AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Oracle inequalities for high-dimensional panel data models

Anders Bredahl Kock

CREATES Research Paper 2013-20

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

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

ORACLE INEQUALITIES FOR HIGH-DIMENSIONAL PANEL DATA MODELS

ANDERS BREDAHL KOCK
AARHUS UNIVERSITY AND CREATES

Abstract. This paper is concerned with high-dimensional panel data models where the number of regressors can be much larger than the sample size. Under the assumption that the true parameter vector is sparse we establish finite sample upper bounds on the estimation error of the Lasso under two different sets of conditions on the covariates as well as the error terms. Upper bounds on the estimation error of the unobserved heterogeneity are also provided under the assumption of sparsity. Next, we show that our upper bounds are essentially optimal in the sense that they can only be improved by multiplicative constants. These results are then used to show that the Lasso can be consistent in even very large models where the number of regressors increases at an exponential rate in the sample size. Conditions under which the Lasso does not discard any relevant variables asymptotically are also provided.

In the second part of the paper we give lower bounds on the probability with which the adaptive Lasso selects the correct sparsity pattern in finite samples. These results are then used to give conditions under which the adaptive Lasso can detect the correct sparsity pattern asymptotically. We illustrate our finite sample results by simulations and apply the methods to search for covariates explaining growth in the G8 countries.

Key words: Panel data, Lasso, Adaptive Lasso, Oracle inequality, Nonasymptotic bounds, High-dimensional models, Sparse models, Consistency, Variable selection, Asymptotic sign consistency.
JEL classifications: C01, C13, C23.

1. Introduction

When building an econometric model one of the first decisions one has to make is which variables are to be included in the model and which are to be left out. Often this decision is made based on economic theory but different theories might suggest different explanatory variables and this leaves the researcher with a large set of potential variables. In fact, one may often have access to many more variables than observations rendering standard techniques inapplicable. Since this kind of high-dimensional data is becoming increasingly available, the last 10-15 years have witnessed a great deal of research into procedures that can handle such data sets. In particular, a lot of attention has been given to penalized estimators. The Lasso of Tibshirani (1996) is the most prominent of these procedures and a lot of subsequent research has focussed on investigating the theoretical properties of the Lasso, see Zhao and Yu (2006), Meinshausen and Bühlmann (2006), Bickel et al. (2009), Belloni and Chernozhukov (2011) and Bühlmann and Van De Geer (2011) to mention just a few. Many other procedures have been investigated as

Date: June 12, 2013.
I would like to thank seminar participants at PUC-Rio for helpful comments and suggestions. Financial support from the Danish National Research Foundation (DNRF78) is gratefully acknowledged.

well: the SCAD of Fan and Li (2001), the Adaptive LASSO of Zou (2006), the Bridge and Marginal Bridge estimators of Huang et al. (2008), the Dantzig selector of Candes and Tao (2007), the Sure Independence Screening of Fan and Lv (2008) and the square root LASSO of Belloni et al. (2011). These procedures have become popular since they are computationally feasible and perform variable selection and parameter estimation at the same time.

Most focus in the literature has been on the standard linear regression model. However, often objects (such as individuals, firms or countries) are sampled repeatedly over time resulting in a panel data set. Since these data sets may often contain many variables it is important to have procedures that can deal with them in a theoretically sound and computationally feasible manner. In this paper we make a step in that direction by investigating the properties of the Lasso and the adaptive Lasso in the linear fixed effects panel data model

$$y_{i,t} = x_{i,t}'\beta^* + c_i^* + \epsilon_{i,t},\ i = 1, ..., N,\ t = 1, ..., T \tag{1}$$

where $x_{i,t}$ is a $p_{N,T} \times 1$ vector of covariates and where $p_{N,T}$ is indexed by N and T to indicate that the number of covariates can increase in the sample size. In the sequel we shall omit this indexation. The c_i^*s are the unobserved time homogeneous heterogeneitys (such as intelligence of a person) while the $\epsilon_{i,t}$ are the error terms about which we shall be more specific later. Even though economic theory may guide the researcher towards a set of potential explanatory variables to be included in $x_{i,t}$, large data sets are becoming increasingly available nowadays and one may not want to take a strong stand a priori on which variables to include in the model and which to leave out. This implies that $x_{i,t}$ can be a very long vector – potentially much longer than the sample size. On the other hand, only a few variables in $x_{i,t}$ might be relevant for explaining $y_{i,t}$ meaning that the vector β^* is sparse.

Oftent the unobserved heterogeneity $c_{i,t}$ is simply removed by a differencing or demeaning procedure. However, just like β^*, $c^* = (c_1^*, ..., c_N^*)$ might be a sparse vector. Example of this could be intelligence only having an effect for certain individuals when modeling income or the culture of a country when modeling its growth. It is our goal to investigate the properties of the Lasso for fixed effects panel data models in such situations. We shall see that the Lasso can estimate the two parameter vectors almost as precisely as if the true sparsity pattern had been known and only the relevant variables had been included from the outset. For the adaptive Lasso we show that it selects the correct sparsity pattern with high probability. In particular, we

(1) provide *nonasymptotic* oracle inequalities for the estimation error of the Lasso for β^* and c^* under different sets of moment/tail assumptions on the covariates and the error terms. More precisely, for a given sample size we provide upper bounds on the estimation error which hold with at least a certain probability. In the first of our settings we allow for much heavier tails than the usual sub-gaussian ones.
(2) show that our bounds are optimal in the sense that they can at most be improved by a multiplicative constant.
(3) use the nonasymptotic bounds to give a set of sufficient conditions under which the Lasso estimates β^* and c^* consistently. It turns out that the Lasso can be consistent in even very high-dimensional models. We also provide conditions under which the Lasso does not discard any relevant variables, i.e. conditions

under which it can be used as a strong initial screening device removing irrelevant variables and thus reducing the dimension of the model.

(4) establish nonasymptotic lower bounds on the probability with which the adaptive Lasso unveils the correct sparsity pattern.
(5) use the nonasymptotic bounds to give conditions under which the adaptive Lasso detects the correct sparsity pattern asymptotically.
(6) propose an efficient algorithm to implement the Lasso and the adaptive Lasso in panel data models which reduces the estimation problem to a standard Lasso one.
(7) introduce a new restricted eigenvalue condition similar in spirit to Bickel et al. (2009) and show how this can be valid even for data with non-gaussian, non-independent rows, hence extending the work of Raskutti et al. (2010) and Vershynin (2011). The proof of our Theorem 1 is also different than the one for the plain cross sectional model due to the presence of two parameter vectors which have to be treated separately.
(8) illustrate the methods by means of simulations and a real data example.

We believe that these results will be very useful for applied researchers since they provide tools with which very large panel data sets can be handled in a theoretically sound way without reducing the dimension of the model in an ad hoc way prior to estimation.

The rest of the paper is organized as follows: Section 2 introduces relevant notation and the panel Lasso. Section 3 provides a range of non-asymptotic oracle inequalities for the Lasso while Section 4 uses these inequalities to give asymptotic results for it. Next, Section 5 is concerned with finite sample probabilities of the adaptive Lasso selecting the correct sparsity pattern. It also gives sufficient conditions for when this probability tends to one asymptotically. Section 6 provides a simulation study while Section 7 contains an application to growth in the G8 countries. Finally, Section 8 concludes while all proofs are deferred to the appendix.

2. Setup and notation

Let $J_1 = \{j : \beta_j^* \neq 0\} \subseteq \{1, ..., p\}$ and $J_2 = \{i : c_i^* \neq 0\} \subseteq \{1, ..., N\}$ be the sets of active covariates and unobserved heterogeneities, respectively. $\beta_{\min} = \min\{|\beta_j^*| : j \in J_1\}$ and $c_{\min} = \min\{|c_j^*| : j \in J_2\}$ are the smallest nonzero entries of β^* and c^*, respectively. Denote by $\gamma^* = (\beta^{*\prime}, c^{*\prime})'$ and $J = J_1 \cup J_2 \subseteq \{1, ..., N + p\}$[1]. For any set A, $|A|$ denotes its cardinality while A^c denotes its complement. In particular, $|J_1| = s_1$, $|J_2| = s_2$ and $|J| = s$.

For any $x \in \mathbb{R}^n$, $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$, $\|x\|_{\ell_1} = \sum_{i=1}^n |x_i|$ and $\|x\|_{\ell_\infty} = \max_{1 \leq i \leq n} |x_i|$ denote ℓ_2, ℓ_1 and ℓ_∞ norms, respectively. For a random variable U, $\|U\|_{L_r} = (E|U|^r)^{1/r}$ denotes its L_r-norm and for a symmetric square matrix M, $\phi_{\min}(M)$ and $\phi_{\max}(M)$ denote the minimal and maximal eigenvalues of M.

For any vector $x \in \mathbb{R}^n$ and subset A of $\{1, ..., n\}$, x_J denotes the vector in $\mathbb{R}^{|J|}$ only consisting of the elements indexed by A. For a matrix R, R_A denotes the submatrix only containing the columns indexed by A while $R_{A,B}$ denotes the submatrix with rows indexed by A and columns indexed by B. Next, for any two real numbers a and b, $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For any $x \in \mathbb{R}^n$, $\text{sign}(x)$ denotes the sign function applied to each component of x.

[1]Here $J_1 \cup J_2$ is understood as $J_1 \cup (J_2 + p)$ where $J_2 + p = \{s = r + p : r \in J_2\}$ such that $J_1 \cup J_2 \subseteq \{1, ..., p + N\}$. J shall be used to index $p + N \times 1$ vectors.

Since our primary focus is high-dimensional models we shall sometimes tacitly assume that $p, N \geq e$ for the sole reason of keeping the presentation simple.

Define $X_i = (x_{i,1}, ..., x_{i,T})'$ and $X = (X_1', ..., X_N')'$. Letting ι denote the $T \times 1$ vector of ones, set $D = I_N \otimes \iota$ (where $\otimes$ denotes the Kronecker product) and define the $NT \times (p+k)$ matrix $Z = (X, D)$. We shall refer to the jth column of X by $x_j,\ j = 1, ..., p$ and to the ith column of D by $d_i,\ i = 1, ..., N$. Defining $y_i = (y_{i,1}, ..., y_{i,T})'$ and $\epsilon_i = (\epsilon_{i,1}, ..., \epsilon_{i,T})'$ for $i = 1, ..., N$ and setting $y = (y_1', ..., y_N')'$ as well as $\epsilon = (\epsilon_1', ..., \epsilon_N')'$ one may equivalently write (1) as

$$y = Z\gamma^* + \epsilon.$$

The properly scaled Gram matrix of Z will turn out to play an important role in the sequel.

2.1. **The panel Lasso.** The panel Lasso estimates $\gamma^* = (\beta^{*\prime}, c^{*\prime})'$ by minimizing the following objective function

$$L(\beta, c) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t} - x_{i,t}'\beta - c_i\right)^2 + 2\lambda_{N,T}\sum_{k=1}^{p}|\beta_k| + 2\mu_{N,T}\sum_{i=1}^{N}|c_i| \tag{2}$$

$$= \|y - Z\gamma\|^2 + 2\lambda_{N,T}\|\beta\|_{\ell_1} + 2\mu_{N,T}\|c\|_{\ell_1}. \tag{3}$$

The Lasso estimator, denoted $\hat{\gamma} = (\hat{\beta}', \hat{c}')'$, is the solution of a minimization problem which is the sum of the usual least squares objective function plus two terms that penalize β_k and c_i for being different from 0. The size of the penalty is determined by the sequences $\lambda_{N,T}$ and $\mu_{N,T}$. The larger these are, the more will the entries of $\hat{\beta}$ and $\hat{c}$ be shrunk towards zero. As will be seen later, two different regularization sequences ($\lambda_{N,T}$ and $\mu_{N,T}$) are needed to establish desirable properties of $\hat{\gamma} = (\hat{\beta}', \hat{c}')'$. On an intuitive level this is due to the fact that the number of effective observations for each $\beta_k,\ k = 1, ..., p$ is NT while it only is T for each $c_i\ i = 1, ..., N$.

2.2. **The panel restricted eigenvalue condition.** Since we are primarily interested high-dimensional models the properly scaled Gram matrix of Z will often be ill-behaved or even singular. However, Bickel et al. (2009) observed for the standard linear regression model that the Lasso does not need the smallest eigenvalue of the scaled Grammian of Z to be strictly positive on order to derive useful upper bounds on the estimation error. In particular, it suffices that a so-called *restricted eigenvalue* is bounded away from 0. We shall see next that a similar, though slightly more involved, observation can be made for the panel Lasso.

Let $S = \begin{pmatrix} \sqrt{NT}\mathbf{I}_p & 0 \\ 0 & \sqrt{T}\mathbf{I}_N \end{pmatrix}$ and set $\psi_{N,T} = S^{-1}Z'ZS^{-1}$. If $p + N > NT$ it is well known that

$$\min_{\delta \in \mathbb{R}^{p+N}\setminus\{0\}} \frac{\delta'\Psi_{N,T}\delta}{\|\delta\|^2} = \min_{\delta \in \mathbb{R}^{p+N}\setminus\{0\}} \frac{\|ZS^{-1}\delta\|^2}{\|\delta\|^2} = 0.$$

In this case ordinary least squares is infeasible. However, for the Lasso it turns out that we do not need to minimize the above Rayleigh-Ritz ratio over all of $\mathbb{R}^{p+N}$ – it suffices to minimize over a subset implying that the minimum can be non-zero even when $\Psi_{N,T}$ is not of full rank. More precisely, letting δ^1 be $p \times 1$ and δ^2 be $N \times 1$ with $\delta = (\delta^{1\prime}, \delta^{2\prime})'$

and $R_1 \subseteq \{1, ..., p\}$ as well as $R_2 \subseteq \{1, ..., N\}$ we define the $RE(r_1, r_2)$ panel restricted eigenvalue as

$$\kappa^2_{\psi_{N,T}}(r_1, r_2) = \min \left\{ \frac{\left\| ZS^{-1}\delta \right\|^2}{\|\delta\|^2} : \delta \in \mathbb{R}^{p+N} \setminus \{0\}, \ |R_1| \leq r_1, \ |R_2| \leq r_2, \right.$$

$$\left. \frac{\lambda_{N,T}}{\sqrt{NT}} \left\| \delta^1_{R_1^c} \right\|_{\ell_1} + \frac{\mu_{N,T}}{\sqrt{T}} \left\| \delta^2_{R_2^c} \right\|_{\ell_1} \leq 3\frac{\lambda_{N,T}}{\sqrt{NT}} \left\| \delta^1_{R_1} \right\|_{\ell_1} + 3\frac{\mu_{N,T}}{\sqrt{T}} \left\| \delta^2_{R_2} \right\|_{\ell_1} \right\} > 0. \tag{4}$$

The panel restricted eigenvalue condition looks similar to the one introduced in Bickel et al. (2009). It extends it in that it allows for different penalty sequences for the two groups of parameters. Similarly, for

$$\Gamma = \begin{pmatrix} E\left(\frac{X'X}{NT}\right) & 0 \\ 0 & \mathbf{I}_N \end{pmatrix}$$

define

$$\kappa^2(r_1, r_2) = \min \left\{ \frac{\delta'\Gamma\delta}{\|\delta\|^2} : \delta \in \mathbb{R}^{p+N} \setminus \{0\}, \ |R_1| \leq r_1, \ |R_2| \leq r_2, \right.$$

$$\left. \frac{\lambda_{N,T}}{\sqrt{NT}} \left\| \delta^1_{R_1^c} \right\|_{\ell_1} + \frac{\mu_{N,T}}{\sqrt{T}} \left\| \delta^2_{R_2^c} \right\|_{\ell_1} \leq 3\frac{\lambda_{N,T}}{\sqrt{NT}} \left\| \delta^1_{R_1} \right\|_{\ell_1} + 3\frac{\mu_{N,T}}{\sqrt{T}} \left\| \delta^2_{R_2} \right\|_{\ell_1} \right\}.$$

Note that for $\kappa^2 > 0$ it suffices that Γ is of full rank which is a rather standard assumption and independent of whether $p + N < NT$ or not. I turns out that in order to get tight upper bounds on the estimation error of the Lasso $\kappa^2_{\Psi_{N,T}}$ should be as large as possible. In Lemma 5 in the appendix we show that $\kappa^2_{\Psi_{N,T}}$ is close to κ^2 if $\Psi_{N,T}$ is close to Γ. Hence, it suffices that κ^2 is bounded away from zero and that $\Psi_{N,T}$ is close to Γ in order to bound $\kappa^2_{\Psi_{N,T}}$ away from 0 with high probability. In Lemmata 6 and 7 in the Appendix lower bounds on the probability with which $\kappa^2_{\Psi_{N,T}} > \kappa^2/2$ are provided using this idea for heavy- and light-tailedness assumptions on the covariates and the error terms. While the results for light-tailed (sub-gaussian) variables in Lemma 7 are to be expected in the light of previous results in the literature (see e.g. Vershynin (2011)) the results on more heavy-tailed random variables in Lemma 6 are to our knowledge new.

3. Results for the Lasso

Before stating our first result we introduce the following two sets

$$\mathcal{A}_{N,T} = \left\{ \|X'\epsilon\|_{\ell_\infty} \leq \frac{\lambda_{N,T}}{2}, \ \|D'\epsilon\|_{\ell_\infty} \leq \frac{\mu_{N,T}}{2} \right\} \text{ and } \mathcal{B}_{N,T} = \left\{ \kappa^2_{\Psi_{N,T}} \geq \kappa^2/2 \right\}.$$

The set $\mathcal{A}_{N,T}$ is the set where none of the covariates X or D are too highly correlated with the error term. This requirement limits the number of variables in X and D. Working on the set $\mathcal{B}_{N,T}$ means restricting attention to settings where the restricted eigenvalue of $\Psi_{N,T}$ is not too small.

Theorem 1 gives upper bounds on the estimation error of the Lasso on $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}$ and will be our main tool to derive further bounds under more specific assumptions on the covariates and the error terms. It is worth emphasizing that it is a purely algebraic result without any probabilities attached to it yet.

Theorem 1. *On $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}$ with $\kappa^2 > 0$ one has for any positive sequences $\lambda_{N,T}$ and $\mu_{N,T}$*

$$\left\|\hat{\beta} - \beta^*\right\| \leq \frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT} + \frac{4\mu_{N,T}\sqrt{s_2}}{\kappa^2\sqrt{N}T} \tag{5}$$

and

$$\left\|\hat{c} - c^*\right\| \leq \frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 T} + \frac{4\lambda_{N,T}\sqrt{s_1}}{\kappa^2\sqrt{N}T}. \tag{6}$$

We stress that the claims in Theorem 1 are deterministic. Probabilities will be attached to the bounds once we have made statistical assumptions on the covariates and the error terms.

The bounds in Theorem 1 reveal that the further κ^2 is away from zero the more precisely can one estimate the parameters of the model. This is reasonable since it means that the problem is in some sense far from a singular one. However, the set $\mathcal{B}_{N,T}$ is clearly decreasing in κ^2, revealing a tradeoff between the sharpness of the upper bounds on the estimation error and the size of the set on which the bounds hold. The same tradeoff is present for $\lambda_{N,T}$ and $\mu_{N,T}$ – the set $\mathcal{A}_{N,T}$ is increasing in both of these but the same is true for the upper bounds on the estimation error. Put differently, small values of $\lambda_{N,T}$ and $\mu_{N,T}$ give tight bounds on the estimation error but the bounds are only valid on a smaller set. Our next two theorems investigate the tradeoff further under different sets of assumptions on the tail behaviour of the covariates and the error terms. First, we shall put forward the statistical assumptions of the panel data model:

A1 a) $\{X_i, \epsilon_i\}_{i=1}^N$ are identically and independently distributed
b) X_i and ϵ_i are independent for $i = 1, ..., N$
c) $\{\epsilon_{1,t}\}_{t=1}^T$ are independent with mean zero.

Assumption A1a) is standard in the panel data literature, see e.g. Wooldridge (2002) or Arellano (2003). Part b) is also relatively standard but slightly stronger than $E(\epsilon_{it}|X_i) = 0$ which is often assumed. However, for most applied work involving panel data it is hard to come up with realistic examples where $E(\epsilon_{it}|X_i) = 0$ but X_i and ϵ_i are not independent[2]. A1c) is standard. Note that we are *not* assuming that $\{\epsilon_{1,t}\}_{t=1}^T$ are identically distributed. In particular, they may be heteroscedastic. Put differently, for every $i = 1, ..., N$ $(\epsilon_{i,1}, ..., \epsilon_{i,T})$ is distributed the same way, but the marginal distributions of the individual elements may be non-identical.

Furthermore, the upper bounds on the estimation errors in (5) and (6) as well as the probability with which they hold, depend on the number of moments the error terms and covariates possess. We shall give results under two different sets of conditions.

A2a) $E(|x_{1,t,k}|^r)$, $E(|\epsilon_{1,t}|^r) < \infty$ for some $r \geq 2$ and $t = 1, ..., T$, $k = 1, ..., p$. Actually, we shall assume $\max_{1\leq t\leq T} E|x_{1,t,k}|^r \leq 1$ for all $k = 1, ..., p$.

Assumption A2a) is a moment assumption stating that the covariates as well as the error terms possess r moments. $\max_{1\leq t\leq T} E|x_{1,t,k}| \leq 1$ for all $k = 1, ..., p$ is merely a normalization for technical convenience and to keep expressions simple. All results remain valid without this normalization.

[2]Of course it is possible to construct examples where $E(\epsilon_{it}|X_i) = 0$ but X_i and ϵ_i are not independent. See e.g. Stoianov (1997).

A2b) $x_{1,t,k}$ and $\epsilon_{1,t}$ are uniformly subgaussian, i.e. there exist constants C and K such that $P(|x_{1,t,k}| \geq t)$, $P(|\epsilon_{1,t}| \geq t) \leq \frac{1}{2}Ke^{-Ct^2}$ for all $1 \leq t \leq T$ and $1 \leq k \leq p$.

Assumption A2b) controls the tail behaviour of the covariates and the error terms (and hence also its moments). It is a standard assumption in the high-dimensional statistics literature and much more restrictive than A2a) which only assumes the existence of r moments. However, we will see that the dimension of the models considered can be a lot larger under A2b) than under A2a).

We are now ready to transform the deterministic statement in Theorem 1 into probabilistic ones. We stress that the bounds below are *finite sample* bounds, i.e. for a given sample size we provide upper bounds on the estimation error that hold with at least a certain probability. First, we work under assumption A2a):

Theorem 2. *Let assumption A1) and A2a) be satisfied and assume that $\kappa^2 > 0$. Then, choosing $\lambda_{N,T} = 4a_{N,T}p^{1/r}(NT)^{1/2}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}$ and $\mu_{N,T} = 4a_{N,T}N^{1/r}T^{1/2}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}$ for any positive sequence $a_{N,T}$ one has $P(\mathcal{A}_{N,T}\cap\mathcal{B}_{N,T}) \geq 1-2\left(\frac{C_r}{a_{N,T}}\right)^r - D_r\frac{(p^2+Np)(s_1+s_2)^{r/2}(\frac{p}{N}\vee\frac{N}{p})}{\kappa^r N^{r/4}}$ for constants C_r and D_r only depending on r. Furthermore, with at least this probability (i.e. on $\mathcal{A}_{N,T}\cap\mathcal{B}_{N,T}$),*

$$\left\|\hat{\beta}-\beta^*\right\| \leq \frac{\xi_{N,T}}{\sqrt{NT}} \tag{7}$$

and

$$\left\|\hat{c}-c^*\right\| \leq \frac{\xi_{N,T}}{\sqrt{T}} \tag{8}$$

where $\xi_{N,T} = 32a_{N,T}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}\left(p^{1/r}\sqrt{s_1}+N^{1/r}\sqrt{s_2}\right)/\kappa^2$.

First, note that the more moments the covariates and the error terms possess (r large) the smaller can $\lambda_{N,T}$ and $\mu_{N,T}$ be chosen and hence the upper bounds on the estimation error are smaller in accordance with Theorem 1. $\xi_{N,T}$ may be interpreted as the punishment on the convergence rate for not knowing the true model. Since $a_{N,T}$ will in general be chosen to be an increasing sequence one sees that in the setting of fixed T, p, s_1 and s_2 the upper bound on $\left\|\hat{\beta}-\beta^*\right\|$ is of the order $a_{N,T}N^{1/r-1/2}$ (if κ^2 is bounded away from zero) which is not far from $1/\sqrt{N}$ if r is large and $a_{N,T}$ is increasing slowly.

If $\epsilon_{1,t}$ is uniformly bounded in L_r, which is the case if they are e.g. identically distributed, then the term $\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}$ can be disregarded in asymptotic considerations. Furthermore, (8) confirms the well known fact that T must be large in order to estimate c^* precisely since there are only T observation per c_i^*, $i = 1, ..., N$.

We also stress that Theorem 2 does not require sub-gaussianity of the covariates and the error terms and in this respect it relaxes one of the standard assumptions in the high-dimensional modeling literature. The next theorem is similar in spirit to Theorem 2 but strengthens the existence of r moments to sub-gaussian tails of the covariates as well as the error terms, i.e. we invoke A2b) instead of A2a).

Theorem 3. *Let assumption A1) and A2b) be satisfied and assume that $\kappa^2 > 0$. Then, choosing $\lambda_{N,T} = \sqrt{4NT\log(p)^3\log(a_{N,T})^3}$ and $\mu_{N,T} = \sqrt{4T\log(N)^3\log(a_{N,T})^3}$ for any sequence $a_{N,T} \geq e$ one has $P(\mathcal{A}_{N,T}\cap\mathcal{B}_{N,T}) \geq 1 - Ap^{1-B\log(a_{N,T})} - AN^{1-B\ln(a_{N,T})} - A(p^2+Np)e^{-B(t^2N)^{1/3}}$ for absolute constants A and B, $t = \frac{\kappa^2}{(s_1+s_2)\left(\frac{\ln(p)}{\ln(N)}\vee\frac{\ln(N)}{\ln(p)}\right)^3}$ and*

$Nt^2 \geq 1$. Furthermore, with at least this probability (i.e. on $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}$),

$$\left\|\hat{\beta} - \beta^*\right\| \leq \frac{\xi_{N,T}}{\sqrt{NT}} \tag{9}$$

and

$$\left\|\hat{c} - c^*\right\| \leq \frac{\xi_{N,T}}{\sqrt{T}} \tag{10}$$

where $\xi_{N,T} = 16\log(a_{N,T})^{3/2}\left[\log(p)^{3/2}\sqrt{s_1} + \log(N)^{3/2}\sqrt{s_2}\right]/\kappa^2$.

The form of the upper bounds on the estimation errors is the same as in Theorem 2. However, the definition of $\xi_{N,T}$ has changed. In particular, $\xi_{N,T}$ is now increasing slower in the number of variables, p, in X and N in D, respectively. In the case where T, p, s_1 and s_2 are bounded the upper bound on $\left\|\hat{\beta} - \beta^*\right\|$ is of order $\ln(a_{N,T})^{3/2}\ln(N)^{3/2}/\sqrt{N}$ (if κ^2 is bounded away from 0). In other words, the punishment for not knowing the true model is now merely logartithmic in the sample size.

In lower bounding the probability of $\mathcal{A}_{N,T}$ in Theorem 3 we have used a concentration inequality for unbounded martingales due to Lesigne and Volnỳ (2001) which they show is optimal.

So far, we have focussed on providing upper bounds on the estimation error. An obvious question is now how tight these bounds are. It turns out that the established bounds are indeed tight. In particular, we show next that no improvements can be made beyond multiplicative constants. First, note that Theorem 3 implies that

$$\left\|S_T(\hat{\gamma} - \gamma^*)\right\| \leq 2\xi_{N,T} \tag{11}$$

with high probability[3]. The following theorem shows that the upper bound in (11) cannot be improved in the case of gaussian error terms.

Theorem 4. *Let A1) and A2b) be satisfied and assume that κ^2 is bounded away from zero. Assume that ϵ_i is $N(0, \sigma^2 I_T)$ and $\phi_{\min}(\Gamma_{J,J}), \kappa^2$ are bounded from below and $\phi_{\max}(\Gamma_{J,J})$ is bounded from above. Choose $\lambda_{N,T}$ and $\mu_{N,T}$ as in Theorem 3. Then when the Lasso detects the correct sparsity pattern, it holds with probability at least $1 - \exp(-c_1|J|) - A(p^2 + Np)e^{-B(t^2N)^{1/3}}$ that*

$$\left\|S_{J,J}(\hat{\gamma}_J - \gamma_J^*)\right\| \geq c_2\xi_{N,T} \tag{12}$$

for absolute constants c_1, c_2, A and B and $t = \frac{\kappa^2}{(s_1+s_2)\left(\frac{\ln(p)}{\ln(N)} \vee \frac{\ln(N)}{\ln(p)}\right)^3}$ as long as $Nt^2 \geq 1$ where $\xi_{N,T}$ is as in Theorem 3.

Inequality (12) is the reverse inequality of (11) and shows that one cannot improve the bounds in Theorem 3 except for multiplicative constants. Hence, our results are sharp and we turn next towards the asymptotic implications of our finite sample bounds.

4. Asymptotic properties of the Lasso

In this section we show that the Lasso can estimate β^* and c^* consistently in even very high-dimensional settings where the number of covariates increases exponentially in N. It is also shown that no relevant variables will be discarded from the model as long as $\beta_{\min}$

[3]To be precise, with at least the lower bound on $P(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T})$ provided in Theorem 3.

and $c_{\min}$ do not tend to zero too fast. Let $a_{N,T} = N, T = N^a, p = e^{N^b}$ and $s_1 = s_2 = N^c$ for $a, b, c \geq 0$. Then we have the following result which builds upon Theorem 3.

Theorem 5. *Let assumptions A1) and A2b) be satisfied and assume that κ^2 is bounded away from zero. Then, if $9b + 2c \leq 1$ as $N \to \infty$ one has with probability tending to one*

(1)

$$\left\|\hat{\beta} - \beta^*\right\| \to 0 \text{ if } 3b + c < 1 + a$$
$$\left\|\hat{c} - c^*\right\| \to 0 \text{ if } 3b + c < a$$

(2) *$\hat{\beta}_j$ will not be classified as zero for any $j \in J_1$ if $\beta_{\min} > \xi_{N,T}/\sqrt{NT}$. Similarly, no $\hat{c}_i$ will be classified as zero for $i \in J_2$ if $c_{\min} > \xi_{N,T}/\sqrt{T}$.*

The first part of Theorem 5 shows that even when p increases exponentially in N, it is possible for the Lasso to be consistent for β^* as well as c^*. Put differently, the Lasso can be consistent in even ultra high-dimensional models. However, and as can be expected, one must have $a > 0$ in order to estimate c^* consistently since only $T = N^a$ observations are available to estimate each $c_i^*,\ i = 1, ..., N$. In the case of standard *large* N asymptotics ($a = 0$), the Lasso can still be consistent for β^* as long as $9b + 2c \leq 1$. This is clearly satisfied in the standard setting of fixed p, s_1 and s_2 ($b = c = 0$).

The second part of the theorem reveals that the Lasso can be used as a strong screening device since no relevant variables will be excluded from the model if their coefficients are not too close to zero. The necessity of such a "beta-min" (or "c-min") condition is not surprising since one cannot expect to be able to distinguish non-zero parameters from zero ones if the distance between these is too small. It is not difficult to see that in the standard large N setting of $a = b = c = 0$, the "beta-min" condition requires $\beta_{\min} \geq \log(N)^3/\sqrt{N}$. Hence, all non-zero parameters outside a disc centered at zero with radius $\log(N)/\sqrt{N}$ will also be classified as non-zero by the Lasso. In the same setting, the "c-min" condition requires $c_{\min} \geq \ln(N)^3$ implying that in the limit only $c_i \geq \ln(N)^3,\ i \in J_2$ can be guaranteed to be classified as non-zero. Put differently, only large c_i^* can be guaranteed to be classified as non-zero. One must have $a > 3b + c$ in order for this disc to have a radius which tends to zero, i.e. to make sure that any non-zero c_i^* will be classified as non-zero in fixed parameter asymptotics. The necessity of the non-zero parameters being bounded away from zero is not surprising in the light of the work of Pötscher and Leeb (2009) who document some of the limitations of the Lasso-type estimators.

It is also worth mentioning that the conditions of Theorem 5 are merely sufficient. For example it is also possible to let κ^2 tend to zero at the price of slower growth rates in the other variables without sacrificing consistency. Furthermore, one could also use Theorem 2 instead of Theorem 3 to deduce a theorem in the spirit of Theorem 5. Of course, the models sizes would no longer be allowed to increase as fast as above.

5. The Adaptive Lasso

So far we have focussed on deriving upper bounds on the estimation error that hold with high probability. Next, we turn to variable selection. The Lasso penalizes all parameters equally much. This implies that it can only recover the correct sparsity pattern under rather stringent assumptions. If one could penalize the truly zero parameters more than the non-zero ones, one would expect a better performance. This idea was utilized by Zou

(2006) to propose the adaptive Lasso in the standard linear regression model with a fixed number of non-random regressors. He established that the adaptive Lasso can detect the correct sparsity pattern *asymptotically* in such a setting. This motivates us to modify the adaptive Lasso to make it applicable in the linear panel data model and to derive lower bounds on the finite sample probabilities with which it selects the correct sparsity pattern. The adaptive Lasso estimates β^* and c^* by minimizing the following objective function:

$$\tilde{L}(\beta,c)=\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}-x_{i,t}'\beta-c_i\right)^2+\lambda_{N,T}\sum_{k\in J_1(\hat{\beta})}\frac{|\beta_k|}{|\hat{\beta}_k|}+\mu_{N,T}\sum_{i\in J_2(\hat{c})}\frac{|c_i|}{|\hat{c}_i|} \tag{13}$$

where $J_1(\hat{\beta})=\{j:\hat{\beta}_j\neq 0\}$ and $J_2(\hat{c})=\{i:\hat{c}_i\neq 0\}$. Denote the minimizers of L by $\tilde{\beta}$ and $\tilde{c}$, respectively. Note that if $\hat{\beta}_j$ or $\hat{c}_i$ equal zero, the corresponding variable is entirely excluded from the model in the second step. Hence, the dimension of the second step estimation can be of a much smaller order of magnitude than the first step estimation. If $\beta_j^*=0$ then it follows by Theorems 2 and 3 that $\hat{\beta}_j$ is likely to be small (or even 0) and so the penalty on β_j in (13) is large implying that $\tilde{\beta}$ is likely to be classified as being zero. The reverse logic applies when $\beta_j^*\neq 0$ (and similarly for c_i^*). Put differently, the adaptive Lasso is a two-step estimator which uses more intelligent weights than the ordinary Lasso. We shall see next, that these more intelligent weights imply that the adaptive Lasso can select the correct sparsity pattern. As for the Lasso, we start with a purely deterministic result to which we then attach probabilities by adding assumptions A1) and A2a) or A2b). First, define the sets

$$\mathcal{C}_{1,N,T}=\left\{\max_{k\in J_1^c}\max_{l\in J_1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l}\vee\max_{i\in J_2}\max_{k\in J_1^c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\leq K_{1,N,T}\right\},$$

$$\mathcal{C}_{2,N,T}=\left\{\max_{i\in J_2^c}\max_{k\in J_1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\leq K_{2,N,T}\right\}\text{ and }\mathcal{D}_{N,T}=\left\{\phi_{\min}(\Psi_{J,J})\geq\phi_{\min}(\Gamma_{J,J})/2\right\}.$$

$\mathcal{C}_{1,N,T}$ may be interpreted as the set where none of the irrelevant x_j's has a too big inner product (in ℓ_2), or covariance, with any of the relevant x_j's or dummies in D. Similarly $\mathcal{C}_{2,N,T}$ is the set where none of the relevant dummies is too highly correlated with any of the relevant x_j's (all dummies are orthogonal by construction so no condition is needed on their interdependence). On these sets, the problem is well-posed in the sense that the relevant and irrelevant variables are not too highly correlated and hence we can distinguish between them as we will see below. On the set $\mathcal{D}_{N,T}$, one basically has that $\Psi_{J,J}$ is bounded away from singularity. With these definitions in place we may state the following theorem.

Theorem 6. *On* $\mathcal{A}_{N,T}\cap\mathcal{C}_{1,N,T}\cap\mathcal{D}_{N,T}\cap\{\|\hat{\beta}-\beta^*\|\leq\beta_{\min}/2\}\cap\{\|\hat{c}-c^*\|\leq c_{\min}/2\}$ *one has* $\operatorname{sign}(\tilde{\beta})=\operatorname{sign}(\beta^*)$ *if*

$$\frac{2\sqrt{|J|}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}}\vee\frac{\mu_{N,T}}{2\sqrt{T}}+\frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}}\vee\frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right)\leq\sqrt{NT}\beta_{\min} \tag{14}$$

$$\frac{2|J|K_{1,N,T}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}}\vee\frac{\mu_{N,T}}{2\sqrt{T}}+\frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}}\vee\frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right)+\frac{\lambda_{N,T}}{2}\leq\frac{\lambda_{N,T}}{\|\hat{\beta}-\beta^*\|}. \tag{15}$$

Similarly, on $\mathcal{A}_{N,T} \cap \mathcal{C}_{2,N,T} \cap \mathcal{D}_{N,T} \cap \{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\} \cap \{\|\hat{c} - c^*\| \leq c_{\min}/2\}$ *one has* $\operatorname{sign}(\tilde{c}) = \operatorname{sign}(c^*)$ *if*

$$\frac{2\sqrt{|J|}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}} + \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right) \leq \sqrt{T}c_{\min} \tag{16}$$

$$\frac{2|J|K_{2,N,T}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}} + \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right) + \frac{\mu_{N,T}}{2} \leq \frac{\mu_{N,T}}{\|\hat{c} - c^*\|}. \tag{17}$$

Note that just as Theorem 1, Theorem 6 is purely deterministic. Inequality (14) is sufficient to ensure that no relevant x_j's are excluded from the model. It is sensible that the smaller $\beta_{\min}$ is the more difficult it is to avoid excluding relevant variables. This is reflected in (14) in that the left hand side is deceasing in $\beta_{\min}$ while the right hand side is increasing. Larger $\lambda_{N,T}$ and $\mu_{N,T}$ also make it harder to satisfy the inequality since too much shrinkage can result in relevant variables being discarded. On the other hand, as in Theorem 1, the size of $\mathcal{A}_{N,T}$ is increasing in these two quantities revealing the same tradeoff as discussed previously.

Inequality (15) gives a sufficient condition for not classifying any irrelevant x_js as relevant. Note that the more precise the initial Lasso estimator is the larger is the right hand side and hence the more likely it is that the inequality is satisfied. Increasing $K_{1,N,T}$ allows for larger dependence between relevant and irrelevant variables and thus makes it harder to distinguish between these. Hence, it is sensible that the left hand side of (15) is increasing in $K_{1,N,T}$. On the other hand, the size of $\mathcal{C}_{1,N,T}$ is increasing in $K_{1,N,T}$. The intuition behind inequalities (16) and (17) is the same for the preceding two inequalities. At this point it is also worth mentioning that Theorem 6 does not assume the use of the Lasso as initial estimator. The estimators $\hat{\beta}$ and $\hat{c}$ could be any estimators for which an upper bound on the estimation error is available and – as can be seen – more precise initial estimators will make the conditions of Theorem 6 more likely to be satisfied.

Next, we use the above theorem to give lower bounds on the probability with which the adaptive Lasso selects the correct sparsity pattern by invoking assumptions A1) and A2a) or A2b), respectively.

Corollary 1. (1) *Let assumptions A1 and A2a) be satisfied and assume that (14)-(15) are valid with* $\lambda_{N,T}$ *and* $\mu_{N,T}$ *as in Theorem 2 and* $K_{1,N,T} = |J_1^c|^{2/r}|J_1|^{2/r}(NT)^{1/2}a_{N,T}$. *Assume that* $\beta_{\min} \geq 2\frac{\xi_{N,T}}{\sqrt{NT}}$ *and* $c_{\min} \geq 2\frac{\xi_{N,T}}{\sqrt{T}}$ *with* $\xi_{N,T}$ *as in Theorem 2. Then,* $\operatorname{sign}(\tilde{\beta}) = \operatorname{sign}(\beta^*)$ *with probability at least* $1 - 2\left(\frac{C_r}{a_{N,T}}\right)^r - D_r\frac{(p^2+Np)(s_1+s_2)^{r/2}(\frac{p}{N}\vee\frac{N}{p})}{\kappa^r N^{r/4}} - \frac{2}{a_{N,T}^{r/2}}$ *for constants* C_r *and* D_r *only depending on* r. *Similarly, if (16)-(17) are valid with* $K_{2,N,T} = |J_1|^{1/r}|J_2^c|^{1/r}T^{1/2}a_{N,T}$ *then* $\operatorname{sign}(\tilde{c}) = \operatorname{sign}(c^*)$ *with probability at least* $1 - 2\left(\frac{C_r}{a_{N,T}}\right)^r - D_r\frac{(p^2+Np)(s_1+s_2)^{r/2}(\frac{p}{N}\vee\frac{N}{p})}{\kappa^r N^{r/4}} - \frac{1}{a_{N,T}^{r/2}}$.

(2) *Let assumptions A1 and A2b) be satisfied and assume that (14)-(15) are valid with* $\lambda_{N,T}$ *and* $\mu_{N,T}$ *as in Theorem 3 and* $K_{1,N,T} = A\log(1 + |J_1^c|)\log(e + |J_1|)\sqrt{NT}\log(a_{N,T})$ *for* $A > 0$. *Assume that* $\beta_{\min} \geq 2\frac{\xi_{N,T}}{\sqrt{NT}}$ *and* $c_{\min} \geq 2\frac{\xi_{N,T}}{\sqrt{T}}$ *with* $\xi_{N,T}$ *as in Theorem 3. Then,* $\operatorname{sign}(\tilde{c}) = \operatorname{sign}(c)$ *with probability at least* $1 - Ap^{1-B\log(a_{N,T})} - AN^{1-B\ln(a_{N,T})} - A(p^2+Np)e^{-B(t^2N)^{1/3}} - \frac{4}{a_{N,T}}$ *for absolute constants* A *and* B *and* $t = \frac{\kappa^2}{(s_1+s_2)\left(\frac{\ln(p)}{\ln(N)}\vee\frac{\ln(N)}{\ln(p)}\right)^3}$ *as long as* $Nt^2 \geq 1$. *Similarly, if*

(16)-(17) are valid with $K_{2,N,T} = A\log(1+|J_1|)\log(1+|J_2^c|)\sqrt{T}\log(a_{N,T})$ *then* $\text{sign}(\tilde{c}) = \text{sign}(c^*)$ *with probability at least* $1 - Ap^{1-B\log(a_{N,T})} - AN^{1-B\ln(a_{N,T})} - A(p^2+Np)e^{-B(t^2N)^{1/3}} - \frac{2}{a_{N,T}}$.

Corollary 1 gives lower bounds on the probability with which the adaptive Lasso detects the correct sparsity pattern under the two sets of assumptions employed in Theorems 2 and 3, respectively. Corollary 1 can also be used to derive a crude lower bound on $P(\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*), \text{sign}(\tilde{c}) = \text{sign}(c^*))$. A tighter bound can be derived by optimizing the proof slightly.

In order to get a feeling for the size of the models that the adaptive Lasso can detect the correct sparsity pattern in, we shall use part (2) of the Corollary 1 to establish the following asymptotic result. As with Theorem 5 we shall consider the asymptotic setting where $a_{N,T} = N, T = N^a, p = e^{N^b}$ and $s_1 = s_2 = N^c$ for $a, b, c \geq 0$.

Theorem 7. *Let assumptions A1 and A2b) be satisfied and let* κ, $\beta_{\min}$ *and* $c_{\min}$ *be bounded away from 0. Assume furthermore, that* $9b + 2c \leq 1$. *Then,*

(1) $P\left(\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)\right) \to 1$ *if* $5b + 3c < 1 + a$
(2) $P\left(\text{sign}(\tilde{c}) = \text{sign}(c^*)\right) \to 1$ *if* $6b + 3c < a$.

Part one of Theorem 7 reveals that p may increase at a sub-exponential rate while the number of relevant variables cannot increase faster than the square root of the sample size (set $b = 0$ in $9b + 2c \leq 1$ to conclude that $c \leq 1/2$) if the adaptive Lasso is to detect the correct sparsity pattern asymptotically. Actually, for $a < 1/2$ the number of relevant variables must increase even slower. It is also worth noticing that sign consistency can be achieved in a fixed T setting ($a = 0$). This is in opposition to part 2 of the theorem: for the adaptive Lasso to be sign consistent for c^* one needs $a > 0$. This is of course sensible in the light of Theorem 5 since $a > 0$ is needed for the first step Lasso estimator to be consistent.

6. Monte Carlo

In this section we investigate the finite sample properties of the Lasso as well as the adaptive Lasso by means of Monte Carlo experiments. The Lasso is implemented using the publicly available `glmnet` package for R. Since $\mu_{N,T}/\lambda_{N,T}$ is roughly equal to $1/\sqrt{N}$ in Theorems 2 and 3 we can reduce the optimization problem to a search over only one tuning parameter in the following way:

(1) Define $\tilde{D} = \sqrt{N}D$.
(2) Minimize $\left\|y - X\beta - \tilde{D}c\right\|^2 + \lambda_{N,T}\sum_{k=1}^{p}|\beta_k| + \lambda_{N,T}\sum_{i=1}^{N}|c_i|$ wrt. (β, c) by `glmnet` and denote the minimizer by $(\hat{\beta}, \hat{c})$.
(3) Return $(\hat{\beta}, \hat{c}) = (\hat{\beta}, \sqrt{N}\hat{c})$.

In step 2 above $\lambda_{N,T}$ is chosen by BIC. It is our experience that more time consuming procedures such as cross validation do not improve the results. The adaptive Lasso is implemented in the following way:

(1) Define $\tilde{x}_j = x_j\hat{\beta}_j$ $j = 1, ..., p$ and $\tilde{d}_i = \sqrt{N}\hat{c}_i d_i$, $i = 1, ..., N$.
(2) Minimize $\left\|y - \sum_{j=1}^{p}\tilde{x}_j\beta - \sum_{i=1}^{N}\tilde{d}_i c_i\right\|^2 + \lambda_{N,T}\sum_{k=1}^{p}|\beta_k| + \lambda_{N,T}\sum_{i=1}^{N}|c_i|$ wrt. (β, c) by `glmnet` and denote the minimizer by $(\tilde{\tilde{\beta}}, \tilde{\tilde{c}})$.
(3) Return $\tilde{\beta}_j = \tilde{\tilde{\beta}}_j\hat{\beta}_j, j = 1, ..., p$ and $\tilde{c}_i = \sqrt{N}\hat{c}_i\tilde{\tilde{c}}_i$.

As for the Lasso, $\lambda_{N,T}$ is chosen by BIC. The above implementation of the adaptive Lasso is similar in spirit to the one described in Zou (2006). To provide a benchmark for the Lasso and the adaptive Lasso, least squares including all variables is also implemented whenever feasible. This procedure is denoted OLSA. At the other extreme, least squares *only* including the relevant variables is applied to provide an infeasible target which we are ideally aiming at. This procedure is called the OLS Oracle (OLSO). We measure the performance of the proposed estimators along the following dimensions

(1) The average root mean square error of the parameter estimates of β^* and c^*, i.e. the average ℓ_2 estimation error.
(2) How often is the true model included in the model chosen. This is relevant since even if the true model is not selected a good procedure should not exclude too many relevant variables. This measure is reported for β^* as well as c^*.
(3) How often is the correct sparsity pattern uncovered, i.e. how often is exactly the correct model chosen. This measure is reported for β^* as well as c^*.
(4) What is the mean number of non-zero parameters in the estimated model. This measures how much the dimension of the model is reduced and is reported for β^* as well as c^*.

The following experiments are carried out to gauge the performance along the above dimensions (the number of Monte Carlo replications is always 1000).

- Experiment A: N=T=10 with β^* having five entries of 1 and 20 of zero. The non-zero entries are equidistant. c^* has $floor(N^{1/3}) = 2$ entries of 1 and the rest zeros. The correlation between the ith and jth column of X is $0.75^{|i-j|}$ and the covariates in X possess two moments only.
- Experiment B: As experiment A but with $N = 100$ and c^* having $floor(N^{1/3}) = 4$.
- Experiment C: As experiment A but with $T = 100$.
- Experiment D: As experiment A but with gaussian covariates.
- Experiment E: As experiment B but with gaussian covariates.
- Experiment F: As experiment C but with gaussian covariates.
- Experiment G: As experiment A but now β^* has five entries of one and 245 entries of zero. The non-zero entries are equidistant.
- Experiment H: As experiment G but with gaussian covariates.
- Experiment I: N=T=10 with β^* having 10 entries of 1 and 490 of zero. The non-zero entries are equidistant. c^* has $floor(N^{1/3}) = 2$ entries of 1 and the rest zeros. The correlation between the ith and jth column of X is $0.75^{|i-j|}$ and the covariates in X re gaussian.

Experiments A-C are meant to illustrate Theorem 2 and part 1 of Corollary 1. Note that tails of the covariates and the error terms are extremely heavy in these experiments since they merely allow for the existence of two moments. Similarly, Experiments D-F are meant to illustrate Theorem 3 and part 2 of Corollary 1 as the tails of the covariates and error terms are now subgaussian (in fact they are exactly gaussian) allowing the existence of all (polynomial) moments. Experiments G-H intend to investigate the performance of the Lasso and the adaptive Lasso in settings with more variables than observations and various moment assumptions on the covariates and the error terms.

6.1. **Results.** Experiment A reveals that the Lasso as well as the adaptive Lasso estimate β^* and c^* at a precision which lies in between the one of least squares including all variables and the least squares oracle. The adaptive Lasso retains all non-zero β^*s in

		MSE(β)	MSE(c)	Sub(β)	Sub(c)	Spar(β)	Spar(c)	$\#\beta$	$\#c$
Exp A	Lasso	1.02	1.16	0.87	0.42	0.01	0.09	9.17	2.47
	ALasso	0.87	1.31	0.75	0.34	0.31	0.13	5.90	1.84
	OLSO	0.39	0.64	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA	2.04	1.89	1.00	1.00	0.00	0.00	25.00	10.00
Exp B	Lasso	0.33	2.15	1.00	0.00	0.04	0.00	8.06	2.26
	ALasso	0.14	2.83	1.00	0.00	0.89	0.00	5.12	2.02
	OLSO	0.12	0.97	1.00	1.00	1.00	1.00	5.00	4.00
	OLSA	0.54	5.45	1.00	1.00	0.00	0.00	25.00	100.00
Exp C	Lasso	0.29	0.43	1.00	0.99	0.02	0.49	9.08	2.72
	ALasso	0.14	0.28	1.00	0.99	0.91	0.84	5.11	2.16
	OLSO	0.12	0.22	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA	0.51	0.54	1.00	1.00	0.00	0.00	25.00	10.00

TABLE 1. MSE(β) and MSE(c) are the average root mean square errors of the parameter estimates. Sub(β) and Sub(c) indicate the fraction of times the estimated model contains all the relevant variables (in X and D) while Spar(β) and Spar(c) show how often exactly the correct subset of variables is chosen. Finally, $\#\beta$ and $\#c$ give the average number of non-zero βs and cs, respectively.

75% of the instances while only including 5.9 variables on average (recall that there are 5 relevant variables).

Increasing N to 100, Experiment B shows that β^* is now estimated more precisely while the opposite is the case for c^*. It is to be expected, however, that the mean square error of $\hat{c}$ increases since the vector now has 100 entries to be estimated as opposed to only 10 in Experiment A. The adaptive Lasso always retains all non-zero β^*s while detecting exactly the right sparsity pattern in 89% of the cases. This is never the case for c^*, the reason being the same as mentioned above.

In Experiment C, T is increased to 100 while $N = 10$. This results in a higher precision of all estimators. In particular, the adaptive Lasso estimates β^* and c^* almost as precisely as the least squares oracle. The number of selected variables is also close to the ideal number.

Experiments D-F use gaussian covariates and error terms instead of ones with only two moments. Comparing the results to those in Experiments A-C reveals that the Lasso and the adaptive Lasso perform better now. Note for example, in Experiment D, the adaptive Lasso does not estimate β^* much less precisely than the least squares oracle while in Experiment A it was more than twice as imprecise. Furthermore, all non-zero c^* are classified as such by the Lasso in 81% of the Monte Carlo replications while in the corresponding number in Experiment A was only 42%.

Moving from Experiment D to E all measures pertaining to β^* improve – the parameters are estimated more precisely (the adaptive Lasso is actually as precise as the least squares oracle) and the correct sparsity pattern is selected more than 9 out of ten times. As can be expected all measures pertaining to c^* worsen since the number of parameters to be estimated ten-doubles.

In Experiment F, the Lasso and the adaptive Lasso perform well along all dimensions.

		MSE(β)	MSE(c)	Sub(β)	Sub(c)	Spar(β)	Spar(c)	$\#\beta$	$\#c$
Exp D	Lasso	0.57	0.78	1.00	0.81	0.01	0.24	9.56	3.14
	ALasso	0.34	0.72	1.00	0.74	0.62	0.41	5.55	2.30
	OLSO	0.23	0.41	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA	1.14	1.14	1.00	1.00	0.00	0.00	25.00	10.00
Exp E	Lasso	0.19	1.55	1.00	0.26	0.05	0.06	8.17	3.63
	ALasso	0.08	1.40	1.00	0.23	0.93	0.09	5.08	3.17
	OLSO	0.07	0.59	1.00	1.00	1.00	1.00	5.00	4.00
	OLSA	0.31	3.20	1.00	1.00	0.00	0.00	25.00	100.00
Exp F	Lasso	0.17	0.25	1.00	1.00	0.02	0.56	9.05	2.65
	ALasso	0.07	0.14	1.00	1.00	0.96	0.92	5.04	2.08
	OLSO	0.07	0.12	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA	0.29	0.31	1.00	1.00	0.00	0.00	25.00	10.00

TABLE 2. MSE(β) and MSE(c) are the average root mean square errors of the parameter estimates. Sub(β) and Sub(c) indicate the fraction of times the estimated model contains all the relevant variables (in X and D) while Spar(β) and Spar(c) show how often exactly the correct subset of variables is chosen. Finally, $\#\beta$ and $\#c$ give the average number of non-zero βs and cs, respectively.

Experiments G-H are the truly high-dimensional ones where the number of variables is (much) larger than the sample size. Hence, we do not implement least squares using all variables. Experiment G illustrates a rather difficult setting with many heavy-tailed covariates. The Lasso does a decent job in reducing dimensionality without being overwhelming either. The average number of non-zero $\hat{\beta}$s is 36.97 which is still larger than the five true non-zero coefficients. The adaptive Lasso removes ten more variables without discarding (many) more relevant ones so the second step seems worth implementing.

In Experiment H the covariates are gaussian and the Lasso and the adaptive Lasso perform much better than in the heavy-tailed Experiment G. The estimation error of $\hat{\beta}$ is more than halved compared to Experiment G and all relevant variables are retained in the model. This does not come at the price of bigger models since the average number of non-zero coefficients is now smaller than before. The adaptive Lasso only classifies 17.64 βs as zero (of which five are truly non-zero) resulting in a significant dimension reduction.

Experiment I doubles the number of variables in X compared to Experiment H. In this light, it is reasonable that the estimation error of $\hat{\beta}$ roughly doubles. Almost all non-zero β^* are also classified as such but unfortunately, though not unexpectedly, the total number of βs classified as non-zero also roughly doubles. However, the adaptive Lasso still manages to reduce the number of variables to less than one tenth of the original number of variables.

7. Empirical illustration

In this section we illustrate the use of the panel (adaptive) Lasso on a large data set for the G8 countries. In particular, we try to determine which variables are relevant for explaining economic growth in these countries. The neoclassical growth model predicts that higher initial wealth should lead to lower growth rates. The primary mechanism

		MSE(β)	MSE(c)	Sub(β)	Sub(c)	Spar(β)	Spar(c)	$\#\beta$	$\#c$
Exp G	Lasso	1.73	1.42	0.67	0.30	0.00	0.05	36.97	2.77
	ALasso	1.66	1.51	0.62	0.26	0.05	0.05	26.93	2.45
	OLSO	0.37	0.67	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA								
Exp H	Lasso	0.87	1.05	1.00	0.49	0.01	0.24	24.46	2.28
	ALasso	0.66	0.94	1.00	0.48	0.20	0.25	17.54	2.13
	OLSO	0.22	0.40	1.00	1.00	1.00	1.00	5.00	2.00
	OLSA								
Exp I	Lasso	1.43	1.03	0.97	0.55	0.00	0.14	63.95	2.90
	ALasso	1.20	0.99	0.93	0.49	0.04	0.17	38.27	2.43
	OLSO	0.33	0.43	1.00	1.00	1.00	1.00	10.00	2.00
	OLSA								

TABLE 3. MSE(β) and MSE(c) are the average root mean square errors of the parameter estimates. Sub(β) and Sub(c) indicate the fraction of times the estimated model contains all the relevant variables (in X and D) while Spar(β) and Spar(c) show how often exactly the correct subset of variables is chosen. Finally, $\#\beta$ and $\#c$ give the average number of non-zero βs and cs, respectively.

behind this prediction is that countries with low capital to labor ratios tend to have a higher marginal return to capital, Barro (1991). In this section we shall investigate whether this prediction is true for some of the biggest economies in the world.

The data set has been obtained from the data bank of world development indicators. The panel that we analyse consists of 8 countries with 20 annual observations for each country for the period 1992-2011. The number of explanatory variables (excluding the eight individual effects dummies) is 161. Hence, the number of variables is large compared to the number of observations and the Lasso-type estimators come to use since they offer a non ad hoc way of choosing the variables. Put differently, one can handle a much larger conditioning set of variables than previous methods.

The variables cover broad categories such as economical, health, demographical and technological ones. The GDP level is treated specially in the sense that it enters the right hand side of the model with a lag of one year to enable us to test whether initial GDP is related (negatively) to GDP growth. All right hand side variables are standardised to have an ℓ_2-norm equal to the sample size. The Lasso as well as the adaptive Lasso are implemented by the `glmnet` as in the Monte Carlo section.

Table 4 contains the results of the estimation. In the first round $\lambda_{N,T}$ is chosen by BIC for the Lasso as well as the adaptive Lasso. Then it is gradually reduced by choosing decreasing fractions of this initial choice. This is done as a kind of sensitivity check to verify the robustness of the sparsity. As can be seen from Table 4 the Lasso and the adaptive Lasso indeed choose very sparse models when $\lambda_{N,T}$ is chosen by BIC. In particular, they include three and two variables, respectively. Note that all variables chosen are annual growth rates. This is sensible since we are trying to explain the *annual growth rate* of GDP. Furthermore, it is seen that initial GDP does not enter as an explanatory variable. Hence, we find no support for the neoclassical growth hypothesis. However, it should

be said that this hypothesis might be more relevant at explaining differences in growth between developed and less developed countries while all countries in our sample are rather developed. Furthermore, we use the GDP of the previous year as initial GDP which is a choice that might not leave enough time for the transmission mechanisms to function properly.

As can be expected, lowering $\lambda_{N,T}$ results in more variables being included in the model. This is manifested in Table 4 by the models becoming gradually larger as $\lambda_{N,T}$ is decreased. But only for $\lambda_{N,T} = 0.1 \cdot \lambda_{BIC}$ a dummy is included in the model by the Lasso (for the United Kingdom).

$\lambda_{N,T}$	Lasso	Adaptive Lasso
λ_{BIC}	Exports of goods and services (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)	Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)
$0.75 \cdot \lambda_{BIC}$	Exports of goods and services (annual % growth) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)	Exports of goods and services (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)
$0.5 \cdot \lambda_{BIC}$	Exports of goods and services (annual % growth) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth) Market capitalization of listed companies (% of GDP)	Exports of goods and services (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)
$0.25 \cdot \lambda_{BIC}$	Exports of goods and services (annual % growth) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth) Market capitalization of listed companies (% of GDP)	Exports of goods and services (annual % growth) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth)
$0.1 \cdot \lambda_{BIC}$	Exports of goods and services (annual % growth) Final consumption expenditure, etc. (annual % growth) Forest rents (% of GDP) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Household final consumption expenditure (annual % growth) Household final consumption expenditure, PPP (constant 2005 international $) Inflation, GDP deflator (annual %) Market capitalization of listed companies (% of GDP) UK dummy	Exports of goods and services (annual % growth) Final consumption expenditure, etc. (annual % growth) General government final consumption expenditure (annual % growth) Gross fixed capital formation (annual % growth) Inflation, GDP deflator (annual %)

TABLE 4. The table shows which variables were chosen by the Lasso and the adaptive Lasso for various choices of $\lambda_{N,T}$. λ_{BIC} indicates that $\lambda_{N,T}$ was chosen by BIC while the other sections of the table contain the results for $\lambda_{N,T}$ being a certain fraction of λ_{BIC}.

8. Conclusion

High-dimensional data is becoming increasingly available and one of the first choices one has to make when building a model is which variables to include. Furthermore, panel data models are a work horse tool for microeconometric analysis. For these reasons we have studied the performance of the panel Lasso and adaptive Lasso in high-dimensional panel data models. In particular, this paper has established finite sample upper bounds on the estimation error of the panel Lasso estimator that hold with high probability. We have also shown that the upper bounds are optimal in a sense made clear in Theorem 4. Conditions for consistency in even very high-dimensional models were also provided.

Next, the panel adaptive Lasso was analyzed and we gave lower bounds on the probability with which it selects the correct sign pattern in finite samples. These results were then used to deduce asymptotic results.

The results were proven under various assumptions on the moment/tail behavior of the covariates and the error terms. In particular we allowed for non-subgaussian behavior in some of our theorems.

The methods were then applied to finding the variables that explain growth in the G8 countries over the last 20 years. A rather sparse model was found to explain the growth.

In this paper we have used BIC to select the tuning parameters but ideally one would like a data driven way with theoretical guarantees. We leave this as an interesting avenue for future research.

9. Appendix

We start with the following Lemma which is similar in spirit to Lemma B.1 in Bickel et al. (2009).

Lemma 1. *On $\mathcal{A}_{N,T}$ the following inequalities are valid.*

(18)
$$\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2+\lambda_{N,T}\left\|\hat{\beta}-\beta^*\right\|_{\ell_1}+\mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1}\leq 4\lambda_{N,T}\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1}+4\mu_{N,T}\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1}$$

and

(19)
$$\lambda_{N,T}\left\|\hat{\beta}_{J_1^c}-\beta^*_{J_1^c}\right\|_{\ell_1}+\mu_{N,T}\left\|\hat{c}_{J_2^c}-c^*_{J_2^c}\right\|_{\ell_1}\leq 3\lambda_{N,T}\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1}+3\mu_{N,T}\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1}$$

Proof. By the minimizing property of $\hat{\gamma}$ it follows that

$$\left\|y-Z\hat{\gamma}\right\|^2+2\lambda_{N,T}\left\|\hat{\beta}\right\|_{\ell_1}+2\mu_{N,T}\left\|\hat{c}\right\|_{\ell_1}\leq\left\|y-Z\gamma^*\right\|^2+2\mu_{N,T}\left\|\beta^*\right\|_{\ell_1}+2\mu_{N,T}\left\|c^*\right\|_{\ell_1}$$

which, using that $y=Z\gamma^*+\epsilon$, yields

$$\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2-2\epsilon' Z(\hat{\gamma}-\gamma^*)+2\lambda_{N,T}\left\|\hat{\beta}\right\|_{\ell_1}+2\mu_{N,T}\left\|\hat{c}\right\|_{\ell_1}\leq 2\lambda_T\left\|\beta^*\right\|_{\ell_1}+2\mu_{N,T}\left\|c^*\right\|_{\ell_1}$$

Or, equivalently

$$\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2\leq 2\epsilon' Z(\hat{\gamma}-\gamma^*)+2\lambda_{N,T}\left(\left\|\beta^*\right\|_{\ell_1}-\left\|\hat{\beta}\right\|_{\ell_1}\right)+2\mu_{N,T}\left(\left\|c^*\right\|_{\ell_1}-\left\|\hat{c}\right\|_{\ell_1}\right) \tag{20}$$

So to bound $\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2$ one must bound $2\epsilon' Z(\hat{\gamma}-\gamma^*)$. Note that on $\mathcal{A}_{N,T}$ one has

$$\begin{aligned}2\epsilon' Z(\hat{\gamma}-\gamma^*) &= 2\epsilon' X(\hat{\beta}-\beta^*) + 2\epsilon' D(\hat{c}-c^*)\\ &\leq 2\left\|\epsilon' X\right\|_{\ell_\infty}\|\hat{\beta}-\beta^*\|_{\ell_1} + 2\left\|\epsilon' D\right\|_{\ell_\infty}\|\hat{c}-c^*\|_{\ell_1}\\ &\leq \lambda_{N,T}\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} + \mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1}\end{aligned}$$

Putting things together, on $\mathcal{A}_{N,T}$,

$$\begin{aligned}&\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2\\ &\leq \lambda_{N,T}\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} + 2\lambda_{N,T}\left(\left\|\beta^*\right\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1}\right) + \mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1} + 2\mu_{N,T}\left(\|c^*\|_{\ell_1} - \|\hat{c}\|_{\ell_1}\right)\end{aligned}$$

Adding $\lambda_{N,T}\left\|\hat{\beta}-\beta^*\right\|_{\ell_1}$ and $\mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1}$ yields

$$\begin{aligned}&\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2 + \lambda_{N,T}\left\|\hat{\gamma}-\gamma^*\right\|_{\ell_1} + \mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1}\\ &\leq 2\lambda_{N,T}\left(\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} + \left\|\beta^*\right\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1}\right) + 2\mu_{N,T}\left(\|\hat{c}-c^*\|_{\ell_1} + \|c^*\|_{\ell_1} - \|\hat{c}\|_{\ell_1}\right)\end{aligned} \tag{21}$$

Notice that

$$\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} + \left\|\beta^*\right\|_{\ell_1} - \left\|\hat{\beta}\right\|_{\ell_1} = \left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1} + \left\|\beta^*_{J_1}\right\|_{\ell_1} - \left\|\hat{\beta}_{J_1}\right\|_{\ell_1}$$

In addition, $\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1} + \left\|\beta^*_{J_1}\right\|_{\ell_1} - \left\|\hat{\beta}_{J_1}\right\|_{\ell_1} \leq 2\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1}$ by continuity of the norm. By exactly the same arguments $\|\hat{c}-c^*\|_{\ell_1} + \|c^*\|_{\ell_1} - \|\hat{c}\|_{\ell_1} \leq 2\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1}$. Using these estimates in (21) yields inequality (18). Next notice that (18) gives

$$\lambda_{N,T}\left\|\hat{\beta}-\beta^*\right\|_{\ell_1} + \mu_{N,T}\left\|\hat{c}-c^*\right\|_{\ell_1} \leq 4\lambda_{N,T}\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1} + 4\mu_{N,T}\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1}$$

which is equivalent to

$$\lambda_{N,T}\left\|\hat{\beta}_{J_1^c}-\beta^*_{J_1^c}\right\|_{\ell_1} + \mu_{N,T}\left\|\hat{c}_{J_2^c}-c^*_{J_2^c}\right\|_{\ell_1} \leq 3\lambda_{N,T}\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1} + 3\mu_{N,T}\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1}$$

and establishes inequality (19). □

Proof of Theorem 1. By (18) of Lemma 1 (which is valid on $\mathcal{A}_{N,T}$)

$$\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2 \leq 4\lambda_{N,T}\left\|\hat{\beta}_{J_1}-\beta^*_{J_1}\right\|_{\ell_1} + 4\mu_{N,T}\left\|\hat{c}_{J_2}-c^*_{J_2}\right\|_{\ell_1} \tag{22}$$

Next, note that for $b = S^{-1}\delta$ where b is partitioned as $b = (b^{1\prime}, b^{2\prime})'$ with b^1 being a $p\times 1$ vector and b^2 an $N\times 1$ vector, the restricted eigenvalue condition (4) may be formulated equivalently as

$$\begin{aligned}\kappa^2_{\psi_{N,T}}(r_1,r_2) = \min\Bigg\{&\frac{\|Zb\|^2}{\|Sb\|^2} : b\in\mathbb{R}^{p+N}\setminus\{0\},\ |R_1|\leq r_1,\ |R_2|\leq r_2,\\ &\lambda_{N,T}\left\|b^1_{R_1^c}\right\|_{\ell_1} + \mu_{N,T}\left\|b^2_{R_2^c}\right\|_{\ell_1} \leq 3\lambda_{N,T}\left\|b^1_{R_1}\right\|_{\ell_1} + 3\mu_{N,T}\left\|b^2_{R_2}\right\|_{\ell_1}\Bigg\} > 0\end{aligned}$$

Hence, the restricted eigenvalue condition (which is applicable due to (19) yields

$$\left\|Z(\hat{\gamma}-\gamma^*)\right\|^2 \geq \kappa^2_{\Psi_{N,T}}\left\|S(\hat{\gamma}-\gamma^*)\right\|^2 \geq \kappa^2/2\left[NT\left\|\hat{\beta}-\beta^*\right\|^2 + T\left\|\hat{c}-c^*\right\|^2\right] \tag{23}$$

where the last estimate holds on $\mathcal{B}_{N,T}$. By Jensen's inequality

$$4\lambda_{N,T}\|\hat{\beta}_{J_1}-\beta^*_{J_1}\|_{\ell_1}+4\mu_{N,T}\|\hat{c}_{J_2}-c^*_{J_2}\|_{\ell_1}\leq 4\lambda_{N,T}\sqrt{s_1}\|\hat{\beta}_{J_1}-\beta^*_{J_1}\|+4\mu_{N,T}\sqrt{s_2}\|\hat{c}_{J_2}-c^*_{J_2}\|$$

$$\leq 4\lambda_{N,T}\sqrt{s_1}\|\hat{\beta}-\beta^*\|+4\mu_{N,T}\sqrt{s_2}\|\hat{c}-c^*\| \tag{24}$$

Inserting (23) and (24) into (22) yields

$$\frac{\kappa^2}{2}\left[NT\|\hat{\beta}-\beta^*\|^2+T\|\hat{c}-c^*\|^2\right]\leq 4\lambda_{N,T}\sqrt{s_1}\|\hat{\beta}-\beta^*\|+4\mu_{N,T}\sqrt{s_2}\|\hat{c}-c^*\|$$

or equivalently,

$$\|\hat{\beta}-\beta^*\|^2+\frac{1}{N}\|\hat{c}-c^*\|^2-\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}\|\hat{\beta}-\beta^*\|-\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 NT}\|\hat{c}-c^*\|\leq 0$$

For $x=\|\hat{\beta}-\beta^*\|$ and $y=\|\hat{c}-c^*\|$ this can be written as a quadratic inequality in two variables:

$$x^2-ax+by^2-cy\leq 0,\ x,y\geq 0 \tag{25}$$

[4] with $a=\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}$, $b=\frac{1}{N}$ and $c=\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 NT}$. First bound $x=\|\hat{\beta}-\beta^*\|$. For every y the values of x that satisfy (25) form an interval in $\mathbb{R}_+$. The right end point of this interval is the desired upper bound on x. Clearly this right end point is a decreasing function in by^2-cy. Hence, we first minimize the polynomial by^2-cy. This yields $y=\frac{c}{2b}$ and the corresponding value of by^2-cy is $-\frac{c^2}{4b}$. Hence, our desired upper bound on x is the largest solution of $x^2-ax-\frac{c^2}{4b}\leq 0$. By the standard solution formula for the roots of a quadratic polynomial this yields

$$\|\hat{\beta}-\beta^*\|=x\leq\frac{a+\sqrt{a^2+c^2/b}}{2} \tag{26}$$

Switching the roles of x and y, one gets a similar bound on $y=\|\hat{c}-c^*\|$, namely

$$\|\hat{c}-c^*\|=y\leq\frac{c+\sqrt{c^2+ba^2}}{2b} \tag{27}$$

Inserting the definitions of a,b and c into (26) yields

$$\|\hat{\beta}-\beta^*\|\leq\frac{\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}+\sqrt{\left(\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}\right)^2+\left(\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 NT}\right)^2 N}}{2}\leq\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}+\frac{4\mu_{N,T}\sqrt{s_2}}{\kappa^2\sqrt{N}T}$$

by subadditivity of $x\mapsto\sqrt{x}$. Similarly,

$$\|\hat{c}-c^*\|\leq\frac{\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 NT}+\sqrt{\left(\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 NT}\right)^2+\frac{1}{N}\left(\frac{8\lambda_{N,T}\sqrt{s_1}}{\kappa^2 NT}\right)^2}}{2/N}\leq\frac{8\mu_{N,T}\sqrt{s_2}}{\kappa^2 T}+\frac{4\lambda_{N,T}\sqrt{s_1}}{\kappa^2\sqrt{N}T}$$

□

Before stating the next lemma we shall remark that when no further distinction between subscripts i and t is needed we shall sometimes use $x_{j,k}$ to denote the jth entry of the kth variable $x_k=(x_{1,1,k},x_{1,2,k},...,x_{1,T,k},x_{2,1,k},...,x_{N,T,k})'$ with $1\leq j\leq NT$. Similarly, we will write ϵ_j for the jth entry of $\epsilon=(\epsilon_{1,1},\epsilon_{1,2},...,\epsilon_{1,T},\epsilon_{2,1},...,\epsilon_{N,T})'$ $1\leq j\leq NT$ where

[4] Note that this inequality is trivially satisfied by $x=y=0$, corresponding to no estimation error. However, we are looking for an upper bound on x and y.

Lemma 2. *Let* $\lambda_{N,T} = 4a_{N,T}p^{1/r}(NT)^{1/2}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}$ *and* $\mu_{N,T} = 4a_{N,T}N^{1/r}T^{1/2}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}$ *for some sequence* $a_{N,T}$*. Then, under assumption A1) and A2a)*

$$P\left(\mathcal{A}_{N,T}^c\right) = P\left(\left\{\|X'\epsilon\|_{\ell_\infty} > \frac{\lambda_{N,T}}{2}\right\}\cup\left\{\|D'\epsilon\|_{\ell_\infty} > \frac{\mu_{N,T}}{2}\right\}\right) \leq 2\left(\frac{C_r}{a_{N,T}}\right)^r \tag{28}$$

Proof of Lemma 2. First bound $\left\|\max_{1\leq k\leq p}\left|\sum_{j=1}^{NT}x_{j,k}\epsilon_j\right|\right\|_{L_r}$. To this end, note that for any collection of random variables $\{U_k\}_{k=1}^p \subseteq L_r$,

$$\left\|\max_{1\leq k\leq p}U_k\right\|_{L_r} = [E(|\max_{1\leq k\leq p}U_k|^r)]^{1/r} \leq \left[E\left(\sum_{k=1}^p|U_k|^r\right)\right]^{1/r} \leq p^{1/r}\max_{1\leq k\leq p}\|U_k\|_{L_r}$$

Next, bound $\left\|\sum_{j=1}^{NT}x_{j,k}\epsilon_j\right\|_{L_r}$ uniformly in $1\leq k\leq p$. Denote by $\mathcal{F}_n = \sigma\left(\left\{X,\epsilon_j,\ 1\leq j\leq n\right\}\right)$ the σ-field generated by X and $\epsilon_j,\ 1\leq j\leq n$ and set $S_{n,k} = \sum_{j=1}^n x_{j,k}\epsilon_j$. Then $\{(S_{n,k},\mathcal{F}_n),\ 1\leq n\leq NT\}$ is a martingale for all $1\leq k\leq p$ under assumptions A1 and the given moment assumptions. Hence, by Rosenthal's inequality for martingales (see Hitczenko (1990) or Hall and Heyde (1980)) for a constant C_r depending only on r,[5]

$$\begin{aligned}
\left\|\sum_{j=1}^{NT}x_{j,k}\epsilon_j\right\|_{L_r} &\leq C_r\left[\left(E\left(\sum_{j=1}^{NT}E(x_{j,k}^2\epsilon_j^2|\mathcal{F}_{j-1})\right)^{r/2}\right)^{1/r} + \left(E\left[\max_{1\leq j\leq NT}|x_{j,k}\epsilon_j|^r\right]\right)^{1/r}\right]\\
&\leq C_r\left[\left(E\left(\sum_{j=1}^{NT}x_{j,k}^2\|\epsilon_j\|_{L_2}^2\right)^{r/2}\right)^{1/r} + \left(E\left[\sum_{j=1}^{NT}|x_{j,k}\epsilon_j|^r\right]\right)^{1/r}\right]\\
&\leq C_r\left[\left((NT)^{r/2-1}\sum_{j=1}^{NT}E|x_{j,k}|^r\|\epsilon_j\|_{L_2}^r\right)^{1/r} + (NT)^{1/r}\max_{1\leq t\leq T}\|x_{1,t,k}\|_{L_r}\|\epsilon_{1,t}\|_{L_r}\right]\\
&\leq C_r\left[\left((NT)^{r/2-1}NT\max_{1\leq t\leq T}E|x_{1,t,k}|^r\|\epsilon_{1,t}\|_{L_2}^r\right)^{1/r} + (NT)^{1/r}\max_{1\leq t\leq T}\|x_{1,t,k}\|_{L_r}\|\epsilon_{1,t}\|_{L_r}\right]\\
&\leq C_r\left[(NT)^{1/2}\max_{1\leq t\leq T}\|x_{1,t,k}\|_{L_r}\|\epsilon_{1,t}\|_{L_2} + (NT)^{1/r}\max_{1\leq t\leq T}\|x_{1,t,k}\|_{L_r}\|\epsilon_{1,t}\|_{L_r}\right]\\
&\leq 2C_r(NT)^{1/2}\max_{1\leq t\leq T}\|\epsilon_{1,t}\|_{L_r}
\end{aligned}$$

In the above display we have used Loeve's c_r-inequality and by Hitczenko (1990) we know that $C_r\leq 10r$. Hitczenko (1990) actually shows that the optimal constant $C_r\in O(r/\ln(r))$

[5]By independence of $x_{j,k}$ and ϵ_j their product is in L_r and Rosenthal's inequality yields a nontrivial upper bound.

as $r \to \infty$. Hence,

$$\left\| \max_{1\leq k\leq p} \left| \sum_{j=1}^{NT} x_{j,k}\epsilon_j \right| \right\|_{L_r} \leq \max_{1\leq k\leq p} p^{1/r} \left\| \sum_{j=1}^{NT} x_{j,k}\epsilon_j \right\|_{L_r} \leq p^{1/r} 2C_r (NT)^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r}$$

By Markov's inequality,

$$P\left(\max_{1\leq k\leq p} \left| \sum_{j=1}^{NT} x_{j,k}\epsilon_j \right| > \frac{\lambda_{N,T}}{2} \right) \leq \frac{1}{\left(\lambda_{N,T} / (4p^{1/r} C_r (NT)^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r}) \right)^r} = \left(\frac{C_r}{a_{N,T}} \right)^r$$

In a similar way as above it follows by Rosenthal's inequality

$$\begin{aligned} \left\| \sum_{t=1}^{T} \epsilon_{i,t} \right\|_{L_r} &\leq C_r \left[\left(\sum_{t=1}^{T} E(\epsilon_{i,t}^2) \right)^{1/2} + \left(E\left(\max_{1\leq t\leq T} |\epsilon_{i,t}|^r \right) \right)^{1/r} \right] \\ &\leq C_r \left[T^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_2} + T^{1/r} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r} \right] \\ &\leq 2C_r T^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r} \end{aligned}$$

This implies that

$$\left\| \max_{1\leq i\leq N} \sum_{t=1}^{T} \epsilon_{i,t} \right\|_{L_r} \leq \max_{1\leq i\leq N} N^{1/r} \left\| \sum_{t=1}^{T} \epsilon_{i,t} \right\|_{L_r} \leq N^{1/r} 2C_r T^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r}$$

And so, by Markov's inequality,

$$P\left(\max_{1\leq i\leq N} \sum_{t=1}^{T} \epsilon_{i,t} > \frac{\mu_{N,T}}{2} \right) \leq \frac{1}{\left(\mu_{N,T} / (4N^{1/r} C_r T^{1/2} \max_{1\leq t\leq T} \left\| \epsilon_{1,t} \right\|_{L_r}) \right)^r} = \left(\frac{C_r}{a_{N,T}} \right)^r$$

It follows that

$$P\left(\left\{ \left\| X'\epsilon \right\|_{\ell_\infty} > \frac{\lambda_{N,T}}{2} \right\} \cup \left\{ \left\| D'\epsilon \right\|_{\ell_\infty} > \frac{\mu_{N,T}}{2} \right\} \right) \leq 2 \left(\frac{C_r}{a_{N,T}} \right)^r$$

□

Lemma 3. *Let $\{U_i, \mathcal{F}_i\}_{i=1}^n$ be a martingale difference sequence and assume that there exist $\delta, M > 0$ such that $E \exp(\delta |U_i|) \leq M$ for all $i = 1, ..., n$. Then, there exists positive constants A and B such that for all $x \geq a/\sqrt{n}$*

$$P\left(\left| \sum_{i=1}^{n} U_i \right| > nx \right) < Ae^{-B(x^2 n)^{1/3}} \tag{29}$$

Proof. In the proof of their Theorem 3.2 Lesigne and Volnỳ (2001) show that if $E\exp(|U_i|) \leq M$ for all $i = 1, ..., n$, then for any $x > 0$ and $t \in (0, 1)$ one has[6]

$$P\left(\left|\sum_{i=1}^{n} U_i\right| > nx\right)$$

(30)

$$< \left(2 + \frac{M}{(1-t)^2}\left[\frac{1}{4}t^{4/3}(x^{-2}n^{-1})^{1/3} + t^{2/3}(x^{-2}n^{-1})^{2/3} + 2x^{-2}n^{-1}\right]\right) e^{-(1/2)t^{2/3}(x^2n)^{1/3}}$$

But note that $P\left(|\sum_{i=1}^{n} U_i| > nx\right) = P\left(|\sum_{i=1}^{n}(\delta U_i)| > n(\delta x)\right)$ where $\{\delta U_i\}_{i=1}^{n}$, by assumption, now satisfy the conditions of Theorem 3.2 in Lesigne and Volnỳ (2001) and so replacing x by δx in (30) yields

$$P\left(\left|\sum_{i=1}^{n} U_i\right| > nx\right)$$

$$< \left(2 + \frac{M}{(1-t)^2}\left[\frac{1}{4}t^{4/3}\delta^{-2/3}(x^{-2}n^{-1})^{1/3} + t^{2/3}\delta^{-4/3}(x^{-2}n^{-1})^{2/3} + 2\delta^{-2}x^{-2}n^{-1}\right]\right) e^{-(1/2)t^{2/3}\delta^{2/3}(x^2n)^{1/3}}$$

Restricting x to be greater than $a/\sqrt{n}$, implying that $x^{-2}n^{-1} \leq 1/a^2$, and using that M, t and δ are constants the conclusion of the lemma follows. □

For the proof of Lemma 4 below, we shall use Orlicz norms as defined in Van Der Vaart and Wellner (1996): Let ψ be a non-decreasing convex function with $\psi(0) = 0$. Then, the Orlicz norm of a random variable X is given by

$$\|X\|_{\psi} = \inf\left\{C > 0 : E\psi\left(|X|/C\right) \leq 1\right\}$$

where, as usual, $\inf \emptyset = \infty$. We will use Orlicz norms for $\psi(x) = \psi_p(x) = e^{x^p} - 1$ for $p = 1, 2$.

Lemma 4. *Assume that assumptions A1 and A2b are satisfied. Then, for $a_{N,T} \geq e$*

$$P\left(\|X'\epsilon\|_{\ell_\infty} \geq \lambda_{N,T}/2\right) \leq Ap^{1-B\log(a_{N,T})} \text{ for } \lambda_{N,T} = \sqrt{4NT\log(p)^3\log(a_{N,T})^3}$$

and

$$P\left(\|D'\epsilon\|_{\ell_\infty} \geq \mu_{N,T}/2\right) \leq AN^{1-B\ln(a_{N,T})} \text{ for } \mu_{N,T} = \sqrt{4T\log(N)^3\log(a_{N,T})^3}$$

Proof. First note that for all $1 \leq j \leq NT$ and $1 \leq k \leq p$ one has for all $t > 0$

$$P\left(|x_{j,k}\epsilon_j| > t\right) \leq P\left(|x_{j,k}| > \sqrt{t}\right) + P\left(|\epsilon_j| > \sqrt{t}\right) \leq K\exp(-Ct)$$

and so it follows from Lemma 2.2.1 in Van Der Vaart and Wellner (1996) that $\|x_{j,k}\epsilon_j\|_{\psi_1} \leq \frac{1+K}{C}$ and so $E\exp\left(\frac{C}{1+K}|x_{j,k}\epsilon_j|\right) \leq 2$ by the definition of the Orlicz-norm. Hence, $\delta = \frac{C}{1+K}$ works in Lemma 3 for all $1 \leq k \leq p$. Next, denote by $\mathcal{F}_n = \sigma\left(\{X, \epsilon_j,\ 1 \leq j \leq n\}\right)$ the σ-field generated by X and $\epsilon_j,\ 1 \leq j \leq n$ and set $S_{n,k} = \sum_{j=1}^{n} x_{j,k}\epsilon_j$. Then it is clear that $\{(S_{n,k}, \mathcal{F}_n),\ 1 \leq n \leq NT\}$ is a martingale for all $1 \leq k \leq p$. From a union bound it follows from Lemma 3 (with $a = 1$) that[7]

$$P\left(\|X'\epsilon\|_{\ell_\infty} \geq \lambda_{N,T}/2\right) = P\left(\|X'\epsilon\|_{\ell_\infty} \geq \frac{\lambda_{N,T}/2}{NT}NT\right) \leq pAe^{-B\left(\frac{\lambda_{NT}^2}{4NT}\right)^{1/3}} = Ap^{1-B\log(a_{N,T})}$$

[6]See the last expression in the proof of their Theorem 3.2.

[7]Lemma **?** is applicable since $a_{N,T}$ and p are assumed greater than e.

Next, by the subgaussianity of $\epsilon_{i,t}$, $1 \leq i \leq N$, $1 \leq t \leq T$, it follows from Lemma 2.2.1 in Van Der Vaart and Wellner (1996) that $\|\epsilon_{i,t}\|_{\psi_2} \leq \left(\frac{1+K/2}{C}\right)^{1/2}$, and so $\|\epsilon_{i,t}\|_{\psi_1} \leq \left(\frac{1+K/2}{C}\right)^{1/2} \log(2)^{-1/2}$ by the second to last inequality on page 95 in Van Der Vaart and Wellner (1996). Hence, $E\exp\left(\left(\frac{C}{1+K/2}\right)^{1/2}\log(2)^{1/2}|\epsilon_{i,t}|\right) \leq 2$.[8] Furthermore, for all $i = 1, ..., N$, $\{\epsilon_{i,t}\}_{t=1}^{T}$ are independent and so by the union bound and Lemma 3 [9] (with $a = 1$)

$$P\left(\|D'e\|_{\ell_\infty} \geq \mu_{N,T}/2\right) \leq NP\left(\|D'e\|_{\ell_\infty} \geq \frac{\mu_{N,T}/2}{T}T\right) \leq NAe^{-B\left(\frac{\mu_{N,T}^2}{4T}\right)^{1/3}} \leq AN^{1-B\log(a_{N,T})}$$

□

Lemma 5. *Let A and B be two positive semi-definite $(p+N)\times(p+N)$ matrices and assume that A satisfies the restricted eigenvalue condition RE(s_1, s_2) for some κ_A. Then, for $\delta = \max_{1\leq i,j\leq p+N}|A_{i,j} - B_{i,j}|$, one also has $\kappa_B^2 \geq \kappa_A^2 - 16\delta(s_1+s_2)m_{N,T}^2$ where $m_{N,T} = \frac{\lambda_{N,T}}{\sqrt{N}\mu_{N,T}} \vee \frac{\sqrt{N}\mu_{N,T}}{\lambda_{N,T}}$*

Proof. The proof is similar to Lemma 10.1 in Van De Geer and Bühlmann (2009). Let x_1 be $p\times 1$, x_2 be $N\times 1$ and define $x = (x_1', x_2')'$ and assume that $\frac{\lambda_{N,T}}{\sqrt{NT}}\|x_{1J_1^c}\|_{\ell_1} + \frac{\mu_{N,T}}{\sqrt{T}}\|x_{2J_2^c}\|_{\ell_1} \leq 3\frac{\lambda_{N,T}}{\sqrt{NT}}\|x_{1J_1}\|_{\ell_1} + 3\frac{\mu_{N,T}}{\sqrt{T}}\|x_{2J_2}\|_{\ell_1}$. Defining

$$V = \begin{pmatrix} \frac{\lambda_{N,T}}{\sqrt{NT}}I_{|J_1|} & 0 \\ 0 & \frac{\mu_{N,T}}{\sqrt{T}}I_{|J_2|} \end{pmatrix} \text{ and } V_c = \begin{pmatrix} \frac{\lambda_{N,T}}{\sqrt{NT}}I_{|J_1^c|} & 0 \\ 0 & \frac{\mu_{N,T}}{\sqrt{T}}I_{|J_2^c|} \end{pmatrix}$$

this can also be expressed as $\|V_c x_{J^c}\|_{\ell_1} \leq 3\|Vx_J\|_{\ell_1}$. For any (non-zero) $(p+N)\times 1$ vector x satisfying this restriction one has

$$\|x_{J^c}\|_{\ell_1} = \|V_c^{-1}V_c x_{J^c}\|_{\ell_1} \leq \|V_c^{-1}\|_{\ell_1}\|V_c x_{J^c}\|_{\ell_1} \leq 3\|V_c^{-1}\|_{\ell_1}\|Vx_J\|_{\ell_1} \leq 3\|V_c^{-1}\|_{\ell_1}\|V\|_{\ell_1}\|x_J\|_{\ell_1}$$

Since

$$\|V_c^{-1}\|_{\ell_1}\|V\|_{\ell_1} = \frac{\lambda_{N,T}}{\sqrt{N}\mu_{N,T}} \vee \frac{\sqrt{N}\mu_{N,T}}{\lambda_{N,T}} = m_{N,T}$$

one gets

$$\begin{aligned}|x'Ax - x'Bx| = |x'(A-B)x| &\leq \|x\|_{\ell_1}\|(A-B)x\|_{\ell_\infty} \leq \delta\|x\|_{\ell_1}^2 \leq \delta(\|x_J\|_{\ell_1} + \|x_{J^c}\|_{\ell_1})^2 \\ &\leq \delta(1+3m_{N,T})^2\|x_J\|_{\ell_1}^2 \leq 16\delta(s_1+s_2)m_{N,T}^2\|x_J\|^2 \leq 16\delta(s_1+s_2)m_{N,T}^2\|x\|^2\end{aligned}$$

where the last estimate follows from the fact that $m_{N,T} \geq 1$ and Jensen's inequality. Hence,

$$x'Bx \geq x'Ax - 16\delta(s_1+s_2)m_{N,T}^2\|x\|^2$$

[8] We note that this estimate is slightly suboptimal since we are not taking full advantage of the subgaussianity of the $\epsilon_{i,t}$ by merely using it to deduce subexponentiality and then invoking Lemma Lesigne and Volnỳ (2001). One could use the full strength of the subgaussianity by strengthening $E\exp(|\epsilon|) \leq K$ to $E\exp(\epsilon^2) \leq K$ in Lemma 3.2 of Lesigne and Volnỳ (2001). Doing so, and adjusting Lemma 3 accordingly yields that the exponent 1/3 in (29) can be increased to 1/2 and hence $\mu_{N,T}$ can in turns be reduced to $\sqrt{4T\log(N)^2\log(a_{N,T})^2}$. As a third route, one could use Hoeffding's inequality in combination with a truncation of the $\epsilon_{i,t}$. This does not reduce $\mu_{N,T}$ significantly either.

[9] In principle, the constants A and B need not be the same as above but by simply using the worst ones they can be chosen to be identical. Also, we have used $a_{N,T}, N \geq e$.

or equivalently,

$$\frac{x'Bx}{x'x} \geq \frac{x'Ax}{x'x} - 16\delta(s_1+s_2)m_{N,T}^2 \geq \kappa_A^2 - 16\delta(s_1+s_2)m_{N,T}^2$$

Minimizing the left hand side over $\{x \in \mathbb{R}^{p+N} \setminus \{0\} : \|V_c x_{J^c}\|_{\ell_1} \leq 3\|Vx_J\|_{\ell_1}\}$ yields the claim. □

In the following two lemmas we shall use Lemma 5 with

$$A = \Gamma = \begin{pmatrix} E\left(\frac{X'X}{NT}\right) & 0 \\ 0 & I_N \end{pmatrix} \text{ and } B = \Psi_{N,T} = \begin{pmatrix} \frac{X'X}{NT} & \frac{X'D}{\sqrt{N}T} \\ \frac{D'X}{\sqrt{N}T} & I_N \end{pmatrix}$$

in order to establish that Ψ_T satisfies the restricted eigenvalue condition with high probability. Furthermore, define

$$\tilde{\mathcal{B}}_{N,T} = \left\{ \max_{1\leq i,j\leq p+N} |\Psi_{T,i,j} - \Gamma_{i,j}| \leq \frac{\kappa^2}{32(s_1+s_2)m_{N,T}^2} \right\} \tag{31}$$

Lemma 6. *Under assumptions A1 and A2a,* $P\left(\kappa_{\Psi_T}^2 \geq \kappa^2/2\right) \geq P(\tilde{\mathcal{B}}_{N,T}) \geq 1 - D_r \frac{(p^2+Np)(s_1+s_2)^{r/2}(\frac{p}{N}\vee\frac{N}{p})}{\kappa^r N^{r/4}}$ *for a constant* D_r *only depending on* r*.*

Proof. By Lemma 5 it follows that $\kappa_{\Psi_{N,T}} \geq \kappa^2/2$ on $\tilde{\mathcal{B}}_{N,T}$ Since the lower right $N \times N$ blocks of $\Psi_{N,T}$ and Γ are identical it suffices to bound the entries of $\frac{X'X}{NT} - E\left(\frac{X'X}{NT}\right)$ and $\frac{X'D}{\sqrt{N}T}$. A typical element of $\frac{X'X}{NT} - E\left(\frac{X'X}{NT}\right)$ is of the form $\frac{1}{N}\sum_{i=1}^N \left(\frac{1}{T}\sum_{t=1}^T \left[x_{i,t,k}x_{i,t,l} - E(x_{i,t,k}x_{i,t,l})\right]\right)$ for some $k,l \in \{1,...,p\}$. Next note that for any sequence of mean zero i.i.d. variables $Z_1,...,Z_N$ in L_r it follows from Rosenthal's inequality that

$$\left\|\sum_{i=1}^N Z_i\right\|_{L_r} \leq C_r\left(\left[\sum_{i=1}^N EZ_i^2\right]^{1/2} + \left[E\max_{1\leq i\leq N}|Z_i|^r\right]^{1/r}\right) \leq C_r\left(N^{1/2}\|Z_1\|_{L_2} + N^{1/r}\|Z_1\|_{L_r}\right)$$

$$\leq 2C_r N^{1/2}\|Z_1\|_{L_r} \tag{32}$$

Furthermore,

$$\begin{aligned}\left\|\frac{1}{T}\sum_{t=1}^T \left[x_{1,t,k}x_{1,t,l} - E(x_{1,t,k}x_{1,t,l})\right]\right\|_{L_{r/2}} &\leq \max_{1\leq t\leq T}\left\|x_{1,t,k}x_{1,t,l} - E(x_{1,t,k}x_{1,t,l})\right\|_{L_{r/2}} \\ &\leq 2\max_{1\leq t\leq T}\left\|x_{1,t,k}x_{1,t,l}\right\|_{L_{r/2}} \leq 2\end{aligned}$$

where the last estimate follows from the Cauchy-Schwarz inequality. Using this in (32) (with r replaced by $r/2$) yields

$$\left\|\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{T}\sum_{t=1}^T\left[x_{i,t,k}x_{i,t,l} - E(x_{i,t,k}x_{i,t,l})\right]\right)\right\|_{L_{r/2}} \leq 4C_{r/2}N^{-1/2}.$$

Markov's inequality yields that for any $\epsilon > 0$

$$P\left(\left|\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{T}\sum_{t=1}^T\left[x_{i,t,k}x_{i,t,l} - E(x_{i,t,k}x_{i,t,l})\right]\right)\right| > \epsilon\right) \leq \frac{(4C_{r/2})^{r/2}}{\epsilon^{r/2}N^{r/4}} \tag{33}$$

Next, consider a typical term in $\frac{X'D}{\sqrt{N}T}$. Such a term is on the form $\frac{\sum_{t=1}^{T} x_{i,t,k}}{\sqrt{N}T}$ for $i = 1, ..., N$ and $k = 1, ..., p$. Since

$$\left\| \frac{1}{\sqrt{N}T} \sum_{t=1}^{T} x_{i,t,k} \right\|_{L_r} \leq \frac{1}{\sqrt{N}} \max_{1 \leq t \leq T} \left\| x_{i,t,k} \right\|_{L_r} \leq \frac{1}{\sqrt{N}}$$

it follows by Markov's inequality that for any $\epsilon > 0$

$$P\left(\left| \frac{1}{\sqrt{N}T} \sum_{t=1}^{T} x_{i,t,k} \right| > \epsilon \right) \leq \frac{1}{\epsilon^r N^{r/2}} = \frac{1}{(\epsilon^{r/2} N^{r/4})^2} \tag{34}$$

Combining (33) and (34) yields via a union bound over $(p^2 + Np)$ terms

$$P\left(\max_{1 \leq i,j \leq p+N} |A_{i,j} - B_{i,j}| > \epsilon \right) \leq (p^2 + Np) \left(\frac{(4C_{r/2})^{r/2}}{\epsilon^{r/2} N^{r/4}} \vee \frac{1}{(\epsilon^{r/2} N^{r/4})^2} \right) \leq D_r \frac{p^2 + Np}{\epsilon^{r/2} N^{r/4}}$$

[10]where the last estimate follows from the fact that without loss of generality (since otherwise the upper bound is greater than one) one may assume $\epsilon^{r/2} N^{r/4} \geq 1$ and so $\epsilon^{r/2} N^{r/4} \leq (\epsilon^{r/2} N^{r/4})^2$. $D_r = ([4C_{r/2}]^{r/2} \vee 1)$ is a constant only depending on r. Using $\epsilon = \frac{\kappa^2}{32(s_1+s_2)m_{N,T}^2}$ yields the lemma upon noting that $m_{N,T} = \left(\frac{p}{N} \vee \frac{N}{p} \right)^{1/r}$ and merging all constants into D_r. □

Lemma 7. *Let* $t = \frac{\kappa^2}{(s_1+s_2)\left(\frac{\ln(p)}{\ln(N)} \vee \frac{\ln(N)}{\ln(p)} \right)^3}$ *and let* $Nt^2 \geq 1$. *Then, under assumptions A1 and A2b),* $P\left(\kappa_{\Psi_T}^2 \geq \kappa^2/2 \right) \geq P(\tilde{\mathcal{B}}_{N,T}) \geq 1 - A(p^2 + Np)e^{-B(t^2 N)^{1/3}}$ *for absolute constants A and B.*

Proof. By Lemma 5 it follows that $\kappa_{\Psi_{N,T}}^2 \geq \kappa^2/2$ on $\tilde{\mathcal{B}}_{N,T}$. Since the lower right $N \times N$ blocks of $\Psi_{N,T}$ and Γ are identical it suffices to bound the entries of $\frac{X'X}{NT} - E\left(\frac{X'X}{NT} \right)$ and $\frac{X'D}{\sqrt{N}T}$. A typical element of $\frac{X'X}{NT} - E\left(\frac{X'X}{NT} \right)$ is of the form $\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{T} \sum_{t=1}^{T} \left[x_{i,t,k} x_{i,t,l} - E(x_{i,t,k} x_{i,t,l}) \right] \right)$ for some $k, l \in \{1, ..., p\}$. First, note that for all $1 \leq i \leq N$, $1 \leq t \leq T$ and $1 \leq k, l \leq p$ one has for all $\epsilon > 1/\sqrt{N}$

$$P\left(|x_{i,t,k} x_{i,t,l}| > \epsilon \right) \leq P\left(|x_{i,t,k} > \sqrt{\epsilon} \right) + P\left(|x_{i,t,l}| > \sqrt{\epsilon} \right) \leq K \exp(-C\epsilon)$$

and so it follows from Lemma 2.2.1 in Van Der Vaart and Wellner (1996) that $\left\| x_{i,t,k} x_{i,t,l} \right\|_{\psi_1} \leq \frac{1+K}{C}$. Next, note that by subadditivity of the Orlicz norm and Jensen's inequality

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \left[x_{i,t,k} x_{i,t,l} - E(x_{i,t,k} x_{i,t,l}) \right] \right\|_{\psi_1} \leq 2 \max_{1 \leq t \leq T} \left\| x_{i,t,k} x_{i,t,l} \right\|_{\psi_1} \leq 2 \frac{1+K}{C}$$

[10]Note that the first estimate in the display may be replaced by the slightly sharper estimate

$$P\left(\max_{1 \leq i,j \leq p+N} |A_{i,j} - B_{i,j}| > \epsilon \right) \leq p^2 \frac{(4C_{r/2})^{r/2}}{\epsilon^{r/2} N^{r/4}} + Np \frac{1}{(\epsilon^{r/2} N^{r/4})^2}$$

However, for $p \geq N$ this will lead to no improvement asymptotically, while the improvement is minor for $N > p$.

Hence, $E\exp(\frac{C}{2(1+K)}|x_{i,t,k}x_{i,t,l}|) \leq 2$. It now follows by the independence across $i = 1, ..., N$ (using Lemma 3) there exists constants A and B such that for any $\epsilon > \frac{1}{32\sqrt{N}}$

$$P\left(\Big|\frac{1}{N}\sum_{i=1}^{N}\big(\frac{1}{T}\sum_{t=1}^{T}\big[x_{i,t,k}x_{i,t,l} - E(x_{i,t,k}x_{i,t,l})\big]\big)\Big| \geq \epsilon\right) \leq Ae^{-B(\epsilon^2 N)^{1/3}} \tag{35}$$

Next, consider a typical term in $\frac{X'D}{\sqrt{NT}}$. Such a term is on the form $\frac{\sum_{t=1}^{T} x_{i,t,k}}{\sqrt{NT}}$ for $i = 1, ..., N$ and $k = 1, ..., p$. Since $\|x_{i,t,k}\|_{\psi_2} \leq \left(\frac{1+K/2}{C}\right)^{1/2}$ by Lemma 2.2.1 in **??** one gets

$$\left\|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T} x_{i,t,k}\right\|_{\psi_2} \leq \frac{1}{\sqrt{N}}\max_{1\leq t\leq T}\|x_{i,t,k}\|_{\psi_2} \leq \frac{1}{\sqrt{N}}\left(\frac{1+K/2}{C}\right)^{1/2} := \frac{M}{\sqrt{N}}.$$

It follows by Markov's inequality and $1 \wedge \psi_2(x)^{-1} = 1 \wedge (e^{x^2} - 1)^{-1} \leq 2e^{-x^2}$ that for any $\epsilon > 0$

$$P\left(\Big|\frac{1}{\sqrt{NT}}\sum_{t=1}^{T} x_{i,t,k}\Big| > \epsilon\right) \leq 1 \wedge \frac{1}{e^{(\epsilon\sqrt{N}/M)^2} - 1} \leq 2e^{-(\epsilon\sqrt{N}/M)^2} \leq Ae^{-B\epsilon^2 N} \tag{36}$$

where the last estimate follows by choosing A and B sufficiently large/small for (35) and (36) both to be valid. Combining (35) and (36) yields via a union bound over $(p^2 + Np)$ terms

$$P\left(\max_{1\leq i,j\leq p+N}|A_{i,j} - B_{i,j}| > \epsilon\right) \leq A(p^2 + Np)\left(e^{-B(\epsilon^2 N)^{1/3}} \vee e^{-B\epsilon^2 N}\right)$$

Using $\epsilon = \frac{\kappa^2}{32(s_1+s_2)m_{N,T}^2}$ with $m_{N,T} = \frac{\ln(p)^{3/2}}{\ln(N)^{3/2}} \vee \frac{\ln(N)^{3/2}}{\ln(p)^{3/2}}$ means that $\epsilon \geq \frac{1}{32\sqrt{N}}$ since $t^2 N \geq 1$. Hence,

$$P\left(\max_{1\leq i,j\leq p+N}|A_{i,j} - B_{i,j}| > \epsilon\right) \leq A(p^2 + Np)\left(e^{-B((1/32)^2 t^2 N)^{1/3}} \vee e^{-B(1/32)^2 t^2 N}\right) = e^{-B(t^2 N)^{1/3}}$$

where the $(1/32)^2$ have been merged into B and we have used that $t^2 N \geq 1$. □

Proof of Theorem 2. $P(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}) \geq 1 - 2\left(\frac{C_r}{a_{N,T}}\right)^r - D_r\frac{(p^2+Np)(s_1+s_2)^{r/2}(\frac{p}{N}\vee\frac{N}{p})}{\kappa^r N^{r/4}}$ follows from Lemmas 2 and 6. Hence, the estimates in Theorem 1 are valid with at least this probability. Inserting the definitions of $\lambda_{N,T}$ and $\mu_{N,T}$ into (5) and (6) yields (7) and (8). □

Proof of Theorem 3. The lower bound on $P(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T})$ by combining Lemmas 4 and 7. Hence, the estimates in Theorem 1 are valid with a probability bounded from below by this estimate. Inserting the definitions of $\lambda_{N,T}$ and $\mu_{N,T}$ into (5) and (6) yields (9) and (10). □

Before we prove Theorem 4 below we define the weighted Lasso as the minimizer of the following objective function,

$$\|y - Z\gamma\|^2 + 2\sum_{j=1}^{p+N} w_j|\gamma_j| \tag{37}$$

where $w_j,\ j = 1, ..., p + N$ are the weights. Note that in the plain Lasso, $w_j = \lambda_{N,T}$ for all $j = 1, ..., p$ and $w_j = \mu_{N,T}$ for $j = p + 1, ..., p + N$. From standard convex analysis we

know that a vector $\tilde{\gamma}$ minimizes (37) if and only if there exists a subgradient v of $\|\gamma\|_{\ell_1}$ such that

$$-Z_j'(y - Z\tilde{\gamma}) + w_j v_j = 0 \text{ for all } j = 1, ..., p + N \tag{38}$$

where $v_j = \text{sign}(\tilde{\gamma}_j)$ if $\tilde{\gamma}_j \neq 0$ and $v_j \in [-1, 1]$ if $\tilde{\gamma}_j = 0$. The following Lemma will be used in the proof of Theorems 4 and 6.

Lemma 8. *Suppose that $|v_j| < 1$ for all $\tilde{\gamma}_j = 0$ in (38) and that $Z_J'Z_J$ is invertible. Then* $\text{sign}(\tilde{\gamma}) = \text{sign}(\gamma^*)$ *if*

$$\text{sign}\left(\gamma_J^* + \left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right]\right) = \text{sign}(\gamma_J^*) \tag{39}$$

(here r is the $(p + N) \times 1$ vector with jth entry $w_j v_j$) and

$$\left|-Z_j'Z_J\left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right] + Z_j'\epsilon\right| < w_j \tag{40}$$

for all $j \in J^c$

Proof. The proof combines ideas from Wainwright (2009) and Zhou et al. (2009). Clearly, $\text{sign}(\tilde{\gamma}) = \text{sign}(\gamma^*)$ if and only if i) $\tilde{\gamma}$ solves (38) and ii) $\text{sign}(\tilde{\gamma}) = \text{sign}(\gamma^*)$. Using $y = Z\gamma^* + \epsilon$ the first order condition (38) is equivalent to

$$Z'Z(\tilde{\gamma} - \gamma^*) - Z'\epsilon + r = 0$$

Using $\tilde{\gamma}_{J^c} = \gamma_{J^c}^* = 0$ it follows by the invertibility of $Z_J'Z_J$ that

$$\tilde{\gamma}_J - \gamma_J^* = \left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right] \tag{41}$$

which yields $\text{sign}(\tilde{\gamma}_J) = \text{sign}(\gamma_J^*)$ under the stated conditions. Furthermore, we have

$$0 = Z_{J^c}'Z_J(\tilde{\gamma}_J - \gamma_J^*) - Z_{J^c}'\epsilon + r_{J^c} = Z_{J^c}'Z_J\left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right] - Z_{J^c}'\epsilon + r_{J^c}$$

Hence, we must have

$$w_j v_j = r_j = -Z_j'Z_J\left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right] + Z_j'\epsilon$$

for all $j \in J^c$ which means (using $|v_j| < 1$)

$$\left|-Z_j'Z_J\left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right] + Z_j'\epsilon\right| < w_j \tag{42}$$

for all $j \in J^c$. Next, $|v_j| < 1$ may be used to show that *any* solution $\bar{\gamma}$ of the minimization problem must have $\bar{\gamma}_j = 0$ if $\tilde{\gamma}_j$=0. This can be done by mimicking the argument in the proof of Lemma 2.1 in Bühlmann and Van De Geer (2011). Finally, using that $\tilde{\gamma}_{J^c} = 0$ and that $Z_J'Z_J$ is invertible (37) is seen to be strictly convex and so $\tilde{\gamma}' = \left(\gamma^{*\prime} + \left(Z_J'Z_J\right)^{-1}\left[Z_J'\epsilon - r_J\right]', 0'\right)$ is indeed the only solution. □

Proof of Theorem 4. By (41) one gets

$$S_{J,J}(\tilde{\gamma}_J - \gamma_J^*) = \left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\left[S_{J,J}^{-1}Z_J'\epsilon - S_{J,J}^{-1}r_J\right]$$

which implies

$$\left\|S_{J,J}(\tilde{\gamma}_J - \gamma_J^*)\right\| \geq \left\|\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}S_{J,J}^{-1}r_J\right\| - \left\|\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}S_{J,J}^{-1}Z_J'\epsilon\right\|$$

Next, note that using arguments similar to those in (5) it is seen that on $\tilde{\mathcal{B}}_{N,T}$ one has $\phi_{\max}\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right) \leq 2\phi_{\max}(\Gamma_{J,J})$ and so

$$\left\|\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}S_{J,J}^{-1}r_J\right\|^2 \geq \phi_{\min}\left[\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\right]\left(\frac{|J_1|\lambda_{N,T}^2}{NT}+\frac{|J_2|\mu_{N,T}^2}{T}\right)$$
$$\geq \frac{1}{\phi_{\max}\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)}\left(\frac{|J_1|\lambda_{N,T}^2}{NT}+\frac{|J_2|\mu_{N,T}^2}{T}\right) \geq \frac{1}{2\phi_{\max}\left(\Gamma_{J,J}\right)}\left(\frac{|J_1|\lambda_{N,T}^2}{NT}+\frac{|J_2|\mu_{N,T}^2}{T}\right)$$

Furthermore, by the independence of Z_J and ϵ and the gaussianity of ϵ, it follows that conditional on Z_J, $\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}S_{J,J}^{-1}Z_J'\epsilon$ is gaussian with mean zero and covariance $\sigma^2\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}$. Hence, for any $s>0$ letting $\tilde{\epsilon}\in\mathbb{R}^{|J|}$ be a standard gaussian vector we have

$$P\left(\left\|\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}S_{J,J}^{-1}Z_J'\epsilon\right\|^2 \leq s\right) = P\left(\tilde{\epsilon}'\sigma^2\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\tilde{\epsilon}\leq s\right)$$

But since $\tilde{\epsilon}'\tilde{\epsilon}$ is $\chi^2(|J|)$ it follows from expression (54a) in Wainwright (2009)[11] that there exists a constant c_1 such that $P(\tilde{\epsilon}'\tilde{\epsilon}\geq 3|J|)\leq \exp\left(-c_1|J|\right)$

$$\tilde{\epsilon}'\sigma^2\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\tilde{\epsilon} \leq \tilde{\epsilon}'\tilde{\epsilon}\phi_{\max}\left(\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\right) = \tilde{\epsilon}'\tilde{\epsilon}\frac{1}{\phi_{\min}\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)}$$

Now by arguments similar to the one in the proof of Lemma 5 one has on $\tilde{\mathcal{B}}_{N,T}$ that $1/\phi_{\min}\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)\leq 2/\phi_{\min}(\Gamma_{J,J})$ and so

$$P\left(\tilde{\epsilon}'\sigma^2\left(S_{J,J}^{-1}Z_J'Z_JS_{J,J}^{-1}\right)^{-1}\tilde{\epsilon}\leq s\right) \geq 1-\exp\left(-c_1|J|\right)-A(p^2+Np)e^{-B(t^2N)^{1/3}}$$

for $s=3|J|\cdot 2/\phi_{\min}(\Gamma_{J,J})$. Hence, with probability at least $1-\exp\left(-c_1|J|\right)-A(p^2+Np)e^{-B(t^2N)^{1/3}}$ for constants d_1,d_2 and c

$$\begin{aligned}
\left\|S_{J,J}(\tilde{\gamma}_J-\gamma_J^*)\right\| &\geq \sqrt{\frac{1}{2\phi_{\max}\left(\Gamma_{J,J}\right)}\left(\frac{|J_1|\lambda_{N,T}^2}{NT}+\frac{|J_2|\mu_{N,T}^2}{T}\right)}-\sqrt{3|J|\cdot 2/\phi_{\min}(\Gamma_{J,J})} \\
&\geq d_1(\sqrt{|J_1|}\lambda_{N,T}/\sqrt{NT}+\sqrt{|J_2|}\mu_{N,T}/\sqrt{T})-d_2(\sqrt{|J_1|}+\sqrt{|J_2|}) \\
&= d_1\sqrt{|J_1|}\lambda_{N,T}/\sqrt{NT}\left(1-\frac{d_2\sqrt{NT}}{d_1\lambda_{N,T}}\right)+d_1\sqrt{|J_2|}\mu_{N,T}/\sqrt{T}\left(1-\frac{d_2\sqrt{T}}{d_1\mu_{N,T}}\right) \\
&\geq c\sqrt{|J_1|}\lambda_{N,T}/\sqrt{NT}+c\sqrt{|J_2|}\mu_{N,T}/\sqrt{T} \\
&= c_2\xi_{N,T}
\end{aligned}$$

where the first estimate used Jensen's inequality on the concave $x\mapsto\sqrt{x}$ for the first (constants merged into d_1) term and the subadditivity of the same function on the second term. The existence of the constants d_1 and d_2 follows from the fact that $\phi_{\max}\left(\Gamma_{J,J}\right)$ and

[11]More precisely, (54a) in Wainwright (2009) states that given a centered χ^2-variable X with d degrees of freedom, then for any $t\in(0,1/2)$ one has $P(X\geq d(1+t))\leq\exp\left(-\frac{3}{16}dt^2\right)$. Hence, for an uncentered χ^2-variable Y with d degrees of freedom

$$P(Y\geq 3d)\leq P(Y\geq d+(1+t)d)=P(Y-d\geq(1+t)d)=P(X\geq(1+t)d)\leq\exp\left(-\frac{3}{16}dt^2\right)=\exp\left(-c_1d\right)$$

where the first estimate follows from $d\in(0,1/2)$ and the last equality by fixing some $t\in(0,1/2)$

$\phi_{\min}\left(\Gamma_{J,J}\right)$ are bounded from above and below, respectively. The last inequality follows by choosing $a_{N,T}$ sufficiently large while the last equality follows from the definitions of $\lambda_{N,T}, \mu_{N,T}$ and $\xi_{N,T}$ and the fact that κ^2 is bounded from below. □

Proof of Theorem 5. We start with the consistency part. The conclusion follows from Theorem 3 if we show that $P\left(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}\right) \to 1$ and that $\xi_{N,T}/\sqrt{NT},\ \xi_{N,T}/\sqrt{T} \to 0$. All notation is as in the statement of Theorem 3. To establish that $P\left(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}\right) \to 1$ it suffices to show that $A(p^2 + Np)e^{-B(t^2N)^{1/3}} \to 0$ Note that, ignoring constants,

$$t^2N = \frac{N}{\left(N^c\right)^2 \left(\frac{N^b}{\ln(N)} \vee \frac{\ln(N)}{N^b}\right)^6} = N^{1-2c-6b}\ln(N)^6 \to \infty$$

because $6b + 2c \leq 9b + 2c \leq 1$. Since $t^2N \to \infty$ and p increases exponentially in N it is enough to show that $p^2e^{-B(t^2N)^{1/3}} \to 0$. But this is the case, since

$$p^2e^{-B(t^2N)^{1/3}} = \exp(2N^b)\exp(-BN^{(1/3-(2/3)c-2b)}\ln(N)^2) \to 0$$

because $9b + 2c \leq 1$. Next, note that, ignoring constants, $\xi_{N,T} = \log(N)^{3/2}N^{(3/2)b}N^{c/2} + \log(N)^3N^{c/2}$ which implies that

$$\xi_{N,T}/\sqrt{NT} = \log(N)^{3/2}N^{(3/2)b+c/2-1/2-(1/2)a} + \log(N)^3N^{c/2-1/2-(1/2)a} \to 0$$

since $3b + c < 1 + a$. Similarly,

$$\xi_{N,T}/\sqrt{T} = \log(N)^{3/2}N^{(3/2)b+c/2-(1/2)a} + \log(N)^3N^{c/2-(1/2)a} \to 0$$

since $3b+c < a$. Regarding the second part we have already established that $P\left(\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}\right) \to 1$ since $9b + 2c \leq 1$. Hence, $\|\hat{\beta} - \beta^*\| \leq \xi_{N,T}/\sqrt{NT}$ with probability tending to one. But $\hat{\beta}_j = 0$ for some $j \in J_1$ implies $\|\hat{\beta} - \beta^*\| > \xi_{N,T}/\sqrt{NT}$. This is a contradiction and so it can't be the case that $\hat{\beta}_j = 0$ for any $j \in J_1$. A similar argument applies to $\hat{c}_i$ for $i \in J_2$. □

Lemma 9. *Under assumption A1) and A2a)*

(1) $P(\mathcal{C}_{1,N,T}) \geq 1 - \frac{2}{a_{N,T}^{r/2}}$ *for* $K_{1,N,T} = |J_1^c|^{2/r}|J_1|^{2/r}(NT)^{1/2}a_{N,T}$

(2) $P(\mathcal{C}_{2,N,T}) \geq 1 - \frac{1}{a_{N,T}^{r}}$ *for* $K_{2,N,T} = |J_1|^{1/r}|J_2^c|^{1/r}T^{1/2}a_{N,T}$

Proof. First, note that

$$\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l}\right\|_{L_{r/2}} \leq \sqrt{NT}\max_{1\leq t\leq T}\left\|x_{1,t,k}x_{1,t,l}\right\|_{L_{r/2}} \leq \sqrt{NT}$$

where the last estimate follows from the Cauchy-Schwarz inequality. Hence, $\left\|\max_{k\in J_1^c}\max_{l\in J_1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x\right.$ $|J_1^c|^{2/r}|J_1|^{2/r}\sqrt{NT}$. It follows from Markov's inequality that

$$P\left(\max_{k\in J_1^c}\max_{l\in J_1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l} \geq K_{1,N,T}\right) \leq \frac{|J_1^c||J_1|(NT)^{r/4}}{K_{1,N,T}^{r/2}} = \frac{1}{a_{N,T}^{r/2}}$$

Next,

$$\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\right\|_{L_r} \leq \sqrt{T}\max_{1\leq t\leq T}\left\|x_{i,t,k}\right\|_{L_r} \leq \sqrt{T}$$

This implies, $\left\|\max_{i\in J_2}\max_{k\in J_1^c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\right\|_{L_r}\leq |J_1^c|^{1/r}|J_2|^{1/r}\sqrt{T}$ and Markov's inequality yields

$$P\left(\max_{i\in J_2}\max_{k\in J_1^c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\geq K_{1,N,T}\right)\leq\frac{|J_1^c||J_2|T^{r/2}}{K_{1,N,T}^r}=\frac{|J_2|}{|J_1^c||J_1|^2N^{r/2}a_{N,T}^r}\leq\frac{1}{a_{N,T}^{r/2}}$$

where the last estimates follows from $|J_2|\leq N^{r/2}$ and $a_{N,T}\geq 1$. The conclusion of the first part of the lemma now follows by a union bound. The second part of the lemma is proved in a similar manner. □

Lemma 10. *Under assumption A1) and A2b)*

(1) $P(\mathcal{C}_{1,N,T})\geq 1-\frac{4}{a_{N,T}}$ *for* $K_{1,N,T}=A\log(1+|J_1^c|)\log(e+|J_1|)\sqrt{NT}\log(a_{N,T})$

(2) $P(\mathcal{C}_{2,N,T})\geq 1-\frac{2}{a_{N,T}}$ *for* $K_{2,N,T}=A\log(1+|J_1|)\log(1+|J_2^c|)\sqrt{T}\log(a_{N,T})$

for a constant $A>0$.

Proof. First, note that

$$\left\|\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l}\right\|_{\psi_1}\leq\sqrt{NT}\max_{1\leq t\leq T}\left\|x_{1,t,k}x_{1,t,l}\right\|_{\psi_1}\leq\sqrt{NT}\frac{1+K}{C}$$

where the last estimate follows from $\left\|x_{1,t,k}x_{1,t,l}\right\|_{\psi_1}\leq\frac{1+K}{C}:=A$ as argued in the proof of Lemma 7. Hence, $\left\|\max_{k\in J_1^c}\max_{l\in J_1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l}\right\|_{\psi_1}\leq A\log(1+|J_1^c|)\log(e+|J_1|)\sqrt{NT}$. By Markov's inequality, the definition of the Orlicz norm, and the fact that $1\wedge\psi(x)^{-1}=1\wedge(e^x-1)^{-1}\leq 2e^{-x}$,

$$\begin{aligned}&P\left(\max_{k\in J_1^c}\max_{l\in J_1}\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{i,t,k}x_{i,t,l}\geq K_{1,N,T}\right)\\&\leq 1\wedge\frac{1}{\exp(K_{1,N,T}/A\log(1+|J_1^c|)\log(1+|J_1|)\sqrt{NT})-1}=\frac{2}{a_{N,T}}\end{aligned}$$

Next, since $x_{i,t,k}$ is subgaussian it is also subexponential, and so there exists a constant $A>0$ such that

$$\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\right\|_{\psi_1}\leq\sqrt{T}\max_{1\leq t\leq T}\left\|x_{i,t,k}\right\|_{\psi_1}\leq\sqrt{T}A$$

This implies, $\left\|\max_{i\in J_2}\max_{k\in J_1^c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\right\|_{\psi_1}\leq A\log(1+|J_1^c|)\log(1+|J_2|)\sqrt{T}$ and Markov's inequality yields by similar arguments as above[12]

$$\begin{aligned}&P\left(\max_{i\in J_2}\max_{k\in J_1^c}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}x_{i,t,k}\geq K_{1,N,T}\right)\\&\leq 1\wedge\frac{1}{\exp(K_{1,N,T}/A\log(1+|J_1^c|)\log(1+|J_2|)\sqrt{T})-1}\\&\leq 2\exp\left(-\frac{A\log(1+|J_1^c|)\log(e+|J_1|)\sqrt{NT}\log(a_{N,T})}{A\log(1+|J_1^c|)\log(1+|J_2|)\sqrt{T}}\right)\leq\frac{2}{a_{N,T}}\end{aligned}$$

[12]The constant A may take different values throughout.

where the last estimate follows from $\log(1+|J_2|) \leq N^{1/2}$, $\log(e+|J_1|) \geq 1$ and $a_{N,T} \geq 1$. The conclusion of the first part of the lemma now follows by a union bound. The second part of the lemma is proved in a similar manner. □

Before we prove Theorem 6 note that $\tilde{\mathcal{B}}_{N,T} \subseteq \mathcal{B}_{N,T}$ (see the definition of $\tilde{B}_{N,T}$ in (31)) as already argued in the proofs of Lemmas 6 and 7. Furthermore, an argument similar to the one in Lemma 5 reveals that $\mathcal{D}_{N,T} = \{\phi_{\min}(\Psi_{J,J}) \geq \frac{1}{2}\phi_{\min}(\Gamma_{J,J})\}$ occurs if the maximal entry of $|\Psi_{J,J} - \Gamma_{J,J}|$ is less than or equal to $\frac{\phi_{\min}(\Gamma_{J,J})}{2(s_1+s_2)}$. But this latter event clearly contains $\tilde{\mathcal{B}}_{N,T}$ and so $\tilde{\mathcal{B}}_{N,T} \subseteq \mathcal{D}_{N,T}$.

Proof of Theorem 6. We shall prove the first part of the theorem since the proof of the second part follows along exactly the same lines (except for replacing $\mathcal{C}_{1,N,T}$ by $\mathcal{C}_{2,N,T}$ in the following arguments). Throughout we work on $\mathcal{A}_{N,T} \cap \mathcal{C}_{1,N,T} \cap \mathcal{D}_{N,T} \cap \{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\} \cap \{\|\hat{c} - c^*\| \leq c_{\min}/2\}$ and verify that (39) and (40) are valid on this set with $w = (w_1', w_2')'$ and $w_{1j} = \lambda_{N,T}/|\hat{\beta}_j|$, $j = 1, ..., p$ as well as $w_{2j} = \mu_{N,T}/|\hat{c}_j|$, $j = 1, ..., N$ and the convention that $1/0 = \infty$. First note that since $S_{J,J}$ is a diagonal matrix with positive entries on the diagonal (39) is equivalent to

$$\operatorname{sign}\left(S_{J,J}\gamma_J^* + S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right]\right) = \operatorname{sign}(\gamma_J^*)$$

Focussing on an X_j with $j \in J_1$ it hence suffices to show that[13]

$$\left|\left(S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right]\right)_j\right| \leq \sqrt{NT}\beta_{\min}$$

The left hand side in the above display may be upper bounded by $\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right\|_{\ell_\infty} \left\|(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right]\right\|_{\ell_\infty}$. Since

$$\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right\|_{\ell_\infty} \leq \sqrt{|J|}\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right\|$$

and on $\mathcal{D}_{N,T}$ one has

$$\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right\| = \phi_{\max}\left(S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right) = \frac{1}{\phi_{\min}(\Psi_{J,J})} \leq \frac{2}{\phi_{\min}(\Gamma_{J,J})} \tag{43}$$

it follows that

$$\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1} S_{J,J}\right\|_{\ell_\infty} \leq \frac{2\sqrt{|J|}}{\phi_{\min}(\Gamma_{J,J})}$$

Furthermore, because $\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2$ (by assumption)

$$|\hat{\beta}_j| \geq \beta_j^* - |\hat{\beta}_j - \beta_j^*| \geq \beta_{\min} - \|\hat{\beta} - \beta^*\| \geq \beta_{\min}/2$$

for all $j \in J_1$. By a similar argument $\hat{c}_j \geq c_{\min}/2$ for all $j \in J_2$. Hence,

$$\left\|(S_{J,J})^{-1} r_J\right\|_{\ell_\infty} = \left\|\frac{\lambda_{N,T}}{\sqrt{NT}\hat{\beta}_{J_1}}\right\|_{\ell_\infty} \vee \left\|\frac{\mu_{N,T}}{\sqrt{T}\hat{c}_{J_2}}\right\|_{\ell_\infty} \leq \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}$$

Next, on $\mathcal{A}_{N,T}$

$$\left\|(S_{J,J})^{-1} Z_J'\epsilon\right\|_{\ell_\infty} \leq \left\|\frac{X_{J_1}'\epsilon}{\sqrt{NT}}\right\|_{\ell_\infty} \vee \left\|\frac{D_{J_2}'\epsilon}{\sqrt{T}}\right\|_{\ell_\infty} \leq \frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}}$$

[13]Here, without causing confusion, we assume that X_j, $j \in J_1$ is indeed the jth variable.

It follows that

$$\begin{aligned}\left\|(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right]\right\|_{\ell_\infty} &\leq \left\|(S_{J,J})^{-1}Z_J'\epsilon\right\|_{\ell_\infty} + \left\|(S_{J,J})^{-1}r_J\right\|_{\ell_\infty} \\ &\leq \frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}} + \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}} \end{aligned} \tag{44}$$

Hence, putting the pieces together, (39) is satisfied for all $j \in J_1$ if

$$\frac{2\sqrt{|J|}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}} + \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right) \leq \sqrt{NT}\beta_{\min}$$

Next, (40) is equivalent to

$$\left|-Z_j'Z_J(S_{J,J})^{-1}S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right] + Z_j'\epsilon\right| < w_j \tag{45}$$

for all $j \in J^c$. The left hand side in the above display is bounded from above by

$$\left\|Z_j'Z_J(S_{J,J})^{-1}S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}\right\|_{\ell_1}\left\|(S_{J,J})^{-1}\left[Z_J'\epsilon - r_J\right]\right\|_{\ell_\infty} + \left|Z_j'\epsilon\right|$$

Assume again that Z_j is an X_j. Then, on $\mathcal{C}_{1,N,T}$ and by (43)

$$\begin{aligned}&\left\|Z_j'Z_J(S_{J,J})^{-1}S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}\right\|_{\ell_1} \leq \sqrt{|J|}\left\|Z_j'Z_J(S_{J,J})^{-1}S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}\right\| \\ &\leq |J|\left\|Z_j'Z_J(S_{J,J})^{-1}\right\|_{\ell_\infty}\left\|S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}\right\| \\ &\leq \frac{2|J|K_{1,N,T}}{\phi_{\min}(\Gamma_{J,J})}\end{aligned}$$

where the second estimate follows by considering $S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}$ as a bounded linear operator from $\ell_2(\mathbb{R}^{|J|}) \to \ell_2(\mathbb{R}^{|J|})$ with induced operator norm given by $\phi_{\max}\left(S_{J,J}\left(Z_J'Z_J\right)^{-1}S_{J,J}\right)$. Putting the pieces together, and using that we are on $\mathcal{A}_{N,T}$ and by (44) the left hand side in (45) may be upper bounded by

$$\frac{2|J|K_{1,N,T}}{\phi_{\min}(\Gamma_{J,J})}\left(\frac{\lambda_{N,T}}{2\sqrt{NT}} \vee \frac{\mu_{N,T}}{2\sqrt{T}} + \frac{2\lambda_{N,T}}{\sqrt{NT}\beta_{\min}} \vee \frac{2\mu_{N,T}}{\sqrt{T}c_{\min}}\right) + \frac{\lambda_{N,T}}{2}$$

Finally, the right hand side in (45) may be bounded from below by $\lambda_{N,T}/\left\|\hat{\beta} - \beta^*\right\|$ and the result follows. □

Proof of Corollary 1. We know from Theorem 6 that $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$ on $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T} \cap \mathcal{C}_{1,N,T} \cap \mathcal{D}_{N,T} \cap \{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\} \cap \{\|\hat{c} - c^*\| \leq c_{\min}/2\}$ if (14)-(15) is satisfied[14]. Furthermore, if $\beta_{\min} \geq \frac{2\xi_{N,T}}{\sqrt{NT}}$ one has $\mathcal{A}_{N,T} \cap \tilde{\mathcal{B}}_{N,T} \cap \mathcal{C}_{1,N,T} \subseteq \mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}\mathcal{C}_{1,N,T} \cap \mathcal{D}_{N,T} \cap \{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\} \cap \{\|\hat{c} - c^*\| \leq c_{\min}/2\}$ [15]. The lower bound on the probability of $\{\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)\}$ now follows by Lemmas 2, 6 and 9 in case of part one of the corollary. In case of part 2 of the corollary Lemmas 4, 7 and 10 are used. A similar argument gives the lower bound on the probability with which $\text{sign}(\tilde{c}) = \text{sign}(c^*)$ by verifying (16)-(17). □

[14]Actually, we know from Theorem 6 that $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$ on the larger set $\mathcal{A}_{N,T} \cap \mathcal{C}_{1,N,T} \cap \mathcal{D}_{N,T} \cap \{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\} \cap \{\|\hat{c} - c^*\| \leq c_{\min}/2\}$. As will be seen, this distinction will turn out not to make any difference for our lower bounds on the probability of the events.

[15]The inclusion follows from the fact that $\tilde{\mathcal{B}}_{N,T} \subseteq \mathcal{B}_{N,T} \cap \mathcal{D}_{N,T}$ as argued prior to the proof of Theorem 6. Also the inclusion has used that on $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}$ one has $\|\hat{\beta} - \beta^*\| \leq \frac{\xi_{N,T}}{\sqrt{NT}}$ and $\|\hat{c} - c^*\| \leq \frac{\xi_{N,T}}{\sqrt{T}}$ such that $\beta_{\min} \geq \frac{2\xi_{N,T}}{\sqrt{NT}}$ and $c_{\min} \geq \frac{2\xi_{N,T}}{\sqrt{T}}$ imply that $\{\|\hat{\beta} - \beta^*\| \leq \beta_{\min}/2\}$ and $\{\|\hat{c} - c^*\| \leq c_{\min}/2\}$, respectively.

Proof of Theorem 7. We proceed by verifying the conditions related to the sign consistency of $\tilde{\beta}$ and $\tilde{c}$ in part 2 of Corollary 1 and showing that the lower bound on the probability with which $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$ and $\text{sign}(\tilde{c}) = \text{sign}(c^*)$ tends to one. We focus on $P\left(\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)\right) \to 1$ since the second part of the theorem follows by identical arguments.

First, we verify that 14 is satisfied asymptotically. To do so it suffices to show that $\frac{\sqrt{|J|}\lambda_{N,T}}{NT} \to 0$ and $\frac{\sqrt{|J|}\mu_{N,T}}{\sqrt{NT}} \to 0$. Now, ignoring constants, and using the definition of $\lambda_{N,T}$

$$\frac{\sqrt{|J|}\lambda_{N,T}}{NT} = \frac{\sqrt{|J|}\log(p)^{3/2}\log {a_{N,T}}^{3/2}}{\sqrt{NT}} = N^{c/2+\frac{3}{2}b-a/2-1/2}\log N^{3/2} \to 0$$

since $3b + c < 1 + a$. Similarly, using the definition of $\mu_{N,T}$

$$\frac{\sqrt{|J|}\mu_{N,T}}{\sqrt{NT}} = \frac{\sqrt{|J|}\log(N)^3}{\sqrt{NT}} = N^{c/2-a/2-1/2}\log(N)^3 \to 0$$

since $c < a + 1$. Next, we verify that (15) is valid asymptotically. To do so it suffices to show that $\frac{|J|K_{1,N,T}}{\sqrt{NT}}\|\hat{\beta} - \beta^*\| \to 0$, $\frac{|J|K_{1,N,T}\mu_{N,T}/\lambda_{N,T}}{\sqrt{T}}\|\hat{\beta} - \beta^*\| \to 0$ and $\|\hat{\beta} - \beta^*\| \to 0$. For this purpose, note that $K_{1,N,T} \leq A\log(1 + |p|)\log(e + |J_1|)\sqrt{NT}\log(a_{N,T})$ which is of order $\log(|p|)\log(|J_1|)\sqrt{NT}\log(a_{N,T}) = N^b c\log(N)^2\sqrt{NT} = N^{b+1/2+a/2}\log(N)^2$. Furthermore, $\|\hat{\beta} - \beta^*\| \leq \xi_{N,T}/\sqrt{NT}$ on $\mathcal{A}_{N,T} \cap \mathcal{B}_{N,T}$ which we are working on in Corollary 1 [16] (where $\xi_{N,T}$ is as defined in Theorem 3). Hence, ignoring constants,

$$\|\hat{\beta} - \beta^*\| \leq \xi_{N,T}/\sqrt{NT} \leq \log(N)^3 N^{\frac{3}{2}b+c/2-1/2-a/2} \to 0 \tag{46}$$

since $3b + c < 1 + a$. Also,

$$\begin{aligned}\frac{|J|K_{1,N,T}}{\sqrt{NT}}\|\hat{\beta} - \beta^*\| &\leq N^c N^{b+1/2+a/2}\log(N)^2\log(N)^3 N^{\frac{3}{2}b+c/2-1/2-a/2}N^{-1/2-a/2} \\ &= N^{\frac{5}{2}b+\frac{3}{2}c-1/2-a/2}\log(N)^5 \to 0\end{aligned}$$

since $5b + 3c < 1 + a$. Similarly, since $\mu_{N,T}/\lambda_{N,T} = \frac{\log(N)^{3/2}}{\sqrt{N}\log(p)^{3/2}} = \frac{\log(N)^{3/2}}{\sqrt{N}N^{\frac{3}{2}b}}$

$$\begin{aligned}\frac{|J|K_{1,N,T}\mu_{N,T}/\lambda_{N,T}}{\sqrt{T}}\|\hat{\beta} - \beta^*\| &= N^c N^{b+1/2+a/2}\log(N)^2\frac{\log(N)^{3/2}}{\sqrt{N}N^{\frac{3}{2}b}}\log(N)^3 N^{\frac{3}{2}b+c/2-1/2-a/2}N^{-a/2} \\ &= N^{b+\frac{3}{2}c-1/2-a/2}\log(N)^{13/2} \to 0\end{aligned}$$

since $2b + 3c < 1 + a$. Furthermore, $\beta_{\min} \geq 2\frac{\xi_{N,T}}{\sqrt{NT}}$ since $\frac{\xi_{N,T}}{\sqrt{NT}} \to 0$ when $3b + c < 1 + a$ as seen from (46) while $\beta_{\min}$ is bounded away from 0. Finally, we note that $9b + 2c \leq 1$ suffices to ensure that the lower bound on the probability in part 2 of Corollary 1 tends to one as was already argued in the proof of Theorem 5. □

References

Arellano, M. (2003). *Panel Data Econometrics*, Volume 1. Oxford University Press, Oxford.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics 106*(2), 407–443.

[16] See the first line of the proof of Corollary 1.

Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. *Inverse Problems and High-Dimensional Estimation*, 121–156.

Belloni, A., V. Chernozhukov, and L. Wang (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*(4), 791–806.

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: Methods, Theory and Applications*. Springer-Verlag, New York.

Candes, E. and T. Tao (2007). The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

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 (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Hall, P. and C. Heyde (1980). *Martingale limit theory and its application*, Volume 142. Academic press New York.

Hitczenko, P. (1990). Best constants in martingale version of rosenthal's inequality. *The Annals of Probability*, 1656–1668.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics 36*(2), 587–613.

Lesigne, E. and D. Volnỳ (2001). Large deviations for martingales. *Stochastic processes and their applications 96*(1), 143–159.

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

Pötscher, B. M. and H. Leeb (2009). On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis 100*(9), 2065–2082.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research 11*, 2241–2259.

Stoianov, Ĭ. (1997). *Counterexamples in probability*. John Wiley & Sons (Chichester).

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 Der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Verlag.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint*.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on 55*(5), 2183–2202.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. The MIT press.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zhou, S., S. Van De Geer, and P. Bühlmann (2009). Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*.

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

Research Papers 2013

2013-03: Stefano Grassi and Paolo Santucci de Magistris: It's all about volatility (of volatility): evidence from a two-factor stochastic volatility model

2013-04: Tom Engsted and Thomas Q. Pedersen: Housing market volatility in the OECD area: Evidence from VAR based return decompositions

2013-05: Søren Johansen and Bent Nielsen: Asymptotic analysis of the Forward Search

2013-06: Debopam Bhattacharya, Pascaline Dupasand Shin Kanaya: Estimating the Impact of Means-tested Subsidies under Treatment Externalities with Application to Anti-Malarial Bednets

2013-07: Sílvia Gonçalves, Ulrich Hounyo and Nour Meddahi: Bootstrap inference for pre-averaged realized volatility based on non-overlapping returns

2013-08: Katarzyna Lasak and Carlos Velasco: Fractional cointegration rank estimation

2013-09: Roberto Casarin, Stefano Grassi, Francesco Ravazzolo and Herman K. van Dijk: Parallel Sequential Monte Carlo for Efficient Density Combination: The Deco Matlab Toolbox

2013-10: Hendrik Kaufmann and Robinson Kruse: Bias-corrected estimation in potentially mildly explosive autoregressive models

2013-11: Robinson Kruse, Daniel Ventosa-Santaulària and Antonio E. Noriega: Changes in persistence, spurious regressions and the Fisher hypothesis

2013-12: Martin M. Andreasen, Jesús Fernández-Villaverde and Juan F. Rubio-Ramírez: The Pruned State-Space System for Non-Linear DSGE Models: Theory and Empirical Applications

2013-13: Tom Engsted, Stig V. Møller and Magnus Sander: Bond return predictability in expansions and recessions

2013-14: Charlotte Christiansen, Jonas Nygaard Eriksen and Stig V. Møller: Forecasting US Recessions: The Role of Sentiments

2013-15: Ole E. Barndorff-Nielsen, Mikko S. Pakkanen and Jürgen Schmiegel: Assessing Relative Volatility/Intermittency/Energy Dissipation

2013-16: Peter Exterkate, Patrick J.F. Groenen, Christiaan Heij and Dick van Dijk: Nonlinear Forecasting With Many Predictors Using Kernel Ridge Regression

2013-17: Daniela Osterrieder: Interest Rates with Long Memory: A Generalized Affine Term-Structure Model

2013-18: Kirstin Hubrich and Timo Teräsvirta: Thresholds and Smooth Transitions in Vector Autoregressive Models

2013-19: Asger Lunde and Kasper V. Olesen: Modeling and Forecasting the Volatility of Energy Forward Returns - Evidence from the Nordic Power Exchange

2013-20: Anders Bredahl Kock: Oracle inequalities for high-dimensional panel data models