. (2016). \(\chi^2\)-Confidence Sets in High-Dimensional Regression. *Volume 11 of the series Abel Symposia*, pages 279–306.

van der Vaart, A. (2000). Asymptotic statistics. *Cambridge University Press*.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York.
Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76:217–242.
A Appendix

The section is organized as follows.

1. Several preliminary results are stated in Section A.1.
2. Proofs for Section 2 (generalized linear models) can be found in Section A.3.
3. Proofs for Section 4 (general models) are contained in Section A.2.
4. Proofs for Section 3 (nodewise regression) are given in Section A.4.
5. Proofs for Section 5 (examples) are contained in Section A.5.

For two sequences, \( f_n, g_n \), we write \( f_n \lesssim g_n \) if there exists a constant \( C > 0 \) such that \( f_n \leq C g_n \) for all \( n \).

A.1 Preliminary material

We define

\[
Z(F) := \sup_{f \in F} |(P_n - \mathbb{P})f|,
\]

and

\[
Z'(F) := \sup_{f \in F} \sum_{i=1}^{n} \epsilon_i f(x_i).
\]

**Theorem A.1** (see e.g. van der Vaart (2000)).

\[
EZ(F) \leq 2EZ'(F).
\]

**Theorem A.2** (Dudley’s inequality).

\[
EZ'(F) \leq C_0 \inf_{\delta \geq 0} \mathbb{E} \left[ \hat{R} \int_{\delta}^{1} \sqrt{\log N(u\hat{R}, \mathcal{F}, \| \cdot \|_n)} du / \sqrt{n} + \delta \hat{R} \right],
\]

where \( \hat{R} := \sup_{f \in \mathcal{F}} \| f \|_n \).

For reader’s convenience, we recall the Nemirovski inequality.

**Theorem A.3** (Nemirovski inequality, see e.g. Bühlmann and van de Geer (2011)). For \( m \geq 1 \) and \( p \geq e^{m-1} \), we have

\[
\mathbb{E} \max_{1 \leq j \leq p} \left( \sum_{i=1}^{n} \gamma_j(z_i) - \mathbb{E}\gamma_j(Z_i) \right)^m \leq (8 \log(2p))^{m/2} \mathbb{E} \left( \max_{1 \leq j \leq p} \sum_{i=1}^{n} \gamma_j^2(z_i) \right)^{m/2}.
\]

A.2 Proofs for Section 4 (General high-dimensional models)

**Proof of Theorem 4.1** Consider the decomposition

\[
\hat{\beta}_j - \beta_0 - \Theta_j \mathbb{P}n \psi_{\beta_0} = \Theta_j \mathbb{P}n \psi_{\beta_0} - \Theta_j (\mathbb{P}n - P)(\psi_{\beta} - \psi_{\beta_0}) + \hat{\beta}_j - \beta_0 - \Theta_j P(\psi_{\beta} - \psi_{\beta_0}).
\]

By assumption, we have \( i = \Theta_j (\mathbb{P}n - P)(\psi_{\beta} - \psi_{\beta_0}) = o_P(1/\sqrt{n}) \). Next we treat the term \( ii \). By condition (D1) we have \( \| \Theta_j \|_1 \leq \sqrt{s} \mathcal{O}(1) \). Condition (D2) and assumption \( d^2(\hat{\beta}, \beta_0) = o_P \left( \frac{1}{\sqrt{n}} \right) \) then yield

\[
ii = \hat{\beta}_j - \beta_0 - \Theta_j P(\psi_{\beta} - \psi_{\beta_0}) = \mathcal{O}(\| \Theta_j \|_1 d^2(\hat{\beta}, \beta_0)) = \mathcal{O}_p(\sqrt{s}d^2(\hat{\beta}, \beta_0)) = o_P(1/\sqrt{n}).
\]
Hence we conclude
\[ \hat{\beta}_j - \beta^0_j - \Theta_j T F_n \psi_\beta = -\Theta_j T F_n \psi_{\beta_0} + o_P(1/\sqrt{n}). \tag{18} \]

The sparsity condition (10) and condition (D3) imply
\[ \| (\hat{\Theta}_j - \Theta_j)^T F_n \psi_\beta \|_\infty \leq \| \hat{\Theta}_j - \Theta_j \|_1 \| F_n \psi_\beta \|_\infty = O_P(s \lambda^2) = o_P(1/\sqrt{n}). \]

Hence (18) holds with \( \Theta_j^T F_n \psi_{\beta_0} \) replaced by \( \hat{\Theta}_j^T F_n \psi_\beta \). That means
\[ \hat{\beta}_j - \beta^0_j - \hat{\Theta}_j^T F_n \psi_\beta = -\Theta_j^T F_n \psi_{\beta_0} + o_P(1/\sqrt{n}). \tag{19} \]

By Condition (D5) the term \( \Theta_j^T F_n \psi_{\beta_0} \) is asymptotically normal \( \mathcal{N}(0,1) \) when normalized by the square-root of variance
\[ \text{Var}(\Theta_j^T F_n \psi_{\beta_0}) = \Theta_j^T F \psi_{\beta_0}^T \Theta_j / n. \]

The result follows.

**Proof of Theorem 4.2.** By Theorems A.1 and A.2 (upper-bounding the entropy integral by taking its lower bound equal to zero), it follows that
\[ \mathbb{E} \sup_{f \in F} |(F_n - P) f| \leq 2 C_0 \mathbb{E} \left[ \int_0^\hat{R} \sqrt{\log N(u, F, \| \cdot \|_n) \| \cdot \|_n} du / \sqrt{n} \right], \]
where \( \hat{R} := \sup_{f \in F} \| f \|_n \). Let \( F(x) = \sup_{f \in F} |f(x)| \) be the envelope function of \( F \). Note that \( \hat{R} \leq \| F \|_n \) by the definition of the envelope function \( F \). Moreover, note that \( N(u, F, \| \cdot \|_n) \leq N(u/4, F, \| \cdot \|_n) \). Thus, and by the assumed entropy condition, we obtain
\[ \mathbb{E} \sup_{f \in F} |(F_n - P) f| \leq 2 C_0 \left[ \int_0^\hat{R} \sqrt{\log (u/4, F, \| \cdot \|_n) \| \cdot \|_n} du / \sqrt{n} \right], \]
\[ \leq 2 C_0 \left[ \sqrt{\log p + \log (4\| F \|_n / u)} \right] du / \sqrt{n} \]
\[ \leq 2 C_0 \left[ \sqrt{\log (4\| F \|_n / u)} \right] du / \sqrt{n}, \]
where we also used that for any \( x, y > 0 \) it holds that \( \sqrt{x+y} \leq \sqrt{x} + \sqrt{y} \). One can show the following upper bound using integration per partes:
\[ \int_0^\delta \log(1/u) du = \delta \log \left( \frac{1}{\delta} \right) \leq \frac{1}{\log \frac{1}{\delta}} \cdot \frac{1}{\log \frac{1}{\delta}}. \]

Then for \( \delta \leq 1/4 \) it holds that \( \frac{1}{\log \frac{1}{\delta}} \leq \frac{1}{\log \frac{1}{\delta}} = C_1 \), whence \( \int_0^\delta \log(1/u) du \leq C_1 \delta \log \frac{1}{\delta} \). We then have
\[ \int_0^\hat{R} \sqrt{\log (4\| F \|_n / u)} du \leq 4\| F \|_n \int_0^\hat{R} / (4\| F \|_n) \sqrt{\log(1/u)} du \]
\[ \leq 4 C_1 \| F \|_n \hat{R} / (4\| F \|_n) \log \left( \frac{4\| F \|_n}{\hat{R}} \right) \]
\[ = C_1 \hat{R} \log \left( \frac{4\| F \|_n}{\hat{R}} \right). \]

We now show that \( \| F \|_n \leq \sqrt{n} \). This follows since
\[ \| F \|_n^2 = \frac{1}{n} \sum_{i=1}^n F(X_i)^2 \leq \max_{f \in F} \sup_{f \in F} F(X_i)^2, \]
and
\[ \hat{R}^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 \geq \sup_{f \in \mathcal{F}} \max_{i=1,\ldots,n} f(X_i)^2 / n. \]

Hence we obtain
\[ \int_{0}^{\hat{R}} \sqrt{\log(4\|F\|_n/u)} du \leq C_1 \hat{R} \log(4\sqrt{n}). \]

Then it follows
\[ \mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f| \leq 2C_0 \mathbb{E} \hat{R} \left( \sqrt{s \log p/n} + C_1 \log(4\sqrt{n}) \sqrt{s/n} \right) \leq C_3 \mathbb{E} \hat{R} \log(4\sqrt{n}) \sqrt{s \log p/n} \]

We next apply the Dudley’s inequality to the class \( \mathcal{F}^2 := \{ f^2 : f \in \mathcal{F} \} \). First observe that
\[ \log N(\epsilon, \mathcal{F}^2, \| \cdot \|_n) \leq \log N(\epsilon / \|F\|_{n,\infty}, \mathcal{F}, \| \cdot \|_n) \leq s \log p + s \log \left( \frac{\|F\|_{n,\infty}}{\epsilon} \right). \]  

Let \( \hat{R}_2^2 := \sup_{f \in \mathcal{F}^2} \| f \|_n^2 = \sup_{f \in \mathcal{F}^2} \frac{1}{n} \sum_{i=1}^{n} (2f(X_i))^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)^4 \). Then since
\[ \mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2| = \mathbb{E} \sup_{f \in \mathcal{F}^2} |(\mathbb{P}_n - P)f|, \]

we obtain by Dudley’s inequality and the entropy bound (20) for \( \mathcal{F}^2 \) that
\[ \mathbb{E} \sup_{f \in \mathcal{F}^2} |(\mathbb{P}_n - P)f| \leq \mathbb{E} \int_{0}^{\hat{R}_2} \sqrt{s \log p + s \log(4\|F\|_n\|F\|_{n,\infty}/u)} du / \sqrt{n} \leq \mathbb{E} \hat{R}_2 \sqrt{s \log p/n} + \mathbb{E} \int_{0}^{\hat{R}_2} \sqrt{s \log(4\|F\|_n\|F\|_{n,\infty}/u)} du / \sqrt{n} \]

We then have
\[ i = \int_{0}^{\hat{R}_2} \sqrt{\log(\|F\|_n\|F\|_{n,\infty}/u)} du \]
\[ = 4\|F\|_n\|F\|_{n,\infty} \int_{0}^{\hat{R}_2/(4\|F\|_n\|F\|_{n,\infty})} \sqrt{\log(1/u)} du \]
\[ \leq \|F\|_n\|F\|_{n,\infty} \hat{R}_2 / \|F\|_{n,\infty} \log \left( \frac{4\|F\|_n\|F\|_{n,\infty}}{\hat{R}_2} \right) \]
\[ = \hat{R}_2 \log \left( \frac{4\|F\|_n\|F\|_{n,\infty}}{\hat{R}_2} \right). \]

We now show that \( \frac{\|F\|_n\|F\|_{n,\infty}}{\hat{R}_2} \leq \sqrt{n} \). To this end observe that \( \|F\|_n \leq \|F\|_{n,\infty} \), and
\[ \|F\|_{n,\infty}^2 \leq \max_{i=1,\ldots,n} \sup_{f \in \mathcal{F}} f(X_i)^2, \]

and moreover
\[ \hat{R}_2^2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)^4 \geq \sup_{f \in \mathcal{F}} \max_{i=1,\ldots,n} f(X_i)^4 / n. \]

Then it follows that
\[ \frac{\|F\|_n\|F\|_{n,\infty}}{\hat{R}_2} \leq \frac{\|F\|_{n,\infty}^2}{\hat{R}_2} \leq \sqrt{n}. \]
Hence we obtain
\[
i = \int_0^{\hat{R}_2} \sqrt{\log(4\|F\|_n\|F\|_{n,\infty}/u)} \, du \leq \hat{R}_2 \log(4\sqrt{n}).
\]
Thus
\[
\mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2| \leq \mathbb{E} \hat{R}_2 \log(4\sqrt{n}) \sqrt{s \log p/n}
\]
By the triangle inequality, we obtain
\[
\hat{R}_2 \leq \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2| + \sup_{f \in \mathcal{F}} Pf^2.
\]
Then
\[
\mathbb{E} \hat{R} \leq \mathbb{E} \left( \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2| + \sup_{f \in \mathcal{F}} Pf^2 \right) \leq \mathbb{E} \left( \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2| + \sqrt{\sup_{f \in \mathcal{F}} Pf^2} \right).
\]
Using Jensen’s inequality, we have
\[
\mathbb{E}(\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2|)^{1/2} \leq (\mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2|)^{1/2}.
\]
Hence
\[
\mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f| \leq \mathbb{E} \hat{R} \log n \sqrt{s \log p/n}
\]
\[
\leq \left( (\mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f^2|)^{1/2} + \sup_{f \in \mathcal{F}} (\mathbb{E} f^2)^{1/2} \right) \log n \sqrt{s \log p/n}
\]
\[
\leq \mathbb{E} \hat{R}_2^{1/2} (\log n \sqrt{s \log p/n})^{1/2} + \sup_{f \in \mathcal{F}} \sqrt{\mathbb{E} f^2} \log n \sqrt{s \log p/n}
\]
\[
\leq \mathbb{E} \hat{R}_2^{1/2} (\log n)^{3/2} (\log p/n)^{1/4} \sqrt{s \log p/\sqrt{n}}
\]
\[
+ \sup_{f \in \mathcal{F}} \sqrt{\mathbb{E} f^2} \log n \sqrt{s \log p/n}.
\]
Next observe that by Jensen’s inequality
\[
\mathbb{E} \sup_{f \in \mathcal{F}} \|f\|_{n,4} = \mathbb{E} \left( \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i)^4 \right)^{1/4}
\]
\[
\leq \left( \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i)^4 \right)^{1/4}
\]
\[
= \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E} \sup_{f \in \mathcal{F}} f(X_i)^4 \right)^{1/4}.
\]
By assumption, we have \(\max_{i=1,\ldots,n} (\mathbb{E} \sup_{f \in \mathcal{F}} f(X_i)^4)^{1/4} (\log n \sqrt{s \log p/n})^{3/2} = o_P(1/\sqrt{n})\) and
\[
\sqrt{\mathbb{E} f^2} \log n \sqrt{s \log p/n} = o(1/\sqrt{n}).
\]
Hence
\[
\mathbb{E} \sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f| = o(1/\sqrt{n}).
\]
Then
\[ \mathbb{E} \sup_{f \in \mathcal{F}} |G_n f| = o(1). \]
By Markov’s inequality it follows that \( \sup_{f \in \mathcal{F}} |G_n f| = o_P(1). \)

## A.3 Proofs for Section 2 (High-dimensional generalized linear models)

We first need the following preliminary lemmas before proving the statement of Theorem 2.1.

### A.3.1 Preliminary lemmas

**Lemma A.1.** Consider the class
\[ \mathcal{F} := \{ \Theta_j^T x(\mathbf{w}_\beta - \mathbf{w}_{\beta_0}) : \| \beta - \beta_0 \|_1 \leq s\lambda, \mathbb{E}(x^T(\beta - \beta_0))^2 \leq s\lambda^2 \}. \]
Assume conditions [A1], [B2]. Then
\[ \sup_{f \in \mathcal{F}} \| f \| = O \left( s^{3/4} \lambda \right). \]

**Proof of Lemma A.1.** By the Lipschitz property in condition [B2], we have \( |w_{\beta} - w_{\beta_0}| \leq L|x^T(\beta - \beta_0)| \). By condition [A1] and by Hölder’s inequality, we obtain \( |x(\beta - \beta_0)| \leq K_X \| \beta - \beta_0 \|_1 \). By condition [A1], we further have \( \mathbb{E}(\Theta_j^T x)^4 = O(1) \). Hence
\[ P f^2 = \mathbb{E}(\Theta_j^T x(w_{\beta} - w_{\beta_0}))^2 \leq (\mathbb{E}(\Theta_j^T x)^4)^{1/2}(\mathbb{E}(w_{\beta} - w_{\beta_0})^4)^{1/2} \leq (\mathbb{E}(\Theta_j^T x)^4)^{1/2}(\mathbb{E}(w_{\beta} - w_{\beta_0})^4)^{1/2} \leq O(1)s\lambda(\mathbb{E}(x^T(\beta - \beta_0))^2)^{1/2} = O \left( s^{3/2} \lambda^2 \right). \]

**Lemma A.2.** Consider the class
\[ \mathcal{F} := \{ \Theta_j^T x(\mathbf{w}_\beta - \mathbf{w}_{\beta_0}) : \| \beta - \beta_0 \|_1 \leq s\lambda, \mathbb{E}(x^T(\beta - \beta_0))^2 \leq s\lambda^2 \}, \]
where \( \lambda \asymp \sqrt{\log p/n} \). Suppose that condition [B2] is satisfied, assume that \( P|\Theta_j^T x|^4 = O(1) \) and \( s^3(\log p)^2(\log n)^2/n = o(1) \). Then the condition
\[ P \sup_{f \in \mathcal{F}} f^4 s^3(\log p)^3(\log n)^6 = o(1) \]
is satisfied.

**Proof of Lemma A.2.** By Lipschitz property of \( w_{\beta} \) in condition [B2], we have \( |w_{\beta} - w_{\beta_0}| \leq L|x(\beta - \beta_0)| \), and hence
\[ P \sup_{f \in \mathcal{F}} f^4 \leq P|\Theta_j^T x|^4 \sup_{f \in \mathcal{F}} |w_{\beta} - w_{\beta_0}|^4 \leq P|\Theta_j^T x|^4 L\| \beta - \beta_0 \|_1^4 \leq O(1)(s\lambda)^4. \]
Then under \( s^3(\log p)^2(\log n)^2/n = o(1) \), the claim follows. \( \square \)
Lemma A.3. Assume conditions [A1] and [C1]. Then the Condition [D2] is satisfied with $d^2(\beta, \beta_0) = \mathbb{E}[x^T(\beta - \beta_0)^2$.

Proof of Lemma A.3 First note that $P_{\beta} = \int x \int w(y, x \beta) dP_{\epsilon|x} dP_{\epsilon}$ and $(P_{\beta})'_{\beta} = \int x G'(u) u=\beta dP_{X}$.

\[
P(\psi_\beta(y, x) - \psi_\beta(y, x)) = P(w(y, x^T \beta) - w(y, x^T \beta_0))x = \int (w(y, x^T \beta) - w(y, x^T \beta_0)) x dP = \int x \int (w(y, x^T \beta) - w(y, x^T \beta_0)) dP_{\epsilon|x} dP_{\epsilon} = \int x G'(x^T \beta_0) x^T (\beta - \beta_0) dP_{X} + \int x (G'(\tilde{x}) - G'(x^T \beta_0)) x^T (\beta - \beta_0) dP_{X}
\]

where $\tilde{x} \in [x^T \beta_0, x^T \beta]$. Then for the remainder we have

\[
\|\text{rem}\|_{\infty} = \|\int x (G'(\tilde{x}) - G'(x^T \beta_0)) x^T (\beta - \beta_0) dP_{X}\|_{\infty} \\
\leq \int \|x\|_{\infty} |G'(\tilde{x}) - G'(x^T \beta_0)| |x^T (\beta - \beta_0)| dP_{X} \\
\leq \int K_{X} L |x^T (\beta - \beta_0)|^2 dP_{X}
\]

\[
\Box
\]

Lemma A.4. Denote $h(\beta) := P_{\beta}$. Assume that $\beta \mapsto \partial^2 (P_{\beta})_{L}$ is bounded, i.e. $\frac{\partial^2 h_j}{\partial \beta_k \partial \beta_\ell}$ is bounded for all $j = 1, \ldots, p$. Then Condition [D2] is satisfied with $d(\beta, \beta_0) = \|\beta - \beta_0\|_1$.

Proof of Lemma A.4 By the mean-value theorem,

\[
(P_{\beta} - P_{\beta_0})_j = ((P_{\beta_0})_j)'_{\beta=\beta_0}(\beta - \beta_0) + ((P_{\beta_0})_j)'_{\beta=\tilde{\beta}} - ((P_{\beta_0})_j)'_{\beta=\beta_0}(\beta - \beta_0)
\]

Then for the remainder, we obtain

\[
\|((P_{\beta_0})_j)'_{\beta=\beta} - ((P_{\beta_0})_j)'_{\beta=\beta_0}(\beta - \beta_0)\|_1 \\
\leq \|((P_{\beta_0})_j)'_{\beta=\beta} - ((P_{\beta_0})_j)'_{\beta=\beta_0}\|_1 \|\beta - \beta_0\|_1
\]

Finally,

\[
|((P_{\beta_0})_j)'_{\beta=\tilde{\beta}} - ((P_{\beta_0})_j)'_{\beta=\beta_0}| = \|((P_{\beta_0})_j)'_{\beta=\tilde{\beta}}(\beta - \beta_0)\|_1 \\
\leq \||((P_{\beta_0})_j)'_{\beta=\tilde{\beta}}\|_1 \|\beta - \beta_0\|_1 \\
\leq L \|\beta - \beta_0\|_1
\]

\[
\Box
\]

Proof of Lemma 2.4 By the Lipschitz property [B2] it follows that

\[
|w_{\beta} - w_{\beta_0}| \leq L |x^T(\beta - \beta_0)|,
\]

hence

\[
N(F, \|\cdot\|_n, \epsilon) \leq N(H, \|\cdot\|_n, \epsilon/L),
\]

29
where $H = \{x \mapsto \Theta^T_{\beta}xx^T(\beta - \beta_0) : \|\beta\|_0 \leq Cs\}$. Let $V_i \subset \{1, \ldots, p\}$ for $i = 1, \ldots, \binom{p}{s}$ be all subsets of $\{1, \ldots, p\}$ of size $s$. We can rewrite

$$H = \bigcup_{i=1}^{\binom{p}{s}} H_i,$$

where $H_i := \{(x, y) \mapsto \Theta^T_{\beta}xx^T(\beta - \beta_0) : \beta \in \mathbb{R}^p, \beta_{V_i} = 0\}$. The collection $H_i$ has the same VC-index as an $s$-dimensional real vector space, which is $s + 2$ by Lemma 2.6.15 in [van der Vaart and Wellner (1996)].

Then by Theorem 2.6.7 in [van der Vaart and Wellner (1996)] and since the covering number of a union of sets is upper bounded by sum of the covering numbers, we obtain

$$N(\epsilon \|H\|_n, H, \|\cdot\|_n) \leq \binom{p}{s} N(\epsilon \|H\|_n, H_i, \|\cdot\|_n) \leq \binom{p}{s} Ks(16\epsilon)^s \left(\frac{1}{\epsilon}\right)^{2(s-1)},$$

where $K$ is a universal constant and $0 < \epsilon < 1$. Then

$$\binom{p}{s} Ks(16\epsilon)^s \left(\frac{1}{\epsilon}\right)^{2(s-1)} \leq \frac{p^s}{s!} Ks(16\epsilon)^s \left(\frac{1}{\epsilon}\right)^2 \leq \frac{p^s}{(s-1)!} K \left(\frac{16\epsilon}{\epsilon}\right)^{2s} \leq Kp^s \left(\frac{16\epsilon}{\epsilon}\right)^{2s}.$$

Hence

$$N(H, \|\cdot\|_n, \epsilon \|H\|_n) \lesssim \left(\frac{p}{\epsilon}\right)^s.$$

Since $\|F\|_n \leq \|H\|_n$ and by the Lipschitz property of $w_\beta$ we have

$$N(F, \|\cdot\|_n, \epsilon \|F\|_n) \leq N(F, \|\cdot\|_n, \epsilon \|H\|_n) \leq N(H, \|\cdot\|_n, \epsilon \|H\|_n/L) \lesssim \left(\frac{p}{\epsilon}\right)^s.$$
Furthermore, by Lemma 2.3, we have

\[ (\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0) = O(s\lambda^2) = o(1/\sqrt{n}) \]

under the condition \( s^3(\log p)^2/n = o(1) \).

Condition (D3) is satisfied under condition (A2).

Condition (D4). We now show that the entropy condition of Theorem 4.2 is satisfied. Consider the class of functions

\[ \mathcal{F} = \{ \Theta^T_j x(w_{\beta} - w_{\beta_0}) : \| \beta - \beta_0 \|_1 \leq s\lambda, \| \beta \|_0 \leq s, E|x^T(\beta - \beta_0)|^2 \leq s\lambda^2 \}, \]

where \( \lambda \asymp \sqrt{\log p/n} \). Under the condition (A2), it follows that \( f_\beta \in \mathcal{F} \) with high probability.

We proceed to check the entropy condition of Theorem 4.2. Under condition (B2), Lemma 2.1 implies that the entropy bound is satisfied for the class \( \mathcal{F} \). Finally, we check condition (B1) of Theorem 4.2. By Lemma A.1 it follows that \( R = O(s^{3/2}/\sqrt{\log p/n}) \).

Hence under \( s^3(\log p)^2(\log n)^2/n = o(1) \) it holds that \( R \sqrt{s\log p\log n} = o(1) \). The condition

\[ P \sup_{f \in \mathcal{F}} f^4 s^3(\log p)^3(\log n)^6/n = o(1) \]

is satisfied under the sparsity condition \( s^3(\log p)^2(\log n)^2/n = o(1) \) by Lemma A.2.

Condition (D5) by the assumption \( E(\Theta^T_j X_j)^4 = O(1) \) and by (D5), we can apply the central limit theorem to conclude the asymptotic normality as required in condition (D5).

The above implies that

\[ \sqrt{n}(\hat{\beta}_j - \beta_0)/\sqrt{\Theta^T_j P \psi_{\beta_0} \psi^T_{\beta_0} \Theta_j} \sim N(0,1), \]

which shows the first claim of the theorem.

Next by Theorem 3.1 it follows that

\[ \| \hat{\Theta}_j - \Theta_j \|_1 = O_P(s^{3/2}/\sqrt{\log p/n}). \]

Furthermore, by Lemma 2.3 we have

\[ |\Theta^T_j P_n \psi_{\beta} \psi^T_{\beta} \hat{\Theta}_j - \Theta^T_j P \psi_{\beta_0} \psi^T_{\beta_0} \Theta_j| = o_P(1). \]

But then

\[ (\hat{\beta} - \beta_0 - \hat{\Theta}^T_j P_n \psi_{\beta_\hat{\beta}})/\sqrt{\hat{\Theta}^T_j P_n \psi_{\beta_\beta} \hat{\Theta}^T_j} = (\hat{\beta} - \beta_0 - \Theta^T_j P_n \psi_{\beta_\beta})/\sqrt{\Theta^T_j P \psi_{\beta_0} \psi^T_{\beta_0} \Theta_j} - (\hat{\Theta}_j - \Theta_j)^T P_n \psi_{\beta} \psi^T_{\beta} \hat{\Theta}_j. \]

\[ \square \]

Proof of Theorem 2.2. We apply Theorems 4.1 and 4.2. Under condition (C1) by Lemma A.3 it follows that Condition (D2) is satisfied with \( d^2(\hat{\beta}, \beta_0) = E|x^T(\beta - \beta_0)|^2 \). Then \( d^2(\hat{\beta}, \beta_0) = (\hat{\beta} - \beta_0)^T \Sigma (\hat{\beta} - \beta_0) = O(s\lambda^2) = o(1/\sqrt{n}) \), under the condition \( s^3(\log p)^2/n = o(1) \).

By inspection of proof of Lemma A.1 we have by boundedness of \( w_\beta \) (condition (C1)) that \( |w_\beta - w_{\beta_0}| \leq |w_\beta| + |w_{\beta_0}| = O(1) \) and hence

\[ Pf^2 = E (\Theta^T_j x(w_\beta - w_{\beta_0}))^2 \leq (E|\Theta^T_j x|^4)^{1/2} (E|w_\beta - w_{\beta_0}|^4)^{1/2} \leq (E|\Theta^T_j x|^4)^{1/2} (E|w_\beta - w_{\beta_0}|^4)^{1/2} \leq O(1)(E|w_\beta - w_{\beta_0}|^2)^{1/2}. \]

By condition (C2) we have \( E|w_\beta - w_{\beta_0}|^2 = O \left( \frac{1}{s^3(\log p)^2(\log n)^2} \right) \) which implies that \( R \log n \sqrt{s\log p} = o(1) \).
We have
\[ \max_{i=1, \ldots, n} \mathbb{E} \sup_{f \in \mathcal{F}} f^4(x_i, Y_i) \leq \max_{i=1, \ldots, n} \mathbb{E}[\Theta_j^T x_i]^4 \sup_{f \in \mathcal{F}} |w_\beta(x_i, Y_i) - w_{\beta_0}(x_i, Y_i)|^4 \]
\[ \leq \max_{i=1, \ldots, n} \mathbb{E}[\Theta_j^T x_i]^4 \leq O(1), \]
where we used the boundedness of \( w_\beta \) from condition \([C1]\) and the assumption
\[ \max_{i=1, \ldots, n} \mathbb{E}[\Theta_j^T x_i]^4 = O(1) \]
from condition \([A1]\). The condition
\[ \max_{i=1, \ldots, n} \mathbb{E} \sup_{f \in \mathcal{F}} f(x_i, Y_i)^4 s^3(\log p)^3(\log n)^6/n = o(1) \]
is then satisfied under \( s^3(\log p)^3(\log n)^6/n = o(1) \).

\[ \square \]

**A.3.2 Proofs for Section 2.2.4: Sparsity of the Lasso**

**Proof of Lemma 2.3** The KKT conditions for \( \hat{\beta} \) give
\[ \mathbb{P}_n \psi_\beta + \lambda \hat{Z} = 0, \]
where \( \psi_\beta(y_i, x_i) = w(y_i, x_i^T \beta) \). This can be rewritten as
\[ \mathbb{P}_n (\psi_\beta - \psi_{\beta_0}) = -\lambda \hat{Z} - \mathbb{P}_n \psi_{\beta_0}. \quad (21) \]
Then we further separate the empirical process part
\[ P(\psi_\beta - \psi_{\beta_0}) = -\lambda \hat{Z} - \mathbb{P}_n \psi_{\beta_0} - (\mathbb{P}_n - P)(\psi_\beta - \psi_{\beta_0}). \quad (22) \]
Taking the \( \ell_2 \)-norm of the left-hand side of (21) and by the mean-value theorem we obtain
\[ \|P(\psi_\beta - \psi_{\beta_0})\|_2^2 = \|P_x(w(y_i, x_i^T \tilde{\beta}) - w(y_i, x_i^T \beta_0))\|_2^2 \]
\[ = \|P_x G'(y_i, x_i^T \beta) x_i x_i^T (\tilde{\beta} - \beta_0)\|_2^2 \]
\[ = (\tilde{\beta} - \beta_0)^T (P_x G'(y_i, x_i^T \tilde{\beta}) x_i x_i^T)^2 (\tilde{\beta} - \beta_0) \]
For all \( u \in \mathbb{R}^p \) we have (since \( G'(y_i, x_i^T \tilde{\beta}) \geq 0 \) by assumption of convexity of the loss function)
\[ u^T P_x G'(y_i, x_i^T \tilde{\beta}) x_i x_i^T u \leq C u^T P_x x_i x_i^T u \leq C \Lambda_{\max}(P_x x_i x_i^T) u^T u \leq C_2 u^T u, \]
for some constant \( C_2 > 0 \). Thus it must necessarily hold that \( \Lambda_{\max}(P_x G'(y_i, x_i^T \tilde{\beta}) x_i x_i^T) \leq C_2 \). But then we obtain
\[ \|P(\psi_\beta - \psi_{\beta_0})\|_2^2 \leq \Lambda_2^2 \Lambda_{\max}(P_x G'(y_i, x_i^T \tilde{\beta}) x_i x_i^T) \|\tilde{\beta} - \beta_0\|_2^2 \]
\[ = C_2 |\tilde{\beta} - \beta_0|_2^2 \]
\[ = O_P(|\tilde{\beta} - \beta_0|_2^2), \]
where \( W_{\beta_0} := \text{diag}(w_{\beta_0}(x_1, y_1), \ldots, w_{\beta_0}(x_n, y_n)) \). Next we consider the right-hand side of (21). First, by equation (9), which is assumed in the conditions, we have for the empirical process part
\[ (\mathbb{P}_n - P)(\psi_\beta - \psi_{\beta_0}) = O_P(\lambda_0), \]
for \( \lambda_0 \approx \sqrt{\log p/n} \). This follows analogously as in the proof of Theorem 2.2. We further have,
\[ \mathbb{P}_n \psi_{\beta_0} = O_P(\lambda_0). \]
Denote \( \hat{s} := \| \hat{\beta} \|_0 \). Then taking the \( \ell_2 \)-norm of the right-hand side of (21)

\[
\| - \lambda \hat{Z} - P_n \psi_{\beta_0} - (P_n - P)(\psi_\beta - \psi_{\beta_0}) \|_2^2 \geq (\lambda - \lambda_0)^2 \hat{s}.
\]

Hence we obtain

\[
\hat{s} \leq \mathcal{O}_P(\| W_{\beta_0} X (\hat{\beta} - \beta_0) \|_2^2) \leq \mathcal{O}_P(s).
\]

\[\square\]

### A.3.3 Proofs for Section 2.2.5: Estimation of asymptotic variance

#### Proof of Lemma 2.3

For the first term, we have by Hölder’s inequality

\[
| \hat{\Theta}_j^T \hat{\beta} \hat{\Theta}_j - \hat{\Theta}_j^T P \psi_{\beta_0} \hat{\Theta}_j | \leq \| \hat{\Theta}_j \|_2^2 |(P_n - P)\psi_{\beta_0} \|_\infty = \mathcal{O}_P(s \sqrt{\log p/n}).
\]

For the second term, we have

\[
| P(\psi_{\beta_0} \psi_{\beta_0}^T - \hat{\beta} \hat{\beta}^T ) | = | P(w^2_{\beta} - w^2_{\beta_0}) xx^T | \leq \| x \|_{\infty}^2 | P(w^2_{\beta} - w^2_{\beta_0}) | = \mathcal{O}(\mathbb{E} x^T (\hat{\beta} - \beta_0)) = \mathcal{O}_P(\| (\mathbb{E} x^T (\hat{\beta} - \beta_0) )^2 \|^{1/2}) = \mathcal{O}_P(\sqrt{s} \lambda).
\]

Then

\[
| i j | \leq \| \hat{\Theta}_j \|_1^2 \| x \|_{\infty}^2 | P(w^2_{\beta} - w^2_{\beta_0}) | = \mathcal{O}_P(s \sqrt{s} \lambda) = \mathcal{O}_P(s^{3/2} \lambda) = \omega_P(1).
\]

\[\square\]

### A.4 Proofs for section 3 (Nodewise regression for estimation of precision matrices)

#### Lemma A.5

Suppose that conditions (A1), (A2), (E1), (E2), (E3) are satisfied. Let \( \lambda \asymp \sqrt{\log p/n} \). Then it holds that

\[
\| \hat{\gamma}_{\beta,j} - \gamma_{0,j} \|_1 = \mathcal{O}_P(s^{3/2} \sqrt{\log p/n}).
\]

#### Proof of Lemma A.5

We denote \( W_{\beta} := \text{diag}(v_3(y_1, x_1), \ldots, v_3(y_n, x_n)) \). Further denote by

\[
\eta_{\beta_0,j} := W_{\beta_0}(X_j - X_{-j}\gamma_{0,j}).
\]

We have the basic inequality

\[
\| X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) \|_2^2 \leq n + 2 \lambda_j \| \hat{\gamma}_{\beta,j} \|_1 \leq 2 \eta_{\beta_0,j} W_{\beta_0}^{-1} W_{\beta} X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j})/n + 2 \lambda_j \| \gamma_{\beta_0,j} \|_1.
\]

First we have by the Cauchy-Schwarz inequality

\[
| \eta_{\beta_0,j} W_{\beta_0}^{-1} W_{\beta} X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j})/n - \eta_{\beta_0,j} W_{\beta_0}^{-1} X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j})/n | \leq \| (W_{\beta}^2 W_{\beta}^{-2} - I) \eta_{\beta_0,j} \|_2 \sqrt{n} \| X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) \|_2 \sqrt{n}
\]

\[
| (W_{\beta}^2 W_{\beta}^{-2} - I) \eta_{\beta_0,j} \|_2 \sqrt{n} \| X_{-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) \|_2 \sqrt{n}
\]
We bound the term $\|(W_\beta^2 W_\beta^{-2} - I)\eta_{j,0,j}\|^2_2/n$. We have by the Cauchy-Schwarz inequality

$$\|(W_\beta^2 W_\beta^{-2} - I)\eta_{j,0,j}\|^2_2/n = \frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 v_{\beta,i}^{-2} - 1)\eta_{j,0,i}^2 \leq \left( \frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 v_{\beta,i}^{-2} - 1) \right)^4 \left( \frac{1}{n} \sum_{i=1}^n \eta_{j,0,i}^4 \right).$$

By condition (A1) and (E1) we have $E\eta_{j,0,i}^4 = O(1)$. Then, and by the law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^n \eta_{j,0,i}^4 = O_P(1).$$

For the other term we have since $v_{\beta,i}^{-2} = O(1)$ by condition (E2) that

$$\frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 v_{\beta,i}^{-2} - 1) = \frac{1}{n} \sum_{i=1}^n v_{\beta,i}^{-2} (v_{\beta,i}^2 - v_{\beta,i}^{-2})^4 = O \left( \frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 - v_{\beta,i}^{-2})^4 \right).$$

On this term we apply the Lipschitz property of $v_{\beta,i}$ to have

$$|v_{\beta,i}^2 - v_{\beta,i}^{-2}| = |(v_{\beta,i}^2 - v_{\beta,i}^{-2})^2 + 2v_{\beta,i} (v_{\beta,i} - v_{\beta,i}^{-2})| \leq L|X_i^T (\hat{\beta} - \beta_0)| (L|X_i^T (\hat{\beta} - \beta_0)| + 2).$$

Now observe that

$$|L|X_i^T (\hat{\beta} - \beta_0)| + 2 \leq 2 + \|X_i\|_\infty \|\hat{\beta} - \beta_0\|_1 = O_P(1).$$

Then we obtain

$$\frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 - v_{\beta,i}^{-2})^4 = O_P(1) \frac{1}{n} \sum_{i=1}^n (L|X_i^T (\hat{\beta} - \beta_0)|)^4 \leq O_P \left( \max_{i=1,...,n} \|X_i\|_\infty^2 \|\hat{\beta} - \beta_0\|_1^2 \frac{1}{n} \sum_{i=1}^n (L|X_i^T (\hat{\beta} - \beta_0)|)^2 \right)^2,$$

where in the last step we applied Hölder’s inequality to one part of the term. This then gives by the result $\|X(\hat{\beta} - \beta_0)\|^2_2/n = O_P(s\lambda^2)$, by $\max_{i=1,...,n} \|X_i\|_\infty^2 = O(K^2)$ and by the $\ell_1$ rates $\|\hat{\beta} - \beta_0\|_1 = O_P(s\lambda)$ that

$$\frac{1}{n} \sum_{i=1}^n (v_{\beta,i}^2 - v_{\beta,i}^{-2})^4 \leq O_P(1) \max_{i=1,...,n} \|X_i\|_\infty^2 \|\hat{\beta} - \beta_0\|_1^2 \frac{1}{n} \sum_{i=1}^n (L|X_i^T (\hat{\beta} - \beta_0)|)^2 \leq O_P(s^3\lambda^4).$$

This is turn implies

$$\|(W_\beta^2 W_\beta^{-2} - I)\eta_{j,0,j}\|^2_2/n = O_P(s^{3/2} \lambda^2).$$

Hence, and returning to the basic inequality, we obtain

$$\|X_{\beta,-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta,j})\|^2_2/n + 2\lambda_j \|\hat{\gamma}_{\beta,j}\|_1 \leq 2\eta_{j,0,j}^\top X_{j,0,-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta,j})/n + O_P(s^{3/4}\lambda) \|X_{\beta,0,-j}(\hat{\gamma}_{\beta,j} - \gamma_{\beta,j})\|_2/\sqrt{n} + 2\lambda_j \|\gamma_{\beta,j}\|_1.$$

For arbitrary $\delta > 0$, we have

$$2ab \leq \delta a^2 + b^2/\delta.$$
Applying this claim we get

\[ 2\mathcal{O}(s^{3/4} \lambda) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n \leq \mathcal{O}_P(s^{3/2} \lambda^2) / \delta + \delta \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n. \]

Now we use that, under conditions on \( \nu_\beta \) in condition [E2]

\[ \| X_{\beta_0 - j} (\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) \|_2 / n = \mathcal{O}_P(\| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n). \]

Hence we get

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + 2\lambda_j \| \hat{\gamma}_{\hat{\beta}_j} \|_1 \leq 2\eta_{\beta_0,j} X_{\beta_0 - j} (\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) / n + \mathcal{O}_P(s^{3/2} \lambda^2) + 2\lambda_j \| \gamma_{\beta_0,j} \|_1.
\]

Now we have by Hölder’s inequality

\[ \eta_{\beta_0,j} X_{\beta_0 - j} (\hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j}) / n \leq \| \eta_{\beta_0,j} X_{\beta_0 - j} \|_\infty / n \| \hat{\gamma}_{\beta,j} - \gamma_{\beta_0,j} \|_1. \]

Under [A1] by Nemirovski’s inequality (Theorem A.3) it follows

\[ \| \eta_{\beta_0,j} X_{\beta_0 - j} \|_\infty / n \leq \lambda_j, \]

for \( \lambda \approx \sqrt{\log p} / n. \) Hence

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + 2\lambda_j \| \hat{\gamma}_{\hat{\beta}_j} \|_1 \leq \lambda_j \| \hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j} \|_1 + \mathcal{O}_P(s^{3/2} \lambda^2) + 2\lambda_j \| \gamma_{\beta_0,j} \|_1.
\]

Using triangle inequality we can get from the above

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + \lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 \leq 3\lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1 + \mathcal{O}_P(s^{3/2} \lambda^2).
\]

**Case i)**

If \( \lambda \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1 \geq \mathcal{O}_P(s^{3/2} \lambda^2) \) then

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + \lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 \leq 4\lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1.
\]

Then we continue the chain of calculations

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + \lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 \leq 4\lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1
\leq 4\lambda_j \sqrt{s} \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_2
\leq 4\lambda_j \sqrt{s} \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / \sqrt{n}
\leq 16\lambda^2 / \delta
+ \delta \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n.
\]

That implies

\[
(1 - 2\delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + \lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 \leq 16\lambda^2 / \delta.
\]

But then

\[ \| \hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j} \|_1 = \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1 + \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 = \mathcal{O}_P(s \lambda^2). \]

**Case ii)**

If \( \lambda \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1 \leq \mathcal{O}_P(s^{3/2} \lambda^2) \) then

\[
(1 - \delta) \| X_{\hat{\beta}_j - j} (\hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j}) \|_2 / n + \lambda_j \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 \leq \mathcal{O}_P(s^{3/2} \lambda^2).
\]

But then

\[ \| \hat{\gamma}_{\hat{\beta}_j} - \gamma_{\beta_0,j} \|_1 = \| \hat{\gamma}_{\hat{\beta}_j,S} - \gamma_{\beta_0,j,S} \|_1 + \| \hat{\gamma}_{\hat{\beta}_j,S} \|_1 = \mathcal{O}_P(s^{3/2} \lambda^2). \]

\qed
Lemma A.6. Suppose that conditions [A1] [A2] [E1] [E2] [E3] are satisfied. Let \( \lambda \asymp \sqrt{\log p/n} \). Then it holds that

\[
|\hat{\beta}_j^2 - \tau_j^2| = O_P(K_X \sqrt{s \log p/n}).
\]

Proof. By the definition of \( \hat{\beta}_j^2 \)

\[
\hat{\beta}_j^2 = X^T_{\beta,j} (X_{\beta,j} - X_{\beta,-j} \hat{\beta}_j) / n = X^T_{\beta} W^2_{\beta} (X_j - X_{-j} \hat{\beta}_j) / n.
\]

We have

\[
\hat{\beta}_j^2 - \tau_j^2 = \underbrace{X^T_{\beta} (W^2_{\beta} - W^2_{\beta_0}) (X_j - X_{-j} \hat{\beta}_j)}_{\text{I}} / n + \underbrace{X^T_{\beta} W^2_{\beta_0} (X_j - X_{-j} \hat{\beta}_j)}_{\text{II}} / n - \tau_j^2.
\]

We treat the two terms separately. We have

\[
|\text{I}| = X^T_{\beta} W^2_{\beta} (X_j - X_{-j} \hat{\beta}_j) / n - \tau_j^2 = \frac{||W_{\beta_0} (X_j - X_{-j} \gamma_{0,j})||_2^2 / n - \tau_j^2}{O_P(1/\sqrt{n})} + \frac{\gamma_{0,j}^T X^T_{\beta} W^2_{\beta} (X_j - X_{-j} \gamma_{0,j}) / n}{O_P(\sqrt{\lambda})} + \frac{X^T_{\beta} W^2_{\beta_0} (X_j - X_{-j} \hat{\beta}_j)}{O_P(\lambda s^{3/2})}.
\]

Hence, \( \text{II} = O_P(\sqrt{s \log p/n}) \). For the first term, we have

\[
|\text{I}| = |X^T_{\beta} (W^2_{\beta} - W^2_{\beta_0}) (X_j - X_{-j} \hat{\beta}_j) / n| = \frac{1}{n} \sum_{i=1}^n X_{j,i} (v^2_{\beta,i} - v^2_{\beta_0,i}) \hat{\beta}_j,i,
\]

where \( \hat{\beta}_j := X_j - X_{-j} \hat{\beta}_j \). Then

\[
|\text{I}| = \frac{1}{n} \sum_{i=1}^n X_{j,i} (v^2_{\beta,i} - v^2_{\beta_0,i}) \hat{\beta}_j,i \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (v^2_{\beta,i} - v^2_{\beta_0,i})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n X^2_{j,i} \hat{\beta}_j,i} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (v^2_{\beta,i} - v^2_{\beta_0,i})^2} K_X \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{\beta}_j^2,i}.
\]

Then we have

\[
\frac{1}{n} \sum_{i=1}^n (v^2_{\beta,i} - v^2_{\beta_0,i})^2 = \frac{1}{n} \sum_{i=1}^n ((v^2_{\beta,i} - v^2_{\beta_0,i})^2 + 2v_{\beta_0}(v^2_{\beta,i} - v^2_{\beta_0,i})) \leq \frac{1}{n} \sum_{i=1}^n (v^2_{\beta,i} - v^2_{\beta_0,i})^2 \leq \frac{1}{n} \sum_{i=1}^n (x^T_{i} (\hat{\beta} - \beta_0))^2 \lesssim s \log p/n.
\]
Therefore, 
\[
\hat{\eta}^T_i \hat{\eta}_j / n \quad = \quad \|X_j - X_j \hat{\gamma}_j\|^2 / n
\]
\[
= \|X_j - X_j \hat{\gamma}_j\|^2 / n + 2(X_j - X_j \hat{\gamma}_j)^T X_j (\hat{\gamma} - \gamma_0) / n
\]
\[
+ \|X_j (\hat{\gamma} - \gamma_0)\|^2 / n.
\]
Let \( \eta_j := X_j - X_j \gamma_0 \). Then we have \( \eta_j^T \eta_j / n = O_P(\tau_j^2) \). Further we have
\[
|(X_j - X_j \gamma_0)^T X_j (\hat{\gamma} - \gamma_0) / n| \quad \leq \quad \eta_j^T \eta_j / 2 / \sqrt{n} \|X_j (\hat{\gamma} - \gamma_0)\| / \sqrt{n}
\]
\[
= \quad O_P(\tau_j^2) a_P(s^{\nu/4} \lambda) = O_P(1).
\]
Hence
\[
\hat{\eta}^T_i \hat{\eta}_j / n = O_P(\tau_j^2) = O_P(1).
\]
Therefore,
\[
|I| \quad \leq \quad \sqrt{\frac{1}{n} \sum_{i=1}^n (v_i^2 - v_i^2_{\alpha,i})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{\eta}^2_{j,i}}.
\]
\[
\approx \quad K \sqrt{s \log p / n}.
\]

**Proof of Theorem** We use Lemmas A.5 and A.6 to obtain
\[
\|\hat{\Theta}_j - \Theta_j^0\|_1 = \|\hat{\Gamma}_j / \hat{\tau}_j^2 - \Gamma_j / \tau_j^2\|_1 \leq \|\hat{\gamma}_j - \gamma_j^0\|_1 / \hat{\tau}_j^2 + \|\tau_j^0\|_1 (1 / \hat{\tau}_j^2 - 1 / \tau_j^2).
\]

We have \( i = O_P(s^{3/2} \sqrt{\log p / n}) \) and \( ii = O_P(s \sqrt{\log p / n}) \).

**Lemma A.7.** Suppose the generalized linear model setting from Example 2.1: \( Y_i = g(x_i^T \beta_0) + \epsilon_i \), where \( X_i \) and \( \epsilon_i \) are independent for \( i = 1, \ldots, n \). Let \( \psi_{\beta}(x, y) := w(y - g(x^T \beta)) x \). Then
\[
(E \psi_{\beta}(x, y))_{\beta} = \sum_{\beta_0} E_Y w(u, y)_{\beta_0} x x^T.
\]

**Proof of Lemma** We have
\[
E \psi_{\beta}(y, x) = E_{(x, Y)} w(x^T \beta, y) x
\]
\[
= \sum_{\beta_0} E_Y w(x^T \beta, y)_{\beta_0} x x^T.
\]
Let \( G(u) := E_Y w(u, y) \). We use a Taylor expansion of \( G \) around \( u_0 = x^T \beta_0 \)
\[
G(u) = G(u_0) + G'(u_0)(u - u_0) + \frac{1}{2} G''(u)|_{u = \hat{u}}(u - u_0)^2,
\]
where \( \hat{u} = \alpha u_0 + (1 - \alpha) u \) for some \( 0 \leq \alpha \leq 1 \). This yields
\[
E_Y (w(x^T \beta, y)|x) = \sum_{\beta_0} E_Y (w(x^T \beta_0)|x) + G'(u)|_{u = x^T \beta_0} x^T (\beta - \beta_0)
\]
\[
+ \frac{1}{2} G''(u)|_{u = \hat{u}} (x^T (\beta - \beta_0))^2.
\]
Differentiating this with respect to \( \beta \) we obtain
\[
(E_Y (w(x^T \beta, y)|x))_{\beta} = -G'(u)|_{u = x^T \beta_0} x
\]
\[
+ (\frac{1}{2} G''(u)|_{u = \hat{u}})_{\beta} (x^T (\beta - \beta_0))^2
\]
\[
+ \frac{1}{2} G''(u)|_{u = \hat{u}} 2 (x^T (\beta - \beta_0)) x^T.
\]
Plugging in $\beta = \beta_0$ we obtain
\[
(\mathbb{E}_Y (w(x^T \beta, y)|x))'_{\beta=\beta_0} = -G'(u)|_{u=x^T \beta_0 x}.
\]
Now we have
\[
(\mathbb{E}_Y (w(x^T \beta, y)|x))'_{\beta=\beta_0} = \mathbb{E}_x (E_Y (w(x^T \beta, y)|x))'_{\beta=\beta_0} x.
\]
Hence
\[
(\mathbb{E}_Y (w(x^T \beta, y)|x))'_{\beta} = -\mathbb{E}_x G'(u)|_{u=x^T \beta_0 x} x x^T = -\mathbb{E}_x (E_Y w(y, u)|x)'_{u=x^T \beta_0 x} x x^T.
\]

A.5 Proofs for Section 5 (Examples)

A.5.1 Proofs for Section 5.2 ($\ell_1$-penalized LAD estimator)

Proof of Lemma 5.1 The necessary conditions for the problem (16) read
\[
P_n s_\beta + \lambda \hat{Z} = 0,
\]
where $s_\beta$ is the subdifferential of $\beta \mapsto |y - x^T \beta|$ and $\hat{Z}$ is the subdifferential of the $\ell_1$ norm.

One can show that if $Y_i$ is absolutely continuous conditional on $X_i$, then with probability one,
\[
\{i : y_i = x_i^T \hat{\beta}\} = \{\hat{\beta}\} (i.e. exact interpolation happens exactly $\|\hat{\beta}\|_0$ times). The estimator $\hat{\beta}$ satisfies the condition of Lemma 2.2, therefore, $\|\hat{\beta}\|_0 = O_P(s)$. Hence,
\[
\|P_n \psi_{\hat{\beta}}\|_\infty = \|P_n \psi_{\hat{\beta}} - P_n s_{\hat{\beta}}\|_\infty + \|P_n s_{\hat{\beta}}\|_\infty \\
\leq \left\| \frac{1}{n} \sum_{i : y_i \neq x_i^T \hat{\beta}} (\psi_{\hat{\beta}} - s_{\hat{\beta}}) \right\|_\infty + \left\| \frac{1}{n} \sum_{i : y_i = x_i^T \hat{\beta}} (\psi_{\hat{\beta}} - s_{\hat{\beta}}) \right\|_\infty + O_P(\lambda) \\
\leq s/n + O_P(\lambda) = O_P(1),
\]
where we used that $\|s_{\hat{\beta}}\|_\infty \leq \|\psi_{\hat{\beta}}\|_\infty = O_P(1)$.

Proof of Lemma 5.2 Let $V_i \subset \{1, \ldots, p\}$ for $i = 1, \ldots, \binom{p}{s}$ be all subsets of $\{1, \ldots, p\}$ of size $s$.

We can rewrite
\[
\mathcal{F} = \bigcup_{i=1}^{\binom{p}{s}} \mathcal{F}_i,
\]
where $\mathcal{F}_i := \{(x, y) \mapsto \Theta_f^T x (1_{y \leq x^T \beta} - 1_{y > x^T \beta}) : \beta \in \mathbb{R}^p, \beta V_i = 0\}$. We now show that the class $\mathcal{F}_i$ has VC-index $V(\mathcal{F}_i)$ of order $s$.

The VC-index of a class of functions $\mathcal{F}_i$ is defined as the VC-index of the collection of sets $\{(x, y, t) : t < f(x, y)\}$, where $t \in \mathbb{R}, f \in \mathcal{F}_i$. The collection $\{(x, y) \mapsto y - x^T \beta : \beta V_i = 0\}$ has the same VC-index as an $s$-dimensional real vector space, which is $s + 2$ by Lemma 2.6.15 in van der Vaart and Wellner (1996). Hence it follows that the VC-index of the collection $\{(x, y, t) : t < f(x, y)\}$ is of order $s$.

Then by Theorem 2.6.7 in van der Vaart and Wellner (1996) and since the covering number of a union of sets is upper bounded by sum of the covering numbers, we obtain
\[
N(\epsilon\|F\|_n, \mathcal{F}_i, \| \cdot \|_n) \leq \left( \frac{p}{s} \right)^{\binom{p}{s}} N(\epsilon\|F\|_n, \mathcal{F}_i, \| \cdot \|_n) \\
\leq \left( \frac{p}{s} \right)^{\binom{p}{s}} KV(\mathcal{F}_i)(16e)^{V(\mathcal{F}_i)} \left( \frac{1}{\epsilon} \right)^{2(V(\mathcal{F}_i)-1)} \\
= \left( \frac{p}{s} \right)^{\binom{p}{s}} K s(16e)^s \left( \frac{1}{\epsilon} \right)^{2(s-1)},
\]
where $K$ is a universal constant and $0 < \epsilon < 1$. 

\[
\left(\frac{p}{s}\right) K s(16\epsilon)^s \left(\frac{1}{\epsilon}\right)^{2(s-1)} \leq \frac{p^s}{s!} K s(16\epsilon)^s \left(\frac{1}{\epsilon}\right)^{2s} \leq \frac{p^s}{(s-1)!} K \left(\frac{16\epsilon}{\epsilon}\right)^{2s} \leq K p^s \left(\frac{16\epsilon}{\epsilon}\right)^{2s}.
\]

Hence

\[
\log N(\epsilon \|F\|_n, \|\cdot\|_n) \leq s \log p + 2s \log \left(\frac{16\epsilon}{\epsilon}\right).
\]

\[\square\]

**Proof of Theorem 5.1.** We apply Theorem 2.2 and take $\psi_\beta(x, y) := \text{sign}(y-x\beta)x$. Then by Lemma 5.1 it follows that $\|\psi_\beta\|_\infty = O_p(\lambda)$.

First note that Condition (C1) is satisfied, which can be seen by direct calculation as follows. We have

\[
u \rightarrow G(\nu) = \int w(\nu, y) dP_{\epsilon|x} = \int \text{sign}(y-\nu) dP_{\epsilon|x} = 1 - 2f_\epsilon(\nu).
\]

Then $G'(\nu) = -f_\epsilon(\nu)$. Hence by the assumed Lipschitz property of $f_\epsilon$, it follows that

\[|G'(\nu) - G'(v)| = |f_\epsilon(\nu) - f_\epsilon(v)| \leq L|\nu - v|,
\]

thus $G'$ is Lipschitz.

Next we need to show that $E(w_\beta - w_{\beta_0})^2 = o(n/(s^3(\log p)^2(\log n)^4))$. First we calculate the expectation conditioned on $x$:

$E_{\epsilon|x}(w_\beta - w_{\beta_0})^2 = E_{\epsilon|x}(\text{sign}(y-x^T\beta) - \text{sign}(y-x^T\beta_0))^2$

$= \int (1_{\epsilon < 0} - 1_{\epsilon > x^T(\beta - \beta_0)})^2 dP_{\epsilon|x}$

$= \int_0^\infty dP_{\epsilon|x} = |F_\epsilon(0) - F_\epsilon(x^T(\beta - \beta_0))| \leq L|x^T(\beta - \beta_0)|$.

Then for $\beta$ satisfying $\|X(\beta - \beta_0)\|_2^2/n = O(s\lambda^2)$ and $\|\beta - \beta_0\|_1 = O(s\lambda)$ we have

$E(w_\beta - w_{\beta_0})^2 = E_{x}(E_{\epsilon|x}(w_\beta - w_{\beta_0})^2|x)$

$\leq E_{x} L|x^T(\beta - \beta_0)| \leq L\sqrt{E_{x}|x^T(\beta - \beta_0)|^2}$

$= L\sqrt{(\beta - \beta_0)^T E_{x} xx^T(\beta - \beta_0)}$

$= O(\sqrt{s}\lambda)$.

Then

$P f^2 = P|\Theta_\beta^T x|^2|w_\beta - w_{\beta_0}|^2 \leq s E(w_\beta - w_{\beta_0})^2 = O(s^{3/2}\lambda)$.

Then under $s^5(\log p)^3(\log n)^4/n = o(1)$, it follows that $R\sqrt{s\log p}\log n = o(1)$.

By boundedness of $w_\beta$, we have

$P\sup_{\beta} |w_\beta - w_{\beta_0}|^4 = O(1)$

and hence by the Cauchy-Schwarz inequality, the condition

$P\sup_{\beta} f^4 s^3(\log p)^3(\log n)^6 = o(1)$

39
is satisfied.

We have that \( \hat{\Theta}' \) is an estimate of \( \Theta' \) constructed using nodewise regression with the matrix \( \hat{\Sigma} := X^T X/n \). Then since \( \Lambda_{\max}(\mathbb{E}xx^T) = O(1) \), we have by Theorem 2.4 in \cite{van2014},

\[
\| \hat{\Theta}' - \Theta' \|_1 = \mathcal{O}_P(s\lambda).
\]

Then since \( f_x(0) \geq c > 0 \) where \( c \) is a universal constant, we get

\[
\| \hat{\Theta}_j - \Theta_0^j \|_1 = \mathcal{O}_P(s\lambda).
\]

Finally, the entropy condition is satisfied by Lemma 5.2.