Electronic Journal of Statistics
Vol. 13 (2019) 2970–3008
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1599

On the asymptotic variance of the debiased Lasso

Sara van de Geer

Seminar for Statistics
ETH Zürich
Rämistrasse 101
8092 Zürich
Switzerland
e-mail: geer@stat.math.ethz.ch

Abstract: We consider the high-dimensional linear regression model $Y = X\beta^0 + \epsilon$ with Gaussian noise ϵ and Gaussian random design X. We assume that $\Sigma := \mathbb{E}X^T X/n$ is non-singular and write its inverse as $\Theta := \Sigma^{-1}$. The parameter of interest is the first component β_1^0 of β^0. We show that in the high-dimensional case the asymptotic variance of a debiased Lasso estimator can be smaller than $\Theta_{1,1}$. For some special such cases we establish asymptotic efficiency. The conditions include β^0 being sparse and the first column Θ_1 of Θ being not sparse. These sparsity conditions depend on whether Σ is known or not.

MSC 2010 subject classifications: Primary 62J07; secondary 62E20.
Keywords and phrases: Asymptotic efficiency, asymptotic variance, Cramér Rao lower bound, debiasing, Lasso, sparsity.

Received August 2018.

Contents

1. Introduction

Let Y be an n-vector of observations and $X \in \mathbb{R}^{n \times p}$ an input matrix. The linear model is

$$Y = X\beta^0 + \epsilon,$$

where $\beta^0 \in \mathbb{R}^p$ is a vector of unknown coefficients and $\epsilon \in \mathbb{R}^n$ is unobservable noise. We examine the high-dimensional case with $p \gg n$. The parameter of interest in this paper is a component of β^0, say the first component β^0_1. We consider the asymptotic properties of debiased estimators of the one-dimensional parameter β^0_1 under scenarios where certain sparsity assumptions fail to hold.

The paper shows that the asymptotic variance of the debiased estimator can be smaller than the "usual" value for the low-dimensional case. For simplicity we examine Gaussian data: the rows of $(X, Y) \in \mathbb{R}^{n \times (p+1)}$ are i.i.d. copies of a zero-mean Gaussian row vector $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{p+1}$, where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ has covariance matrix $\Sigma := \mathbb{E}\mathbf{x}^T\mathbf{x}$. We assume the inverse of Σ exists and write it as $\Theta := \Sigma^{-1}$. The vector β^0 of regression coefficients is $\beta^0 = \Theta\mathbb{E}\mathbf{x}^T\mathbf{y}$. We denote the first column of Θ by $\Theta_1 \in \mathbb{R}^p$ and the first element of this vector by $\Theta_{1,1}$. Our main aim is to present examples where lack of sparsity in Θ_1 can result in a small asymptotic variance of a suitably debiased estimator. In particular, the asymptotic variance can be smaller than $\Theta_{1,1}$. For the case Σ known, this means applying for instance a (noiseless) node-wise Lasso, instead of an exact orthogonalization of the first variable with respect to the others, may reduce the asymptotic variance (as follows from combining Theorem 2.1 with Lemma 3.2). If Σ is unknown, the high dimensionality of the problem already excludes exact empirical projections for orthogonalization. The (noisy) Lasso is designed to deal with approximate orthogonalization in the high-dimensional case. Using the node-wise Lasso, we find that one may again profit from non-sparsity of the (now unknown) vector Θ_1 (see Theorem 4.1).

We look at specific examples or constructions of covariance matrices Σ. The results illustrate that asymptotic efficiency claims require some caution. The high-dimensional situation exhibits new phenomena that do not occur in low dimensions.

Throughout, the minimal eigenvalue of Σ, denoted by Λ^2_{min}, is required to stay away from zero, i.e. $1/\Lambda^2_{\text{min}} = \mathcal{O}(1)$. We further consider only Σ's with all 1's on the diagonal and assume for simplicity that $\sigma^2_\epsilon := \mathbb{E}(\mathbf{y} - \mathbf{x}\beta^0)^2$ is known and that its value is $\sigma^2_\epsilon = 1$.

Let a given subset $\mathcal{B}$ of $\mathbb{R}^p$ be the model class for β^0. An interesting research goal is to construct for the model $\mathcal{B}$ a regular estimator of β_1^0 with asymptotic variance that achieves the asymptotic Cramér Rao lower bound (given here in Proposition 1.1). One then needs to decide which model class $\mathcal{B}$ one considers as relevant. In high-dimensional statistics it is commonly assumed that β^0 is sparse in some sense. Let $0 < r \le 1$, define for a vector $b \in \mathbb{R}^p$ its ℓ_r-"norm" $\|b\|_r$ by $\|b\|_r^r := \sum_{j=1}^p |b_j|^r$ and let $\|b\|_0^0$ be its number of non-zero entries. A sparse model is for example

$$\mathcal{B} := \{b \in \mathbb{R}^p : \ \|b\|_0^0 \le s\} \tag{1.1}$$

for some ("small") $s \in \mathbb{N}$. Alternatively one may believe only in ℓ_1-sparsity. Then

$$\mathcal{B} := \{b \in \mathbb{R}^p : \ \|b\|_1 \le \sqrt{s}\} \tag{1.2}$$

for some $s > 0$. These are the two extremes of weakly sparse models of the form

$$\mathcal{B} := \{b \in \mathbb{R}^p : \|b\|_r^r \le s^{\frac{2-r}{2}}\}, \tag{1.3}$$

for some $s > 0$ and $0 \le r \le 1$. Throughout, the value of s is allowed to depend on n, but r is fixed for all n.

Constructing estimators that achieve the asymptotic Cramér Rao lower bound for model (1.1), (1.2), (1.3) or some other sparse model is to our understanding quite ambitious, especially if one wants to do this for all possible covariance matrices Σ. See e.g. Example 1.1 for some details concerning model (1.1). However, for special cases of Σ's the problem can be solved. One such special case is where the first row of Θ_1 of Θ is sparse in an appropriate sense. This is the situation considered in previous work such as [26] and [25] where Σ is unknown. In this paper we consider known and unknown Σ and in both cases do not require sparsity of Θ_1. The paper [11] also does not require sparsity of Θ_1 when Σ is known and it turns out that for certain non-sparse vectors Θ_1 their estimator is not asymptotically efficient, for example under the model (1.2) with $s = o(n/\log p)$ and with a matrix Σ of a certain form (see Theorem 2.1 or Remark 2.6 following this theorem).

The debiased Lasso defined in this paper in equation (1.5) below is based on a direction $\tilde{\Theta}_1 \in \mathbb{R}^p$ where $\tilde{\Theta}_1$ is thought of as some estimate of Θ_1. As we do not assume sparsity of Θ_1 a reliable estimator of Θ_1 may not be available. Nevertheless, we show that this does not rule out good theoretical performance. We present a class of covariance matrices Σ for which a debiased Lasso has asymptotic variance smaller than $\Theta_{1,1}$. This phenomenon is tied to the high-dimensional situation, see Remark 2.3. For special cases, we establish that an asymptotic Cramér Rao bound smaller than $\Theta_{1,1}$ can be achieved. In other words, there exist cases where a debiased Lasso profits from sparsity of β^0 combined with non-sparsity of Θ_1. This is good news: the asymptotic variance can be small for two reasons: either Θ_1 is sparse in which case the asymptotic variance $\Theta_{1,1}$ is the inverse of the residuals of regressing the first variable on only a few of the other variables, or Θ_1 is not sparse but then the asymptotic variance can be smaller than $\Theta_{1,1}$. This paper presents cases where the latter situation indeed occurs.

1.1. The Lasso, debiased Lasso and sparsity assumptions

The Lasso ([23]) is defined as

$$\hat{\beta} := \arg\min_{b\in\mathbb{R}^p}\left\{ \|Y - Xb\|_2^2/n + 2\lambda\|b\|_1 \right\} \tag{1.4}$$

with $\lambda > 0$ a tuning parameter. (We will throughout take of order $\sqrt{\log p/n}$ but not too small.)

A debiased Lasso is given by

$$\hat{b}_1 = \hat{\beta}_1 + \tilde{\Theta}_1^T X^T (Y - X\hat{\beta})/n. \tag{1.5}$$

The p-dimensional vector $\tilde{\Theta}_1$ is some estimate of the first column Θ_1 of the precision matrix Θ, but in our case it will rather be estimating a sparse approximation. We refer to $\tilde{\Theta}_1$ as a direction. The estimator $\hat{\beta}$ is commonly taken to be the Lasso given in (1.4) although this is not a must.

The debiased Lasso (1.5) was introduced in [26] and further developed in [9] and [25] for example. Related work is [1] and [2].

The theory for the Lasso (1.4) and debiased Lasso (1.5) requires some form of sparsity of β^0. Consider for some $s \in \mathbb{N}$ one of the sparsity models (1.1), (1.2) or, more generally, model (1.3). Two prevalent sparsity assumptions are:

(i) $s = o(n/\log p)$,
(ii) $s = o(\sqrt{n}/\log p)$.

Sparsity variant (i) is for example invoked to establish ℓ_2-consistency of the Lasso $\hat{\beta}$ (see [4] or the monographs [12], [5] and [7], and their references).

Which sparsity variant is needed to establish asymptotic normality of the debiased Lasso (1.5) depends to a large extent on whether Σ is known or not. In [10], [6], [20]. [11] one can find refined results on this issue.

This case of Σ known is treated in Section 2. We introduce and apply there the concept of an eligible pair, see Definition 2.1. An eligible pair is a sparse approximation of Θ_1 together with a parameter describing the order of approximation and sparsity. We allow for sparsity variant (i) in model (1.1) as in [11], see Example 2.1. Sparsity variant (i) will also be allowed for the models (1.2) and (1.3), see Examples 2.2 and 2.3.

Eligible pairs will also play a crucial role in Section 4 where Σ is unknown. Let us discuss some of the literature for this case and for the sparsity model (1.1). From the papers [6] and [20] we know that for the minimax bias of an estimator of β_1^0 to be of order $1/\sqrt{n}$, the assumption $s = O(\sqrt{n/\log p})$ is necessary. Thus, up to log-terms this needs the second sparsity variant. When considering asymptotic Cramér Rao lower bounds, one also needs to restrict oneself to a certain class of estimators, for instance estimators with bias of small order $1/\sqrt{n}$

or asymptotically linear estimators. In [8] such restrictions are studied. One can show asymptotic linearity of the debiased Lasso under model (1.1) with sparsity variant (ii) and in addition $\|\Theta_1\|_0 = o(\sqrt{n}/\log p)$. If Θ_1 is not sparse nor can be approximated by a sparse vector, then it is unclear whether an asymptotically linear estimator exists. We refer to Remark 4.4 for more details. In summary, modulo log-terms, sparsity variant (ii) cannot be relaxed as far as minimax rates for the bias are concerned, and sparsity variant (ii) with in addition sparsity of order $o(\sqrt{n}/\log p)$ for Θ_1 or its sparse approximation appears to be needed for establishing asymptotic linearity. We note that the paper [11] establishes asymptotic normality under (among others) the assumption

$$\min(s, \|\Theta_1\|_0^0) = o(\sqrt{n}/\log p). \tag{1.6}$$

Bias and asymptotic linearity are not considered (these issues are not within the scope of that paper). In our setup however, Θ_1 is not sparse at all, so variant (ii) is in line with (1.6).

Tables 2 and 3 presented in Subsection 1.4 summarizes the sparsity conditions applied in this paper. One sees that models (1.1) and (1.2) are special cases of model (1.3), with $r = 0$ and $r = 1$ respectively. However, when $r = 0$ the asymptotic efficiency depends on β^0 and also quite severely on the value of s. For the case Σ unknown, model (1.2) is too large.

1.2. The asymptotic Cramér Rao lower bound

We briefly review the Cramér Rao lower bound and refer to [8] for details. Let the model be $\beta^0 \in \mathcal{B}$, where $\mathcal{B}$ is a given class of regression coefficients. Let $\mathcal{H}_{\beta^0} := \{h \in \mathbb{R}^p : \ \beta^0 + h/\sqrt{n} \in \mathcal{B}\}$. We call $\mathcal{H}_{\beta^0}$ the set of model directions. An estimator T (or actually: sequence of estimators) is called regular at β^0 if for all fixed $\rho > 0$ and $R > 0$ not depending on n, and all sequences $h \in \mathcal{H}_{\beta^0}$ with $|h_1| \ge \rho$ and $h^T \Sigma h \le R^2$, it holds that

$$\sqrt{n}\left(\frac{T - (\beta_1^0 + h_1/\sqrt{n})}{V_{\beta^0}}\right) \xrightarrow{\mathcal{D}_{\beta^0 + h/\sqrt{n}}} \mathcal{N}(0,1)$$

where $V_{\beta^0}^2 = \mathcal{O}(1)$ is some constant (depending on n and possibly on β^0, but not depending on ρ, R or h), called the asymptotic variance (it is defined up to smaller order terms). Regularity is important in practice. It means that the asymptotics is not just pointwise but uniformly over neighbourhoods.

Remark 1.1. *The class $\mathcal{B}$ is not always a cone. This is the reason why we do not restrict ourselves to model directions $h \in \mathcal{H}_{\beta^0}$ with $h_1 = 1$ (say).*

Remark 1.2. *One may also opt for defining the set of possible directions $\mathcal{H}_{\beta^0}$ differently, say $\mathcal{H}_{\beta^0} := \mathcal{B}$. Then regularity concerns parameter values that fall outside the parameter space. For example under model (1.1) one then has to deal with sparsity $2s$ instead of s. With our choice of $\mathcal{H}_{\beta^0}$ we stay inside the parameter space but the set of possible directions then depends on β^0. In model*

(1.2) or (1.3) one can move away from this dependence when β^0 is a proper "interior point" of $\mathcal{B}$ (see Examples 1.2 and 1.3).

Proposition 1.1. *Suppose T is asymptotically linear at β^0 with influence function $\mathbf{i}_{\beta^0} : \mathbb{R}^{p+1} \to \mathbb{R}$:*

$$T - \beta_1^0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{i}_{\beta^0}(X_i, Y_i) + o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n})$$

where $\mathbb{E}_{\beta^0}\mathbf{i}_{\beta^0}(\mathbf{x}, \mathbf{y}) = 0$ and $V_{\beta^0}^2 := \mathbb{E}_{\beta^0}\mathbf{i}_{\beta^0}^2(\mathbf{x}, \mathbf{y}) = \mathcal{O}(1)$. Assume the Lindeberg condition

$$\lim_{n\to\infty} \mathbb{E}_{\beta^0}\mathbf{i}_{\beta^0}^2(\mathbf{x}, \mathbf{y})\mathrm{l}\left\{\mathbf{i}_{\beta^0}^2(\mathbf{x}, \mathbf{y}) > \eta n V_{\beta^0}^2\right\} = 0 \ \forall\ \eta > 0.$$

Assume further that T is regular at β^0. Then for all fixed $\rho > 0$ and $R > 0$

$$V_{\beta^0}^2 + o(1) \geq \max_{h\in\mathcal{H}_{\beta^0}:\ |h_1|\geq\rho,\ h^T\Sigma h\leq R^2} \frac{h_1^2}{h^T\Sigma h}.$$

This proposition is as Theorem 9 in [8] but tailored for the particular situation. A proof is given in Section 6. We remark that such results are not a direct consequence of the Le Cam theory, as we are dealing with triangular arrays.

The next corollary is our main tool to arrive at asymptotic efficiency for some special Σ's.

Corollary 1.1. *Assume the conditions of Proposition 1.1 and that for some fixed $\rho > 0$ and $R > 0$ and some sequence $h \in \mathcal{H}_{\beta^0}$, with $|h_1| \geq \rho$ and $h^T\Sigma h \leq R^2$, it is true that*

$$V_{\beta^0}^2 = \frac{h_1^2}{h^T\Sigma h} + o(1).$$

Then T is asymptotically efficient.

The restriction to directions in $\mathcal{H}_{\beta^0}$ means that the Cramér Rao lower bound for the asymptotic variance $V_{\beta^0}^2$ can be orders of magnitude smaller than $\Theta_{1,1}$.

Example 1.1. *Under the sparse model (1.1)*[1] *we have*

$$\underline{\mathcal{H}}_{\beta^0} \subseteq \mathcal{H}_{\beta^0},$$

where

$$\underline{\mathcal{H}}_{\beta^0} := \{\|h\|_0^0 \leq s - s_0\}$$

and $s_0 = |S_0|$ with $S_0 := \{\beta_j^0 \neq 0\}$ being the set of active coefficients of β^0.

Some special cases

a) If $\Theta_1 \in \underline{\mathcal{H}}_{\beta^0}$, then $V_{\beta^0}^2 + o(1) = \Theta_{1,1}$. Note that the condition on Θ_1 depends on β^0 (via s_0).

[1] A more natural model in this context might be $\mathcal{B} := \{b \in \mathbb{R}^p :\ \|b_{-1}\|_0^0 \leq s - 1\}$ where, for $b \in \mathbb{R}^p$, $b_{-1} := (b_2, \ldots, b_p)^T$.

b) Suppose $\{1\} \in S_0$, $s = s_0$ and that the following "betamin" condition holds: $|\beta_j^0| > m_n/\sqrt{n}$ for all $j \in S_0$, where m_n is some sequence satisfying $m_n \to \infty$. Then we see that

$$\mathcal{H}_{\beta^0} \cap \{h^T \Sigma h \le R^2\} = \{\|h_{-S_0}\|_0 = 0\} \cap \{h^T \Sigma h \le R^2\}.$$

The lower bound is then

$$V_{\beta^0}^2 + o(1) \ge (\Sigma_{S_0,S_0}^{-1})_{1,1},$$

where Σ_{S_0,S_0} is the matrix of covariances of the variables in S_0. This lower bound corresponds to the case S_0 is known. The bound could be achieved if one has a consistent estimate $\hat{S}$ of S_0. For this one needs betamin conditions in order to have no false negatives. Applying least squares with variables in $\hat{S}$, where $\hat{S}$ is an estimator of S_0 results in an estimator of β_1^0 which is not regular. There is a series of papers on this issue, e.g. [15], [14], [17], [18], [19]. Imposing further conditions beyond model (1.1), for example betamin conditions, will diminish the lower bound.

c) More generally, if $\{1\} \in S_0$ and $|\beta_j^0| > m_n/\sqrt{n}$ for all $j \in S_0$ and some sequence $m_n \to \infty$ then the lower bound corresponds to knowing the set S_0 up to $s - s_0$ additional variables.

d) Suppose that β^0 is an interior point in the sense that it stays away from the boundary: for some fixed $0 < \eta < 1$ not depending on n, it holds that $s_0 \le (1-\eta)s$ (so that $1 - s_0/s$ stays away from zero). Then

$$\{h : \ \|h\|_0^0 \le \eta s\} \subset \mathcal{H}_{\beta^0}.$$

The lower bound can then still depend on η. A rescaling argument as applied in the next examples, Example 1.2 and more generally Example 1.3, does not work here. This signifies that model (1.1) less suitable in our context: the exact value of s plays a too prominent role.

Example 1.2. *Consider model (1.2). The situation is then more like a classical one. Suppose β^0 stays away from the boundary of the parameter space, i.e., for a fixed $0 < \eta < 1$ not depending on n, it holds that $\|\beta^0\|_1 \le (1-\eta)\sqrt{s}$. Then*

$$\{\|h\|_1 \le \eta\sqrt{ns}\} \subset \mathcal{H}_{\beta^0}.$$

By a rescaling argument the dependence on η in the left hand side plays no role in the lower bound: for all $M > 0$ fixed (not depending on n)

$$V_{\beta^0}^2 + o(1) \ge \left(\min_{c \in \mathbb{R}^{p-1}: \ \|c\|_1 \le M\sqrt{ns}} \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}c)^2 \right)^{-1},$$

where $\mathbf{x}_1$ is the first entry of $\mathbf{x}$ and $\mathbf{x}_{-1} := (\mathbf{x}_2, \ldots, \mathbf{x}_p)$. In other words, unlike in model (1.1), the exact value of s does not play a prominent role. We thus see that in order to be able to improve over $\Theta_{1,1}$ we must have that Θ_1 is rather non-sparse: $\|\Theta_1\|_1$ should be of larger order than $\sqrt{ns}$. For example if $s = o(n/\log p)$, *say $s = n/(m_n \log p)$ for some sequence $m_n \to \infty$ slowly, the model directions have ℓ_1-norm of order $n/\sqrt{m_n \log p}$. To improve over $\Theta_{1,1}$ one thus must have $\|\Theta_1\|_1$ of larger order, say $\|\Theta_1\|_1 = n/\sqrt{\log p}$. We get back to this in Example 2.2.*

Example 1.3. *We now turn to model (1.3). Using the same arguments as in Example 1.2 one sees that if β^0 stays away from the boundary, one may use model directions with $\| \cdot \|_r^r$ of order $\sqrt{n^r s^{2-r}}$. The lower bound is then*

$$V_{\beta^0}^2 + o(1) \geq \left(\min_{c \in \mathbb{R}^{p-1}:\ \|c\|_r^r \leq M\sqrt{n^r s^{2-r}}} \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1} c)^2 \right)^{-1}$$

where $M > 0$ is any fixed constant (not depending on n).

It is clear, and illustrated by Examples 1.1, 1.2 and 1.3, that the lower bound of Proposition 1.1 depends on the model $\mathcal{B}$. The sparse model (1.1) is perhaps too stringent. One may want to take the model $\mathcal{B}$ as the largest set for which a regular estimator exists. This points in the direction of model (1.2). We will see that when Σ is known this model is indeed useful but when Σ is unknown it is too large.

1.3. Notations and definitions

We consider an asymptotic framework with triangular arrays of observations. Thus, unless explicitly stated otherwise, all quantities depend on n although we do not (always) express this in our notation.

The order symbols refer to asymptotics with sample size $n \to \infty$. Thus, for sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, the notation $a_n = \mathcal{O}(b_n)$ means that $\limsup_{n\to\infty} a_n/b_n < \infty$ and $a_n = o(b_n)$ means that $\lim_{n\to\infty} a_n/b_n = 0$. Moreover, $a_n \asymp b_n$ means that both $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$ hold. Finally, $a_n \gg 0$ means that the sequence $\{a_n\}$ stays away from zero, i.e. that $1/a_n = \mathcal{O}(1)$.

Let $\mathbf{x}_1$ be the first entry of $\mathbf{x}$ and $\mathbf{x}_{-1} := (\mathbf{x}_2, \ldots, \mathbf{x}_p)$ be this vector with the first entry excluded, so that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_{-1})$. Write $\Sigma_{-1,-1} := \mathbb{E}\mathbf{x}_{-1}^T \mathbf{x}_{-1} \in \mathbb{R}^{(p-1)\times(p-1)}$. For vectors $b \in \mathbb{R}^p$ be use a similar notation: $b_1 \in \mathbb{R}$ is the first coefficient and $b_{-1} \in \mathbb{R}^{p-1}$ forms the rest of the coefficients. Apart from the regression (projection) $\mathbf{x}\beta^0$ of $\mathbf{y}$ on $\mathbf{x}$, we consider the regression $\mathbf{x}_{-1}\gamma^0$ of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$. Thus $\gamma^0 = \Sigma_{-1,-1}^{-1} \mathbb{E}\mathbf{x}_{-1}^T \mathbf{x}_1$. Note that γ^0 is proportional to $-(\Theta_1)_{-1}$: $\gamma^0 = -(\Theta_1)_{-1}/\Theta_{1,1}$. Moreover $1/\Theta_{1,1} = \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^0)^2$ is the squared residual of the projection of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$.

We define when possible an approximation $\gamma^\sharp$ of γ^0 which is accompanied by a parameter $\lambda^\sharp$ to form an "eligible pair" $(\gamma^\sharp, \lambda^\sharp)$, see Definition 2.1. When Σ is known we can invoke the noiseless Lasso γ_{Lasso} with tuning parameter λ_{Lasso} (an approximate projection of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$) to approximate γ^0. See (3.1) for its definition. For the case Σ is unknown we apply the notation $\hat{\Sigma} := X^T X/n$. We let $X_1 \in \mathbb{R}^n$ be the first column of X, and the matrix $X_{-1} \in \mathbb{R}^{n\times(p-1)}$ be the remaining columns and we write $\hat{\Sigma}_{-1,-1} := X_{-1}^T X_{-1}/n$. We do an approximate regression of X_1 on X_{-1} invoking the noisy Lasso $\hat{\gamma}$ with tuning parameter λ^{Lasso} as given in (4.1). The various vectors of coefficients and their "lambda parameter" are summarized in Table 1. Here we also add the Lasso $\hat{\beta}$ for the estimation of β^0, as defined in (1.4) with tuning parameter λ.

Table 1
The various coefficients and lambda parameters.

	coefficients	lambda parameter
projection $\mathbf{y}$ on $\mathbf{x}$	β^0	0
noisy Lasso	$\hat{\beta}$	$\lambda \asymp \sqrt{\log p/n}$
projection $\mathbf{x}_1$ on $\mathbf{x}_{-1}$	γ^0	0
eligible pair	$\gamma^\sharp$	$\lambda^\sharp$
noiseless Lasso	γ_{Lasso}	λ_{Lasso}
noisy Lasso	$\hat{\gamma}$	$\lambda^{\text{Lasso}} \asymp \sqrt{\log p/n}$

We write for $S \in \{1, \ldots, p\}$, $\mathbf{x}_S := \{\mathbf{x}_j\}_{j \in S}$ and $\mathbf{x}_{-S} := \{\mathbf{x}_j\}_{j \notin S,\ j \neq 1}$. We further let for $S \subset \{1, \ldots, p\}$ with cardinality s, the matrix $\Sigma_{S,S} := \mathbb{E}\mathbf{x}_S^T \mathbf{x}_S \in \mathbb{R}^{\text{s} \times \text{s}}$ be the covariance sub-matrix formed by the variables in S and $\Sigma_{-S,-S} := \mathbb{E}\mathbf{x}_{-S}^T \mathbf{x}_{-S} \in \mathbb{R}^{(p-1-\text{s}) \times (p-1-\text{s})}$ and $\Sigma_{S,-S} := \mathbb{E}\mathbf{x}_S^T \mathbf{x}_{-S} =: \Sigma_{-S,S}^T \in \mathbb{R}^{\text{s} \times (p-1-\text{s})}$.

Note that the vector of coefficients γ^0 of the regression of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$ is a vector in $\mathbb{R}^{p-1}$. We will index its entries by $\{2, \ldots, p\}$: $\gamma^0 = (\gamma_2^0, \ldots, \gamma_p^0)^T$. Also the various other "gamma" parameters" γ will be indexed by $\{2, \ldots, p\}$. It should be clear from the context when this indexing applies. For $c = (c_2, \ldots, c_p)$ and $S \subset \{2, \ldots, p\}$ we write $c_S := \{c_j :\ j \in S\}$ and $c_{-S} := \{c_j :\ j \notin S,\ j \neq 1\}$. We use the same notation for the $(p-1)$-dimensional vector c_S which has the entries not in S set to zero, and we then let $c_{-S} = c - c_S$.

For a positive semi-definite matrix A we let $\Lambda_{\min}^2(A)$ be its smallest eigenvalue and $\Lambda_{\max}^2(A)$ be its largest eigenvalue. The smallest eigenvalue of Σ is written shorthand as $\Lambda_{\min}^2 := \Lambda_{\min}^2(\Sigma)$. Recall we assume throughout that $\Lambda_{\min}^2$ stays away from zero: $1/\Lambda_{\min}^2 = \mathcal{O}(1)$. We use the shorthand notation $\gg 0$ for "strictly positive and staying away from zero". Thus throughout we assume $\Lambda_{\min}^2 \gg 0$.

In order to be able to construct confidence intervals one needs some uniformity in unknown parameters. We give the following definition (see also [3], Definition 1 on page 18).

Definition 1.1. *Let $\mathcal{B}$ be the model for β^0. Let $\{\mathbf{Z}_n\}$ be a sequence of real-valued random variables depending on (X, Y) and $\{r_n\}$ a sequence of positive numbers. We say that $\{\mathbf{Z}_n\}$ is $\mathcal{O}_{\mathbf{P}_{\beta^0}}(r_n)$ uniformly in $\beta^0 \in \mathcal{B}$ if*

$$\lim_{M \to \infty} \limsup_{n \to \infty} \sup_{\beta^0 \in \mathcal{B}} \mathbb{P}_{\beta^0}(|\mathbf{Z}_n| > M r_n) = 0.$$

We say that $\mathbf{Z}_n = o_{\mathbf{P}_{\beta^0}}(r_n)$ uniformly in $\beta^0 \in \mathcal{B}$ if

$$\lim_{n \to \infty} \sup_{\beta_0 \in \mathcal{B}} \mathbb{P}_{\beta^0}\biggl(|\mathbf{Z}_n| > \eta r_n\biggr) = 0,\ \forall\ \eta > 0.$$

1.4. *Organization of the rest of the paper*

Section 2 contains the results for the case of Σ known and applying a debiased Lasso using sample splitting. Here we also introduce the concept of an eligible pair $(\gamma^\sharp, \lambda^\sharp)$ in Definition 2.1. Section 3 contains results and constructions for eligible pairs. Section 4 considers the case Σ unknown and a debiased Lasso (without sample splitting). Section 5 concludes and Section 6 collects the proofs. Section 7 (included for completeness) contains some elementary probability inequalities for products of Gaussians, which are applied in Section 4.

In Tables 2 and 3 we summarize the (sparsity) conditions we use, see Examples 2.1, 2.2 and 2.3 for the case Σ known and Examples 4.1 and 4.2 for the case Σ unknown. The particular cases $r = 0$ and $r = 1$ follow from the general case $0 \le r \le 1$ when Σ is known. When Σ is unknown the case $r = 0$ also follows from the general case $0 \le r < 1$. With $r = 1$ the model is then too large. We have displayed the extreme cases separately so that the conditions for these can be read off directly. In particular for $r = 0$ one sees the standard sparsity conditions known from the literature. For $r = 1$ (Σ known) one sees that unlike in the other cases there is no logarithmic gap between conditions for asymptotic normality and asymptotic efficiency.

TABLE 2

The conditions used to prove asymptotic normality, linearity and efficiency when Σ is known. Throughout, $(\gamma^\sharp, \lambda^\sharp)$ is required to be an eligible pair (see Definition 2.1), i.e. $\|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty \le \lambda^\sharp$ (and $\lambda^\sharp\|\gamma^\sharp\|_1 \to 0$). Asymptotic efficiency is established when β^0 stays away from the boundary of $\mathcal{B}$. In the case $\mathcal{B} = \{\|b\|_0^0 \le s\}$ the conditions on $\gamma^\sharp$ for asymptotic efficiency depend on β^0.

	Σ known $\mathcal{B} = \{\|b\|_0^0 \le s\}$	Σ known $\mathcal{B} = \{\|b\|_1 \le \sqrt{s}\}$	Σ known $\mathcal{B} = \{\|b\|_r^r \le \sqrt{s^{2-r}}\}$
asymptotic normality	$s = o(\frac{n}{\log p})$ $\lambda^\sharp s \log^{\frac{1}{2}} p = o(1)$ $\lambda^\sharp\|\gamma^\sharp\|_1 = o(1)$	$s = o(\frac{n}{\log p})$ $\lambda^\sharp\sqrt{ns} = o(1)$ $\lambda^\sharp\|\gamma^\sharp\|_1 = o(1)$	$s = o(\frac{n}{\log p})$ $\lambda^\sharp n^{\frac{r}{2}} s^{\frac{2-r}{2}} \log^{\frac{1-r}{2}} p = o(1)$ $\lambda^\sharp\|\gamma^\sharp\|_1 = o(1)$
asymptotic linearity	yes	yes	yes
asymptotic efficiency	$\|\gamma^\sharp\|_0^0 = \mathcal{O}(s)$	$\|\gamma^\sharp\|_1 = \mathcal{O}(\sqrt{ns})$	$\|\gamma^\sharp\|_r^r = \mathcal{O}(n^{\frac{r}{2}} s^{\frac{2-r}{2}})$

2. The case of Σ known

Before presenting "eligible pairs" in Definition 2.1, we provide the motivation that led us to this concept.

Recall the debiased Lasso given in (1.5). If Σ is known we choose the direction

TABLE 3

The conditions used to prove asymptotic normality, linearity and efficiency when Σ is unknown. Throughout, $(\gamma^\sharp, \lambda^\sharp)$ is required to be an eligible pair (see Definition 2.1), i.e. $\|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty \le \lambda^\sharp$ (and $\lambda^\sharp \|\gamma^\sharp\|_1 \to 0$). The value of r may be different from r. It is assumed to be fixed and $0 \le \mathrm{r} \le 1$. Asymptotic efficiency is established when β^0 stays away from the boundary of $\mathcal{B}$. In the case $\mathcal{B} = \{\|b\|_0^0 \le s\}$ the conditions on $\gamma^\sharp$ for asymptotic efficiency depend on β^0.

	Σ unknown $\mathcal{B} = \{\|b\|_0 \le s\}$	Σ unknown $\mathcal{B} = \{\|b\|_r^r \le \sqrt{s^{2-r}}\}$ $0 \le r < 1$
asymptotic normality	$s = o(\frac{\sqrt{n}}{\log p})$ $\lambda^\sharp = \mathcal{O}(\sqrt{\frac{\log p}{n}})$ $\sqrt{\frac{\log p}{n}} \|\gamma^\sharp\|_1 = o(1)$	$s = o(n^{\frac{1-r}{2-r}} / \log p)$ $\lambda^\sharp = \mathcal{O}(\sqrt{\frac{\log p}{n}})$ $\sqrt{\frac{\log p}{n}} \|\gamma^\sharp\|_1 = o(1)$
asymptotic linearity	$\|\gamma^\sharp\|_{\mathrm{r}}^{\mathrm{r}} = o(n^{\frac{1-\mathrm{r}}{2}} / \log^{\frac{2-\mathrm{r}}{2}} p)$	$\|\gamma^\sharp\|_{\mathrm{r}}^{\mathrm{r}} = o(n^{\frac{1-\mathrm{r}}{2}} / \log^{\frac{2-\mathrm{r}}{2}} p)$
asymptotic efficiency	$\|\gamma^\sharp\|_0^0 = \mathcal{O}(s)$	$\|\gamma^\sharp\|_r^r = \mathcal{O}(n^{\frac{r}{2}} s^{\frac{2-r}{2}})$

$\tilde{\Theta}_1 =: \Theta_1^\sharp$ non-random, depending on Σ. We invoke the decomposition

$$\hat{b}_1 - \beta_1^0 = \Theta_1^{\sharp T} X^T \epsilon / n + \underbrace{(\mathrm{e}_1 - \hat{\Sigma}\Theta^\sharp)^T (\hat{\beta} - \beta^0)}_{\text{remainder}}.$$

The remainder is

$$(\mathrm{e}_1 - \hat{\Sigma}\Theta^\sharp)^T (\hat{\beta} - \beta^0) = \underbrace{\Theta_1^{\sharp T} (\Sigma - \hat{\Sigma})(\hat{\beta} - \beta^0)}_{:=(i)} + \underbrace{(\mathrm{e}_1 - \Sigma\Theta^\sharp)^T (\hat{\beta} - \beta^0)}_{:=(ii)}.$$

The first term (i) can be handled assuming $\Theta_1^{\sharp T} \Sigma \Theta_1^\sharp = \mathcal{O}(1)$ and $\|\Sigma^{1/2}(\hat{\beta} - \beta^0)\|_2 = o_{\mathbf{P}}(1)$. This goes along the lines of techniques as in [11], applying the conditions used there. One then arrives at $(i) = o_{\mathbf{P}}(1/\sqrt{n})$. (We will however alternatively use a sample splitting technique later on in Theorem 2.1 to simplify the derivations.)

The second term (ii) is additional bias and will be our major concern. If $\Theta_1^\sharp = \Theta_1$ this term vanishes. However, as we will see it is useful to apply instead of Θ_1 some sparse approximation of Θ_1. In fact, we aim at a sparse approximation $\Theta_1^\sharp$ with $\Theta_{1,1}^\sharp$ being smaller than $\Theta_{1,1}$ and their difference not vanishing.

We will assume conditions that ensure the additional bias is negligible and invoke that

$$|(\mathrm{e}_1 - \Sigma\Theta_1^\sharp)^T (\hat{\beta} - \beta^0)| \le \|\Sigma(\Theta_1^\sharp - \Theta_1)\|_\infty \|\hat{\beta} - \beta^0\|_1 \tag{2.1}$$

(recall that by the definition of Θ_1 it is true that $\mathrm{e}_1 = \Sigma\Theta_1$).

Remark 2.1. *One may think of applying instead the Cauchy-Schwarz inequality*

$$|(\mathrm{e}_1 - \Sigma\Theta_1^\sharp)^T(\hat\beta - \beta^0)| \le \|\Sigma^{1/2}(\Theta_1^\sharp - \Theta_1)\|_2 \|\Sigma^{1/2}(\hat\beta - \beta^0)\|_2.$$

This leads to requiring that $\|\Sigma^{1/2}(\Theta_1^\sharp - \Theta_1)\|_2 \to 0$ fast enough. But we actually want $\|\Sigma^{1/2}(\Theta_1^\sharp - \Theta_1)\|_2 \not\to 0$ in order to be able to arrive at an improvement over the asymptotic variance.

Remark 2.2. *Consider now for some $\mathrm{p} \ge 1$ the general dual norm inequality*

$$|(\mathrm{e}_1 - \Sigma\Theta^\sharp)^T(\hat\beta - \beta^0)| \le \|\Sigma(\Theta_1^\sharp - \Theta_1)\|_{\mathrm{p}} \|\hat\beta - \beta^0\|_{\mathrm{q}}$$

where $1/\mathrm{p} + 1/\mathrm{q} = 1$. Choosing $\mathrm{p} \le 2$ here again works against our aim to improve the asymptotic variance. Thus we need to choose $p > 2$ (and therefore $\mathrm{q} < 2$). This certifies the choice $\mathrm{p} = \infty$ as being quite natural.

Remark 2.3. *We note that $\|\Sigma^{1/2}(\Theta_1^\sharp - \Theta_1)\|_2 \le \|\Sigma(\Theta_1^\sharp - \Theta_1)\|_2/\Lambda_{\min}$. This means we want $\|\Sigma(\Theta_1^\sharp - \Theta_1)\|_2 \not\to 0$ as we assume that $\Lambda_{\min}$ stays away from zero. Now it is clear that if for some vector $v \in \mathbb{R}^p$, it holds that $\|v\|_\infty \le \lambda_0^\sharp$, then $\|v\|_2 \le \sqrt{p}\lambda_0^\sharp$. So $\|v\|_2 \not\to 0$ implies $p\lambda_0^{\sharp 2} \not\to 0$. In other words, we can only improve the asymptotic variance in the high-dimensional case.*

Taking the dual norm inequality (2.1) as starting point we now need

$$\|\Sigma(\Theta_1^\sharp - \Theta_1)\|_\infty \le \lambda_0^\sharp \tag{2.2}$$

for some constant $\lambda_0^\sharp$ small enough, such that uniformly in $\beta^0 \in \mathcal{B}$

$$\lambda_0^\sharp \|\hat\beta - \beta^0\|_1 = o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n}).$$

With the above considerations as motivation, we now concentrate on an ℓ_∞-condition as given in inequality (2.2). We settle for some $\lambda^\sharp$ and construct vectors $\Theta_1^\sharp$ for which inequality (2.2) holds. It is based on replacing the vector of coefficients γ^0 of the regression of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$ by a sparse approximation $\gamma^\sharp$.

Definition 2.1. *Let $\gamma^\sharp \in \mathbb{R}^{p-1}$ and $\lambda^\sharp > 0$. We say that the pair $(\gamma^\sharp, \lambda^\sharp)$ is eligible if*

$$\|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty \le \lambda^\sharp \tag{2.3}$$

and

$$\lambda^\sharp \|\gamma^\sharp\|_1 \to 0. \tag{2.4}$$

Clearly, if for a vector $\gamma^\sharp \in \mathbb{R}^{p-1}$,

$$\|\gamma^\sharp\|_1 \|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty \to 0,$$

then $(\gamma^\sharp, \|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty)$ is an eligible pair. However, as we will see in the last statement in the next lemma, we aim in Definition 2.1 at eligible pairs $(\gamma^\sharp, \lambda^\sharp)$ with $\lambda^\sharp$ a large value (instead of the smallest value) such that (2.3) and (2.4) are met.

The conditions in Definition 2.1 will allow us to arrive at (2.2) as is shown in the next lemma.

Lemma 2.1. *Suppose $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. Then*

$$1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp \geq \Lambda_{\min}^2 - o(1)$$

i.e., $1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp \gg 0$ eventually. Let (for n sufficiently large)

$$\Theta_1^\sharp := \begin{pmatrix} 1 \\ -\gamma^\sharp \end{pmatrix} / (1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp). \tag{2.5}$$

Then we have

$$\|\Sigma(\Theta_1^\sharp - \Theta_1)\|_\infty \leq \lambda_0^\sharp$$

where

$$\lambda_0^\sharp := \lambda^\sharp / (1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp) \asymp \lambda^\sharp.$$

Moreover,

$$\begin{aligned} \Theta_1^{\sharp T}\Sigma\Theta_1^\sharp &= \Theta_{1,1}^\sharp + o(1) \\ &\leq \Theta_{1,1} + o(1). \end{aligned}$$

Finally, in order to have a non-vanishing improvement of $\Theta_{1,1}^\sharp$ over $\Theta_{1,1}$ it must be true that γ^0 is not sparse, in the sense that

$$\lambda^\sharp \|\gamma^0\|_1 \gg 0.$$

Remark 2.4. *The first condition (2.3) of Definition 2.1 can be rewritten as*

$$\mathbf{x}_{-1}\gamma^0 = \mathbf{x}_{-1}\gamma^\sharp + \xi^0, \ |\mathbb{E}\mathbf{x}_j\xi^0| \leq \lambda^\sharp \ \forall \ j \in \{2, \ldots, p\}.$$

The second condition[2] (2.4) in this definition can be thought of as a sparsity condition on $\gamma^\sharp$. The two conditions together require that the regression of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$ is sparse when one relaxes the orthogonality condition of residuals to approximate orthogonality. One may think of γ^0 as a "least squares estimate" of $\gamma^\sharp$ in a noisy regression model. This leads to a very natural interpretation of eligible pairs. We refer to Subsection 3.5.2 for details. Further, for $\Theta_1^\sharp$ defined in (2.5) one has the equivalence

$$\Theta_{1,1} - \Theta_{1,1}^\sharp \gg 0 \ \Leftrightarrow \ \mathbb{E}(\xi^0)^2 \gg 0.$$

To have $\Theta_{1,1}^\sharp$ improving over $\Theta_{1,1}$ we see from the above lemma that we aim at a situation where γ^0, and hence Θ_1, is not sparse, but where γ^0 can be replaced by a sparse vector $\gamma^\sharp$. For some special Σ's, we give examples of eligible pairs in Section 3. That section also discusses for a given $\lambda^\sharp$ uniqueness of the vectors $\gamma^\sharp$ for which the pair $(\gamma^\sharp, \lambda^\sharp)$ is eligible. Moreover, we show cases where $\mathbf{x}_{-1}\gamma^\sharp$ is an approximation of the projection of $\mathbf{x}_1$ on a subset of $\mathbf{x}_S$ of the other

[2] Condition (2.4) is of the same nature as the condition $\lambda\|\beta^0\|_1 = o(1)$ (which follows from the classical condition $\lambda^2\|\beta^0\|_0^0 = o(1)$ if $\|\beta^0\|_2 = \mathcal{O}(1)$) when applying the Lasso (1.4) with tuning parameter λ.

variables for some $S \subset \{2, \ldots, p\}$, see Lemma 3.5. This is why the Cramér Rao lower bound can be achieved in those cases.

Lemma 2.1 has all the ingredients to prove asymptotic normality of the debiased Lasso (1.5) with direction $\tilde{\Theta}_1 = \Theta_1^\sharp$ and $\Theta_1^\sharp$ given in (2.5) in this lemma. It can be done along the lines of Theorem 3.8 in [11], assuming the conditions stated there. However, as the authors point out, when using instead the sample splitting approach their Assumption (iii) is not needed. It is also mathematically less involved in the present context. Sampling splitting techniques date back at least to [22]. We use the following.

Assume the sample size n is even. Define the matrices

$$\begin{aligned} (X_I, Y_I) &:= \{X_{i,1}, \ldots, X_{i,p}, Y_i\}_{1 \le i \le n/2} \in \mathbb{R}^{n/2 \times (p+1)}, \\ (X_{II}, Y_I) &:= \{X_{i,1}, \ldots, X_{i,p}, Y_i\}_{n/2 < i \le n} \in \mathbb{R}^{n/2 \times (p+1)}. \end{aligned}$$

Let $\hat{\beta}_I$ be an estimator of β^0 based on the first half (X_I, Y_I) of the sample, for instance the Lasso estimator $\arg\min\{\|Y_I - X_I b\|_2^2/n + \lambda\|b\|_1\}$. Similarly, let $\hat{\beta}_{II}$ be an estimator of β^0 based on the second half (X_{II}, Y_{II}) of the sample. Let $(\gamma^\sharp, \lambda^\sharp)$ be an eligible pair. We then define the two debiased estimators

$$\begin{aligned} \hat{b}_{I,1}^\sharp &:= \hat{\beta}_{II,1} + 2\Theta_1^{\sharp T} X_I^T \Big(Y_I - X_I \hat{\beta}_{II} \Big)/n \\ \hat{b}_{II,1}^\sharp &:= \hat{\beta}_{I,1} + 2\Theta_1^{\sharp T} X_{II}^T \Big(Y_{II} - X_{II} \hat{\beta}_{I} \Big)/n \end{aligned}$$

where $\Theta_1^\sharp$ is given in (2.5) in Lemma 2.1. The final estimator $\hat{b}_1^\sharp$ is obtained by averaging these two:

$$\hat{b}_1^\sharp := \frac{\hat{b}_{I,1}^\sharp + \hat{b}_{II,1}^\sharp}{2}. \tag{2.6}$$

Let now $\mathcal{B}$ be a given model class for the unknown vector of regression coefficients β^0.

Theorem 2.1. *Let $(\gamma^\sharp, \lambda^\sharp)$ be an eligible pair, $\Theta_1^\sharp$ be given in (2.5) and $\hat{b}_1^\sharp$ be given in (2.6) with $\hat{b}_{I,1}^\sharp$ and $\hat{b}_{II,1}^\sharp$ the debiased estimators based on $\Theta_1^\sharp$ using the splitted sample. Suppose that uniformly in $\beta^0 \in \mathcal{B}$*

$$\|\Sigma^{1/2}(\hat{\beta}_I - \beta^0)\|_2 = o_{\mathbf{P}_{\beta^0}}(1), \ \ \|\Sigma^{1/2}(\hat{\beta}_{II} - \beta^0)\|_2 = o_{\mathbf{P}_{\beta^0}}(1) \tag{2.7}$$

and

$$\sqrt{n}\lambda^\sharp \|\hat{\beta}_I - \beta^0\|_1 = o_{\mathbf{P}_{\beta^0}}(1), \ \ \sqrt{n}\lambda^\sharp \|\hat{\beta}_{II} - \beta^0\|_1 = o_{\mathbf{P}_{\beta^0}}(1). \tag{2.8}$$

Then, uniformly in $\beta^0 \in \mathcal{B}$,

$$\hat{b}_1^\sharp - \beta_1^0 = \Theta_1^{\sharp T} X^T \epsilon/n + o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n}),$$

and

$$\lim_{n\to\infty} \sup_{\beta_0 \in \mathcal{B}} \mathbb{P}_{\beta^0} \left(\frac{\sqrt{n}(\hat{b}_1^\sharp - \beta_1^0)}{\Theta_{1,1}^\sharp} \le z \right) = \Phi(z), \ \ \forall \ z \in \mathbb{R}.$$

Corollary 2.1. *Theorem 2.1 shows that under its conditions the estimator* $\hat{b}_1^\sharp$ *is uniformly asymptotically linear and regular. It means that for this estimator the Cramér Rao lower bound as given in Subsection 1.2 is relevant. Depending on* $\mathcal{B}$ *and* Σ *it does or does not reach the Cramér Rao lower bound, see Remark 2.5.*

Remark 2.5. *Lemmas 3.7, 3.8 and 3.10 in Subsection 3.5 present examples of eligible pairs* $(\gamma^\sharp, \lambda^\sharp)$ *where* $\Theta_{1,1} - \Theta_{1,1}^\sharp$ *(with* $\Theta_{1,1}^\sharp$ *given in (2.5)) is non-vanishing:* $\Theta_{1,1} - \Theta_{1,1}^\sharp \gg 0$*. Thus for those cases Theorem 2.1 shows an asymptotic variance remaining strictly smaller than* $\Theta_{1,1}$*. Moreover, the constructions of Lemmas 3.7, 3.8 and 3.10 allow for directions* $\Theta_1^\sharp$ *(depending on* $\|\beta^0\|_0^0$*) in* $\mathcal{H}_{\beta^0}$ *in model (1.1). In models (1.2) and (1.3) the direction* $\Theta_1^\sharp$ *lies in* $\mathcal{H}_{\beta^0}$ *after scaling where only the scaling depends on* β^0*. For these cases (special constructions of* Σ*) the Cramér Rao lower bound is therefore achieved.*

We now discuss the requirements (2.7) and (2.8) for the models (1.1), (1.2) and (1.3). The overall picture is summarized in Table 2.

Example 2.1. *Consider model model (1.1) with* $s = o(n/\log p)$ *(sparsity variant (i)). For the Lasso estimator* $\hat{\beta}$ *given in (1.4), with appropriate choice of the tuning parameter* $\lambda \asymp \sqrt{\log p/n}$*, one has uniformly in* $\beta^0 \in \mathcal{B}$

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^0)\|_2^2 = \mathcal{O}_{\mathbf{P}_{\beta^0}}(s \log p/n), \ \|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbf{P}_{\beta^0}}(s\sqrt{\log p/n}).$$

This follows from e.g. Theorem 6.1 in [5], together with results for Gaussian quadratic forms as given in (4.3). So for $\hat{\beta}_I$ *and* $\hat{\beta}_{II}$ *Lasso's based on the splitted sample with suitable tuning parameter* λ*, the requirement on* $\lambda^\sharp$ *becomes*

$$\lambda^\sharp s\sqrt{\log p} = o(1).$$

This is guaranteed when $\lambda^\sharp = \mathcal{O}(\sqrt{\log p}/n)$*. The smaller the sparsity* s*, the more room there is for improvement (i.e., the larger the collection of covariance matrices* Σ *that allow for improvement over* $\Theta_{1,1}$*). For example, if in fact* $s = o(\sqrt{n}/\log p)$ *(sparsity variant (ii)), we can take* $\lambda^\sharp = \mathcal{O}(\sqrt{\log p/n})$*. Finally, if* $\|\gamma^\sharp\|_0^0 = s_\sharp$ *with* $s_\sharp \le s - s_0 - 1$ *the estimator* $\hat{b}_1^\sharp$ *is asymptotically efficient.*

Remark 2.6. *In view of Theorem 2.1 and the statements of Example 2.1 and Remark 2.5, we see that for model (1.1) the debiased estimator of Theorem 3.8 in [11], which uses* $\tilde{\Theta}_1 = \Theta_1$*, is in certain cases asymptotically inefficient as a choice* $\tilde{\Theta}_1 = \Theta_1^\sharp \ne \Theta_1$ *can give an improvement in the asymptotic variance (and is then efficient for certain such cases). We see this happening in the next example (Example 2.2), where the model is (1.2):* $\mathcal{B} := \{b \in \mathbb{R}^p : \ \|b\|_1 \le \sqrt{s}\}$ *with* $0 < s = o(n/\log p)$ *and the matrix* Σ *is constructed following one of the Lemmas 3.7, 3.8 or 3.10.*

Example 2.2. *In this example we take the model (1.2) with* $0 < s = o(n/\log p)$ *(sparsity variant (i)). Let* $\hat{\beta}$ *be again the Lasso estimator given in (1.4) with* *appropriate choice of the tuning parameter $\lambda \asymp \sqrt{\log p/n}$. One may use a "slow rates" result: uniformly in $\beta^0 \in \mathcal{B}$ it is true that*

$$\|\Sigma^{1/2}(\hat\beta - \beta^0)\|_2 = o_{\mathbf{P}_{\beta^0}}(1),\ \|\hat\beta - \beta^0\|_1 = \mathcal{O}_{\mathbf{P}_{\beta^0}}(\sqrt{s}),$$

see for example [5], Theorem 6.3, and combine this with results for quadratic forms as given in (4.3). (The arguments for establishing these "slow rates" are in fact as in Lemma 4.1 for the node-wise Lasso.) Taking $\hat\beta_I$ and $\hat\beta_{II}$ again as appropriate Lasso's, condition (2.8) of Theorem 2.1 holds if

$$\lambda^\sharp\sqrt{ns} = o(1).$$

Then for $\|\gamma^\sharp\|_1 = \mathcal{O}(\sqrt{ns})$ we get an eligible pair $(\gamma^\sharp, \lambda^\sharp)$. Since when β^0 stays away from the boundary, such $\gamma^\sharp$ form a model direction after scaling, we see that the Cramér Rao lower bound is achieved. Recall now that in Example 1.2 we concluded that, in view of Corollary 1.1, in order to be able to improve over $\Theta_{1,1}$ we must have $\|\gamma^0\|_1$ of order larger than $\sqrt{ns}$. As pointed out in Remark 2.5 one may apply one of the Lemmas 3.7, 3.8 or 3.10 to create vectors γ^0 and eligible pairs $(\gamma^\sharp, \lambda^\sharp)$ where $\lambda^\sharp\sqrt{ns} \to 0$, $\|\gamma^\sharp\|_1$ can take any value of order at most $\sqrt{ns}$ and $\Theta_{1,1} - \Theta_{1,1}^\sharp \gg 0$.

Example 2.3. *The results for model (1.2) are a special case of those for model (1.3). One needs again $s = o(n/\log p)$ and one may for example apply Corollary 2.4 in [24], again together with (4.3). For the Lasso $\hat\beta$ in (1.4) with $\lambda \asymp \sqrt{\log p/n}$ appropriately chosen one gets uniformly in $\beta^0 \in \mathcal{B}$*

$$\|\Sigma^{1/2}(\hat\beta - \beta^0)\|_2^2 = \mathcal{O}_{\mathbf{P}_{\beta^0}}(s\log p/n)^{\frac{2-r}{2}},\ \|\hat\beta - \beta^0\|_1 = \mathcal{O}_{\mathbf{P}_{\beta^0}}((\log p/n)^{\frac{1-r}{2}} s^{\frac{2-r}{2}}).$$

The requirement (2.8) on $\lambda^\sharp$ is thus

$$\lambda^\sharp(\log p)^{\frac{1-r}{2}}\sqrt{n^r s^{2-r}} = o(1).$$

If $\|\gamma^\sharp\|_r^r = \mathcal{O}(\sqrt{n^r s^{2-r}})$ then $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair and the Cramér Rao lower bound is achieved whenever β^0 stays away from the boundary. In order to be able to improve over $\Theta_{1,1}$ we now need $\|\gamma^0\|_r^r$ of larger order $\sqrt{n^r s^{2-r}}$ by Corollary 1.1. Remark 2.5 can be taken into the considerations here too.

3. Finding eligible pairs

The main results of this section can be found in Subsection 3.5 where for any $\lambda^\sharp$ we construct eligible pairs $(\gamma^\sharp, \lambda^\sharp)$ by choosing γ^0 appropriately. These results can be seen as existence proofs. Before doing these constructions we discuss uniqueness in Subsection 3.1, in Subsection 3.2 the noiseless Lasso as a practical method for improving over $\Theta_{1,1}$ and in Subsection 3.3 we examine whether or not projections on a subset of the variables can lead to eligible pairs. For the latter we impose rather stringent conditions. We show in Subsection 3.4 that eligible pairs are more flexible than projections. Nevertheless in the final part of this section we return to projections as they come up naturally when imposing non-sparsity constraints on γ^0.

3.1. Uniqueness

Lemma 3.1. *Let $(\gamma^\sharp, \lambda^\sharp)$ and $(\gamma^\flat, \lambda^\sharp)$ be two eligible pairs. Then*

$$\|\Sigma_{-1,-1}^{1/2}(\gamma^\sharp - \gamma^\flat)\|_2^2 \to 0.$$

This lemma tells us that asymptotically it makes no difference to debias using $(\gamma^\sharp, \lambda^\sharp)$ or $(\gamma^\flat, \lambda^\sharp)$.

3.2. Using the Lasso

Consider the noiseless Lasso with tuning parameter $\lambda_{\rm Lasso}$:

$$\gamma_{\rm Lasso} := \arg \min_{c \in \mathbb{R}^{p-1}} \biggl\{ \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}c)^2 + 2\lambda_{\rm Lasso}\|c\|_1 \biggr\}. \tag{3.1}$$

One may verify that $(\gamma_{\rm Lasso}, \lambda_{\rm Lasso})$ is an eligible pair if $\lambda_{\rm Lasso}\|\gamma^0\|_1 \to 0$. But the latter is exactly what we want to avoid. If $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair then the noiseless Lasso can find it if one chooses $\lambda_{\rm Lasso} = \mathcal{O}(\lambda^\sharp)$ larger than $\lambda^\sharp$, as follows from the next lemma. It says that given $\lambda^\sharp$ one may use the noiseless Lasso for constructing a direction, $\Theta_{1,{\rm Lasso}}$ say, with which one has same improvement over $\Theta_{1,1}$ as with $\Theta_1^\sharp$.

Lemma 3.2. *Suppose $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. Let $\lambda_{\rm Lasso} > \lambda^\sharp$ and $\lambda_{\rm Lasso}\|\gamma^\sharp\|_1 \to 0$. Let $\gamma_{\rm Lasso}$ be the noiseless Lasso defined in (3.1). Then*

$$\|\Sigma_{-1,-1}^{1/2}(\gamma_{\rm Lasso} - \gamma^\sharp)\|_2^2 \to 0.$$

In addition, if $\lambda_{\rm Lasso} \ge 2\lambda^\sharp$ (say) then $(\gamma_{\rm Lasso}, \lambda_{\rm Lasso})$ is an eligible pair.

3.3. Using projections

In this subsection we investigate (rather straightforwardly) conditions such that the coefficients of a projection of $\mathbf{x}_{-1}\gamma^0$ can be joined with a $\lambda^\sharp$ to form an eligible pair.

Consider some set $S \subset \{2, \ldots, p\}$ with cardinality $\rm s$. The value of $\rm s$ need not be s, where s is the sparsity used in the model class $\mathcal{B}$. Let $\gamma_S^S := \Sigma_{S,S}^{-1}\mathbb{E}\mathbf{x}_S^T\mathbf{x}_{-1}\gamma^0$ so that $\mathbf{x}_S\gamma_S^S$ is the projection of $\mathbf{x}_{-1}\gamma^0$ on $\mathbf{x}_S$.

Let $\gamma^S \in \mathbb{R}^{p-1}$ be the vector $\gamma_S^S \in \mathbb{R}^{\rm s}$ completed with zeroes. Then $\|\gamma^S\|_1 \le \sqrt{\rm s}\|\gamma^S\|_2 = \mathcal{O}(\sqrt{\rm s})$ so that when

$$\lambda^\sharp\sqrt{\rm s} \to 0 \tag{3.2}$$

we have for γ^S the sparsity condition (2.4) of Definition 2.1: $\lambda^\sharp\|\gamma^S\|_1 \to 0$. Note that $\gamma^{0T}\Sigma_{-1,-1}\gamma^S = \gamma^{ST}\Sigma_{-1,-1}\gamma^S (= \mathbb{E}(\mathbf{x}_S\gamma_S^S)^2)$ exactly in this case. Furthermore

$$v^S := \Sigma_{-1,-1}(\gamma^0 - \gamma^S) = \begin{pmatrix} 0 \\ v_{-S}^S \end{pmatrix}$$

where

$$v^S_{-S} = \mathbb{E}\mathbf{x}^T_{-S}\mathbf{x}_{-1}\gamma^{-S},$$

with

$$\mathbf{x}_{-1}\gamma^{-S} = (\mathbf{x}_{-S}\gamma_{-S})\mathrm{A}\mathbf{x}_S =: \biggl(\mathbf{x}_{-S} - \mathbf{x}^T_S\Sigma^{-1}_{S,S}\Sigma_{S,-S}\biggr)\gamma^0_{-S}$$

being the anti-projection of $\mathbf{x}_{-S}\gamma^0_{-S}$ on $\mathbf{x}_S$. We check whether the pair $(\gamma^S, \lambda^\sharp)$ is an eligible pair, which is the case if $\lambda^\sharp\sqrt{\mathrm{s}} = o(1)$ and $\|v^S_{-S}\|_\infty \le \lambda^\sharp$. We briefly discuss some conditions that may help ensuring the latter.

Let $\|\!|A\|\!|_1 := \max_j \sum_k |a_{j,k}|$ be the ℓ_1-operator norm of the matrix A.

Lemma 3.3. *It holds that*

$$\|v^S_{-S}\|_\infty \le \|\!|\Sigma_{-S,-S} - \Sigma_{-S,S}\Sigma^{-1}_{S,S}\Sigma_{S,-S}\|\!|_1\|\gamma^0_{-S}\|_\infty.$$

To arrive at a sparse approximation of γ^0 one may consider putting its $p - \mathrm{s} - 1$ smallest-in-absolute-value coefficients to zero. To this end, consider the ordered sequence $|\gamma^0|_{(1)} \ge \cdots \ge |\gamma^0|_{(p-1)}$, write $|\gamma^0|_{(j)} =: |\gamma^0_{r_j}|$, $j = 1, \ldots, p-1$, and let $S = \{r_1, \ldots r_{\mathrm{s}}\}$. Then clearly $\|\gamma^0_{-S}\|_\infty \le \|\gamma^0\|_2/\sqrt{\mathrm{s}} = \mathcal{O}(1/\sqrt{s})$. The condition (3.2) excludes $1/\sqrt{\mathrm{s}} = \mathcal{O}(\lambda^\sharp)$. We therefore examine situations where the coefficients in γ^0 decrease at a rate quicker than $1/\sqrt{\mathrm{s}}$. The following definition is analogous to the definition of "effective sparsity" as given in [16].

Definition 3.1. *Let N be some integer and $0 \ne v \in \mathbb{R}^N$ be a vector. We call*

$$\kappa(v) := \frac{\|v\|_1\|v\|_\infty}{\|v\|_2}$$

the sparsity index of v. We say that v is asymptotically sparse if $\kappa(v) \to 0$.

A vector v with $\|v\|_2 = \mathcal{O}(1)$ can have some relatively large coefficients, but it cannot have too many of these. If in addition $\|v\|_1$ is large it cannot have many zeroes either. Asymptotic non-sparseness of the vector v with the large coefficients removed means that there are many very small non-zero coefficients.

Example 3.1. *Let $v_j = 1/\sqrt{N}$ for all j. Then $\|v\|_1 = 1/\|v\|_\infty = \sqrt{N}$. Thus $\kappa(v) = 1$ and v is not (asymptotically) sparse.*

Example 3.2. *Let $\mathrm{s} \le N$ and $v_j = 0$ for $j \le \mathrm{s}$ and $v_j = 1/\sqrt{j\log N}$ for $j > \mathrm{s}$. Then $\kappa(v) \asymp \sqrt{N/(\mathrm{s}\log^2 N)}$. The vector v is not asymptotically sparse if $\mathrm{s} = \mathcal{O}(N/(\log^2 N))$.*

Lemma 3.4. *Suppose that γ^0_{-S} is not asymptotically sparse. Let*

$$\lambda^S = \frac{\kappa(\gamma^0_{-S})}{C\|\gamma^0_{-S}\|_1},$$

where

$$C := \|\!|\Sigma_{-S,-S} - \Sigma_{-S,S}\Sigma^{-1}_{S,S}\Sigma_{S,-S}\|\!|_1.$$

Assume that $\lambda^S\sqrt{\mathrm{s}} \to 0$. Then (γ^S, λ^S) is an eligible pair and $\lambda^S\|\gamma^0\|_1 \not\to 0$.

3.4. Approximate projections

Recall that the first condition (2.3) of Definition 2.1 can be written as

$$\mathbf{x}_{-1}\gamma^0 = \mathbf{x}_{-1}\gamma^\sharp + \xi^0$$

where ξ^0 is "noise" satisfying $\|\mathbb{E}\mathbf{x}_{-1}^T\xi^0\|_\infty \le \lambda^\sharp$. Denote the active set of $\gamma^\sharp$ by $S = \{j \in \{2, \ldots, p\} : \gamma_j^\sharp \ne 0\}$ and its cardinality by $\mathrm{s} := |S|$. Then $\lambda^\sharp\sqrt{\mathrm{s}} \to 0$ implies that $1/\Theta_{1,1}^\sharp$ is asymptotically the squared residual of the projection of $\mathbf{x}_1$ on $\mathbf{x}_S$ as is shown in the next lemma. In that sense, eligible pairs are more flexible than projections.

If s is small enough, then for model (1.1), (1.2) or (1.3) one has that $\Theta_1^\sharp$ is a model direction (after rescaling). The estimator $\hat{b}_1^\sharp$ in Theorem 2.1 then reaches the Cramér Rao lower bound.

Lemma 3.5. *Suppose for some set S with cardinality s that $\gamma^\sharp = \gamma_S^\sharp$ and assume the first condition (2.3) of Definition 2.1. If $\lambda^\sharp\sqrt{\mathrm{s}} \to 0$ then the second condition (2.4) holds too so that $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. Moreover,*

$$1/\Theta_{1,1}^\sharp = \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_S\gamma_S^S)^2 + o(1)$$

where

$$\gamma_S^S = \Sigma_{S,S}^{-1}\mathbb{E}\mathbf{x}_S^T\mathbf{x}_1.$$

3.5. Reverse engineering

In this subsection we fix $\lambda^\sharp$ and then construct vectors $\gamma^0 \in \mathbb{R}^{p-1}$ such that there is an eligible pair $(\gamma^\sharp, \lambda^\sharp)$. These constructions are in a sense equivalent but approach the problem from different angles. In these constructions the vector $\gamma^\sharp$ is having active set $S = \{\gamma_j^\sharp \ne 0\}$ with cardinality $\mathrm{s} := |S|$. The sparsity of $\gamma^\sharp$ is then measured in terms of the value of s. More general constructions are possible, but in this way we can apply the results to any of the models (1.1), (1.2) or (1.3). In view of Lemma 3.5 is means that the constructions correspond to an approximate projection on $\mathbf{x}_S$.

Throughout this subsection, the matrix $\Sigma_{-1,-1}$ is assumed to have 1's on the diagonal and smallest eigenvalue $\Lambda_{\min}^2(\Sigma_{-1,-1}) \gg 0$.

3.5.1. Which γ^0's are allowed?

We let for $\gamma^0 \in \mathbb{R}^p$

$$\Sigma(\gamma^0) := \begin{pmatrix} 1 & \gamma^{0T}\Sigma_{-1,-1} \\ \Sigma_{-1,-1}\gamma^0 & \Sigma_{-1,-1} \end{pmatrix}.$$

Definition 3.2. *We say that the vector γ^0 is allowed if $\Sigma(\gamma^0)$ is positive definite, with $\Lambda_{\min}^2(\Sigma(\gamma^0)) \gg 0$ and $\|\Sigma_{-1,-1}\gamma^0\|_\infty \le 1$.*

Lemma 3.6. *Suppose that*

$$1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2^2 \gg 0.$$

Then γ^0 is allowed.

3.5.2. Regression: γ^0 as least squares estimate of $\gamma^\sharp$

In this subsection we create γ^0 using random noise. We then arrive at an eligible pair "with high probability". Let $N \in \mathbb{N}$ be a given sequence with $N > p$. Take a matrix $Z_{-1} \in \mathbb{R}^{N\times(p-1)}$ which has $Z_{-1}^T Z_{-1}/N = \Sigma_{-1,-1}$. Let $\xi \in \mathbb{R}^N$ have i.i.d. standard Gaussian entries, let $\gamma^\sharp \in \mathbb{R}^{p-1}$ be a vector satisfying $\|\gamma^\sharp\|_1 = o(\sqrt{\log p/N})$ and define

$$Z_1 := Z_{-1}\gamma^\sharp + \xi.$$

Let

$$\gamma^0 := (Z_{-1}^T Z_{-1})^{-1} Z_{-1}^T Z_1$$

be the least squares estimator of $\gamma^\sharp$. Finally let $\lambda^\sharp \asymp \sqrt{\log p/N}$ be appropriately chosen. Then $(\gamma^\sharp, \lambda^\sharp)$ is with high probability an eligible pair. Indeed the first condition (2.3) of Definition 2.1 follows from

$$\|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty = \|Z_{-1}^T\xi/N\|_\infty = \mathcal{O}_{\mathbf{P}}(\sqrt{\log p/N})$$

so that for appropriate $\lambda^\sharp \asymp \sqrt{\log p/N}$, with high probability

$$\|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty \le \lambda^\sharp.$$

The second condition (2.4) of Definition 2.1 follows from the condition $\|\gamma^\sharp\|_1 = o(\sqrt{\log p/N})$ so $\lambda^\sharp\|\gamma^\sharp\|_1 = o(1)$. We also see that if $p/N \gg 0$ then with high probability $\Theta_{1,1}^\sharp$ as given in (2.5) is an improvement over $\Theta_{1,1}$, since

$$\|\Sigma_{-1,-1}^{1/2}(\gamma^0 - \gamma^\sharp)\|_2^2 = \chi_{p-1}^2/N$$

where χ_{p-1}^2 has the chi-squared distribution with $p-1$ degrees of freedom. It follows that

$$\|\Sigma_{-1,-1}^{1/2}(\gamma^0 - \gamma^\sharp)\|_2^2 = \frac{p + \mathcal{O}_{\mathbf{P}}(\sqrt{p})}{N}.$$

With high probability this stays away from zero so that also

$$\Theta_{1,1} - \Theta_{1,1}^\sharp \gg 0.$$

For appropriate N with $1 - p/N \gg 0$ the vector γ^0 is with high probability also allowed by Lemma 3.6. With this choice of N we have $\lambda_\sharp \asymp \sqrt{\log p/p}$. Recall that according to Remark 2.3 it must be true that $p\lambda^{\sharp 2} \gg 0$. In the present context we in fact have $p\lambda^{\sharp 2}/\log p \gg 0$.

3.5.3. Creating γ^0 directly

We first recall that

$$\Theta_{1,1} - \Theta_{1,1}^\sharp \gg 0 \Leftrightarrow \|\Sigma_{-1,-1}^{1/2}(\gamma^0 - \gamma^\sharp)\|_2^2 \gg 0$$

with $\Theta_1^\sharp$ given in (2.5). It will be the case in the constructions of this subsection.

Consider some set $S \subset \{2, \ldots, p\}$ with cardinality s and some $\lambda^\sharp$. We will assume $\lambda^\sharp \sqrt{\mathrm{s}} \to 0$.

Lemma 3.7. *Suppose there exists a vector $z \in \mathbb{R}^{p-1}$ with*

$$\|z\|_\infty \le 1,$$

and

$$1 - \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z\|_2^2 \gg 0,\ \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z\|_2^2 \gg 0. \tag{3.3}$$

Let $\gamma^\sharp$ be a vector in $\mathbb{R}^{p-1}$ with $\gamma_{-S}^\sharp = 0$ (i.e. $\gamma^\sharp = \gamma_S^\sharp$) and

$$1 - \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z\|_2^2 - \|\Sigma_{-1,-1}^{1/2}\gamma_S^\sharp\|_2^2 \gg 0.$$

Define

$$\gamma^0 := \gamma^\sharp + \lambda^\sharp \Sigma_{-1,-1}^{-1} z.$$

Then, if $\lambda^\sharp\sqrt{\mathrm{s}} \to 0$, the pair $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. Moreover, γ^0 is eventually allowed, $\lambda^\sharp\|\gamma^0\|_1 \not\to 0$ and in fact

$$\|\Sigma_{-1,-1}^{1/2}(\gamma^0 - \gamma^\sharp)\|_2^2 \gg 0.$$

Remark 3.1. *Recall that $\Lambda_{\min}^2(\Sigma_{-1,-1}) \gg 0$. To check condition (3.3) for some $\|z\|_\infty \le 1$ one may want to impose in addition that $\Lambda_{\max}(\Sigma_{-1,-1}) = \mathcal{O}(1)$. Then the condition is true if $\|z\|_2 \asymp 1/\lambda^\sharp$ which is true for instance if $1/\lambda^{\sharp 2}$ of the coefficients of z stay away from zero. This is only possible if $p > 1/\lambda^{\sharp 2}$ i.e., in high-dimensional situations (see also Remark 2.3).*

Remark 3.2. *One may also consider taking $z = Z_{-1}^T\xi/(N\lambda^\sharp)$ where Z_{-1}, ξ and $\lambda^\sharp$ are is in Subsection 3.5.2. Then*

$$\lambda^{\sharp 2}\|\Sigma_{-1,-1}^{1/2} z\|_2^2 = \chi_{p-1}^2/N$$

as there and one arrives at eligible pairs "with high probability".

We now examine the following question: can one choose $\gamma_S^\sharp$ in Lemma 3.7 equal to γ_S^0? As we will see this will only be possible if a form of the irrepresentable condition holds. The "usual" irrepresentable condition (that implies the absence of false positives of the Lasso, see [27]) involves the coefficients of the projection of the "large" collection $\mathbf{x}_{-S}$ on the "small" collection $\mathbf{x}_S$. In our case, we reverse the roles of S and $-S$.

Definition 3.3. *Fix a vector $z_{-S} \in \mathbb{R}^{p-\mathrm{s}-1}$ with $\|z_{-S}\|_\infty \le 1$. We say that the reversed irrepresentable condition holds for (S, z_{-S}) if*

$$\|\Sigma_{S,-S}\Sigma_{-S,-S}^{-1} z_{-S}\|_\infty \le 1.$$

Lemma 3.8.
a) Assume the reversed irrepresentable condition holds for (S, z_{-S}), and that in addition

$$1 - \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z_{-S}\|_2^2 \gg 0,\ \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z_{-S}\|_2^2 \gg 0.$$

Let γ_S^0 satisfy

$$1 - \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} z_{-S}\|_2^2 - \|\Sigma_{-1,-1}^{1/2}\gamma_S^0\|_2^2 \gg 0.$$

Define $\gamma_{-S}^0 := \lambda^\sharp \Sigma_{-S,-S}^{-1} z_{-S}$ and $\gamma^\sharp := \gamma_S^0$ Then, if $\lambda^\sharp\sqrt{\mathrm{s}} \to 0$, the pair $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair, γ^0 is eventually allowed and $\lambda^\sharp\|\gamma^0\|_1 \not\to 0$. In fact

$$\|\Sigma_{-1,-1}^{1/2}(\gamma^0 - \gamma^\sharp)\|_2^2 \gg 0.$$

b) Conversely, if for some γ^0 and for $\gamma^\sharp := \gamma_S^0$ the pair $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair, then the reversed irrepresentable condition holds for (S, z_{-S}) with appropriate z_{-S} satisfying $\|z_{-S}\|_\infty \le 1$.

3.5.4. Creating γ^0 using a non-sparsity restriction

Consider some set $S \subset \{2, \ldots, p\}$ with cardinality s and some $\lambda^\sharp > 0$. Let $w \in \mathbb{R}^{p-1}$ be a vector of strictly positive weights with $\|w\|_\infty \le 1$ and define the matrix W as the diagonal matrix with w on its diagonal. Let

$$c^0 \in \arg\min\left\{\|\Sigma_{-1,-1}^{1/2} c\|_2^2 :\ \lambda^\sharp\|(Wc)_{-S}\|_1 = 1\right\}$$

and

$$\zeta_S := 0,\ \zeta_j := \mathrm{sign}(c_j^0),\ j \notin S.$$

Lemma 3.9. *The random variable $\mathbf{x}_{-1}c^0$ is orthogonal to (i.e. independent of) $\mathbf{x}_S$. Moreover*

$$\|\Sigma_{-1,-1}c^0\|_\infty = \frac{\|w_{-S}\|_\infty}{\lambda^\sharp\|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2}$$

and

$$\|\Sigma_{-1,-1}^{1/2} c^0\|_2^2 = \frac{1}{\lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2}.$$

Lemma 3.10. *Suppose that*

$$1 - (\lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2)^{-1} \gg 0,\ \lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2 = \mathcal{O}(1). \tag{3.4}$$

Let $\gamma^\sharp$ be a vector satisfying $\gamma_{-S}^\sharp = 0$ and

$$0 < 1 - (\lambda^{\sharp 2}\|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2)^{-1} - \|\Sigma_{-1,-1}^{1/2}\gamma_S^\sharp\|_2^2 \gg 0.$$

Define $\gamma^0_{-S} = c^0_{-S}$ and $\mathbf{x}_{-1}\gamma^0 := \mathbf{x}_S\gamma^\sharp_S + (\mathbf{x}_{-S}\mathrm{A}\mathbf{x}_S)\gamma^0_{-S}$. Then $\mathbf{x}_S\gamma^\sharp_S$ is the projection of $\mathbf{x}_{-1}\gamma^0$ on $\mathbf{x}_S$. Moreover, if $\lambda^\sharp\sqrt{\mathrm{s}} \to 0$, the pair $(\gamma^\sharp, \lambda^\sharp)$ is eligible, γ^0 is allowed and $\lambda^\sharp\|\gamma^0\|_1 \not\to 0$. In fact

$$\|\Sigma^{1/2}_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_2^2 \gg 0.$$

Remark 3.3. *As in Remark 3.1, to deal with the requirement (3.4) one may want to impose the condition $\Lambda^2_{\max}(\Sigma_{-1,-1}) = \mathcal{O}(1)$.*

4. The case of Σ unknown

As we will see the concept of an eligible pair will also play a crucial role when Σ is unknown.

We use in this section the noisy Lasso

$$\hat\gamma \in \arg\min_{c\in\mathbb{R}^{p-1}} \left\{ \|X_1 - X_{-1}c\|_2^2/n + 2\lambda^{\mathrm{Lasso}}\|c\|_1 \right\}. \tag{4.1}$$

We require the the tuning parameter λ^{Lasso} to be of order $\sqrt{\log p/n}$. Then in the debiased Lasso given in (1.5) we apply

$$\tilde\Theta_1 := \hat\Theta_1$$

where

$$\hat\Theta_1 := \begin{pmatrix} 1 \\ -\hat\gamma \end{pmatrix} / (\|X_1 - X_{-1}\hat\gamma\|_2^2/n + \lambda^{\mathrm{Lasso}}\|\hat\gamma\|_1). \tag{4.2}$$

Let $(\gamma^\sharp, \lambda^\sharp)$ be an eligible pair, with $\lambda^\sharp = \mathcal{O}(\sqrt{\log p/n})$. In Lemma 4.1, we in fact need λ^{Lasso} to be at least as large as $\lambda^\sharp$. The latter is allowed to be of small order $\sqrt{\log p/n}$, yet we will assume $\lambda^{\mathrm{Lasso}}\|\gamma^\sharp\|_1 \to 0$.

Let $\eta_n \to 0$ be a sequence such that

$$\mathbb{P}\left(\inf_{c:\ \lambda^{\mathrm{Lasso}}\|c\|_1 \le 4\eta_n^2,\ \|\Sigma^{1/2}_{-1,-1}c\|_2 = 1} \|X_{-1}c\|_2^2/n \ge \frac{1}{2} \right) \to 1. \tag{4.3}$$

Such a sequence exists, see for example Chapter 16 in [24] and its references.

Define $\varepsilon := X_1 - X_{-1}\gamma^0$ and $\varepsilon^\sharp := X_1 - X_{-1}\gamma^\sharp = \varepsilon + X_{-1}(\gamma^0 - \gamma^\sharp)$. Choose $\lambda^\sharp_\varepsilon \asymp \sqrt{\log p/n}$ in such a way that

$$\mathbb{P}\left(\|X^T\epsilon^\sharp\|_\infty/n \ge \lambda^\sharp_\varepsilon \right) \to 0$$

(see Lemma 7.3). This implies $\lambda^\sharp_\epsilon \ge \lambda^\sharp$.

The following lemma establishes the so-called "slow rate". The proof is standard (see for example [5], Theorem 6.3), up to the replacement of ε by $\varepsilon^\sharp$. The result is the noisy counterpart of Lemma 3.2.

Lemma 4.1. *Let $\lambda^{\text{Lasso}} \asymp \sqrt{\log p/n}$ satisfy $\lambda^{\text{Lasso}} \geq 2\lambda_\varepsilon^\sharp$ and suppose $\lambda^{\text{Lasso}} \|\gamma^\sharp\|_1 \leq \eta_n^2$. Then we have*

$$\lambda^{\text{Lasso}}\|\hat{\gamma}\|_1 = o_{\mathbf{P}}(1),\ \|X_{-1}(\hat{\gamma}-\gamma^\sharp)\|_2^2/n = o_{\mathbf{P}}(1),\ \|\Sigma_{-1,-1}^{1/2}(\hat{\gamma}-\gamma^\sharp)\|_2^2 = o_{\mathbf{P}}(1).$$

Theorem 4.1. *Let $\hat{b}_1$ be the debaised Lasso given in (1.5) with $\tilde{\Theta}_1$ equal to $\hat{\Theta}_1$ given in (4.2):*

$$\hat{b}_1 := \hat{\beta}_1 + \hat{\Theta}_1 X^T(Y - X\hat{\beta})/n.$$

Assume that $\lambda^\sharp = \mathcal{O}(\sqrt{\log p/n})$, that $\sqrt{\log p/n}\|\gamma^\sharp\|_1 = o(1)$ and $\lambda^{\text{Lasso}} \asymp \sqrt{\log p/n}$ is sufficiently large (depending only on $\lambda^\sharp$). Suppose that uniformly in $\beta^0 \in \mathcal{B}$

$$\sqrt{\log p}\|\hat{\beta} - \beta^0\|_1 = o_{\mathbf{P}_{\beta^0}}(1). \tag{4.4}$$

Then uniformly in $\beta^0 \in \mathcal{B}$

$$\hat{b}_1 - \beta_1^0 = \hat{\Theta}_1^T X^T \epsilon/n + o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n}).$$

Moreover,

$$\lim_{n\to\infty} \sup_{\beta^0 \in \mathcal{B}} \mathbb{P}\left(\sqrt{n}\frac{(\hat{b}_1 - \beta_1^0)}{\sqrt{\hat{\Theta}_1^T\hat{\Sigma}\hat{\Theta}_1}} \leq z\right) = \Phi(z)\ \forall\ z \in \mathbb{R},$$

and for $\Theta_1^\sharp$ given in (2.5)

$$\hat{\Theta}_1^T\hat{\Sigma}\hat{\Theta}_1 = \Theta_{1,1}^\sharp + o_{\mathbf{P}}(1).$$

Remark 4.1. *Recall that Lemmas 3.7, 3.8 and 3.10 present examples of eligible pairs $(\gamma^\sharp, \lambda^\sharp)$ where $\Theta_{1,1}^\sharp$ remains strictly smaller than $\Theta_{1,1}$.*

Remark 4.2. *By Slutsky's Theorem we conclude that the asymptotic variance of $\hat{b}_1$ is (up to smaller order terms) equal to $\Theta_{1,1}^\sharp$.*

Remark 4.3. *Note that Theorem 4.1 only requires sparsity of $\gamma^\sharp$ in ℓ_1-sense: it requires $\|\gamma^\sharp\|_1 = o(\sqrt{n/\log p})$. Indeed, we in fact base the result on "slow rates" for the Lasso $\hat{\gamma}$ given in Lemma 4.1.*

Remark 4.4. *Assume the conditions of Theorem 4.1, and that in fact*

$$\sqrt{\log p}\|\hat{\gamma} - \gamma^\sharp\|_1 = o_{\mathbf{P}}(1). \tag{4.5}$$

Then also $\sqrt{\log p}\|\hat{\Theta}_1 - \Theta_1^\sharp\|_1 = o_{\mathbf{P}}(1)$, which implies

$$\hat{\Theta}_1 X^T\epsilon/n = \Theta_1^\sharp X^T\epsilon/n + o_{\mathbf{P}}(1/\sqrt{n}).$$

Thus, then the estimator $\hat{b}_1$ is asymptotically linear, uniformly in $\beta^0 \in \mathcal{B}$. The uniform asymptotic linearity of $\hat{b}_1$ implies in turn that the Cramér Rao lower bound of Subsection 1.2 applies.

The results of the following two examples are summarized in Table 3.

Example 4.1. *Assume the sparse model (1.1) with* $s = o(\sqrt{n}/\log p)$ *(sparsity variant (ii)). As stated in Example 2.1, for the Lasso estimator* $\hat\beta$ *given in (1.4), with appropriate choice of the tuning parameter* $\lambda \asymp \sqrt{\log p/n}$, *one has uniformly in* $\beta^0 \in \mathcal{B}$

$$\|\hat\beta - \beta^0\|_1 = O_{\mathbf{P}_{\beta^0}}(s\sqrt{\log p/n}) = o_{\mathbf{P}_{\beta^0}}(1/\sqrt{\log p}).$$

Suppose now as in Theorem 4.1 that $\lambda^\sharp = \mathcal{O}(\sqrt{\log p/n})$ *and* $\sqrt{\log p/n}\|\gamma^\sharp\|_1 = o(1)$. *In fact, assume that* $\|\gamma^\sharp\|_0^0$ *is small enough so that* $\Theta_1^\sharp$ *is a model direction. Then one obtains by the same arguments if* $\lambda^{\rm Lasso} \asymp \sqrt{\log p/n}$ *is suitably chosen*

$$\|\hat\gamma - \gamma^\sharp\|_1 = O_{\mathbf{P}}(s\sqrt{\log p/n}) = o_{\mathbf{P}_{\beta^0}}(1/\sqrt{\log p}).$$

Thus then (4.5) is met so that we have asymptotic linearity. It means that the Cramér Rao lower bound applies and is achieved.

The model (1.2) is too large for us to be able to apply when Σ is unknown. We now turn to the model (1.3).

Example 4.2. *For the model (1.3) the rates for the Lasso* $\hat\beta$ *are given in Example 2.3. One sees that for* $0 \le r < 1$ *the requirement on* s *becomes*

$$s = o(n^{\frac{1-r}{2-r}}/\log p).$$

By the same arguments, if for a fixed $0 \le \mathrm{r} \le 1$

$$\|\gamma^\sharp\|_{\mathrm{r}}^{\mathrm{r}} = o(n^{\frac{1-\mathrm{r}}{2}}/\log^{\frac{2-\mathrm{r}}{2}} p)$$

one finds

$$\|\hat\gamma - \gamma^\sharp\|_1 = o_{\mathbf{P}}(1/\sqrt{\log p})$$

which yields asymptotic linearity so that the Cramér Rao lower bound applies. If such $\gamma^\sharp$ *is in addition a model direction after scaling, i.e. if* $\|\gamma^\sharp\|_r^r = \mathcal{O}(n^{\frac{r}{2}}s^{\frac{2-r}{2}})$, *then the Cramer Rao lower bound is achieved whenever* β^0 *stays away from the boundary.*

5. Conclusion

This paper illustrates that $\Theta_{1,1}$ can be larger than the asymptotic Cramér Rao lower bound, that for certain Σ the asymptotic variance of a debiased Lasso is smaller than $\Theta_{1,1}$ and that in special such cases the asymptotic Cramér Rao lower bound is achieved. In Examples 1.2 and 1.3 we showed that if β^0 stays away from the boundary, then the asymptotic Cramér Rao lower bound is

$$\left(\min_{\|c\|_r^r \le M n^{\frac{r}{2}} s^{\frac{2-r}{2}}} \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}c)^2\right)^{-1}$$

where M is any fixed value not depending on n. When Σ is known, Theorem 2.1 shows that up to log-terms this lower bound is achieved as soon as there exists an eligible pair $(\gamma^\sharp, \lambda^\sharp)$, with $\|\gamma^\sharp\|_r^r = \mathcal{O}(n^{\frac{r}{2}} s^{\frac{2-r}{2}})$. When Σ is unknown the situation is more involved and in particular for model (1.1) sparsity variant (i) is replaced by the stronger variant (ii). Model (1.2) is too large for the case Σ unknown and model (1.3) requires more sparsity then model (1.1): if r is larger the sparsity s is required to be smaller. Model (1.1) however appears for both known and unknown Σ too stringent as results depend on the exact value of s, not only on its order. Model (1.2) (with Σ known) or more generally model (1.3) (with $0 < r < 1$ if Σ is unknown) do not suffer from such a dependence as long as β^0 stays away from the boundary.

6. Proofs

6.1. Proof for Section 1

The proof of Proposition 1.1 relies on the results in [8], which allow the arguments to follow those of the low-dimensional case. These arguments are then rather standard.

Proof of Proposition 1.1. Let $h \in \mathcal{H}_{\beta^0}$, $|h_1| \geq \rho$ and $h^T \Sigma h \leq R^2$. The log-likelihood ratio $\mathcal{L}(h)$ for $\beta^0 + h/\sqrt{n}$ with respect to β^0 is

$$\begin{aligned}\mathcal{L}(h) &= h^T X^T \epsilon/\sqrt{n} + h^T \hat{\Sigma} h/2 \\ &= h^T X^T \epsilon/\sqrt{n} + h^T \Sigma h/2 + o_{\mathbf{P}}(1)\end{aligned}$$

since $h^T \Sigma h = \mathcal{O}(1)$. Let for $(x, y) \in \mathbb{R}^{p+1}$

$$l_{\beta_0}(x, y) := \begin{pmatrix} \mathbf{i}_{\beta^0}(x, y) \\ x^T (y - x\beta^0) \end{pmatrix}$$

and define

$$\Omega_{\beta^0} := \mathbb{E}_{\beta^0} l_{\beta^0}(\mathbf{x}, \mathbf{y}) l_{\beta^0}(\mathbf{x}, \mathbf{y})^T.$$

By the Lindeberg condition, we can apply Lindeberg's central limit theorem to conclude that for any sequence $a := (1, c^T)^T \in \mathbb{R}^{p+1}$ with $c^T \Sigma c = \mathcal{O}(1)$ it holds that

$$\frac{a^T \sum_{i=1}^n l_{\beta^0}(X_i, Y_i)}{\sqrt{n a^T \Omega_{\beta^0} a}}) \xrightarrow{\mathcal{D}_{\beta^0}} \mathcal{N}(0, 1).$$

Therefore, by Wold's device

$$\left(\begin{pmatrix} 1 & 0 \\ 0 & h^T \end{pmatrix} \Omega_{\beta^0} \begin{pmatrix} 1 & 0 \\ 0 & h \end{pmatrix} \right)^{-1/2} \begin{pmatrix} 1 & 0 \\ 0 & h^T \end{pmatrix} \sum_{i=1}^n l_{\beta^0}(X_i, Y_i)/\sqrt{n}$$
$$\xrightarrow{\mathcal{D}_{\beta^0}} \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

We now apply a slight modification of Lemmas 16 and 23 in [8], where we drop the assumption of bounded eigenvalues of Σ (which is possible because we have $h^T \Sigma h \le R^2 = \mathcal{O}(1)$). The asymptotic linearity of T and the 2-dimensional central limit theorem just obtained imply that at the alternative $\beta^0 + h/\sqrt{n}$ it holds that

$$\frac{T - (\beta_1^0 + h_1/\sqrt{n}) + h_1 - h^T \mathbb{E}_{\beta^0} \mathbf{i}_{\beta^0}(\mathbf{x}, \mathbf{y}) \mathbf{x}^T (\mathbf{y} - \mathbf{x}\beta^0)}{\sqrt{n V_{\beta^0}^2}} \xrightarrow{\mathcal{D}_{\beta^0 + h/\sqrt{n}}} \mathcal{N}(0,1).$$

As T is assumed to be regular at β^0 we conclude that

$$h^T \mathbb{E}_{\beta^0} \mathbf{i}_{\beta^0}(\mathbf{x}, \mathbf{y}) \mathbf{x}^T (\mathbf{y} - \mathbf{x}\beta^0) = h_1 + o(1).$$

But by the Cauchy-Schwarz inequality

$$\biggl(h^T \mathbb{E}_{\beta^0} \mathbf{i}_{\beta^0}(\mathbf{x}, \mathbf{y}) \mathbf{x}^T (\mathbf{y} - \mathbf{x}\beta^0) \biggr)^2 \le V_{\beta^0}^2 h^T \Sigma h.$$

Moreover,

$$\biggl(h_1 + o(1) \biggr)^2 = h_1^2 + o(1),$$

so that we obtain

$$V_{\beta^0}^2 \ge \frac{h_1^2 + o(1)}{h^T \Sigma h} = \frac{h_1^2}{h^T \Sigma h} + o(1) \tag{6.1}$$

where in the last step we used $h^T \Sigma h \ge \|h\|_2^2 / \Lambda_{\min}^2 \ge \rho^2$ so that $1/h^T \Sigma h = \mathcal{O}(1)$. Since the result is true for all $h \in \mathcal{H}_{\beta^0}$, $|h_1| \ge \rho$ and $h^T \Sigma h \le R^2$, we may maximize the right hand side of (6.1) over all such h. □

6.2. *Proofs for Section 2*

Proof of Lemma 2.1. Because $\mathbf{x}_{-1}\gamma^0$ is the projection of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$ we know that

$$\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^\sharp)^2 \ge \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^0)^2 = 1/\Theta_{1,1}.$$

Moreover

$$\Theta_{1,1} = \mathrm{e}_1^T \Theta \mathrm{e}_1 \le \Lambda_{\max}^2(\Theta) = 1/\Lambda_{\min}^2.$$

Thus

$$\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^\sharp)^2 \ge \Lambda_{\min}^2. \tag{6.2}$$

We now rewrite

$$\begin{aligned} \mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^\sharp)^2 &= 1 + \gamma^{\sharp T} \Sigma_{-1,-1} \gamma^\sharp - 2\mathbb{E}\mathbf{x}_1^T \mathbf{x}_{-1} \gamma^\sharp \\ &= 1 + \gamma^{\sharp T} \Sigma_{-1,-1} \gamma^\sharp - 2\gamma^{0T} \Sigma_{-1,-1} \gamma^\sharp \end{aligned}$$

where in the second equality we used that $\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^0$ is the anti-projection of $\mathbf{x}_1$ on $\mathbf{x}_{-1}$ and hence orthogonal to $\mathbf{x}_{-1}\gamma^\sharp$. For the cross-product we have by the two conditions on the pair $(\gamma^\sharp, \lambda^\sharp)$

$$\begin{aligned} \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp &= \gamma^{\sharp T}\Sigma_{-1,-1}\gamma^\sharp + \underbrace{(\gamma^0 - \gamma^\sharp)^T\Sigma_{-1,-1}\gamma^\sharp}_{|\cdot| \le \|\Sigma_{-1,-1}(\gamma^0-\gamma^\sharp)\|_\infty \|\gamma^\sharp\|_1 \le \lambda^\sharp \|\gamma^\sharp\|_1 = o(1)} \\ &= \gamma^{\sharp T}\Sigma_{-1,-1}\gamma^\sharp + o(1). \end{aligned}$$

Thus

$$\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^\sharp)^2 = 1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp + o(1).$$

Combining this with inequality (6.2) proves the first result of the lemma. The second result:

$$\|\Sigma(\Theta_1^\sharp - \Theta_1)\|_\infty \le \lambda_0^\sharp$$

follows trivially from this. For the third result, we compute and re-use the already obtained results:

$$\begin{aligned} \Theta_1^\sharp \Sigma \Theta_1^\sharp &= \frac{1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp - \gamma^{\sharp T}\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)}{(1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp)^2} \\ &= \underbrace{\frac{1}{1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp}}_{=\Theta_{1,1}^\sharp} + o(1) \\ &= 1/\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^\sharp)^2 + o(1) \\ &\le \underbrace{1/\mathbb{E}(\mathbf{x}_1 - \mathbf{x}_{-1}\gamma^0)^2}_{=\Theta_{1,1}} + o(1). \end{aligned}$$

To show the final statement of the lemma, assume on the contrary that γ^0 is sparse:

$$\lambda^\sharp \|\gamma^0\|_1 \to 0.$$

Then

$$\begin{aligned} \Theta_{1,1}^\sharp &= \frac{1}{1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^\sharp} \\ &= \frac{1}{1 - \gamma^{0T}\Sigma_{-1,-1}\gamma^0 + \underbrace{\gamma^{0T}\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)}_{|\cdot| \le \lambda^\sharp \|\gamma^0\|_1 = o(1)}} \\ &= \Theta_{1,1} + o(1). \end{aligned}$$

□

Proof of Theorem 2.1. We use the decomposition of the beginning of this section applied to $\hat{b}_{I,1}^\sharp$

$$\hat{b}_{I,1}^\sharp - \beta_1^0 = 2\Theta_1^{\sharp T} X_I^T \epsilon_I / n + \underbrace{(\mathrm{e}_1^T - \Theta_1^{\sharp T}\hat{\Sigma}_I)(\hat{\beta}_{II} - \beta_1^0)}_{\text{remainder}_I}$$

where remainder$_I$ is

$$(\mathrm{e}_1^T - \Theta_1^{\sharp T}\hat{\Sigma}_I)(\hat{\beta}_{II} - \beta^0) = \underbrace{\Theta_1^{\sharp T}(\Sigma - \hat{\Sigma}_I)(\hat{\beta}_{II} - \beta^0)}_{:=(i)} + \underbrace{(\mathrm{e}_1^T - \Theta_1^{\sharp T}\Sigma)(\hat{\beta}_{II} - \beta^0)}_{:=(ii)}.$$

Here, $\epsilon_I := Y_I - X_I\beta^0$ and $\hat{\Sigma}_I := 2X_I^TX_I/n$. But, given (X_{II}, Y_{II}),

$$\hat{\Theta}_1^{\sharp T}\hat{\Sigma}_I(\hat{\beta}_{II} - \beta^0) = 2\Theta_1^{\sharp T}X_I^TX_I(\hat{\beta}_{II} - \beta^0)/n$$

is the average of $n/2$ i.i.d. random variables which are the product of a random variable with the $\mathcal{N}(0, \Theta_1^{\sharp T}\Sigma\Theta_1^\sharp)$-distribution and a $\mathcal{N}(0, \|\Sigma_{1/2}(\hat{\beta}_{II} - \beta^0)\|_2^2)$-distributed random variable. Since the variances satisfy $\Theta_1^{\sharp T}\Sigma\Theta_1^\sharp = \Theta_{1,1}^\sharp + o(1) = \mathcal{O}(1)$ and $\|\Sigma_{1/2}(\hat{\beta}_{II} - \beta^0)\|_2^2 = o_{\mathbf{P}_{\beta^0}}(1)$ uniformly in $\beta^0 \in \mathcal{B}$ we see that

$$(i) = o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n})$$

uniformly in $\beta^0 \in \mathcal{B}$. For the term (ii) we use that

$$\|\Sigma\Theta_1^\sharp - \mathrm{e}_1\|_\infty = \mathcal{O}(\lambda^\sharp)$$

and the assumption $\lambda^\sharp\|\hat{\beta}_{II} - \beta^0\|_1 = o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n})$ uniformly in $\beta^0 \in \mathcal{B}$. This gives that uniformly in $\beta^0 \in \mathcal{B}$

$$\hat{b}_{I,1}^\sharp - \beta_1^0 = 2\Theta_1^{\sharp T}X_I^T\epsilon_I/n + o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n}).$$

In the same way one derives that uniformly in $\beta^0 \in \mathcal{B}$

$$\hat{b}_{II,1}^\sharp - \beta_1^0 = 2\Theta_1^{\sharp T}X_{II}^T\epsilon_{II}/n + o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n})$$

with $\epsilon_{II} := Y_{II} - X_{II}\beta^0$. Since $\hat{b}_1^\sharp = (b_{I,1}^\sharp + b_{II,1}^\sharp)/2$ is the average of the two, this proves the asymptotic linearity. Further $\mathrm{var}(\Theta_1^\sharp X^T\epsilon/\sqrt{n}) = \Theta_1^{\sharp T}\Sigma\Theta_1^\sharp = \Theta_{1,1}^\sharp + o(1)$ by Lemma 2.1. The central limit theorem completes the proof. □

6.3. *Proofs for Section 3*

Proof of Lemma 3.1. We have

$$\begin{aligned}(\gamma^\sharp - \gamma^\flat)^T\Sigma_{-1,-1}(\gamma^\sharp - \gamma^\flat) &\le \|\gamma^\sharp - \gamma^\flat\|_1\|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^\flat)\|_\infty \\ &\le 2\lambda^\sharp\|\gamma^\sharp - \gamma^\flat\|_1 \\ &\le 2\lambda^\sharp\|\gamma^\sharp\|_1 + 2\lambda^\sharp\|\gamma^\flat\|_1 \to 0.\end{aligned}$$

□

Proof of Lemma 3.2. By the KKT conditions

$$(\gamma_{\text{Lasso}} - \gamma^\sharp)^T\Sigma_{-1,-1}(\gamma_{\text{Lasso}} - \gamma^0) \le \lambda_{\text{Lasso}}\|\gamma^\sharp\|_1 - \lambda_{\text{Lasso}}\|\gamma_{\text{Lasso}}\|_1.$$

Therefore,

$$\begin{array}{cl} & (\gamma_{\text{Lasso}} - \gamma^\sharp)^T \Sigma_{-1,-1} (\gamma_{\text{Lasso}} - \gamma^\sharp) \\ = & (\gamma_{\text{Lasso}} - \gamma^\sharp)^T \Sigma_{-1,-1} (\ \gamma_{\text{Lasso}} - \gamma^0) + (\gamma_{\text{Lasso}} - \gamma^\sharp)^T \Sigma_{-1,-1} (\gamma^0 - \gamma^\sharp) \\ \le & \lambda_{\text{Lasso}} \|\gamma^\sharp\|_1 - \lambda_{\text{Lasso}} \|\gamma_{\text{Lasso}}\|_1 + \|\gamma_{\text{Lasso}} - \gamma^\sharp\|_1 \|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty \\ \le & \lambda_{\text{Lasso}} \|\gamma^\sharp\|_1 - \lambda_{\text{Lasso}} \|\gamma_{\text{Lasso}}\|_1 + \lambda^\sharp \|\gamma_{\text{Lasso}} - \gamma^\sharp\|_1 \\ \le & (\lambda_{\text{Lasso}} + \lambda^\sharp) \|\gamma^\sharp\|_1 - (\lambda_{\text{Lasso}} - \lambda^\sharp) \|\gamma_{\text{Lasso}}\|_1 . \end{array}$$

Thus

$$(\gamma_{\text{Lasso}} - \gamma^\sharp)^T \Sigma_{-1,-1} (\gamma_{\text{Lasso}} - \gamma^\sharp) + (\lambda_{\text{Lasso}} - \lambda^\sharp) \|\gamma_{\text{Lasso}}\|_1 \le (\lambda_{\text{Lasso}} + \lambda^\sharp) \|\gamma^\sharp\|_1 \to 0$$

where we used that $\lambda_{\text{Lasso}} > \lambda^\sharp$ and $\lambda_{\text{Lasso}} \|\gamma^\sharp\|_1 \to 0$. We also know that by the KKT conditions

$$\|\Sigma_{-1,-1}(\gamma_{\text{Lasso}} - \gamma^0)\|_\infty \le \lambda_{\text{Lasso}}.$$

If $\lambda_{\text{Lasso}} \ge 2\lambda^\sharp$, we obtain from the above

$$\lambda_{\text{Lasso}} \|\gamma_{\text{Lasso}}\|_1 \le 3\lambda_{\text{Lasso}} \|\gamma^\sharp\|_1 / 2 \to 0.$$

So $(\gamma_{\text{Lasso}}, \lambda_{\text{Lasso}})$ is an eligible pair. □

Proof of Lemma 3.3. Note that

$$\begin{array}{rcl} \mathbb{E}\mathbf{x}_{-S}^T \mathbf{x}_{-1} \gamma^{-S} & = & \mathbb{E}\mathbf{x}_{-S}^T \Big[(\mathbf{x}_{-S}\gamma_{-S}^0) \mathrm{A} \mathbf{x}_S \Big] \\ & = & \mathbb{E}\mathbf{x}_{-S}^T \Big[\mathbf{x}_{-S} - \mathbf{x}_S \Sigma_{S,S}^{-1} \Sigma_{S,-S} \Big] \gamma_{-S}^0 \\ & = & \Big[\Sigma_{-S,-S} - \Sigma_{-S,S} \Sigma_{S,S}^{-1} \Sigma_{S,-S} \Big] \gamma_{-S}^0 . \end{array}$$

We therefore have

$$\begin{array}{rcl} \|v_{-S}^S\|_\infty & = & \left\| \Big[\Sigma_{-S,-S} - \Sigma_{-S,S} \Sigma_{S,S}^{-1} \Sigma_{S,-S} \Big] \gamma_{-S}^0 \right\|_\infty \\ & \le & \|\Sigma_{-S,-S} - \Sigma_{-S,S} \Sigma_{S,S}^{-1} \Sigma_{S,-S}\|_1 \|\gamma_{-S}^0\|_\infty . \end{array}$$

□

Proof of Lemma 3.4. We have

$$\|\gamma_{-S}^0\|_\infty = \lambda^S / C,$$

so that by Lemma 3.3

$$\|\Sigma_{-1,-1}\gamma^S\|_\infty \le \lambda^S.$$

Moreover

$$\lambda^S \|\gamma^S\|_1 \le \lambda^S \sqrt{\mathrm{s}} \to 0.$$

Thus (γ^S, λ^S) is an eligible pair. Finally

$$\lambda^S \|\gamma^0\|_1 \ge \lambda^S \|\gamma_{-S}^0\|_1 = \frac{\kappa(\gamma_{-S}^0)}{C} \not\to 0.$$

□

Proof of Lemma 3.5. Write

$$\mathbf{x}_{-1}\gamma^0 = \mathbf{x}_{-1}\gamma^\sharp + \xi^0, \ \ \|\mathbb{E}\mathbf{x}_{-1}^T\xi^0\|_\infty \le \lambda^\sharp .$$

It holds that

$$\gamma_S^S = \Sigma_{S,S}^{-1}\mathbb{E}\mathbf{x}_S^T\mathbf{x}_{-1}\gamma^0 .$$

So

$$\begin{aligned} \|\Sigma_{-1,-1}^{1/2}(\gamma_S^S - \gamma^\sharp)\|_2^2 &= (\mathbb{E}\mathbf{x}_S^T\xi^0)^T\Sigma_{S,S}^{-1}(\mathbb{E}\mathbf{x}_S^T\xi^0) \\ &\le \|\mathbb{E}\mathbf{x}_S^T\xi^0\|_2/\Lambda_{\min}^2 \\ &\le \mathrm{s}\|\mathbb{E}\mathbf{x}_S^T\xi^0\|_\infty^2/\Lambda_{\min}^2 \to 0. \end{aligned}$$

□

Proof of Lemma 3.6. For $a \in \mathbb{R}$ and $c \in \mathbb{R}^{p-1}$ satisfying $a^2 + \|\Sigma_{-1,-1}^{1/2}c\|_2^2 = 1$,

$$\begin{aligned} \begin{pmatrix} a \\ c \end{pmatrix}^T \Sigma(\gamma^0) \begin{pmatrix} a \\ c \end{pmatrix} &= a^2 + 2a\gamma^{0T}\Sigma_{-1,-1}c + c^T\Sigma_{-1,-1}c \\ &= 1 + 2a\gamma^{0T}\Sigma_{-1,-1}c \\ &\ge 1 - 2|a|\|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2\|\Sigma_{-1,-1}^{1/2}c\|_2 \\ &= 1 - 2|a|\sqrt{1-a^2}\|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2 \\ &\ge 1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2 . \end{aligned}$$

But then

$$\begin{aligned} & \frac{1}{a^2 + \|c\|_2^2}\begin{pmatrix} a \\ c \end{pmatrix}^T \Sigma(\gamma^0) \begin{pmatrix} a \\ c \end{pmatrix} \\ = \ & \frac{a^2 + \|\Sigma_{-1,-1}^{1/2}c\|_2^2}{a^2 + \|c\|_2^2}\begin{pmatrix} a \\ c \end{pmatrix}^T \Sigma(\gamma^0) \begin{pmatrix} a \\ c \end{pmatrix} \\ \ge \ & \frac{a^2 + \Lambda_{\min}^2(\Sigma_{-1,-1})\|c\|_2}{a^2 + \|c\|_2}(1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2) \\ \ge \ & (1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2)\Lambda_{\min}^2(\Sigma_{-1,-1}). \end{aligned}$$

Hence $\Lambda_{\min}^2(\Sigma(\gamma^0))$ is positive definite and

$$\Lambda_{\min}^2(\Sigma(\gamma^0)) \ge (1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2)\Lambda_{\min}^2(\Sigma_{-1,-1}).$$

It further holds for all $j \in \{2, \ldots, p\}$ that

$$\mathbb{E}|\mathbf{x}_j^T\mathbf{x}_{-1}\gamma^0| \le \sqrt{\mathbb{E}(\mathbf{x}_{-1}\gamma^0)^2} < 1.$$

□

Proof of Lemma 3.7. By definition

$$\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp) = \lambda^\sharp z,$$

so that

$$\|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty \le \lambda^\sharp.$$

Moreover, $\lambda^\sharp \|\gamma^\sharp\|_1 \le \lambda \sqrt{\mathrm{s}} \|\gamma^\sharp\|_2 \to 0$, since

$$\|\gamma^\sharp\|_2 \le \|\Sigma_{-1,-1}^{1/2}\gamma^\sharp\|_2 / \Lambda_{\min}(\Sigma_{-1,-1}) = \mathcal{O}(1).$$

So $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. We further have

$$(\gamma^0 - \gamma^\sharp)^T \Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp) = \lambda^{\sharp 2} z^T \Sigma_{-1,-1}^{-1} z.$$

Thus

$$\begin{array}{rcl} \gamma^{0T}\Sigma_{-1,-1}\gamma^0 & = & \gamma^{\sharp T}\Sigma_{-1-1}\gamma^\sharp + 2(\gamma^0 - \gamma^\sharp)^T \Sigma_{-1,-1}\gamma^\sharp + \lambda^{\sharp 2} z^T \Sigma_{-1,-1}^{-1} z \\ & < & 1 + o(1) \end{array}$$

and

$$1 - \|\Sigma_{-1,-1}^{1/2}\gamma^0\|_2^2 \gg 0$$

where the positivity is true for large enough n. Therefore, by Lemma 3.6, γ^0 is eventually allowed. Finally,

$$\lambda^\sharp \|\gamma^0 - \gamma^\sharp\|_1 \ge \lambda^\sharp z^T(\gamma^0 - \gamma^\sharp) = \lambda^{\sharp 2} z^T \Sigma_{-1,-1}^{-1} z \gg 0.$$

So, since $\lambda^\sharp \|\gamma^\sharp\|_1 \to 0$, it must be true that $\lambda^\sharp \|\gamma^0\|_1 \gg 0$. We in fact have

$$\gamma^{0T}\Sigma_{-1,-1}\gamma^0 - \gamma^{\sharp T}\Sigma_{-1,-1}\gamma^\sharp = \lambda^{\sharp 2} z^T \Sigma_{-1,-1}^{-1} z + o(1)$$

so that

$$\gamma^{0T}\Sigma_{-1,-1}\gamma^0 - \gamma^{\sharp T}\Sigma_{-1,-1}\gamma^\sharp \gg 0.$$

□

Proof of Lemma 3.8. ($\overset{\text{a)}}{\Rightarrow}$) For $z_S := \Sigma_{S,-S}\Sigma_{-S,-S}^{-1} z_{-S}$ the equality

$$\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp) = \lambda^\sharp z$$

holds. By assumption $\|z_{-S}\|_\infty \le 1$ and by the reversed irrepresentable condition also $\|z_S\|_\infty \le 1$. Thus

$$\|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty \le \lambda^\sharp.$$

Moreover,

$$\lambda^\sharp \|\gamma^\sharp\|_1 = \lambda^\sharp \|\gamma_S^0\|_1 \le \lambda^\sharp \sqrt{\mathrm{s}} \|\gamma_S^0\|_2 \to 0.$$

So $(\gamma^\sharp, \lambda^\sharp)$ is eligible.

To see make sure that γ^0 is allowed we bound $\gamma^{0T}\Sigma_{-1,-1}\gamma^0$:

$$\gamma^{0T}\Sigma_{-1,-1}\gamma^0 \le \gamma_S^{0T}\Sigma_{S,S}\gamma_S^0 + 2\lambda^\sharp \|\gamma_S^0\|_1 + \lambda^{\sharp 2} z_{-S}^T \Sigma_{-1,-1}^{-1} z_{-S}.$$

Therefore, since $\lambda^\sharp \|\gamma_S^0\|_1 \to 0$, and in view of Lemma 3.6, the vector γ^0 is for large enough n allowed. Finally, we have

$$\lambda^\sharp \|\gamma_0\|_1 \quad \ge \quad \lambda^\sharp \|\gamma_{-S}^0\|_1$$

$$\begin{aligned} &\geq \lambda^\sharp z_{-S}^T \gamma_{-S}^0 \\ &= \lambda^{\sharp 2} z_{-S}^T \Sigma_{-S,-S}^{-1} z_{-S} \gg 0. \end{aligned}$$

In fact

$$\gamma^{0T}\Sigma_{-1,-1}\gamma^0 - \gamma^{\sharp T}\Sigma_{-1,-1}\gamma^\sharp = \lambda^{\sharp 2} z_{-S}^T \Sigma_{-S,-S}^{-1} z_{-S} + o(1) \gg 0.$$

($\overset{\text{b)}}{\Leftarrow}$) If $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair we have

$$\|\Sigma_{-1,-1}(\gamma^\sharp - \gamma^0)\|_\infty \leq \lambda^\sharp.$$

Define now $c = \gamma^0 - \gamma^\sharp$ and $z = \Sigma_{-1,-1}^{-1} c / \lambda^\sharp$. Then

$$\Sigma_{-1,-1} c = \lambda^\sharp z$$

and $\|z\|_\infty \leq 1$, $c_S = 0$. It follows that

$$\Sigma_{S,-S}\Sigma_{-S,-S}^{-1} z_{-S} = z_S$$

so that

$$\|\Sigma_{S,-S}\Sigma_{-S,-S}^{-1} z_{-S}\|_\infty \leq 1. \qquad \square$$

Proof of Lemma 3.9. One readily verifies that all c_j^0 with $j \notin S$ are non-zero. One thus has the Lagrangrian

$$\Sigma_{-1,-1} c^0 = \tilde{\lambda} W \zeta$$

where $\tilde{\lambda}$ is the Lagrangian parameter. Since $\zeta_S = 0$ this says that

$$(\Sigma_{-1,-1} c^0)_S = 0$$

so we know that $\mathbf{x}_{-1} c^0$ is orthogonal to $\mathbf{x}_S$.

The restriction gives

$$\|\Sigma_{-1,-1}^{1/2} c^0\|_2^2 = \tilde{\lambda} \|(W c^0)_{-S}\|_\infty = \tilde{\lambda}/\lambda^\sharp,$$

so

$$\tilde{\lambda} = \lambda^\sharp \|\Sigma_{-1,-1}^{1/2} c^0\|_2^2,$$

and inserting this back yields

$$\Sigma_{-1,-1} c^0 = \lambda^\sharp \|\Sigma_{-1,-1}^{1/2} c^0\|_2^2 W \zeta.$$

It follows that

$$c^0 = \lambda^\sharp \|\Sigma_{-1,-1}^{1/2} c^0\|_2^2 \Sigma_{-1,-1}^{-1} W \zeta.$$

But then

$$\|\Sigma_{-1,-1}^{1/2} c^0\|_2 = \lambda^\sharp \|\Sigma_{-1,-1}^{1/2} c^0\|_2^2 \|\Sigma_{-1,-1}^{-1/2} W \zeta\|_2$$

or

$$\|\Sigma_{-1,-1}^{1/2} c^0\|_2 = \frac{1}{\lambda^\sharp \|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2}.$$

So now we have

$$\Sigma_{-1,-1} c^0 = \frac{1}{\lambda^\sharp \|\Sigma_{-1,-1}^{-1/2} \zeta\|_2^2} W\zeta$$

and hence

$$\|\Sigma_{-1,-1} c^0\|_\infty = \frac{\|w_{-S}\|_\infty}{\lambda^\sharp \|\Sigma_{-1,-1}^{-1/2} \zeta\|_2^2}. \qquad \square$$

Proof of Lemma 3.10. It holds that

$$\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp) = \mathbb{E}\mathbf{x}_{-1}^T \mathbf{x}_{-1}(\gamma^0 - \gamma^\sharp) = \mathbb{E}\mathbf{x}_{-1}^T (\mathbf{x}_{-S} \mathrm{A} \mathbf{x}_S)\gamma_{-S}^0 = \Sigma_{-1,-1} c^0.$$

So

$$\|\Sigma_{-1,-1}(\gamma^0 - \gamma^\sharp)\|_\infty = \frac{\|w_{-S}\|_\infty}{\lambda^\sharp \|\Sigma_{-1,-1}^{1/2} W\zeta\|_2^2} = \frac{\lambda^\sharp \|w_{-S}\|_\infty}{\lambda^{\sharp 2} \|\Sigma_{-1,-1}^{1/2} W\zeta\|_2^2} \le \lambda^\sharp.$$

Moreover

$$\lambda^\sharp \|\gamma^\sharp\|_1 \le \sqrt{\mathrm{s}} \|\gamma^\sharp\|_2 \to 0.$$

Thus $(\gamma^\sharp, \lambda^\sharp)$ is an eligible pair. Furthermore

$$\begin{aligned} \gamma^{0T} \Sigma_{-1,-1} \gamma^0 &= \gamma^{\sharp T} \Sigma_{-1,-1} \gamma^\sharp + c^{0T} \Sigma_{-1,-1} c^0 \\ &= \gamma^{\sharp T} \Sigma_{-1,-1} \gamma^\sharp + \frac{1}{\lambda^{\sharp 2} \|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2}. \end{aligned}$$

So by Lemma 3.6 γ^0 is allowed. Finally, $\lambda^\sharp \|\gamma^0\|_1 \ge \lambda^\sharp \|(W\gamma^0)_{-S}\|_1 = 1$ and in fact

$$\gamma^{0T} \Sigma_{-1,-1} \gamma^0 - \gamma^{\sharp T} \Sigma_{-1,-1} \gamma^\sharp = \frac{1}{\lambda^{\sharp 2} \|\Sigma_{-1,-1}^{-1/2} W\zeta\|_2^2} \gg 0. \qquad \square$$

6.4. Proofs for Section 4

Proof of Lemma 4.1. We recall the notation $\hat{\Sigma}_{-1,-1} := X_{-1}^T X_{-1}/n$. The event

$$\left\{ \|X_{-1}^T \epsilon^\sharp\|_\infty / n \le \lambda_\varepsilon^\sharp \right\} \cap \left\{ \inf_{c:\ \lambda^{\text{Lasso}} \|c\|_1 \le 4\eta_n^2,\ c^T \Sigma_{-1,-1} c = 1} c^T \hat{\Sigma}_{-1,-1} c \ge \frac{1}{2} \right\}$$

has probability converging to one so in the rest of the proof we may assume that we are on this event. By the KKT conditions

$$\hat{\Sigma}_{-1,-1}(\hat{\gamma} - \gamma^0) = X_{-1}^T \varepsilon / n - \lambda^{\text{Lasso}} \hat{\zeta}$$

where $\hat{\zeta} \in \partial \|\hat{\gamma}\|_1$, with $\partial \|c\|_1$ the sub-differential of the map $c \mapsto \|c\|_1$. Thus

$$\hat{\Sigma}_{-1,-1}(\hat{\gamma} - \gamma^\sharp) = X_{-1}^T \varepsilon^\sharp / n - \lambda^{\text{Lasso}} \hat{\zeta}.$$

Therefore

$$\begin{array}{rcl}(\hat\gamma-\gamma^\sharp)^T\hat\Sigma_{-1,-1}(\hat\gamma-\gamma^\sharp) & = & (\hat\gamma-\gamma^\sharp)^T X_{-1}^T\varepsilon^\sharp/n-\lambda^{\rm Lasso}(\hat\gamma-\gamma^\sharp)^T\hat\zeta \\ & \le & \lambda_\varepsilon^\sharp\|\hat\gamma-\gamma^\sharp\|_1+\lambda^{\rm Lasso}\|\gamma^\sharp\|_1-\lambda^{\rm Lasso}\|\hat\gamma\|_1 \\ & \le & (\lambda^{\rm Lasso}+\lambda_\varepsilon^\sharp)\|\gamma^\sharp\|_1-(\lambda^{\rm Lasso}-\lambda_\varepsilon^\sharp)\|\hat\gamma\|_1 \\ & \le & (\lambda^{\rm Lasso}+\lambda_\varepsilon^\sharp)\|\gamma^\sharp\|_1 \\ & \le & 3\lambda^{\rm Lasso}\|\gamma^\sharp\|_1/2\le 3\eta_n^2/2.\end{array}$$

It moreover follows from the above that

$$\|\hat\gamma\|_1\le\left(\frac{\lambda^{\rm Lasso}+\lambda_\varepsilon^\sharp}{\lambda^{\rm Lasso}-\lambda_\varepsilon^\sharp}\right)\|\gamma^\sharp\|_1$$

and so

$$\lambda^{\rm Lasso}\|\hat\gamma\|_1\le\lambda^{\rm Lasso}\left(\frac{\lambda^{\rm Lasso}+\lambda_\varepsilon}{\lambda^{\rm Lasso}-\lambda_\varepsilon^\sharp}\right)\le 3\eta_n^2,$$

and also

$$\lambda^{\rm Lasso}\|\hat\gamma-\gamma^\sharp\|_1\le\lambda^{\rm Lasso}\left(\frac{2\lambda^{\rm Lasso}}{\lambda^{\rm Lasso}-\lambda_\varepsilon^\sharp}\right)\|\gamma^\sharp\|_1\le 4\eta_n^2.$$

If $\|\Sigma^{1/2}(\hat\gamma-\gamma^\sharp)\|_2\le 2\eta_n$ we are done. Otherwise, if $\|\Sigma^{1/2}(\hat\gamma-\gamma^\sharp)\|_2\ge 2\eta_n$ it holds that $4\eta_n^2\le 2\eta_n\|\Sigma^{1/2}(\hat\gamma-\gamma^\sharp)\|_2$. But then

$$\begin{array}{rcl}\frac{1}{2}(\hat\gamma-\gamma^\sharp)^T\Sigma_{-1,-1}(\hat\gamma-\gamma^\sharp) & \le & (\hat\gamma-\gamma^\sharp)^T\hat\Sigma_{-1,-1}(\hat\gamma-\gamma^\sharp) \\ & \le & (\lambda^{\rm Lasso}+\lambda_\varepsilon^\sharp)\|\gamma^\sharp\|_1 \\ & \le & 2\lambda^{\rm Lasso}\|\gamma^\sharp\|_1\le 2\eta_n^2.\end{array}$$

□

Proof of Theorem 4.1. We rewrite

$$\hat\beta_1-\beta_1^0=\hat\Theta_1X^T\epsilon/n+\underbrace{({\rm e}_1^T-\hat\Theta_1^T\hat\Sigma)(\hat\beta-\beta^0)}_{\rm remainder}.$$

By the KKT conditions

$$\|{\rm e}_1^T-\hat\Theta_1^T\hat\Sigma\|_\infty\le\lambda^{\rm Lasso}/(\|X_1-X_{-1}\hat\gamma\|_2^2/n+\lambda^{\rm Lasso}\|\hat\gamma\|_1).$$

But by Lemma 4.1

$$\|X_1-X_{-1}\hat\gamma\|_2^2/n+\lambda^{\rm Lasso}\|\hat\gamma\|_1=\mathbb{E}(\mathbf{x}_1-\mathbf{x}_{-1}\gamma^\sharp)^2+o(1)$$

which stays away from zero. Moreover, by assumption, $\sqrt{n}\lambda^{\rm Lasso}\|\hat\beta-\beta_0\|_1=o_{\mathbf{P}_{\beta^0}}(1)$ uniformly in $\beta^0\in\mathcal{B}$. Thus, for the remainder we find

$$|({\rm e}_1^T-\hat\Theta_1^T\hat\Sigma)(\hat\beta-\beta^0)|\le\|{\rm e}_1^T-\hat\Theta_1^T\hat\Sigma\|_\infty\|\hat\beta-\beta^0\|_1=o_{\mathbf{P}_{\beta^0}}(1/\sqrt{n})$$

uniformly in $\beta^0 \in \mathcal{B}$. For the main term, we have after standardization

$$\frac{\hat{\Theta}_1^T X^T \epsilon/\sqrt{n}}{\sqrt{\hat{\Theta}_1^T \hat{\Sigma} \hat{\Theta}_1}} \sim \mathcal{N}(0,1).$$

It further holds that $\hat{\Theta}_1^T \hat{\Sigma} \hat{\Theta}_1 = \Theta^\sharp_{1,1} + o_{\mathbf{P}}(1)$ by Lemma 4.1, which stays away from zero. Therefore, for the standardized remainder term

$$\frac{\sqrt{n}|(\mathrm{e}_1^T - \hat{\Theta}_1^T \hat{\Sigma})(\hat{\beta} - \beta^0)|}{\sqrt{\hat{\Theta}_1^T \hat{\Sigma} \hat{\Theta}_1}} = o_{\mathbf{P}_{\beta^0}}(1)$$

uniformly in $\beta^0 \in \mathcal{B}$. The final result thus follows from Slutsky's Theorem. □

7. Probability inequalities

In this section we present some probability inequalities for products of Gaussians. Such results are known (for example as Hanson-Wright inequalities for sub-Gaussians, see [21]) and only presented here for completeness.

Lemma 7.1. *Let U and W be two independent $\mathcal{N}(0,1)$-distributed random variables. Then for all $L > 1$*

$$\mathbb{E}\exp\left[\frac{UW}{L}\right] \le \exp\left[\frac{1}{2L^2 - 2L}\right].$$

Proof. We have for $L > 1$

$$\begin{aligned}
\mathbb{E}\exp\left[\frac{UW}{L}\right] &\le \mathbb{E}\exp\left[\frac{(U+W)^2 - (U-W)^2}{4L}\right] \\
&= \mathbb{E}\exp\left[\frac{(U+W)^2 - 2}{4L}\right]\exp\left[\frac{2-(U-W)^2}{4L}\right] \\
&= \mathbb{E}\exp\left[\frac{(U+W)^2 - 2}{4L}\right]\mathbb{E}\exp\left[\frac{2-(U-W)^2}{4L}\right] \\
&\le \exp\left[\frac{1}{4L^2 - 4L}\right]\mathbb{E}\exp\left[\frac{2-(U-W)^2}{4L}\right],
\end{aligned}$$

(see Lemma 1 and its proof in [13]) or Section 8.4 in [24]). But

$$\begin{aligned}
\mathbb{E}\exp\left[\frac{2-(U-W)^2}{4L}\right] &= \frac{1}{\sqrt{2\pi}}\int \exp\left[\frac{1-v^2}{2L}\right]\exp\left[-\frac{v^2}{2}\right]dv \\
&= \exp\left[\frac{1}{2L}\right]\frac{1}{\sqrt{2\pi}}\int \exp\left[-\frac{1}{2}\left(1+\frac{1}{L}\right)v^2\right]dv \\
&= \exp\left[\frac{1}{2L}\right]\left(1+\frac{1}{L}\right)^{-1/2}
\end{aligned}$$

$$= \exp\left[\frac{1}{2L} - \frac{1}{2}\log(1+1/L)\right].$$

Since

$$\log(1+1/L) \geq 1/L - \tfrac{1}{2}(1/L^2)$$

we obtain

$$\begin{aligned} \mathbb{E}\exp\left[\frac{2-(U+W)^2}{4L}\right] &\leq \exp\left[\frac{1}{4L^2}\right] \\ &\leq \exp\left[\frac{1}{4L^2-4L}\right]. \end{aligned}$$

It follows that

$$\mathbb{E}\exp\left[\frac{UW}{L}\right] \leq \exp\left[\frac{2}{4L^2-4L}\right] = \exp\left[\frac{1}{2L^2-2L}\right]. \qquad \square$$

Lemma 7.2. *Let $U = (U_1, \ldots, U_n)^T$ and $W = (W_1, \ldots, W_n)^T$ be two independent standard Gaussian n-dimensional random vectors. Then for all $t > 0$*

$$\mathbb{P}\left(U^T W/n \geq \sqrt{2t/n} + t/n\right) \leq \exp[-t].$$

Proof. By Lemma 7.1 and using the independence

$$\mathbb{E}\exp\left[\frac{U^T W}{L}\right] \leq \exp\left[\frac{n}{2L^2-2L}\right].$$

This gives for all $t > 0$

$$\mathbb{P}\left(U^T W \geq \sqrt{2nt} + t\right) \leq \exp[-t]$$

(see e.g. Lemma 8.3 in [24]). □

Lemma 7.3. *Let $\{(U_i, V_i)\}_{i=1}^n$ be i.i.d. two-dimensional Gaussians with mean zero. Suppose $\mathrm{var}(U_1) = 1$. Define $\lambda^\sharp := \mathbb{E}U_1 V_1$ and $\sigma^{\sharp 2} = \mathbb{E}V_1^2$. Then for all $t > 0$*

$$\mathbb{P}\left(U^T V/n \geq \lambda^\sharp + (\sqrt{2}\sigma^\sharp + 2\lambda^\sharp)\sqrt{t/n} + (\sigma^\sharp + 2\lambda^\sharp)t/n\right) \leq 2\exp[-t].$$

Proof. For all i the projection of V_i on U_i is $[\mathbb{E}U_i V_i/\mathrm{var}(U_i)]U_i = \lambda^\sharp U_i$. Hence we may write for all i

$$V_i = \lambda^\sharp U_i + W_i,$$

where W_i is a zero-mean Gaussian random variable independent of U_i. It follows that

$$U^T V/n = \lambda^\sharp \|U\|_2^2/n + U^T W/n.$$

Since $\text{var}(W_i) \le \sigma^{\sharp 2}$ for all i we see from Lemma 7.2 that

$$\mathbb{P}\left(U^T W/n \ge \sqrt{2}\sigma^\sharp \sqrt{t/n} + \sigma^\sharp t/n \right) \le \exp[-t].$$

Moreover (see Lemma in [13], also given in [24] as Lemma 8.6)

$$\mathbb{P}\left(\|U\|_2^2/n - 1 \ge 2\sqrt{t/n} + 2t/n \right) \le \exp[-t].$$

Thus

$$\mathbb{P}\left(U^T V/n \ge \lambda^\sharp + (\sqrt{2}\sigma^\sharp + 2\lambda^\sharp)\sqrt{t/n} + (\sigma^\sharp + 2\lambda^\sharp)t/n \right) \le 2\exp[-t]. \qquad \square$$

References

[1] A. Belloni, V. Chernozhukov, and K. Kato. Uniform postselection inference for LAD regression models. *Biometrika*, 102:77–94, 2015. MR3335097

[2] A. Belloni, V. Chernozhukov, and Y. Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619, 2016. MR3547999

[3] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore, 1993. MR1245941

[4] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, pages 1705–1732, 2009. MR2533469

[5] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011. MR2807761

[6] T. Cai and Z. Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, 45(2):615–646, 2017. MR3650395

[7] C. Giraud. *Introduction to High-Dimensional Statistics*, volume 138. CRC Press, 2014. MR3307991

[8] J. Janková and S. van de Geer. Semi-parametric efficiency bounds for high-dimensional models. *Annals of Statistics*, pages 2356–2359, 2018. MR3845020

[9] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014. MR3277152

[10] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, 2014. MR3265038

[11] A. Javanmard and A. Montanari. Debiasing the Lasso: optimal sample size for Gaussian designs. *Annals of Statistics*, 46:2593–2622, 2018. MR3851749

[12] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 38. Springer Science & Business Media, 2011. MR2829871

[13] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000. MR1805785

[14] H. Leeb and B.M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005. MR2153856

[15] H. Leeb and B.M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142(1):201–211, 2008. MR2394290

[16] Y. Plan and R. Vershynin. One-bit compressed censing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013. MR3069959

[17] B.M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71-A:1–18, 2009. MR2579644

[18] B.M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082, 2009. MR2543087

[19] B.M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electronic Journal of Statistics*, 4:334–360, 2010. MR2645488

[20] Z. Ren, T. Sun, C.-H. Zhang, and H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Annals of Statistics*, 43:991–1026, 2015. MR3346695

[21] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013. MR3125258

[22] A. Schick. On asymptotically efficient estimation in semiparametric models. *Annals of Statistics*, 14(3):1139–1151, 1986. MR0856811

[23] R. Tibshirani. Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996. MR1379242

[24] S. van de Geer. *Estimation and Testing Under Sparsity: Ecole d'Eté de Probabilités de Saint-Flour XLV-2016*. Springer Science & Business Media, 2016. MR3526202

[25] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42:1166–1202, 2014. MR3224285

[26] C.-H. Zhang and S.S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. MR3153940

[27] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006. MR2274449