AARHUS UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Asymptotically Honest Confidence Regions for High Dimensional

Mehmet Caner and Anders Bredahl Kock

CREATES Research Paper 2014-36

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

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

Mehmet Caner* Anders Bredahl Kock†

October 15, 2014

Abstract

While variable selection and oracle inequalities for the estimation and prediction error have received considerable attention in the literature on high-dimensional models, very little work has been done in the area of testing and construction of confidence bands in high-dimensional models. However, in a recent paper van de Geer et al. (2014) showed how the Lasso can be desparsified in order to create asymptotically honest (uniform) confidence band. In this paper we consider the conservative Lasso which penalizes more correctly than the Lasso and hence has a lower estimation error.

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 which in our context provides sharper bounds than Nemirovski's inequality. As opposed to van de Geer et al. (2014) 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. As a stepping stone towards this, we also provide a feasible uniformly consistent estimator of the asymptotic covariance matrix of an increasing number of parameters which is robust against conditional heteroskedasticity. To our knowledge we are the first to do so. Next, we show that our confidence bands are honest over sparse high-dimensional sub vectors of the parameter space and that they contract at the optimal rate. All our results are valid in high-dimensional models. Our simulations reveal that the desparsified conservative Lasso estimates the parameters much more

*North Carolina State University, Department of Economics, 4168 Nelson Hall, Raleigh, NC 27695. Email: mcaner@ncsu.edu.

†Aarhus University and CREATES, Department of Economics and Business, Fuglesangs Alle 4, 8210 Aarhus V, Denmark. Email: akock@creates.au.dk.

precisely than the desparsified Lasso, has much better size properties and produces confidence bands with markedly superior coverage rates.

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

JEL classification: 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. At first, much focus was devoted to establishing the so-called oracle property in models of fixed or increasing dimensions, see e.g. Fan and Li (2001), Zou (2006), and Huang et al. (2008). This entails showing that the procedure asymptotically detects the correct, and only the correct, variables and that the non-zero coefficients have the same asymptotic distribution as if only the relevant variables had been included in the model from the outset.

The oracle property is an asymptotic one and in recent years more focus has been devoted to establishing finite sample oracle inequalities for the estimation and prediction error. That is, finite sample upper bounds on the estimation and prediction error that are valid with high probability. Pioneering work in this direction was done by Bickel et al. (2009). For excellent reviews, see Bühlmann and van de Geer (2011), Fan and Lv (2010), and Belloni and Chernozhukov (2011).

Methods for analyzing high-dimensional data sets have also found increasing use in econometrics as many economic data sets are becoming increasingly high-dimensional. For example, Belloni et al. (2012) have shown how the least absolute shrinkage and selection operator (Lasso) of Tibshirani (1996) can be used to select among many instruments. Furthermore, Belloni et al. (2012) have shown how the Lasso can be used to make inference on treatment effects in the presence of a high-dimensional set of control variables. The authors also allow for non-gaussian heteroskedastic errors and construct uniformly valid confidence bands for the treatment effect. Caner and Zhang (2014) have considered GMM estimation with many parameters while Kock (2013) has considered estimation and inference in large panel data models.

Despite the big progress that has been made in terms of analyzing high-dimensional models very little progress has been made in the area of testing and constructing reliable confidence bands. However, recently van de Geer et al. (2014) 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. The idea behind this procedure is to remove the bias introduced by shrinkage by desparsifying the estimator and constructing a clever approximate inverse to 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 consistent variable selection techniques. By using the 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 an 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 introduce the so-called conservative Lasso to econometrics (see Section 2.3). The important observation here is that, in the presence of an oracle inequality on the plain Lasso, the penalty of 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 simulations strongly indicate that this is the case.

We provide an oracle inequality for the conservative Lasso estimator and use the method of desparsification introduced in van de Geer et al. (2014) to conduct inference without the use of any assumptions regarding the zero and non-zero coefficients being well-separated (no $\beta_{\min}$-condition is imposed). 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 deliver slightly more precise estimates than Nemirovski's inequality.

We also show that hypotheses involving an increasing number of parameters can be tested 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 econometric problems over van de Geer et al. (2014) who focus on low-dimensional hypotheses in the presence of homoskedastic error terms and do not provide a practically feasible estimator of the asymptotic covariance matrix. However, heteroskedasticity is omniscient in econometrics and thus our procedure is of practical interest as it is the first of its kind which is able to handle high-dimensionality and heteroskedasticity simultaneously. Next,

we show that the weak convergence to the normal distribution of our estimator is valid uniformly. More importantly, this is used to show that confidence bands based on the desparsified conservative Lasso are *honest* over the subset of the parameter space consisting of sparse vectors. Thus, there exists a fixed time, *not depending on the true parameter* β_0, from which on our confidence bands have coverage close to the desired coverage probability. This is in stark opposition to *dishonest* confidence intervals. While dishonest confidence intervals might still have the correct coverage rate asymptotically the sample size which is needed in order to achieve this coverage rate may depend on the unknown true parameter β_0. This is unfortunate for the applied researcher who will not know how large a sample is needed in order to achieve a desired coverage rate. Finally, we show that the confidence bands have uniformly the optimal rate of contraction such that their honesty is not bought at the price of them being wide, see Pötscher (2009). As we shall stress again in the discussion of Theorem 3 below honest confidence are remarkable as Bahadur and Savage (1956) have shown that honest confidence intervals can not even exist for the mean based on an iid gaussian sample if one insists on these bands to have a finite length almost surely. This also underscores that one can probably not hope for honesty to be valid over the whole parameter space.

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, 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. Finally, and perhaps most importantly, the confidence intervals based on our procedure have coverage rates much closer to the nominal rate than the confidence bands based on the desparsified plain Lasso. This improvement in the coverage rates comes directly from more precise parameter estimates as well as indirectly through a more precise estimate of the covariance matrix by the use of the conservative Lasso for nodewise regressions instead of the plain Lasso.

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 establishes honest confidence intervals and shows that they contract at an optimal rate. The simulations can be found in Section 5. Section 6 provides a summary. 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 modification of x that places zeros in all entries of x whose index does not belong to S. For an $n \times n$ matrix B let B_S denote submatrix of B consisting only of the rows and columns indexed by S. If $S = \{j\}$ is a singleton set, we use B_j as shorthand for the j'th diagonal element of B.

For any set S, we shall let $|S|$ denote its cardinality and for an $n-$dimensional vector x, $\|x\|_n^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$. $\xrightarrow{d}$ 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 only differ by a multiplicative constant.

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. We assume that the explanatory variables are exogenous and precise assumptions will be made in Assumption 1 below. 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

Before we introduce the precise definition of the conservative Lasso we recall that the plain Lasso of Tibshirani (1996) is defined as

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

where λ_n is a positive tuning parameter determining the size of the penalty attached to non-zero parameters. For λ_n sufficiently large, some parameters will be classified exactly as zero. The plain Lasso attaches the same penalty to all parameters. However, ideally one would like to penalize the truly non-zero parameters less than the truly zero ones. The problem is that one does not know which parameters are zero and which are not. One potential solution to this would be to apply the adaptive Lasso of Zou (2006) which is defined as

$$\hat{\beta}_{AL} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| Y - X\beta \right\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \frac{1}{|\hat{\beta}_{L,j}|} |\beta_j| \,, \tag{3}$$

where $\hat{\beta}_{L,j}$ is the Lasso estimator for the jth coefficient. However, and first of all, unless one imposes a restrictive condition on the minimal size of non-zero coefficients, not even the adaptive Lasso can be guaranteed to penalize truly zero coefficients more than truly non-zero ones. Next, the adaptive Lasso usually relies on the Lasso as an initial estimator to construct its weights. In particular, the adaptive Lasso discards those variables from the second step estimation which have been deemed irrelevant by the first step Lasso estimator. This is unfortunate since we want to construct confidence intervals and tests for every parameter $\beta_{0,j},\ j = 1, ..., p$. The conservative Lasso does not suffer from these two shortcomings and we shall introduce it next. The conservative Lasso estimator is defined as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| Y - X\beta \right\|_n^2 + 2\lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j| \tag{4}$$

with $\hat{w}_j = \frac{\lambda_{prec}}{|\hat{\beta}_{L,j}| \vee \lambda_{prec}}$. Here λ_n and λ_{prec} are positive non-random quantities chosen by the researcher which we shall be specific about shortly. 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 $|\hat{\beta}_j| = 0$ 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) since we have merged one of the tuning parameters into the definition of the weights. However, the difference is merely a matter of parameterization. Furthermore, these authors do not provide any inferential procedure for the conservative Lasso, they merely suggest it as a potential estimator.

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 $\mathcal{C} = \left\{\|\hat{\beta}_L - \beta_0\|_1 \leq \lambda_{prec}\right\}$ is a set with large probability. In Lemma 1 and Theorem 1 below we shall give examples of λ_{prec}. Its order will be the rate of convergence of the Lasso estimator in the l_1-norm. Observe that

1. $\hat{w}_j \leq 1$ for all $j = 1, ..., p$.

   Furthermore, on $\mathcal{C} = \left\{\|\hat{\beta}_L - \beta_0\|_1 \leq \lambda_{prec}\right\}$ we have the following two properties:

2. $\hat{w}_j = 1$ for all $j \in S_0^c$ since $|\hat{\beta}_{L,j}| = |\hat{\beta}_{L,j} - \beta_{0,j}| \leq \lambda_{prec}$ for all $j \in S_0^c$.

3. $\hat{w}_j \to 0$ for $j \in S_0$ if $\frac{|\beta_{0,j}|}{\lambda_{prec}} \to \infty$. This is because $|\hat{\beta}_{L,j}| \geq |\beta_{0,j}| - |\hat{\beta}_{L,j} - \beta_{0,j}| \geq |\beta_{0,j}| - \lambda_{prec} = \lambda_{prec}\left(\frac{|\beta_{0,j}|}{\lambda_{prec}} - 1\right) \geq \lambda_{prec}$ for n large enough. Hence, $\hat{w}_j \leq \frac{\lambda_{prec}}{|\hat{\beta}_{L,j}|} \leq \frac{1}{\frac{|\beta_{0,j}|}{\lambda_{prec}} - 1} \to 0$

Observations 1) and 2) imply that the penalty attached to non-zero coefficients will never be larger than the penalty attached to the truly zero coefficients on $\mathcal{C} = \left\{\|\hat{\beta}_L - \beta_0\|_1 \leq \lambda_{prec}\right\}$. As we shall see below this set has a large probability. Observations 1) and 2) also imply that the conservative Lasso penalizes the non-zero coefficients less than the plain Lasso does since the latter corresponds to $\hat{w}_j = 1$ for all $j = 1, ..., p$. Put differently, the non-zero coefficients will never be penalized more than the truly zero ones. Observation 2) implies that the truly zero coefficients receive the same penalty as they do when the Lasso is applied. By consistency of the Lasso (see Lemma 1 below) $\hat{\beta}_{L,j}$ is often either zero or close to zero for $j \in S_0^c$. Thus even small values of λ_{prec} ensure that Observation 2 is valid, i.e. the zero coefficients will receive a no smaller penalty than the non-zero ones.

In the situation where the non-zero coefficients are bounded away from zero by more than λ_{prec} (the rate of the initial Lasso estimator) observation 3 implies that one even has that the weights attached to the non-zero coefficients tend to zero. We also want to remark that the arguments in observations 2 and 3 actually only rely on λ_{prec} dominating the in general smaller sup-norm instead of the ℓ_1-norm. Thus, the requirement $\frac{|\beta_{0,j}|}{\lambda_{prec}} \to \infty$ in observation 3 can be relaxed since λ_{prec} can be lowered. However, an upper bound on $\|\hat{\beta} - \beta_0\|_\infty$ for the Lasso is only available under rather

strong assumptions, see e.g. Lounici et al. (2008), and we shall stick to the current setting for now.

To sum up, the conservative Lasso is attractive since on a set with high probability it penalizes the zero coefficients more than the non-zero ones. Thus, on that set the weights are more appropriate than those of the Lasso which we shall see results in great performance gains.

As is standard in the literature we assume that the covariates X_i are iid with $\Sigma = E(X_1 X_1')$ satisfying an adaptive restricted eigenvalue type condition: for $|S| \leq s$,

$$\phi_\Sigma^2(s) = \min_{\substack{\delta \in \mathbb{R}^p \setminus \{0\} \\ \|\delta_{S^c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2}} \frac{\delta' \Sigma \delta}{\|\delta_S\|_2^2} > 0, \tag{5}$$

where $S \subseteq \{1, ..., p\}$ and $|S|$ is its cardinality. Instead of minimizing over all of $\mathbb{R}^p$, the minimum in (5) is restricted to those vectors which satisfy $\|\delta_{S^c}\|_1 \leq 3\sqrt{s}\|\delta_S\|_2$ and where S has cardinality at most s.

Notice that the adaptive restricted eigenvalue condition is trivially satisfied if Σ has full rank since $\delta_S' \delta_S \leq \delta' \delta$ for every $\delta \in \mathbb{R}^p$ and so,

$$\frac{\delta' \Sigma \delta}{\|\delta_S\|_2^2} \geq \frac{\delta' \Sigma \delta}{\|\delta\|_2^2} \geq \min_{\delta \in \mathbb{R}^p \setminus \{0\}} \frac{\delta' \Sigma \delta}{\|\delta\|_2^2} > 0.$$

Assuming Σ to be of full rank is a rather innocent assumption as Σ is nothing else than the population covariance matrix of X_1 in the case where $E(X_1)$ is assumed to have mean zero. Since the population covariance matrix is commonly assumed to be of full rank, the adaptive restricted eigenvalue condition is satisfied in particular. We will also see that under Assumption 1, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ does also satisfy a restricted eigenvalue condition if Σ does so as long as p is not too large.

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

Assumption 1. *The covariates X_i, $i = 1, ..., n$ are independently and identically distributed while the error terms u_i, $i = 1, ..., n$ 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. Furthermore, $\phi_\Sigma^2(s_0)$ is bounded away from 0.*

Assumption 1 assumes 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 economic 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 also deliver stronger results. In any case, it would not be difficult to pursue this avenue but we shall not do so here in order to avoid digressions.

Before stating the oracle inequality 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 1. *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* $1 - \frac{D}{M^{r/2}} - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$*, the Lasso satisfies the following inequalities*

$$\|X(\hat{\beta}_L - \beta_0)\|_n^2 \leq 18\frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}, \tag{6}$$

$$\|\hat{\beta}_L - \beta_0\|_1 \leq 24\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}, \tag{7}$$

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

Lemma 1 provides an oracle inequality for the prediction and estimation error of the Lasso under the assumption of uniformly bounded r'th moments of the covariates and the error terms. It is similar in spirit to previous oracle inequalities, however it does not assume subgaussianity. It is included here as it reveals that $\lambda_{prec} \asymp \lambda_n s_0$ will work in connection with observation 3 above to ensure that $\mathcal{C}$ has a high probability. An upper bound on $\|\hat{\beta}_L - \beta_0\|_\infty$ would actually be more useful for the choice of λ_{prec}. Under a quite restrictive assumption of near orthogonality of Σ, Lounici et al. (2008) has shown that $\|\hat{\beta}_L - \beta_0\|_\infty$ is of the order λ_n such that the unknown s_0 could be dropped in the choice of λ_{prec}.

We are now ready to state the oracle inequality for the conservative Lasso.

Theorem 1. *Let Assumption 1 be satisfied, set* $\lambda_n = M\frac{p^{2/r}}{n^{1/2}}$ *for* $M > 0$ *and* $\lambda_{prec} = 24\frac{\lambda_n s_0}{\phi_\Sigma^2(S_0)}$*. Then, with probability at least* $1 - \frac{D}{M^{r/2}} - D\frac{p^2 s_0^{r/2}}{n^{r/4}}$*, the conservative Lasso satisfies the following*

inequalities

$$\|X(\hat{\beta}-\beta_0)\|_n^2 \leq 18\frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}, \tag{8}$$

$$\|\hat{\beta}-\beta_0\|_1 \leq 24\frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}, \tag{9}$$

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

We shall see in Section 5 that the conservative Lasso provides more precise parameter estimates than the plain Lasso since its weights are more intelligent. For establishing the uniform validity of our covariance matrix estimator over $\mathcal{B}_{\ell_0}(s_0) = \left\{\|\beta_0\|_{\ell_0} \leq s_0\right\}$ in Theorem 2 it will turn out to be important that (9) is valid uniformly over this set. Theorem 1 is mainly used as a tool to prove the validity of our inferential procedure but is also of interest in its own right.

3 Inference

In this section we explain how to conduct statistical inference with the conservative Lasso. To do so we first discuss desparsification.

3.1 The Desparsified Conservative Lasso

In order to conduct inference we shall use the idea of desparsification first proposed in van de Geer et al. (2014). The idea is that the shrinkage bias introduced due to the presence of penalization in (4) 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 (4) may be written as

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

$$\|\hat{\kappa}\|_\infty \leq 1,$$

and $\hat{\kappa}_j = sign(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$ for $j = 1, ..., p$. Using the first equation above

$$\lambda \hat{W}\hat{\kappa} = X'(Y - X\hat{\beta})/n. \tag{10}$$

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 $p \times p$ square 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,

$$\hat{\beta} = \beta_0 - \hat{\Theta}\lambda_n \hat{W}\hat{\kappa} + \hat{\Theta}X'u/n - \Delta/n^{1/2}, \tag{11}$$

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. Note also that the bias term $\hat{\Theta}\lambda_n \hat{W}\hat{\kappa}$ resulting from the penalization is known. This suggests removing it by simply adding it to both sides of (11), resulting in the following estimator:

$$\hat{b} = \hat{\beta} + \hat{\Theta}\lambda_n \hat{W}\hat{\kappa} = \beta_0 + \hat{\Theta}X'u/n - \Delta/n^{1/2}. \tag{12}$$

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

$$\sqrt{n}\alpha'\left(\hat{b} - \beta_0\right) = \alpha'\hat{\Theta}X'u/n^{1/2} - \alpha'\Delta \tag{13}$$

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 asymptotic 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 (13), 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, (13) reduces to

$$\sqrt{n}\left(\hat{b}_j - \beta_{0,j}\right) = \left(\hat{\Theta}X'u/n^{1/2}\right)_j - \Delta_j. \tag{14}$$

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 models with more regressors than observations $(p > n)$, but $h/n \to 0$ as $n \to \infty$.

In the next section we shall 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. Note that we can practically compute the desparsified conservative Lasso from the following equation using (10)

$$\hat{b} = \hat{\beta} + \hat{\Theta} X'(Y - X\hat{\beta})/n.$$

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.

Remark: Another way to remove the shrinkage bias would be to perform least squares after model selection (by the Lasso) as in Belloni and Chernozhukov (2013). However, the Lasso is only guaranteed to include all relevant variable if the non-zero coefficients are well separated from the zero ones ($\beta_{\min}$-condition). An assumption we are avoiding. Thus, in that framework there is in general no guarantee that all relevant variables will be included in the second step least squares estimation and hence inference would be difficult for the parameters of these. However, conducting inference by means of unpenalized post-estimation is an interesting avenue for future research.

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 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

$$\hat{\gamma}_{L,j} = \underset{\gamma \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\| X_j - X_{-j}\gamma \right\|_n^2 + 2\lambda_{node,n} \sum_{k \neq j} |\gamma_k| \tag{15}$$

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

$$\hat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{p-1}}{\operatorname{argmin}} \left\| X_j - X_{-j}\gamma \right\|_n^2 + 2\lambda_{node,n} \left\| \hat{\Gamma}_j \gamma \right\|_1 \tag{16}$$

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. 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 2 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

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

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

$$\hat{\tau}_j^2 = \|X_j - X_{-j}\hat{\gamma}_j\|_n^2 + \lambda_{node,n}\|\hat{\Gamma}_j\hat{\gamma}_j\|_1, \tag{17}$$

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

$$\hat{\Theta} = \hat{T}^{-2}\hat{C}. \tag{18}$$

[1] 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, $\hat{\Theta}_j = \hat{C}_j/\hat{\tau}_j^2$. With this notation in place, note that

$$sgn(\hat{\gamma}_j)'\hat{\Gamma}_j\hat{\gamma}_j = \left\|\hat{\Gamma}_j\hat{\gamma}_j\right\|_1. \tag{19}$$

where $sgn(\hat{\gamma}_j) = \left(sgn(\hat{\gamma}_{j,k}),\ k = 1, ..., p, k \neq j\right)$. Therefore, postmultiplying the Karush-Kuhn-Tucker conditions (written as a row vector) of the problem (16) by $\hat{\gamma}_j$ and adding $(X_j - X_{-j}\hat{\gamma}_j)'X_j/n$ to both sides yields

$$\frac{(X_j - X_{-j}\hat{\gamma}_j)'(X_j - X_{-j}\hat{\gamma}_j)}{n} + \lambda_{node,n}\left\|\hat{\Gamma}_j\hat{\gamma}_j\right\|_1 = \frac{(X_j - X_{-j}\hat{\gamma}_j)'X_j}{n}. \tag{20}$$

Next, we recognize the left hand side of (20) as $\hat{\tau}_j^2$ such that

$$\hat{\tau}_j^2 = \frac{(X_j - X_{-j}\hat{\gamma}_j)'X_j}{n}. \tag{21}$$

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

$$1 = \frac{(X_j - X_{-j}\hat{\gamma}_j)'X_j}{\hat{\tau}_j^2 n} = \frac{(X\hat{\Theta}_j)'X_j}{n} = \frac{\hat{\Theta}_j'X'X_j}{n}, \tag{22}$$

[1] 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.

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 (16) 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,

$$\frac{\|X'_{-j} X \hat{\Theta}_j\|_\infty}{n} = \frac{\|\hat{\Gamma}_j \hat{\Gamma}_j^{-1} X'_{-j} X \hat{\Theta}_j\|_\infty}{n} \leq \|\hat{\Gamma}_j\|_{\ell_\infty} \frac{\|\hat{\Gamma}_j^{-1} X'_{-j} X \hat{\Theta}_j\|_\infty}{n} \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}, \tag{23}$$

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 by observation 2 after (4). Of course (23) is equivalent to

$$\frac{\left\|\hat{\Theta}'_j X' X_{-j}\right\|_\infty}{n} \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}. \tag{24}$$

In total, denoting by e_j the j'th $p \times 1$ unit vector, (22) and (24) yield

$$\|\hat{\Theta}'_j \hat{\Sigma} - e'_j\|_\infty \leq \frac{\lambda_{node,n}}{\hat{\tau}_j^2}. \tag{25}$$

Hence, the above display provides an upper bound on the j'th row of $\hat{\Theta}\hat{\Sigma} - I_p$ which, combined with the oracle inequality for $\|\hat{\beta} - \beta_0\|$, will yield an upper bound on Δ_j in (13) by arguments made rigorous in the appendix.

3.3 Properties of the Nodewise Regressions

In order to establish a central limit theorem for $\alpha'\hat{\Theta}X'u/n^{1/2}$ in (13) 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 Σ of 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 (15)). 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

$$\Theta_{j,-j} = -\Theta_{j,j}\gamma_j'. \tag{26}$$

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

$$X_{j,i} = X_{-j,i}\gamma_j + \eta_{j,i} \tag{27}$$

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 (27) is sparse if and only if Θ_j is sparse as can be seen from (26). 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 (27) 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 (16) is close to γ_j. We shall make this claim rigorous in Lemma 2 below. Next, by (27),

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

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 (26). In Lemma 2 below we will show that $\hat{\tau}_j^2$ as defined in (17) 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 now turn to the properties of the nodewise regressions which will be of great importance for the proof of Theorem 2 below. Defining $\bar{s} = \max_{j \in H} s_j$ we introduce the following assumption.

Assumption 2:

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

b) $\frac{p^2 \bar{s}^{r/2}}{n^{r/4}} \to 0$.

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

Assumption 2a) states that the smallest eigenvalue of the *population* covariance matrix is bounded away from zero. It is used to make sure that the τ_j^2 are bounded away from zero, as $\tau_j^2 = 1/\Theta_{j,j} \geq 1/\phi_{\max}(\Theta) = \phi_{\min}(\Sigma)$. 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 $\hat{\Sigma}$ 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 p can grow. On the other hand, we must always have that $\bar{s} = o(\sqrt{n})$ independently of the number of moments assumed to exist. Thus, the inverse covariance matrix must be sparse. This is satisfied if Σ is e.g. block diagonal. In the simulations we shall also see that our methods works well even if $\Theta = \Sigma^{-1}$ is not sparse as long as its entries are not too far from zero. Assumption 2c) is a moment assumption on the error terms from the nodewise regressions.

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

$$\max_{j\in H} \|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2 = O_p\left(\frac{\bar{s}h^{4/r}p^{4/r}}{n}\right). \tag{28}$$

$$\max_{j\in H} \|\hat{\gamma}_j - \gamma_j\|_1 = O_p\left(\frac{\bar{s}h^{2/r}p^{2/r}}{n^{1/2}}\right). \tag{29}$$

$$\max_{j\in H} |\hat{\tau}_j^2 - \tau_j^2| = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{30}$$

$$\max_{j\in H} \left\|\hat{\Theta}_j - \Theta_j\right\|_1 = O_p\left(\bar{s}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{31}$$

$$\max_{j\in H} \|\hat{\Theta}_j - \Theta_j\|_2 = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{32}$$

$$\max_{j\in H} \|\hat{\Theta}_j\|_1 = O_p(\bar{s}^{1/2}). \tag{33}$$

Lemma 2 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. (28) and (29) reduce to inequalities of the type (8) and (9) in Theorem 1 when H is a singleton such that $h = 1$. Note also that (31) can be used to bound the estimation error of each row of $\hat{\Theta}$ for the corresponding row of Θ. Thus, choosing $H = \{1, ..., p\}$, (31) 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 (13) as well as for estimating the variance of the limiting distribution.

Before stating Theorem 2 we introduce the following notation in connection with the asymptotic covariance matrix. Let $\Sigma_{xu} = n^{-1}\sum_{i=1}^n EX_iX_i'u_i^2$ and $\hat{\Sigma}_{xu} = n^{-1}\sum_{i=1}^n X_iX_i'\hat{u}_i^2$, where $\hat{u}_i = Y_i - X_i'\hat{\beta}$. For Theorem 2 we need the following assumptions.

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 must be of order $o(n^{1/2})$. Letting the number of parameters involved in the hypothesis increase with the sample size is a vast generalization of van de Geer et al. (2014) who only mention the possibility of H possessing a 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 $r > 16$. Furthermore, the maximal number of non-zero terms in the nodewise regression, $\bar{s}$, should not increase faster than the square root of the sample size. Assumptions 3a)-d) are trivially satisfied in the classical setting where p, h, s_0 and $\bar{s}$ are fixed. It is of course sensible, that these quantities can not grow too fast if one still wishes to obtain standard normal inference with precise estimation of an asymptotic covariance matrix of increasing dimension. Finally, Assumption 3e) restricts the eigenvalues of Σ and Σ_{xu}.

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

$$\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \xrightarrow{d} N(0,1), \tag{34}$$

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

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \left|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha\right| = o_p(1). \tag{35}$$

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 (34) 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 want to emphasize that the above results allows the error terms to be heteroskedastic and do not assume that they are independent of the covariates. (35) provides a uniformly consistent estimator of the asymptotic variance of $n^{1/2}\alpha'(\hat{b}-\beta_0)$. We are the first to do so in the literature on high-dimensional regressions models in the presence of heteroskedasticity and an increasing number of parameters. The uniformity of (35) will also be used in the proof of Theorem 3 below. (35) is also remarkable as it gives the limit of the variance in the denominator of (34) even as the dimension ($p \times p$) of the matrices involved in the expression increases. Consider the leading special case where

[2] 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.

$H = \{j\}$ such that α reduces to the j'th unit vector e_j of $\mathbb{R}^p$. If, furthermore, the covariates and the error terms are independent and the latter are homoskedastic with variance σ^2 we get that

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

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. In the case where H is a set of fixed cardinality h, (34) reveals that

$$\left\| \left(\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\right)_H^{-1/2} \sqrt{n}\left(\hat{b}_H - \beta_{0,H}\right) \right\|_2^2 \xrightarrow{d} \chi^2(h), \tag{36}$$

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.

4 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. Then, for all $t \in \mathbb{R}$ and $\alpha \in \mathbb{R}^p$ with $\|\alpha\|_2 = 1$,*

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \left| P\left(\frac{n^{1/2}\alpha'(\hat{b} - \beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t \right) - \Phi(t) \right| \to 0. \tag{37}$$

Furthermore, letting $\hat{\sigma}_j = \sqrt{e_j'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'e_j}$ (corresponding to $\alpha = e_j$ in (37)) and $z_{1-\delta/2}$ the $1-\delta/2$ percentile of the standard normal distribution, one has for all $j = 1, ..., p$

$$\liminf_{n\to\infty} \inf_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\beta_{0,j} \in \left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right] \right) \geq 1 - \delta. \tag{38}$$

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

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} \operatorname{diam}\left(\left[\hat{b}_j - z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}, \hat{b}_j + z_{1-\delta/2}\frac{\hat{\sigma}_j}{\sqrt{n}}\right] \right) = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{39}$$

(37) reveals that convergence to the standard normal distribution in Theorem 2 is actually valid uniformly over the ℓ_0-ball of radius at most s_0. Such uniformity is possible in the light of the work of Leeb and Pötscher (2005) since we refrain from model selection: the desparsified conservative Lasso is, as its name says, not sparse. Hence, our result does not contradict the work of these authors. (38) is a consequence of (37) 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). Asymptotic honesty is important to produce practically useful confidence sets as it ensures that there is a known time n, *not depending on* β_0, after which the coverage rate of the confidence set is not much smaller than $1-\delta$. Thus, pointwise confidence bands that are *dishonest*, i.e. which do not satisfy (38) but

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

are of much less practical use since the n from which point and onwards the coverage is close to $1-\delta$ is allowed to depend on the unknown β_0. Of course an honest confidence set S_n could also easily be produced by setting $S_n = \mathbb{R}$ for all $n \geq 1$. Such a confidence set is clearly of little practical use. Thus, (39) 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 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, (39) does not contradict inequality 6 in Theorem 2 of Pötscher (2009) who shows that honest confidence bands based on sparse estimators must be large. Results (38) and (39) are also remarkable in light of the classical result of Bahadur and Savage (1956) stating that even in the problem of constructing confidence intervals for the mean of a gaussian random variable honest confidence bands are not possible in general if one insists on the diameter of the confidence bands to be bounded almost surely. Finally, the above results are valid without any sort of $\beta_{\min}$-condition which requires the absolute value of the smallest non-zero coefficient to be greater than $s_0\lambda_n$.

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.

5 Monte Carlo

In this section we investigate the finite sample performance of the (desparsified) conservative Lasso and compare it to the one the (desparsified) Lasso of van de Geer et al. (2014). The Lasso as well as the conservative Lasso are implemented in R by means of the publicly available `glmnet` package and for both of these λ_n is chosen by BIC, see e.g. (9.4.9) in Davidson (2000). $\lambda_{node,n}$ is also chosen by BIC in the nodewise regressions. 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 discussed in subsection 2.3 λ_{prec} should be chosen of the order of the right hand side of (7) in order for the conservative Lasso to work well. However, this quantity is unknown and we thus choose it along with λ_n by means of BIC. We considered the grid $\{0.01, 0.05, 0.1, 0.5, 1\}$ for λ_{prec}. The motivation for the values in this grid is that ideally λ_{prec} should be bigger than $\hat{\beta}_{L,j}$ when $j \in S^c$ and smaller than $\hat{\beta}_{L,j}$ when $j \in S$. As $\hat{\beta}_{L,j}$ is often either zero or very close to zero for $j \in S_0^c$ (by consistency of the Lasso) it suffices to consider a grid of rather small values for λ_{prec} in order to drive a wedge between the zero and the non-zero coefficients. We also experimented with a wider and denser grid but this did not change the results.

All simulations are carried out with 1,000 replications 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 averaged over the Monte Carlo replications.

2. Size: We evaluate the size of the χ^2-test in (36) for a hypothesis involving more than one parameter.

3. Power: We evaluate the power of the χ^2-test in (36) for a hypothesis involving more than one parameter.

4. Coverage rate: We calculate coverage rate a gaussian confidence interval constructed as in (38). 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 also wish to 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 did only affect 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. The χ^2-tests always involve the two first parameters in β_0 of which we deliberately make sure that first one is 1 and the second one is zero. Thus, $h = 2$ in our simulations. 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[3]. Note that this u_i satisfies our assumption $E(u_i|X_i) = 0$ and has variance conditional on X_i given by $E(\epsilon_i^2)\left(\frac{1}{\sqrt{2}} X_{1,i} + b_x X_{2,i}\right)$. 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

[3] 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.

$\rho = 0, 0.5$ 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.
- Experiment 3b (very high-dimensional setting). As Experiment 3a but with heteroskedastic errors.

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
$\rho = 0$	Lasso	0.668	0.136	0.949	0.852	0.929	0.386	0.383
	CLasso	0.425	0.112	0.966	0.885	0.939	0.354	0.360
$\rho = 0.5$	Lasso	0.709	0.146	0.900	0.852	0.918	0.394	0.409
	CLasso	0.454	0.105	0.907	0.902	0.944	0.417	0.471
$\rho = 0.9$	Lasso	1.392	0.201	0.630	0.820	0.854	0.617	0.738
	CLasso	1.237	0.123	0.479	0.897	0.929	0.841	1.113

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. CLasso: Conservative Lasso.

Table 1 contains the results for Experiment 1a. First, we wish to stress that as predicted in Section 2.3, the conservative Lasso has a lower estimation error than the plain Lasso. This is the case no matter whether $\rho = 0, 0.5$ or 0.9. Furthermore, the conservative Lasso is always less size distorted than the Lasso while having slightly more power except for when $\rho = 0.9$. When $\rho = 0.9$ both procedures have serious power deficiencies. Next, our procedure always has a coverage rate which is closer to the nominal rate of 95%. This is the case for the zero as well as the non-zero parameters. When $\rho = 0$ one even has that the conservative Lasso has better coverage with narrower

bands. Note, however, that both 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 much less pronounced for the conservative Lasso. The reason for this is 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.

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
$\rho = 0$	Lasso	0.738	0.158	0.765	0.854	0.898	0.557	0.563
	CLasso	0.499	0.155	0.793	0.871	0.912	0.536	0.547
$\rho = 0.5$	Lasso	0.780	0.193	0.774	0.828	0.913	0.609	0.534
	CLasso	0.528	0.153	0.782	0.856	0.934	0.617	0.564
$\rho = 0.9$	Lasso	1.484	0.218	0.524	0.792	0.867	0.789	0.835
	CLasso	1.378	0.144	0.420	0.868	0.937	0.990	1.203

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. CLasso: Conservative Lasso.

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 both 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 $\rho = 0$, the estimation error of the conservative Lasso is almost twice as low as the one for the plain Lasso underscoring the prediction in Section 2.3. However, in this case this does not result in a better performance along the other dimensions measured as the two procedures perform similarly there: they both have good size and power properties and the coverage rate is close to the nominal one.

For $\rho = 0.5$ the conservative Lasso is still more precise than the Lasso but now it is also considerably less size distorted than the Lasso and has a power which is around 20%-points higher than the one of the Lasso. This is a considerable improvement which can also be found in the

			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
	CLasso	0.220	0.077	0.926	0.925	0.937	0.383	0.389
$\rho = 0.5$	Lasso	0.337	0.162	0.687	0.928	0.823	0.439	0.436
	CLasso	0.214	0.074	0.867	0.925	0.929	0.445	0.499
$\rho = 0.9$	Lasso	0.451	0.237	0.407	0.841	0.796	0.642	0.748
	CLasso	0.392	0.101	0.450	0.912	0.907	0.843	1.080

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. CLasso: Conservative Lasso.

coverage probability for the zero parameter. When $\rho = 0.9$ the conservative Lasso again remains the most precise estimator with superior size and power properties. However, as in Experiment 1, for none of the procedures the χ^2-test has good power properties. The conservative Lasso has much better coverage rate, being up to ten percentage points larger for the zero parameter. This comes from more precise parameter estimates and wider bands.

When adding heteroskedasticity to Experiment 2a, Table 4 shows that the estimation errors of both procedures increase slightly. The conservative Lasso remains the most precise one. Both procedures produce confidence bands having around the same coverage probability as in the homoskedastic case. In fact, the coverage rates are better in the heteroskedastic setting when $\rho = 0.9$ for both procedure as the bands become quite a bit wider.

The results for the very high-dimensional Experiment 3a are found in Table 5. When the sample size is $n = 100$, the plain Lasso has an estimation error which is 50% larger than the one for the conservative Lasso. Furthermore, the χ^2-test based on the Lasso is so size distorted (the size is 70%) that its usefulness may be questioned. While the conservative Lasso also suffers from size distortion (the size is 32%) it is still *much* more reliable than the Lasso. In terms of power, the χ^2-tests based on the two procedures perform similarly.

Turning to the coverage rates of the confidence intervals of the non-zero coefficients, the Lasso provides such a poor coverage (30%) that it may almost be deemed useless. The conservative Lasso,

			χ^2		Coverage rate		Length	
$n = 100$		ℓ_2	Size	Power	non-zero	zero	non-zero	zero
$\rho = 0$	Lasso	0.445	0.082	0.714	0.923	0.945	0.631	0.634
	CLasso	0.283	0.095	0.744	0.911	0.943	0.583	0.609
$\rho = 0.5$	Lasso	0.391	0.184	0.545	0.918	0.875	0.698	0.587
	CLasso	0.284	0.092	0.686	0.914	0.945	0.670	0.602
$\rho = 0.9$	Lasso	0.512	0.220	0.315	0.879	0.804	0.870	0.862
	CLasso	0.482	0.097	0.354	0.913	0.930	1.028	1.163

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. CLasso: Conservative Lasso.

while not being perfect, still has produced a coverage of 70%. It also performs much better for the truly zero parameter than the Lasso. The superior coverage of conservative Lasso is again due to much more precise estimates and wider confidence bands than the Lasso.

When the sample size is increased to just $n = 150$ the conservative Lasso delivers more than twice as precise parameter estimates as the plain Lasso. Actually, we conclude that it 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 93% (from 70%). The Lasso has also improved. However, it is remarkable that the size of its χ^2-test for $n = 150$ still only corresponds to the one for the conservative Lasso when $n = 100$. Similarly, the coverage rate of the confidence bands for zero as well as non-zero parameters based on the Lasso has only now risen to the coverage rate that the conservative Lasso produced for $n = 100$.

It is also remarkable that for both procedures the length of the confidence bands has actually become wider as n is increased from 100 to 150. This indicates that the undercoverage for $n = 100$ is to a high extent due to too narrow confidence bands as a result of under estimating the variance of the parameters.

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,

$\rho = 0.75$		ℓ_2	χ^2 Size	χ^2 Power	Coverage rate non-zero	Coverage rate zero	Length non-zero	Length zero
$n = 100$	Lasso	1.524	0.700	0.899	0.297	0.813	0.253	0.254
	CLasso	0.960	0.317	0.871	0.695	0.909	0.429	0.504
$n = 150$	Lasso	1.090	0.316	0.794	0.696	0.847	0.360	0.382
	CLasso	0.391	0.081	0.891	0.931	0.940	0.445	0.546
$n = 200$	Lasso	0.868	0.099	0.860	0.902	0.910	0.391	0.428
	CLasso	0.280	0.067	0.942	0.939	0.942	0.399	0.493
$n = 500$	Lasso	0.497	0.080	1.000	0.929	0.919	0.246	0.281
	CLasso	0.150	0.057	1.000	0.936	0.950	0.260	0.320

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. CLasso: Conservative Lasso.

the conservative Lasso also remains the superior procedure but now the Lasso now has coverage of above 90% for both parameters as well.

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).

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.

6 Conclusion

This paper shows how the conservative Lasso can be used to conduct inference in the high-dimensional linear regression model. We are the first in the literature to allow for conditional heteroskedasticity in the error terms and we also show how to consistently estimate the limiting

$\rho = 0.75$		ℓ_2	χ^2 Size	χ^2 Power	Coverage rate non-zero	Coverage rate zero	Length non-zero	Length zero
$n = 100$	Lasso	1.663	0.721	0.918	0.277	0.811	0.290	0.271
	CLasso	1.171	0.400	0.832	0.611	0.889	0.504	0.526
$n = 150$	Lasso	1.225	0.368	0.719	0.646	0.861	0.459	0.421
	CLasso	0.534	0.111	0.765	0.889	0.939	0.588	0.588
$n = 200$	Lasso	0.964	0.125	0.663	0.892	0.921	0.559	0.518
	CLasso	0.357	0.082	0.791	0.924	0.957	0.563	0.561
$n = 500$	Lasso	0.558	0.060	0.967	0.933	0.941	0.379	0.342
	CLasso	0.186	0.060	0.971	0.945	0.950	0.387	0.371

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. CLasso: Conservative Lasso.

high-dimensional covariance matrix. 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.

A Appendix

A.1 Appendix A – auxiliary lemmas

We begin by providing some auxiliary lemmas used for the proofs of the main results in Appendix B. First, 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 $\mathcal{A} = \left\{ \|X'u/n\|_\infty \leq \lambda_n/2 \right\}$ and $\mathcal{B} = \left\{ \phi_\Sigma^2 \geq \phi_\Sigma^2/2 \right\}$. 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.

Lemma A.1. *Assume that* $\|\hat{w}_{g,S_0}\|_2 \leq \sqrt{s_0}$ *and* $\hat{w}_{g,S_0^c}^{min} = \min_{j \in S_0^c} \hat{w}_j = 1$. *Then, on the set* $\mathcal{A} \cap \mathcal{B}$ *the following inequalities are valid.*

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 \leq 18 \frac{\lambda_n^2 s_0}{\phi_\Sigma^2(s_0)}. \tag{A.1}$$

$$\|\hat{\beta}_w - \beta_0\|_1 \leq 24 \frac{\lambda_n s_0}{\phi_\Sigma^2(s_0)}. \tag{A.2}$$

Proof. We begin by establishing (A.1). By the minimizing property of $\hat{\beta}_w$ it follows that

$$\|Y - X\hat{\beta}_w\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\hat{\beta}_{w,j}| \leq \|Y - X\beta_0\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\beta_{0,j}|. \tag{A.3}$$

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

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\hat{\beta}_{w,j}| \leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\beta_{0,j}|. \tag{A.4}$$

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

$$\begin{aligned} \|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \sum_{j \in S_0^c} \hat{w}_{g,j} |\hat{\beta}_{w,j}| &\leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 - 2\lambda_n \sum_{j \in S_0} \hat{w}_{g,j} |\hat{\beta}_{w,j}| + 2\lambda_n \sum_{j=1}^p \hat{w}_{g,j} |\beta_{0,j}| \\ &\leq \lambda_n \|\hat{\beta}_w - \beta_0\|_1 + 2\lambda_n \sum_{j \in S_0} \hat{w}_{g,j} |\hat{\beta}_{w,j} - \beta_{0,j}|. \end{aligned} \tag{A.5}$$

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.5) as

$$\|X(\hat{\beta}_w - \beta_0)\|_n^2 + 2\lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 \leq \lambda_n \|\hat{\beta}_{w,S_0} - \beta_{0,S_0}\|_1 + \lambda_n \|\hat{\beta}_{w,S_0^c}\|_1 + 2\lambda_n \sum_{j \in S_0} \hat{w}_{g,j} |\hat{\beta}_{w,j} - \beta_{0,j}|. \tag{A.6}$$

Subtract $\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1$ from both sides of (A.6) to get

$$\|X(\hat{\beta}_w-\beta_0)\|_n^2+\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1\leq\lambda_n\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_1+2\lambda_n\sum_{j\in S_0}\hat{w}_{g,j}|\hat{\beta}_{w,j}-\beta_{0,j}|. \tag{A.7}$$

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

$$\begin{aligned}\|X(\hat{\beta}_w-\beta_0)\|_n^2+\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1&\leq\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2+2\lambda_n\|\hat{w}_{g,S_0}\|_2\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2\\&=\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2+2\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2\\&=3\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2. \end{aligned}\tag{A.8}$$

(A.8) 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.8) implies

$$\|X(\hat{\beta}_w-\beta_0)\|_n^2+\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1\leq 3\lambda_n\sqrt{s_0}\frac{\|X(\hat{\beta}_w-\beta_0)\|_n}{\phi_{\hat{\Sigma}}(s_0)}. \tag{A.9}$$

Then, using $3uv\leq u^2/2+(9/2)v^2$, with $v=\lambda_n\sqrt{s_0}/\phi_{\hat{\Sigma}}(s_0)$, $u=\|X(\hat{\beta}_w-\beta_0)\|_n$, one gets

$$\|X(\hat{\beta}_w-\beta_0)\|_n^2+\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1\leq\frac{\|X(\hat{\beta}_w-\beta_0)\|_n^2}{2}+\frac{9}{2}\frac{\lambda_n^2 s_0}{\phi_{\hat{\Sigma}}^2(s_0)}. \tag{A.10}$$

Subtracting the first right hand side term in (A.10) from the left and right hand sides of (A.10) and multiplying all terms by 2 yields

$$\|X(\hat{\beta}_w-\beta_0)\|_n^2+2\lambda_n\|\hat{\beta}_{w,S_0^c}\|_1\leq 9\frac{\lambda_n^2 s_0}{\phi_{\hat{\Sigma}}^2(s_0)}, \tag{A.11}$$

which, using that we are on $\mathcal{B}$, implies (A.1).

Next, we turn to proving (A.2). By adding $\lambda_n\left\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\right\|_1$ to both sides of (A.8) and using $\|\hat{\beta}_{w,S_0^c}\|_1+\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_1=\|\hat{\beta}_w-\beta_0\|_1$ one gets

$$\lambda_n\|\hat{\beta}_w-\beta_0\|_1\leq\lambda_n\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_1+3\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2 \tag{A.12}$$

$$\leq 4\lambda_n\sqrt{s_0}\|\hat{\beta}_{w,S_0}-\beta_{0,S_0}\|_2. \tag{A.13}$$

The adaptive restricted eigenvalue condition and inequality (A.1) of this Lemma yield

$$\lambda_n\|\hat{\beta}_w-\beta_0\|_1\leq 4\lambda_n\sqrt{s_0}\frac{\|X(\hat{\beta}_w-\beta_0)\|_n}{\phi_{\hat{\Sigma}}(s_0)}=12\frac{s_0\lambda_n^2}{\phi_{\hat{\Sigma}}^2(s_0)}, \tag{A.14}$$

which, using that we are on $\mathcal{B}$, implies (A.2). □

To prove Lemma 1 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, $r\geq 2$*

$$a_r E\left(\sum_{i=1}^n U_i^2\right)^{r/2}\leq E\left|\sum_{i=1}^n U_i\right|^r\leq b_r E\left(\sum_{i=1}^n U_i^2\right)^{r/2} \tag{A.15}$$

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.15) 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 $\{U_{j,i}\}_{i=1}^n$ 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,*

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

Remark: 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 $\big(8\log(2m)\big)^{r/2}$ 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

$$E\left(\max_{1\leq j\leq m} |S_{j,n}|^r\right) \leq \big(8\log(2m)\big)^{r/2} E\left[\max_{1\leq j\leq m}\sum_{i=1}^{n} U_{j,i}^2\right]^{r/2}. \tag{A.16}$$

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.16) and using Markov's inequality yields

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

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

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 $1 - \frac{C}{M^{r/2}}$.*

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 $\delta = \|\hat{\Sigma} - \Sigma\|_\infty \leq \frac{\phi_\Sigma^2(s_0)}{32s_0}$. 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

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

□

A.2 Appendix B

This appendix provides the proofs of the main theorems.

Proof of Lemma 1. 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 (6) and (7). The uniformity over $\mathcal{B}_{\ell_0}(s_0)$ follows by noting that the right hand sides of (6) and (7) only depend on β_0 through s_0.

□

Proof of Theorem 1. Choose $\lambda_{prec} = 24\frac{\lambda_n s_0}{\phi^2_\Sigma(s_0)}$. Then, by the observations in section 2.3 and Lemma 1 we get $\hat{w}_j \leq 1$ for all $j \in S_0$ while $\hat{w}_j = 1$ for all $j \in S_0^c$ on $\mathcal{A} \cap \mathcal{B}$. The first observation clearly implies that $\|\hat{w}_{S_0}\|_2 \leq \sqrt{s_0}$ while the latter implies that $\hat{w}^{min}_{S^c} = \min_{j \in S_0^c} \hat{w}_j = 1$. Thus Lemma A.1 applies. Combine this with the lower bounds on the probabilities of the sets $\mathcal{A}$ and $\mathcal{B}$ from Lemmas A.4 and A.6, respectively to obtain (8) and (9). The uniformity over $\mathcal{B}_{\ell_0}(s_0)$ follows by noting that the right hand sides of (8) and (9) only depend in β_0 through s_0. □

Proof of Lemma 2. 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

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

By an exact adaptation of the proof of Lemma A.1 it can be shown for each $j \in H$ that

$$\|X_{-j}(\hat{\gamma}_j - \gamma_j)\|_n^2 \leq 18\frac{\lambda^2_{node,n} s_j}{\phi^2_\Sigma(s_j)}, \tag{A.17}$$

$$\|\hat{\gamma}_j - \gamma_j\|_1 \leq 24\frac{\lambda_{node,n} s_j}{\phi^2_\Sigma(s_j)} \tag{A.18}$$

are valid on the set $\mathcal{A}_{node} \cap \mathcal{B}_j$ for $j \in H$. Hence, these inequalities are valid simultaneously for all $j \in H$ on $\mathcal{A}_{node} \cap (\cap_{j \in H} \mathcal{B}_j)$ [4]. 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

$$\max_{j \in H} \|X'_{-j}\eta_j/n\|_\infty = O_p\left(\lambda_{node,n}\right) = O_p\left(\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right) \tag{A.19}$$

by choosing M sufficiently large.

[4] It will turn out later that it is quite important that (A.17) and (A.18) 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.

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 $\left\{\|\hat{\Sigma}_{-j}-\Sigma_{-j}\|_\infty \leq \frac{\phi^2_{\Sigma_{-j}}(s_j)}{32s_j}\right\} \subseteq \left\{\phi^2_{\hat{\Sigma}_{-j}}(s_j) \geq \phi^2_{\Sigma_{-j}}(s_j)/2\right\} = \mathcal{B}_j$. Thus, the relation

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

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 $1-\frac{C}{M^{r/2}}-D\frac{p^2\bar{s}^{r/2}}{n^{r/4}}$

$$\|X_{-j}(\hat{\gamma}_j-\gamma_j)\|_n^2 \leq 18\frac{\lambda^2_{node,n}s_j}{\phi^2_\Sigma(s_j)}. \tag{A.20}$$

$$\|\hat{\gamma}_j-\gamma_j\|_1 \leq 24\frac{\lambda_{node,n}s_j}{\phi^2_\Sigma(s_j)}. \tag{A.21}$$

By choosing M sufficiently large, using $\frac{p^2\bar{s}^{r/2}}{n^{r/4}} \to 0$, and inserting the definition of $\lambda_{node,n}$ (28) and (29) follow upon taking the maximum in the above display and utilizing that the above inequalities are all valid simultaneously on $\mathcal{A}_{node,n} \cap \left(\cap_{j\in H}\mathcal{B}_j\right)$.

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

$$\max_{j\in H}(\hat{v}_j-v_j)'\Sigma(\hat{v}_j-v_j) \leq \max_{j\in H}(\hat{v}_j-v_j)'\hat{\Sigma}(\hat{v}_j-v_j)+\max_{j\in H}\|\hat{\Sigma}-\Sigma\|_\infty\|\hat{v}_j-v_j\|_1^2. \tag{A.22}$$

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

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

by (28). Next, consider the second term in (A.22). 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

$$\|\hat{\Sigma}-\Sigma\|_\infty = O_p\left(\frac{p^{4/r}}{n^{1/2}}\right). \tag{A.23}$$

In combination with (29) this implies (using $\|\hat{\gamma}_j - \gamma_j\|_1 = \|\hat{v}_j - v_j\|_1$)

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

But since

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

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

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

Therefore, by

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

one gets

$$\max_{j\in H} \|\hat{\gamma}_j - \gamma_j\|_2^2 = \max_{j\in H} \|\hat{v}_j - v_j\|_2^2 = O_p\left(\frac{\bar{s} h^{4/r} p^{4/r}}{n}\right). \tag{A.24}$$

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

Next, we consider $|\hat{\tau}_j^2 - \tau_j^2|$. First, by (21) 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

$$\begin{aligned}
\max_{j\in H} |\hat{\tau}_j^2 - \tau_j^2| \leq \max_{j\in H} \left|\frac{\eta_j'\eta_j}{n} - \tau_j^2\right| + \max_{j\in H} |\eta_j' X_{-j}(\hat{\gamma}_j - \gamma_j)/n| \\
+ \max_{j\in H} |\eta_j' X_{-j}\gamma_j/n| + \max_{j\in H} \left|\frac{\gamma_j' X_{-j}' X_{-j}(\hat{\gamma}_j - \gamma_j)}{n}\right|.
\end{aligned} \tag{A.25}$$

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

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

which implies that

$$\max_{j\in H}\left|\frac{\eta_j'\eta_j}{n} - \tau_j^2\right| = O_p\left(\frac{h^{2/r}}{n^{1/2}}\right). \tag{A.26}$$

Next, consider the second term in (A.25). By (29) and (A.19) it follows that

$$\begin{aligned}\max_{j\in H}|\eta_j'X_{-j}(\hat{\gamma}_j-\gamma_j)/n| &\leq \max_{j\in H}\|\eta_j'X_{-j}/n\|_\infty \max_{j\in H}\|\hat{\gamma}_j-\gamma_j\|_1 \\ &= O_p\left(\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)O_p\left(\frac{\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right) \\ &= O_p\left(\left[\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right]^2\right). \end{aligned} \tag{A.27}$$

Before we bound the third term in (A.25) 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 Σ_{-j} consisting of all rows and columns of Σ except the j'th row and column. Then, note that

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

such that

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

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 $E(X_{j,i}^2) = \gamma_j'\Sigma_{-j}\gamma_j + E(\eta_{j,i}^2)$ such that $\gamma_j'\Sigma_{-j}\gamma_j \leq E(X_{j,i}^2) \leq \max_{j\in H}E(X_{j,i}^2)$. 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,

$$\gamma_j'\gamma_j \leq \frac{C^{2/r}}{\phi_{\min}(\Sigma)}. \tag{A.28}$$

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.25),

$$\max_{j\in H}|\eta_j'X_{-j}\gamma_j/n| \leq \max_{j\in H}\|\eta_j'X_{-j}/n\|_\infty \max_{j\in H}\|\gamma_j\|_1 = O_p\left(\sqrt{\bar{s}}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right), \tag{A.29}$$

where we have also used (A.19). It remains to bound the fourth summand in (A.25). 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.19),

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

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

$$\max_{j\in H}\left|\gamma'_j\frac{X'_{-j}X_{-j}}{n}(\hat{\gamma}_j - \gamma_j)\right| = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.30}$$

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.25) is $O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)$. 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

$$\max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2} - \frac{1}{\tau_j^2}\right| = \max_{j\in H}\frac{|\tau_j^2 - \hat{\tau}_j^2|}{\hat{\tau}_j^2\tau_j^2} = O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right). \tag{A.31}$$

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

(29), and (A.31)

$$\max_{j\in H}\|\hat{\Theta}_j-\Theta_j\|_1=\max_{j\in H}\left\|\frac{\hat{C}_j}{\hat{\tau}_j^2}-\frac{C_j}{\tau_j^2}\right\|_1 \tag{A.32}$$

$$\begin{aligned}
&\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|+\max_{j\in H}\left\|\frac{\hat{\gamma}_j}{\hat{\tau}_j^2}-\frac{\gamma_j}{\tau_j^2}\right\|_1\\
&=\max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|+\max_{j\in H}\left\|\frac{\hat{\gamma}_j}{\hat{\tau}_j^2}-\frac{\gamma_j}{\hat{\tau}_j^2}+\frac{\gamma_j}{\hat{\tau}_j^2}-\frac{\gamma_j}{\tau_j^2}\right\|_1\\
&\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|+\max_{j\in H}\frac{\|\hat{\gamma}_j-\gamma_j\|_1}{\hat{\tau}_j^2}+\max_{j\in H}\|\gamma_j\|_1\max_{j\in H}\left(\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|\right)\\
&=O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)+O_p\left(\frac{\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right)+O_p\left(\bar{s}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)\\
&=O_p\left(\frac{\bar{s}h^{2/r}p^{2/r}}{\sqrt{n}}\right).
\end{aligned} \tag{A.33}$$

Next, for later purposes, we also bound $\|\hat{\Theta}_j-\Theta_j\|_2$. By (A.24), and $\max_{j\in H}\|\gamma_j\|_2^2=O(1)$ by (A.28)

$$\begin{aligned}
\max_{j\in H}\|\hat{\Theta}_j-\Theta_j\|_2 &\leq \max_{j\in H}\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|+\max_{j\in H}\frac{\|\hat{\gamma}_j-\gamma_j\|_2}{\hat{\tau}_j^2}+\max_{j\in H}\|\gamma_j\|_2\max_{j\in H}\left(\left|\frac{1}{\hat{\tau}_j^2}-\frac{1}{\tau_j^2}\right|\right)\\
&=O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right)+O_p\left(\frac{\bar{s}^{1/2}h^{2/r}p^{2/r}}{n^{1/2}}\right)+O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right),\\
&=O_p\left(\bar{s}^{1/2}\frac{h^{2/r}p^{2/r}}{\sqrt{n}}\right).
\end{aligned} \tag{A.34}$$

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

$$\max_{j\in H}\|\Theta_j\|_1\leq\max_{j\in H}\frac{1}{\tau_j^2}+\max_{j\in H}\|\gamma_j/\tau_j^2\|_1=O(\bar{s}^{1/2}) \tag{A.35}$$

(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

$$\max_{j\in H}\|\hat{\Theta}_j\|_1\leq\max_{j\in H}\|\hat{\Theta}_j-\Theta_j\|_1+\max_{j\in H}\|\Theta_j\|_1=O_p\left(\frac{\bar{s}h^{2/r}p^{2/r}}{n^{1/2}}\right)+O(\sqrt{\bar{s}})=O_p(\sqrt{\bar{s}}). \tag{A.36}$$

□

Proof of Theorem 2. We show that the ratio

$$t=\frac{n^{1/2}\alpha'(\hat{b}-\beta_0)}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}}, \tag{A.37}$$

is asymptotically standard normal. First, note that one can write. By (13)

$$t=t_1+t_2,$$

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

$$E\left[\frac{\alpha'\Theta X'u/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right] = E\left[\frac{\alpha'\Theta\sum_{i=1}^n X_iu_i/n^{1/2}}{\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}}\right] = 0, \tag{A.38}$$

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

$$\frac{1}{\left(\alpha'\Theta\Sigma_{xu}\Theta'\alpha\right)^{r/4}}\sum_{i=1}^n E\left|\alpha'\Theta X_iu_i/n^{1/2}\right|^{r/2} \to 0. \tag{A.39}$$

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

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

since $\|\alpha\|_2 = 1$ and $\max_{j\in H}\|\Theta_j\|_1 = O(\sqrt{\bar{s}})$ by (A.35). 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,

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

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,

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

by Assumption 3d). Next, we show that $\alpha'\Theta\Sigma_{xu}\Theta'\alpha$ is asymptotically bounded away from zero in (A.39). Clearly,

$$\alpha'\Theta\Sigma_{xu}\Theta'\alpha \geq \phi_{\min}(\Sigma_{xu})\left\|\Theta'\alpha\right\|_2^2 \geq \phi_{\min}(\Sigma_{xu})\phi_{\min}^2(\Theta)\left\|\alpha\right\|_2^2 \geq \phi_{\min}(\Sigma_{xu})\frac{1}{\phi_{\max}^2(\Sigma)}, \tag{A.40}$$

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.40). We first show that the denominators of t_1' and t_1 are asymptotically equivalent, i.e.

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| = o_p(1). \tag{A.41}$$

Set $\tilde{\Sigma}_{xu} = n^{-1}\sum_{i=1}^{n} X_i X_i' u_i^2$. To establish (A.41) it suffices to show the following relations:

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha| = o_p(1). \tag{A.42}$$

$$|\alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha| = o_p(1). \tag{A.43}$$

$$|\alpha'\hat{\Theta}\Sigma_{xu}\hat{\Theta}'\alpha - \alpha'\Theta\Sigma_{xu}\Theta'\alpha| = o_p(1). \tag{A.44}$$

We first prove (A.42).

$$|\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha - \alpha'\hat{\Theta}\tilde{\Sigma}_{xu}\hat{\Theta}'\alpha| \leq \|\hat{\Sigma}_{xu} - \tilde{\Sigma}_{xu}\|_\infty \|\hat{\Theta}'\alpha\|_1^2. \tag{A.45}$$

But by (33) and $\|\alpha\|_2 = 1$

$$\left\|\hat{\Theta}'\alpha\right\|_1 = \left\|\sum_{j\in H}\hat{\Theta}_j\alpha_j\right\|_1 \leq \sum_{j\in H}|\alpha_j|\left\|\hat{\Theta}_j\right\|_1 = O_p\left(\sqrt{h\bar{s}}\right). \tag{A.46}$$

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

$$\hat{\Sigma}_{xu} - \tilde{\Sigma}_{xu} = -\frac{2}{n}\sum_{i=1}^{n} X_i X_i' u_i X_i'(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} X_i X_i'(\hat{\beta} - \beta_0)' X_i X_i'(\hat{\beta} - \beta_0). \tag{A.47}$$

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

$$\max_{1\leq k,l\leq p}\left|\frac{2}{n}\sum_{i=1}^{n} X_{k,i}X_{l,i}u_i X_i'(\hat{\beta} - \beta_0)\right| \leq 2\sqrt{\max_{1\leq k,l\leq p}\frac{1}{n}\sum_{i=1}^{n} X_{k,i}^2 X_{l,i}^2 u_i^2}\cdot\left\|X(\hat{\beta} - \beta_0)\right\|_n. \tag{A.48}$$

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

$$\begin{aligned} E|Z_1^2 Z_2^2 Z_3^2|^{r/6} = E|Z_1^{r/3} Z_2^{r/3} Z_3^{r/3}| &\leq E\left(|Z_1|^{r/2}|Z_2|^{r/2}\right)^{2/3} E\left(|Z_3^r|\right)^{1/3} \\ &\leq E\left(|Z_1^r|\right)^{1/3} E\left(|Z_2^r|\right)^{1/3} E\left(|Z_3^r|\right)^{1/3}. \end{aligned} \tag{A.49}$$

Thus, by Assumption 1, all summands in (A.48) have uniformly bounded $r/6$ moments and therefore Lemma A.3 implies that

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

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

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

Furthermore, since the L^r-norm is non-decreasing in r and since $r\ge 6$ we have, using (A.49) above,

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

which is uniformly bounded by Assumption 1 since the r'th moments of $X_{k,i}$ and u_i are uniformly bounded. Therefore, $\sqrt{\max_{1\le k,l\le p}\frac{1}{n}\sum_{i=1}^{n}X_{k,i}^2X_{l,i}^2u_i^2}=O(1)+O_p\left(\frac{p^{6/r}}{n^{1/4}}\right)$ in (A.48). By Theorem 1 it follows from choosing M sufficiently large

$$\left\|X(\hat\beta-\beta_0)\right\|_n=O_p\left(\frac{p^{2/r}\sqrt{s_0}}{n^{1/2}}\right).\tag{A.50}$$

Thus,

$$\max_{1\le k,l\le p}\left|\frac{2}{n}\sum_{i=1}^{n}X_{k,i}X_{l,i}u_iX_i'(\hat\beta-\beta_0)\right|=O_p\left(\frac{p^{8/r}\sqrt{s_0}}{n^{3/4}}\right)+O_p\left(\frac{p^{2/r}\sqrt{s_0}}{n^{1/2}}\right)\tag{A.51}$$

Regarding the second term in (A.47) note that

$$\max_{1\le k,l\le p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{k,i}X_{l,i}(\hat\beta-\beta_0)'X_iX_i'(\hat\beta-\beta_0)\right|\le\max_{1\le k,l\le p}\max_{1\le i\le n}\left|X_{k,i}X_{l,i}\right|\frac{1}{n}\sum_{i=1}^{n}\left(X_i'(\hat\beta-\beta_0)\right)^2.\tag{A.52}$$

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

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

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

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

Next, note that by Theorem 1

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i'(\hat\beta-\beta_0)\right)^2=\left\|X(\hat\beta-\beta_0)\right\|_n^2=O_p\left(\frac{p^{4/r}s_0}{n}\right),\tag{A.53}$$

such that, using (A.52),

$$\max_{1\le k,l\le p}\left|\frac{1}{n}\sum_{i=1}^{n}X_{k,i}X_{l,i}(\hat\beta-\beta_0)'X_iX_i'(\hat\beta-\beta_0)\right| = O_p\left(p^{4/r}n^{2/r}\right)O_p\left(\frac{p^{4/r}s_0}{n}\right)$$
$$= O_p\left(\frac{p^{8/r}s_0}{n^{(r-2)/r}}\right). \tag{A.54}$$

Then, combining (A.51) and (A.54) implies that

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

Therefore, combining with (A.46) yields

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

by Assumption 3c). This establishes (A.42).

Next, we turn to (A.43). First, note that

$$|\alpha'\hat\Theta\tilde\Sigma_{xu}\hat\Theta'\alpha-\alpha\hat\Theta\Sigma_{xu}\hat\Theta'\alpha|\le\|\tilde\Sigma_{xu}-\Sigma_{xu}\|_\infty\|\hat\Theta'\alpha\|_1^2. \tag{A.55}$$

Furthermore, similarly to (A.49), 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 $t>0$

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

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.55) and (A.46)

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

and Assumption 3b).

Finally, we establish (A.44) to conclude (A.41). By Lemma 6.1 in van de Geer et al. (2014)

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

Note that

$$\begin{aligned}\|(\hat\Theta'-\Theta')\alpha\|_1&=\left\|\sum_{j\in H}\left(\hat\Theta_j-\Theta_j\right)\alpha_j\right\|_1\le\sum_{j\in H}\left\|\hat\Theta_j-\Theta_j\right\|_1|\alpha_j|\le\max_{j\in H}\left\|\hat\Theta_j-\Theta_j\right\|_1\sum_{j\in H}|\alpha_j|\\&=O_p\left(\bar s\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right),\end{aligned} \tag{A.56}$$

by (31) 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,

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

by (32) and $\|\alpha\|_2 = 1$. Therefore, by $\|\Sigma_{xu}\|_\infty \leq \phi_{\max}(\Sigma_{xu})$ with the latter assumed bounded from Assumption 3e),

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

where we used

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

and Assumption 3b (which also implies $\bar{s} = o(n^{1/2})$). The uniformity of (A.41) over $\mathcal{B}_{\ell_0}(s_0)$ follows from simply observing that (A.50) and (A.53) 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.56) above we get, using $h \leq p$, and Assumption 3b

$$\begin{aligned}n^{1/2}|\alpha'\hat{\Theta}X'u/n - \alpha'\Theta X'u/n| &\leq n^{1/2}\left\|\frac{X'u}{n}\right\|_\infty \|\alpha'(\hat{\Theta} - \Theta)\|_1 \\ &= n^{1/2}O_p\left(\frac{p^{2/r}}{\sqrt{n}}\right)O\left(\bar{s}\frac{h^{2/r+1/2}p^{2/r}}{\sqrt{n}}\right) \\ &= O_p\left(\bar{s}\frac{h^{2/r+1/2}p^{4/r}}{\sqrt{n}}\right) \\ &= O_p\left(\bar{s}\frac{h^{1/2}p^{6/r}}{\sqrt{n}}\right) \\ &= o_p(1). \end{aligned} \tag{A.57}$$

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.40) and (A.41). Thus, it suffices to show that the numerator of t_2 vanishes in probability. Note that, by the definition of Δ, and $\|\alpha\|_2 = 1$,

$$|\alpha'\Delta| \leq \max_{j\in H}|\Delta_j|\sum_{j\in H}|\alpha_j| \leq \max_{j\in H}\left|\left(\hat{\Theta}_j'\hat{\Sigma} - e_j\right)\left(\sqrt{n}(\hat{\beta} - \beta_0)\right)\right|\sum_{j\in H}|\alpha_j| \tag{A.58}$$

$$\leq \max_{j\in H}\left\|\left(\hat{\Theta}_j'\hat{\Sigma} - e_j\right)\right\|_\infty \left\|\sqrt{n}(\hat{\beta} - \beta_0)\right\|_1 O\left(\sqrt{h}\right). \tag{A.59}$$

First, it follows from Theorem 1 that $n^{1/2}\|\hat{\beta}-\beta_0\|_1 = O_p\left(s_0 p^{2/r}\right)$. 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.31) and Assumption 3b). Thus, in total we have

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

by Assumption 3a). 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 $\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)} n^{1/2}\|\hat{\beta}-\beta_0\|_1 = O_p\left(s_0 p^{2/r}\right)$ in the above argument and that this is the only place in which β_0 enters these arguments. Thus, for later reference,

$$\sup_{\beta_0\in\mathcal{B}_{\ell_0}(s_0)}\left|\alpha'\Delta\right| = o_p(1). \tag{A.60}$$

□

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.60), (35), (A.57), and $\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}$ being bounded away from zero (by (A.40)) 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.40) 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

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

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,

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

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

$$\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t,\ A_{1,n}, A_{2,n}, A_{3,n}\right) \leq \Phi(t) + \delta + \epsilon. \tag{A.61}$$

Next, using that $\sqrt{\alpha'\Theta\Sigma_{xu}\Theta'\alpha}$ does not depend on β_0 and is bounded away from zero by (A.40) 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

$$\inf_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\frac{\alpha'\hat{\Theta}X'u/n^{1/2}}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} - \frac{\alpha'\Delta}{\sqrt{\alpha'\hat{\Theta}\hat{\Sigma}_{xu}\hat{\Theta}'\alpha}} \leq t,\ A_{1,n}, A_{2,n}, A_{3,n}\right) \geq \Phi(t) - 2\epsilon - \delta. \tag{A.62}$$

By (A.61) and (A.62) and $\sup_{\beta_0 \in \mathcal{B}_{\ell_0}(s_0)} P\left(\cup_{i=1}^3 A_{i,n}^c\right) = P\left(\cup_{i=1}^3 A_{i,n}^c\right) \to 0$ (here we used that none of the sets A_1, A_2, or A_3 depend on β_0) we conclude that

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

To see (38) note that

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

Thus, taking the supremum over $\beta_0 \in \mathcal{B}_{\ell_0}(s_0)$ and letting n tend to infinity yields (38) via (37).

Finally, we turn to (39). By (35) 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). □

References

Bahadur, R. R. and L. J. Savage (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics 27*(4), 1115–1122.

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. and V. Chernozhukov (2011). *High dimensional sparse econometric models: An introduction*. Springer.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*, 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2012). Inference on treatment effects after selection. *arXiv*, 12010224v3.

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.

Caner, M. and H. H. Zhang (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics 32*(1), 30–47.

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

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

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 587–613.

Kock, A. B. (2013). Oracle efficient variable selection in random and fixed effects panel data models. *Econometric Theory 29*(01), 115–152.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(01), 21–59.

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.

Lounici, K. et al. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics 2*, 90–102.

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

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

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

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 (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.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Research Papers 2013

CREATES
Center for Research in Econometric Analysis of Time Series

2014-19: Niels Haldrup and Robinson Kruse: Discriminating between fractional integration and spurious long memory

2014-20: Martyna Marczak and Tommaso Proietti: Outlier Detection in Structural Time Series Models: the Indicator Saturation Approach

2014-21: Mikkel Bennedsen, Asger Lunde and Mikko S. Pakkanen: Discretization of Lévy semistationary processes with application to estimation

2014-22: Giuseppe Cavaliere, Morten Ørregaard Nielsen and A.M. Robert Taylor: Bootstrap Score Tests for Fractional Integration in Heteroskedastic ARFIMA Models, with an Application to Price Dynamics in Commodity Spot and Futures Markets

2014-23: Maggie E. C. Jones, Morten Ørregaard Nielsen and Michael Ksawery Popiel: A fractionally cointegrated VAR analysis of economic voting and political support

2014-24: Sepideh Dolatabadim, Morten Ørregaard Nielsen and Ke Xu: A fractionally cointegrated VAR analysis of price discovery in commodity futures markets

2014-25: Matias D. Cattaneo and Michael Jansson: Bootstrapping Kernel-Based Semiparametric Estimators

2014-26: Markku Lanne, Jani Luoto and Henri Nyberg: Is the Quantity Theory of Money Useful in Forecasting U.S. Inflation?

2014-27: Massimiliano Caporin, Eduardo Rossi and Paolo Santucci de Magistris: Volatility jumps and their economic determinants

2014-28: Tom Engsted: Fama on bubbles

2014-29: Massimiliano Caporin, Eduardo Rossi and Paolo Santucci de Magistris: Chasing volatility - A persistent multiplicative error model with jumps

2014-30: Michael Creel and Dennis Kristensen: ABC of SV: Limited Information Likelihood Inference in Stochastic Volatility Jump-Diffusion Models

2014-31: Peter Christoffersen, Asger Lunde and Kasper V. Olesen: Factor Structure in Commodity Futures Return and Volatility

2014-32: Ulrich Hounyo: The wild tapered block bootstrap

2014-33: Massimiliano Caporin, Luca Corazzini and Michele Costola: Measuring the Behavioral Component of Financial Fluctuations: An Analysis Based on the S&P 500

2014-34: Morten Ørregaard Nielsen: Asymptotics for the conditional-sum-of-squares estimator in multivariate fractional time series models

2014-35: Ulrich Hounyo: Bootstrapping integrated covariance matrix estimators in noisy jump-diffusion models with non-synchronous trading

2014-36: Mehmet Caner and Anders Bredahl Kock: Asymptotically Honest Confidence Regions for High Dimensional

Dragan Te
European