ON ASYMPTOTICALLY OPTIMAL CONFIDENCE REGIONS AND TESTS FOR HIGH-DIMENSIONAL MODELS

By Sara van de Geer, Peter Bühlmann and Ya'acov Ritov

ETH Zürich and The Hebrew University of Jerusalem

We propose a general method for constructing confidence intervals and statistical tests for single or low-dimensional components of a large parameter vector in a high-dimensional model. It can be easily adjusted for multiplicity taking dependence among tests into account. For linear models, our method is essentially the same as from Zhang and Zhang [37]: we analyze its asymptotic properties and establish its asymptotic optimality in terms of semiparametric efficiency. Our method naturally extends to generalized linear models with convex loss functions. We develop the corresponding theory which includes a careful analysis for Gaussian, sub-Gaussian and bounded correlated designs.

1. Introduction. Much progress has been made over the last decade in high-dimensional statistics where the number of unknown parameters greatly exceeds sample size. The vast majority of work has been pursued for point estimation such as consistency for prediction [14, 3], oracle inequalities and estimation of a high-dimensional parameter [7, 6, 36, 33, 23, 2, 25, 16] or variable selection [21, 38, 11, 34]. Other references and exposition to a broad class of models can be found in [12] or [5].

Very few work has been done for constructing confidence intervals, statistical testing and assigning uncertainty in high-dimensional sparse models. A major difficulty of the problem is the fact that sparse estimators such as the Lasso do not have a tractable limiting distribution: already in the low-dimensional setting, it depends on the unknown parameter [17] and hence the convergence to the limit is not uniform. Furthermore, bootstrap and even subsampling techniques are plagued by non-continuity of limiting distribu-

Y.R. gratefully acknowledges financial support from the Forschungsinstitut für Mathematik (FIM) at ETH Zürich.

AMS 2000 subject classifications: Primary 62J07; secondary 62J12,62F25

Keywords and phrases: Central limit theorem, Generalized linear model, Lasso, Linear model, Multiple testing, Semiparametric efficiency, Sparsity

tions. While in the low-dimensional setting, a modified bootstrap scheme has been proposed [8], it is unclear whether such a method can be extended to high-dimensional scenarios.

Some approaches for quantifying uncertainty include the following. The work in [35] implicitly contains the idea of sample splitting and corresponding construction of p-values and confidence intervals, and the procedure has been improved by using multiple sample splitting and aggregation of dependent p-values from multiple sample splits [24]. Stability Selection [22] and its modification [27] provides another route to estimate error measures for false positive selections in general high-dimensional settings. From another and mainly theoretical perspective, the work in [16] presents necessary and sufficient conditions for recovery with the Lasso $\hat{\beta}$ in terms of $\|\hat{\beta} - \beta^0\|_\infty$, where β^0 denotes the true parameter: bounds on the latter, which hold with probability at least say $1 - \alpha$, could be used in principle to construct (very) conservative confidence regions. Other recent work is discussed in Section 1.1 below.

We propose here a method which enjoys optimality properties when making assumptions on the sparsity and design matrix of the model. For a linear model, the procedure is largely the same as the one in [37] and closely related to the method in [15]. It is based on the Lasso and is "inverting" the corresponding KKT conditions. This yields a non-sparse estimator which has a Gaussian (limiting) distribution. We show, within a sparse linear model setting, that the estimator is optimal in the sense that it reaches the semiparametric efficiency bound. Our procedure can be used and is analyzed for high-dimensional sparse linear and generalized linear models and for regression problems with general convex (robust) loss functions.

1.1. *Related work.* Our work is closest to [37] (and also [15], see below) who proposed the semiparametric approach for distributional inference in a high-dimensional linear model. We take here a slightly different viewpoint, namely by inverting the KKT conditions from the Lasso, while relaxed projections are used in [37]: we describe in Section 2.4 the exact relations. Furthermore, our paper extends the results in [37] by: (i) treating generalized linear models and general convex loss functions; (ii) for linear models, we give conditions under which the procedure achieves the semiparametric efficiency bound and our analysis allows for rather general Gaussian, sub-Gaussian and bounded design. A related approach as in [37] was proposed in [4] based on Ridge regression which is clearly sup-optimal and inefficient with a detection rate (statistical power) larger than $n^{-1/2}$.

Very recently, and developed independently, the work in [15] provides a detailed analysis for linear models (but not covering generalized linear models as we do here) by considering a very similar procedure as in [37] and in our paper. They show that the detection limit is indeed in the $1/\sqrt{n}$-range and they provide a minimax test result; furthermore, they present extensive simulation results indicating that the Ridge-based method in [4] is overly conservative, which is in line with the theoretical results. Their optimality results are interesting and are complementary to the semiparametric optimality established here. Our results cover a substantially broader range of non-Gaussian designs in linear models, and we provide a rigorous analysis for correlated designs with covariance matrix $\Sigma \neq I$: the SDL-test in [15] assumes that Σ is known while we carefully deal with the issue when Σ^{-1} has to be estimated (and arguing why e.g. GLasso is not good for our purpose).

Another way and method to achieve distributional inference for high-dimensional models is given in [1] (claiming semiparametric efficiency). They use a two-stage procedure with a so-called post-double-selection as first and least squares estimation as second stage: as such, their methodology is radically different from ours.

2. High-dimensional linear models.

Consider a high-dimensional linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon, \tag{1}$$

with $n \times p$ design matrix $\mathbf{X} = [X_1, \ldots, X_p]$ ($n \times 1$ vectors X_j), $\varepsilon \sim \mathcal{N}_n(0, \sigma_\varepsilon^2 I)$ and unknown regression $p \times 1$ vector β^0. Throughout the paper, we assume that $p > n$. We denote by $S_0 = \{j;\ \beta_j^0 \neq 0\}$ the active set of variables and its cardinality by $s_0 = |S_0|$.

Our main goal is pointwise statistical inference for the components of the parameter vector β_j^0 ($j = 1, \ldots, p$) but we also discuss simultaneous inference for parameters $\beta_G^0 = \{\beta_j^0;\ j \in G\}$ where $G \subseteq \{1, \ldots, p\}$ is any group. To exemplify, we might want to test statistical hypotheses of the form $H_{0,j}:$ $\beta_j^0 = 0$ or $H_{0,G}:\ \beta_j^0 = 0$ for all $j \in G$, and when pursuing many tests, we aim for an efficient multiple testing adjustment taking dependence into account and being less conservative than say the Bonferroni-Holm procedure.

2.1. *The method: de-sparsifying the Lasso.*

The main idea is to invert the Karush-Kuhn-Tucker characterization of the Lasso. We will discuss in Sections 2.4 some alternative representations.

The Lasso [29] is defined as:

$$(2) \qquad \hat{\beta} = \hat{\beta}(\lambda) = \text{argmin}_{\beta \in \mathbb{R}^p}(\|Y - \mathbf{X}\beta\|_2^2/(2n) + \lambda\|\beta\|_1).$$

It is well-known that the estimator in (2) must fulfill the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} &-\mathbf{X}^T(Y - \mathbf{X}\hat{\beta}) + \lambda\hat{\tau} = 0, \\ &\|\hat{\tau}\|_\infty \le 1, \text{ and } \hat{\tau}_j = \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0. \end{aligned}$$

The vector $\hat{\tau}$ is arising from the sub-differential of $\|\beta\|_1$: using the first equation, we can always represent it as

$$(3) \qquad \lambda\hat{\tau} = \mathbf{X}^T(Y - \mathbf{X}\hat{\beta}).$$

The KKT conditions can be re-written with the notation $\hat{\Sigma} = n^{-1}\mathbf{X}^T\mathbf{X}$:

$$\hat{\Sigma}(\hat{\beta} - \beta^0) + \lambda\hat{\tau} = \mathbf{X}^T\varepsilon/n.$$

The idea is now to use a "relaxed form" of an inverse of $\hat{\Sigma}$. Suppose that $\hat{\Theta}$ is a reasonable approximation for such an inverse, then:

$$(4) \quad \hat{\beta} - \beta^0 + \hat{\Theta}\lambda\hat{\tau} = \hat{\Theta}\mathbf{X}^T\varepsilon/n - \Delta, \text{ where } \Delta = (\hat{\Theta}\hat{\Sigma} - I)(\hat{\beta} - \beta^0).$$

We will show in Theorem 2.2 that Δ is asymptotically negligible under some sparsity assumptions. This suggests the following estimator:

$$(5) \qquad \hat{b} = \hat{\beta} + \hat{\Theta}\lambda\hat{\tau} = \hat{\beta} + \hat{\Theta}\mathbf{X}^T(Y - \mathbf{X}\hat{\beta})/n,$$

using (3) in the second equation. This is essentially the same estimator as in [37], as discussed in Section 2.4, and it is of the same form as the SDL-procedure in [15], when plugging in the estimate $\hat{\Theta}$ for the population quantity $\Theta = \Sigma^{-1}$. With (4), we immediately obtain an asymptotic pivot when $\sqrt{n}\Delta$ is negligible, as is justified in Theorem 2.2 below:

$$(6) \qquad \sqrt{n}(\hat{b} - \beta^0) = W + o_P(1), \quad W|\mathbf{X} \sim \mathcal{N}_p(0, \sigma_\varepsilon^2\hat{\Theta}\hat{\Sigma}\hat{\Theta}^T).$$

An asymptotic pointwise confidence interval for β_j^0, when conditioning on $\mathbf{X}$ (or for fixed $\mathbf{X}$), is then given by:

$$\begin{aligned} &[\hat{b}_j - c(\alpha, n, \sigma_\varepsilon), \hat{b}_j + c(\alpha, n, \sigma_\varepsilon)], \\ &c(\alpha, n, \sigma_\varepsilon) = \Phi^{-1}(1 - \alpha/2)n^{-1/2}\sigma_\varepsilon\sqrt{(\hat{\Theta}\hat{\Sigma}\hat{\Theta}^T)_{jj}} \end{aligned}$$

If σ_ε is unknown, we replace it by a consistent estimator as discussed in Section 2.5.1.

2.1.1. *The Lasso for nodewise regression.* A prime example to construct the approximate inverse $\hat{\Theta}$ is given by the Lasso for the nodewise regression on the design $\mathbf{X}$ [21]: we use the Lasso p times for each regression problem X_j versus $\mathbf{X}_{-j}$, where the latter is the design sub-matrix without the jth column. For each $j = 1\ldots, p$,

$$\hat{\gamma}_j = \mathrm{argmin}_{\gamma}(\|X_j - \mathbf{X}_{-j}\gamma\|_2^2/(2n) + \lambda_j\|\gamma\|_1), \tag{7}$$

with components of $\hat{\gamma}_j = \{\hat{\gamma}_{j,k};\ k = 1, \ldots, p,\ k \neq j, \}$. Denote by

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

and by

$$\hat{T}^2 = \mathrm{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2), \quad \hat{\tau}_j^2 = (X_j - \mathbf{X}_{-j}\hat{\gamma}_j)^T X_j/n.$$

Then, define

$$\hat{\Theta}_{\mathrm{Lasso}} = \hat{T}^{-2}\hat{C}. \tag{8}$$

Not that although $\hat{\Sigma}$ is self-adjoint, its relaxed inverse, $\hat{\Theta}_{\mathrm{Lasso}}$, is not, In the sequel, we denote by

$$\hat{b}_{\mathrm{Lasso}} = \text{ the estimator in (5) with the nodewise Lasso from (8).} \tag{9}$$

We consider the jth row of $\hat{\Theta}$, denoted by $\hat{\Theta}_j$ (as a $p \times 1$ vector), and analogously for $\hat{C}_j$. Then, $\hat{\Theta}_{\mathrm{Lasso},j} = \hat{C}_j/\hat{\tau}_j^2$. Furthermore, because of the choice of $\hat{\tau}_j^2$ we have

$$X_j^T\mathbf{X}\hat{\Theta}_{\mathrm{Lasso},j}/n = 1. \tag{10}$$

Moreover, by the KKT conditions for (7):

$$\|\mathbf{X}_{-j}^T\mathbf{X}\hat{\Theta}_{\mathrm{Lasso},j}/n\|_\infty \leq \lambda_j/\hat{\tau}_j^2.$$

Hence we have

$$\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty \leq \lambda_j/\hat{\tau}_j^2,$$

where e_j is the j-th unit vector. We call this the extended KKT conditions.

We note that using e.g. the GLasso estimator for $\hat{\Theta}$ seems not optimal because (10) would fail and does not directly lead to desirable componentwise properties of the estimator $\hat{b}$ in (5) as established in Section 2.3.

2.2. *Theoretical result for fixed design.* We provide here a first result for fixed design $\mathbf{X}$. A crucial identifiability assumption on the design is the so-called compatibility condition [30]. For a $p \times 1$ vector β and a subset $S \subseteq \{1, \dots, p\}$, define β_S by:

$$\beta_{S,j} = \beta_j I(j \in S),\ \ j = 1, \dots, p.$$

Thus, β_S has zeroes for the components outside the set S. The compatibility condition requires a positive constant $\phi_0 > 0$ such that for all β satisfying $\|\beta_{S_0^c}\|_1 \le 3\|\beta_{S_0}\|_1$:

$$\|\beta_{S_0}\|_1^2 \le \frac{s_0}{\phi_0^2}\beta^T\hat{\Sigma}\beta.$$

The value ϕ_0^2 is called the compatibility constant. We make the following assumption:

(A1) The compatibility condition holds (for $\hat{\Sigma}$) with compatibility constant $\phi_0^2 > 0$. Furthermore, $\hat{\Sigma}_{jj} \le M^2 < \infty$ for some $0 < M < \infty$.

The assumption (A1) is briefly discussed in Section 2.3.2. We then obtain the following result.

THEOREM 2.1. *Consider the linear model in (1) with Gaussian error $\varepsilon \sim \mathcal{N}_n(0, \sigma_\varepsilon^2 I)$, and assume (A1). When using the Lasso for nodewise regression in (8) with $\lambda_j \equiv \lambda_{\max}\ \forall j$ and the Lasso in (2) with $\lambda \ge 2M\sigma_\varepsilon\sqrt{\frac{t^2+2\log(p)}{n}}$, we have:*

$$\hat{b}_{\text{Lasso}} - \beta^0 = \hat{\Theta}_{\text{Lasso}}\mathbf{X}^T\varepsilon/n + \Delta,$$
$$\mathbb{P}[\|\Delta\|_\infty \le \|\hat{T}^{-2}\|_\infty\lambda_{\max}4\lambda s_0/\phi_0^2] \ge 1 - 2\exp(-t^2/2),$$

where $\|A\|_\infty = \max_{j,k}|A_{j,k}|$ is the element-wise sup-norm for a matrix A. A proof is given in Section 5.2.

Theorem 2.1 gives a probabilistic bound for the error $\|\Delta\|_\infty$. Note that since the design is fixed, $\hat{\Theta}_{\text{Lasso}}$ is fixed and non-random, and $\|\hat{T}^{-2}\|_\infty$ is observed, and hence we compare $\|\Delta\|_\infty$ to a known constant. We will show in the proof of Theorem 2.2, see Lemma 5.3 in Section 5, that $\|\hat{T}^{-2}\|_\infty$ is asymptotically bounded and that the correct normalization factor for $\hat{b}_{\text{Lasso}}$ is $\sqrt{n}$. Thus, asymptotically, when choosing $\lambda_{\max} \asymp \lambda \asymp \sqrt{\log(p)/n}$, and if $s_0\log(p)n^{-1/2} = o(1)$, the error term $\|\Delta\|_\infty = o_{\mathbb{P}}(n^{-1/2})$ is negligible and $\sqrt{n}(\hat{b}_{\text{Lasso}} - \beta^0) \approx \mathcal{N}_p(0, \sigma_\varepsilon^2\hat{\Theta}_{\text{Lasso}}\hat{\Sigma}\hat{\Theta}^T)$. The details are discussed next.

2.3. *Conditioning on random design and optimality.* In order to make some further theoretical statements than in Theorem 2.1, we consider an asymptotic framework with random design, but where the analysis for the statistical inference is pursued conditioning on the design. The latter is the common approach for low-dimensional linear models and implemented as the standard procedure in software packages. Some results then follow for random design as well.

The asymptotic framework uses a scheme where $p = p_n \geq n \to \infty$ in model (1) and thus, $Y = Y_n$, $\mathbf{X} = \mathbf{X}_n$, $\beta^0 = \beta^0_n$ and $\sigma^2_\varepsilon = \sigma^2_{\varepsilon,n}$ are all (potentially) depending on n. In the sequel, we usually suppress the index n. We make the following assumption.

(A2) The rows of $\mathbf{X}$ are i.i.d. realizations from a Gaussian distribution P_X whose p-dimensional covariance matrix Σ has smallest eigenvalue $\Lambda^2_{\min} \geq L > 0$, and $\|\Sigma\|_\infty = \max_{j,k} |\Sigma_{jk}| = \mathcal{O}(1)$.

The Gaussian assumption is relaxed in Section 2.3.4. We will assume below some sparsity with respect to rows of $\Theta = \Sigma^{-1}$ and define:

$$s_j = \sum_{k \neq j} I(\Theta_{jk} \neq 0).$$

We then have the following main result.

Theorem 2.2. *Consider the linear model (1) with Gaussian error $\varepsilon \sim \mathcal{N}_n(0, \sigma^2_\varepsilon I)$. Assume (A2) and the sparsity assumptions $s_0 = o(n^{1/2}/\log(p))$ and $s_j \leq s_{\max} = o(n/\log(p))\ \forall j$. Consider the choice of the regularization parameters $\lambda \asymp \sqrt{\log(p)/n}$ for the Lasso in (2) and $\lambda_j \equiv \lambda_{\max} \asymp \sqrt{\log(p)/n}\ \forall j$ for the Lasso for nodewise regression in (8). Then:*

$$\begin{aligned}
&\sqrt{n}(\hat{b}_{\text{Lasso}} - \beta^0) = W_n + \Delta_n, \\
&W_n|\mathbf{X} \sim \mathcal{N}_p(0, \sigma^2_\varepsilon \Omega_n), \quad \Omega_n = \hat{\Theta}\hat{\Sigma}\hat{\Theta}^T, \\
&\|\Delta_n\|_\infty = o_{\mathbb{P}}(1).
\end{aligned}$$

Furthermore, $\|\Omega_n - \Sigma^{-1}\|_\infty = o_{\mathbb{P}}(1)$ as $n \to \infty$.

A proof is given in Section 5.5.

Theorem 2.2 has various implications. For one-dimensional components, we obtain for all $x \in \mathbb{R}$:

(11) $\mathbb{P}[\sqrt{n}(\hat{b}_{\text{Lasso};j} - \beta^0_j)/(\sigma_\varepsilon \sqrt{\Sigma^{-1}_{jj}}) \leq x|\mathbf{X}] - \Phi(x) = o_{\mathbb{P}}(1)\ (n \to \infty),$

where $\Phi(\cdot)$ denotes the c.d.f. of $\mathcal{N}(0,1)$. Since the the limiting distribution is independent of $\mathbf{X}$, the statement also holds unconditionally for random design. Furthermore, for any group $G \subseteq \{1,\dots,p\}$ which is potentially large, we have that for all $x \in \mathbb{R}$:

$$\mathbb{P}[\max_{j \in G} |\sqrt{n}(\hat{b}_{\mathrm{Lasso};j} - \beta_j^0)| \le x|\mathbf{X}] - \mathbb{P}[\max_{j \in G} |W_{n;j}| \le x|\mathbf{X}] = o_{\mathbb{P}}(1).$$

Therefore, the asymptotic distribution of $\max_{j \in G} \sqrt{n}|\hat{b}_{\mathrm{Lasso};j}||\mathbf{X}$ under the null-hypothesis $H_{0,G};\ \beta_j^0 = 0\ \forall j \in G$ is asymptotically equal to the maximum of dependent Gaussian variables $\max_{j \in G} |W_{n;j}||\mathbf{X}$ whose distribution can be easily simulated since Ω_n is known, see also Section 2.5.

Remark 2.1. *Theorem 2.2 can be extended to allow for non-Gaussian errors:* $\sqrt{n}(\hat{b}_{\mathrm{Lasso}} - \beta^0) = W_n + \Delta_n$*, with* $\|\Delta_n\|_\infty = o_{\mathbb{P}}(1)$*,* $Cov(W_n|\mathbf{X}) = \sigma_\varepsilon^2 \Omega_n$ *but* $W_n|\mathbf{X}$ *generally not exactly Gaussian. Often though, a central limit theorem argument can be used to obtain approximate Gaussianity of low-dimensional components of* $W_n|\mathbf{X}$*, see also Section 3.2.*

2.3.1. *Uniform convergence.* The statements of Theorem 2.2 also hold in a uniform sense, and thus, the derived confidence intervals and tests in Section 2.5 below are honest [19]. We consider the set of parameters

$$\mathcal{B}(s) = \{\beta \in \mathbb{R}^p;\ \|\beta\|_0^0 \le s\}.$$

We then have the following for $\hat{b}_{\mathrm{Lasso}}$ in (9).

Corollary 2.1. *Assume the conditions of Theorem 2.2 with* $\beta^0 \in \mathcal{B}(s_0)$ *and* $s_0 = o(n^{1/2}/\log(p))$*. Then, when using* $\lambda \asymp \sqrt{\log(p)/n}$ *for the Lasso in (2), and* $\lambda_j \equiv \lambda_{\max} \asymp \sqrt{\log(p)/n}\ \forall j$ *for the Lasso for nodewise regression in (8):*

$$\begin{aligned}
&\sqrt{n}(\hat{b}_{\mathrm{Lasso}} - \beta^0) = W_n + \Delta_n, \\
&W_n|\mathbf{X} \sim \mathcal{N}_p(0, \sigma_\varepsilon^2 \Omega_n), \quad \Omega_n = \hat{\Theta}\hat{\Sigma}\hat{\Theta}^T, \\
&\sup_{\beta^0 \in \mathcal{B}(s_0)} \|\Delta\|_\infty = o_{\mathbb{P}}(1).
\end{aligned}$$

Moreover, as in Theorem 2.2, $\|\Omega_n - \Sigma^{-1}\|_\infty = o_{\mathbb{P}}(1)\ (n \to \infty)$*.*

The proof is exactly the same as for Theorem 2.2 by simply noting that $\sup_{\beta^0 \in \mathcal{B}(s_0)} \|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(s_0\sqrt{\log(p)/n})$ (with high probability, the compatibility constant is still bounded from below by $L/2$). □

Corollary 2.1 implies that for $j \in \{1, \ldots, p\}$:

$$\sup_{\beta^0 \in \mathcal{B}(s)} |\mathbb{P}[\sqrt{n}(\hat{b}_{\text{Lasso};j} - \beta^0_j)/(\sigma_\varepsilon \sqrt{(\Sigma^{-1})_{jj}}) \le x|\mathbf{X}] - \Phi(x)| = o_{\mathbb{P}}(1) \ (n \to \infty).$$

2.3.2. *Discussion of the assumptions.* The compatibility condition (A1) is weaker than many others which have been proposed such as assumptions on restricted or sparse eigenvalues [31]: a slight relaxation by a constant factor has recently been given in [28]: we could work with this slightly less established condition without changing the asymptotic behavior of our results. Assumption (A2) is rather weak as it concerns the population covariance matrix.

Regarding the sparsity assumption for s_0 in Theorem 2.1, our technique crucially uses the ℓ_1-norm bound $\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(s_0\sqrt{\log(p)/n})$, see Lemma 5.1. In order that this ℓ_1-norm converges to zero, the sparsity constraint $s_0 = o(\sqrt{n/\log(p)})$ is usually required. Our sparsity assumption is slightly stricter by the factor $\log(p)^{-1/2}$ (because the normalization factor is $\sqrt{n}$), namely $s_0 = o(\log(p)^{-1/2}\sqrt{n/\log(p)}) = o(n^{1/2}/\log(p))$.

2.3.3. *Optimality and semiparametric efficiency.* Corollary 2.1 establishes, in fact, that for any j, $\hat{b}_{\text{Lasso},j}$ is an asymptotically efficient estimator of β^0_j, in the sense that it is asymptotically normal, with asymptotic variance converging, as $n \to \infty$ to the variance of the best estimator. Consider, the *one*-dimensional sub-model,

$$Y = \beta_j(X_j - \mathbf{X}_{-j}\gamma^0_j) + \mathbf{X}_{-j}(\beta^0_{-j} + \beta^0_j\mathbf{X}_{-j}\gamma^0_j) + \varepsilon, \tag{12}$$

where γ^0_j is the population analog of $\hat{\gamma}_j$, i.e., $X_j - \mathbf{X}_{-j}\gamma^0_j$ is the projection of X_j to the subspace orthogonal to $\mathbf{X}_{-j}$. Clearly, this is a linear submodel of the general model (1), passing through the true point. The Gauss-Markov theorem argues that the best variance of an unbiased estimator of β_j in (12) is given by $\sigma^2_\varepsilon/\text{Var}(X_j - \mathbf{X}_{-j}\gamma^0_j)$. However, Corollary 2.1 shows that this is the asymptotic variance of $\hat{b}_{\text{Lasso},j}$. Thus, $\hat{b}_{\text{Lasso},j}$ is asymptotically normal, with the variance of the best possible unbiased estimator. Note, that any regular estimator (regular at least on parametric sub-models) must be asymptotically unbiased.

The main difference between this and most of the other papers on complex models is that usually the Lasso is considered as solving a nonparametric model with parameter whose dimension p is increasing to infinity, while we consider the problem as a semiparametric model in which we concentrate on a low dimensional model of interest, e.g., β_j^0, while the rest of the parameters, β_{-j}^0, are considered as nuisance parameters. That is, we consider the problem as a semiparametric one. In the rest of this discussion we try to put the model in a standard semiparametric framework in which there is an infinite dimensional population model.

Without loss of generality, the parameter of interest is β_1^0, i.e., the first component. Consider the random design model

$$Y = \beta_1^0 X_1 + K(Z) + \varepsilon,\ \ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2), \tag{13}$$

where Z is a Gaussian process. Suppose that with sample size n, we observe a sample from $Y, X_1^n, \ldots, X_{p_n}^n$ such that

$$\begin{aligned}
&Y = \sum_{j=1}^{p} \beta_j^n X_j^n + \varepsilon^n, \quad \varepsilon^n \text{ independent of } X_1^n, \ldots, X_p^n \\
&K(Z) - \sum_{j=2}^{p} \beta_j^n X_j^n \xrightarrow{\mathbb{P}} 0, \\
&\mathbb{E}[X_1|Z] - \sum_{j=2}^{p} \gamma_j^n X_j^n \xrightarrow{\mathbb{P}} 0 \\
&\big(K(Z) - \sum_{j=2}^{p} \beta_j^n X_j^n\big)\big(\mathbb{E}[X_1|Z] - \sum_{j=2}^{p} \gamma_j^n X_j^n\big) = o_{\mathbb{P}}(n^{-1/2}).
\end{aligned} \tag{14}$$

Theorem 2.3. *Suppose* (14) *and the conditions of Theorem* 2.2 *are satisfied, then*

$$\hat{b}_{\text{Lasso};1} = \beta_1^0 + \frac{1}{n} \sum_{i=1}^{n} \big(X_{1i} - \mathbb{E}[X_{1i}|Z_i]\big)\varepsilon_i + o_{\mathbb{P}}(n^{-1/2}).$$

In particular, the limiting variance of $\sqrt{n}(\hat{b}_{\text{Lasso};1} - \beta_1^0)$ *reaches the information bound* $\sigma_\varepsilon^2 / \mathbb{E}[(X_1 - \mathbb{E}[X_1|Z])^2]$. *Furthermore,* $\hat{b}_{\text{Lasso};1}$ *is regular at the one-dimensional parametric sub-model with component* β_1 *and hence,* $\hat{b}_{\text{Lasso};1}$ *is asymptotically efficient for estimating* β_1^0.

A proof is given in Section 6.1.

As a concrete example, condition (14) and the conditions of Theorem 2.2 are satisfied when:

$$Y = \sum_{j=1}^{\infty} \beta_j^0 X_j + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2),\ \beta^0 \in \mathcal{B},\ (X_j)_{j \in \mathbb{N}} \sim P_X^0 \in \mathcal{P},$$

$$X_j^n \equiv X_j \ \forall j = 1, \ldots, p_n, \tag{15}$$

$$\begin{aligned} \text{where} \qquad & \mathcal{B} = \{(\beta_j)_{j\in\mathbb{N}};\ \|\beta\|_0^0 < \infty\}, \\ & \mathcal{P} = \{P_X \text{ Gaussian on } \mathbb{R}^\infty; \\ & \qquad 0 < \min_{S \subset \mathbb{N}} \nu_{\min}(\Sigma_{S,S}),\ \max_{j\in\mathbb{N}} |\Sigma_{j,j}| < \infty, \\ & \qquad \text{and } \forall j,\ \mathbb{E}[X_j] = 0,\ \|\gamma_j(P_X)\|_0^0 < \infty\}. \end{aligned}$$

Thereby, $\Sigma_{S,S}$ is the covariance matrix of $\{X_j;\ j \in S\}$, and $\Lambda_{\min}^2(\cdot)$ denotes the minimal eigenvalue. Furthermore, $\gamma_j(P_X) = \text{argmin}_\gamma \mathbb{E}_{P_X}[(X_j - \sum_{k \neq j} \gamma_k X_k)^2]$ are the coefficients from the projection of X_j on all other infinitely many variables $\{X_k;\ k \neq j\}$. The assumption about the covariances is equivalent to saying that $(X_j)_{j\in\mathbb{N}}$ has a positive definite covariance function. A proof that this example fulfills the required assumptions is given in Section 6.1.

2.3.4. *Non-Gaussian design.* We extend here Theorem 2.2 to allow for non-Gaussian designs. Besides covering a broader range of designs for linear models, the result is important for the treatment of generalized linear models in Section 3.

Consider a random design matrix $\mathbf{X}$ with i.i.d. rows having mean zero and population covariance Σ with its inverse (assumed to exist) $\Theta = \Sigma^{-1}$. Denote by $\gamma_j = \text{argmin}_\gamma \mathbb{E}[\|X_j - \mathbf{X}_{-j}\gamma\|_2^2]$ (which was denoted by $\gamma_j(P_X)$ in the previous subsection). Define the error $\eta_j := X_j - \mathbf{X}_{-j}\gamma_j$ with variance $\tau_j^2 = \mathbb{E}[\|\eta_j\|_2^2/n] = 1/\Theta_{jj}$. We make the following assumptions.

(B1) The design $\mathbf{X}$ has either i.i.d. sub-Gaussian rows or i.i.d. rows and for some $K \geq 1$, $\|\mathbf{X}\|_\infty = \max_{i,j} |\mathbf{X}_{ij}| = \mathcal{O}(K)$. The latter we call the bounded case. The strongly bounded case assumes in addition that $\|\mathbf{X}_{-j}\gamma_j\|_\infty = \mathcal{O}(K)$. We write $K_0 = 1$ in the sub-Gaussian case and $K_0 = K$ in the (strongly) bounded case (where K_0 appears in some of the conditions below).

(B2) It holds that $K_0^2 s_j \sqrt{\log(p)/n} = o(1)$. In the sub-Gaussian case we relax this to $\sqrt{s_j \log(p)/n} = o(1)$.

(B3) The smallest eigenvalue of Σ satisfies $0 < L \leq \Lambda_{\min}^2$ and $\|\Sigma\|_\infty = \mathcal{O}(1)$.

(B4) It holds that $\mathbb{E}\eta_{1,j}^4 = \mathcal{O}(K_0^4)$.

We note that the strongly bounded case in (B1) follows from the bounded case if $\|\gamma_j\|_1 = \mathcal{O}(1)$, and the sub-Gaussian assumption is stronger. Furthermore, in the sub-Gaussian case or the strongly bounded case, the assumption (B4) follows automatically. Assumption (B2) is a standard sparsity assumption for Θ. Finally, assumption (B3) implies that $\|\Theta_j\|_2 \leq \Lambda_{\min}^{-4} \leq L^{-2} = \mathcal{O}(1)$ uniformly in j (see (34)), so that in particular $\tau_j^2 = 1/\Theta_{jj}$ stays away from zero. Note that (B3) also implies $\tau_j^2 \leq \Sigma_{jj} = \mathcal{O}(1)$ uniformly in j.

Theorem 2.4. *Suppose the conditions (B1)-(B4) hold. Denote by $\hat{\Theta}$ and $\hat{\tau}_j^2$ the estimates from the nodewise Lasso in (8). Then for $\lambda_j \asymp K_0\sqrt{\log(p)/n}$ we have:*

$$\|\hat{\Theta}_j - \Theta_j\|_1 = \mathcal{O}_{\mathbb{P}}\left(K_0 s_j \sqrt{\frac{\log(p)}{n}}\right), \quad \|\hat{\Theta}_j - \Theta_j\|_2 = \mathcal{O}_{\mathbb{P}}\left(K_0 \sqrt{\frac{s_j \log(p)}{n}}\right),$$

$$|\hat{\tau}_j^2 - \tau_j^2| = \mathcal{O}_{\mathbb{P}}\left(K_0 \sqrt{\frac{s_j \log(p)}{n}}\right).$$

Furthermore,

$$|\hat{\Theta}_j^T \Sigma \hat{\Theta}_j - \Theta_{jj}| \leq \|\Sigma\|_\infty \|\hat{\Theta}_j - \Theta_j\|_1^2 \wedge \Lambda_{\max}^2 \|\hat{\Theta}_j - \Theta_j\|_2^2 + 2|\hat{\tau}_j^2 - \tau_j^2|,$$

where $\Lambda_{\max}^2$ is the maximal eigenvalue of Σ. If the conditions hold uniformly in j then in the sub-Gaussian or strongly bounded case the results are also uniform in j.
Finally, for the sub-Gaussian or strongly bounded case, if the conditions hold uniformly in j, $K_0 s_{\max}\sqrt{\log(p)/n} = o(1)$ and $s_0 = o(n^{1/2}/\log(p))$, the statements of Theorem 2.2 hold.

A proof is given in Section 5.6.

2.4. *Linear projection estimator and bias correction with the Lasso.* We discuss here a relation to the method in [37]. The estimator in (5) has an (almost) equivalent representation. Consider any $n \times 1$ score vector Z_j, for $j = 1\ldots,p$. Pursuing a linear projection of Y onto Z_j we obtain:

$$Z_j^T Y = Z_j^T X_j \beta_j^0 + \sum_{k \neq j} Z_j^T X_k \beta_k^0 + Z_j^T \varepsilon/n.$$

Therefore,

$$\frac{Z_j^T Y}{Z_j^T X_j} - \beta_j^0 = \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0 + \frac{Z_j^T \varepsilon}{Z_j^T X_j},$$

with a bias term $\sum_{k\neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \beta_k^0$. When estimating the bias with the Lasso, we obtain the following estimator:

$$\hat{b}_{\text{proj};j} = \frac{Z_j^T Y}{Z_j^T X_j} - \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} \hat{\beta}_k. \tag{16}$$

Similarly as in (4), we can derive an approximate pivot (for fixed design):

$$\begin{aligned} &\sqrt{n}(\hat{b}_{\text{proj}} - \beta^0) = W - \sqrt{n}\Delta_{\text{proj}} \\ &W = \mathcal{N}_p(0, \sigma_\varepsilon^2 \Omega_{\text{proj}}), \quad (\Omega_{\text{proj}})_{jk} = \frac{Z_j^T Z_k}{(Z_j^T X_j)(Z_k^T X_k)}, \\ &\Delta_{\text{proj};j} = \sum_{k \neq j} \frac{Z_j^T X_k}{Z_j^T X_j} (\hat{\beta}_k - \beta_k^0) \quad (j = 1, \dots, p). \end{aligned} \tag{17}$$

A typical choice of a score vector is $Z_j = X_j - \mathbf{X}_{-j}\hat{\gamma}_j$ where $\hat{\gamma}$ is an estimated vector of coefficients when regressing X_j versus $\mathbf{X}_{-j}$: a prominent example is the nodewise Lasso in (7) in Section 2.1.1

$$Z_{\text{Lasso};j} = X_j - \mathbf{X}_{-j}\hat{\gamma}_{\text{Lasso};j}. \tag{18}$$

Another choice is using $\hat{\Theta}$ from Section 2.1.

$$\mathbf{Z} = \mathbf{X}\hat{\Theta}^T, \tag{19}$$

whose jth column vector build the $n \times 1$ score vectors Z_j $(j = 1, \dots, p)$.

The estimators in (5) and (16) are equivalent whenever the vectors Z_j in (19) and in (16) coincide, and if $Z_j^T X_j / n = 1$ for all j. The latter is true when enforcing, for each j,

$$(\hat{\Theta}\hat{\Sigma})_{jj} = 1 \tag{20}$$

for the estimator in (5); for (16), we can always re-scale Z_j such that $Z_j^T X_j / n = 1$ (by re-scaling $Z_j \leftarrow Z_j / (Z_j^T X_j / n)$). Thus, the main condition to make the estimators equal is (20), and it holds for the nodewise Lasso as shown in (10).

2.5. *Confidence intervals and hypothesis testing.* We assume in this section an estimator $\hat{b}$ which satisfies:

$$\sqrt{n}(\hat{b} - \beta^0) = W_n + \Delta_n,$$

$$\|\Delta_n\|_\infty = o_{\mathbb{P}}(1), \quad W_n|\mathbf{X} \sim \mathcal{N}_p(0, \sigma_\varepsilon^2 \Omega). \tag{21}$$

This holds for $\hat{b}_{\text{Lasso}}$ in (5) assuming sparsity and design conditions as discussed in Theorem 2.2.

Confidence intervals and statistical hypothesis tests, when conditioning on $\mathbf{X}$, can be immediately derived from such an approximate pivot. For one-dimensional parameters β_j^0, the two-sided confidence interval and statistical test for $H_{0,j}: \ \beta_j^0 = 0$ are given by

$$I_j = [\hat{b}_j - \Phi^{-1}(1-\alpha/2)\sigma_\varepsilon \sqrt{\Omega_{jj}}, \hat{b}_j + \Phi^{-1}(1-\alpha/2)\sigma_\varepsilon \sqrt{\Omega_{jj}}]$$

and the p-value

$$P_j = 2\left(1 - \Phi(\frac{\hat{b}_{j;\text{observ}}}{\sigma_\varepsilon \sqrt{\Omega_{jj}}})\right), \tag{22}$$

where $\hat{b}_{j;\text{observ}}$ denotes the observed value of the statistic $\hat{b}_j$.

For simultaneous inference, we focus on testing $H_{0,G}: \beta_j^0 = 0$ for all $j \in G$ versus the alternative $H_{A,G}: \ \beta_j^0 \neq 0$ for at least one $j \in G$, where $G \subseteq \{1, \ldots, p\}$ is an arbitrary set. As a concrete test-statistic, consider

$$T_G = \max_{j \in G} |\hat{b}_j|/\sigma_\varepsilon$$

whose distribution under $H_{0,G}$ can be approximated by $T_{W,G} = \max_{j \in G} |W_j|/\sigma_\varepsilon$. Its distribution $J_G(c) = \mathbb{P}[\max_{j \in G} |W_j|/\sigma_\varepsilon \leq c]$ can be easily simulated by generating Gaussian variables $\tilde{W} \sim \mathcal{N}_{|G|}(0, n^{-1}\Omega_{G,G})$ which do not involve σ_ε. Denote by $\hat{\gamma}_{G;\text{observ}} = \max_{j \in G} \sqrt{n}|\hat{b}_{j;\text{observ}}/\sigma_\varepsilon|$. Then, the p-value for $H_{0,G}$, against the alternative being the complement $H_{0,G}^c$, is defined as

$$P_G = 1 - J_G(\hat{\gamma}_{G;\text{observ}}). \tag{23}$$

2.5.1. *Estimation of σ_ε^2.* For the construction of confidence intervals and hypothesis tests, we need an estimate for σ_ε. The scaled Lasso [28] yields a consistent estimator for this quantity, under the assumptions made for Theorem 2.2. We then simply plug-in an estimate $\hat{\sigma}_\varepsilon$ into (22) or into $\hat{\gamma}_G$ for (23).

2.5.2. *Multiple testing adjustment.* Based on many single p-values, we can use standard procedures for multiple testing adjustment to control for various type I error measures. The representation from Theorem 2.1 or 2.2 with $\|\Delta\|_\infty$ being sufficiently small allows to construct a multiple testing adjustment which takes the dependence in terms of the covariance Ω (see Theorem 2.2) into account: the exact procedure is described in [4]. Especially when having strong dependence among the p-values, the method is much less conservative than the Bonferroni-Holm procedure for strongly controlling the familywise error rate.

3. Generalized linear models and general convex loss functions. We show here that the idea of de-sparsifying ℓ_1-norm penalized estimators and corresponding theory from Section 2 carries over to models with convex loss functions such as generalized linear models (GLMs).

3.1. *The setting and de-sparsifying the ℓ_1-norm regularized estimator.* We consider the following framework with $1 \times p$ vectors of covariables $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ and univariate responses $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ for $i = 1, \ldots, n$. As before, we denote by $\mathbf{X}$ the design matrix with ith row equal to x_i. At the moment, we do not distinguish whether $\mathbf{X}$ is random or fixed (e.g. when conditioning on $\mathbf{X}$).

For $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ being a $1 \times p$ vector, we have a loss function

$$\rho_\beta(y, x) = \rho(y, x\beta) \quad (\beta \in \mathbb{R}^p)$$

which is assumed to be a strictly convex function in $\beta \in \mathbb{R}^p$. We now define

$$\dot{\rho}_\beta := \frac{\partial}{\partial \beta}\rho_\beta, \ \ddot{\rho}_\beta := \frac{\partial}{\partial\beta\partial\beta^T}\rho_\beta,$$

where we implicitly assume that the derivatives exist. For a function $g : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, we write $P_n g := \sum_{i=1}^n g(y_i, x_i)/n$ and $Pg := \mathbb{E}P_n g$. Moreover, we let $\|g\|_n^2 := P_n g^2$ and $\|g\|^2 := Pg^2$.

The ℓ_1-norm regularized estimator is

$$\hat{\beta} = \text{argmin}_\beta (P_n \rho_\beta + \lambda\|\beta\|_1). \tag{24}$$

As in Section 2.1, we de-sparsify the estimator. For this purpose, define

$$\hat{\Sigma} := P_n \ddot{\rho}_{\hat{\beta}}. \tag{25}$$

Note that in general, $\hat{\Sigma}$ depends on $\hat{\beta}$ (an exception being the squared error loss). We construct $\hat{\Theta} = \hat{\Theta}_{\text{Lasso}}$ by doing a nodewise Lasso with $\hat{\Sigma}$ as input as detailed below in (29). We then define

$$\hat{b} := \hat{\beta} - \hat{\Theta} P_n \dot{\rho}_{\hat{\beta}}. \tag{26}$$

The estimator in (5) is a special case of (26) with squared error loss.

3.1.1. *Lasso for nodewise regression with matrix input.* Denote by $\hat{\Sigma}$ a matrix which we want to approximately invert using the nodewise Lasso. For every row j, we consider the optimization

$$\hat{\gamma}_j = \text{argmin}_{\gamma_j} (\hat{\Sigma}_{jj} - 2\hat{\Sigma}_{j,\setminus j}\gamma_j + \gamma_j^T \hat{\Sigma}_{\setminus j,\setminus j}\gamma_j + \lambda_j \|\gamma_j\|_1), \tag{27}$$

where $\hat{\Sigma}_{j,\setminus j}$ denotes the jth row of $\hat{\Sigma}$ without the diagonal element (j,j), and $\hat{\Sigma}_{\setminus j,\setminus j}$ is the submatrix without the jth row and jth column. We note that for the case where $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}/n$, $\hat{\gamma}$ is the same as in (7).

Based on $\hat{\gamma}_j$ from (27), we compute

$$\hat{\tau}_j^2 = \hat{\Sigma}_{jj} - \hat{\Sigma}_{j,\setminus j}\hat{\gamma}_j. \tag{28}$$

Having $\hat{\gamma}_j$ and $\hat{\tau}_j^2$ from (27) and (28), we define the nodewise Lasso as

$$\hat{\Theta}_{\text{Lasso}} \text{ as in (8) using (27)-(28) from matrix input } \hat{\Sigma} \text{ in (25).} \tag{29}$$

Moreover, we denote by

$$\hat{b}_{\text{Lasso}} = \hat{b} \text{ from (26) using the nodewise Lasso from (29).}$$

Computation of (27) and hence of $\hat{\Theta}$ can be done efficiently via coordinate descent using the KKT conditions to characterize the zeroes. Furthermore, an active set strategy leads to additional speed-up. See for example [13] and [20].

3.2. *Theoretical results.* We show here that the components of the estimator $\hat{b}$ in (26), when normalized with the easily computable standard error, converge to a standard Gaussian distribution. Based on such a result, the construction of confidence intervals and tests is straightforward.

Let $\beta^0 \in \mathbb{R}^p$ be the unique minimizer of $P\rho_\beta$ with s_0 denoting the number of non-zero coefficients. We use analogous notation as in Section 2.3

but with modifications for the current context. The asymptotic framework, which allows for Gaussian approximation of averages, is as in Section 2.3 for $p = p_n \geq n \to \infty$ and thus, $y_1, \ldots, y_n = Y_n$, $\mathbf{X} = \mathbf{X}_n$, $\beta^0 = \beta^0_n$ and underlying parameters are all (potentially) depending on n. As before, we usually suppress the corresponding index n.

We make the following assumptions which are discussed in Section 3.3.1. Thereby, we assume (C3), (C5), (C6) and (C8) for some constant $K \geq 1$ and furthermore, λ, λ_* and s_* are positive constants.

(C1) The derivatives

$$\dot\rho(y,a) := \frac{d}{da}\rho(y,a),\ \ddot\rho(y,a) := \frac{d^2}{da^2}\rho(y,a),$$

exist for all y, a, and for some δ-neighborhood ($\delta > 0$), $\ddot\rho(y,a)$ is Lipschitz:

$$\max_{a_0 \in \{x_i\beta^0\}} \sup_{|a-a_0|\vee|\hat a - a_0| \leq \delta} \sup_{y \in \mathcal{Y}} \frac{|\ddot\rho(y,a) - \ddot\rho(y,\hat a)|}{|a - \hat a|} \leq 1.$$

Moreover

$$\max_{a_0 \in \{x_i\beta^0\}} \sup_{|a-a_0| \leq \delta} \sup_{y \in \mathcal{Y}} |\ddot\rho(y,a)| = \mathcal{O}(1).$$

(C2) It holds that $\|\hat\beta - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(s_0\lambda)$, $\|\mathbf{X}(\hat\beta - \beta^0)\|^2 = \mathcal{O}_{\mathbb{P}}(s_0\lambda^2)$, and $\|\mathbf{X}(\hat\beta - \beta^0)\|_n^2 = \mathcal{O}_{\mathbb{P}}(s_0\lambda^2)$.

(C3) It holds that $\|\mathbf{X}\|_\infty = \max_{i,j}|\mathbf{X}_{ij}| = \mathcal{O}(K)$.

(C4) It holds that $\|P_n\ddot\rho_{\hat\beta}\hat\Theta_j - e_j\|_\infty = \mathcal{O}_{\mathbb{P}}(\lambda_*)$.

(C5) It holds that $\|\mathbf{X}\hat\Theta_j\|_\infty = \mathcal{O}_{\mathbb{P}}(K)$ and $\|\hat\Theta_j\|_1 = \mathcal{O}_{\mathbb{P}}(\sqrt{s_*})$.

(C6) It holds that $\|(P_n - P)\dot\rho_{\beta^0}\dot\rho^T_{\beta^0}\|_\infty = \mathcal{O}_{\mathbb{P}}(K^2\lambda)$.

(C7) For every j, the random variable

$$\frac{\sqrt{n}(\hat\Theta P_n \dot\rho_{\beta^0})_j}{\sqrt{(\hat\Theta P \dot\rho_{\beta^0}\dot\rho^T_{\beta^0}\hat\Theta^T)_{jj}}}$$

converges weakly to a $\mathcal{N}(0,1)$-distribution.

(C8) It holds that

$$Ks_0\lambda^2 = o(n^{-\frac{1}{2}}),\ \lambda_*\lambda s_0 = o(n^{-\frac{1}{2}}),\ \text{and } K^2 s_*\lambda + K^2\sqrt{s_0}\lambda = o(1).$$

We note that often the regularization parameters in (27) are the same and λ_* can be chosen as $\lambda_* = \lambda_{\max} \equiv \lambda_j$, see also Section 3.3.1. Furthermore, when assuming that a population $\Theta = \Sigma^{-1}$ exists for $\Sigma = P\ddot{\rho}_{\beta^0}$, s_* can be chosen as $s_{\max}$ which is the maximal row sparsity of Θ. The following main result holds for fixed or random design according to whether the assumptions hold for one or the other case.

Theorem 3.1. *Assume (C1)-(C8). For the estimator in (26), we have for each $j \in \{1, \dots, p\}$:*

$$\sqrt{n}(\hat{b}_j - \beta^0_j)/\hat{\sigma}_j = W_j + o_{\mathbf{P}}(1),$$

where $W_j \sim \mathcal{N}(0,1)$ and $\hat{\sigma}^2_j = (\hat{\Theta} P_n \dot{\rho}_{\hat{\beta}} \dot{\rho}^T_{\hat{\beta}} \hat{\Theta}^T)_{jj}$.

A proof is given in Section 5. Assumption (C1) of Theorem 3.1 means that we regress to the classical conditions for asymptotic normality in the one-dimensional case as in for example [9]. Assumption (C8) is a sparsity assumption: for $K = O(1)$ and choosing $\lambda_* (= \lambda_{\max}) \asymp \lambda \asymp \sqrt{\log(p)/n}$ the condition reads as $s_0 = o(\sqrt{n}/\log(p))$ (as in Theorem 2.2) and $s_*(= s_{\max}) = o(\sqrt{n/\log(p)})$. All the other assumptions (C2)-(C7) follow essentially from the conditions of Corollary 3.1 presented later, with the exception that (C3) is straightforward to understand. For more details see Section 3.3.1.

3.3. *About nodewise regression with certain random matrices.* We justify in this section most of the assumptions for Theorem 3.1 when using the nodewise Lasso estimator $\hat{\Theta} = \hat{\Theta}_{\text{Lasso}}$ as in (29) and when the matrix input is parameterized by $\hat{\beta}$ as for standard generalized linear models. For notational simplicity, we drop the subscript "Lasso" in $\hat{\Theta}$. Let w_β be an n-vector with entries $w_{i,\beta} = w_\beta(y_i, x_i)$. We consider the matrix $\mathbf{X}_\beta := W_\beta \mathbf{X}$ where $W_\beta = \text{diag}(w_\beta)$. We define $\hat{\Sigma}_\beta := \mathbf{X}^T_\beta \mathbf{X}_\beta / n$. We consider $\hat{\Theta}_{\hat{\beta},j}$ as the jth row of the nodewise regression $\hat{\Theta} = \hat{\Theta}_{\hat{\beta}}$ in (29) based on the matrix input $\hat{\Sigma}_{\hat{\beta}}$.

We let $\Sigma_\beta = \mathbb{E}[\mathbf{X}^T_\beta \mathbf{X}_\beta / n]$ and define $\Theta := \Theta_{\beta^0} := \Sigma^{-1}_{\beta^0}$ (assumed to exist). Let $s_j := s_{\beta^0,j} := \|\Theta_{\beta^0,j}\|^0_0$. Analogous to Section 2.3.4, we let $\mathbf{X}_{\beta^0,-j}\gamma_{\beta^0,j}$ be the projection of $\mathbf{X}_{\beta^0,j}$ on $\mathbf{X}_{\beta^0,-j}$ using the inner products in the matrix Σ_{β^0} and let $\eta_{\beta^0,j} := \mathbf{X}_{\beta^0,j} - \mathbf{X}_{\beta^0,-j}\gamma_{\beta^0,j}$. We then make the following assumptions.

(D1) The pairs of random variables $\{(y_i, x_i)\}^n_{i=1}$ are i.i.d. and $\|\mathbf{X}\|_\infty = \max_{i,j} |\mathbf{X}_{ij}| = \mathcal{O}(K)$ and $\|\mathbf{X}\gamma^0_{\beta^0,j}\|_\infty = \mathcal{O}(K)$ for some $K \geq 1$.

(D2) It holds that $K^2 s_j \sqrt{\log(p)/n} = o(1)$.

(D3) The smallest eigenvalue of Σ_{β^0} is bounded away from zero and moreover $\|\Sigma_{\beta^0}\|_\infty = \mathcal{O}(1)$. (The latter is ensured by requiring that the largest eigenvalue is bounded from above).

(D4) For some $\delta > 0$ and all $\|\beta - \beta^0\|_1 \leq \delta$, it holds that w_β stays away from zero and that $\|w_\beta\|_\infty = \mathcal{O}(1)$. We further require that for all such β and all x and y

$$|w_{\hat{\beta}}(y,x) - w_{\beta^0}(y,x)| \leq |x(\hat{\beta} - \beta^0)|.$$

(D5) It holds that

$$\|\mathbf{X}(\hat{\beta} - \beta^0)\|_n = \mathcal{O}_{\mathbb{P}}(\lambda\sqrt{s_0}),\ \|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda s_0).$$

Note that (D5) typically holds when $\lambda\sqrt{s_0} = o(1)$ with $\lambda \asymp \sqrt{\log(p)/n}$ since the compatibility condition is then inherited from (D3). We have the following result.

Theorem 3.2. *Assume the conditions (D1)-(D5). Then, using $\lambda_j \asymp K\sqrt{\log(p)/n}$ for the nodewise Lasso $\hat{\Theta}_{\hat{\beta}}$:*

$$\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_1 = \mathcal{O}_{\mathbb{P}}\Big(K s_j \sqrt{\log(p)/n}\Big) + \mathcal{O}_{\mathbb{P}}\Big(K^2 s_0((\lambda^2/\sqrt{\log(p)/n}) \vee \lambda)\Big),$$

$$\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_2 = \mathcal{O}_{\mathbb{P}}\Big(K\sqrt{s_j \log(p)/n}\Big) + \mathcal{O}_{\mathbb{P}}\Big(K^2\sqrt{s_0}\lambda\Big),$$

and

$$|\hat{\tau}^2_{\hat{\beta},j} - \tau^2_{\beta^0,j}| = \mathcal{O}_{\mathbb{P}}\Big(K\sqrt{s_j \log(p)/n}\Big) + \mathcal{O}_{\mathbb{P}}\Big(K^2\sqrt{s_0}\lambda\Big).$$

Moreover,

$$|\hat{\Theta}^T_{\hat{\beta},j}\Sigma_{\beta^0}\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,jj}| \leq \|\Sigma_{\beta^0}\|_\infty \|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_1^2 \wedge \Lambda^2_{\max}\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_2^2 + 2|\hat{\tau}^2_{\hat{\beta},j} - \tau^2_{\beta^0,j}|$$

where $\Lambda^2_{\max}$ is the maximal eigenvalue of Σ_{β^0}.

A proof, using ideas for establishing Theorem 2.4, is given in Section 6.2.

Corollary 3.1. *Assume the conditions of Theorem 3.2, with* $\lambda \asymp \sqrt{\log(p)/n}$, $K \asymp 1$, $s_j = o(\sqrt{n}/\log(p))$ *and* $s_0 = o(\sqrt{n}/\log(p))$. *Then*

$$\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_1 = o_{\mathbb{P}}\Big(1/\sqrt{\log(p)}\Big),$$

$$\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_2 = o_{\mathbb{P}}(n^{-1/4}),$$

and

$$|\hat{\Theta}_{\hat{\beta},j}^T \Sigma_{\beta^0} \hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,jj}| = o_{\mathbb{P}}\Big(1/\log(p)\Big).$$

Lemma 3.1. *Assume the conditions of Corollary 3.1. Let for* $i = 1, \ldots, n$, ξ_i *be a real-valued random variable and* $x_i^T \in \mathbb{R}^p$, *and let* $(x_i, \xi_i)_{i=1}^n$ *be i.i.d. Assume* $\mathbb{E}x_i^T \xi_i = 0$ *and that* $|\xi_i| \le 1$. *Then*

$$\hat{\Theta}_{\hat{\beta},j}^T \sum_{i=1}^n x_i^T \xi_i / n = \Theta_{\beta^0,j}^T \sum_{i=1}^n x_i^T \xi_i / n + o_{\mathbb{P}}(n^{-1/2}).$$

Let $A := \mathbb{E}x_i^T x_i \xi_i^2$ *(assumed to exist). Assume that* $\|A\Theta_j^0\|_\infty = \mathcal{O}(1)$ *and that* $1/((\Theta_j^0)^T A \Theta_j^0) = \mathcal{O}(1)$. *Then*

$$\hat{\Theta}_{\hat{\beta},j}^T A \hat{\Theta}_{\hat{\beta},j} = \Theta_{\beta^0,j}^T A \Theta_{\beta^0,j} + o_{\mathbb{P}}(1).$$

Moreover, then

$$\frac{\hat{\Theta}_{\hat{\beta},j}^T \sum_{i=1}^n x_i^T \xi_i / \sqrt{n}}{\sqrt{\hat{\Theta}_{\hat{\beta},j}^T A \hat{\Theta}_{\hat{\beta},j}}}$$

convergences weakly to a $\mathcal{N}(0,1)$*-distribution.*

A proof is given in Section 6.2

3.3.1. *Discussion of the assumptions for GLMs.* Assumption (C1) is classical [9] and (C3) is easy to understand. All the other assumptions (C2)-(C8) follow essentially from the conditions of Corollary 3.1 with $\Sigma_\beta := P\ddot{\rho}_\beta$ and $w_\beta(y,x) := \ddot{\rho}(y, x\beta)$, provided we take $\hat{\Theta}$ as the nodewise Lasso in (29), and $s_* = s_j$ and $\lambda_* \asymp \lambda_j$. We will discuss this for the case $\|\mathbf{X}\beta^0\|_\infty = \mathcal{O}(1)$ (that is we assume $K = 1$ for simplicity) and $|\dot{\rho}(y - x\beta^0)| = \mathcal{O}(1)$ uniformly in x and y. We also need to assume $1/(\Theta_{\beta^0,j}^T P \dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \Theta_{\beta^0,j}) = \mathcal{O}(1)$. Note that for the case of canonical loss, $P\dot{\rho}_{\beta^0}\dot{\rho}_{\beta^0}^T = \Sigma_{\beta^0}$, and hence $1/(\Theta_{\beta^0,j}^T P \dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \Theta_{\beta^0,j}) = \tau_{\beta^0,j}^2$.

Condition (C2) holds because the compatibility condition is met as Σ_{β^0} is non-singular and

$$\|\hat{\Sigma} - \Sigma_{\beta^0}\|_\infty = \mathcal{O}_{\mathbb{P}}(\lambda_*).$$

The condition that $\dot{\rho}(y, x\beta^0)$ is bounded ensures that $\rho(y, a)$ is locally Lipschitz, so that we can control the empirical process $(P_n - P)(\rho_{\hat{\beta}} - \rho_{\beta^0})$ as in [33] (see also [5] or [32]). (In the case of a GLM with canonical loss (e.g. least squares loss) we can relax the condition of a locally bounded derivative because the empirical process is then linear). Condition (C3) is assumed to hold with $\|\mathbf{X}\|_\infty = \mathcal{O}(1)$, and Condition (C4) holds with $\lambda_* \asymp \sqrt{\log p/n}$. This is because in the node-wise regression construction, the $1/\hat{\tau}_j^2$ are consistent estimators of $(\Sigma_{\beta^0}^{-1})_{jj}$ (see Theorem 3.2). Condition (C5) holds as well. Indeed, $\|\Theta_{\beta^0,j}\|_1 = \mathcal{O}(\sqrt{s_j})$, and $\|\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j}\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda_j s_j) = \mathcal{O}_{\mathbb{P}}(\sqrt{s_j})$. Condition (C6) holds as well, since we assume that $|\dot{\rho}| = \mathcal{O}(1)$ as well as $\|\mathbf{X}\|_\infty = \mathcal{O}(1)$. As for Condition (C7), this follows from Lemma 3.1, since $|\Theta_{\beta^0,j}^T \dot{\rho}_{\beta^0}(y, x)| = |\Theta_{\beta^0,j}^T x^T \dot{\rho}(y, x\beta^0)| = \mathcal{O}(1)$, which implies for $A := P\dot{\rho}_{\beta^0}\dot{\rho}_{\beta^0}^T$ that $\|A\Theta_{\beta^0,j}\|_\infty = \mathcal{O}(1)$.

4. Conclusions.

We derive confidence regions and statistical tests for low-dimensional components of a large parameter in high-dimensional models. We propose a general principle which is based on "inverting" the KKT conditions from ℓ_1-penalized estimators. The method easily allows for multiple testing adjustment which takes the dependence structure into account.

For linear models, the procedure is (essentially) the same as the projection method in [37]: we prove its asymptotic optimality in terms of semiparametric efficiency, assuming certain sparsity conditions. For generalized linear models with convex loss functions, we develop a substantial body of new theory which in turn justifies the general "KKT inversion" principle.

The conditions we impose seem rather tight for our method. We require ℓ_0-sparsity of the underlying regression coefficient $s_0 = o(\sqrt{n}/\log(p))$ (assuming bounded design for GLMs). This is essentially the condition for ℓ_1-norm convergence, and our method requires such an ℓ_1-norm bound; the additional factor $1\sqrt{\log(p)}$ stems from the fact that the asymptotic pivot $\hat{b} - \beta^0$ is normalized with $\sqrt{n}$. Regarding the design, we assume that the minimal eigenvalue of the population covariance matrix Σ is bounded from below, and that the row-sparsity of $\Theta = \Sigma^{-1}$ satisfies $s_{\max} = o(n/\log(p))$ for Gaussian or sub-Gaussian design or $s_{\max} = o(\sqrt{n/\log(p)})$ for bounded design, respectively. More generally, our methods needs two elements: the ℓ_1-norm convergence of the Lasso estimator $\hat{\beta}$ and the ℓ_∞-norm of the jth row of $\hat{\Theta}\hat{\Sigma} - I$, see (4) or (30). However, since the analysis is conditional on $\mathbf{X}$, the latter is an observable quantity, and hence its bound in practice does not depend on the assumptions but is observed from data.

4.1. *Empirical results.*

We have done some empirical validation for two-sided testing of individual hypotheses $H_{0,j} : \beta_j^0 = 0$ and corresponding multiple testing adjustment. Due to the length of this paper, we don't include the results but rather give a short summary of the findings (and we intend to provide an R-package with corresponding illustrations).

We compared our de-sparsified estimator $\hat{b}_{\text{Lasso}}$ with a bias-corrected Ridge estimator which has been proposed in [4]. As an overall conclusion, we find that our $\hat{b}_{\text{Lasso}}$ estimator has more power than the Ridge-type method while still controlling type I error measures reasonably well, whereas the Ridge procedure yields conservative type I error control for a broader class of designs. We also considered the mean-squared error for estimating a single parameter β_j^0, and we compared $\hat{b}_{\text{Lasso}}$ with the standard Lasso and the bias-corrected Ridge estimator [4]. We found that $\hat{b}_{\text{Lasso}}$ is clearly better than the Ridge-type method. For the standard Lasso, we observe a "super-efficiency" phenomenon, namely that it estimates the zero coefficients often very accurately while estimation for the non-zero parameters is poor in comparison to $\hat{b}_{\text{Lasso}}$.

5. Proofs and materials needed.

5.1. *Bounds for $\|\hat{\beta} - \beta^0\|_1$ with fixed design.*

The following known result gives a bound for the ℓ_1-norm estimation accuracy.

Lemma 5.1. *Assume a linear model as in (1) with Gaussian error and fixed design $\mathbf{X}$ which satisfies the compatibility condition with compatibility constant ϕ_0^2 and with $\hat{\Sigma}_{jj} \le M^2 < \infty$ for all j. Consider the Lasso with regularization parameter $\lambda \ge 2M\sigma_\varepsilon\sqrt{\frac{t^2+2\log(p)}{n}}$. Then, with probability at least $1 - 2\exp(-t^2/2)$:*

$$\|\hat{\beta} - \beta^0\|_1 \le 8\lambda \frac{s_0}{\phi_0^2} \text{ and } \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n \le 8\lambda^2 \frac{s_0}{\phi_0^2}.$$

A proof follows directly from the arguments in [5, Th.6.1] which can be modified to treat the case with unequal values of $\hat{\Sigma}_{jj}$ for various j. □

5.2. *Proof of Theorem 2.1.* It is straightforward to see that

$$\|\Delta\|_\infty = \|(\hat{\Theta}_{\text{Lasso}}\hat{\Sigma} - I)(\hat{\beta} - \beta^0)\|_\infty \le \|(\hat{\Theta}_{\text{Lasso}}\hat{\Sigma} - I)\|_\infty \|\hat{\beta} - \beta^0\|_1, \tag{30}$$

where $\|A\|_\infty = \max_{j,k} |A_{j,k}|$ is the element-wise sup-norm for a matrix A.

For bounding $\|(\hat{\Theta}_{\text{Lasso}}\hat{\Sigma} - I)\|_\infty$ we invoke the KKT conditions for the Lasso for nodewise regression in (7):

$$\|\mathbf{X}_{-j}^T(X_j - \mathbf{X}_{-j}\hat{\gamma}_{\text{Lasso},j})\|_\infty \le \lambda_j \ (j = 1, \ldots, p),$$

and thus we obtain for $\hat{\Theta}_{\text{Lasso}}$ from (8):

$$\|(\hat{\Theta}_{\text{Lasso}}\hat{\Sigma} - I)\|_\infty \le \|\Lambda\hat{T}^{-2}\|_\infty. \tag{31}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$. The right-hand side of (31) can be bounded by

$$\|\Lambda\hat{T}^{-2}\|_\infty \le \lambda_{\max}\|\hat{T}^{-2}\|_\infty.$$

Therefore, using the latter bound in (30) and the bound from Lemma 5.1 completes the proof. □

5.3. *Random design: bounds for compatibility constant and $\|T^{-2}\|_\infty$.* The compatibility condition with constant ϕ_0^2 being bounded away from zero is ensured by a rather natural condition about sparsity. We have the following result.

Lemma 5.2. *Assume that P_X is Gaussian satisfying (A2). Furthermore, assume that $s_0 = o(n/\log(p))$ $(n \to \infty)$. Then, with probability tending to one, the compatibility condition holds with compatibility constant*

$$\phi_0 \ge L/2 > 0.$$

A proof follows directly as in [26, Th.1].

Lemma 5.1 and (5.2) say that we have a bound

$$\begin{aligned} &\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(s_0\sqrt{\log(p)/n}), \\ &\|\mathbf{X}(\hat{\beta} - \beta^0)\|_n^2 := \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = \mathcal{O}_{\mathbb{P}}(s_0\log(p)/n), \end{aligned} \tag{32}$$

when assuming (A2) for a Gaussian distribution P_X and sparsity $s_0 = o(n/\log(p))$.

When using the Lasso for nodewise regression in (8), we would like to have a bound for $\|\hat{T}_{\text{Lasso}}^{-2}\|_\infty$ appearing in Theorem 2.1.

LEMMA 5.3. *Assume (A2) with row-sparsity for $\Theta = \Sigma^{-1}$ bounded by*

$$s_j \le s_{\max} = o(n/\log(p)) \text{ for all } j = 1, \dots, p.$$

Then, when choosing the regularization parameters $\lambda_j \equiv \lambda_{\max} \asymp \sqrt{\log(p)/n}$,

$$\|\hat{T}_{\text{Lasso}}^{-2}\|_\infty = \mathcal{O}_{\mathbb{P}}(1) \ (n \to \infty).$$

A proof follows using standard arguments. The compatibility assumption holds for each nodewise regression with corresponding compatibility constant bounded from below by $L/2$, as in Lemma 5.2. Furthermore, the population error variance $\tau_j^2 = \mathbb{E}[(X_j - \sum_{k \ne j} \gamma_{j,k} X_k)^2]$, where $\gamma_{j,k}$ are the population regression coefficients of X_j versus $\{X_k;\ k \ne j\}$ satisfy: for all j, $\tau_j^2 = \frac{1}{\Theta_{jj}} \ge \Lambda_{\min}^2 \ge L > 0$ (see formula (35)) and $\tau_j^2 \le \mathbb{E}[\|X_j\|^2/n] = \Sigma_{jj} \le \|\Sigma\|_\infty = \mathcal{O}(1)$, thereby invoking assumption (A2). Thus, all the error variances behave nicely and therefore, each nodewise regression satisfies $\|X_j - \mathbf{X}_{-j}\hat{\gamma}_j\|_2^2/n = \mathcal{O}_{\mathbb{P}}(s_j \log(p)/n)$ (see Lemma 5.1 or (32) and hence the statement follows. □

5.4. *Bounds for $\|\hat{\beta} - \beta^0\|_2$ with random design.* As argued in Lemma 2, assuming $s_0 = s_0 = o(n/\log(p))$ $(n \to \infty)$, the compatibility condition holds with probability tending to one. Therefore, the weaker restricted eigenvalue condition [2] holds as well and assuming (A2) we have the bound (see [2]):

$$\|\hat{\beta} - \beta^0\|_2 = \mathcal{O}_{\mathbb{P}}(\sqrt{s_0 \log(p)/n}). \tag{33}$$

5.5. *Proof of Theorem 2.2.* Invoking Theorem 2.1 and Lemma 5.3 we have that

$$n^{1/2}\|\Delta\|_\infty \le \mathcal{O}_{\mathbb{P}}(s_0 \log(p) n^{-1/2}) = o_{\mathbb{P}}(1)$$

where the last bound follows by the sparsity assumption on s_0.

What remains to be shown is that $\|\Omega_n - \Theta\|_\infty = o_{\mathbb{P}}(1)$, as detailed by the following Lemma.

LEMMA 5.4. *Assume*

$$\max_j \|\hat{\Theta}_j - \Theta_j\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda_{\max} s_{\max}), \quad \max_j \|\hat{\Theta}_j - \Theta_j\|_2 = \mathcal{O}_{\mathbb{P}}(\lambda_{\max}\sqrt{s_{\max}}).$$

Suppose that $\lambda_{\max}^2 s_{\max} = o(1)$. Then

$$\|\Omega - \Theta\|_\infty = o_{\mathbb{P}}(1).$$

Proof: We first show that for $L > 0$ from (A2):

$$\max_{j=1,\dots,p} \|\Theta_j\|_2 \le L^{-2} < \infty. \tag{34}$$

Clearly, we have:

$$\Theta_{jj} \le \max_{\alpha} \frac{\alpha^T \Theta \alpha}{\|\alpha\|_2^2} = \Lambda_{\min}^{-2}. \tag{35}$$

Furthermore,

$$\|\Theta_j\|_2^2 \le \frac{\Theta_j^T \Sigma \Theta_j}{\Lambda_{\min}^2} = \frac{\Theta_{jj}}{\Lambda_{\min}^2} \le \Lambda_{\min}^{-4} \le L^{-2} < \infty,$$

where we used (35) and assumption (A2) in the two last inequalities. Therefore, (34) holds.

Using standard arguments, analogous to (32) and using Lemma 5.3, we have that

$$\max_j \|\hat{\Theta}_j - \Theta_j\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda_{\max} s_{\max})$$

Hence, uniformly in j:

$$\|\hat{\Theta}_j\|_1 = \|\Theta_j\|_1 + \mathcal{O}_{\mathbb{P}}(\lambda_{\max} s_{\max}) = \mathcal{O}_{\mathbb{P}}(\sqrt{s_{\max}}). \tag{36}$$

Furthermore, we have

$$\Omega = \hat{\Theta}\hat{\Sigma}\hat{\Theta}^T = (\hat{\Theta}\hat{\Sigma} - I)\hat{\Theta}^T + \hat{\Theta}^T \tag{37}$$

and

$$\begin{aligned} \|(\hat{\Theta}\hat{\Sigma} - I)\hat{\Theta}^T\|_\infty &\le \lambda_{\max} \|\hat{T}^{-2}\|_\infty \max_j \|\hat{\Theta}_j\|_1 \\ &= \mathcal{O}_{\mathbb{P}}(\lambda_{\max}\sqrt{s_{\max}}) = o_{\mathbb{P}}(1), \end{aligned} \tag{38}$$

where the second-last bound follows from Lemma 5.3 and (36). Finally, we have using standard arguments for the ℓ_2-norm bounds, see also (33):

$$\|\hat{\Theta} - \Theta\|_\infty \le \max_j \|\hat{\Theta}_j - \Theta_j\|_2 \le \lambda_{\max}\sqrt{s_{\max}} = o_{\mathbb{P}}(1). \tag{39}$$

Using (37)–(39) we complete the proof. □

The proof of Theorem 2.2 is now complete using the fact that the sparsity assumptions and (A2) automatically imply that the compatibility condition holds for every nodewise regression (see also Lemma 5.2), and also the restricted eigenvalue condition holds [2] which allows for bounding $\max_j \|\hat{\Theta}_j - \Theta_j\|_2$ as worked out in [26, Th.1]. From this we deduce that the conditions in Lemma 5.4 about $\max_j \|\hat{\Theta}_j - \Theta_j\|_q$ $(q = 1, 2)$ hold. □

5.6. *Proof of Theorem 2.4.* Under the sub-Gaussian assumption we know that η_j is also sub-Gaussian. So then $\|\eta_j^T X_{-j}/n\|_\infty = \mathcal{O}_{\mathbb{P}}(\sqrt{\log(p)/n})$. If $\|X\|_\infty = \mathcal{O}(K)$, we can use the work in [10] to conclude that

$$\|\eta_j^T \mathbf{X}_{-j}/n\|_\infty = \mathcal{O}_{\mathbb{P}}(K\sqrt{\log(p)/n}).$$

However, this result does not hold uniformly in j. Otherwise, in the strongly bounded case, we have

$$\|\eta_j\|_\infty \leq \|X_j\|_\infty + \|\mathbf{X}_{-j}\gamma_j^0\|_\infty = \mathcal{O}(K).$$

So then $\|\eta_j^T \mathbf{X}_{-j}/n\|_\infty = \mathcal{O}_{\mathbb{P}}(K\sqrt{\log(p)/n}) + \mathcal{O}_{\mathbb{P}}(K^2 \log(p)/n)$, which is uniform in j.

Then by standard arguments (see e.g. [2], and see [5] which complements the concentration results in [18] for the case of errors with only second moments) for $\lambda_j \asymp K_0\sqrt{\log(p)/n}$ (recall that $K_0 = 1$ in the sub-Gaussian case and $K_0 = K$ in the (strongly) bounded case)

$$\|X_j(\hat{\gamma}_j - \gamma_j^0)\|_n^2 = \mathcal{O}_{\mathbb{P}}(s_j\lambda_j),\ \|\hat{\gamma}_j - \gamma_j^0\|_1 = \mathcal{O}(s_j\lambda_j).$$

The condition $K^2 s_j\sqrt{\log(p)/n}$ is used in the (strongly) bounded case to be able to conclude that the empirical compatibility condition holds (see [5], Section 6.12). In the sub-Gaussian case, we use that $\sqrt{s_j \log(p)/n} = o(1)$ and an extension of Theorem 1 in [26] from the Gaussian case to the sub-Gaussian case. This gives again that the empirical compatibility condition holds.

We further find that

$$\|\hat{\gamma}_j - \gamma_j^0\|_2 = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{s_j \log(p)/n}).$$

To show this, we first introduce the notation $\beta^T \Sigma^0 \beta := \|X\beta\|^2$. Then in the (strongly) bounded case

$$\left| \|X\beta\|_n^2 - \|X\beta\|^2 \right| \leq \|\hat{\Sigma} - \Sigma^0\|_\infty \|\beta\|_1^2 = \mathcal{O}_{\mathbb{P}}(K^2\sqrt{\log(p)/n})\|\beta\|_1^2.$$

Since $\|\hat{\gamma}_j - \gamma_j^0\|_1 = \mathcal{O}_{\mathbb{P}}(K_0 s_j\sqrt{\log(p)/n})$ and the smallest eigenvalue Λ_{min}^2 of Σ stays away from zero, this gives

$$\begin{aligned} & \mathcal{O}_{\mathbb{P}}(K_0^2 s_j \log(p)/n) = \|X_j(\hat{\gamma}_j - \gamma_j^0)\|_n^2 \\ \geq\ & \Lambda_{\text{min}}\|\hat{\gamma}_j - \gamma_j^0\|_2^2 - \mathcal{O}_{\mathbb{P}}(K_0^4 s_j^2 (\log(p)/n)^{3/2}) \\ \geq\ & \Lambda_{\text{min}}\|\hat{\gamma}_j - \gamma_j^0\|_2^2 - o_{\mathbb{P}}(K_0^2 \log(p)/n), \end{aligned}$$

where we again used that $K_0^2 s_j \sqrt{\log(p)/n} = o(1)$. In the sub-Gaussian case, the result for the $\|\cdot\|_2$-estimation error follows by similar arguments invoking again a sub-Gaussian extension of Theorem 1 in [26].

We moreover have

$$\begin{aligned}|\hat{\tau}_j^2 - \tau_j^2| &= \underbrace{|\eta_j^T \eta_j/n - \tau_j^2|}_{I} + \underbrace{|\eta_j^T \mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)/n|}_{II} \\ &+ \underbrace{|\eta_j^T \mathbf{X}_{-j}\gamma_j^0/n|}_{III} + \underbrace{|(\gamma_j^0)^T \mathbf{X}_{-j}^T \mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)/n|}_{IV}.\end{aligned}$$

Now, since we assume fourth moments of the errors,

$$I = \mathcal{O}_{\mathbb{P}}(K_0^2 n^{-1/2}).$$

Moreover,

$$II = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{\log(p)/n})\|\hat{\gamma}_j - \gamma_j^0\|_1 = \mathcal{O}_{\mathbb{P}}(K_0^2 s_j \log(p)/n).$$

As for III, we have

$$III = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{\log(p)/n})\|\gamma_j^0\|_1 = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{s_j \log(p)/n})$$

since $\|\gamma_j^0\|_1 \leq \sqrt{s_j}\|\gamma_j^0\|_2 = \mathcal{O}(\sqrt{s_j})$. Finally by the KKT conditions

$$\|\mathbf{X}_{-j}^T \mathbf{X}_{-j}(\hat{\gamma}_j - \gamma_j^0)\|_\infty = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{\log(p)/n}),$$

and hence

$$IV = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{\log(p)/n})\|\gamma_j^0\|_1 = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{s_j \log(p)/n}).$$

So now we have shown that

$$|\hat{\tau}_j^2 - \tau_j^2| = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{s_j \log(p)/n}).$$

Since $1/\tau_j^2 = \mathcal{O}(1)$, this implies that also

$$1/\hat{\tau}_j^2 - 1/\tau_j^2 = \mathcal{O}_{\mathbb{P}}(K_0\sqrt{s_j \log(p)/n}).$$

We conclude that

$$\|\hat{\Theta}_j - \Theta_j^0\|_1 = \|\hat{C}_j/\hat{\tau}_j^2 - C_j^0/\tau_j^2\|_1 \leq \underbrace{\|\hat{\gamma}_j - \gamma_j^0\|_1/\hat{\tau}_j^2}_{i} + \underbrace{\|\gamma_j^0\|_1(1/\hat{\tau}_j^2 - 1/\tau_j^2)}_{ii},$$

where

$$i = \mathcal{O}_{\mathbb{P}}(K_0 s_j \sqrt{\log(p)/n})$$

since $\hat{\tau}_j^2$ is a consistent estimator of τ_j^2 and $1/\tau_j^2 = \mathcal{O}(1)$, and also

$$ii = \mathcal{O}_{\mathbb{P}}(K_0 s_j \sqrt{\log(p)/n}),$$

since $\|\gamma_j^0\|_1 = \mathcal{O}(\sqrt{s_j})$.

Recall that

$$\|\hat{\gamma}_j - \gamma_j^0\|_2 = \mathcal{O}_{\mathbb{P}}(K_0 \sqrt{s_j \log(p)/n}).$$

But then

$$\|\hat{\Theta}_j - \Theta_j^0\|_2 \leq \|\hat{\gamma}_j - \gamma_j^0\|_2/\hat{\tau}_j^2 + \|\gamma_j^0\|_2(1/\hat{\tau}_j^2 - 1/\tau_j^2)$$

$$= \mathcal{O}_{\mathbb{P}}(K_0 \sqrt{s_j \log(p)/n}).$$

For the last part, we write

$$\hat{\Theta}_j^T \Sigma^0 \hat{\Theta}_j - \Theta_{jj} = (\hat{\Theta}_j - \Theta_j^0)^T \Sigma^0 (\hat{\Theta}_j - \Theta_j^0) + 2(\Theta_j^0)^T \Sigma^0 (\hat{\Theta}_j - \Theta_j^0) + (\Theta_j^0)^T \Sigma^0 \Theta_j^0 - \Theta_{jj}$$

$$= (\hat{\Theta}_j - \Theta_j^0)^T \Sigma^0 (\hat{\Theta}_j - \Theta_j^0) + 2(1/\hat{\tau}_j^2 - 1/\tau_j^2),$$

since $(\Theta_j^0)^T \Sigma^0 = e_j^T$, $(\Theta_j^0)^T \Sigma^0 \Theta_j^0 = \Theta_{jj}$, $\hat{\Theta}_{jj} = 1/\hat{\tau}_j^2$, and $\Theta_{jj} = 1/\tau_j^2$. But

$$(\hat{\Theta}_j - \Theta_j^0)^T \Sigma^0 (\hat{\Theta}_j - \Theta_j^0) \leq \|\Sigma^0\|_\infty \|\hat{\Theta}_j - \Theta_j^0\|_1.$$

We may also use

$$(\hat{\Theta}_j - \Theta_j^0)^T \Sigma^0 (\hat{\Theta}_j - \Theta_j^0) \leq \Lambda_{\max}^2 \|\hat{\Theta}_j - \Theta_j^0\|_2^2.$$

□

5.7. *Proof of Theorem 3.1.* Note that

$$\dot{\rho}(y, x_i\hat{\beta}) = \dot{\rho}(y, x_i\beta^0) + \ddot{\rho}(y, \tilde{a}_i)x_i(\hat{\beta} - \beta^0),$$

where $\tilde{a}_i$ is a point intermediating $x_i\hat{\beta}$ and $x_i\beta^0$, so that $|\tilde{a}_i - x_i\hat{\beta}| \leq |x_i(\hat{\beta} - \beta^0)|$.

We find by the Lipschitz condition on $\ddot{\rho}$ (Condition)

$$|\ddot{\rho}(y, \tilde{a}_i)x_i(\hat{\beta} - \beta^0) - \ddot{\rho}(y, x_i\hat{\beta})x_i(\hat{\beta} - \beta^0)|$$

$$= |\tilde{a}_i - x_i\hat{\beta}||x_i(\hat{\beta} - \beta^0)||x_i(\hat{\beta} - \beta^0)|^2.$$

Thus, using that by Condition (C5) $|x_i\hat{\Theta}_j| = \mathcal{O}_{\mathbb{P}}(K)$ uniformly in i,

$$\hat{\Theta}_j^T P_n \dot{\rho}_{\hat{\beta}} = \hat{\Theta}_j^T P_n \dot{\rho}_{\beta^0} + \hat{\Theta}_j P_n \ddot{\rho}_{\hat{\beta}}(\hat{\beta} - \beta^0) + \mathrm{Rem}_1,$$

where

$$\mathrm{Rem}_1 = \mathcal{O}_{\mathbb{P}}(K)\sum_{i=1}^n |x_i(\hat{\beta} - \beta^0)|^2/n = \mathcal{O}(K)\|X(\hat{\beta} - \beta^0)\|_n^2$$
$$= \mathcal{O}_{\mathbb{P}}(Ks_0\lambda^2) = o_{\mathbb{P}}(1)$$

where we used Condition (C2) and in the last step Condition (C8).

We know that by Condition (C4)

$$\|\hat{\Theta}_j^T P_n \ddot{\rho}_{\hat{\beta}} - e_j^T\|_\infty = \mathcal{O}(\lambda_*).$$

It follows that

$$\begin{aligned} & b_j - \beta_j^0 = \hat{\beta}_j - \beta_j^0 - \hat{\Theta}_j^T P_n \dot{\rho}_{\hat{\beta}} \\ = \quad & \hat{\beta}_j - \beta_j^0 - \hat{\Theta}_j^T P_n \dot{\rho}_{\beta^0} - \hat{\Theta}_j^T P_n \ddot{\rho}_{\hat{\beta}}(\hat{\beta} - \beta^0) - \mathrm{Rem}_1 \\ = \quad & \hat{\Theta}_j^T P_n \dot{\rho}_{\beta^0} - (\hat{\Theta}_j^T P_n \ddot{\rho}_{\hat{\beta}} - e_j^T)(\hat{\beta} - \beta^0) - \mathrm{Rem}_1 = \hat{\Theta}_j^T P_n \dot{\rho}_{\beta^0} - \mathrm{Rem}_2, \end{aligned}$$

where

$$|\mathrm{Rem}_2| \le |\mathrm{Rem}_1| + \mathcal{O}(\lambda_*)\|\hat{\beta} - \beta^0\|_1 = o_{\mathbb{P}}(n^{-1/2}) + \mathcal{O}_{\mathbb{P}}(s_0\lambda\lambda_*) = o_{\mathbb{P}}(n^{-1/2})$$

since by Condition (C2) $\|\hat{\beta} - \beta^0\|_1 = \mathcal{O}_{\mathbb{P}}(\lambda s_0)$, and by the second part of Condition (C8) also $\lambda_*\lambda s_0 = o(n^{-1/2})$.

We now have to show that our estimator of the variance is consistent. We find

$$\begin{aligned} & |(\hat{\Theta} P \dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \hat{\Theta}^T)_{jj} - (\hat{\Theta} P_n \dot{\rho}_{\hat{\beta}} \dot{\rho}_{\hat{\beta}}^T \hat{\Theta}^T)_{jj}| \\ \le \quad & \underbrace{|(\hat{\Theta}(P_n - P)\dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \hat{\Theta}^T)_{jj}|}_{I} + \underbrace{|(\hat{\Theta} P \dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \hat{\Theta}^T)_{jj} - (\hat{\Theta} P \dot{\rho}_{\hat{\beta}} \dot{\rho}_{\hat{\beta}}^T \hat{\Theta}^T)_{jj}|}_{II}. \end{aligned}$$

But, writing $\varepsilon_{k,l} := (P_n - P)\dot{\rho}_{k,\beta^0}\dot{\rho}_{l,\beta^0}$, we see that

$$I = |\hat{\Theta}(P_n - P)\dot{\rho}_{\beta^0} \dot{\rho}_{\beta^0}^T \hat{\Theta}^T)_{jj}| = |\sum_{k,l} \hat{\Theta}_{j,k}\hat{\Theta}_{j,l}\varepsilon_{k,l}| \le \|\hat{\Theta}_j\|_1^2 \|\varepsilon\|_\infty = \mathcal{O}_{\mathbb{P}}(s_* K^2 \lambda),$$

where we used Conditions (C5) and (C6).

Next, we will handle II. We have

$$\dot\rho_{\hat\beta}(y,x)\dot\rho_{\hat\beta}^T(y,x) - \dot\rho_{\hat\beta}(y,x)\dot\rho_{\hat\beta}^T(y,x) = [\dot\rho^2(y-x\hat\beta) - \dot\rho^2(y-x\beta^0)]x^Tx := w(y,x)x^Tx,$$

with

$$|w(y,x)| := |\dot\rho^2(y-x\hat\beta) - \dot\rho^2(y-x\beta^0)| = \mathcal{O}_{\mathbb{P}}(1)|x(\hat\beta-\beta^0)|,$$

where we use that $\ddot\rho$ is locally bounded (Condition (C1)). It follows from Condition (C2) that

$$P|w| \le \sqrt{P|w|^2} = \mathcal{O}(\lambda\sqrt{s_0}).$$

Moreover by Condition (C5)

$$\|\hat\Theta_j^T x^T\|_\infty = \mathcal{O}_{\mathbb{P}}(K)$$

so that

$$|(\hat\Theta w(x,y)x^Tx\hat\Theta^T)_{jj}| \le \mathcal{O}(K^2)|w(y,x)|.$$

Thus

$$|(\hat\Theta P\dot\rho_{\beta^0}\dot\rho_{\beta^0}^T\hat\Theta^T)_{jj} - (\hat\Theta P\dot\rho_{\hat\beta}\dot\rho_{\hat\beta}^T\hat\Theta^T)_{jj}| = \mathcal{O}_{\mathbb{P}}(K^2\sqrt{s_0}\lambda).$$

It follows that

$$I + II = \mathcal{O}_{\mathbb{P}}(K^2 s_*\lambda) + \mathcal{O}_{\mathbb{P}}(K^2\sqrt{s_0}\lambda) = o_{\mathbb{P}}(1)$$

by the last part of Condition (C8). □

References.

[1] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection, 2012. arXiv:1201.0224v3.

[2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[3] P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34: 559–583, 2006.

[4] P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 2013. To appear.

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

[6] F. Bunea, A. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[7] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, 35:2313–2404, 2007.

[8] A. Chatterjee and S.N. Lahiri. Bootstrapping Lasso estimators. *Journal of the American Statistical Association*, 106:608–625, 2011.

[9] H. Cramer. *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[10] L. Dümbgen, S.A. van de Geer, M.C. Veraar, and J.A. Wellner. Nemirovski's inequalities revisited. *The American Mathematical Monthly*, 117:138–160, 2010.

[11] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B*, 70:849–911, 2008.

[12] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[13] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[14] E. Greenshtein and Y. Ritov. Persistence in high-dimensional predictor selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.

[15] A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory, 2013. arXiv:1301.4240v1.

[16] A. Juditsky, F. Kilinç-Karzan, A. Nemirovski, and Boris Polyak. Accuracy guarantees for ℓ_1 recovery of block-sparse signals. *Annals of Statistics*, 2013. To appear.

[17] K. Knight and W. Fu. Asymptotics of Lasso-type estimators. *The Annals of Statistics*, 28:1356–1378, 2000.

[18] J. Lederer and S. van de Geer. New concentration inequalities for suprema of empirical processes, 2011. arXiv:1111.3486.

[19] K.-C. Li. Honest confidence regions for nonparametric regression. *Annals of Statistics*, 17:1001–1008, 1989.

[20] L. Meier, S. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society Series B*, 70:53–71, 2008.

[21] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[22] N. Meinshausen and P. Bühlmann. Stability Selection (with discussion). *Journal of the Royal Statistical Society Series B*, 72:417–473, 2010.

[23] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

[24] N. Meinshausen, L. Meier, and P. Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104:1671–1681, 2009.

[25] S.N. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27:538–557, 2012.

[26] G. Raskutti, M.J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[27] R.D. Shah and R.J. Samworth. Variable selection with error control: another look at Stability Selection. *Journal of the Royal Statistical Society Series B*, 75:55–80, 2013.

[28] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012.

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

[30] S. van de Geer. The deterministic Lasso. In *JSM proceedings, 2007, 140*. American Statistical Association, 2007.

[31] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[32] S. van de Geer and P. Müller. Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science*, 27:469–480, 2012.

[33] S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614–645, 2008.

[34] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

[35] L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of Statistics*, 37:2178–2201, 2009.

[36] C.-H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

[37] C.-H. Zhang and S.S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data, 2011. arXiv:1110.2563v1.

[38] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

6. Supplemental Section.

6.1. *Proofs for results in Section 2.3.3.*

6.1.1. *Proof of Theorem 2.3.* Arguing as in the beginning of Section 2.3.3, we consider the submodel parameterized by β_1 only and in which the mean of Y is shifted by $\beta_1\big(X_1 - \mathbb{E}[X_1|Z]\big)$. In this submodel, the efficient estimator is asymptotically normal with variance $\sigma_\varepsilon^2/\mathrm{Var}\big(X_1 - \mathbb{E}[X_1|Z]\big)$. Now, Corollary 2.1, ensures that $\hat{b}_{\mathrm{Lasso},j}$ is asymptotically normal with mean $\sigma_\varepsilon^2 \Omega_{n,jj}$. However, by (14) and Corollary 2.1, $\Omega_{n,jj} \to \mathrm{Var}\big(X_1 - \mathbb{E}[X_1|Z]\big)$. Moreover, by Corollary 2.1 again, the convergence is uniform and hence regularity for one-dimensional submodels follows.

6.1.2. *Proof that model (15) satisfies (14) and conditions of Theorem 2.2.* We use $Z = (X_j)_{j=2}^\infty$ and $K(Z) = \sum_{j=2}^p \beta_j^0 X_j$. The parameter β^n equals

$$\beta^n = \alpha(p_n, \beta^0, P_X^0) = \mathrm{argmin}_\alpha \mathbb{E}[(Y - \sum_{j=1}^{p_n} \alpha_j X_j)^2].$$

Since $\beta^0 \in \mathcal{B}$ has finite support $S(\beta^0)$, there exists $n(\beta^0)$ such that

$$\beta_j^n = \alpha(p_n, \beta^0, P_X^0)_j = \beta_j^0 \ \forall j = 1, \ldots, p_n \ \forall n \geq n(\beta^0).$$

In fact, we can choose $n(\beta^0) = \min\{n; \{1, \ldots, p_n\} \supseteq S(\beta^0)\}$. This implies the first condition in (14), namely that $K(Z) - \sum_{j=1}^{p_n} \beta_j^n X_j = 0$ for all $n \geq n(\beta^0)$.

Using $\gamma_n = \text{argmin}_\kappa \mathbb{E}[(X_1 - \sum_{j=2}^{p_n} \kappa_j X_j)^2]$ and the fact that $\mathbb{E}[X_1|Z] = \sum_{j=2}^{\infty} \gamma_j^0 X_j$ where $\gamma^0 = \gamma(P_X^0)$ has finite sparsity, we can use exactly the same argument as above to conclude that the second condition in (14) holds, namely $\mathbb{E}[X_1|Z] - \sum_{j=1}^{p_n} \gamma_j^n X_j = 0$ for all n greater than some $n(\gamma^0)$. The last condition in (14) follows then as well.

Finally, since β^0 and also γ^0 have finite sparsity, the projected parameters β^n and γ^n have finite sparsity as well (because the projected are equal to the non-projected values for n sufficiently large). Hence, the conditions of Theorem 2.2 hold. □

6.2. *Proofs for results in Section 3.3.*

6.2.1. *Proof of Theorem 3.2.* We can write

$$X_{\beta^0,j} = \mathbf{X}_{\beta^0,-j}\gamma^0_{\beta^0,j} + \eta_{\beta^0,j}.$$

Hence

$$X_{\hat\beta,j} = \mathbf{X}_{\hat\beta,-j}\gamma^0_{\beta^0,j} + W_{\hat\beta}W_{\beta^0}^{-1}\eta_{\beta^0,j}.$$

By definition

$$\hat\gamma_{\hat\beta,j} = \arg\min_\gamma \Big\{ \|X_{\hat\beta,j} - \mathbf{X}_{\hat\beta,-j}\gamma\|_n^2 + \lambda_j\|\gamma\|_1 \Big\}.$$

This implies

$$\begin{aligned} & \|\mathbf{X}_{\hat\beta,-j}(\hat\gamma_{\hat\beta,j} - \gamma_{\beta^0,j})\|_n^2 + \lambda_j\|\hat\gamma_{\hat\beta,1}\|_1 \\ \le\ & 2(W_{\hat\beta}W_{\beta^0}^{-1}\eta_{\beta^0,j}, \mathbf{X}_{\hat\beta,-j}(\hat\gamma_{\hat\beta,j} - \gamma^0_{\beta^0,j}))_n + \lambda_j\|\gamma^0_{\beta^0,j}\|_1. \end{aligned}$$

But by the Cauchy-Schwarz inequality

$$\begin{aligned} & \Big| (W_{\hat\beta}W_{\beta^0}^{-1}\eta_{\beta^0,j}, \mathbf{X}_{\hat\beta,-j}(\hat\gamma_{\hat\beta,j} - \gamma^0_{\beta^0,j}))_n - (\eta_{\beta^0,j}, \mathbf{X}_{\beta^0,-j}(\hat\gamma_{\hat\beta,j} - \gamma^0_{\beta^0,j}))_n \Big| \\ \le\ & \|(W_{\hat\beta}^2 W_{\beta^0}^{-2} - I)\eta_{\beta^0,j}\|_n \|\mathbf{X}_{\beta^0,-j}(\hat\gamma_{\hat\beta,j} - \gamma^0_{\beta^0,j})\|_n. \end{aligned}$$

Since

$$\|\eta_{\beta^0,j}\|_\infty \le \|X_{\beta^0,j}\|_\infty + \|\mathbf{X}_{\beta^0,-j}\gamma^0_{\beta^0,j}\|_\infty = \mathcal{O}(K),$$

we get

$$\begin{aligned} \|(W_{\hat\beta}^2 W_{\beta^0}^{-2} - I)\eta_{\beta^0,j}\|_n^2 &= \frac{1}{n}\sum_{i=1}^n \left(\frac{\hat w_{i,\hat\beta}^2 - w_{i,\beta^0}^2}{w_{i,\beta^0}^2} \right)^2 \mathcal{O}(K^2) \\ &= \mathcal{O}(K^2)\|X(\hat\beta - \beta^0)\|_n^2 = \mathcal{O}_{\mathbb{P}}(K^2\lambda^2 s_0), \end{aligned}$$

where in the last step, we used the third (Lipschitz) part of the conditions on the weights, as well as the conditions on the rate of convergence of $\|\mathbf{X}(\hat{\beta} - \beta^0)\|_n^2$. Now, for arbitrary $\delta > 0$ we have

$$2ab \leq \delta a^2 + b^2/\delta.$$

Hence, we get for arbitrary $0 < \delta < 1$

$$\begin{aligned} & (1-\delta)\|\mathbf{X}_{\hat{\beta},-j}(\hat{\gamma}_{\hat{\beta},j} - \gamma_{\beta^0,j})\|_n^2 + \lambda_j \|\hat{\gamma}_{\hat{\beta},1}\|_1 \\ \leq \quad & 2(\eta_{\beta^0,j}, \mathbf{X}_{\beta^0,-j}(\hat{\gamma}_{\hat{\beta},j} - \gamma^0_{\beta^0,j}))_n + \lambda_j \|\gamma^0_{\beta^0,j}\|_1 + \mathcal{O}_{\mathbb{P}}(K^2 \lambda^2 s_0). \end{aligned}$$

Here, we invoked that $\|\mathbf{X}_{\beta^0}(\hat{\beta} - \beta^0)\|_n^2 = \mathcal{O}_{\mathbb{P}} \|\mathbf{X}_{\hat{\beta}}(\hat{\beta} - \beta^0)\|_n$ since the weights stay away from zero and infinity.

This implies by the same arguments as in Theorem 2.4

$$\|\hat{\gamma}_{\hat{\beta},j} - \gamma_{\beta^0,j}\|_1 = \mathcal{O}_{\mathbb{P}}(K^2 s_j \sqrt{\log(p)/n}) + \mathcal{O}_{\mathbb{P}}(K^2 \lambda^2 s_0/\lambda_j)$$

and

$$\|\hat{\gamma}_{\hat{\beta},j} - \gamma_{\beta^0,j}\|_2 = \mathcal{O}_{\mathbb{P}}(K \sqrt{s_j \log(p)/n}) + \mathcal{O}_{\mathbb{P}}(K \lambda \sqrt{s_0}).$$

Indeed, it is easy to see that it the compatibility condition holds with Σ_{β^0} since it is non-singular with smallest eigenvalue staying away from zero. Since $K^2 s_j \sqrt{\log(p)/n} = o(1)$, the compatibility also holds for $\hat{\Sigma}_{\beta^0}$ with a slightly smaller compatibility constant. But then it also holds for $\hat{\Sigma}_{\hat{\beta}}$ because the weights stay away from zero and infinity. This argument can then used as well to obtain the rate in ℓ_2, as in Theorem 2.4. Next, by definition

$$\hat{\tau}^2_{\hat{\beta},j} := X^T_{\hat{\beta},j}(X_{\hat{\beta},j} - \mathbf{X}_{\hat{\beta},-j}\hat{\gamma}_{\hat{\beta},j})/n.$$

Insert

$$X_{\hat{\beta},j} = W_{\hat{\beta}} W^{-1}_{\beta^0}(\mathbf{X}_{\beta^0,-j}\gamma^0_{\beta^0,j} + \eta_{\beta^0,j})$$

and

$$(X_{\hat{\beta},j} - \mathbf{X}_{\hat{\beta},-j}\hat{\gamma}_{\hat{\beta},j}) = W_{\hat{\beta}} W^{-1}_{\beta^0}(\eta_{\beta^0,j} + X_{\beta^0,j}(\hat{\gamma}_{\hat{\beta},j} - \gamma^0_{\beta^0,j})).$$

We then get

$$\hat{\tau}^2_{\hat{\beta},j} - \tau^2_{\beta^0,j} = (i) + (ii),$$

where

$$\begin{aligned} (i) &:= X^T_{\beta^0,j}(X_{\beta^0,j} - \mathbf{X}_{\beta^0,-j}\hat{\gamma}_{\hat{\beta},j}) - \tau^2_{\beta^0,j}/n, \\ (ii) &:= X^T_{\beta^0,j}\Big(W^2_{\hat{\beta}} W^{-2}_{\beta^0} - I\Big)(X_{\beta^0,j} - \mathbf{X}_{\beta^0,-j}\hat{\gamma}_{\hat{\beta},j})/n. \end{aligned}$$

We can treat (i) in the same way as in Theorem 2.4 to find

$$(i) = \mathcal{O}_{\mathbb{P}}(K\sqrt{s_j \log(p)/n}).$$

As for (ii), since $\|X_{\beta^0,j}\|_\infty = \mathcal{O}(K)$ and

$$\|\mathbf{X}_{\beta^0,-j}\hat{\gamma}_{\hat{\beta},j}\|_\infty \leq \|\mathbf{X}_{\beta^0,-j}\gamma^0_{\beta^0,j}\|_\infty + \mathcal{O}(K)\|\hat{\gamma}_{\hat{\beta},j} - \gamma^0_{\beta^0,j}\|_1 = \mathcal{O}_{\mathbb{P}}(K)$$

we get

$$(ii) = \mathcal{O}_{\mathbb{P}}(K^2)\sum_{i=1}^n \left| \frac{w^2_{i,\hat{\beta}} - w^2_{i,\beta^0}}{w^2_{i,\beta^0}} \right| = \mathcal{O}_{\mathbb{P}}(K^2\sqrt{s_0}\lambda).$$

So we arrive at

$$|\hat{\tau}^2_{\hat{\beta},j} - \tau^2_{\beta^0,j}| = \mathcal{O}_{\mathbb{P}}(K\sqrt{s_j \log(p)/n}) + \mathcal{O}_{\mathbb{P}}(K^2\sqrt{s_0}\lambda).$$

The rest of the proof goes along the lines of the proof of Theorem 2.4. □

The last result of Theorem 3.2 is actually a direct corollary of the following simple lemma.

Lemma 6.1. *Let A be a symmetric $(p \times p)$-matrix with largest eigenvalue Λ^2_A and $\hat{v}$ and $v \in \mathbb{R}^p$. Then*

$$\begin{aligned} |\hat{v}^T A\hat{v} - v^T Av| &\leq \Big(\|A\|_\infty \|\hat{v} - v\|_1^2\Big) \wedge \Big(\Lambda^2_A\|\hat{v} - v\|_2^2\Big) \\ + \quad & 2\Big(\|Av\|_\infty\|\hat{v} - v\|_1\Big) \wedge \Big(\|Av\|_2\|\hat{v} - v\|_2\Big). \end{aligned}$$

Proof. It is clear that

$$\hat{v}^T A\hat{v} - v^T Av = (\hat{v} - v)^T A(\hat{v} - v) + 2v^T A(\hat{v} - v).$$

The result follows therefore from

$$\begin{aligned} &|(\hat{v} - v)^T A(\hat{v} - v)| \leq \Big(\|A\|_\infty\|\hat{v} - v\|_1^2\Big) \wedge \Big(\Lambda^2_A\|\hat{v} - v\|_2^2\Big), \\ &|v^T A(\hat{v} - v)| \leq \|Av\|_\infty\|\hat{v} - v\|_1 \end{aligned}$$

and

$$|v^T A(\hat{v} - v)| \leq \|Av\|_2\|\hat{v} - v\|_2.$$

□

6.2.2. *Proof of Lemma 3.1.* It holds that

$$|(\hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j})^T \sum_{i=1}^n x_i^T \xi_i / n| \leq \| \sum_{i=1}^n \xi_i x_i \|_\infty \| \hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j} \|_1 / n$$

$$= \mathcal{O}_{\mathbb{P}}(\sqrt{\log(p)/n}) \| \hat{\Theta}_{\hat{\beta},j} - \Theta_{\beta^0,j} \|_1 = o_{\mathbb{P}}(n^{-1/2})$$

by Corollary 3.1. For the second result, we use Lemma 6.1. We get

$$|\hat{\Theta}_{\hat{\beta},j}^T A \hat{\Theta}_j - (\Theta_{\beta^0,j})^T A \Theta_{\beta^0,j}| \leq \|A\|_\infty o_{\mathbb{P}}(1/\log(p)) + \|A\Theta_{\beta^0,j}\|_\infty o_{\mathbb{P}}(1/\sqrt{\log(p)}) = o_{\mathbb{P}}(1).$$

We thus have that

$$\frac{\hat{\Theta}_{\hat{\beta},j}^T \sum_{i=1}^n x_i^T \xi_i / \sqrt{n}}{\sqrt{\hat{\Theta}_{\hat{\beta},j}^T A \hat{\Theta}_{\hat{\beta},j}}} = \frac{(\Theta_{\beta^0,j})^T \sum_{i=1}^n x_i^T \xi_i / \sqrt{n}}{\sqrt{(\Theta_{\beta^0,j})^T A \Theta_j^0}} + o_{\mathbb{P}}(1).$$

The random variables $(\Theta_{\beta^0,j})^T x_i^T \xi_i$ are bounded, since $|\xi_i| \leq 1$ and $|x_i \Theta_{\beta^0,j}| = \mathcal{O}(1)$, and the $x_i^T \xi_i$'s are i.i.d.: thus, the Lindeberg condition is fulfilled and the asymptotic normality follows from this. □

Seminar für Statistik, ETH Zürich
E-mail: geer@stat.math.ethz.ch,
Seminar für Statistik, ETH Zürich
E-mail: buhlmann@stat.math.ethz.ch,
The Hebrew University of Jerusalem
E-mail: yaacov.ritov@gmail.com