Asymptotically Honest Confidence Regions for High Dimensional Parameters by the Desparsified Conservative Lasso

Mehmet Caner* Anders Bredahl Kock†

November 19, 2017

Abstract

In this paper we consider the conservative Lasso which we argue penalizes more correctly than the Lasso and show how it may be desparsified in the sense of van de Geer et al. (2014) in order to construct asymptotically honest (uniform) confidence bands. In particular, we develop an oracle inequality for the conservative Lasso only assuming the existence of a certain number of moments. This is done by means of the Marcinkiewicz-Zygmund inequality. We allow for heteroskedastic non-subgaussian error terms and covariates. Next, we desparsify the conservative Lasso estimator and derive the asymptotic distribution of tests involving an increasing number of parameters. Our simulations reveal that the desparsified conservative Lasso estimates the parameters more precisely than the desparsified Lasso, has better size properties and produces confidence bands with superior coverage rates.

Keywords and phrases: conservative Lasso, honest inference, high-dimensional data, uniform inference, confidence intervals, tests.

JEL codes: C12, C13, C21.

1 Introduction

In recent years we have seen a burgeoning literature on high-dimensional problems where the number of parameters is much greater than the sample size. Statistical inference in the sense of constructing tests and confidence bands in the high-dimensional linear regression model were considered in a seminal series of papers by Belloni et al. (2010, 2012, 2011b, 2014, 2011a). These authors showed how a cleverly constructed (double)

*Ohio State University, 452 Arps Hall, Department of Economics, Translational Data Analytics, Department of Statistics, OH 43210. Email:caner.12@osu.edu.

†Oxford University, Aarhus University and CREATES, Department of Economics, Manor Road, Oxford, OX1 3UQ, UK. Email: anders.kock@economics.ox.ac.uk. We would like to thank Victor Chernozhukov and Andrea Montanari for pointing us to relevant related research. The paper has also benefited tremendously from insightful comments by the co- and associate editor as well as the referees. Financial support from the Danish National Research Foundation is gratefully acknowledged by the second author (grant DNRF78). First version: October 2014.

post selection estimator can be used to construct uniformly valid confidence intervals for the parameter of interest in instrumental variable and treatment effect models allowing for imperfect model selection in the first step. Also Fan et al. (2015) show how to set up test statistics in high dimensions with power enhancing components against sparse alternatives. Nickl and van de Geer (2013) consider honest adaptive inference when $p > n$. This can be obtained as long as the rate of sparse estimation does not exceed $n^{-1/4}$. Hoffmann and Nickl (2011) consider the existence of honest adaptive confidence bands for an unknown density function. They show that this is possible if the non-parametric hypotheses for the null and alternative are asymptotically consistently distinguishable. Berk et al. (2013) propose a conservative post selection inference method. The idea is simultaneous inference in all models' submodels and this results in very wide confidence intervals. Taylor and Tibshirani (2015) discuss a practical way of taking into account the model selection's effect on post selection inference. Tibshirani (2011) provides a nice summary of developments in the literature while Lockhart et al. (2014) provide a computation based significance test for Lasso estimators. Also Zou and Li (2008) and Fan et al. (2014) used adaptive weights in Lasso type estimators that enhance model selection properties.

The paper closest in spirit to ours is van de Geer et al. (2013, 2014) who cleverly showed how the classical Lasso estimator may be *desparsified* to construct asymptotically valid confidence bands for a low-dimensional subset of a high-dimensional parameter vector. This paper in turn is related to Zhang and Zhang (2014), Javanmard and Montanari (2013) and Javanmard and Montanari (2014). The idea behind desparsification is to remove the bias introduced by shrinkage via desparsifying the estimator using a cleverly constructed approximate inverse of the non-invertible empirical Gram matrix. Furthermore, these confidence bands do not suffer from the critique of Pötscher (2009) regarding the overly large size of confidence bands based on variable selection consistent estimators. By using the desparsified Lasso to construct confidence bands and tests, van de Geer et al. (2014) strike a middle ground between classical low dimensional inference, which relies heavily on testing, and Lasso-type techniques which perform estimation and variable selection in one step without any testing.

In the framework of the high-dimensional linear regression model and inspired by the work of van de Geer et al. (2014) we study the so-called conservative Lasso. The important observation here is that, in the presence of an oracle inequality on the plain Lasso, the penalty of the conservative Lasso on the non-zero parameters will be no larger than the one for the Lasso while the penalty on the zero parameters will be the same as the one induced by the plain Lasso. Hence, the conservative Lasso may be expected to deliver more precise parameter estimates (in finite samples) than the Lasso. And indeed, our theoretical results and simulations strongly indicate that this is the case. Also note that recently Fan et al. (2014) proposed a weighted ℓ_1 penalized estimator with very similar weights. Their focus is on strong oracle optimality and we show that a variant of our conservative Lasso possesses the strong oracle optimality property.

We provide an oracle inequality for the conservative Lasso estimator and use the method of desparsification introduced in van de Geer et al. (2014). This approach has the advantage that the zero and non-zero

coefficients do not have to be well-separated (no $\beta_{\min}$-condition is imposed) in order to conduct valid inference. We only assume the existence of r moments as opposed to the classical sub-gaussianity assumption. The oracle inequalities rely on the use of the Marcinkiewicz-Zygmund inequality which we argue delivers slightly more precise estimates than Nemirovski's inequality.

We also show that hypotheses involving an increasing number of parameters can be tested (we are considering a *fixed* sequence of hypotheses) which generalizes the results on hypotheses involving a bounded number of parameters in van de Geer et al. (2014). Furthermore, we allow for heteroskedastic error terms and provide a uniformly consistent estimator of the high-dimensional asymptotic covariance matrix. This is an important generalization in practical problems as heteroskedasticity is omniscient in econometrics and statistics. A similar approach could be of interest in large linear panel data models under strict exogeneity.

The simulations show that vast improvements can be obtained by using the desparsified conservative Lasso as opposed to the plain desparsified Lasso. To be precise, the true parameter β_0 is in general estimated much more precisely and χ^2-tests based on the desparsified conservative Lasso have much better size properties (and often also higher power) than their counterparts based on the desparsified Lasso.

When implementing Lasso-type estimators the choice of tuning parameter is important. Thus, in Theorem 5 in the appendix, we show how the method of Fan and Tang (2013) can be used to choose the tuning parameter of the variant of the conservative Lasso when the objective is consistent model selection in high dimensions.

The rest of the paper is organized as follows. Section 2 introduces the model and the conservative Lasso. Section 3 introduces nodewise regression, desparsification, and the approximate inverse to the empirical Gram matrix. Section 4 introduces inference and establishes honest confidence intervals and shows that they contract at the optimal rate. The simulations can be found in Section 5. Section 6 concludes the paper. All proofs are deferred to the appendix

2 The Model

Before stating the model setup we introduce some notation used throughout the paper.

2.1 Notation

For any real vector x, we let $\|x\|_q$ denote the ℓ_q-norm. We will primarily use the ℓ_1-, ℓ_2-, and the ℓ_∞-norm. For any $m \times n$ matrix A, we define $\|A\|_\infty = \max_{1\leq i\leq m, 1\leq j\leq n} |A_{i,j}|$. Occasionally we shall also use the induced ℓ_∞-norm. This will be denoted by $\|A\|_{\ell_\infty}$ and equals the maximum absolute row sum of A. For any symmetric matrix B, let $\phi_{\min}(B)$ and $\phi_{\max}(B)$ denote the smallest and largest eigenvalue of B, respectively. If $x \in \mathbb{R}^n$ and S is a subset of $\{1, ..., n\}$ we let x_S be the subvector of x that places picks out only those elements indexed by S.

For any set S, let $|S|$ denote its cardinality and for $x \in \mathbb{R}^n$ its prediction norm is defined as $\|x\|_n =$

$\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$. $\xrightarrow{d}$ will indicate convergence in distribution and $o_p(a_n)$ as well as $O_p(b_n)$ are used in their usual meaning for sequences a_n and b_n. $a_n \asymp b_n$ means that these sequences differ at most by strictly positive multiplicative constants.

2.2 The model

We consider the model

$$Y = X\beta_0 + u, \tag{1}$$

where X is the $n \times p$ matrix of explanatory variables and u is a vector of error terms. β_0 is the $p \times 1$ population regression coefficient which we shall assume to be sparse. However, the location of the non-zero coefficients is unknown and potentially p could be much greater than n. The sparsity assumption can be replaced by a weak sparsity assumption as we shall make precise after Theorem 1 below. We assume that the explanatory variables are exogenous and precise assumptions will be made in Assumption 1 below. Let $S_0 = \{j : \beta_{0,j} \neq 0\}$ and $s_0 = |S_0|$. For later purposes define X_j as the j'th column of X and X_{-j} as all columns of X except for the j'th one.

2.3 The conservative Lasso and comparison to (adaptive) Lasso

The conservative Lasso is a two-step estimator defined as the weighted Lasso

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \left\| Y - X\beta \right\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \hat{w}_j \, |\beta_j| \} \tag{2}$$

with weights $\hat{w}_j = \frac{\lambda_{prec}}{|\hat{\beta}_{L,j}| \vee \lambda_{prec}}$ where $\hat{\beta}_L$ is the plain Lasso estimator which is used to construct the weights $\hat{w}_j$. The plain Lasso corresponds to $w_j = 1$ for $j = 1, ..., p$ in (2). Here λ_n and λ_{prec} are positive non-random quantities chosen by the researcher which we shall be specific about shortly. In Lemma A.7 and the simulation section we show that λ_{prec} can be chosen as an estimable multiple of λ_n. Hence, the only tuning parameter is λ_n. We choose λ_n by either BIC or the Generalized Information Criterion (GIC) of Fan and Tang (2013). Details are provided in the Monte Carlo section. A theorem tying GIC to model selection consistency of a variant of our conservative Lasso (which will be described in the next subsection) is at the end of Appendix B.

As opposed to the adaptive Lasso, the conservative Lasso gives variables that were excluded by the first step initial Lasso estimator a second chance — even if $|\hat{\beta}_{L,j}| = 0$ one has $\hat{w}_j = 1$ instead of an "infinitely" large penalty. Hence, the name "conservative" Lasso. The adaptive Lasso usually performs its worst when a relevant variable has been left out by the initial Lasso estimator. The conservative Lasso rules out such a situation while still using more intelligent weights than the Lasso as we shall see shortly. Note that our definition of the conservative Lasso is at first glance slightly different from the one on page 205 in Bühlmann and van de Geer (2011).

We shall choose λ_{prec} to equal an upper bound on the estimation error of the first step Lasso for reasons to be made clear next. In particular, assume that λ_{prec} is such that $\mathcal{C}_1 = \left\{\|\hat{\beta}_L - \beta_0\|_\infty \leq \lambda_{prec}\right\}$ is a set with large probability. Lemma A.7 in the Appendix provides a concrete choice of λ_{prec} ensuring that $\mathcal{C}_1$ occurs with high probability. In Theorem 1 below we shall give examples of λ_{prec}.

Recently Fan et al. (2014) proposed a one step solution to folded concave penalized estimation of which a subcase is the SCAD of Fan and Li (2001). This weighted ℓ_1 penalty approach is similar to our conservative Lasso. Unlike our fractional weight structure their weights are normalized and truncated by a multiple of λ_n. Like us, Fan et al. (2014) also solve the zero denominator issue of the adaptive Lasso, as pointed out by Fan and Lv (2008) and Fan and Lv (2010). However, their paper's emphasis is on strong oracle optimality, which we shall discuss in more details when introducing our variant of the conservative Lasso in the next subsection, while we are interested in constructing tests and confidence bands.

As is standard in the literature we assume that the covariates X_i are i.i.d. with $\Sigma = E(X_1 X_1')$ satisfying an adaptive restricted eigenvalue condition:

$$\phi_\Sigma^2(s) = \min_{\substack{\delta \in \mathbb{R}^p \setminus \{0\} \\ \|\delta_{S^c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2}} \frac{\delta' \Sigma \delta}{\|\delta_S\|_2^2} > 0, \tag{3}$$

where $S \subseteq \{1, ..., p\}$. Instead of minimizing over all of $\mathbb{R}^p$, the minimum in (3) is restricted to those vectors which satisfy $\|\delta_{S^c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2$. Thus, the adaptive restricted eigenvalue condition is satisfied in particular when Σ has full rank.

In order to establish an oracle inequality for the conservative Lasso we shall assume the following.

Assumption 1. *The covariates $X_i \in \mathbb{R}^p$, $i = 1, ..., n$ are independently and identically distributed while the error terms $u_i \in \mathbb{R}$, $i = 1, ..., n$ are independently distributed with $E(u_i|X_i) = 0$. Furthermore, $\max_{1 \leq j \leq p} E|X_{1,j}|^r \leq C$ and $\max_{1 \leq i \leq n} E|u_i|^r \leq C$ for some $r \geq 2$ and a positive universal constant C. $\phi_\Sigma^2(s_0)$ is bounded away from 0.*

Assumption 1 states that the covariates are independently and identically distributed with uniformly bounded r'th moments. The assumption of identical distribution of the covariates is mainly made to keep expressions simple but could be relaxed. We will comment in more detail on this later. The error terms are allowed to be non-identically distributed and may, in particular, be conditionally heteroskedastic. Thus, many applications of interest are covered. At this point it is also worth mentioning that in the literature one often assumes that the covariates as well as the error terms are uniformly sub-gaussian. This is a much stronger assumption than the one imposed here and rules out data with heavy tails. However, strengthening our assumption to sub-gaussianity would not cause any trouble and deliver stronger results. In particular, all powers of p below could be replaced by powers of $\log(p)$ which are asymptotically much smaller. A third route which is sometimes taken is to assume the covariates to be bounded, the error terms to possess bounded second moments and then use Nemirovski's inequality to obtain oracle inequalities which only depend on p through its logarithm.

Define $\Theta = \Sigma^{-1}$, and $\|\hat{w}_{S_0}\|_\infty = \max_{j \in S_0} |\hat{w}_j|$, which is the maximal weight among all the relevant

variables. We are now ready to state the oracle inequality for the weighted Lasso estimator in (2).

Theorem 1. *Let Assumption 1 be satisfied, set* $\lambda_n = M\frac{p^{2/r}}{n^{1/2}}$ *for* $M > 0$ *and* $\lambda_{prec} = \frac{9\lambda_n}{4}\|\Theta\|_{l_\infty}$. *Then, with probability at least* $1 - \frac{C}{M^{r/2}} - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$, *the conservative Lasso satisfies the following inequalities*

$$\|X(\hat{\beta} - \beta_0)\|_n^2 \leq 2(2\|\hat{w}_{S_0}\|_\infty + 1)^2 \frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}, \tag{4}$$

$$\|\hat{\beta} - \beta_0\|_1 \leq 4(\|\hat{w}_{S_0}\|_\infty + 1)(2\|\hat{w}_{S_0}\|_\infty + 1)\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}, \tag{5}$$

for universal constants $C, D > 0$. *Furthermore, these bounds are valid uniformly over the* ℓ_0*-ball* $\mathcal{B}_{\ell_0}(s_0) = \{\|\beta_0\|_{\ell_0} \leq s_0\}$.

Remarks.

1. For the Lasso $\|\hat{w}_{S_0}\|_\infty = 1$. Thus, the upper bound in (4) takes the value $18\frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}$. For the conservative Lasso we always have $0 < \|\hat{w}_{S_0}\|_\infty \leq 1$ such that the upper bound is no worse than for the Lasso. In fact, the multiplicative constant 18 can be considerably improved in certain settings. To give a concrete example consider Lemma 1(iii) where $\|\hat{w}_{S_0}\|_\infty \to 0$ on $\mathcal{C}_1 = \{\|\hat{\beta}_L - \beta_0\|_\infty \leq \lambda_{prec}\}$ [1] Therefore, the upper bound in (4) approaches $2\frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}$ which is 9 times smaller than the bound for the Lasso. Note that Lemma 1 (iii) relies on a $\beta_{\min}$-type condition. However, even without this condition, one has $\|\hat{w}_{S_0}\|_\infty \leq 1$ implying upper bounds for the conservative Lasso that are no worse than the ones for the plain Lasso.

2. Next consider ℓ_1 error bounds for the estimation error. For the Lasso the upper bound in (5) is $24\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}$ and when $\max_{j\in S_0}\hat{\omega}_j \to 0$, (5) approaches $4\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}$ for the conservative Lasso by Lemma 1 (iii).

3. To simplify the notation in future lemmas and proofs, define $d_{n1} = 2(2\|\hat{w}_{S_0}\|_\infty + 1)^2$ and $d_{n2} = 4(\|\hat{w}_{S_0}\|_\infty + 1)(2\|\hat{w}_{S_0}\|_\infty + 1)$.

4. $\lambda_{prec} = \frac{9\lambda_n}{4}\|\Theta\|_{l_\infty}$, and in the simulation section we provide a consistent estimator of $\|\Theta\|_{\ell_\infty}$. This choice of λ_{prec} is motivated by Lemma A.7 in the appendix which shows that λ_{prec} is a high probability upper bound on the ℓ_∞ estimation error of the Lasso.

Finally, the sparsity assumption on β_0 can be replaced by a bound on $\sum_{j=1}^p |\beta_{0,j}|^q$ for $0 < q < 1$ as it is not difficult to establish oracle inequalities in such a "weakly sparse" setting. Thus, none of the entries of β_0 need to equal exactly zero but we stick to the classical ℓ_0-sparsity here.

Define $S_j = \{k = 1, ..., p : \Theta_{j,k} \neq 0\}$ as the indices of the non-zero entries of the jth row of Θ_j. Let $s_j = |S_j|$. Define also $\eta_j = X_j - X_{-j}\gamma_j$, which is a $n \times 1$ vector.

Assumption 2:

a) $\phi_{\min}(\Sigma)$ is bounded away from zero.

b) $\frac{p^2(\max(s_0, \max_{1\leq j\leq p} s_j))^{r/2}}{n^{r/4}} \to 0$.

c) $E(|\eta_{j,i}|^r)$ uniformly bounded over $i = 1, ..., n$ and $j = 1, ..., p$.

[1] $\mathcal{C}_1$ occurs whenever the event in Theorem 1 having probability at least $1 - \frac{C}{M^{r/2}} - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$ occurs. Thus, we are not working on a smaller event than for the Lasso.

Assumption 2a) states that the smallest eigenvalue of the *population* covariance matrix is bounded away from zero. It is used to make sure that $\tau_j^2 = 1/\Theta_{j,j} \geq 1/\phi_{\max}(\Theta) = \phi_{\min}(\Sigma)$ are bounded away from zero. Part b) is needed to show that $\left\|\hat{\Sigma} - \Sigma\right\|_\infty$ converges to zero sufficiently fast to conclude that the adaptive restricted eigenvalue of $\hat{\Sigma} = \frac{1}{n}X'X$ is close to the one of Σ. It implies an upper bound on how fast the dimension, p, of the model can increase. The more moments one assumes the covariates and the error terms to possess, the faster can p grow. From Assumption 2b), it is clear that since $p \geq \max(s_0, \max_{1\leq j\leq p} s_j)$, $\max(s_0, \max_{1\leq j\leq p} s_j) = o(n^{\frac{r}{2(r+4)}})$. This restricts the number of non-zero coefficients in the model and each row of Θ. On the other hand, if the error terms and covariates are subgaussian, it is not difficult to show that this requirement is relaxed to $\max(s_0, \max_{1\leq j\leq p} s_j) = o(\sqrt{n})$. Intuitively this can be seen by letting $r \to \infty$. Nevertheless, the inverse covariance matrix must be sparse. This is satisfied if Σ is e.g. block diagonal or has the Toeplitz structure $\Sigma_{i,j} = \rho^{|i-j|}$, $-1 < \rho < 1$. In the simulations we shall also see that our method works well even if $\Theta = \Sigma^{-1}$ is not sparse as long as its entries are not too far from zero. This is not surprising as the sparsity assumption can easily be relaxed to the weak sparsity assumption of $\sum_{l=1}^p |\Omega_{j,l}|^q$ not being too large for any $j \in H$ for some $0 < q < 1$ as similar bounds to the ones in Lemma A.9 below remain valid under this assumption. Thus, no entry of Θ needs to be zero as long as each row can be well approximated by a sparse vector. This observation was also made in Yuan (2010) for a different estimator of Θ.

Observe that $\hat{w}_j \leq 1$ for all $j = 1, ..., p$. We now provide a lemma that shows desirable properties of the weights of the conservative Lasso. We caution that the fact that the weights of the non-zero coefficients approaching zero in (iii) comes at the price of a $\beta_{\min}$ type of condition which rules out very small coefficients. This kind of assumption may not be suitable in economics. Without this type of condition the result in (iii) will not be true, however the weights of the non-zero coefficients of the conservative Lasso will still be smaller than for the plain Lasso.

Lemma 1. *Under Assumptions 1-2, with* $\lambda_n = M\frac{p^{2/r}}{n^{1/2}}$ *for* $M > 0$, *and* $\lambda_{prec} = \frac{9\lambda_n}{4}\|\Theta\|_{\ell_\infty}$. *Then,*

(i).

$$\lambda_{prec} \to 0,$$

and on $\mathcal{C}_1 = \left\{\|\hat{\beta}_L - \beta_0\|_\infty \leq \lambda_{prec}\right\}$, *with* $P(\mathcal{C}_1) \to 1$, *the following two statements hold*

(ii).

$$\min_{j\in S_0^c} \hat{w}_j = 1,$$

(iii). In addition, if $\min_{j\in S_0} |\beta_{0,j}|/\lambda_{prec} \to \infty$ *then*

$$\max_{j\in S_0} \hat{w}_j \to 0,$$

Remarks.

1. Lemma 1 shows that $\lambda_{prec} = O(\lambda_n\sqrt{\max_{1\leq j\leq p} s_j}) = o(1)$. Its proof reveals that even if we replace $\|\Theta\|_{\ell_\infty}$ by $\|\hat{\Theta}_L\|_{\ell_\infty}$ we still get $\lambda_{prec} \xrightarrow{p} 0$. Note that $\hat{\Theta}_L$ represents the Lasso nodewise regression estimate of the inverse matrix of Σ^{-1}. It is a subcase of our conservative nodewise regression in Section 3.2, and explained in footnote 4 there.

2. An even better result can be achieved in terms of the conditions needed for λ_{prec} to converge to zero. If, for example, Σ is an equicorrelation matrix then $\|\Theta\|_{\ell_\infty} = O(1)$ by Example 2.5.1 of van de Geer (2014). Thus, $\lambda_{prec} = O(\lambda_n) = o(1)$. The same is the case if $\Sigma_{i,j} = \rho^{|i-j|}$ for some $-1 < \rho < 1$ which is an often considered structure.

3. Parts (ii) and (iii) of the lemma show that asymptotically no penalty is applied to the coefficients of the relevant variables while the same penalty as for the Lasso is applied to the coefficients of the irrelevant variables. In particular, it is guaranteed that all non-zero coefficients are penalized less than all zero coefficients.

We shall see in Section 5 that the above advantages of the conservative Lasso over the plain Lasso materialize in better performance also in the simulations.

2.4 A Variant of the Conservative Lasso

In this section we introduce a variant of the conservative Lasso estimator. This variant possess the property of strong oracle optimality under slightly stronger conditions than Assumptions 1 and 2, see Theorem 4 at the end of Appendix B. Strong oracle optimality is defined as an estimator being equal to the oracle estimator with probability approaching one (p.822 of Fan et al. (2014)). As the plain Lasso is generally not strongly oracle optimal this shows superiority of the variant of the conservative Lasso to the former. First, we define the variant of the conservative Lasso as

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|\}, \tag{6}$$

where $\tilde{w}_j = 1_{\{|\hat{\beta}_{L,j}| \leq \lambda_{prec}\}}$. In this new variant the weights only take the values 0 or 1. Importantly, this is a variant of the conservative Lasso since all variables still get a second chance after the first step Lasso estimation.

Strong oracle optimality of (6) is established in Theorem 4 at the end of Appendix B. Theorem 4 (i) shows that $\min_{j \in S_0^c} \tilde{w}_j = 1$ and $\max_{j \in S_0} \tilde{w}_j = 0$ with with probability approaching one. Thus, in this variant of the conservative Lasso the weights pertaining to the non-zero coefficients will be *exactly* equal to zero with probability approaching one while for the conservative Lasso these weights only converge to zero with probability approaching one. This slightly stronger property contributes to obtaining the strong oracle property for the variant of the conservative Lasso in the proof of Theorem 4. Note, however, that Theorem 4(i) comes with a $\beta_{\min}$ type of condition similar to the one in Lemma 1(iii). However, under Assumption 1, Theorem 1 above holds also when $\tilde{\beta}$ replaces $\hat{\beta}$. The same holds for Theorems 2 and 3 if one desparsifies $\tilde{\beta}$ instead of $\hat{\beta}$.

3 Desparsification

3.1 The Desparsified Conservative Lasso

In order to conduct inference we shall use the idea of desparsification proposed in van de Geer et al. (2014). The idea is that the shrinkage bias introduced due to the presence of penalization in (2) will show up in the properly scaled limiting distribution of $\hat{\beta}_j$. Hence, we remove this bias prior to conducting statistical inference. Letting $\hat{W} = diag\left(\hat{w}_1, ..., \hat{w}_p\right)$ be a $p \times p$ diagonal matrix containing the weights of the conservative Lasso, the first order condition of (2) may be written as

$$-X'(Y - X\hat{\beta})/n + \lambda_n \hat{W}\hat{\kappa} = 0,$$

with $\|\hat{\kappa}\|_\infty \leq 1$, and $\hat{\kappa}_j = sign(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$ for $j = 1, ..., p$. Thus,

$$\lambda \hat{W}\hat{\kappa} = X'(Y - X\hat{\beta})/n. \tag{7}$$

Next, as $Y = X\beta_0 + u$ and defining $\hat{\Sigma} = X'X/n$, the above display yields

$$\lambda_n \hat{W}\hat{\kappa} + \hat{\Sigma}(\hat{\beta} - \beta_0) = X'u/n.$$

In order to isolate $\hat{\beta} - \beta_0$ we need to invert $\hat{\Sigma}$. However, when $p > n$, $\hat{\Sigma}$ is not invertible. Thus, the idea is now to construct an approximate inverse, $\hat{\Theta}$, to $\hat{\Sigma}$ and control the error term resulting from this approximation. We shall be explicit about the construction of $\hat{\Theta}$ in the next section. For any $p \times p$ matrix we may write, by multiplying the above equation by $\hat{\Theta}$, and adding $\hat{\beta} - \beta_0$ to each side of the above equation,

$$\hat{\beta} = \beta_0 - \hat{\Theta}\lambda_n \hat{W}\hat{\kappa} + \hat{\Theta}X'u/n - \Delta/n^{1/2}, \tag{8}$$

where

$$\Delta = \sqrt{n}(\hat{\Theta}\hat{\Sigma} - I_p)(\hat{\beta} - \beta_0),$$

is the error resulting from using an approximate inverse, $\hat{\Theta}$, as opposed to an exact inverse. We shall show that Δ is asymptotically negligible. Adding $\hat{\Theta}\lambda_n \hat{W}\hat{\kappa}$ to both sides of (8) results in the following estimator by using (7)

$$\begin{aligned} \hat{b} &= \hat{\beta} + \hat{\Theta}\lambda_n \hat{W}\hat{\kappa} = \hat{\beta} + \hat{\Theta}X'(Y - X\hat{\beta})/n & (9) \\ &= \beta_0 + \hat{\Theta}X'u/n - \Delta/n^{1/2}. & (10) \end{aligned}$$

Hence, for any $p \times 1$ vector α with $\|\alpha\|_2 = 1$ we can consider

$$\sqrt{n}\alpha'\left(\hat{b} - \beta_0\right) = \alpha'\hat{\Theta}X'u/n^{1/2} - \alpha'\Delta, \tag{11}$$

such that a central limit theorem for $\alpha'\hat{\Theta}X'u/n^{1/2}$ and a verification of asymptotic negligibility of $\alpha'\Delta$ will yield asymptotically gaussian inference. Furthermore, we provide a uniformly consistent estimator of the asymptotic variance of $\sqrt{n}\alpha'\left(\hat{b} - \beta_0\right)$ which makes inference practically feasible. In connection with

Theorem 2 we shall see that a central limit theorem for $\alpha'\hat{\Theta}X'u/n^{1/2}$ puts certain limitations on the number of non-zero entries of α in (11), i.e. the number of parameters involved in a hypothesis. A leading special case of the above setting is of course $\alpha = e_j$ where e_j is the j'th unit vector for $\mathbb{R}^p$. Then, (11) reduces to

$$\sqrt{n}\left(\hat{b}_j - \beta_{0,j}\right) = \left(\hat{\Theta}X'u/n^{1/2}\right)_j - \Delta_j. \tag{12}$$

In general, let $H = \left\{j = 1, ..., p : \alpha_j \neq 0\right\}$ with cardinality $h = |H|$. Thus, H contains the indices of the coefficients involved in the hypothesis being tested. We shall allow for $h \to \infty$ as the first in the literature on inference in high-dimensional regression models with more regressors than observations ($p > n$) but require $h/n \to 0$ as $n \to \infty$.

In the next section we construct the approximate inverse $\hat{\Theta}$ which enters in both terms in the above display and thus plays a crucial role for the limiting inference. The above desparsification procedure is similar in spirit to the one outlined in van de Geer et al. (2014). However, $\hat{\beta}$ is used instead of $\hat{\beta}_L$. Furthermore, the construction of the approximate inverse $\hat{\Theta}$ in the next section relies on the conservative Lasso as opposed to the plain Lasso.

3.2 Constructing the Approximate Inverse of the Gram Matrix: $\hat{\Theta}$

In this subsection we construct the approximate inverse $\hat{\Theta}$ of $\hat{\Sigma}$. This is done by nodewise regression a la Meinshausen and Bühlmann (2006) and van de Geer et al. (2014). To formally define the nodewise regression recall that X_j is the j'th column in X and X_{-j} all columns of X except for the j'th one. First, along the lines of van de Geer et al. (2014) we define the Lasso nodewise regression estimates

$$\hat{\gamma}_{L,j} = \underset{\gamma \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\|X_j - X_{-j}\gamma\right\|_n^2 + 2\lambda_{node,n} \sum_{k \neq j} |\gamma_k| \tag{13}$$

for each $j = 1, ..., p$. We use these estimates to construct the weights of the conservative Lasso nodewise regression which is defined as follows

$$\hat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\|X_j - X_{-j}\gamma\right\|_n^2 + 2\lambda_{node,n} \left\|\hat{\Gamma}_j\gamma\right\|_1, \tag{14}$$

where $\hat{\Gamma}_j = diag\left(\frac{\lambda_{prec}}{|\hat{\gamma}_{L,l}| \vee \lambda_{prec}},\ l = 1, ..., p, l \neq j\right)$ is a $(p-1) \times (p-1)$ matrix of weights. [2]

Note that we choose $\lambda_{node,n}$ to be the same in all of the nodewise regressions. This is needed for the uniform results in Lemma A.9 below to be valid. Thus, the conservative Lasso is run p times as an intermediate step to construct $\hat{\Theta}$. Using the notation $\hat{\gamma}_j = \left\{\hat{\gamma}_{j,k};\ k = 1, ..., p,\ k \neq j\right\}$ we define

$$\hat{C} = \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \cdots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \cdots & -\hat{\gamma}_{2,p} \\ \cdots & \cdots & \ddots & \cdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \cdots & 1 \end{pmatrix}. \tag{15}$$

[2] For the variant of the conservative Lasso we have $\tilde{\Gamma}_j = diag\left(1_{\{|\hat{\gamma}_{L,l}| \leq \lambda_{prec}\}},\ l = 1, ..., p, l \neq j\right)$.

To define $\hat{\Theta}$ we introduce $\hat{T}^2 = diag(\hat{\tau}_1^2, \cdots, \hat{\tau}_p^2)$ which is a $p \times p$ diagonal matrix with

$$\hat{\tau}_j^2 = \|X_j - X_{-j}\hat{\gamma}_j\|_n^2 + \lambda_{node,n}\|\hat{\Gamma}_j\hat{\gamma}_j\|_1, \tag{16}$$

for all $j = 1, ..., p$. We now define

$$\hat{\Theta} = \hat{T}^{-2}\hat{C}. \tag{17}$$

[3] [4] It remains to be shown that this $\hat{\Theta}$ is close to being an inverse of $\hat{\Sigma}$. To this end, we define $\hat{\Theta}_j$ as the j'th row of $\hat{\Theta}$ but understood as a $p \times 1$ vector and analogously for $\hat{C}_j$. Thus, $\hat{\Theta}_j = \hat{C}_j/\hat{\tau}_j^2$. Denoting by e_j the j'th $p \times 1$ unit vector, arguments detailed in appendix C show that

$$\|\hat{\Theta}_j'\hat{\Sigma} - e_j'\|_\infty \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}. \tag{18}$$

Hence, the above display provides an upper bound on the maximal absolute entry of the j'th row of $\hat{\Theta}\hat{\Sigma} - I_p$ which, combined with the oracle inequality for $\|\hat{\beta} - \beta_0\|_1$, will yield an upper bound on Δ_j in (11) by arguments made rigorous in the appendix.

Before stating Assumption 3 we introduce the following notation in connection to the asymptotic covariance matrix. Set $\bar{s} = \max_{j \in H} s_j$, $\Sigma_{xu} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n EX_iX_i'u_i^2$ and $\hat{\Sigma}_{xu} = n^{-1}\sum_{i=1}^n X_iX_i'\hat{u}_i^2$, where $\hat{u}_i = Y_i - X_i'\hat{\beta}$.

Assumption 3.

Let $r \geq 6$ and

a) $s_0\frac{h^{2/r+1/2}p^{4/r}}{n^{1/2}} \to 0$.

b) $\frac{p^{8/r}h\bar{s}}{n^{1/2}} \to 0$.

c) $\frac{p^{2/r}\sqrt{s_0}h\bar{s}}{n^{1/2}} \to 0$, $\frac{p^{8/r}\sqrt{s_0}h\bar{s}}{n^{3/4}} \to 0$ and $\frac{p^{8/r}s_0h\bar{s}}{n^{(r-2)/r}} \to 0$.

d) $\frac{(h\bar{s})^{r/4+1} \wedge (h\bar{s})^{r/4}p}{n^{r/4-1}} \to 0$.

e) $\phi_{\min}(\Sigma_{xu})$ is bounded away from 0 and $\phi_{\max}(\Sigma_{xu})$ is uniformly bounded. $\phi_{\max}(\Sigma)$ is bounded from above.

Assumptions 3a)-d) all restrict the rate at which the size of the model (p), the number of relevant variables (s_0) as well as the number of coefficients involved in the hypothesis being tested (h) are allowed to increase. However, part b) of Assumption 3 reveals that the number of $\beta_{0,j}$ involved in the hypothesis being tested must be of order $o(n^{1/2})$. Letting the number of parameters involved in the hypothesis increase with the sample size is a useful generalization of van de Geer et al. (2014) who only mention the possibility of H possessing a

[3] A practical benefit is that the nodewise regressions actually only have to be run for $j \in H$ and not all $j = 1, ..., p$ as we only need to estimate the covariance matrix of those parameters involved in the hypothesis being tested.

[4] Denote by $\hat{\Theta}_L$ the nodewise regression estimate of Θ based on the Lasso. This can be obtained by using $\hat{\gamma}_L$ from (13) instead of $\hat{\gamma}$ in (15)-(17).

fixed or growing number of elements. Part b) also reveals that if one encounters a situation where p increases faster than the sample size, then one needs $r > 16$ for our theory. If one is willing to assume subgaussianity of the covariates and the error terms the powers of p in Assumption 3 could be replaced by powers of $\log(p)$. Furthermore, in a different context, Belloni et al. (2012, 2014) have used moderate deviation inequalities for self-normalized sums to get results which are of the same order as if subgaussianity was imposed but only assuming certain moments to exist for the covariates and the error terms. In that case, as usual, p can grow almost exponentially in n. Assumptions 3a)-d) are trivially satisfied in the classical setting where p, h, s_0 and $\bar{s}$ are fixed. Finally, Assumption 3e) restricts the eigenvalues of Σ and Σ_{xu}.

4 Inference

This section has two main results. The first result provides sufficient conditions for asymptotically gaussian inference to be valid for linear combinations of the entries of desparsified conservative Lasso $\hat{b}$. The second result shows that the resulting confidence bands are uniformly valid and contract at the optimal rate.

Theorem 2. *Let Assumptions 1-3* [5] *be satisfied. Then,*

$$\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \xrightarrow{d} N(0,1), \tag{19}$$

where α is a $p \times 1$ vector with $\|\alpha\|_2 = 1$. Furthermore,

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \left|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha\right| = o_p(1). \tag{20}$$

Theorem 2 provides sufficient conditions for asymptotically gaussian inference to be valid. We stress again that the number of $\beta_{0,j}$, h, involved in the statistic in (19) is allowed to increase as the sample size tends to infinity as long as this does not happen too fast. Furthermore, these results can be valid in the presence of more variables than observations ($p > n$).

We also emphasize that the above results allow the error terms to be heteroskedastic. (20) provides a uniformly consistent estimator of the asymptotic variance of $n^{1/2}\alpha'(\hat{b}-\beta_0)$. The uniformity of (20) will also be used in the proof of Theorem 3 below. (20) is also interesting as it gives the limit of the variance in the denominator of (19) even as the dimension ($p \times p$) of the matrices involved in the expression increases.

Note that while d_{n1} and d_{n2} do not directly enter in the first order asymptotic result of Theorem 2, equations (A.75), (A.78), (A.79) and (A.82) in the appendix still reveal that the desparsified conservative Lasso is likely to result in more precise inference than the plain desparsified Lasso. The effect comes directly from more precise parameter estimates as well as through more precise covariance matrix estimation using the nodewise regressions and is clearly seen in the simulations.

[5] Assumption 2b) is of course implied by Assumption 3b) but to keep the statement clean we shall simply assume all of Assumption 2 to be valid.

In the case where H is a set of fixed cardinality h, (19) implies that

$$\left\| \left(\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\right)^{-1/2}_{H,H} \sqrt{n}\left(\hat{b}_H - \beta_{0,H}\right) \right\|_2^2 \xrightarrow{d} \chi^2(h), \tag{21}$$

as it is asymptotically a sum of h independent standard normal random variables. Thus, asymptotically valid χ^2-inference can be performed in order to test a hypothesis on h parameters simultaneously. Wald tests of general restrictions of the type $H_0 : g(\beta_0) = 0$ (where $g : \mathbb{R}^p \to \mathbb{R}^h$ is differentiable in an open neighborhood around β_0 and has derivative matrix of rank h) can now also be constructed in the usual manner, see e.g. Davidson (2000) Chapter 12, even when $p > n$ which has hitherto been impossible.

Consider the leading special case where $H = \{j\}$ such that α reduces to the j'th unit vector e_j of $\mathbb{R}^p$ and $h = 1$. As a corollary to the previous theorem we consider testing a hypothesis about a single coefficient. The number of regressors is assumed to be a positive multiple of the sample size and the maximal number of non-zero entries in the jth row of the inverse population covariance matrix is bounded. This is satisfied when, eg, Σ is a Toeplitz matrix. The important thing to notice is that all dimensionality assumptions from Assumptions 1-3 are automatically satisfied in the setting considered in Corollary 1.

Corollary 1. *Let Assumptions 1, 2a, 2c and 3e be satisfied with $p = an,\ a > 0$, with $r > 16$, $s_0 = O(n^{1/4})$, $\bar{s} = O(1)$. Then,*

$$\frac{n^{1/2}(\hat{b}_j - \beta_{j0})}{\sqrt{(\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}')_{j,j}}} \xrightarrow{d} N(0,1), \tag{22}$$

Furthermore,

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \left| (\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}')_{j,j} - (\Theta\Sigma_{xu}\Theta')_{j,j} \right| = o_p(1). \tag{23}$$

If, furthermore, the covariates and the error terms are independent and the latter are homoskedastic with variance σ^2 we get that

$$\alpha'\Theta\Sigma_{xu}\Theta'\alpha = e_j'\Sigma^{-1}\sigma^2\Sigma\Sigma^{-1}e_j = \sigma^2(\Sigma^{-1})_{j,j},$$

which is nothing else than the standard formula for the asymptotic variance of the least squares estimator of the j'th coefficient $\hat{\beta}_{OLS,j}$ in a fixed dimensional linear regression model. Thus, there is no efficiency loss. Corollary 1 is in the spirit of Robinson (1988) who constructed a $\sqrt{n}$ consistent estimator of the coefficients pertaining to the linear part of a semiparametric model. See also van de Geer et al. (2014) Section 2.3.3 for more discussion and relations to the semiparametric framework. In the context of uniformly valid confidence bands for a single parameter the work of Belloni et al. (2014) is also relevant. These authors consider inference on treatment effects using a post-double-selection procedure.

4.1 Uniform Convergence

The next theorem shows that the confidence bands based on the desparsified conservative Lasso are honest and that they contract at the optimal rate. Recall that $\mathcal{B}_{\ell_0}(s_0) = \left\{\|\beta_0\|_{\ell_0} \leq s_0\right\}$.

Theorem 3. *Let Assumptions 1-3 be satisfied and let $\alpha = \alpha_n \in \mathbb{R}^p$ denote any fixed sequence of vectors satisfying $\|\alpha\|_2 = 1$. Then we have*

$$\sup_{t\in\mathbb{R}} \sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)} \left| P\left(\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t \right) - \Phi(t) \right| \to 0. \tag{24}$$

Furthermore, letting $\hat{\sigma}_j = \sqrt{e_j'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'e_j}$ (corresponding to $\alpha = e_j$ in (24)) and $z_{1-\delta/2}$ the $1-\delta/2$ percentile of the standard normal distribution, one has for all $j = 1, ..., p$

$$\lim_{n\to\infty} \inf_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)} P\left(\beta_{0,j} \in \left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}} \right] \right) = 1-\delta. \tag{25}$$

Finally, letting $\mathrm{diam}([a,b]) = b-a$ *be the length of an interval $[a,b]$ in the real line, we have that*

$$\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)} \mathrm{diam}\left(\left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}} \right] \right) = O_p\left(\frac{1}{\sqrt{n}} \right). \tag{26}$$

(24) reveals that convergence to the standard normal distribution is actually valid uniformly over the ℓ_0-ball of radius at most s_0. We stress, however, that (24) ceases to be valid if one also takes the supremum over all α sequences satisfying the assumptions of the theorem. Thus, the asymptotics are uniform over β_0 but pointwise in α. (25) is a consequence of (24) and entails that the confidence band $\left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right]$ is *asymptotically honest* for $\beta_{0,j}$ over $\mathcal{B}(s_0)$ in the sense of Li (1989).

(26) is important as it reveals that the confidence band $\left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right]$ has the optimal rate of contraction $1/\sqrt{n}$. Furthermore, these confidence bands are uniformly narrow over $\mathcal{B}_{\ell_0}(s_0)$ such that for all $\epsilon > 0$ there exists an $M > 0$, not depending on β_0, with the property that

$$\mathrm{diam}\left(\left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}} \right] \right) \leq M/\sqrt{n},$$

for all $\beta_0 \in \mathcal{B}_{\ell_0}(s_0)$ with probability at least $1-\epsilon$. Here it is vital that at the same time the confidence intervals are asymptotically honest. Since the desparsified conservative Lasso is not a sparse estimator, (26) does not contradict inequality 6 in Theorem 2 of Pötscher (2009) who shows that honest confidence bands based on sparse estimators must be large.

Finally, the above results are valid without any sort of $\beta_{\min}$-condition. In total, Theorem 3 reveals that the inference of our procedure is very robust as the confidence bands are honest and contract *uniformly* at the optimal rate.

We provide a brief overview of the proofs here. Lemmata A.1-A.3 in the appendix are crucial ingredients of our main theorems. Lemma A.1 provides an oracle inequality for general weighted Lasso type estimators subject to a condition on the smallest weight of the truly zero coefficients. Lemmata A.2 and A.3 are very important to get maximal inequalities for certain sums that determine the order of λ_n in our setting of regressors and error terms with bounded rth moments. Our use of the Marcinkiewicz-Zygmund inequality provides sharper bounds than Nemirowski's inequality. The technical details are given in the remarks after Lemma A.3. Thus, Lemma A.3 is a novel contribution in high dimensional statistics. Lemma A.7 provides an ℓ_∞ bound for the estimation error of the Lasso in the case of error terms and regressors with bounded rth

moments and heteroskedasticity. Furthermore, Lemma A.7 provides a theoretical choice of λ_{prec}. Theorem 2 is key in getting a heteroskedasticity consistent estimate of the variance of linear combinations of the parameters involved in the hypothesis being tested and is new in the literature. At the end of Appendix B we also establish strong oracle optimality for the variant of the conservative Lasso and a way to choose λ_n for consistent variable selection for this estimator.

5 Monte Carlo

In this section we investigate the finite sample performance of the (desparsified) conservative Lasso and compare it to the one of the (desparsified) Lasso of van de Geer et al. (2014). We also implement the procedure of Javanmard and Montanari (2014) using the authors' code[6]. The Lasso as well as the conservative Lasso are implemented in R by means of the publicly available `glmnet` package.

To choose λ_n as well as $\lambda_{node,n}$, we follow Fan and Tang (2013) and use their Generalized Information Criterion (GIC). In the regression equation (1)

$$\lambda_n^* = \underset{\lambda_n \in \{\lambda_l, \ldots, \lambda_u\}}{\operatorname{argmin}} GIC(\lambda_n),$$

where

$$GIC(\lambda_n) = \log(\|Y - X\hat{\beta}_{\lambda_n}\|_n^2) + \frac{\log\log(n)\log(p)|S_{\lambda_n}|}{n},$$

and λ_l, λ_u are lower and upper bounds for λ_n while $\hat{\beta}_{\lambda_n}$ is the conservative Lasso estimate pertaining to λ_n. Finally, $|S_{\lambda_n}|$ denotes the number of non-zero entries in $\hat{\beta}_{\lambda_n}$. The same procedure is used to choose λ_n for the variant of the conservative Lasso as well as in the nodewise regressions to choose $\lambda_{node,n}$. At the end of Appendix B we provide a theorem stating that for the variant of the conservative Lasso choosing the tuning parameters by GIC leads to consistent model selection.

We compare GIC to choosing the tuning parameters by BIC, see e.g. (9.4.9) in Davidson (2000). Of course one could also use cross validation to choose λ_n but in our experience this does not improve the quality of the results while being considerably slower. All data will be generated from the model (1).

As argued in Section 2.3 a good choice of λ_{prec} should be a high probability bound on $\|\hat{\beta}_L - \beta_0\|_{\ell_\infty}$. Lemma A.7 in the appendix shows that $\lambda_{prec} = \frac{9\lambda_n}{4}\|\Theta\|_{\ell_\infty}$ is exactly such a bound. Next, Lemma A.9 in the appendix justifies using the plug-in estimate $\|\hat{\Theta}_L\|_{\ell_\infty}$ for $\|\Theta\|_{\ell_\infty}$ in the choice of λ_{prec}. However, we find that in practice one might as well use $\lambda_{prec} = \frac{9\lambda_n}{4}$ which corresponds to $\Theta = I_p$. This choice also has the additional computational advantage of avoiding running all p nodewise regressions. Furthermore, it is the fallback option used in Javanmard and Montanari (2014) in case any of their optimizations needed to get $\hat{\Theta}$ fails. Thus, we shall use $\lambda_{prec} = \frac{9\lambda_n}{4}$ which, however, does not come with theoretical performance guarantees.

[6] Available at `https://web.stanford.edu/~montanar/sslasso/code.html`.

The following algorithm summarizes how to implement the desparsified conservative Lasso and how to conduct inference with it.

Algorithm to implement the desparsified conservative Lasso

1. For each $\lambda_n \in \{\lambda_l, ..., \lambda_u\}$ implement the Lasso $\hat{\beta}_L$ by imposing $\hat{w}_j = 1$ in (2). $\{\lambda_l, ..., \lambda_u\}$ is constructed by the `glmnet` package in R to ensure that models of many sizes are implemented. Use either BIC or GIC to select $\lambda_n \in \{\lambda_l, ..., \lambda_u\}$.

2. Construct $\hat{w}_j = \frac{\lambda_{prec}}{|\hat{\beta}_{L,j}| \vee \lambda_{prec}}$ $j = 1, ..., p$ with $\lambda_{prec} = \frac{9}{4}\lambda_n$ and implement the conservative Lasso $\hat{\beta}$ as in (2). Use either BIC or GIC to select $\lambda_n \in \{\lambda_l, ..., \lambda_u\}$.

3. For each $j \in H$ construct the jth element of the desparsified conservative Lasso by the following steps.

   a) Run the nodewise Lasso in (13) with $\lambda_{node} = \lambda_n$ to get $\hat{\gamma}_{L,j}$.

   b) Construct the weights for the nodewise conservative Lasso: $\hat{\Gamma}_j = diag\left(\frac{\lambda_{prec}}{|\hat{\gamma}_{L,l}| \vee \lambda_{prec}},\ l = 1, ..., p, l \neq j\right)$.

   c) Run the nodewise conservative Lasso as in (14) using $\hat{\Gamma}_j$ from step 3b above.

   d) Construct $\hat{C}_j$, the jth row of $\hat{C}$, as in as in (15) and obtain $\hat{\tau}_j^2$ as in (16).

   e) Let $\hat{\Theta}_j = \hat{C}_j / \hat{\tau}_j^2$ be the jth row of $\hat{\Theta}$.

   f) Construct the jth element of the desparsified conservative Lasso (9) which is $\hat{b}_j = \hat{\beta}_j + \hat{\Theta}_j X'(Y - X\hat{\beta})/n$.

4. χ^2-tests are constructed as in (21) while the confidence bands are constructed as in (25).

The variant of the conservative Lasso goes through steps 1-4 using $\tilde{w}_j$ instead of $\hat{w}_j$ and $\tilde{\Gamma}_j$ instead of $\hat{\Gamma}_j$.

All simulations are carried out with 1,000 replications unless stated otherwise and we consider the following performance measures for each of the procedures:

1. Estimation error: We compute the ℓ_2-estimation error of the Lasso and the conservative Lasso and its variant averaged over the Monte Carlo replications.

2. Size: We evaluate the size of the χ^2-test in (21) for a hypothesis involving more than one parameter.

3. Power: We evaluate the power of the χ^2-test in (21) for a hypothesis involving more than one parameter.

4. Coverage rate: We calculate the coverage rate of a gaussian confidence interval constructed as in (25). This is done for a non-zero as well as a zero parameter.

5. Length of confidence interval: We calculate the length of the two confidence intervals considered in point 4, above.

In the simulations we investigate the performance of the conservative Lasso in moderate, high, and very high-dimensional settings. The covariance matrices of the covariates are always chosen to have a Toeplitz

structure with (i,j)'th entry equal to $\rho^{|i-j|}$ for some $0 \leq \rho < 1$ to be made precise below. The covariates and the error terms are assumed to be t-distributed with 10 degrees of freedom. At this point we remark that all experiments reported below were also carried out with the covariates possessing a block diagonal covariance matrix and/or gaussian error terms (all combinations were tried). This only affected the findings in the simulations marginally and we shall not report these results here.

All tests are carried out at a 5% significance level and all confidence intervals are at the 95% level. Unless mentioned otherwise, the χ^2-tests involve the two first parameters in β_0 of which we deliberately make sure that the first one is 1 and the second one is zero. Thus, $h = 2$ in our Experiments 1-3. For measuring the size of the χ^2-test, we test the true hypothesis $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$. For measuring the power of the χ^2-test, we test the false hypothesis $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$. Thus, the hypothesis is only false on the second entry of β_0. Similarly, we construct confidence intervals for the first two parameters of β_0 such that the coverage rate can be compared between non-zero and zero parameters.

As our theory allows for heteroskedastic error terms we also investigate the effect of this. To be precise, we consider error terms of the form $u_i = \epsilon_i \left(\frac{1}{\sqrt{2}} X_{1,i} + b_x X_{2,i}\right)$ where $\epsilon_i \sim t(10)$ is independent of the covariates and b_x is chosen such that the unconditional variance of u_i is still that of a t-distribution with 10 degrees of freedom[7]. Note that this u_i satisfies our assumption $E(u_i|X_i) = 0$ and has variance conditional on X_i given by $E(\epsilon_i^2)\left(\frac{1}{\sqrt{2}} X_{1,i} + b_x X_{2,i}\right)^2$. The reason we ensure that the unconditional variance of u_i is still that of a $t(10)$-distribution is that we do not want any findings to be driven by a plain change in the unconditional variance. It is also deliberate that we choose the conditional heteroskedasticity to depend on $X_{1,i}$ and $X_{2,i}$ as these are the variables involved in the χ^2-tests and the confidence intervals.

- Experiment 1a (moderate-dimensional setting). β_0 is 50×1 with 10 ones and 40 zeros. The 10 ones are equidistant in the parameter vector. Thus, $p = 50$ and $s_0 = 10$. We consider $\rho = 0, 0.5$ and 0.9 and $n = 100$.

- Experiment 1b (moderate-dimensional setting). As Experiment 1a but with heteroskedastic errors.

- Experiment 2a (high-dimensional setting). β_0 is 104×1 with the first four entries being $(1, 0, 1, 0.1)$ and the remaining 100 entries being zero. Thus, $p = 104$ and $s_0 = 3$. We consider $\rho = 0, 0.5$ and 0.9 and $n = 100$.

- Experiment 2b (high-dimensional setting). As Experiment 2a but with heteroskedastic errors.

- Experiment 3a (very high-dimensional setting). β_0 is 1000×1 with 10 ones and 990 zeros. The 10 ones are equidistant in the parameter vector. Thus, $p = 1000$ and $s_0 = 10$. $\rho = 0.75$. This experiment is carried out for $n = 100, 150, 200, 500$ to gauge the effect of an increasing sample size. We also experimented with different values of ρ but this did not qualitatively alter our findings. The number

[7] To ensure that u_i still has the variance of $\epsilon_i \sim t(10)$ a small calculation shows that it suffices to choose $b_x = \frac{-\sqrt{2}\rho + \sqrt{2\rho^2+2}}{2}$. Thus, the higher the correlation between $X_{1,i}$ and $X_{2,i}$, the smaller b_x should be chosen.

of replications is 100 as the procedure of Javanmard and Montanari (2014) is rather time consuming in high dimensions.

- Experiment 3b (very high-dimensional setting). As Experiment 3a but with heteroskedastic errors.
- Experiment 4: As Experiment 2a with $\rho = 0.5$ but testing a hypothesis involving the first ten parameters to investigate the properties of the proposed procedures when many parameters are involved in the hypothesis being tested. When gauging power, the only deviation from the true parameter vector is that the second entry of β_0 is hypothesized to be 0.4 (as in all other power calculations).

5.1 Results

Most often, using BIC or GIC to choose λ_n is not overly important for our performance measures. However, BIC tends to perform better when p is large compared to n and, unless mentioned otherwise, we will focus on the results for BIC in the sequel. We also note that a general finding is that the conservative Lasso performs better than its variant when p is small compared to n while this ordering reverses when p is large compared to n.

Table 1 contains the results for Experiment 1a. First, as predicted in Section 2.3, both versions of the conservative Lasso have a lower estimation error than the plain Lasso due to more intelligent weights. The variant of the conservative Lasso fares particularly well for $\rho = 0$ and $\rho = 0.5$. Furthermore, the conservative Lasso is always less size distorted than the Lasso while having slightly more power except for when $\rho = 0.9$. The procedure of Javanmard and Montanari (2014) has even less size distortion but the price is very low power. When $\rho = 0.9$ all procedures have serious power deficiencies. Next, our procedure (both versions) always has a coverage rate which is closer to the nominal rate of 95% than the plain desparsified Lasso. Note, however, that all Lasso-based procedures still have a slight tendency towards undercoverage (a phenomenon which disappears as the sample size is increased (not reported here)). This is the case in particular for the plain Lasso and less pronounced for the conservative Lasso. The reasons for this are that the confidence intervals produced by the Lasso are too narrow compared to the more accurate ones produced by the conservative Lasso and that the latter produces more precise parameter estimates. The confidence intervals of Javanmard and Montanari (2014) have good coverage but are very wide.

Next, Table 2 adds heteroskedasticity to the results of Experiment 1a. The main message of this table is that qualitatively the results of Experiment 1a remain unchanged as all procedures only suffer slightly from the introduction of heteroskedasticity in the error terms.

Table 3 contains the results for Experiment 2a) in which the number of variables is slightly larger than the sample size. For $\rho = 0.5$ both versions of the conservative Lasso are more precise than the Lasso, have less size distortion and higher power. This is the case in particular for the variant of the conservative Lasso with indicator function weights. The coverage probability for the zero parameter is also higher. The procedure of Javanmard and Montanari (2014) is rather size distorted. When $\rho = 0.9$ the power of the

χ^2-test decreases for all Lasso based procedures. The procedure of Javanmard and Montanari (2014) suffers from severe size distortion. The conservative Lasso has a much better coverage rate, sometimes being more than ten percentage points larger for the zero parameter than the competitors. This comes from more precise parameter estimates and wider bands.

When adding heteroskedasticity to Experiment 2a, Table 4 shows that the estimation errors of all procedures increase slightly. The coverage rate of all procedures is roughly unchanged but the bands become wider.

The results for the very high-dimensional Experiment 3a are found in Table 5. Here GIC performs quite badly (for low values of n) for all methods and we thus focus on the results for BIC. When the sample size is $n = 100$, the plain Lasso has an estimation error which is 50% larger than the one of the conservative Lasso. Furthermore, the χ^2-test based on the Lasso is so size distorted (the size is 76%) that its usefulness may be questioned. While the conservative Lasso also suffers from size distortion (the size is 22%) it is still *much* more reliable than the Lasso. The version of the conservative Lasso lies in between in terms of estimation error and size of the χ^2-test. The procedure of Javanmard and Montanari (2014) is severely size distorted when $n = 100$ but this gradually improves as the sample size is increased.

Turning to the coverage rates of the confidence intervals of the non-zero coefficients, the Lasso provides such a poor coverage (25 %) that it may almost be deemed useless. The conservative Lasso, while not being perfect, still has a coverage of 83%. It also performs much better for the truly zero parameter than the Lasso. The superior coverage of conservative Lasso is due to much more precise parameter estimates and wider confidence bands than the Lasso. The coverage of the version of the conservative Lasso is higher than for the Lasso but lower than for the conservative Lasso.

When the sample size is increased to just $n = 150$ the conservative Lasso performs well along all dimensions even in this high-dimensional setting. The size distortion has disappeared and the coverage for the non-zero parameter has increased to 96% (from 83%). The Lasso has also improved. However, it is remarkable that the size of its χ^2-test for $n = 150$ is still higher than the one for the conservative Lasso when $n = 100$. Similarly, the coverage rate of the confidence bands for the zero as well as the non-zero parameters based on the Lasso is still lower than the one the conservative Lasso produced for $n = 100$.

Next, for $n = 200$, the conservative Lasso still estimates the parameters much more precisely than the plain Lasso. It also has better size and power properties but the gap has narrowed as these quantities approach their asymptotic values of 0.05 and 1, respectively. Regarding the coverage rate, the conservative Lasso also remains the superior procedure. The variant of the conservative Lasso now actually delivers the lowest estimation error which is in accordance with our initial observation of the variant performing relatively well as p/n decreases.

Finally, for $n = 500$, both procedures work very well, but the conservative Lasso remains by far the most precise estimator in terms of ℓ_2-estimation error (three times as precise as the plain Lasso). The size distortion of the procedure of Javanmard and Montanari (2014) is now only moderate while its confidence

bands still undercover the non-zero coefficient.

Table 6 adds heteroskedasticity to the results in Table 5. Qualitatively nothing changes in the sense that the rankings between the Lasso and the conservative Lasso remain the same in terms of estimation precision, size, power and coverage for all sample sizes. The conservative Lasso again estimates the parameters more precisely and has much better size and coverage properties. For $n = 500$ both procedures work well but as usual the conservative Lasso remains the most precise estimator in terms of ℓ_2-estimation error.

Table 7 considers the effect of testing a hypothesis involving many parameters. The results should be compared to those of Table 3. The main message is that the size of the Lasso based tests only inflates slightly compared to the case where only two parameters were involved in the hypothesis. Among the Lasso based tests the inflation is largest for the variant of the conservative Lasso. The size of Javanmard and Montanari (2014) increases by much more. Furthermore, the conservative Lasso is still found to slightly outperform the plain Lasso in terms of size and power.

6 Conclusion

This paper shows how the conservative Lasso can be used to conduct inference in the high-dimensional linear regression model. We allow for conditional heteroskedasticity in the error terms and also show how to consistently estimate the population covariance matrix in this case. In fact, the convergence is uniform over sparse sub vectors of the parameter space. Next, we show that the confidence bands based on the desparsified conservative are honest and that they contract at the optimal rate. This rate of contraction is also uniform over sparse sub vectors of the parameter space. χ^2-inference is also briefly discussed. Our simulations show that the conservative Lasso provides much more precise parameter estimates than the plain Lasso and that tests based on it have superior size properties. Furthermore, confidence intervals based on the desparsified conservative Lasso have better coverage rates than the ones based on the desparsified plain Lasso. Future work may include bootstrapping the desparsified conservative Lasso to gain further finite sample improvements.

Appendix

In Appendix A we begin by providing some auxiliary lemmas used for the proofs of the main results in Appendix B. The details of (18) can be found in Appendix C.

Appendix A – auxiliary lemmas

First, we provide the proof of Lemma 1 in the main text.

Proof of Lemma 1. (i). Note that by (A.55) with $H = \{1, ..., p\}$ it follows under Assumptions 1 and 2 that

$$\|\Theta\|_{\ell_\infty} = \max_{1 \leq j \leq p} \|\Theta_j\|_1 = O(\sqrt{\max_{1 \leq j \leq p} s_j}), \tag{A.1}$$

It actually also follows from (A.56) that (since $\hat{\Theta}_L$ is a subcase of $\hat{\Theta}$)

$$\|\hat{\Theta}_L\|_{\ell_\infty} = \max_{1\le j\le p} \|\hat{\Theta}_{L,j}\|_1 = O_p(\sqrt{\max_{1\le j\le p} s_j}), \tag{A.2}$$

Thus,

$$\lambda_{prec} = O(\lambda_n \sqrt{\max_{1\le j\le p} s_j}) \tag{A.3}$$

where

$$\lambda_n \sqrt{\max_{1\le j\le p} s_j} = \frac{Mp^{2/r}}{\sqrt{n}} \sqrt{\max_{1\le j\le p} s_j} = M \left[\frac{p^2 (\max_{1\le j\le p} s_j)^{r/2}}{n^{r/4}} \right]^{1/r} \frac{1}{n^{1/4}} \to 0, \tag{A.4}$$

by Assumption 2b. Therefore, we get $\lambda_{prec} \to 0$. Note that replacing $\|\Theta\|_{\ell_\infty}$ by $\|\hat{\Theta}_L\|_{\ell_\infty}$ in the definition of λ_{prec} makes no difference since by (A.2) we still get $\lambda_{prec} \xrightarrow{p} 0$.

(ii). By Lemma A.7 the set $\mathcal{C}_1 = \{\|\hat{\beta}_L - \beta_0\|_\infty \le \lambda_{prec}\}$ has probability approaching one. First, note that on $\mathcal{C}_1$ one has $\max_{j\in S_0^c} |\hat{\beta}_{L,j}| = \max_{j\in S_0^c} |\hat{\beta}_{L,j} - \beta_{0,j}| \le \lambda_{prec}$. Thus, $\min_{j\in S_0^c} \hat{w}_j = 1$ on $\mathcal{C}_1$.

(iii). On $\mathcal{C}_1$

$$\min_{j\in S_0} |\hat{\beta}_{L,j}| \quad \ge \quad |\beta_{0,j}| - |\hat{\beta}_{L,j} - \beta_{0,j}| \ge \min_{j\in S_0} (|\beta_{0,j}| - \lambda_{prec}) = \lambda_{prec} \min_{j\in S_0} \left[\left| \frac{\beta_{0,j}}{\lambda_{prec}} \right| - 1 \right]. \tag{A.5}$$

Thus, since $\min_{j\in S_0} |\beta_{0,j}|/\lambda_{prec} \to \infty$ we have that $\min_{j\in S_0} |\hat{\beta}_{L,j}| \ge \lambda_{prec}$ for n sufficiently large. Hence, by (A.5), on $\mathcal{C}_1$ which has probability tending to one,

$$\max_{j\in S_0} \hat{w}_j = \frac{\lambda_{prec}}{\min_{j\in S_0} |\hat{\beta}_{L,j}| \vee \lambda_{prec}} = \frac{\lambda_{prec}}{\min_{j\in S_0} |\hat{\beta}_{L,j}|} \le \frac{1}{\min_{j\in S_0} \frac{|\beta_{0,j}|}{\lambda_{prec}} - 1} \to 0. \tag{A.6}$$

□

Now, we provide an oracle inequality for a general weighted Lasso which satisfies certain assumptions and then utilize that the plain Lasso and the conservative Lasso satisfy these assumptions. Define

$$\hat{\beta}_w = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left(\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\beta_j| \right),$$

where $\hat{w}_{g,j}$ denotes a general weight. When $\hat{w}_{g,j} = 1$ one recovers the Lasso, when $\hat{w}_{g,j} = \hat{w}_j$ the result is the conservative Lasso. In particular, we shall work on the intersection of $\mathcal{A} = \left\{ \|X'u/n\|_\infty \le \lambda_n/2 \right\}$ and $\mathcal{B} = \left\{ \phi_{\hat{\Sigma}}^2 \ge \phi_\Sigma^2/2 \right\}$. On these sets we have a handle on the maximal empirical "correlation" between the covariates and the error terms, and a lower bound on the empirical adaptive restricted eigenvalue, respectively. Define $a_n = \|\hat{w}_{S_0}\|_\infty$.

Lemma A.1. *Let $\hat{w}_{g,S_0^c}^{min} = \min_{j\in S_0^c} \hat{w}_j = 1$ and $a_n \le 1$. Then, on the set $\mathcal{A} \cap \mathcal{B}$ the following inequalities are valid.*

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 \le 2(2a_n + 1)^2 \frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}. \tag{A.7}$$

$$\|\hat{\beta}_w - \beta_0\|_1 \le 4(a_n + 1)(2a_n + 1) \frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}. \tag{A.8}$$

Proof. We begin by establishing (A.7). By the minimizing property of $\hat{\beta}_w$ it follows that

$$\|Y - X\hat{\beta}_w\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\hat{\beta}_{w,j}| \le \|Y - X\beta_0\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\beta_{0,j}|. \tag{A.9}$$

Inserting $Y = X\beta_0 + u$, using Hölder's inequality, and using that we are on the set $\mathcal{A}$ we arrive at

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \hat{w}_{g,j}|\hat{\beta}_{w,j}| \leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 + 2\lambda_n \sum_{j=1}^{p} \hat{w}_{g,j}|\beta_{0,j}|. \tag{A.10}$$

Then, using $\|\hat{\beta}_w\|_1 = \|\hat{\beta}_{w,S_0}\|_1 + \|\hat{\beta}_{w,S_0^c}\|_1$ one gets

$$\begin{aligned}\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \sum_{j\in S_0^c} \hat{w}_{g,j}|\hat{\beta}_{w,j}| &\leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 - 2\lambda_n \sum_{j\in S_0} \hat{w}_{g,j}|\hat{\beta}_{w,j}| + 2\lambda_n \sum_{j=1}^{p} \hat{w}_{g,j}|\beta_{0,j}| \\ &\leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 + 2\lambda_n \sum_{j\in S_0} \hat{w}_{g,j}|\hat{\beta}_{w,j} - \beta_{0,j}|. \end{aligned} \tag{A.11}$$

Noting that $\|\hat{\beta}_w - \beta_0\|_1 = \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1 + \|\hat{\beta}_{w,S_0^c}\|_1$ and $\sum_{j\in S_0^c} \hat{w}_{g,j}|\hat{\beta}_{w,j}| \geq \hat{w}_{S_0^c}^{min}\|\hat{\beta}_{w,S_0^c}\|_1 = \|\hat{\beta}_{w,S_0^c}\|$ rewrite (A.11) as

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq \lambda_n \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 + 2\lambda_n \sum_{j\in S_0} \hat{w}_{g,j}|\hat{\beta}_{w,j} - \beta_{0,j}|. \tag{A.12}$$

Subtract $\lambda_n \|\hat{\beta}_{w,S_0^c}\|_1$ from both sides of (A.12) to get

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq \lambda_n \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1 + 2\lambda_n \sum_{j\in S_0} \hat{w}_{g,j}|\hat{\beta}_{w,j} - \beta_{0,j}|. \tag{A.13}$$

Next, use the Cauchy-Schwarz inequality, $\|.\|_1 \leq \sqrt{s_0}\|.\|_2$, as well as $\|\hat{w}_{g,S_0}\|_2 \leq a_n\sqrt{s_0}$, and $0 < a_n \leq 1$ to get

$$\begin{aligned}\|X(\hat{\beta}_w - \beta_0)\|_n^2 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 &\leq \lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2 + 2\lambda_n \|\hat{w}_{g,S_0}\|_2 \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2 \\ &\leq (2a_n + 1)\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2 \qquad \text{(A.14)} \\ &\leq 3\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2. \qquad \text{(A.15)}\end{aligned}$$

(A.15) implies that

$$\|\hat{\beta}_{w,S_0^c}\|_1 \leq 3\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2.$$

Hence, by the adaptive restricted eigenvalue condition, (A.14) implies

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq (2a_n + 1)\lambda_n\sqrt{s_0}\frac{\|X(\hat{\beta}_w - \beta_0)\|_n}{\phi_{\hat{\Sigma}}(s_0)}. \tag{A.16}$$

Then, using $(2a_n + 1)uv \leq u^2/2 + (2a_n + 1)^2v^2/2$, with $v = \lambda_n\sqrt{s_0}/\phi_{\hat{\Sigma}}(s_0)$, $u = \|X(\hat{\beta}_w - \beta_0)\|_n$, one gets

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq \frac{\|X(\hat{\beta}_w - \beta_0)\|_n^2}{2} + \frac{(a_n + 1)^2}{2}\frac{\lambda_n^2 s_0}{\phi_{\hat{\Sigma}}^2(s_0)}. \tag{A.17}$$

Subtracting the first right hand side term in (A.17) from the left and right hand sides of (A.17) and multiplying all terms by 2 yields

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq (2a_n + 1)^2\frac{\lambda_n^2 s_0}{\phi_{\hat{\Sigma}}^2(s_0)}, \tag{A.18}$$

which, using that we are on $\mathcal{B}$, implies (A.7).

Next, we turn to proving (A.8). By adding $\lambda_n \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1$ to both sides of (A.14) and using $\|.\|_1 \leq \sqrt{s_0}\|.\|_2$ one gets

$$\begin{aligned}\lambda_n \|\hat{\beta}_w - \beta_0\|_1 &\leq \lambda_n \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1 + (2a_n + 1)\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2 \qquad \text{(A.19)} \\ &\leq 2(a_n + 1)\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_2. \qquad \text{(A.20)}\end{aligned}$$

The adaptive restricted eigenvalue condition and inequality (A.7) of this Lemma yield (using that $\mathcal{B}$ occurs)

$$\|\hat{\beta}_w - \beta_0\|_1 \leq 2(a_n+1)\sqrt{s_0}\frac{\|X(\hat{\beta}_w - \beta_0)\|_n}{\phi_{\hat{\Sigma}}(s_0)} \leq \frac{4(a_n+1)(2a_n+1)s_0\lambda_n}{\phi^2_{\hat{\Sigma}}(s_0)}, \tag{A.21}$$

which is (A.8). □

To prove Lemma A.8 and Theorem 1 it suffices to provide a lower bound on the probabilities of $\mathcal{A}$ and $\mathcal{B}$. To do so, recall the Marcinkiewicz-Zygmund inequality:

Lemma A.2. *[Marcinkiewicz-Zygmund inequality, see Lin and Bai (2010), result 9.7.a] Let $\{U_i\}_{i=1}^n$ be a sequence of independent mean zero real random variables with finite r'th moment. Then, for positive constants a_r and b_r, only depending on r, $r \geq 2$*

$$a_r E\left(\sum_{i=1}^n U_i^2\right)^{r/2} \leq E\left|\sum_{i=1}^n U_i\right|^r \leq b_r E\left(\sum_{i=1}^n U_i^2\right)^{r/2} \tag{A.22}$$

Note in particular that, by an application of the summation version of Jensen's inequality on the convex map $x \mapsto x^{r/2}$, (A.22) implies that

$$E\left|\sum_{i=1}^n U_i\right|^r \leq b_r n^{r/2} E\left(\frac{1}{n}\sum_{i=1}^n U_i^2\right)^{r/2} \leq b_r n^{r/2-1}\sum_{i=1}^n E|U_i|^r \leq b_r n^{r/2}\max_{1\leq i\leq n} E\left|U_i\right|^r.$$

Hence, by a union bound and Markov's inequality we arrive at the following result which we shall use frequently throughout the appendix.

Lemma A.3. *For each $j \in \{1,...,m\}$ let $\left\{U_{j,i}\right\}_{i=1}^n$ be a sequence of independent mean zero real random variables with finite r'th moment and define $S_{j,n} = \sum_{i=1}^n U_{j,i}$. Then,*

$$P\left(\max_{1\leq j\leq m}|S_{j,n}| \geq t\right) \leq b_r m\frac{n^{r/2}\max_{1\leq j\leq m}\max_{1\leq i\leq n}E|U_{j,i}|^r}{t^r}.$$

Remarks: 1. In Lemma A.3 above we used the Marcinkiewicz-Zygmund inequality. Another common approach is using Nemirowski's inequality, see van de Geer et al. (2014). We show that application of Nemirowski's inequality will bring an additional $\left(8\log(2m)\right)^{r/2}$ in Lemma A.3. To make this point clear, for $r \geq 2$, note that Nemirovski's inequality in Lemma 14.24 of van de Geer et al. (2014) yields

$$E\left(\max_{1\leq j\leq m}|S_{j,n}|^r\right) \leq \left(8\log(2m)\right)^{r/2} E\left[\max_{1\leq j\leq m}\sum_{i=1}^n U_{j,i}^2\right]^{r/2}. \tag{A.23}$$

Thus, we need to bound $E\left[\max_{1\leq j\leq m}\sum_{i=1}^n U_{j,i}^2\right]^{r/2}$. By convexity of $x \mapsto x^{r/2}$ and Jensen's inequality

$$\begin{aligned} E\left[\max_{1\leq j\leq m}\sum_{i=1}^n U_{j,i}^2\right]^{r/2} &= n^{r/2}E\max_{1\leq j\leq m}\left[\frac{1}{n}\sum_{i=1}^n U_{j,i}^2\right]^{r/2} \leq n^{r/2}E\max_{1\leq j\leq m}\frac{1}{n}\sum_{i=1}^n |U_{j,i}|^r \\ &\leq n^{r/2-1}E\sum_{j=1}^m\sum_{i=1}^n |U_{j,i}|^r \leq n^{r/2}m\max_{1\leq j\leq m}\max_{1\leq i\leq n}E|U_{j,i}|^r. \end{aligned}$$

Inserting the above display into (A.23) and using Markov's inequality yields

$$P\left(\max_{1\leq j\leq m}|S_{j,n}| \geq t\right) \leq \frac{\left(8\log(2m)\right)^{r/2} n^{r/2}m\max_{1\leq j\leq m}\max_{1\leq i\leq n}E|U_{j,i}|^r}{t^r}.$$

Note that the above bound, relying on Nemirovski's inequality, is larger by a factor $\left(8\log(2m)\right)^{r/2}$ (which increases in m) than the bound in Lemma A.3. Thus, Lemma A.3 results in lower choices of the tuning parameter and hence sharper bounds. This is a new theoretical contribution of the paper.

2. In a seminal paper about optimal instrumental variable selection, Belloni et al. (2012) use self-normalized moderate deviation results to get the tuning parameter and its rate. They propose a heteroskedasticity consistent penalty term unlike our data dependent penalty which focuses on creating a wedge between zero and nonzero parameters. Condition RF (iii) in the analysis of Belloni et al. (2012) results in $\log^3(p)/n = o(1)$. However, our rate for λ_n will require $p^{2/r}/n^{1/2} \to 0$. The reason for this is we are interested in maxima of sums, as in the previous lemma, unlike Belloni et al. (2012) who use maxima of self normalized sum (i.e. sum normalized by the ℓ_2 norm of the vector of variables) which provides their rate.

We are now ready to provide a lower bound on the probability of $\mathcal{A}$.

Lemma A.4. *Let $M > 0$ be an arbitrary positive number. Then, under Assumption 1, for $\lambda_n = M\frac{p^{2/r}}{\sqrt{n}}$ the set $\mathcal{A} = \left\{\|X'u/n\|_\infty \leq \lambda_n/2\right\}$ has probability at least $1 - \frac{C}{M^{r/2}}$, for a universal constant $C > 0$.*

Proof. For each $j \in \{1, ..., p\}$, $\{X_{j,i}u_i\}_{i=1}^n$ is a sequence of independent mean zero random variables with $(r/2)'th$ moment $E|X_{j,i}u_i|^{r/2} \leq \sqrt{E|X_{j,i}|^r E|u_i|^r} \leq C$. Hence, Lemma A.3 yields

$$P(\mathcal{A}^c) = P\left(\|X'u\|_\infty > n\lambda_n/2\right) \leq p\frac{b_{r/2}Cn^{r/4}}{(n\lambda_n/2)^{r/2}} = \frac{C}{M^{r/2}},$$

where the last equality follows from the choice of λ_n and has merged the constants. □

The next two lemmas will provide a lower bound on the probability of set $\mathcal{B}$.

Lemma A.5. *Let A and B be two positive semi-definite $p \times p$ matrices and assume that A satisfies the restricted eigenvalue condition RE(s) for some $\phi_A(s) > 0$. Then, for $\delta = \max_{1\leq i,j\leq p}|A_{i,j} - B_{i,j}|$, one also has $\phi_B^2 \geq \phi_A^2 - 16s\delta$.*

Proof. The proof is similar to Lemma 10.1 in van de Geer and Bühlmann (2009). For any (non-zero) $p \times 1$ vector v such that $\|v_{S^c}\|_1 \leq 3\sqrt{s}\|v_S\|_2$ one has

$$\begin{aligned} v'Av - v'Bv \leq |v'Av - v'Bv| = |v'(A-B)v| &\leq \|v\|_1 \|(A-B)v\|_\infty \leq \delta\|v\|_1^2 \\ &= \delta\left(\|v_S\|_1 + \|v_{S^c}\|_1\right)^2 \leq \delta 16s\|v_S\|_2^2. \end{aligned}$$

Hence, rearranging the above, yields

$$v'Bv \geq v'Av - 16s\delta\|v_S\|_2^2,$$

or equivalently,

$$\frac{v'Bv}{v_S'v_S} \geq \frac{v'Av}{v_S'v_S} - 16s\delta.$$

Minimizing over $\{v \in \mathbb{R}^n \setminus \{0\} : \|v_{S^c}\|_1 \leq 3\sqrt{s}\|v_S\|_2\}$ and using the adaptive restricted eigenvalue condition yields the claim. □

In order to verify the restricted eigenvalue condition we present the following lemma.

Lemma A.6. *Let Assumption 1 be satisfied. Then, the set $\mathcal{B} = \left\{\phi_{\hat{\Sigma}}^2 \geq \phi_\Sigma^2/2\right\}$ has probability at least $1 - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$ for a universal constant $D > 0$.*

Proof. By Lemma A.5, with $s = s_0$, it suffices to show that $\delta = \|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{\phi^2_\Sigma(s_0)}{32 s_0}$. The (k,l) entry of $\hat{\Sigma} - \Sigma$ is given by $\frac{1}{n}\sum_{i=1}^n \left(X_{k,i}X_{l,i} - E(X_{k,i}X_{l,i})\right)$. Each summand has mean zero and $E\left|X_{k,i}X_{l,i} - E(X_{k,i}X_{l,i})\right|^{r/2}$ is bounded by a universal constant D by the Cauchy-Schwarz inequality. Hence, merging constants, Lemma A.3 yields

$$P(\mathcal{B}^c) \leq P\left(\|\hat{\Sigma} - \Sigma\|_\infty > \frac{\phi^2_\Sigma(s_0)}{32 s_0}\right) \leq p^2 \frac{D n^{r/4}}{(\frac{n}{s_0})^{r/2}} = D\frac{p^2 s_0^{r/2}}{n^{r/4}}.$$

□

Lemma A.7. *Let Assumption 1 be satisfied. Then on $\mathcal{A} \cap \mathcal{B}$ (defined prior to Lemma A.1)*

$$\|\hat{\beta}_L - \beta_0\|_\infty \leq (\frac{9\lambda_n}{4})\|\Theta\|_{\ell_\infty}, \tag{A.24}$$

and $\mathcal{A} \cap \mathcal{B}$ occurs with probability at least $1 - \frac{C}{M^{r/2}} - \frac{D p^2 s_0^{r/2}}{n^{r/4}}$.

Proof. By Lemma 2.5.1 of van de Geer (2014)

$$\|\hat{\beta}_L - \beta_0\|_\infty \leq \|\Theta\|_{\ell_\infty}\left[\frac{\|X'u\|_\infty}{n} + \|\hat{\Sigma} - \Sigma\|_\infty \|\hat{\beta}_L - \beta_0\|_1 + \lambda_n\right]. \tag{A.25}$$

Using Lemma A.8 (see below) we get that on $\mathcal{A} \cap \mathcal{B}$

$$\|\hat{\beta}_L - \beta_0\|_\infty \leq \|\Theta\|_{\ell_\infty}\left[\frac{\lambda_n}{2} + \left(\frac{\phi^2_\Sigma(s_0)}{32 s_0}\right)\left(\frac{24\lambda_n s_0}{\phi^2_\Sigma(s_0)}\right) + \lambda_n\right], \tag{A.26}$$

which provides the result after some simple algebra and upon using that Lemmas A.4, A.6 give the lower bound on the probability of $\mathcal{A} \cap \mathcal{B}$. □

Appendix B

This appendix provides the proofs of the main theorems.

We state the following result on the Lasso. It is very similar to the classical oracle inequality for the Lasso that assumes subgaussianity of the error terms in Bickel et al. (2009). However, it is tailored to our Assumption 1 which only assumes r moments of the covariates and the error terms and hence we still mention it here. Furthermore, the result is needed in order to guide our choice of λ_{prec} for the conservative Lasso.

Lemma A.8. *Let Assumption 1 be satisfied and set $\lambda_n = M\frac{p^{2/r}}{n^{1/2}}$ for $M > 0$. Then, with probability at least $1 - \frac{C}{M^{r/2}} - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$, the Lasso satisfies the following inequalities*

$$\|X(\hat{\beta}_L - \beta_0)\|_n^2 \leq 18\frac{\lambda_n^2 s_0}{\phi^2_\Sigma(s_0)}, \tag{A.27}$$

$$\|\hat{\beta}_L - \beta_0\|_1 \leq 24\frac{\lambda_n s_0}{\phi^2_\Sigma(s_0)}, \tag{A.28}$$

for universal constants $C, D > 0$. Furthermore, these bounds are valid uniformly over the ℓ_0-ball $\mathcal{B}_{\ell_0}(s_0) = \left\{\|\beta_0\|_{\ell_0} \leq s_0\right\}$.

Proof of Lemma A.8. The Lasso corresponds to $\hat{w}_j = 1$ for all $j = 1, ..., p$. Thus, Lemma A.1 combined with the lower bounds on the probabilities of the sets $\mathcal{A}$ and $\mathcal{B}$ from Lemmas A.4 and A.6 yields (A.27) and (A.28). The uniformity over $\mathcal{B}_{\ell_0}(s_0)$ follows by noting that the right hand sides of (A.27) and (A.28) only depend on β_0 through s_0.

□

Proof of Theorem 1. The oracle inequalities will follow upon verifying the conditions of Lemma A.1 and showing that $\mathcal{A} \cap \mathcal{B}$ has high probability. As all weights of the conservative Lasso as are less than or equal to one it remains to show that $\min_{j \in S_0^c} \hat{w}_j = 1$. To this end Lemma A.7 (which uses only Assumption 1) shows that $\max_{j \in S_0^c} |\hat{\beta}_{L,j}| \leq \lambda_{prec} = \frac{9\lambda_n}{4} \|\Theta\|_{\ell_\infty}$ on $\mathcal{A} \cap \mathcal{B}$ such that $\min_{j \in S_0^c} \hat{w}_j = 1$. The lower bound on $\mathcal{A} \cap \mathcal{B}$ follows from Lemmas A.4 and A.6. The uniformity over $\mathcal{B}_{\ell_0}(s_0)$ follows by noting that the right hand sides of (4) and (5) only depend on β_0 through s_0. □

Θ's relation to the regression coefficients

In order to establish a central limit theorem for $\alpha' \hat{\Theta} X' u / n^{1/2}$ in (11) we need to understand the asymptotic properties of $\hat{\Theta}$. To do so we relate $\hat{\Theta}$ to $\Theta := \Sigma^{-1}$. First, let $\Sigma_{-j,-j}$ represent the $(p-1) \times (p-1)$ submatrix of Σ where the jth row and column have been removed. $\Sigma_{j,-j}$ is the jth row of Σ with jth element of that row removed. $\Sigma_{-j,j}$ represent the j th column of Σ with its jth element removed. By Section 2.1 of Yuan (2010) we know that

$$\Theta_{j,j} = \left(\Sigma_{j,j} - \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} \right)^{-1}$$

and

$$\Theta_{j,-j} = - \left(\Sigma_{j,j} - \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} \right)^{-1} \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} = -\Theta_{j,j} \Sigma_{j,-j} \Sigma_{-j,-j}^{-1}$$

Next, let $X_{j,i}$ denote the ith element of X_j and $X_{-j,i}$ the ith element of X_{-j} (recall the definition of X_j and X_{-j} just prior to (13)). Now, defining γ_j as the value of γ minimizing,

$$E \left(X_{j,i} - X_{-j,i} \gamma \right)^2$$

implies that

$$\gamma_j' = \Sigma_{j,-j} \Sigma_{-j,-j}^{-1}$$

such that

$$\Theta_{j,-j} = -\Theta_{j,j} \gamma_j'. \tag{A.29}$$

Thus, for $\eta_{j,i} := X_{j,i} - X_{-j,i} \gamma_j$, it follows from the definition of γ_j as an L^2-projection that all entries of $X_{-j,i} \eta_{j,i}$ have mean zero such that

$$X_{j,i} = X_{-j,i} \gamma_j + \eta_{j,i} \tag{A.30}$$

is a regression model with covariates orthogonal in L^2 to the error terms for all $j = 1, ..., p$ and $i = 1, ..., n$. Let Θ_j be the j'th row of Θ written as a column vector. Then the crux is that (A.30) is sparse if and only if Θ_j is sparse as can be seen from (A.29). Let $S_j = \left\{ k = 1, ..., p : \Theta_{j,k} \neq 0 \right\}$ with cardinality $s_j = |S_j|$ denote the indices of the non-zero terms of Θ_j. Then, the regression model (A.30) will also be sparse with γ_j possessing s_j non-zero entries. Thus, with Theorem 1 in mind it is sensible that the estimator $\hat{\gamma}_j$ resulting from (14) is close to γ_j. We make this claim rigorous in Lemma A.9. Next, by (A.30),

$$\Sigma_{j,j} = E(X_{j,i}^2) = \gamma_j' \Sigma_{-j,-j} \gamma_j + E(\eta_{j,i}^2) = \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} + E(\eta_{j,i}^2),$$

such that

$$\tau_j^2 := E(\eta_{j,i}^2) = \Sigma_{j,j} - \Sigma_{j,-j} \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} = \frac{1}{\Theta_{j,j}}.$$

Thus, defining

$$C = \begin{pmatrix} 1 & -\gamma_{1,2} & \cdots & -\gamma_{1,p} \\ -\gamma_{2,1} & 1 & \cdots & -\gamma_{2,p} \\ \cdots & \cdots & \ddots & \cdots \\ -\gamma_{p,1} & -\gamma_{p,2} & \cdots & 1 \end{pmatrix},$$

and $T^2 = diag(\tau_1^2, \cdots, \tau_p^2)$ we can write $\Theta = T^{-2}C$ using (A.29). In Lemma A.9 we show that $\hat{\tau}_j^2$ as defined in (16) is close to τ_j^2 such that $\hat{\Theta}_j$ is close to Θ_j when $\hat{\gamma}_j$ is close to γ_j.

Remark: The above arguments have relied on X_i being i.i.d. such that $\Sigma = E\left(X_i X_i'\right)$ is constant and does not depend on $i = 1, ..., n$. At the cost of more involved notation and proofs the arguments above would also be valid in the case of non-identically distributed covariates if we consider $\Sigma = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i X_i'\right)$ instead of $E(X_1 X_1')$. However, we shall not pursue this generalization here.

We can now state the asymptotic properties of $\hat{\Theta}$.

Lemma A.9. *Let Assumptions 1 and 2 be satisfied and set* $\lambda_{node,n} \asymp \frac{h^{2/r}p^{2/r}}{n^{1/2}}$. *Then,*

$$\max_{j\in H} \|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2 = O_p\left(\frac{d_{n1}\bar{s}h^{4/r}p^{4/r}}{n}\right). \tag{A.31}$$

$$\max_{j\in H} \|\hat{\gamma}_j - \gamma_j\|_1 = O_p\left(\frac{d_{n2}\bar{s}h^{2/r}p^{2/r}}{n^{1/2}}\right). \tag{A.32}$$

$$\max_{j\in H} |\hat{\tau}_j^2 - \tau_j^2| = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.33}$$

$$\max_{j\in H} \left\|\hat{\Theta}_j - \Theta_j\right\|_1 = O_p\left(d_{n2}\bar{s}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.34}$$

$$\max_{j\in H} \|\hat{\Theta}_j - \Theta_j\|_2 = O_p\left(\sqrt{d_{n1}}\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.35}$$

$$\max_{j\in H} \|\hat{\Theta}_j\|_1 = O_p(\bar{s}^{1/2}). \tag{A.36}$$

Remark. Clearly we see that divergences d_{n1} and d_{n2} between the Lasso and the conservative Lasso influence the upper bounds in the nodewise regressions. The roles of d_{n1} and d_{n2} are explained in detail in Remark 3 after Theorem 1. Clearly we see that the conservative nodewise regression Lasso can have smaller errors in prediction norm, ℓ_1 and ℓ_2 errors for estimates than the its Lasso counterpart since $d_{n1} = 18$ for the Lasso and as low as nearly 2 for the former. Furthermore, d_{n2} is 24 in the Lasso nodewise regression and as small as almost 4 in conservative Lasso nodewise regression as also explained in the Remarks to Theorem 1.

Lemma A.9 is an auxiliary lemma which will be of great importance in the proof of Theorem 2 below. Note that all bounds provided are uniform in H with upper bounds tending to zero even when $h = |H| \to \infty$ as long as this does not happen too fast. (A.31) and (A.32) reduce to inequalities of the type (4) and (5) in Theorem 1 when H is a singleton such that $h = 1$. Note also that (A.34) can be used to bound the estimation error of each row of $\hat{\Theta}$ for the corresponding row of Θ. Thus, choosing $H = \{1, ..., p\}$, (A.34) provides a bound on $\|\hat{\Theta} - \Theta\|_{\ell_\infty}$. Finally, we remark that the uniformity of the above results is crucial for establishing the limiting distribution of $\alpha'\hat{\Theta}X'u/n^{1/2}$ in (11) as well as for estimating the variance of the limiting distribution.

Proof of Lemma A.9. We start by establishing the order of magnitude of $\|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2$ and $\|\hat{\gamma}_j - \gamma_j\|_1$.

For concreteness, consider nodewise regression j. Define

$$\mathcal{A}_{node} = \left\{\max_{j\in H} \|X'_{-j}\eta_j\|_\infty \leq \lambda_{node,n}/2\right\} \text{ and } \mathcal{B}_j = \left\{\phi^2_{\hat{\Sigma}_{-j,-j}}(s_j) \geq \phi^2_{\Sigma_{-j,-j}}(s_j)/2\right\}.$$

By an exact adaptation of the proof of Lemma A.1 it can be shown for each $j \in H$ that with definition of $d_{n1} = 2(2a_n+1)^2$, and $d_{n2} = 4(a_n+1)(2a_n+1)$, $0 < a_n \leq 1$

$$\|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2 \leq d_{n1}\frac{\lambda^2_{node,n}s_j}{\phi^2_\Sigma(s_j)}, \tag{A.37}$$

$$\|\hat{\gamma}_j - \gamma_j\|_1 \leq d_{n2}\frac{\lambda_{node,n}s_j}{\phi^2_\Sigma(s_j)} \tag{A.38}$$

are valid on the set $\mathcal{A}_{node} \cap \mathcal{B}_j$ for $j \in H$.

Note that (A.37) and (A.38) are valid simultaneously for all $j \in H$ on $\mathcal{A}_{node} \cap (\cap_{j\in H}\mathcal{B}_j)$ [8]. Thus, we establish a lower bound on the probability of this set. First, consider $\mathcal{A}_{node}$. Since $\eta_{j,i}$ is the residual from the L^2-projection of $X_{j,i}$ on the linear span of the elements of $X_{-j,i}$ it follows that $E(X_{-j,i}\eta_{j,i}) = 0$ for all $i = 1, ..., n$ and all $j \in H$. Furthermore, by the Cauchy-Schwarz inequality, every entry of $X_{-j,i}\eta_{j,i}$ has bounded $r/2$-norm via Assumption 2c. The maximum in the definition of $\mathcal{A}_{node}$ is over $h(p-1)$ terms. Thus, merging constants and choosing $\lambda_{node,n} = M\frac{h^{2/r}p^{2/r}}{\sqrt{n}}$ for some $M > 0$, Lemma A.3 yields,

$$P(\mathcal{A}^c_{node}) = P\left(\max_{j\in H}\|X'_{-j}\eta_j\|_\infty > n\lambda_{node,n}/2\right) \leq hp\frac{b_r C^2 n^{r/4}}{(n\lambda_{node,n}/2)^{r/2}} = \frac{C}{M^{r/2}},$$

which also shows that

$$\max_{j\in H}\|X'_{-j}\eta_j/n\|_\infty = O_p\left(\lambda_{node,n}\right) = O_p\left(\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right) \tag{A.39}$$

by choosing M sufficiently large.

Next, we provide a lower bound on the probability of the set $\cap_{j\in H}\mathcal{B}_j$. We know by Lemma A.5 that $\left\{\|\hat{\Sigma}_{-j,-j} - \Sigma_{-j,-j}\|_\infty \leq \frac{\phi^2_{\Sigma_{-j,-j}}(s_j)}{32s_j}\right\} \subseteq \left\{\phi^2_{\hat{\Sigma}_{-j,-j}}(s_j) \geq \phi^2_{\Sigma_{-j,-j}}(s_j)/2\right\} = \mathcal{B}_j$. Thus, the relation

$$\|\hat{\Sigma}_{-j,-j} - \Sigma_{-j,-j}\|_\infty \leq \|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{\phi^2_\Sigma(\bar{s})}{32\bar{s}} \leq \frac{\phi^2_{\Sigma_{-j,-j}}(s_j)}{32s_j}$$

implies that $\left\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{\phi^2_\Sigma(\bar{s})}{32\bar{s}}\right\} \subseteq \mathcal{B}_j$ for all $j \in H$ and therefore $\left\{\|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{\phi^2_\Sigma(\bar{s})}{32\bar{s}}\right\} \subseteq \cap_{j\in H}\mathcal{B}_j$.

Next, by arguments exactly parallel to those in Lemma A.6, it follows that

$$P\left(\left(\cap_{j\in H}\mathcal{B}_j\right)^c\right) \leq P\left(\|\hat{\Sigma} - \Sigma\|_\infty > \frac{\phi^2_\Sigma(\bar{s})}{32\bar{s}}\right) \leq D\frac{p^2\bar{s}^{r/2}}{n^{r/4}}.$$

Hence, with probability at least $1 - \frac{C}{M^{r/2}} - D\frac{p^2\bar{s}^{r/2}}{n^{r/4}}$

$$\|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2 \leq d_{n1}\frac{\lambda^2_{node,n}s_j}{\phi^2_\Sigma(s_j)}. \tag{A.40}$$

$$\|\hat{\gamma}_j - \gamma_j\|_1 \leq d_{n2}\frac{\lambda_{node,n}s_j}{\phi^2_\Sigma(s_j)}. \tag{A.41}$$

By choosing M sufficiently large, using $\frac{p^2\bar{s}^{r/2}}{n^{r/4}} \to 0$, and inserting the definition of $\lambda_{node,n}$ (A.31) and (A.32) follow upon taking the maximum in the above display and utilizing that the above inequalities are all valid simultaneously on $\mathcal{A}_{node,n} \cap \left(\cap_{j\in H}\mathcal{B}_j\right)$.

[8] It will turn out later that it is quite important that (A.37) and (A.38) are valid simultaneously for all $j \in H$ since this will give us a vital uniformity when bounding $\hat{\tau}_j^2$ away from 0. If one is only interested in one nodewise regression the outer maximum in the definition of $\mathcal{A}_{node}$ can be omitted.

We shall also need an upper bound on $\max_{j\in H}\|\hat{\gamma}_j-\gamma_j\|_2$ in the proof of Theorem 2. Let $\hat{v}_j$ and v_j be $p\times 1$ vectors containing 0 in the j'th position and the elements of $\hat{\gamma}_j$ and γ_j, respectively, in the remaining positions in the same order as they appear in $\hat{\gamma}_j$ and γ_j. Thus, $\max_{j\in H}\|\hat{\gamma}_j-\gamma_j\|_2=\max_{j\in H}\|\hat{v}_j-v_j\|_2$. Thus,

$$|(\hat{v}_j-v_j)'\hat{\Sigma}(\hat{v}_j-v_j)-(\hat{v}_j-v_j)'\Sigma(\hat{v}_j-v_j)|\leq\|\hat{\Sigma}-\Sigma\|_\infty\|\hat{v}_j-v_j\|_1^2$$

such that

$$\max_{j\in H}(\hat{v}_j-v_j)'\Sigma(\hat{v}_j-v_j)\leq\max_{j\in H}(\hat{v}_j-v_j)'\hat{\Sigma}(\hat{v}_j-v_j)+\max_{j\in H}\|\hat{\Sigma}-\Sigma\|_\infty\|\hat{v}_j-v_j\|_1^2. \tag{A.42}$$

Next, we bound each term on the right hand side of the above display. First,

$$\max_{j\in H}(\hat{v}_j-v_j)'\hat{\Sigma}(\hat{v}_j-v_j)=\max_{j\in H}\left\|X(\hat{v}_j-v_j)\right\|_n^2=\max_{j\in H}\left\|X_{-j}(\hat{\gamma}_j-\gamma_j)\right\|_n^2=O_p\left(\frac{d_{n1}\bar{s}h^{4/r}p^{4/r}}{n}\right),$$

by (A.31). Next, consider the second term in (A.42). To this end, apply Lemma A.3 and Assumption 1, for any $t>0$ to get

$$P\left(\|\hat{\Sigma}-\Sigma\|_\infty>t\right)=P\left(\max_{1\leq k,l\leq p}\left|\frac{1}{n}\sum_{i=1}^n\left(X_{k,i}X_{l,i}-E(X_{k,i}X_{l,i})\right)\right|>t\right)\leq b_{r/2}\frac{p^2n^{r/4}C}{(tn)^{r/2}}.$$

Thus, choosing $t=M\frac{p^{4/r}}{n^{1/2}}$ for $M>0$ sufficiently large yields

$$\|\hat{\Sigma}-\Sigma\|_\infty=O_p\left(\frac{p^{4/r}}{n^{1/2}}\right). \tag{A.43}$$

In combination with (A.32) this implies (using $\|\hat{\gamma}_j-\gamma_j\|_1=\|\hat{v}_j-v_j\|_1$)

$$\max_{j\in H}\|\hat{\Sigma}-\Sigma\|_\infty\|\hat{v}_j-v_j\|_1^2=O_p\left(\frac{p^{4/r}}{n^{1/2}}\right)O_p\left(\frac{d_{n2}^2\bar{s}^2h^{4/r}p^{4/r}}{n}\right)=O_p\left(\frac{d_{n2}^2\bar{s}^2h^{4/r}p^{8/r}}{n^{3/2}}\right).$$

But since d_{n2}^2 is bounded by constants

$$O_p\left(\frac{d_{n2}^2\bar{s}^2h^{4/r}p^{8/r}}{n^{3/2}}\right)=O_p\left(d_{n2}^2\frac{\bar{s}p^{4/r}}{n^{1/2}}\frac{\bar{s}h^{4/r}p^{4/r}}{n}\right)=o_p\left(\frac{\bar{s}h^{4/r}p^{4/r}}{n}\right),$$

as $\frac{\bar{s}p^{4/r}}{n^{1/2}}=\left(\frac{p^2\bar{s}^{r/2}}{n^{r/4}}\right)^{2/r}\to 0$ by Assumption 2b) we conclude

$$\max_{j\in H}(\hat{v}_j-v_j)'\Sigma(\hat{v}_j-v_j)\leq O_p\left(\frac{d_{n1}\bar{s}h^{4/r}p^{4/r}}{n}\right).$$

Therefore, by

$$\max_{j\in H}\phi_{\min}(\Sigma)\|\hat{v}_j-v_j\|_2^2\leq\max_{j\in H}(\hat{v}_j-v_j)'\Sigma(\hat{v}_j-v_j)\leq O_p\left(d_{n1}\frac{\bar{s}h^{4/r}p^{4/r}}{n}\right),$$

one gets

$$\max_{j\in H}\|\hat{\gamma}_j-\gamma_j\|_2^2=\max_{j\in H}\|\hat{v}_j-v_j\|_2^2=O_p\left(d_{n1}\frac{\bar{s}h^{4/r}p^{4/r}}{n}\right). \tag{A.44}$$

since $\phi_{\min}(\Sigma)$ is bounded away from zero by Assumption 2a).

Next, we consider $|\hat{\tau}_j^2-\tau_j^2|$. First, by (A.101) and $X_j=X_{-j}\gamma_j+\eta_j$,

$$\begin{aligned}\hat{\tau}_j^2&=\frac{(X_j-X_{-j}\hat{\gamma}_j)'X_j}{n}\\&=\frac{[\eta_j-X_{-j}(\hat{\gamma}_j-\gamma_j)]'[X_{-j}\gamma_j+\eta_j]}{n}\\&=\frac{\eta_j'\eta_j}{n}+\frac{\eta_j'X_{-j}\gamma_j}{n}-\frac{(\hat{\gamma}_j-\gamma_j)'X_{-j}'X_{-j}\gamma_j}{n}-\frac{(\hat{\gamma}_j-\gamma_j)'X_{-j}'\eta_j}{n}.\end{aligned}$$

Using the above expression one gets

$$\max_{j\in H}|\hat{\tau}_j^2-\tau_j^2| \leq \max_{j\in H}\Big|\frac{\eta_j'\eta_j}{n}-\tau_j^2\Big| + \max_{j\in H}|\eta_j'X_{-j}(\hat{\gamma}_j-\gamma_j)/n| \\ + \max_{j\in H}|\eta_j'X_{-j}\gamma_j/n| + \max_{j\in H}\left|\frac{\gamma_j'X_{-j}'X_{-j}(\hat{\gamma}_j-\gamma_j)}{n}\right|. \tag{A.45}$$

Since $\frac{\eta_j'\eta_j}{n}-\tau_j^2=\frac{1}{n}\sum_{i=1}^n\left(\eta_{j,i}^2-E(\eta_{j,i}^2)\right)$ is a sum of mean zero terms with $r/2$ moments uniformly bounded by a constant C (the latter is seen by means of the Cauchy-Schwarz inequality and Assumption 2c) it follows from Lemma A.3

$$P\left(\max_{j\in H}\Big|\frac{\eta_j'\eta_j}{n}-\tau_j^2\Big|>Mh^{2/r}/n^{1/2}\right)=P\left(\max_{j\in H}\Big|\frac{1}{n}\sum_{i=1}^n\left(\eta_{j,i}^2-E(\eta_{j,i}^2)\right)\Big|>Mh^{2/r}/n^{1/2}\right)\leq\frac{b_rC}{M^{r/2}},$$

which implies that

$$\max_{j\in H}\Big|\frac{\eta_j'\eta_j}{n}-\tau_j^2\Big|=O_p\left(\frac{h^{2/r}}{n^{1/2}}\right). \tag{A.46}$$

Next, consider the second term in (A.45). By (A.32) and (A.39) it follows that

$$\begin{aligned}\max_{j\in H}|\eta_j'X_{-j}(\hat{\gamma}_j-\gamma_j)/n| &\leq \max_{j\in H}\|\eta_j'X_{-j}/n\|_\infty\max_{j\in H}\|\hat{\gamma}_j-\gamma_j\|_1\\ &= O_p\left(\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)O_p\left(\frac{d_{n2}\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right)\\ &= O_p\left(\left[\sqrt{d_{n2}}\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right]^2\right).\end{aligned} \tag{A.47}$$

Before we bound the third term in (A.45) we show that $\max_{j\in H}\|\gamma_j\|_1=O(\sqrt{\bar{s}})$. To this end, define the $(p-1)\times(p-1)$ matrix $\Sigma_{-j,-j}$ consisting of all rows and columns of Σ except the j'th row and column. Then, note that

$$\frac{\gamma_j'\Sigma_{-j,-j}\gamma_j}{\gamma_j'\gamma_j}\geq\phi_{\min}(\Sigma_{-j,-j})\geq\phi_{\min}(\Sigma),$$

such that

$$\gamma_j'\gamma_j\leq\frac{\gamma_j'\Sigma_{-j,-j}\gamma_j}{\phi_{\min}(\Sigma)}.$$

Since $X_{j,i}=X_{-j,i}\gamma_j+\eta_{j,i}$ it follows from the orthogonality in L^2 of each entry in $X_{-j,i}$ to $\eta_{j,i}$ that $E(X_{j,i}^2)=\gamma_j'\Sigma_{-j,-j}\gamma_j+E(\eta_{j,i}^2)$ such that $\gamma_j'\Sigma_{-j,-j}\gamma_j\leq E(X_{j,i}^2)\leq\max_{j\in H}E(X_{j,i}^2)$. Since $\left(E(X_{j,i}^2)\right)^{1/2}\leq\left(E(X_{j,i}^r)\right)^{1/r}\leq C^{1/r}$ for all $j\in H$ one has $\max_{j\in H}E(X_{j,i}^2)\leq C^{2/r}$. Hence,

$$\gamma_j'\gamma_j\leq\frac{C^{2/r}}{\phi_{\min}(\Sigma)}. \tag{A.48}$$

Thus, by Assumption 2a), $\gamma_j'\gamma_j$ is bounded by a constant not depending on j which implies that $\max_{j\in H}\|\gamma_j\|_1=O(\sqrt{\bar{s}})$. Hence, returning to the third term of (A.45),

$$\max_{j\in H}|\eta_j'X_{-j}\gamma_j/n|\leq\max_{j\in H}\|\eta_j'X_{-j}/n\|_\infty\max_{j\in H}\|\gamma_j\|_1=O_p\left(\sqrt{\bar{s}}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right), \tag{A.49}$$

where we have also used (A.39). It remains to bound the fourth summand in (A.45). By the Karush-Kuhn-Tucker conditions for the conservative lasso nodewise regression one has

$$\lambda_{node,n}\hat{\Gamma}_j\hat{\kappa}_j+\frac{X_{-j}'X_{-j}\hat{\gamma}_j}{n}-\frac{X_{-j}'X_j}{n}=0,$$

which, using $X_j = X_{-j}\gamma_j + \eta_j$, is equivalent to

$$\lambda_{node,n}\hat{\Gamma}_j\hat{\kappa}_j + \frac{X'_{-j}X_{-j}\hat{\gamma}_j}{n} - \frac{X'_{-j}\eta_j}{n} - \frac{X'_{-j}X_{-j}\gamma_j}{n} = 0.$$

The above equation can be rewritten as

$$\frac{X'_{-j}X_{-j}}{n}(\hat{\gamma}_j - \gamma_j) = \frac{X'_{-j}\eta_j}{n} - \lambda_{node,n}\hat{\Gamma}_j\hat{\kappa}_j.$$

This implies

$$\left\|\frac{X'_{-j}X_{-j}}{n}(\hat{\gamma}_j - \gamma_j)\right\|_\infty \leq \left\|\frac{X'_{-j}\eta_j}{n}\right\|_\infty + \|\lambda_{node,n}\hat{\Gamma}_j\hat{\kappa}_j\|_\infty.$$

The second term on the right hand side in the above display can be bounded as

$$\|\lambda_{node,n}\hat{\Gamma}_j\hat{\kappa}_j\|_\infty \leq \|\lambda_{node,n}\hat{\Gamma}_j\|_{\ell_\infty}\|\hat{\kappa}_j\|_\infty \leq \lambda_{node,n},$$

for all $j \in H$ since $\|\hat{\kappa}_j\|_\infty \leq 1$ and $\|\hat{\Gamma}_j\|_{\ell_\infty} \leq 1$. Hence, using (A.39),

$$\max_{j\in H}\left\|\frac{X'_{-j}X_{-j}}{n}(\hat{\gamma}_j - \gamma_j)\right\|_\infty = O_p(\lambda_{node,n}) + O_p(\lambda_{node,n}) = O_p\left(\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)$$

This means, using $\max_{j\in H}\|\gamma_j\|_1 = O(\bar{s}^{1/2})$,

$$\max_{j\in H}\left|\gamma'_j\frac{X'_{-j}X_{-j}}{n}(\hat{\gamma}_j - \gamma_j)\right| = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.50}$$

Since $h \leq p$, Assumption 2b) implies that

$$\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}} \leq \bar{s}^{1/2}\frac{p^{4/r}}{\sqrt{n}} = \frac{1}{\bar{s}^{1/2}}\left(\frac{\bar{s}^{r/2}p^2}{n^{r/4}}\right)^{2/r} \to 0,$$

such that the dominant term in (A.45) is $O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)$ given d_{n2}. Thus,

$$\max_{j\in H}|\hat{\tau}_j^2 - \tau_j^2| = O_p(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{n^{1/2}}).$$

Next, note that $\tau_j^2 = 1/\Theta_{j,j} \geq 1/\phi_{\max}(\Theta) = \phi_{\min}(\Sigma)$ for all $j = 1, ..., p$ with $\phi_{\min}(\Sigma)$ bounded away from zero by Assumption 2. Thus, $\min_{1\leq j\leq p}\tau_j^2$ is bounded away from zero, and so

$$\min_{1\leq j\leq p}\hat{\tau}_j^2 = \min_{1\leq j\leq p}[\hat{\tau}_j^2 - \tau_j^2 + \tau_j^2] \geq \min_{1\leq j\leq p}\tau_j^2 - \max_{1\leq j\leq p}|\hat{\tau}_j^2 - \tau_j^2|$$

is bounded away from zero with probability tending to one using $\max_{j\in H}|\hat{\tau}_j^2 - \tau_j^2| = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right) = o_p(1)$. This implies

$$\max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| = \max_{j\in H}\frac{|\tau_j^2 - \hat{\tau}_j^2|}{\hat{\tau}_j^2\tau_j^2} = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.51}$$

We are now ready to bound $\max_{j\in H}\|\hat{\Theta}_j - \Theta_j\|_1$. Recall that $\hat{\Theta}_j$ is formed by dividing $\hat{C}_j$ by $\hat{\tau}_j^2$. Let Θ_j denote the j'th row of Θ written as a column vector. Then, Θ_j is formed by dividing C_j (j'th row of C

written as a column vector) by τ_j^2. Therefore, using $\max_{j\in H}\|\gamma_j\|_1 = O(\bar{s}^{1/2})$, (A.32), and (A.51)

$$\begin{aligned}
\max_{j\in H}\left\|\hat{\Theta}_j - \Theta_j\right\|_1 &= \max_{j\in H}\left\|\frac{\hat{C}_j}{\hat{\tau}_j^2} - \frac{C_j}{\tau_j^2}\right\|_1 \qquad (A.52)\\
&\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| + \max_{j\in H}\left\|\frac{\hat{\gamma}_j}{\hat{\tau}_j^2} - \frac{\gamma_j}{\tau_j^2}\right\|_1\\
&= \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| + \max_{j\in H}\left\|\frac{\hat{\gamma}_j}{\hat{\tau}_j^2} - \frac{\gamma_j}{\hat{\tau}_j^2} + \frac{\gamma_j}{\hat{\tau}_j^2} - \frac{\gamma_j}{\tau_j^2}\right\|_1\\
&\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| + \max_{j\in H}\frac{\|\hat{\gamma}_j - \gamma_j\|_1}{\hat{\tau}_j^2} + \max_{j\in H}\|\gamma_j\|_1 \max_{j\in H}\left(\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right|\right)\\
&= O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right) + O_p\left(\frac{d_{n2}\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right) + O_p\left(\bar{s}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)\\
&= O_p\left(\frac{d_{n2}\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right). \qquad (A.53)
\end{aligned}$$

Next, for later purposes, we also bound $\|\hat{\Theta}_j - \Theta_j\|_2$. By (A.44), and $\max_{j\in H}\|\gamma_j\|_2^2 = O(1)$ by (A.48)

$$\begin{aligned}
\max_{j\in H}\|\hat{\Theta}_j - \Theta_j\|_2 &\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| + \max_{j\in H}\frac{\|\hat{\gamma}_j - \gamma_j\|_2}{\hat{\tau}_j^2} + \max_{j\in H}\|\gamma_j\|_2 \max_{j\in H}\left(\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right|\right)\\
&= O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right) + O_p\left(\frac{\sqrt{d_{n1}}\bar{s}^{1/2}h^{2/r}p^{2/r}}{n^{1/2}}\right) + O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right),\\
&= O_p\left(\sqrt{d_{n1}}\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \qquad (A.54)
\end{aligned}$$

Finally, we show that $\max_{j\in H}\|\hat{\Theta}_j\|_1 = O_p(\sqrt{\bar{s}})$. To this end,

$$\max_{j\in H}\|\Theta_j\|_1 \leq \max_{j\in H}\frac{1}{\tau_j^2} + \max_{j\in H}\|\gamma_j/\tau_j^2\|_1 = O(\bar{s}^{1/2}) \qquad (A.55)$$

(as τ_j^2 is uniformly bounded away from zero). Then, as $h \leq p$ implies $\frac{\bar{s}h^{2/r}p^{2/r}}{n^{1/2}} \leq [p^2\bar{s}^{r/2}/n^{r/4}]^{2/r} \to 0$ by Assumption 2b, we get

$$\max_{j\in H}\|\hat{\Theta}_j\|_1 \leq \max_{j\in H}\|\hat{\Theta}_j - \Theta_j\|_1 + \max_{j\in H}\|\Theta_j\|_1 = O_p\left(\frac{d_{n2}\bar{s}h^{2/r}p^{2/r}}{n^{1/2}}\right) + O(\sqrt{\bar{s}}) = O_p(\sqrt{\bar{s}}). \qquad (A.56)$$

□

Proof of Theorem 2. We show that the ratio

$$t = \frac{n^{1/2}\alpha'(\hat{b} - \beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}, \qquad (A.57)$$

is asymptotically standard normal. First, note that one can write. By (11)

$$t = t_1 + t_2,$$

where

$$t_1 = \frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \text{ and } t_2 = -\frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}.$$

It suffices to show that t_1 is asymptotically standard normal and $t_2 = o_p(1)$.
Step 1. We first show that t_1 is asymptotically standard normal.

a) To show that t_1 is asymptotically standard normal we first show that

$$t_1' = \frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}$$

converges in distribution to a standard normal where $\Sigma_{xu} = n^{-1}\sum_{i=1}^n E(X_iX_i'u_i^2)$. Then we show that t_1' and t_1 are asymptotically equivalent. Note that, using $E(u_i|X_i) = 0$ for all $i = 1, ..., n$, we obtain

$$E\left[\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right] = E\left[\frac{\alpha'\Theta\sum_{i=1}^n X_iu_i/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right] = 0, \tag{A.58}$$

and

$$E\left[\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right]^2 = E\left[\frac{\alpha'\Theta\sum_{i=1}^n X_iu_i/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right]^2 = 1.$$

Hence, in order to apply Lyapounov's condition in central limit theorem for independent random variables, it suffices to show that

$$\frac{1}{\left(\alpha'\Theta\Sigma_{xu}\Theta'\alpha\right)^{r/4}}\sum_{i=1}^n E\left|\alpha'\Theta X_iu_i/n^{1/2}\right|^{r/2} \to 0. \tag{A.59}$$

First, using the symmetry of Θ, we get (recall that Θ_j is the j'th row of Θ written as a column vector)

$$\left\|\alpha'\Theta\right\|_1 = \left\|\Theta\alpha\right\|_1 = \left\|\sum_{j\in H}\Theta_j\alpha_j\right\|_1 \leq \sum_{j\in H}|\alpha_j|\left\|\Theta_j\right\|_1 = O\left(\sqrt{h\bar{s}}\right),$$

since $\|\alpha\|_2 = 1$ and $\max_{j\in H}\|\Theta_j\|_1 = O(\sqrt{\bar{s}})$ by (A.55). Note also that

$$\alpha'\Theta = \left(\Theta\alpha\right)' = \left(\sum_{j\in H}\Theta_j\alpha_j\right)'$$

such that the non-zero entries of $\alpha'\Theta$ must be contained in $\bar{S} = \cup_{j\in H}S_j$ which has cardinality at most $|\bar{S}| = h\bar{s}\wedge p$, where $S_j = \{\Theta_{j,i} \neq 0\}$. Thus,

$$\begin{aligned}
E\left|\alpha'\Theta X_iu_i/n^{1/2}\right|^{r/2} &\leq E\left(\left\|\alpha'\Theta\right\|_1^{r/2}\max_{k\in\bar{S}}\left|X_{k,i}u_i/n^{1/2}\right|^{r/2}\right)\\
&\leq O\left(\left(\frac{h\bar{s}}{n}\right)^{r/4}\right)\left(h\bar{s}\wedge p\right)\max_{k\in\bar{S}}E|X_{k,i}u_i|^{r/2}\\
&\leq O\left(\left(\frac{h\bar{s}}{n}\right)^{r/4}\left(h\bar{s}\wedge p\right)\right)\\
&= O\left(\frac{(h\bar{s})^{r/4+1}\wedge(h\bar{s})^{r/4}p}{n^{r/4}}\right),
\end{aligned}$$

where the third inequality follows from the Cauchy-Schwarz inequality and using that $X_{k,i}$ and u_i have uniformly bounded $r'th$ moments. Hence,

$$\sum_{i=1}^n E\left|\alpha'\Theta X_iu_i/n^{1/2}\right|^{r/2} = O\left(\frac{(h\bar{s})^{r/4+1}\wedge(h\bar{s})^{r/4}p}{n^{r/4-1}}\right) = o(1),$$

by Assumption 3d). Next, we show that $\alpha'\Theta\Sigma_{xu}\Theta'\alpha$ is asymptotically bounded away from zero in (A.59). Clearly,

$$\alpha'\Theta\Sigma_{xu}\Theta'\alpha \geq \phi_{\min}(\Sigma_{xu})\left\|\Theta'\alpha\right\|_2^2 \geq \phi_{\min}(\Sigma_{xu})\phi_{\min}^2(\Theta)\left\|\alpha\right\|_2^2 = \phi_{\min}(\Sigma_{xu})\frac{1}{\phi_{\max}^2(\Sigma)}, \tag{A.60}$$

which is bounded away from zero since $\phi_{\min}(\Sigma_{xu})$ is bounded away from zero and $\phi_{\max}(\Sigma)$ is bounded from above. Hence, the Lyapounov condition is satisfied and t_1' converges in distribution to a standard normal.

b) We now show that $t_1' - t_1 = o_p(1)$. To do so it suffices that the numerators as well as the denominators of t_1' and t_1 are asymptotically equivalent since $\alpha'\Theta\Sigma_{xu}\Theta'\alpha$ is bounded away from 0 by (A.60). We first show that the denominators of t_1' and t_1 are asymptotically equivalent, i.e.

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| = o_p(1). \tag{A.61}$$

Set $\tilde{\Sigma}_{xu} = n^{-1}\sum_{i=1}^n X_i X_i' u_i^2$. To establish (A.61) it suffices to show the following relations:

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha| = o_p(1). \tag{A.62}$$

$$|\alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha| = o_p(1). \tag{A.63}$$

$$|\alpha'\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| = o_p(1). \tag{A.64}$$

We first prove (A.62).

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha| \leq \|\hat{\Sigma}_{xu} - \tilde{\Sigma}_{xu}\|_\infty \|\hat{\Theta}'\alpha\|_1^2. \tag{A.65}$$

But by (A.56) and $\|\alpha\|_2 = 1$

$$\left\|\hat{\Theta}'\alpha\right\|_1 = \left\|\sum_{j\in H}\hat{\Theta}_j\alpha_j\right\|_1 \leq \sum_{j\in H}|\alpha_j|\left\|\hat{\Theta}_j\right\|_1 = O_p\left(\sqrt{h\bar{s}}\right). \tag{A.66}$$

To proceed, we bound $\|\hat{\Sigma}_{xu} - \tilde{\Sigma}_{xu}\|_\infty$. Using $\hat{u}_i = u_i - X_i'(\hat{\beta} - \beta_0)$ in the definition of $\hat{\Sigma}_{xu}$ we get

$$\hat{\Sigma}_{xu} - \tilde{\Sigma}_{xu} = -\frac{2}{n}\sum_{i=1}^n X_i X_i' u_i X_i'(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^n X_i X_i'(\hat{\beta} - \beta_0)' X_i X_i'(\hat{\beta} - \beta_0). \tag{A.67}$$

We bound each sum separately. First, by the Cauchy-Schwarz inequality,

$$\max_{1\leq k,l\leq p}\left|\frac{2}{n}\sum_{i=1}^n X_{k,i}X_{l,i}u_i X_i'(\hat{\beta} - \beta_0)\right| \leq 2\sqrt{\max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^n X_{k,i}^2 X_{l,i}^2 u_i^2}\cdot\left\|X(\hat{\beta} - \beta_0)\right\|_n. \tag{A.68}$$

Now for any three random variables Z_1, Z_2 and Z_3 with finite r'th moment it follows from two applications of Hölder's inequality

$$\begin{aligned} E|Z_1^2 Z_2^2 Z_3^2|^{r/6} = E|Z_1^{r/3} Z_2^{r/3} Z_3^{r/3}| &\leq E\left(|Z_1|^{r/2}|Z_2|^{r/2}\right)^{2/3} E\left(|Z_3^r|\right)^{1/3} \\ &\leq E\left(|Z_1^r|\right)^{1/3} E\left(|Z_2^r|\right)^{1/3} E\left(|Z_3^r|\right)^{1/3}. \end{aligned} \tag{A.69}$$

Thus, by Assumption 1, all summands in (A.68) have uniformly bounded $r/6$ moments and therefore Lemma A.3 implies that

$$P\left(\max_{1\leq k,l\leq p}\left|\frac{1}{n}\sum_{i=1}^n\left(X_{k,i}^2X_{l,i}^2u_i^2 - E(X_{k,i}^2X_{l,i}^2u_i^2)\right)\right| > t\right) \leq b_{r/6}\frac{Cp^2n^{r/12}}{(tn)^{r/6}}.$$

Hence, choosing $t = M\frac{p^{12/r}}{n^{1/2}}$ for $M > 0$ sufficiently large shows that

$$\max_{1\leq k,l\leq p}\left|\frac{1}{n}\sum_{i=1}^n\left(X_{k,i}^2X_{l,i}^2u_i^2 - E(X_{k,i}^2X_{l,i}^2u_i^2)\right)\right| = O_p\left(\frac{p^{12/r}}{n^{1/2}}\right).$$

Furthermore, since the L^r-norm is non-decreasing in r and since $r \geq 6$ we have, using (A.69) above,

$$\begin{aligned} \max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^n E\left(X_{k,i}^2X_{l,i}^2u_i^2\right) &\leq \max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^n\left(E\left(X_{k,i}^2X_{l,i}^2u_i^2\right)^{r/6}\right)^{6/r} \\ &\leq \max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^n\left[\left(E|X_{k,i}|^r\right)^{1/3}\left(E|X_{l,i}|^r\right)^{1/3}\left(E|u_i|^r\right)^{1/3}\right]^{6/r}, \end{aligned}$$

which is uniformly bounded by Assumption 1 since the r'th moments of $X_{k,i}$ and u_i are uniformly bounded. Therefore, $\sqrt{\max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^n X_{k,i}^2X_{l,i}^2u_i^2} = O(1) + O_p\left(\frac{p^{6/r}}{n^{1/4}}\right)$ in (A.68). By Theorem 1 it follows from choosing M sufficiently large

$$\left\|X(\hat{\beta}-\beta_0)\right\|_n = O_p\left(\frac{\sqrt{d_{n1}}p^{2/r}\sqrt{s_0}}{n^{1/2}}\right). \tag{A.70}$$

Thus,

$$\max_{1\leq k,l\leq p}\left|\frac{2}{n}\sum_{i=1}^n X_{k,i}X_{l,i}u_iX_i'(\hat{\beta}-\beta_0)\right| = O_p\left(\frac{p^{8/r}\sqrt{s_0}}{n^{3/4}}\right) + O_p\left(\frac{\sqrt{d_{n1}}p^{2/r}\sqrt{s_0}}{n^{1/2}}\right). \tag{A.71}$$

Regarding the second term in (A.67) note that

$$\max_{1\leq k,l\leq p}\left|\frac{1}{n}\sum_{i=1}^n X_{k,i}X_{l,i}(\hat{\beta}-\beta_0)'X_iX_i'(\hat{\beta}-\beta_0)\right| \leq \max_{1\leq k,l\leq p}\max_{1\leq i\leq n}\left|X_{k,i}X_{l,i}\right|\frac{1}{n}\sum_{i=1}^n\left(X_i'(\hat{\beta}-\beta_0)\right)^2. \tag{A.72}$$

By the Cauchy-Schwarz inequality, $X_{k,i}X_{l,i}$ has uniformly bounded $r/2$ moments. Hence, by the union bound and Markov's inequality, for any $t>0$ we get via Lemma A.3

$$P\left(\max_{1\leq i\leq n}\max_{1\leq k,l\leq p}\left|X_{k,i}X_{l,i}\right| > t\right) \leq np^2\frac{C}{t^{r/2}}.$$

Therefore, choosing $t = Mp^{4/r}n^{2/r}$ for $M>0$ sufficiently large reveals that

$$\max_{1\leq i\leq n}\max_{1\leq k,l\leq p}\left|X_{k,i}X_{l,i}\right| = O_p\left(p^{4/r}n^{2/r}\right).$$

Next, note that by Theorem 1

$$\frac{1}{n}\sum_{i=1}^n\left(X_i'(\hat{\beta}-\beta_0)\right)^2 = \left\|X(\hat{\beta}-\beta_0)\right\|_n^2 = O_p\left(d_{n1}\frac{p^{4/r}s_0}{n}\right), \tag{A.73}$$

such that, using (A.72),

$$\begin{aligned}\max_{1\leq k,l\leq p}\left|\frac{1}{n}\sum_{i=1}^n X_{k,i}X_{l,i}(\hat{\beta}-\beta_0)'X_iX_i'(\hat{\beta}-\beta_0)\right| &= O_p\left(p^{4/r}n^{2/r}\right)O_p\left(\frac{d_{n1}p^{4/r}s_0}{n}\right)\\ &= O_p\left(\frac{d_{n1}p^{8/r}s_0}{n^{(r-2)/r}}\right).\end{aligned} \tag{A.74}$$

Then, combining (A.71) and (A.74) implies that

$$\left\|\hat{\Sigma}_{xu}-\tilde{\Sigma}_{xu}\right\|_\infty = O_p\left(\frac{p^{8/r}\sqrt{s_0}}{n^{3/4}}\right) + O_p\left(\frac{\sqrt{d_{n1}}p^{2/r}\sqrt{s_0}}{n^{1/2}}\right) + O_p\left(\frac{d_{n1}p^{8/r}s_0}{n^{(r-2)/r}}\right).$$

Therefore, combining with (A.66) yields

$$\left|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha\right| = O_p\left(\frac{p^{8/r}\sqrt{s_0}h\bar{s}}{n^{3/4}}\right) + O_p\left(\frac{\sqrt{d_{n1}}p^{2/r}\sqrt{s_0}h\bar{s}}{n^{1/2}}\right) + O_p\left(\frac{d_{n1}p^{8/r}s_0h\bar{s}}{n^{(r-2)/r}}\right) = o_p(1), \tag{A.75}$$

by Assumption 3c) and since d_{n1} is bounded by constants. This establishes (A.62).

Next, we turn to (A.63). First, note that

$$|\alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha| \leq \|\tilde{\Sigma}_{xu}-\Sigma_{xu}\|_\infty\|\hat{\Theta}'\alpha\|_1^2. \tag{A.76}$$

Furthermore, similarly to (A.69), three applications of Hölder's inequality reveal that $X_{k,i}X_{l,i}u_i^2$ have uniformly bounded $r/4$ moments. Hence, by Lemma A.3, for any $t > 0$

$$P\left(\|\tilde{\Sigma}_{xu} - \Sigma_{xu}\|_\infty > t\right) = P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{k,i}X_{l,i}u_i^2 - E\left(X_{k,i}X_{l,i}u_i^2\right)\right| > t\right) \leq b_{r/4}\frac{p^2 C n^{r/8}}{(tn)^{r/4}}.$$

Thus, choosing $t = M\frac{p^{8/r}}{n^{1/2}}$ for $M > 0$ sufficiently large shows that

$$\|\tilde{\Sigma}_{xu} - \Sigma_{xu}\|_\infty = O_p\left(\frac{p^{8/r}}{n^{1/2}}\right).$$

By (A.76) and (A.66)

$$|\alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha| = O_p\left(\frac{p^{8/r}h\bar{s}}{n^{1/2}}\right) = o_p(1),$$

and Assumption 3b).

Finally, we establish (A.64) to conclude (A.61). By Lemma 6.1 in van de Geer et al. (2014)

$$\begin{aligned}|\alpha'\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| &\leq \|\Sigma_{xu}\|_\infty\|\hat{\Theta}'\alpha - \Theta'\alpha\|_1^2 + 2\|\Sigma_{xu}\Theta'\alpha\|_2\|\hat{\Theta}'\alpha - \Theta'\alpha\|_2 \\ &\leq \|\Sigma_{xu}\|_\infty\|(\hat{\Theta}' - \Theta')\alpha\|_1^2 + 2\phi_{\max}(\Sigma_{xu})\left\|\Theta'\alpha\right\|_2\|(\hat{\Theta}' - \Theta')\alpha\|_2.\end{aligned}$$

Note that

$$\begin{aligned}\|(\hat{\Theta}' - \Theta')\alpha\|_1 &= \left\|\sum_{j\in H}\left(\hat{\Theta}_j - \Theta_j\right)\alpha_j\right\|_1 \leq \sum_{j\in H}\left\|\hat{\Theta}_j - \Theta_j\right\|_1|\alpha_j| \leq \max_{j\in H}\left\|\hat{\Theta}_j - \Theta_j\right\|_1\sum_{j\in H}|\alpha_j| \\ &= O_p\left(d_{n2}\bar{s}\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right), \qquad (A.77)\end{aligned}$$

by (A.34) and $\|\alpha\|_2 = 1$. Furthermore, using the symmetry of Θ,

$$\left\|\Theta'\alpha\right\|_2 \leq \phi_{\max}(\Theta)\|\alpha\|_2 = \frac{1}{\phi_{\min}(\Sigma)},$$

which is bounded by Assumption 2a). Finally,

$$\begin{aligned}\|(\hat{\Theta}' - \Theta')\alpha\|_2 &= \left\|\sum_{j\in H}\left(\hat{\Theta}_j - \Theta_j\right)\alpha_j\right\|_2 \leq \sum_{j\in H}\left\|\hat{\Theta}_j - \Theta_j\right\|_2|\alpha_j| \leq \max_{j\in H}\left\|\hat{\Theta}_j - \Theta_j\right\|_2\sum_{j\in H}|\alpha_j| \\ &= O_p\left(\sqrt{d_{n1}}\sqrt{\bar{s}}\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right),\end{aligned}$$

by (A.35) and $\|\alpha\|_2 = 1$. Therefore, by $\|\Sigma_{xu}\|_\infty \leq \phi_{\max}(\Sigma_{xu})$ with the latter assumed bounded from Assumption 3e),

$$|\alpha'\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| = O_p\left(d_{n2}^2\bar{s}^2\frac{h^{4/r+1}p^{4/r}}{n}\right) + O_p\left(\sqrt{d_{n1}}\sqrt{\bar{s}}\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right) = o_p(1), \qquad (A.78)$$

where we used

$$\frac{\bar{s}^2h^{(4/r)+1}p^{4/r}}{n} \leq \frac{\bar{s}(h\bar{s})p^{8/r}}{n} = \frac{\bar{s}}{n^{1/2}}\cdot\frac{(h\bar{s})p^{8/r}}{n^{1/2}} \to 0,$$

and Assumption 3b (which also implies $\bar{s} = o(n^{1/2})$), and d_{n1}, d_{n2} being bounded by constants. The uniformity of (A.61) over $\mathcal{B}_{\ell_0}(s_0)$ follows from simply observing that (A.70) and (A.73) above are actually valid uniformly over this set and that this is the only place in which β_0 enters in the above arguments.

We now turn to showing that the numerators of t_1' and t_1 are asymptotically equivalent, i.e.

$$|\alpha'\hat{\Theta}X'u/n^{1/2} - \alpha'\Theta X'u/n^{1/2}| = o_p(1).$$

By Lemma A.4 and (A.77) above we get, using $h \leq p$, and Assumption 3b, d_{n2} being bounded by constants

$$\begin{aligned} n^{1/2}|\alpha'\hat{\Theta}X'u/n - \alpha'\Theta X'u/n| &\leq n^{1/2}\left\|\frac{X'u}{n}\right\|_\infty \|\alpha'(\hat{\Theta}-\Theta)\|_1 \\ &= n^{1/2}O_p\left(\frac{p^{2/r}}{\sqrt{n}}\right) O\left(d_{n2}\bar{s}\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right) \\ &= O_p\left(d_{n2}\bar{s}\frac{h^{2/r+1/2}p^{4/r}}{\sqrt{n}}\right) \\ &= O_p\left(d_{n2}\bar{s}\frac{h^{1/2}p^{6/r}}{\sqrt{n}}\right) \\ &= o_p(1). \end{aligned} \tag{A.79}$$

Step 2. It remains to be shown that $t_2 = o_p(1)$. The denominators of t_1 and t_2 are identical. Hence, the denominator of t_2 is asymptotically bounded away from zero with probability approaching one by (A.60) and (A.61). Thus, it suffices to show that the numerator of t_2 vanishes in probability. Note that, by the definition of Δ, and $\|\alpha\|_2 = 1$,

$$|\alpha'\Delta| \leq \max_{j\in H}|\Delta_j| \sum_{j\in H}|\alpha_j| = \max_{j\in H}\left|\left(\hat{\Theta}_j'\hat{\Sigma} - e_j\right)\left(\sqrt{n}(\hat{\beta}-\beta_0)\right)\right| \sum_{j\in H}|\alpha_j| \tag{A.80}$$

$$\leq \max_{j\in H}\left\|\left(\hat{\Theta}_j'\hat{\Sigma} - e_j\right)\right\|_\infty \left\|\sqrt{n}(\hat{\beta}-\beta_0)\right\|_1 O\left(\sqrt{h}\right). \tag{A.81}$$

First, it follows from Theorem 1 that $n^{1/2}\|\hat{\beta}-\beta_0\|_1 = O_p\left(d_{n2}s_0p^{2/r}\right)$. Next, we consider

$$\max_{j\in H}\left\|\left(\hat{\Theta}_j'\hat{\Sigma} - e_j\right)\right\|_\infty \leq \max_{j\in H}\frac{\lambda_{node,n}}{\hat{\tau}_j^2} = O_p\left(\frac{h^{2/r}p^{2/r}}{n^{1/2}}\right),$$

where we have used the definition of $\lambda_{node,n}$ and $\max_{j\in H} 1/\hat{\tau}_j^2 = O_p(1)$ by (A.51) and Assumption 3b). Thus, in total we have

$$\left|\alpha'\Delta\right| = O_p\left(\frac{h^{2/r}p^{2/r}}{n^{1/2}}\right) O_p\left(d_{n2}s_0p^{2/r}\right) O\left(\sqrt{h}\right) = O_p\left(d_{n2}s_0\frac{h^{2/r+1/2}p^{4/r}}{n^{1/2}}\right) = o_p(1), \tag{A.82}$$

by Assumption 3a), and d_{n2} being bounded by constants. The fact that $\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|\alpha'\Delta\right| = o_p(1)$ follows from the observation that Theorem 1 actually yields that $\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)} n^{1/2}\|\hat{\beta}-\beta_0\|_1 = O_p\left(d_{n2}s_0p^{2/r}\right)$ in the above argument and that this is the only place in which β_0 enters these arguments. Thus, for later reference,

$$\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|\alpha'\Delta\right| = o_p(1). \tag{A.83}$$

□

Proof of Theorem 3. For $\epsilon > 0$ define

$$A_{1,n} := \left\{\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|\alpha'\Delta\right| < \epsilon\right\},\ A_{2,n} := \left\{\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|\frac{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} - 1\right| < \epsilon\right\},$$

and

$$A_{3,n} := \left\{\left|\alpha'\hat{\Theta}X'u/n^{1/2} - \alpha'\Theta X'u/n^{1/2}\right| < \epsilon\right\}.$$

By, (A.83), (20), (A.79), and $\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}$ being bounded away from zero (by (A.60)) the probabilities of these three sets all tend to one. Thus, for every $t \in \mathbb{R}$,

$$\begin{aligned}
&\left| P\left(\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t \right) - \Phi(t) \right| \\
&= \left| P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t \right) - \Phi(t) \right| \\
&\leq \left| P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t, A_{1,n}, A_{2,n}, A_{3,n} \right) - \Phi(t) \right| + P\left(\cup_{i=1}^{3} A_{i,n}^c \right).
\end{aligned}$$

Using that $\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}$ does not depend on β_0 and is bounded away from zero by (A.60) there exists a positive constant D such that

$$\begin{aligned}
&P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t, A_{1,n}, A_{2,n}, A_{3,n} \right) \\
&= P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} \leq t\frac{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}, A_{1,n}, A_{2,n}, A_{3,n} \right) \\
&\leq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} \leq t(1+\epsilon) + \frac{\epsilon+\epsilon}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} \right) \\
&\leq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} \leq t(1+\epsilon) + 2D\epsilon \right).
\end{aligned}$$

Thus, as the right hand side in the above display does not depend on β_0

$$\begin{aligned}
&\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t, A_{1,n}, A_{2,n}, A_{3,n} \right) \\
&\leq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}} \leq t(1+\epsilon) + 2D\epsilon \right).
\end{aligned}$$

In step 1a) of the proof of Theorem 2 we established the asymptotic normality of $\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}$. Therefore, for n sufficiently large,

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t,\ A_{1,n}, A_{2,n}, A_{3,n} \right) \leq \Phi\left(t(1+\epsilon) + 2D\epsilon \right) + \epsilon.$$

As the above arguments are valid for all $\epsilon > 0$ we can use the continuity of $q \mapsto \Phi(q)$ to conclude that for any $\delta > 0$ we can choose ϵ sufficiently small to conclude that

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t,\ A_{1,n}, A_{2,n}, A_{3,n} \right) \leq \Phi(t) + \delta + \epsilon. \tag{A.84}$$

Next, using that $\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}$ does not depend on β_0 and is bounded away from zero by (A.60) there exists

a positive constant D such that

$$
\begin{aligned}
&P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}-\frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}\leq t,\ A_{1,n},A_{2,n},A_{3,n}\right)\\
&=P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}-\frac{\alpha'\Delta}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\leq t\frac{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}},\ A_{1,n},A_{2,n},A_{3,n}\right)\\
&\geq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\leq t(1-\epsilon)-\frac{\epsilon+\epsilon}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}},\ A_{1,n},A_{2,n},A_{3,n}\right)\\
&\geq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\leq t(1-\epsilon)-2D\epsilon,\ A_{1,n},A_{2,n},A_{3,n}\right)\\
&\geq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\leq t(1-\epsilon)-2D\epsilon\right)+P\left(\cap_{i=1}^{3}A_{i,n}\right)-1.
\end{aligned}
$$

Thus, as the right hand side in the above display does not depend on β_0 and since $P\left(\cap_{i=1}^{3}A_{i,n}\right)$ can be made arbitrarily close to one by choosing n sufficiently we conclude

$$
\begin{aligned}
&\inf_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}-\frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}\leq t,\ A_{1,n},A_{2,n},A_{3,n}\right)\\
&\geq P\left(\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\leq t(1-\epsilon)-2D\epsilon\right)-\epsilon,
\end{aligned}
$$

for n sufficiently large. In step 1a) of the proof of Theorem 2 we established the asymptotic normality of $\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}$. Thus, for n sufficiently large,

$$
\inf_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}-\frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}\leq t,\ A_{1,n},A_{2,n},A_{3,n}\right)\geq\Phi\left(t(1-\epsilon)-2D\epsilon\right)-2\epsilon.
$$

As the above arguments are valid for all $\epsilon>0$ we can use the continuity of $q\mapsto\Phi(q)$ to conclude that for any $\delta>0$ we can choose ϵ sufficiently small to conclude that

$$
\inf_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}-\frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}\leq t,\ A_{1,n},A_{2,n},A_{3,n}\right)\geq\Phi(t)-2\epsilon-\delta. \tag{A.85}
$$

By (A.84) and (A.85) and $\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}P\left(\cup_{i=1}^{3}A_{i,n}^c\right)=P\left(\cup_{i=1}^{3}A_{i,n}^c\right)\to0$ (here we used that none of the sets A_1,A_2, or A_3 depend on β_0) we conclude that

$$
\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|P\left(\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}\leq t\right)-\Phi(t)\right|\to0.
$$

To see (25) note that

$$
\begin{aligned}
&P\left(\beta_{0,j}\notin\left[\hat{b}_j-z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}},\hat{b}_j+z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right]\right)\\
&=P\left(\left|\frac{\sqrt{n}\left(\hat{b}_j-\beta_{0,j}\right)}{\hat{\sigma}_j}\right|>z_{1-\alpha/2}\right)\\
&=P\left(\frac{\sqrt{n}\left(\hat{b}_j-\beta_{0,j}\right)}{\hat{\sigma}_j}>z_{1-\alpha/2}\right)+P\left(\frac{\sqrt{n}\left(\hat{b}_j-\beta_{0,j}\right)}{\hat{\sigma}_j}<-z_{1-\alpha/2}\right)\\
&\leq1-P\left(\frac{\sqrt{n}\left(\hat{b}_j-\beta_{0,j}\right)}{\hat{\sigma}_j}\leq z_{1-\alpha/2}\right)+P\left(\frac{\sqrt{n}\left(\hat{b}_j-\beta_{0,j}\right)}{\hat{\sigma}_j}\leq-z_{1-\alpha/2}\right).
\end{aligned}
$$

Thus, taking the supremum over $\beta_0 \in \mathcal{B}_{\ell_0}(s_0)$ and letting n tend to infinity yields an inequality in (25) via (24). The reverse inequality follows upon noting that
$P\left(\beta_{0,j} \notin [\hat{b}_j - z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}]\right) \geq 1 - P\left(\frac{\sqrt{n}(\hat{b}_j - \beta_{0,j})}{\hat{\sigma}_j} \leq z_{1-\alpha/2}\right) + P\left(\frac{\sqrt{n}(\hat{b}_j - \beta_{0,j})}{\hat{\sigma}_j} \leq -z_{1-\alpha/2-\delta_1}\right)$
for any $\delta_1 > 0$.

Finally, we turn to (26). By (20) we know $\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \left|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha\right| = o_p(1)$. Hence, choosing $\alpha = e_j$ and $\phi_{\max}(\Theta) = 1/\phi_{\min}(\Sigma)$,

$$\begin{aligned}
&\sqrt{n} \sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \operatorname{diam}\left(\left[\hat{b}_j - z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\alpha/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right]\right) = \sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} 2\hat{\sigma}_j z_{1-\alpha/2} \\
&= 2\left(\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \sqrt{e_j'\Theta\Sigma_{xu}\Theta' e_j} + o_p(1)\right) z_{1-\alpha/2} \\
&\leq 2\left(\sqrt{\phi_{\max}(\Sigma_{xu})}\frac{1}{\phi_{\min}(\Sigma)} + o_p(1)\right) z_{1-\alpha/2} \\
&= O_p(1),
\end{aligned}$$

as $\phi_{\max}(\Sigma_{xu})$ is bounded from above and $\phi_{\min}(\Sigma)$ is bounded from below by Assumptions 2a) and 3e). □

Strong oracle optimality of the variant of the Conservative Lasso

We provide a strong oracle optimality result for $\tilde{\beta}$; the variant of the conservative Lasso estimator. Recall that

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}}\{\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|\},$$

with $\tilde{w}_j = 1_{\{|\hat{\beta}_{L,j}| \leq \lambda_{prec}\}}$. Define the oracle estimator as

$$\hat{\beta}^{oracle} = (\hat{\beta}_{S_0}^{oracle}, 0) = \underset{\beta, \beta_{S_0^c}=0}{\operatorname{argmin}}[\|Y - X\beta\|_n^2]. \tag{A.86}$$

which we assume to be unique as in (Fan et al. (2014)). Strong oracle optimality of $\tilde{\beta}$ means it is equal to the oracle estimator with probability approaching one (Fan et al. (2014)).

Introduce the events

$$\mathcal{C}_1 = \{\|\hat{\beta}_L - \beta_0\|_\infty \leq \lambda_{prec}\} \tag{A.87}$$

and

$$\mathcal{C}_2 = \{\|(\bigtriangledown_{S_0^c}\|Y - X\hat{\beta}^{oracle}\|_n^2)\|_\infty < 2\lambda_n\}. \tag{A.88}$$

where $\bigtriangledown_{S_0^c}$ denotes the gradient with respect to the entries of β that are indexed by S_0^c. Next, we introduce the $n \times (p - s_0)$ matrix

$$\tilde{X} = M_{S_0} X_{S_o^c},$$

with $M_{S_0} = I_n - X_{S_0}(X_{S_0}'X_{S_0})^{-1}X_{S_0}'$, and $X_{S_0^c}, X_{S_0}$ are $(n \times (p - s_0), n \times s_0$ matrices).

Theorem 4. *Impose Assumptions 1-2 and*

(i). With probability approaching one

$$\min_{j \in S_0^c} \tilde{w}_j = 1,$$

and with added $\min_{j \in S_0} |\beta_{0,j}| > 2\lambda_{prec}$,

$$\max_{j \in S_0} \tilde{w}_j = 0.$$

(ii). If, furthermore, $E|\tilde{X}_{j,i}|^r < C$ *for a universal constant* C *then for all* $\epsilon > 0$ *there exists an* n *sufficiently large such that*

$$P(\tilde{\beta} = \hat{\beta}^{oracle}) \geq 1 - \epsilon.$$

Remarks.

1.The first part of Theorem 4 is similar to Lemma 1 (ii)-(iii). However, the important difference is that the new variant of the conservative Lasso ensures that the weights pertaining to the non-zero coefficients will be exactly *equal* to zero with probability approaching one. Lemma 1 only guarantees that these weights *converge* to zero for the conservative Lasso. The same caveat before Lemma 1 applies regarding the restrictiveness of the result since we use $\beta-\min$ condition.

2. Note that $\lambda_{prec}\to 0$ under Assumptions 1-2 also for the variant of the conservative Lasso.

3. Part (ii) of Theorem 4 is the strong oracle optimality of $\tilde{\beta}$.

Proof. Throughout we assume that $\Xi=\mathcal{C}_1\cap\mathcal{C}_2$ occurs and show at the end of the proof that this is indeed the case with probability approaching one. First, on $\mathcal{C}_1$

$$\max_{j\in S_0^c}|\hat{\beta}_{L,j}|=\max_{j\in S_0^c}|\hat{\beta}_{L,j}-\beta_{0,j}|\leq\lambda_{prec}.$$

This shows that

$$\min_{j\in S_0^c}\tilde{w}_j=1_{\{\max_{j\in S_0^c}|\hat{\beta}_{L,j}|\leq\lambda_{prec}\}}=1, \tag{A.89}$$

Next we consider $j\in S_0$.

$$\min_{j\in S_0}|\hat{\beta}_{L,j}|\geq\min_{j\in S_0}|\beta_{0,j}|-\max_{j\in S_0}|\hat{\beta}_{L,j}-\beta_{0,j}|>2\lambda_{prec}-\lambda_{prec}=\lambda_{prec}.$$

Thus,

$$\max_{j\in S_0}\tilde{w}_j=1_{\{\min_{j\in S_0}|\hat{\beta}_{L,j}|\leq\lambda_{prec}\}}=0. \tag{A.90}$$

Now we show that $\tilde{\beta}=\hat{\beta}^{oracle}$ on Ξ. Note that

$$\tilde{\beta}=\underset{\beta}{\text{argmin}}\{\|Y-X\beta\|_n^2+2\lambda_n\sum_{j=1}^{p}\tilde{w}_j|\beta_j|\}=\underset{\beta}{\text{argmin}}\{\|Y-X\beta\|_n^2+2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j|\}, \tag{A.91}$$

since $\tilde{w}_j=0$ for $j\in S_0$ on $\mathcal{C}_1$. By convexity of $\|Y-X\beta\|_n^2$ in β

$$\begin{aligned}\|Y-X\beta\|_n^2 &\geq \|Y-X\hat{\beta}^{oracle}\|_n^2+\sum_{j=1}^{p}\bigtriangledown_j\|Y-X\hat{\beta}^{oracle}\|_n^2(\beta_j-\hat{\beta}_j^{oracle})\\ &= \|Y-X\hat{\beta}^{oracle}\|_n^2+\sum_{j\in S_0^c}\bigtriangledown_j\|Y-X\hat{\beta}^{oracle}\|_n^2(\beta_j-\hat{\beta}_j^{oracle}),\end{aligned} \tag{A.92}$$

where $\sum_{j\in S_0}(\bigtriangledown_j\|Y-X\hat{\beta}^{oracle}\|_n^2)=0$ by the first order conditions for a minimum. Add $2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j|$ to both sides of (A.92) and note that $\hat{\beta}_j^{oracle}=0$ for $j\in S_0^c$ from oracle estimator definition,

$$\|Y-X\beta\|_n^2+2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j|\geq\|Y-X\hat{\beta}^{oracle}\|_n^2+2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j|+\sum_{j\in S_0^c}\bigtriangledown_j\|Y-X\hat{\beta}^{oracle}\|_n^2\beta_j. \tag{A.93}$$

Now subtract $\|Y-X\hat{\beta}^{oracle}\|_n^2$ from both sides of (A.93) and add $2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j\hat{\beta}_j^{oracle}=0$ (which is zero since $\hat{\beta}_j^{oracle}=0$, for $j\in S_0^c$ by the definition of the oracle estimator) to the left side of (A.93) to get

$$\begin{aligned}\|Y-X\beta\|_n^2+2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j| &- \{\|Y-X\hat{\beta}^{oracle}\|_n^2+2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\hat{\beta}_j^{oracle}|\}\\ &\geq [2\lambda_n\sum_{j\in S_0^c}\tilde{w}_j|\beta_j|+\sum_{j\in S_0^c}\bigtriangledown_j\|Y-X\hat{\beta}^{oracle}\|_n^2\beta_j].\end{aligned} \tag{A.94}$$

Note that $\tilde{w}_j = 0$ for all $j \in S_0$ by (A.90). Using this fact, add $2\lambda_n \sum_{j\in S_0} \tilde{w}_j|\beta_j| = 0$ and subtract $2\lambda_n \sum_{j\in S_0} \tilde{w}_j|\hat{\beta}_j^{oracle}| = 0$ from the left side of (A.94).

$$\begin{aligned}
\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j|\beta_j| &- \{\|Y - X\hat{\beta}^{oracle}\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j|\hat{\beta}_j^{oracle}|\} \\
&\geq [2\lambda_n \sum_{j\in S_0^c} \tilde{w}_j|\beta_j| + \sum_{j\in S_0^c} \bigtriangledown_j \|Y - X\hat{\beta}^{oracle}\|_n^2 \beta_j] \\
&= \sum_{j\in S_0^c} [2\lambda_n + \bigtriangledown_j \|Y - X\hat{\beta}^{oracle}\|_n^2 sgn(\beta_j)]|\beta_j|, \qquad \text{(A.95)}
\end{aligned}$$

where we use (A.89) in the last equality and $sgn(\beta_j)|\beta_j| = \beta_j$. Next, if $sgn(\beta_j) = 1$, then

$$\sum_{j\in S_0^c} [2\lambda_n + \bigtriangledown_j \|Y - X\hat{\beta}^{oracle}\|_n^2]|\beta_j| > 0$$

while if $sgn(\beta_j) = -1$, since $\mathcal{C}_2$ is assumed to occur,

$$\sum_{j\in S_0^c} [2\lambda_n - \bigtriangledown_j \|Y - X\hat{\beta}^{oracle}\|_n^2]|\beta_j| > 0.$$

By these inequalities and (A.95) we conclude

$$\|Y - X\beta\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j|\beta_j| - \{\|Y - X\hat{\beta}^{oracle}\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \tilde{w}_j|\hat{\beta}_j^{oracle}|\} \geq 0. \qquad \text{(A.96)}$$

Strict inequality in (A.96) is true, unless $\beta_j = 0$, for all $j \in S_0^c$. We now turn to verifying that the probability of Ξ tends to one. By the above display $\tilde{\beta} = \hat{\beta}^{oracle}$ on $\Xi = \mathcal{C}_1 \cap \mathcal{C}_2$ since $\beta \mapsto \|Y - X\beta\|_n^2$ is assumed to be uniquely minimized at $\hat{\beta}^{oracle}$.

Lemma A.7 proves $P(\mathcal{C}_1^c) \to 0$ under Assumptions 1-2, which also establishes part (i) of the theorem since the desired properties of the weights have been established on $\mathcal{C}_1$

To establish (ii) of the theorem it remains to show that $P(\mathcal{C}_2^c) \geq 1 - \epsilon$ for any $\epsilon > 0$. As in the proof of Theorem 3 in Fan et al. (2014) by definition of the oracle estimator in (A.86) via simple matrix algebra

$$\sum_{j\in S_0^c} (\bigtriangledown_j \|Y - X\hat{\beta}^{oracle}\|_n^2) = \frac{2}{n} X'_{S_0^c} M_{S_0} u = \frac{2}{n} \tilde{X}' u.$$

Next, $E|\tilde{X}_{ij}u_i|^{r/2} \leq \sqrt{E|\tilde{X}_{i,j}|^r E|u_i|^r} \leq C$ such that Lemma A.3 yields

$$\begin{aligned}
P[\|\tilde{X}'u\|_\infty \geq (n\lambda_n)] &\leq \frac{b_{r/2}(p - s_0)n^{r/4} \max_{j\in S_0^c} \max_{1\leq i\leq n} E|\tilde{X}_{i,j}u_i|^{r/2}}{(n\lambda_n)^{r/2}} \\
&\leq \frac{b_{r/2} p n^{r/4} C}{(n\lambda_n)^{r/2}} = \frac{C}{M^{r/2}}, \qquad \text{(A.97)}
\end{aligned}$$

where we used $\lambda_n = Mp^{2/r}/n^{1/2}$, and combined the constants $b_{r/2}$ and C into C. Choosing M sufficiently large we can make the right hand side of (A.97) less than ϵ.

□

Choice of Tuning Parameter λ_n In this part we state a theorem for tuning parameter choice that guarantees variable selection consistency of the variant of the conservative Lasso. We discuss the assumptions needed in detail. Basically, we show that the variant of the conservative Lasso in (6) fits into Corollary 1 of Fan and Tang (2013). For this we assume deterministic regressors and gaussian error terms which simplifies

the conditions of the following theorem a bit. The case of non-gaussianity can be handled as in Condition 3, p.544 of Fan and Tang (2013) but brings more notation.

Denote the set of λ_n that result in an underfit by

$$\Omega_- = \{\lambda_n \in [\lambda_l, \lambda_u] : S_{\lambda_n} \not\supset S_{\lambda_0}\},$$

where λ_0 represents an ideal tuning parameter that provides the correct model. Thus, $S_{\lambda_0} = S_0$. λ_l and λ_u can be chosen as described on p.540 in Fan and Tang (2013). Denote the set of λ_n that result in an overfit by

$$\Omega_+ = \{\lambda_n \in [\lambda_l, \lambda_u] : S_{\lambda_n} \supset S_{\lambda_0}, S_{\lambda_n} \neq S_{\lambda_0}\}.$$

The following theorem shows that the λ_n choice that minimizes GIC will ensure that the variant of the conservative Lasso detects the correct model with probability approaching one. The conditions for the theorem are discussed in detail after the theorem statement.

Theorem 5. *Under Conditions 1-7 below*

$$P\{\inf_{\lambda_n \in \Omega_- \cup \Omega_+} GIC(\lambda_n) > GIC(\lambda_0)\} \to 1.$$

Theorem 5 yields that the λ_n chosen by GIC will neither result in an underfit nor an overfit. Hence, consistent model selection is achieved.

The penalty function for each parameter is defined as $\rho_{\lambda_n}(|\beta_j|) = \lambda_n \tilde{w}_j |\beta_j|$ for the variant of the conservative Lasso. The partial derivative of the penalty function with respect to $\beta_j,\ j \in S_0$ evaluated at $\beta_{0,j}$ is

$$sgn(\beta_{0,j})\rho'_{\lambda_n}(|\beta_{0,j}|). \tag{A.98}$$

Condition 1. For each λ_n, $\rho'_{\lambda_n}(t)$ is non-increasing over $t \in (0, \infty)$.

Condition 2. There is a $\lambda_0 \in [\lambda_l, \lambda_u]$ such that $S_{\lambda_0} = S_0$, and

$$\|\tilde{\beta}_{\lambda_0} - \beta_0\|_2 = O_p(n^{-\pi}),$$

with $0 < \pi < 1/2$.

Condition 3. $n^\pi \min_{j \in S_0} |\beta_{0,j}| \to \infty$, as $n \to \infty$.

Condition 4. $\rho'_{\lambda_0}(\frac{1}{2} \min_{j \in S_0} |\beta_{0,j}|) = o(s_0^{-1/2} n^{-1/2} [\log\log(n) \log(p)]^{1/2})$.

Condition 5. For any $S \subset \{1, 2, ..., p\}$ such that $|S| \leq K_1$, $K_1 > s_0$, $K_1 = o(n)$ the minimum eigenvalue of $n^{-1} X_S' X_S$ is bounded from below by $c_1 > 0$, and the maximum eigenvalue is bounded from above by $1/c_1$.

Condition 6. The design matrix satisfies $\|X\|_\infty = O(n^{1/2-\tau_1})$ with $\tau_1 \in (1/3, 1/2]$ and $\log(p) = O(n^{\kappa_1})$, for some $0 < \kappa_1 < 1$.

Condition 7. Let δ_n be as in (3.2) of Fan and Tang (2013).We assume $\delta_n K_1^{-1} \sqrt{n/\log(p)} \to \infty$, and

$$n\delta_n / (s_0 \log\log(n) \log(p)) \to \infty.$$

Conditions 1-3 are Condition 4 in p.544 of Fan and Tang (2013). Our Condition 4 is in the statement of Proposition 1 on p.535 of Fan and Tang (2013). Condition 5 here is Condition 2 on p.544 of Fan and Tang (2013). Condition 6 is a condition on p.537 of Theorem 2 of Fan and Tang (2013). Condition 7 is in p.539, Corollary 1 of Fan and Tang (2013). δ_n is a measure of the smallest signal strength of the truly relevant covariates. Conditions 5-7 are related to the linear model and have already been verified in Fan and Tang (2013).

Conditions 1-7 here replace Assumptions 1-2, and the beta-min type condition in Lemma 1, and Theorem 4. Conditions 1-7 are more restrictive than Assumptions 1-2.

Further discussion of Conditions 1-7 We now discuss Conditions 1-7 in more detail in our setting to better understand when Corollary 1 in Fan and Tang (2013) applies.

Let us start by verifying Condition 1. For all $t \in (0, \infty)$, the variant of the conservative lasso

$$\rho'_{\lambda_n}(t) = \lambda_n 1_{\{|\hat{\beta}_{L,j}| \leq \lambda_{prec}\}}.$$

which is constant in t.

Regarding Condition 2, as Theorem 1 applies to the variant of conservative Lasso as well, we get that

$$\|\tilde{\beta} - \beta_0\|_2 \leq \|\hat{\beta} - \beta_0\|_1 = O_p(\lambda_n s_0).$$

In the case of deterministic regressors, and Gaussian random errors, $\lambda_n = O(\sqrt{\log(p)/n})$, so Condition 2 will be fulfilled if $\sqrt{\log(p)/n} s_0 = O(1/n^\pi)$ for $0 < \pi < 1/2$.

Condition 3 is a refinement of a beta-min type condition and restricts the size of the smallest absolute value of the non-zero coefficients.

Condition 4 is the following in case of the variant of conservative lasso,

$$\rho'_{\lambda_0}(\frac{1}{2} \min_{j \in S_0} |\beta_{0,j}|) = \lambda_0 1_{\{\frac{1}{2} \min_{j \in S_0} |\beta_{0,j}| \leq \lambda_{prec}\}}.$$

With the beta-min condition in Theorem 4, $\min_{j \in S_0} |\beta_{0,j}| > 2\lambda_{prec}$, we have $\frac{1}{2} \min_{j \in S_0} |\beta_{0,j}| > \lambda_{prec}$, so the indicator is always zero such that $\rho'_{\lambda_0}(\frac{1}{2} \min_{j \in S_0} |\beta_{0,j}|) = 0$ implying that Condition 4 is trivially satisfied.

Conditions 5-6 are about design of the regression and are used by Fan and Tang (2013) in the least squares case. They are more restrictive than our Assumption 1. Condition 7 is related to underfit of a model in least squares.

Appendix C

We first show why $\hat{\Theta}$ constructed by nodewise regressions is an approximate inverse of $\hat{\Sigma}$. Then we link the inverse of the population covariance matrix Θ to linear regression.

We show that

$$\|\hat{\Theta}'_j \hat{\Sigma} - e'_j\|_\infty \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}.$$

for $j = 1, ..., p$ as claimed in (18). First, note that

$$sgn(\hat{\gamma}_j)' \hat{\Gamma}_j \hat{\gamma}_j = \left\|\hat{\Gamma}_j \hat{\gamma}_j\right\|_1, \tag{A.99}$$

where $sgn(\hat{\gamma}_j) = \big(sgn(\hat{\gamma}_{j,k}),\ k = 1, ..., p, k \neq j\big)$. Therefore, postmultiplying the Karush-Kuhn-Tucker conditions (written as a row vector) of the problem (14) by $\hat{\gamma}_j$ and adding $(X_j - X_{-j}\hat{\gamma}_j)' X_j / n$ to both sides yields

$$\frac{(X_j - X_{-j}\hat{\gamma}_j)'(X_j - X_{-j}\hat{\gamma}_j)}{n} + \lambda_{node,n} \left\|\hat{\Gamma}_j \hat{\gamma}_j\right\|_1 = \frac{(X_j - X_{-j}\hat{\gamma}_j)' X_j}{n}. \tag{A.100}$$

Next, we recognize the left hand side of (A.100) as $\hat{\tau}_j^2$ such that

$$\hat{\tau}_j^2 = \frac{(X_j - X_{-j}\hat{\gamma}_j)' X_j}{n}. \tag{A.101}$$

Dividing each side of the above display by $\hat{\tau}_j^2$ (we shall later rigorously argue that $\hat{\tau}_j^2$ is bounded away from zero with high probability) and using the definition of $\hat{\Theta}_j$ implies that

$$1 = \frac{(X_j - X_{-j}\hat{\gamma}_j)'X_j}{\hat{\tau}_j^2 n} = \frac{(X\hat{\Theta}_j)'X_j}{n} = \frac{\hat{\Theta}_j' X' X_j}{n}, \tag{A.102}$$

which shows that the j'th diagonal element of $\hat{\Theta}\hat{\Sigma}$ equals exactly one. It remains to consider the off-diagonal elements of $\hat{\Theta}\hat{\Sigma}$. To this end, note that the Karush-Kuhn-Tucker conditions for the problem (14) can be written as

$$\hat{\kappa}_j = \frac{\hat{\Gamma}_j^{-1} X_{-j}'(X_j - X_{-j}\hat{\gamma}_j)}{n\lambda_{node,n}}.$$

Using $\|\hat{\kappa}_j\|_\infty \leq 1$ yields

$$\left\| \frac{\hat{\Gamma}_j^{-1} X_{-j}'(X_j - X_{-j}\hat{\gamma}_j)}{n\lambda_{node,n}} \right\|_\infty = \|\hat{\kappa}_j\| \leq 1,$$

which is equivalent to

$$\frac{\|\hat{\Gamma}_j^{-1} X_{-j}' X \hat{C}_j\|_\infty}{n} \leq \lambda_{node,n},$$

since $(X_j - X_{-j}\hat{\gamma}_j) = X\hat{C}_j$. Then, dividing both sides of the above display by $\hat{\tau}_j^2$ and using that $\hat{\Theta}_j = \frac{\hat{C}_j}{\hat{\tau}_j^2}$ implies that

$$\frac{\|\hat{\Gamma}_j^{-1} X_{-j}' X \hat{\Theta}_j\|_\infty}{n} \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}.$$

Thus,

$$\frac{\|X_{-j}' X \hat{\Theta}_j\|_\infty}{n} = \frac{\|\hat{\Gamma}_j \hat{\Gamma}_j^{-1} X_{-j}' X \hat{\Theta}_j\|_\infty}{n} \leq \|\hat{\Gamma}_j\|_{\ell_\infty} \frac{\|\hat{\Gamma}_j^{-1} X_{-j}' X \hat{\Theta}_j\|_\infty}{n} \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}, \tag{A.103}$$

where we have used that $\|\hat{\Gamma}_j\|_{\ell_\infty}$ equals the largest diagonal element of $\hat{\Gamma}_j$ since $\hat{\Gamma}_j$ is diagonal and that all diagonal elements are less than one. Of course (A.103) is equivalent to

$$\frac{\left\|\hat{\Theta}_j' X' X_{-j}\right\|_\infty}{n} \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}. \tag{A.104}$$

In total, denoting by e_j the j'th $p \times 1$ unit vector, (A.102) and (A.104) yield

$$\|\hat{\Theta}_j'\hat{\Sigma} - e_j'\|_\infty \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}.$$

References

Belloni, A., D. Chen, V. Chernozhukov, and H. Christian (2010). Sparse models and methods for optimal instruments with an application to eminent domain. *arXiv preprint arXiv:1010.4345*.

Belloni, A., D. Chen, V. Chernozhukov, and H. Christian (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica 80*, 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen (2011a). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., V. Chernozhukov, and C. Hansen (2011b). Inference on treatment effects after selection among high-dimensional controls. *arXiv*, 1201.0224v3.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berk, R., L. Buja, A. Zhang, and L. Zhao (2013). Valid post selection inference. *The Annals of Statistics 41*(2), 802–837.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High Dimensional Data*. Springer Verlag.

Davidson, J. (2000). *Econometric Theory*. Blackwell Publishers.

Fan, J., Y. Fan, and E. Barut (2014). Adaptive robust variable selection. *Annals of Statistics 42*, 324–351.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high dimensional cross section tests. *Econometrica 83*, 1497–1541.

Fan, J. and J. Lv (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of Royal Statistical Society Series B*, 849–911.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 101–148.

Fan, J., L. Xue, and H. Zou (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics 42*, 819–849.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of Royal Statistical Society Series B 75*, 531–552.

Hoffmann, M. and R. Nickl (2011). On adaptive inference and confidence bands. *The Annals of Statistics 39*, 2833–2409.

Javanmard, A. and A. Montanari (2013). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv preprint arXiv:1301.4240*.

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

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 1001–1008.

Lin, Z. and Z. Bai (2010). *Probability inequalities*. Springer.

Lockhart, R., J. Taylor, R. Tibshirani, and R. Tibshirani (2014). A significance test for the lasso. *The Annals of Statistics 42*, 413–430.

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

Nickl, R. and S. van de Geer (2013). Confidence sets in sparse regression. *The Annals of Statistics 41*(6), 2852–2876.

Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 1–18.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Taylor, J. and R. Tibshirani (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences 112*, 7629–7634.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B 73*, 273–282.

van de Geer, S. (2014). *Statistical Theory for High Dimensional Models*. Lecture Notes.

van de Geer, S. and P. Bühlmann (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics 3*, 1360–1392.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*.

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

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research 11*, 2261–2286.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.

Zou, H. and R. Li (2008). One step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics 36*, 1509–1533.

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	0.668	0.136	0.949	0.852	0.929	0.386	0.383
	LassoGIC	0.721	0.116	0.936	0.873	0.940	0.411	0.404
	CLasso	0.516	0.097	0.953	0.888	0.952	0.381	0.383
$\rho = 0$	CLassoGIC	0.589	0.100	0.944	0.889	0.955	0.396	0.396
	CLassoInd	0.361	0.083	0.964	0.906	0.950	0.364	0.371
	CLassoIndGIC	0.371	0.077	0.962	0.913	0.951	0.369	0.375
	J&M	0.824	0.007	0.383	0.989	0.990	0.787	0.776
	Lasso	0.709	0.146	0.900	0.852	0.918	0.394	0.409
	LassoGIC	0.741	0.138	0.860	0.867	0.921	0.411	0.422
	CLasso	0.491	0.093	0.917	0.888	0.954	0.397	0.417
$\rho = 0.5$	CLassoGIC	0.540	0.092	0.892	0.889	0.956	0.405	0.423
	CLassoInd	0.392	0.086	0.941	0.897	0.953	0.387	0.415
	CLassoIndGIC	0.378	0.083	0.945	0.907	0.958	0.388	0.413
	J&M	0.867	0.012	0.300	0.993	0.991	0.896	0.992
	Lasso	1.392	0.201	0.630	0.820	0.854	0.617	0.738
	LassoGIC	1.392	0.199	0.634	0.815	0.855	0.608	0.722
	CLasso	1.214	0.137	0.529	0.885	0.922	0.772	0.961
$\rho = 0.9$	CLassoGIC	1.224	0.132	0.524	0.887	0.927	0.769	0.947
	CLassoInd	1.395	0.136	0.483	0.881	0.912	0.828	1.121
	CLassoIndGIC	1.362	0.130	0.478	0.882	0.921	0.838	1.134
	J&M	1.532	0.025	0.126	0.978	0.978	1.561	2.093

Table 1: Summary statistics for Experiment 1a. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	0.738	0.158	0.765	0.854	0.898	0.557	0.563
	LassoGIC	0.790	0.143	0.735	0.869	0.914	0.582	0.588
	CLasso	0.610	0.132	0.755	0.875	0.933	0.567	0.581
$\rho = 0$	CLassoGIC	0.685	0.130	0.734	0.875	0.932	0.578	0.591
	CLassoInd	0.450	0.120	0.776	0.890	0.938	0.562	0.579
	CLassoIndGIC	0.494	0.113	0.759	0.887	0.942	0.567	0.584
	J&M	0.904	0.012	0.289	0.984	0.981	1.000	1.002
	Lasso	0.780	0.193	0.774	0.828	0.913	0.609	0.534
	LassoGIC	0.815	0.183	0.737	0.835	0.925	0.630	0.554
	CLasso	0.593	0.148	0.778	0.860	0.960	0.631	0.553
$\rho = 0.5$	CLassoGIC	0.656	0.143	0.769	0.860	0.960	0.642	0.564
	CLassoInd	0.477	0.134	0.821	0.864	0.962	0.629	0.551
	CLassoIndGIC	0.485	0.130	0.813	0.868	0.968	0.638	0.557
	J&M	0.952	0.013	0.258	0.978	0.985	1.130	1.138
	Lasso	1.484	0.218	0.524	0.792	0.867	0.789	0.835
	LassoGIC	1.482	0.225	0.523	0.790	0.870	0.784	0.823
	CLasso	1.364	0.151	0.457	0.847	0.928	0.928	1.051
$\rho = 0.9$	CLassoGIC	1.384	0.148	0.453	0.849	0.926	0.928	1.041
	CLassoInd	1.511	0.158	0.432	0.855	0.925	0.973	1.212
	CLassoIndGIC	1.483	0.151	0.428	0.860	0.932	0.987	1.228
	J&M	1.634	0.035	0.132	0.963	0.975	1.807	2.323

Table 2: Summary statistics for Experiment 1b. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
$\rho = 0$	Lasso	0.398	0.058	0.901	0.946	0.931	0.435	0.412
	LassoGIC	0.425	0.051	0.902	0.959	0.933	0.444	0.414
	CLasso	0.375	0.061	0.905	0.949	0.930	0.428	0.408
	CLassoGIC	0.413	0.057	0.901	0.958	0.934	0.439	0.413
	CLassoInd	0.315	0.076	0.906	0.929	0.925	0.486	0.467
	CLassoIndGIC	0.368	0.070	0.911	0.941	0.934	0.422	0.406
	J&M	0.348	0.135	0.973	0.862	0.955	0.373	0.360
$\rho = 0.5$	Lasso	0.337	0.162	0.687	0.928	0.823	0.439	0.436
	LassoGIC	0.354	0.189	0.613	0.937	0.790	0.451	0.442
	CLasso	0.315	0.142	0.720	0.924	0.846	0.435	0.437
	CLassoGIC	0.343	0.173	0.650	0.930	0.813	0.448	0.441
	CLassoInd	0.282	0.096	0.849	0.911	0.916	0.419	0.431
	CLassoIndGIC	0.334	0.131	0.774	0.919	0.881	0.432	0.434
	J&M	0.310	0.429	0.919	0.787	0.767	0.316	0.301
$\rho = 0.9$	Lasso	0.451	0.237	0.407	0.841	0.796	0.642	0.748
	LassoGIC	0.456	0.275	0.381	0.844	0.768	0.637	0.728
	CLasso	0.513	0.163	0.458	0.878	0.900	0.784	0.942
	CLassoGIC	0.527	0.175	0.428	0.873	0.895	0.779	0.915
	CLassoInd	0.556	0.076	0.386	0.926	0.935	0.916	1.228
	CLassoIndGIC	0.647	0.071	0.359	0.932	0.934	0.944	1.251
	J&M	0.440	0.652	0.908	0.491	0.597	0.292	0.302

Table 3: Summary statistics for Experiment 2a. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC.J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	0.445	0.082	0.714	0.923	0.945	0.631	0.634
	LassoGIC	0.472	0.070	0.701	0.932	0.950	0.642	0.641
	CLasso	0.430	0.088	0.715	0.920	0.950	0.626	0.634
$\rho = 0$	CLassoGIC	0.465	0.075	0.704	0.929	0.950	0.639	0.642
	CLassoInd	0.396	0.085	0.713	0.914	0.946	0.696	0.702
	CLassoIndGIC	0.445	0.083	0.712	0.917	0.950	0.624	0.639
	J&M	0.395	0.136	0.771	0.848	0.954	0.567	0.573
	Lasso	0.391	0.184	0.545	0.918	0.875	0.698	0.587
	LassoGIC	0.406	0.202	0.501	0.922	0.861	0.715	0.599
	CLasso	0.381	0.167	0.587	0.906	0.898	0.695	0.589
$\rho = 0.5$	CLassoGIC	0.403	0.186	0.528	0.912	0.877	0.711	0.600
	CLassoInd	0.392	0.150	0.658	0.888	0.940	0.681	0.588
	CLassoIndGIC	0.425	0.170	0.607	0.887	0.927	0.696	0.596
	J&M	0.370	0.504	0.787	0.804	0.840	0.565	0.480
	Lasso	0.512	0.220	0.315	0.879	0.804	0.870	0.862
	LassoGIC	0.514	0.245	0.301	0.877	0.777	0.869	0.846
	CLasso	0.586	0.143	0.343	0.885	0.914	0.979	1.034
$\rho = 0.9$	CLassoGIC	0.597	0.148	0.317	0.882	0.896	0.978	1.011
	CLassoInd	0.698	0.083	0.316	0.934	0.953	1.104	1.324
	CLassoIndGIC	0.765	0.081	0.304	0.936	0.957	1.132	1.349
	J&M	0.500	0.674	0.824	0.633	0.636	0.527	0.483

Table 4: Summary statistics for Experiment 2b. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$\rho = 0.75$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	1.551	0.760	0.880	0.250	0.730	0.232	0.229
	LassoGIC	3.066	0.060	0.040	0.960	0.860	1.541	1.580
	CLasso	1.006	0.220	0.780	0.830	0.910	0.479	0.494
$n = 100$	CLassoGIC	3.066	0.060	0.040	0.960	0.850	1.551	1.588
	CLassoInd	1.419	0.370	0.750	0.590	0.870	1.646	1.049
	CLassoIndGIC	3.066	0.070	0.060	0.970	0.860	1.631	1.653
	J&M	1.514	0.930	0.980	0.220	0.810	0.247	0.242
	Lasso	1.099	0.320	0.780	0.670	0.800	0.336	0.361
	LassoGIC	1.400	0.090	0.340	0.960	0.840	0.579	0.616
	CLasso	0.798	0.050	0.770	0.960	0.880	0.416	0.454
$n = 150$	CLassoGIC	1.418	0.080	0.320	0.960	0.840	0.595	0.632
	CLassoInd	0.875	0.270	0.820	0.710	0.910	0.537	0.433
	CLassoIndGIC	1.432	0.090	0.410	0.960	0.880	0.669	0.720
	J&M	0.937	0.830	0.990	0.400	0.740	0.204	0.205
	Lasso	0.876	0.060	0.860	0.880	0.930	0.394	0.436
	LassoGIC	1.036	0.070	0.710	0.930	0.930	0.450	0.489
	CLasso	0.694	0.040	0.910	0.950	0.930	0.391	0.437
$n = 200$	CLassoGIC	1.002	0.060	0.750	0.930	0.930	0.458	0.497
	CLassoInd	0.397	0.080	0.910	0.910	0.920	0.439	0.507
	CLassoIndGIC	0.864	0.100	0.740	0.900	0.870	0.496	0.556
	J&M	0.746	0.490	1.000	0.530	0.890	0.204	0.209
	Lasso	0.494	0.080	1.000	0.930	0.960	0.246	0.282
	LassoGIC	0.552	0.070	1.000	0.940	0.950	0.254	0.289
	CLasso	0.254	0.060	1.000	0.920	0.970	0.250	0.295
$n = 500$	CLassoGIC	0.307	0.050	1.000	0.930	0.970	0.252	0.295
	CLassoInd	0.139	0.080	1.000	0.930	0.970	0.263	0.329
	CLassoIndGIC	0.139	0.080	1.000	0.930	0.970	0.263	0.329
	J&M	0.420	0.150	1.000	0.770	0.930	0.193	0.217

Table 5: Summary statistics for Experiment 3a. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$\rho = 0.75$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	1.667	0.680	0.880	0.300	0.870	0.297	0.271
	LassoGIC	3.074	0.080	0.050	0.950	0.860	1.664	1.616
	CLasso	1.225	0.370	0.790	0.640	0.920	0.557	0.512
$n = 100$	CLassoGIC	3.074	0.080	0.050	0.950	0.860	1.673	1.623
	CLassoInd	1.578	0.380	0.730	0.600	0.890	1.814	1.120
	CLassoIndGIC	3.092	0.090	0.060	0.950	0.870	1.765	1.711
	J&M	1.610	0.860	0.980	0.330	0.840	0.360	0.330
	Lasso	1.206	0.370	0.710	0.690	0.900	0.465	0.424
	LassoGIC	1.693	0.120	0.290	0.950	0.930	0.841	0.800
	CLasso	0.906	0.100	0.690	0.910	0.960	0.592	0.550
$n = 150$	CLassoGIC	1.703	0.110	0.280	0.950	0.930	0.850	0.812
	CLassoInd	1.070	0.370	0.800	0.640	0.910	0.527	0.478
	CLassoIndGIC	1.708	0.090	0.360	0.900	0.940	0.910	0.910
	J&M	1.040	0.810	0.980	0.480	0.910	0.352	0.325
	Lasso	0.978	0.150	0.680	0.850	0.930	0.548	0.517
	LassoGIC	1.170	0.120	0.520	0.880	0.920	0.622	0.587
	CLasso	0.856	0.110	0.730	0.850	0.950	0.561	0.531
$n = 200$	CLassoGIC	1.150	0.100	0.520	0.900	0.930	0.632	0.598
	CLassoInd	0.628	0.120	0.750	0.860	0.970	0.586	0.590
	CLassoIndGIC	1.067	0.090	0.600	0.890	0.960	0.667	0.671
	J&M	0.842	0.610	0.990	0.550	0.910	0.340	0.304
	Lasso	0.548	0.100	0.980	0.880	0.940	0.380	0.332
	LassoGIC	0.610	0.100	0.970	0.890	0.950	0.389	0.341
	CLasso	0.316	0.100	1.000	0.890	0.950	0.381	0.341
$n = 500$	CLassoGIC	0.378	0.100	1.000	0.890	0.940	0.384	0.343
	CLassoInd	0.177	0.100	0.990	0.910	0.950	0.387	0.367
	CLassoIndGIC	0.182	0.100	0.990	0.910	0.950	0.387	0.367
	J&M	0.473	0.190	1.000	0.780	0.930	0.327	0.272

Table 6: Summary statistics for Experiment 3b. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).

			χ^2		Coverage rate		Length	
$\rho = 0.5$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
	Lasso	0.337	0.174	0.640	0.928	0.823	0.439	0.436
	LassoGIC	0.354	0.187	0.600	0.937	0.790	0.451	0.442
	CLasso	0.315	0.160	0.678	0.924	0.846	0.435	0.437
$n = 100$	CLassoGIC	0.343	0.181	0.629	0.930	0.813	0.448	0.441
	CLassoInd	0.282	0.161	0.807	0.911	0.916	0.419	0.431
	CLassoIndGIC	0.334	0.200	0.766	0.919	0.881	0.432	0.434
	J&M	0.310	0.597	0.930	0.787	0.767	0.316	0.301

Table 7: Summary statistics for Experiment 4. ℓ_2: average ℓ_2-estimation error, χ^2: Size and Power report the size and power of the hypotheses $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0, 1, 0.1, 0, 0, 0, 0, 0, 0)$ and $H_0 : (\beta_{0,1}, \beta_{0,2}) = (1, 0.4, 1, 0.1, 0, 0, 0, 0, 0, 0)$, respectively. Coverage rate: the actual coverage rate of the asymptotically gaussian 95% confidence interval for $\beta_{0,1}$ and $\beta_{0,2}$. Length: the length of the two confidence intervals mentioned above. Lasso: Lasso with BIC. LassoGIC: Lasso with GIC. CLasso: Conservative Lasso with BIC. CLassoGIC: Conservative Lasso with GIC. CLassoInd: Variant of Conservative Lasso with BIC. CLassoIndGIC: Variant of Conservative lasso with GIC. J&M: Procedure of Javanmard and Montanari (2014).