The Annals of Statistics
2018, Vol. 46, No. 5, 2336–2359
https://doi.org/10.1214/17-AOS1622

SEMIPARAMETRIC EFFICIENCY BOUNDS FOR HIGH-DIMENSIONAL MODELS

BY JANA JANKOVÁ AND SARA VAN DE GEER

ETH Zürich

Asymptotic lower bounds for estimation play a fundamental role in assessing the quality of statistical procedures. In this paper, we propose a framework for obtaining semiparametric efficiency bounds for sparse high-dimensional models, where the dimension of the parameter is larger than the sample size. We adopt a semiparametric point of view: we concentrate on one-dimensional functions of a high-dimensional parameter. We follow two different approaches to reach the lower bounds: asymptotic Cramér–Rao bounds and Le Cam's type of analysis. Both of these approaches allow us to define a class of asymptotically unbiased or "regular" estimators for which a lower bound is derived. Consequently, we show that certain estimators obtained by de-sparsifying (or de-biasing) an ℓ_1-penalized M-estimator are asymptotically unbiased and achieve the lower bound on the variance: thus in this sense they are asymptotically efficient. The paper discusses in detail the linear regression model and the Gaussian graphical model.

1. Introduction. Following the development of numerous methods for high-dimensional estimation, more recently the need for statistical inference has emerged. A number of papers have since studied the problem and proposed constructions of estimators which are asymptotically normally distributed, and hence lead to inference. These results naturally give rise to the question of their optimality. This motivates us to study the question whether we can establish asymptotic efficiency bounds in high-dimensional models and whether we can construct an estimator achieving these bounds.

To introduce the setting, suppose that we observe a sample $X^{(1)}, \ldots, X^{(n)}$ which is distributed according to a probability distribution P_β that depends on an unknown high-dimensional parameter $\beta \in \mathcal{B} \subset \mathbb{R}^p$. The dimension p of the parameter can be much larger than the sample size n. A major structural assumption we consider in this paper is sparsity in the high-dimensional parameter. In these sparse high-dimensional settings, a common approach to estimation is based on regularized M-estimators, where the regularization is in terms of the ℓ_1-penalty. This approach has been studied extensively, and under several settings, it produces near-oracle estimators of β under certain sparsity conditions (and some further

Received June 2016; revised August 2017.
MSC2010 subject classifications. Primary 62J07; secondary 62F12.
Key words and phrases. Asymptotic efficiency, high-dimensional, sparsity, Lasso, linear regression, graphical models, Cramér–Rao bound, Le Cam's lemma.

conditions). However, the oracle properties of the regularized estimators come at a price: the regularization introduces bias by shrinking the estimated coefficients towards zero. Hence, the regularized approach does not easily yield estimators which are asymptotically normally distributed. This makes it difficult to establish results for statistical inference.

Several streams of work have emerged that studied "post-regularization inference", which focused on construction of methodology for inference, with some preliminary use of regularized estimators. This was mostly considered for estimation of low-dimensional parameters of the high-dimensional vector. One stream of work concentrates on "de-sparsifying" or "de-biasing" procedures, which were studied for the linear model [Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014a, 2014b, 2015)], for generalized linear models [van de Geer et al. (2014)] and some special cases of nonlinear models, such as undirected graphical models [Janková and van de Geer (2014, 2017)]; we also refer here to the book van de Geer (2016). This approach uses the ℓ_1-regularized M-estimator as an initial estimator and implements a bias correction step which may be interpreted as one iteration using the Newton–Raphson method. Another stream of work studies the use of orthogonalizing conditions to define a new post-regularization estimator; this approach was considered for general models under high-level conditions in Chernozhukov, Hansen and Spindler (2015). Further examples of high-dimensional inference include the works Ren et al. (2015), Gao, Ma and Zhou (2017) or data splitting methods [Meinshausen and Yu (2009)]. The work in essence shows an important result: an asymptotically normal estimator for low-dimensional parameters can be constructed in several of the common models.

Further key questions that were studied concern optimality properties of these de-sparsified estimators. In particular, what are lower bounds on the rate of convergence in the supremum norm? These questions have been investigated for the linear regression with random design [Cai and Guo (2017)] and for Gaussian graphical models [Ren et al. (2015)] and other special cases of nonlinear models [Gao, Ma and Zhou (2017)]. The results in these settings reveal several important findings, which we discuss for the linear regression and graphical models. The minimax rates for estimation of single elements (of the vector of regression coefficients or the precision matrix) are shown to satisfy

$$\inf_{T} \sup_{\beta \in \mathcal{B}} \mathbb{E}_\beta \big| T\big(X^{(1)}, \ldots, X^{(n)}\big) - \beta_i \big| \geq C(1/\sqrt{n} + s \log p/n), \tag{1}$$

for some constant $C > 0$, where $\beta_i \in \mathbb{R}$ is a single regression coefficient or a single entry in a precision matrix and the *unknown* sparsity s is the number of nonzero entries in the regression vector or, in the case of Gaussian graphical models, in rows of a precision matrix. The infimum in (1) is taken over all estimators T. The statement (1) further requires some mild regularity conditions [see Cai and Guo (2017), Ren et al. (2015)]. Naturally, (1) implies that the parametric rate is optimal: it cannot be improved in order. On the other hand, if there is insufficient sparsity,

in particular when the sparsity s satisfies $s \gg n/\log p$, the minimax lower bounds diverge. This is no surprise as the oracle inequalities for certain M-estimators have only been shown under the condition $s = o(n/\log p)$. In the intermediate sparsity regime when $\sqrt{n}/\log p \leq s < n/\log p$, the parametric rate cannot be achieved.

As for the upper bounds, the parametric rate $1/\sqrt{n}$ can be achieved for estimation of single entries. This basically follows directly from the asymptotic normality of the de-sparsified estimators, if sparsity of β is of small order $\sqrt{n}/\log p$. This sparsity condition is stronger than the condition necessary for oracle inequalities [$s = o(n/\log p)$]. However, as we discuss in Section 8.6, the sparsity condition $s = o(\sqrt{n}/\log p)$ is essentially necessary for asymptotically normal estimation. To summarize the findings, the analysis of the minimax rates revealed that under sufficient sparsity of small order $\sqrt{n}/\log p$, the parametric rate of order $1/\sqrt{n}$ is optimal, and the de-sparsified estimator achieves it (in the above mentioned cases).

In this paper, we attempt to answer further questions that arise concerning the optimality of asymptotically normal estimators in high-dimensional settings. The analysis on minimax rates does not address an important question. The derived lower bound (1) does not reveal any explicit lower bounds on the (asymptotic) variance. The question of efficiency in the spirit of the famous Cramér–Rao result thus remains open in the high-dimensional setting. This motivates us to pose the following questions. Can we establish lower bounds on the variance, similar to the Cramér–Rao bounds in the (semi)parametric setting, also in the high-dimensional setting? And if yes, can we construct an estimator that achieves these bounds? We give an affirmative answer to these questions.

2. Our contributions. Asymptotic efficiency of estimators was thoroughly studied in the traditional settings; we refer the reader to the books van der Vaart (1998), Bickel et al. (1993) and the references therein. These results are however developed for *fixed* models which do not change with n, and hence they cannot be applied to high-dimensional settings where the dimension of the parameter may grow with the sample size.

In this paper, we develop a framework for establishing asymptotic efficiency of estimators in high-dimensional models changing with n. We concentrate on two approaches towards deriving the lower efficiency bounds: asymptotic Cramér–Rao bounds and Le Cam's approach.

First, we develop an asymptotic version of a semiparametric Cramér–Rao lower bound for sparse high-dimensional linear and graphical models. To this end, we propose a strong asymptotic unbiasedness assumption. Loosely speaking, this unbiasedness assumption measures the rate at which the bias vanishes in shrinking neighbourhoods of the true distribution of "size" $1/\sqrt{n}$. We consider the linear model and the Gaussian graphical model and for each of them, we establish lower bounds on the variance of any asymptotically unbiased estimator. The proposed framework might be applicable to other high-dimensional models in a similar spirit.

Consequently, for linear regression and Gaussian graphical models, we show that the de-sparsified estimator is an asymptotically unbiased estimator and is asymptotically efficient, that is, it reaches the derived lower bound. Thus, compared to previous results, which only showed asymptotic normality or minimaxity (up to order in n) of the de-sparsified estimator, we show that it is in terms of variance the best among all asymptotically unbiased estimators: thus in this sense asymptotically efficient.

In the second approach, we extend some of the classical results of Le Cam on local asymptotic normality to the high-dimensional setting. The result underlies a likelihood expansion analysis and involves a careful adjustment of Le Cam's arguments to the high-dimensional setting. The result obtained gives us the limiting distribution of an asymptotically linear estimator under a small perturbation of the parameter. We next show for the linear model that the de-sparsified estimator is regular: it converges locally uniformly to the limiting normal distribution with zero mean, and among all regular estimators it has the smallest asymptotic variance.

The two approaches above are strongly related, but one does not clearly dominate the other. A more detailed comparison is discussed in Section 11.

As a by-product of our analysis, we establish new oracle results for the Lasso. Typical analysis considers oracle inequalities for the prediction error and the ℓ_1-error which hold with high-probability. We strengthen these oracle inequalities by showing that they also hold for the mean ℓ_1-error and for higher orders of this error. These oracle inequalities are needed to claim strong asymptotic unbiasedness of the de-sparsified estimators.

3. Relation to prior work. As pointed out in Section 2, the traditional results as in, for instance, van der Vaart (1998) or Bickel et al. (1993), are not directly applicable to the high-dimensional setting. We extend the traditional approach to semiparametric efficiency to the context of high-dimensional models which requires adjustment of the arguments to a model changing with n and the sparsity of the model is required to keep remainders in approximate expansions under control. Our main results show that the lower bounds for high-dimensional models are analogous to those for parametric models, however, a new message for high-dimensional models is that to obtain the parametric lower bound, we require that the "worst possible sub-direction" is sparse. Without this condition, we are unable to claim asymptotic efficiency of the de-sparsified Lasso estimator.

Regarding the upper bounds, to construct asymptotically efficient estimators, our work follows the methodology from the works van de Geer et al. (2014) and Janková and van de Geer (2017), where de-sparsified Lasso estimators are proposed for the linear regression and for undirected graphical models. We borrow these constructions with some small adjustments. However, the upper bounds derived for the de-sparsified estimators in the mentioned papers are not sufficient for the present analysis: we need to show a stronger oracle bound which holds in expectation. Moreover, we extend the results for estimation of single entries as

considered in van de Geer et al. (2014) and Janková and van de Geer (2017) to linear functionals.

Asymptotic efficiency of estimators in high-dimensional settings changing with n was first considered in the paper van de Geer et al. (2014). The paper provides a formulation of asymptotic efficiency of entries of the de-biased Lasso. The approach is based on embedding the high-dimensional model into a fixed (i.e., not changing with n) infinite-dimensional model, for which semiparametric efficiency bounds are available [see van der Vaart (1998)]. However, such an embedding requires a very special model structure. In the present paper, we do not use an embedding but instead directly develop the theory for models changing with n.

4. Organization of the paper. The particular sections of the paper are divided as follows. In Section 7, we state preliminary results on oracle inequalities for the mean ℓ_1-error of the Lasso estimator. In Section 6, we propose a strong asymptotic unbiasedness assumption. Section 8 gives lower and upper bounds on the variance of asymptotically unbiased estimators in the linear model, considering random design in Section 8.3 and fixed design in Section 8.4. In Section 9, we derive lower and upper bounds on the variance of asymptotically unbiased estimators in Gaussian graphical models. Section 10 contains an extension of Le Cam's lemma to the high-dimensional setting, which is applicable to general nonlinear models. Section 11 summarizes the results, conclusions and some open questions. Finally, the proofs are contained in the Supplementary Material [Janková and van de Geer (2018)].

5. Notation. For a vector $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, we denote its ℓ_p norm by $\|x\|_p := (\sum_{i=1}^p x_i^p)^{1/p}$ for $p \geq 1$. We further let $\|x\|_\infty := \max_{i=1,\ldots,p} |x_i|$ and $\|x\|_0 = |\{i : i \in \{1, \ldots, p\}, x_i \neq 0\}|$. For a vector $x \in \mathbb{R}^n$, we denote $\|x\|_n^2 := \|x\|_2^2/n$ (with some abuse of notation). By e_i, we denote a p-dimensional vector of zeros with a one at position i. For a matrix $A \in \mathbb{R}^{m \times n}$, we denote its (i, j)th entry by A_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Further, we let $\|A\|_\infty := \max_{i=1,\ldots,m, j=1,\ldots,n} |A_{ij}|$, $|||A_1||| := \max_{i=1,\ldots,m} \sum_{j=1}^n |A_{ij}|$ and we let $\|A\|_F$ denote the Frobenius norm of A. We denote its jth column by A_j. By $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ we denote the minimum and maximum eigenvalue of a symmetric matrix A, respectively. We use $\operatorname{tr}(A)$ to denote the trace of the matrix A. We recall here that for symmetric matrices $A, B \in \mathbb{R}^{p \times p}$ it holds that $\operatorname{vec}(A)^T \operatorname{vec}(B) = \operatorname{tr}(AB)$, where $\operatorname{vec}(A)$ is the vectorized version of a matrix A obtained by stacking columns of A on each other.

For real sequences f_n, g_n, we write $f_n = \mathcal{O}(g_n)$ or $f_n \lesssim g_n$ if $|f_n| \leq C|g_n|$ for some $C > 0$ independent of n for all n. We write $f_n \asymp g_n$ if both $f_n = \mathcal{O}(g_n)$ and $1/f_n = \mathcal{O}(1/g_n)$ hold. Finally, $f_n = o(g_n)$ if $\lim_{n\to\infty} f_n/g_n = 0$. For a sequence of random variables X_n, we write $X_n = \mathcal{O}_P(f_n)$ if X_n/f_n is bounded in probability. We write $X_n = o_P(1)$ if X_n converges to zero in probability. We use $\rightsquigarrow$ to denote the convergence in distribution. By 1_T, we denote the indicator function of the set T. The identity matrix is denoted by I.

6. Asymptotic unbiasedness. This section defines the concept of strong asymptotic unbiasedness that will be needed for the linear and graphical model. We turn to the linear model in the next section. Consider a probability distribution P_β on some observation space $\mathcal{X}$, where the parameter β lies is a p-dimensional parameter space $\mathcal{B} \subset \mathbb{R}^p$. We consider the parameter set

$$\mathcal{B}(d_n) := \{\beta \in \mathcal{B} : \|\beta\|_0 \leq d_n, \|\beta\|_2 \leq C\}, \tag{2}$$

where $C > 0$ is some universal constant and d_n is a *known* sequence that will be specified later. We further define an ℓ_2-neighbourhood of a point $\beta \in \mathcal{B}(d_n)$ as follows:

$$B(\beta, \varepsilon) := \{\tilde{\beta} \in \mathcal{B}(d_n) : \|\tilde{\beta} - \beta\|_2 \leq \varepsilon\}. \tag{3}$$

We remark that all the parameter vectors appearing in this paper are *sequences* depending on n. In general, we omit the index n, except for situations where omitting the index could lead to confusion.

Let $g : \mathcal{B} \to \mathbb{R}$ and let the parameter of interest be $g(\beta)$. Our goal is to derive an asymptotic lower bound for the variance of an estimator T_n of $g(\beta)$, which is in some sense asymptotically unbiased. To this end, we define *strong asymptotic unbiasedness* as follows.

DEFINITION 1. Let m_n be a sequence such that $n = o(m_n)$. We say that T_n is a strongly asymptotically unbiased estimator of $g(\beta)$ at β_0 (in a neighbourhood of size c) with a rate m_n if it holds that $\operatorname{var}_{\beta_0}(T_n) = \mathcal{O}(1/n)$ and for every $\beta \in B(\beta_0, \frac{c}{\sqrt{m_n}})$ it holds

$$\lim_{n\to\infty} \sqrt{m_n}\big(\mathbb{E}_\beta T_n - g(\beta)\big) = 0.$$

REMARK 1. The reason for requiring unbiasedness to hold in a local neighbourhood $\beta \in B(\beta_0, \frac{c}{\sqrt{m_n}})$ in Definition 1 is that it leads to a broader class of estimators. If β was allowed to be in a neighbourhood without the sparsity constraint, a "strongly asymptotically unbiased" estimator might not exist.

7. Strong oracle inequalities for the Lasso. We present new results on oracle inequalities for the Lasso estimator in linear regression which will be needed in subsequent sections, but can also be of independent interest. Typical high-dimensional analysis derives oracle inequalities for the Lasso which hold with high probability [see Bühlmann and van de Geer (2011) for an overview of such results]. The paper Bellec and Tsybakov (2016) derives bounds on the expectation of the prediction error. Here, we derive oracle inequalities for the ℓ_1-estimation error that hold in expectation.

Consider the linear model

$$Y = X\beta_0 + \epsilon, \tag{4}$$

where X is the $n \times p$ design matrix with independent rows $X^{(i)}$, $i = 1, \ldots, n$, Y is the $n \times 1$ vector of observations and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$ is the (unobservable) error. The error satisfies $\mathbb{E}\epsilon = 0$ and its components ϵ_i are independent for $i = 1, \ldots, n$. Moreover, the error ϵ and the design matrix X are independent. We further denote the Gram matrix by $\hat{\Sigma} := X^T X/n$. The vector $\beta_0 = (\beta_1^0, \ldots, \beta_p^0) \in \mathbb{R}^p$ is unknown. The unknown number of nonzero entries of β_0 is denoted by $s := \|\beta_0\|_0$ and is called the sparsity of β_0.

The Lasso estimator with a tuning parameter $\lambda > 0$ is defined as follows:

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_n^2 + 2\lambda\|\beta\|_1. \tag{5}$$

The known results on oracle inequalities for the Lasso (5) give high-probability bounds for the prediction error and the ℓ_1-error (or under some conditions, for the ℓ_q-error for $1 \le q \le 2$). In particular, for the tuning parameter $\lambda \asymp \sqrt{\log p/n}$ and under further conditions that may be found in Bühlmann and van de Geer (2011), it holds

$$\|X(\hat{\beta} - \beta_0)\|_n^2 + \lambda\|\hat{\beta} - \beta_0\|_1 = \mathcal{O}_P(s\lambda^2).$$

Bellec and Tsybakov (2016) show analogous results for the expected prediction error $\mathbb{E}\|X(\hat{\beta} - \beta_0)\|_n$ for the case of fixed design. We show such results may be obtained for the expected ℓ_1-error, under almost identical conditions. In particular, Theorem 1 presented below implies that the mean ℓ_1-error, $\mathbb{E}_{\beta_0}\|\hat{\beta} - \beta_0\|_1$, is up to a logarithmic factor of the same order as the oracle error $\mathbb{E}_{\beta_0}\|\beta_{\text{ora}} - \beta_0\|_1 = \mathcal{O}(s/\sqrt{n})$, where β_{ora} is the oracle maximum likelihood estimator (i.e., a maximum likelihood estimator applied with the knowledge of true nonzero entries of β_0). Theorem 1 actually shows a more general result since it considers also higher-order errors, namely the kth order error $\mathbb{E}_{\beta_0}\|\hat{\beta} - \beta_0\|_1^k$ for any fixed $k \in \{1, 2, \ldots\}$.

We consider the situation when the errors ϵ_i are independent and sub-Gaussian (with a universal constant) and the design X has independent sub-Gaussian rows (with a universal constant). To this end, we recall a sub-Gaussianity assumption on random variables and vectors [see Section 14 in Bühlmann and van de Geer (2011)].

DEFINITION 2. We say that a random vector $Z \in \mathbb{R}^m$ has sub-Gaussian entries with constants $K, K_2 > 0$ if

$$\mathbb{E}e^{Z_j^2/K^2} \le K_2, \qquad j = 1, \ldots, m.$$

We say that a random vector $Z \in \mathbb{R}^m$ is sub-Gaussian with constants $K, K_2 > 0$ if for all $\alpha \in \mathbb{R}^m$ such that $\|\alpha\|_2 = 1$ it holds that

$$\mathbb{E}e^{(\alpha^T Z)^2/K^2} \le K_2.$$

In our further analysis, we typically require that the sub-Gaussianity condition as in Definition 2 is satisfied with universal constants $K, K_2 > 0$. A prime example of a sub-Gaussian random vector with a universal constant is a Gaussian random vector with zero mean and covariance matrix Σ_0 that satisfies $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$. We formulate the conditions on the error and the design in the following:

(A1) Assume the linear model (4), where the errors ϵ_i are independent sub-Gaussian random variables with universal constants and with $\mathbb{E}\epsilon_i = 0$.

(A2) Assume that X is a random $n \times p$ matrix independent of ϵ with independent rows $X^{(i)}$, $i = 1, \ldots, n$, with mean zero and with sub-Gaussian entries with universal constants. We let $\Sigma_0 := \mathbb{E}\hat{\Sigma}$ and suppose that $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$.

(A2*) Assume that X is a random $n \times p$ matrix independent of ϵ with independent sub-Gaussian rows $X^{(i)}$, $i = 1, \ldots, n$, with universal constants, with mean zero. We let $\Sigma_0 := \mathbb{E}\hat{\Sigma}$ and suppose that $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$.

Under conditions (A2) or (A2*), we denote the inverse covariance matrix by $\Theta_0 := \Sigma_0^{-1}$ and by Θ_j^0 we denote its jth column ($j = 1, \ldots, p$).

THEOREM 1. *Suppose that conditions* (A1), (A2) *are satisfied. Suppose that* $\|\beta_0\|_2 = \mathcal{O}(1)$, $s\sqrt{\log p/n} = o(1)$ *and let* $k \in \{1, 2, \ldots\}$ *be fixed. Consider the Lasso estimator* $\hat{\beta}$ *defined in* (5) *with a tuning parameter* $\lambda \geq c\tau\sqrt{\log p/n}$, *where* $c > 0$ *is a sufficiently large universal constant and* $\tau > 1$ *satisfies* $\tau^2 > 2k \log((\sqrt{s}\lambda^2)^{-1})/\log p$. *Then there exists a universal constant* C_1 *such that*

$$\big(\mathbb{E}_{\beta_0}\|\hat{\beta} - \beta_0\|_1^k\big)^{1/k} \leq C_1 s\lambda.$$

Taking $k = 1$, under the conditions of Theorem 1 we obtain

$$\mathbb{E}_{\beta_0}\|\hat{\beta} - \beta_0\|_1 \leq C_1 s\lambda.$$

Theorem 1 can also be easily extended to fixed design, under a compatibility condition [see Section 13 in the Supplementary Material, Janková and van de Geer (2018)] on the Gram matrix $\hat{\Sigma}$, which substitutes the condition $\Lambda_{\min}(\Sigma_0) \geq L > 0$, and under the condition $\|\hat{\Sigma}\|_\infty = \mathcal{O}(1)$.

We comment on conditions (A1), (A2) and $\|\beta_0\|_2 = \mathcal{O}(1)$, $\|\beta_0\|_0 = o(\sqrt{n/\log p})$ assumed in Theorem 1. Condition $\|\beta_0\|_0 = o(\sqrt{n/\log p})$ together with conditions (A1), (A2) was used to apply the high-probability oracle results for Lasso as in Bühlmann and van de Geer (2011) to the case of random design. Condition $\|\beta_0\|_2 = \mathcal{O}(1)$ can be justified under an assumption on the boundedness of the "signal-to-noise ratio". The "signal-to-noise ratio" is defined as the ratio of the variance of the signal (observations) and the variance of the noise, that is, $\sum_{i=1}^n \operatorname{var}_{\beta_0}(Y_i)/\sum_{i=1}^n \operatorname{var}(\epsilon_i) = 1 + \beta_0^T \Sigma_0 \beta_0/\sigma_\epsilon^2$, where $\sigma_\epsilon^2 := \frac{1}{n}\sum_{i=1}^n \operatorname{var}(\epsilon_i)$. Hence, under upper-boundedness of $1/\Lambda_{\min}(\Sigma_0)$, the signal-to-noise ratio is up to a constant lower-bounded by $\|\beta_0\|_2^2/\sigma_\epsilon^2$. If we assume that the signal-to-noise

ratio remains bounded and the variance of the noise σ_ϵ^2 is bounded [as implied by condition (A1)], then the ℓ_2-norm of β_0 must also remain bounded.

Finally, the condition $\tau^2 > 2k\log((s\lambda^2)^{-1})/\log p$ only guarantees that we choose sufficiently large regularization parameter $\lambda \geq c\tau\sqrt{\log p/n}$ by choosing τ large enough compared to the order k of the error that we want to control. If $p \geq n$ and $\lambda = c\tau\sqrt{\log p/n}$, the condition reduces to $\tau \geq C\sqrt{k}$ for some constant $C > 0$. Then clearly, this condition means that the higher order of error we want to control, the stronger regularization must be chosen.

8. The de-sparsified Lasso.

8.1. *Methodology.* As an initial estimator, we consider the Lasso estimator (5). The Lasso estimator is well understood in terms of prediction and estimation error bounds, and was shown minimax optimal in terms of the prediction error and ℓ_1-error. However, due to the inclusion of the ℓ_1-penalty, the estimator is biased and its limiting distribution can accumulate a positive mass at zero [Knight and Fu (2000)]. In view of statistical inference, a de-sparsified or de-biased version of the Lasso was then considered [see Zhang and Zhang (2014), van de Geer et al. (2014), Javanmard and Montanari (2014a, 2014b, 2015)], which was shown to be asymptotically normal for estimation of β_j^0.

To construct the de-biased estimator, we further need to construct a surrogate inverse of $\hat{\Sigma}$, or, in other words, we need to construct an estimator of the inverse covariance matrix $\Theta_0 = \Sigma_0^{-1}$. We define $\hat{\Theta}_j$ as an estimate of the column Θ_j^0 obtained by solving the following program, that will be referred to as *nodewise regression* [see Meinshausen and Bühlmann (2006), van de Geer et al. (2014)]. Recall that X is the design matrix with rows $X^{(i)}$, $i = 1, \ldots, n$. The columns of the design matrix X will be denoted by X_j, $j = 1, \ldots, p$, and by X_{-j} we denote the $n \times (p-1)$ matrix obtained by removing the jth column from X. For $j = 1, \ldots, p$, we let

$$\begin{aligned} \hat{\gamma}_j &:= \arg\min_{\gamma \in \mathbb{R}^{p-1}} \|X_j - X_{-j}\gamma\|_n^2 + 2\lambda_j\|\gamma\|_1, \\ \hat{\tau}_j^2 &:= \|X_j - X_{-j}\hat{\gamma}_j\|_n^2 + \lambda_j\|\hat{\gamma}_j\|_1, \end{aligned} \tag{6}$$

and we denote the jth column of the nodewise Lasso estimator by

$$\hat{\Theta}_j := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,j-1}, 1, -\hat{\gamma}_{j,j+1}, \ldots, -\hat{\gamma}_{j,p})^T/\hat{\tau}_j^2, \tag{7}$$

where $\lambda_j \asymp \sqrt{\log p/n}$ for $j = 1, \ldots, p$, uniformly in j. We denote the nodewise Lasso estimator by $\hat{\Theta} := (\hat{\Theta}_1, \ldots, \hat{\Theta}_p)$. The necessary Karush–Kuhn–Tucker conditions corresponding to the nodewise regression (obtained by replacing derivatives by sub-differentials) imply the condition $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty = \mathcal{O}_P(\lambda_j/\hat{\tau}_j^2)$ [see van de Geer et al. (2014)], which will be needed later. We now define the de-sparsified Lasso introduced in van de Geer et al. (2014),

$$\hat{b} := \hat{\beta} + \hat{\Theta}^T X^T (Y - X\hat{\beta})/n, \tag{8}$$

and we let $\hat{b}_j$ denote its jth entry. The motivation for the definition (8) comes from updating the initial Lasso estimator $\hat{\beta}$ by removing the bias due to the ℓ_1-penalty. We briefly summarize the main results on $\hat{b}$ as derived in van de Geer et al. (2014). The estimator $\hat{b}_j$ can be shown to be asymptotically linear with a remainder term of small order $1/\sqrt{n}$, in particular, under conditions (A1), (A2*) and

$$s = o(\sqrt{n}/\log p), \qquad \max_{j=1,\ldots,p} s_j = o(\sqrt{n}/\log p)$$

it holds

$$\hat{b} - \beta_0 = \Theta_0 X^T \epsilon/n + \Delta,$$

where $s_j := \|\Theta_j^0\|_0$ and $\|\Delta\|_\infty = o_P(1/\sqrt{n})$. Thus, after normalization by $\sqrt{n}$ and by the (estimated) standard deviation, asymptotic normality of entries of $\hat{b}$ with zero mean and unit variance follows by the central limit theorem. We now investigate the question of "regularity" and asymptotic efficiency of this estimator.

We first show that the de-sparsified estimator $\hat{b}_j$ satisfies the strong asymptotic unbiasedness condition from Definition 1 in Section 6. We then show that $\hat{b}_j$ achieves the lower bound on the variance of any strongly asymptotically unbiased estimator. Thus, in this sense the de-sparsified estimator is asymptotically efficient. In Section 8.3, we investigate the case of a random Gaussian design matrix and in Section 8.4 the case of a fixed design matrix.

8.2. *Strong asymptotic unbiasedness of the de-sparsified Lasso.* We consider estimation of linear functionals $g(\beta) = \xi^T \beta$, where $\xi \in \mathbb{R}^p$ is a known vector. We define an estimator of $g(\beta) = \xi^T \beta$ as a linear combination ξ of the de-sparsified estimator $\hat{b}$. This yields

$$\hat{b}_\xi := \xi^T \hat{b} = \xi^T \hat{\beta} + \xi^T \hat{\Theta} X^T (Y - X\hat{\beta})/n. \tag{9}$$

Then we have the following lemma, which shows strong asymptotic unbiasedness of $\hat{b}_\xi$ for estimation of $\xi^T \beta$.

LEMMA 1. *Suppose that conditions* (A1), (A2*) *are satisfied,* $\beta_0 \in \mathcal{B}(d_n)$ *where* $d_n = o(\sqrt{n}/\log p)$, $\max_j s_j \le d_n$, $\|\xi\|_1 = \mathcal{O}(1)$ *and* $\|\Sigma_0\|_\infty = \mathcal{O}(1)$. *Let* $\hat{b}_\xi$ *be the estimator defined in* (9) *with tuning parameters of the Lasso and nodewise regression* $\lambda \asymp \lambda_j \asymp \sqrt{\log p/n}$ *uniformly in* $j = 1, \ldots, p$. *Then* $\hat{b}_\xi$ *is a strongly asymptotically unbiased estimator of* $\xi^T \beta$ *at* β_0.

8.3. *Main results for random design.* We derive lower bounds for the variance of a strongly asymptotically unbiased estimator. We consider the following conditions on the error distribution and the design matrix X.

(B1) Assume the linear model (4) with $\epsilon \sim \mathcal{N}(0, I)$.

(B2) Assume that X is a random $n \times p$ matrix independent of ϵ with independent rows $X^{(i)} \sim \mathcal{N}(0, \Sigma_0)$ for $i = 1, \ldots, n$. Suppose that the inverse covariance matrix $\Theta_0 := \Sigma_0^{-1}$ exists, $1/\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$ and $\|\Sigma_0\|_\infty = \mathcal{O}(1)$.

THEOREM 2. *Suppose that conditions* (B1), (B2) *are satisfied. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta)$ at $\beta_0 \in \mathcal{B}(d_n)$ with a rate m_n. Let $h \in \mathbb{R}^p$ satisfy $h^T \Sigma_0 h = 1$ and $\beta_0 + h/\sqrt{m_n} \in B(\beta_0, \frac{c}{\sqrt{m_n}})$ for a sufficiently large universal constant c. Assume, moreover, that for some $\dot{g}(\beta_0) \in \mathbb{R}^p$ it holds*

$$\sqrt{m_n}\big(g(\beta_0 + h/\sqrt{m_n}) - g(\beta_0)\big) = h^T \dot{g}(\beta_0) + o(1). \tag{10}$$

Then

$$n \operatorname{var}_{\beta_0}(T_n) \geq \big[h^T \dot{g}(\beta_0)\big]^2 - o(1).$$

Theorem 2 yields a lower bound $[h^T \dot{g}(\beta_0)]^2 - o(1)$ on the variance of an estimator which is a strongly asymptotically unbiased estimator in a direction h, such that $\beta_0 + h/\sqrt{m_n}$ remains within the model. By maximizing $[h^T \dot{g}(\beta_0)]^2$ over all feasible h, we obtain the following corollary.

COROLLARY 1. *If $\beta_0 + \Theta_0 \dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0) m_n} \in B(\beta_0, \frac{c}{\sqrt{m_n}})$, then the lower bound from Theorem 2 is maximized at the value*

$$h_0 := \Theta_0 \dot{g}(\beta_0)/\sqrt{\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)},$$

and under the conditions of Theorem 2, we get and under the conditions of Theorem 2, we get

$$n \operatorname{var}_{\beta_0}(T_n) \geq \dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0) - o(1).$$

DEFINITION 3. Let g be differentiable at β_0 with derivative $\dot{g}(\beta_0)$. We call

$$c_0 := \Theta_0 \dot{g}(\beta_0)/\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)$$

the worst possible sub-direction for estimating $g(\beta_0)$.

The motivation for the terminology *worst possible sub-direction* in Definition 3 is given by Corollary 1. The normalization by $\dot{g}(\beta_0)^T \Theta_0 \dot{g}(\beta_0)$ is arbitrary but natural from a projection theory point of view.

As a special case, consider estimation of $g(\beta) = \beta_j$ for some fixed value of $j \in \{1, \ldots, p\}$. Then $\dot{g}(\beta) = e_j$, the jth unit vector in $\mathbb{R}^p$. Clearly, $\Theta_0 \dot{g}(\beta_0) = \Theta_0 e_j = \Theta_j^0$ and $g(\beta_0)^T \Theta_0 \dot{g}(\beta_0) = e_j^T \Theta_0 e_j = \Theta_{jj}^0$, where Θ_j^0 is the jth column of Θ_0 and Θ_{jj}^0 is its jth diagonal element. It follows that $c_j^0 = \Theta_j^0/\Theta_{jj}^0$ is the worst

possible sub-direction for estimating β_j. If $\beta_0 + \Theta_j^0/\sqrt{\Theta_{jj}^0 m_n} \in B(\beta_0, \frac{c}{\sqrt{m_n}})$, then Corollary 1 implies the lower bound

$$\operatorname{var}_{\beta_0}(T_n) \geq \Theta_{jj}^0/n + o(1/n).$$

REMARK 2. To establish the lower bound, it is crucial that the worst possible sub-direction lies within the model. For illustration, consider the situation with the parameter of interest being $g(\beta) = \beta_1$. When Θ_1^0 is not sufficiently sparse, we are not allowed to take the global maximizer $h = \Theta_1^0/\sqrt{\Theta_{11}^0}$ in the maximum and the lower bound might thus become smaller. In that case, the lower bound is given via a sparse approximation of the (non-sparse) precision matrix. For a set $M \subset \{1, \ldots, p\}$ and a vector $v \in \mathbb{R}^p$, we denote v_M as a p-dimensional vector with entries not in M set to zero. Then we may write

$$\begin{aligned}&\max_{\beta_0 + h/\sqrt{m_n} \in B(\beta_0, c/\sqrt{m_n})} \frac{h^T e_1}{h^T \Sigma_0 h} \\ &\quad \geq \max_{\substack{M \subset \{1,\ldots,p\}: \\ |M| = d_n - \|\beta_0\|_0}} \max_{\substack{h \in \mathbb{R}^p: \|h_M\|_2 \leq c, \\ \|\beta_M^0 + h_M/\sqrt{m_n}\|_2 \leq C}} \frac{[h_M^T e_1]^2}{h_M^T \Sigma_0 h_M}.\end{aligned}$$

But if $h := (\Sigma_{M,M}^0)^{-1} e_1$ satisfies $\|h\|_2 \leq c$ and $\|\beta_M^0 + h/\sqrt{m_n}\|_2 \leq C$, then the lower bound is

$$\max_{M \subset \{1,\ldots,p\}: |M| = d_n - \|\beta_0\|_0} (\Sigma_{M,M}^0)_{11}^{-1} - o(1),$$

where $\Sigma_{M,M}^0$ is the reduction of Σ_0 obtained by keeping only columns and rows belonging to the set M. If Θ_1^0 has sparsity $d_n - \|\beta_0\|_0$, then this lower bound coincides with $(\Sigma_0)_{11}^{-1} - o(1)$ as before. If Θ_1^0 is not sufficiently sparse, then the lower bound is given via a sparse approximation of the precision matrix. Finally, as will be seen in the following sections, without assuming the sparsity condition on the worst possible sub-direction, we would not be able to conclude asymptotic efficiency of the de-sparsified Lasso estimator.

Finally, we show that the de-sparsified estimator $\hat{b}_j$ achieves the lower bound on the variance. Thus, the de-sparsified estimator is strongly asymptotically unbiased and has the smallest variance among all strongly asymptotically unbiased estimators. We assume Gaussianity of the error and the design matrix, as the lower bounds have only been derived for this case.

THEOREM 3. *Suppose that conditions* (B1), (B2) *are satisfied*, $\beta_0 \in \mathcal{B}(d_n)$ *with* $d_n = o(\sqrt{n}/\log p)$ *and* $\max_j s_j \leq d_n$. *Assume that* $\|\xi\|_1 = \mathcal{O}(1)$. *Let* $\hat{b}_\xi$ *be the estimator defined in* (9) *with tuning parameters of the Lasso and nodewise regression* $\lambda \asymp \lambda_j \asymp \sqrt{\log p/n}$, *uniformly in* $j = 1, \ldots, p$. *Then* $\hat{b}_\xi$ *is a strongly*

asymptotically unbiased estimator of $\xi^T\beta$ at β_0. Let T be any strongly asymptotically unbiased estimator of $\xi^T\beta$ at β_0 and assume that $\beta_0+\Theta_0\xi/(\xi^T\Theta_0\xi m_n)^{1/2} \in B(\beta_0, c/\sqrt{m_n})$ where $n = o(m_n)$. Then it holds

$$\operatorname{var}_{\beta_0}(T) \geq \frac{\xi^T\Theta_0\xi + o(1)}{n}, \qquad \operatorname{var}_{\beta_0}(\hat{b}_\xi) = \frac{\xi^T\Theta_0\xi + o(1)}{n}.$$

To obtain the result of Theorem 3, we assumed that $\beta_0+\Theta_0\xi/(\xi^T\Theta_0\xi m_n)^{1/2} \in B(\beta_0, c/\sqrt{m_n})$, which guarantees that the worst possible sub-direction stays within the model. Further we assumed that the sparsity in β_0 satisfies $s = o(\sqrt{n}/\log p)$ and that the sparsity in the rows of Θ_0 is of small order $\sqrt{n}/\log p$. Thus, to be able to claim asymptotic efficiency of the de-sparsified Lasso, we not only require sparsity in β_0, but also sufficient sparsity in the precision matrix. Note that the sparsity condition on β_0 is almost a necessary condition as discussed in Section 8.6 below.

8.4. *Main results for fixed design.* In this section, we assume that the design matrix X is fixed (nonrandom). Recall that $\hat{\Sigma} = X^TX/n$ is the Gram matrix. The following theorem is an analogy of Theorem 2 for fixed design.

THEOREM 4. *Let X be a fixed $n \times p$ matrix and suppose that condition* (B1) *is satisfied. Let $h \in \mathbb{R}^p$ be such that $h^T\hat{\Sigma}h = \mathcal{O}(1)$ and $\beta_0 + h/\sqrt{m_n} \in B(\beta_0, c/\sqrt{m_n})$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\beta)$ at β_0 in the direction h with rate m_n. Assume moreover that for some $\dot{g}(\beta_0) \in \mathbb{R}^p$ it holds that*

$$\sqrt{m_n}\big(g(\beta_0 + h/\sqrt{m_n}) - g(\beta_0)\big) = h^T\dot{g}(\beta_0) + o(1). \tag{11}$$

Then

$$n \operatorname{var}_{\beta_0}(T_n) \geq \big[h^T\dot{g}(\beta_0)\big]^2 - o(1).$$

For fixed design, the matrix $\hat{\Sigma}$ is not invertible, and thus we cannot use the reasoning as in Section 8.3. We can however try to remedy this by proposing an approximate worst possible sub-direction. To this end, we may use an estimator $\hat{\Theta}$, which acts as a surrogate inverse of $\hat{\Sigma}$ in a certain sense. Such an estimate can be obtained in the same way as for the random design, using the nodewise regression (7). The necessary Karush–Kuhn–Tucker conditions of the nodewise regression (obtained by replacing derivatives by sub-differentials) again imply the condition $\|\hat{\Sigma}\hat{\Theta}_j - e_j\|_\infty = \mathcal{O}_P(\lambda_j/\hat{\tau}_j^2)$. The de-sparsified estimator can then be defined in the same way as for the random design, as in equation (8).

We consider estimation of $g(\beta_0) := \beta_j^0$, although one could further consider estimation of linear functionals, similarly as for the random design. Strong asymptotic unbiasedness of $\hat{b}_j$ for estimation of β_j then follows similarly as in Lemma 1

[with $g(\beta) = \beta_j$] for all $\beta \in \mathcal{B}(d_n)$, under $d_n = o(\sqrt{n}/\log p)$, if the compatibility condition is satisfied for $\hat{\Sigma}$ with a universal constant and $\|\hat{\Sigma}\|_\infty = \mathcal{O}(1)$. For the definition of the compatibility condition, see Definition 4 in Section 13 of the Supplementary Material [Janková and van de Geer (2018)]. We formulate the asymptotic efficiency of $\hat{b}_j$ for $g(\beta) := \beta_j$ in the following theorem.

THEOREM 5. *Assume that condition* (B1) *is satisfied and* $\beta_0 \in \mathcal{B}(d_n)$ *with* $d_n = o(\sqrt{n}/\log p)$. *Let* $j \in \{1, \dots, p\}$ *and let* $\hat{\Theta}_j$ *be obtained using the nodewise regression as in* (7) *with* $\lambda_j \asymp \sqrt{\log p/n}$. *Suppose that* $\beta_0 + \hat{\Theta}_j/(\hat{\Theta}_{jj} m_n)^{1/2} \in B(\beta_0, c/\sqrt{m_n})$ *with* $n = o(m_n)$, $\|\hat{\Theta}_j\|_2 = \mathcal{O}(1)$, *the compatibility condition is satisfied for* $\hat{\Sigma}$ *with a universal constant and* $\|\hat{\Sigma}\|_\infty = \mathcal{O}(1)$. *Then* $\hat{b}_j$ *defined in* (8) *using* $\hat{\Theta}_j$ *and with* $\lambda \asymp \sqrt{\log p/n}$ *is a strongly asymptotically unbiased estimator of* β_j *at* β_0 *and for any strongly asymptotically unbiased estimator* T *of* β_j *at* β_0 *it holds*

$$\operatorname{var}_{\beta_0}(T) \geq \frac{\hat{\Theta}_{jj} + o(1)}{n}, \qquad \operatorname{var}_{\beta_0}(\hat{b}_j) = \frac{\hat{\Theta}_{jj} + o(1)}{n}.$$

The condition $\beta_0 + \hat{\Theta}_j/\sqrt{\hat{\Theta}_{jj} m_n} \in \mathcal{B}(d_n)$ implies that $\|\hat{\Theta}_j\|_0 = \mathcal{O}(d_n)$. To this end, we refer to Lemma 12 in Section 14 of the Supplementary Material [Janková and van de Geer (2018)], which shows that sparsity in $\hat{\Theta}_j$ constructed using nodewise regression is guaranteed under random design. The condition $\|\hat{\Theta}_j\|_2 = \mathcal{O}(1)$ replaces the eigenvalue condition we needed in the case of random design.

8.5. *Le Cam's bounds.* In this section, we provide an alternative approach, which makes another choice in the formulation of asymptotic efficiency. This approach is based on Le Cam's arguments [see, e.g., van der Vaart (1998)] rather than the Cramér–Rao bounds, and it allows us to show that the convergence of the de-sparsified estimator to the limiting normal distribution with smallest possible variance is locally uniform in the underlying unknown parameter, and the asymptotic variance of the de-sparsified estimator is smallest among the class of asymptotically linear estimators. Furthermore, the result identifies the asymptotic bias of asymptotically linear estimators. A detailed comparison of the two approaches for deriving the lower bounds is deferred to Section 11.

We consider the setting from Section 8.3, where the design matrix X is random with the parameter of interest being $g(\beta) = \beta_j$.

THEOREM 6. *Assume that conditions* (B1), (B2) *are satisfied,* $\beta_0 \in \mathcal{B}(d_n)$ *with* $d_n = o(\sqrt{n}/\log p)$, $\|\Theta_j^0\|_0 \leq d_n$ *and* $\Lambda_{\max}(\Sigma_0) = \mathcal{O}(1)$. *Assume that* $\hat{b}_j$ *is defined in* (8) *with tuning parameters* $\lambda \asymp \lambda_j \asymp \sqrt{\log p/n}$. *Then for every* $\tilde{\beta}_n \in B(\beta_0, \frac{c}{\sqrt{n}})$ *it holds*

$$\frac{\sqrt{n}(\hat{b}_j - \tilde{\beta}_n)}{(\Theta_{jj}^0)^{1/2}} \overset{\tilde{\beta}_n}{\rightsquigarrow} \mathcal{N}(0, 1).$$

Let T_n be an asymptotically linear estimator with an influence function l_{β_0}:

$$T_n - \beta_j^0 = \frac{1}{n}\sum_{i=1}^{n} l_{\beta_0}(X^{(i)}, Y^{(i)}) + o_{P_{\beta_0}}(n^{-1/2}), \tag{12}$$

where $\mathbb{E} l_{\beta_0}(X^{(i)}, Y^{(i)}) = 0$ and $\operatorname{var}(l_{\beta_0}(X^{(i)}, Y^{(i)})) =: V_{\beta_0} < \infty$. Assume that for all $h \in \mathbb{R}^p$ and $i = 1, \ldots, n$ it holds

$$\mathbb{E} l_{\beta_0}(X^{(i)}, Y^{(i)})\epsilon_i h^T X^{(i)} - h_j = o(1). \tag{13}$$

Then

$$V_{\beta_0} \geq \Theta_{jj}^0 + o(1).$$

8.6. *Discussion of the conditions.* We briefly discuss the conditions assumed to obtain the above results. To establish asymptotic efficiency of the de-sparsified estimator, we considered conditions analogous to the conditions assumed in van de Geer et al. (2014). These include a sparsity condition on the parameter β_0 of order $o(\sqrt{n}/\log p)$, conditions on the covariance matrix $\Lambda_{\min}(\Sigma_0) = \mathcal{O}(1)$, $\|\Sigma_0\|_\infty = \mathcal{O}(1)$, sparsity of the precision matrix and a Gaussianity assumption on the rows on the precision matrix. Unlike in van de Geer et al. (2014), we assume Gaussianity of the design matrix and the error; this condition was needed to derive the lower bounds. In addition to the conditions from van de Geer et al. (2014), we also assume boundedness of ℓ_2-norm of β_0, which follows if the signal to noise ratio is bounded as argued in Section 7. Condition (13) from Theorem 6 is a variant of asymptotic unbiasedness which is known to be satisfied in many traditional settings [see, e.g., van der Vaart (1998)]. The condition is discussed in more detail in Section 10 below.

Our analysis requires the sparsity condition $s = o(\sqrt{n}/\log p)$. This condition is essentially necessary in the linear regression setting for construction of an asymptotically normal estimator, as argued in the following. First, observe that if the (slightly weaker) condition $s = \mathcal{O}(\sqrt{n}/\log p)$ is not satisfied, then there cannot exist an estimator T_n of $\beta_j \in \mathbb{R}$ and a sequence $\sigma_n = \mathcal{O}(1)$ such that

$$\sqrt{n}(T_n - \beta_j^0)/\sigma_n \rightsquigarrow \mathcal{N}(0, 1). \tag{14}$$

Suppose that there exists an estimator T_n that satisfies (14). Then necessarily $\sqrt{n}(T_n - \beta_j^0)/\sigma_n = \mathcal{O}_P(1)$. By similar reasoning as in Ren et al. (2015), we have under the conditions assumed the minimax rates for $\mathbb{E}|T_n - \beta_j^0|$ of order $\frac{1}{\sqrt{n}} + \frac{s \log p}{n}$. But then necessarily $s \log p/n = \mathcal{O}(1/\sqrt{n})$, which gives $s = \mathcal{O}(\sqrt{n}/\log p)$. This is only slightly weaker than the condition we require, $s = o(\sqrt{n}/\log p)$.

Furthermore, for simplicity of presentation, we assumed that the variance of the noise is fixed at $\sigma_\epsilon = 1$. In general, we can include the parameter σ_ϵ as an unknown parameter in the model, and by orthogonality of the score corresponding to this parameter and the score corresponding to β, we can easily extend the arguments. The noise variance will then appear in both lower and upper bounds.

9. Gaussian graphical models. In this part, we consider efficient estimation of edge weights in undirected Gaussian graphical models. Gaussian graphical models have become a popular tool for representing dependencies within large sets of variables and have found application in areas such as neuroscience, biology and climate data analysis. In particular, Gaussian graphical models encode conditional dependencies between variables (nodes in the graph) by including an edge between two variables if and only if they are not independent given all the other variables. This corresponds to the problem of estimation of the precision matrix of a multivariate normal distribution, which we now introduce.

(C1) Assume that the $n \times p$ matrix X has independent rows $X^{(i)}$, $i = 1, \ldots, n$ which are $\mathcal{N}_p(0, \Sigma_0)$-distributed.

Denote the precision matrix by $\Theta_0 := \Sigma_0^{-1}$, where the inverse of Σ_0 is assumed to exist. The matrix $\Theta_0 \in \mathbb{R}^{p \times p}$ is unknown, but we assume bounds on its row-sparsity (column-sparsity) $s_j := \|\Theta_j^0\|_0$, where Θ_j^0 is the jth column of the precision matrix.

9.1. *Methodology.* There have been several methods proposed for estimation of the precision matrix in the high-dimensional setting when $p \gg n$ [see Friedman, Hastie and Tibshirani (2008), Meinshausen and Bühlmann (2006)]. These methods are based on regularization techniques and lead to estimators that are biased. De-biasing was then studied similarly as in the linear regression, and it was shown that de-biasing leads to estimators which are asymptotically normal. For our further analysis, we consider the de-sparsified nodewise Lasso estimator proposed in Janková and van de Geer (2017). We show that this estimator is strongly asymptotically unbiased and reaches the lower bound on the variance derived in the previous section.

To introduce the methodology, consider again the nodewise Lasso estimator $\hat{\Theta} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_p)$ defined in (7). Define the de-sparsified nodewise Lasso [see Janková and van de Geer (2017)]

$$\hat{T} := \hat{\Theta} + \hat{\Theta}^T - \hat{\Theta}\hat{\Sigma}\hat{\Theta}. \tag{15}$$

Furthermore, we write $\hat{T}_{ij} := \hat{\Theta}_{ij} + \hat{\Theta}_{ji} - \hat{\Theta}_i^T \hat{\Sigma} \hat{\Theta}_j$ for $i, j = 1, \ldots, p$. The method and its asymptotic properties were studied in Janková and van de Geer (2017). The estimator $\hat{\Theta}_j$ can be shown to be asymptotically linear with a remainder term of small order $1/\sqrt{n}$, in particular, under condition (C1) and under $\max_{j=1,\ldots,p} s_j = o(\sqrt{n}/\log p)$ it holds

$$\hat{T} - \Theta_0 = -\Theta_0^T (\hat{\Sigma} - \Sigma_0)\Theta_0 + \Delta,$$

where $\|\Delta\|_\infty = o_P(1/\sqrt{n})$. Thus, after normalization by $\sqrt{n}$ and by the (estimated) standard deviation, it follows that it is asymptotically standard normal and minimax optimal [see Ren et al. (2015), Janková and van de Geer (2017)]. We investigate the question of "regularity" and asymptotic efficiency of the proposed estimator.

9.2. *Strong asymptotic unbiasedness of the de-sparsified nodewise Lasso.* Suppose that the parameter Θ ranges over a parameter space $T \subset \mathbb{R}^{p\times p}$. We then define the parameter set

$$\mathcal{G}(d_1,\dots,d_p) := \big\{\Theta \in T : \Theta = \Theta^T, \|\Theta_j\|_0 \le C_1 d_j, j = 1,\dots,p, \\ 1/\Lambda_{\min}(\Theta) \le C_2, \Lambda_{\max}(\Theta) \le C_3\big\},$$

for some universal constants $C_1, C_2, C_3 > 0$. We also need to readjust the definition of a neighbourhood from (3); hence in this section we let

$$B(\Theta,\epsilon) := \big\{\tilde{\Theta} \in \mathcal{G}(d_1,\dots,d_p) : \|\tilde{\Theta} - \Theta\|_F \le \epsilon\big\}.$$

The following lemma shows that $\hat{T}_{ij}$ is strongly asymptotically unbiased for estimation of Θ^0_{ij}.

LEMMA 2. *Let $i, j \in \{1,\dots,p\}$, assume that condition* (C1) *is satisfied and $\Theta_0 \in \mathcal{G}(d_1,\dots,d_p)$ with $\max(d_i,d_j) = o(\sqrt{n}/\log p)$. Let $\hat{T}_{ij}$ be defined in* (15)*, where $\hat{\Theta}_i$, $\hat{\Theta}_j$ are the ith and jth columns of the nodewise Lasso estimator with tuning parameters $\lambda_i \asymp \lambda_j \asymp \sqrt{\log p/n}$. Then $\hat{T}_{ij}$ is a strongly asymptotically unbiased estimator for Θ^0_{ij}.*

9.3. *Main results.* We first derive an asymptotic lower bound for the variance of T_n when T_n is strongly asymptotically unbiased. We restrict our attention to estimation of linear functionals of the precision matrix Θ_0, $h(\Theta_0) = \operatorname{tr}(\Psi\Theta_0)$, where $\Psi \in \mathbb{R}^{p\times p}$ is a known matrix. We shall consider the case when Ψ is of rank one, say $\Psi = \xi_1\xi_2^T$ for some vectors $\xi_1, \xi_2 \in \mathbb{R}^p$. This corresponds to estimation of $g(\Theta_0) = \xi_1^T\Theta_0\xi_2$, where $\xi_1, \xi_2 \in \mathbb{R}^p$ are known vectors.

Contrary to previous sections, the high-dimensional parameter is a matrix, therefore, instead of a vector direction h we shall write the capital letter H to denote a matrix direction in $\mathbb{R}^{p\times p}$.

THEOREM 7. *Assume condition* (C1)*, assume that $\Theta_0 \in \mathcal{G}(d_1,\dots,d_p)$ where $\max_{j=1,\dots,p} d_j = o(\sqrt{n}/\log p)$ and $\Theta_0 + H/\sqrt{m_n} \in B(\Theta_0, c/\sqrt{m_n})$ where $n = o(m_n)$. Suppose that T_n is a strongly asymptotically unbiased estimator of $g(\Theta) = \xi_1^T\Theta\xi_2$ at $\Theta_0 \in \mathcal{G}(d_1,\dots,d_p)$ in the direction $H := \Theta_0(\xi_1\xi_2^T + \xi_2\xi_1^T)\Theta_0/\sigma$, where*

$$\sigma^2 := \xi_1^T\Theta_0\xi_1\xi_2^T\Theta_0\xi_2 + \big(\xi_1^T\Theta_0\xi_2\big)^2.$$

Then it holds

$$\operatorname{var}_{\Theta_0}(T_n) \ge \frac{\sigma^2 - o(1)}{n}.$$

As a corollary, consider estimation of $g(\Theta_0) = \Theta^0_{ij}$ for some fixed $(i, j) \in \{1, \ldots, p\}^2$. Then the worst sub-direction is given by $H := (\Theta^0_i(\Theta^0_j)^T + \Theta^0_j(\Theta^0_i)^T)/\sigma$ where $\sigma^2 := (\Theta^0_{ij})^2 + \Theta^0_{ii}\Theta^0_{jj}$ and the corresponding lower bound is $((\Theta^0_{ij})^2 + \Theta^0_{ii}\Theta^0_{jj})/n + o(1/n)$.

We now show that the de-sparsified estimator $\hat{T}_{ij}$ reaches the lower bound on the variance for the parameter of interest $g(\Theta_0) = \Theta^0_{ij}$.

THEOREM 8. *Suppose that condition* (C1) *holds*, $\Theta_0 \in \mathcal{G}(d_1, \ldots, d_p)$ *where* $\max(d_i, d_j) = o(\sqrt{n}/\log p)$. *Suppose that* $\Theta_0 + H/\sqrt{m_n} \in B(\Theta_0, c/\sqrt{m_n})$ *for* $H := (\Theta^0_i(\Theta^0_j)^T + \Theta^0_j(\Theta^0_i)^T)/\sigma$, *where* $n = o(m_n)$. *Let* $\hat{T}_{ij}$ *be defined in* (15), *where* $\hat{\Theta}_i$, $\hat{\Theta}_j$ *are the* i*th and* j*th columns of the nodewise Lasso estimator with tuning parameters* $\lambda_i \asymp \lambda_j \asymp \sqrt{\log p/n}$. *Then* $\hat{T}_{ij}$ *is a strongly asymptotically unbiased estimator of* Θ_{ij} *at* Θ_0 *and for any strongly asymptotically unbiased estimator* T *of* Θ_{ij} *at* Θ_0 *it holds*

$$\operatorname{var}_{\Theta_0}(T) \geq \frac{\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2 + o(1)}{n},$$

$$\operatorname{var}_{\Theta_0}(\hat{T}_{ij}) = \frac{\Theta^0_{ii}\Theta^0_{jj} + (\Theta^0_{ij})^2 + o(1)}{n}.$$

The condition $\Theta_0 + H/\sqrt{m_n} \in B(\Theta_0, c/\sqrt{m_n})$ for $H = (\Theta^0_i(\Theta^0_j)^T + \Theta^0_j(\Theta^0_i)^T)/\sigma$ ensures that perturbation of Θ_0 along the worst possible sub-direction H lies within the model. This also implies that $\|H_k\|_0 \leq 2C_1 d_k$, $k = 1, \ldots, p$, which in turn implies that necessarily $\|\Theta^0_i\|_0 = \mathcal{O}(d_k)$, $\|\Theta^0_j\|_0 = \mathcal{O}(d_k)$ for $k = 1, \ldots, p$. Note that we only require sparsity in the ith and jth column of the precision matrix. Furthermore, we must have $\|H\|_F \leq c$. This is satisfied under the eigenvalue conditions noting that $\|H\|_F^2 = \operatorname{tr}(H^T H)$ and $\|\Theta^0_k\|_2 = \mathcal{O}(1)$ for $k = i, j$.

9.4. *Discussion of the conditions.* We comment on the conditions used to obtain the above results. The conditions under which we show asymptotic efficiency only include eigenvalue conditions on the true precision matrix, sparsity conditions on columns/rows of the precision matrix and Gaussianity of the observations $X^{(i)}$, $i = 1, \ldots, n$. These conditions are almost identical to conditions in van de Geer et al. (2014) and Janková and van de Geer (2017), with the exception of Gaussianity which was used for deriving the lower bounds. In particular, the condition on row sparsity required is the same as for the linear model: $s = o(\sqrt{n}/\log p)$. In view of the results on minimax rates for estimation of elements of precision matrices [which are derived in Ren et al. (2015)], the condition $s = o(\sqrt{n}/\log p)$ is necessary for asymptotically normal estimation, which follows by similar reasoning as for the linear regression.

10. Le Cam's bounds for general models. In this section, we provide an extension to general nonlinear models and a general parameter of interest. This is achieved via adjustment of Le Cam's arguments on asymptotic efficiency to the high-dimensional setting. Let $X^{(1)}, \ldots, X^{(n)}$ be i.i.d. with distribution $P_{\beta_{n,0}}$: $\beta_{n,0} \in \mathcal{B}$ where $\mathcal{B}$ is an open convex subset of $\mathbb{R}^p$. We consider the parameter set

$$\mathcal{B}(d_n) := \{\beta \in \mathcal{B} : \|\beta\|_0 \leq C_1 d_n, \|\beta\|_2 \leq C_2\},$$

where $C_1, C_2 = \mathcal{O}(1)$ and d_n is a known sequence that will be specified later. Suppose that the parameter of interest is $g(\beta)$ for some function $g : \mathcal{B} \to \mathbb{R}$. Assume that for an estimator T_n of $g(\beta_{n,0})$, we can show asymptotic linearity: there exists a real-valued function $l_{\beta_{n,0}}$ on $\mathcal{X}$ (an influence function) and some sequence $\beta_{n,0}$ such that

$$T_n - g(\beta_{n,0}) = \frac{1}{n}\sum_{i=1}^{n} l_{\beta_{n,0}}(X^{(i)}) + o_{P_{\beta_{n,0}}}(n^{-1/2}),$$

where $P_{\beta_{n,0}} l_{\beta_{n,0}} = 0$ and the variance $V_{\beta_{n,0}} := P_{\beta_{n,0}} l^2_{\beta_{n,0}} < \infty$. Under the conditions of the central limit theorem, the asymptotic linearity implies that

$$\sqrt{n}(T_n - g(\beta_{n,0}))/V^{1/2}_{\beta_{n,0}} \overset{\beta_{n,0}}{\rightsquigarrow} \mathcal{N}(0,1). \tag{16}$$

For asymptotically linear estimators, we thus have the "asymptotic variance" $V_{\beta_{n,0}} = P_{\beta_{n,0}} l^2_{\beta_{n,0}}$. We shall need some conditions on the differentiability of g and the score function. Furthermore, we shall need a Lindeberg's condition related to the influence and score function. Assume that P_β is dominated by some σ-finite measure for all β in the parameter space and denote the corresponding probability densities by p_β. We denote the log-likelihood by $\ell_\beta(x) := \log p_\beta(x)$ and the score function by $s_\beta(x) := \frac{\partial \ell_\beta(x)}{\partial \beta}$ for all $x \in \mathcal{X}$.

(D1) *(Differentiability of g)* Suppose that for a given $\tilde{\beta}_n \in B(\beta_{n,0}, \frac{c}{\sqrt{n}})$ it holds

$$\sqrt{n}(g(\tilde{\beta}_n) - g(\beta_{n,0})) = h^T \dot{g}(\beta_{n,0}) + o(1),$$

where $h = \sqrt{n}(\tilde{\beta}_n - \beta_{n,0})$.

(D2) *(Differentiability of the score)* Suppose that the score function $\beta \mapsto s_\beta$ is twice differentiable and the second derivative satisfies $\|\ddot{s}_\beta\|_\infty \leq L$ for some universal constant $L > 0$ and for all $\beta \in \mathcal{B}(d_n)$. Let $I_{\beta_{n,0}} := P_{\beta_{n,0}} s_{\beta_{n,0}} s^T_{\beta_{n,0}}$ and assume that $\Lambda_{\max}(I_{\beta_{n,0}}) = \mathcal{O}(1)$, $1/\Lambda_{\min}(I_{\beta_{n,0}}) = \mathcal{O}(1)$ and

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \dot{s}_{\beta_{n,0}} + I_{\beta_{n,0}} \right\|_\infty = \mathcal{O}_P(\lambda), \tag{17}$$

for some $\lambda > 0$. Suppose that $d_n = o(\max\{1/\lambda, n^{1/3}\})$.

(D3) *(Lindeberg's condition)* Denote $f_{\beta_{n,0}}(x) := l_{\beta_{n,0}}(x) + h^T s_{\beta_{n,0}}(x)$ for $x \in \mathbb{R}^p$. Suppose that for all $\epsilon > 0$

$$\lim_{n\to\infty} P_{\beta_{n,0}} f^2_{\beta_{n,0}} \mathbf{1}_{|f_{\beta_{n,0}}|>\epsilon\sqrt{n}} = 0, \tag{18}$$

and assume that $V_{\beta_{n,0}} := P_{\beta_{n,0}} l^2_{\beta_{n,0}} = \mathcal{O}(1)$ and $1/V_{\beta_{n,0}} = \mathcal{O}(1)$.

Condition (D1) is a differentiability condition on g; an analogous condition is assumed in the first approach through Cramér–Rao bounds. Condition (D2) is a differentiability condition on the score, which is used to obtain a Taylor expansion of the likelihood. Furthermore, condition (17) guarantees that $-\frac{1}{n}\sum_{i=1}^n \dot{s}_{\beta_{n,0}}(X^{(i)})$ is a good estimator of the Fisher information in supremum norm. This can be verified, for example, for linear regression with $\lambda \asymp \sqrt{\log p/n}$. Condition (D2) further assumes the sparsity $d_n = o(\max\{1/\lambda, n^{1/3}\})$, which guarantees that the likelihood ratio expansion approximately holds. Finally, condition (D3) is a Lindeberg's condition which is needed to conclude asymptotic normality of certain quantities, since in Theorem 9 below we do not require any distributional assumption. This condition can be verified for particular models.

THEOREM 9. *Let $g : \mathcal{B} \to \mathbb{R}$ and suppose that for some fixed sequence $\beta_{n,0} \in \mathcal{B}(d_n)$ it holds*:

$$T_n - g(\beta_{n,0}) = \frac{1}{n}\sum_{i=1}^n l_{\beta_{n,0}}(X^{(i)}) + o_{P_{\beta_{n,0}}}(n^{-1/2}), \tag{19}$$

where $P_{\beta_{n,0}} l_{\beta_{n,0}} = 0$. For some fixed constant $c > 0$, let $\tilde{\beta}_n \in B(\beta_{n,0}, \frac{c}{\sqrt{n}})$ and denote $h := \sqrt{n}(\tilde{\beta}_n - \beta_{n,0})$. Suppose that conditions (D1), (D2) *and* (D3) *are satisfied. Then it holds*:

$$\frac{\sqrt{n}(T_n - g(\beta_{n,0} + \frac{h}{\sqrt{n}})) - (P_{\beta_{n,0}}(l_{\beta_{n,0}} h^T s_{\beta_{n,0}}) - h^T \dot{g}(\beta_{n,0}))}{V^{1/2}_{\beta_{n,0}}} \overset{\beta_{n,0}+\frac{h}{\sqrt{n}}}{\rightsquigarrow} \mathcal{N}(0,1).$$

The result of Theorem 9 contains a bias term $P_{\beta_{n,0}}(l_{\beta_{n,0}} h^T s_{\beta_{n,0}}) - h^T \dot{g}(\beta_{n,0})$ which depends on h. Now consider that the bias term in the result of the theorem above vanishes, that is, the following condition on the score function $s_{\beta_{n,0}}$ and the function $l_{\beta_{n,0}}$ is satisfied: for every $h \in \mathbb{R}^p$ it holds that

$$P_{\beta_{n,0}}(l_{\beta_{n,0}} h^T s_{\beta_{n,0}}) - h^T \dot{g}(\beta_{n,0}) = o(1). \tag{20}$$

The condition (20) is a variant of asymptotic unbiasedness which is known to be satisfied in many traditional settings. If condition (20) is satisfied, then the Cauchy–Schwarz inequality implies

$$(h^T \dot{g}(\beta_{n,0}))^2 \le V_{\beta_{n,0}} h^T I_{\beta_{n,0}} h + o(V^{1/2}_{\beta_{n,0}}(h^T I_{\beta_{n,0}} h)^{1/2}).$$

Hence this implies a lower bound on the asymptotic variance $V_{\beta_{n,0}}$ of an asymptotically linear estimator as follows:

$$V_{\beta_{n,0}} \geq \big(h^T \dot{g}(\beta_{n,0})\big)^2 / h^T I_{\beta_{n,0}} h + o\big(V_{\beta_{n,0}}^{1/2} / (h^T I_{\beta_{n,0}} h)^{1/2}\big). \tag{21}$$

Assuming that the inverse of $I_{\beta_{n,0}}$ exists, the right-hand side of (21) is maximized at $h = I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0})$, provided that $\beta_{n,0} + h/\sqrt{n} \in B(\beta_{n,0}, c/\sqrt{n})$. Hence we obtain the following lower bound on the asymptotic variance

$$V_{\beta_{n,0}} \geq \dot{g}(\beta_{n,0})^T I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0}) + o\big(V_{\beta_{n,0}}^{1/2} \big(\dot{g}(\beta_{n,0})^T I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0})\big)^{1/2}\big).$$

We summarize this simple claim in the lemma below.

LEMMA 3. *Let T_n satisfy (19) with $V_{\beta_{n,0}} = \mathcal{O}(1)$, $1/\Lambda_{\min}(I_{\beta_{n,0}}) = \mathcal{O}(1)$ and for every $h \in \mathbb{R}^p$ it holds that*

$$P_{\beta_{n,0}}\big(l_{\beta_{n,0}} h^T s_{\beta_{n,0}}\big) - h^T \dot{g}(\beta_{n,0}) = o(1), \tag{22}$$

then if $\beta_{n,0} + I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0})/\sqrt{n} \in B(\beta_{n,0}, c/\sqrt{n})$, it holds that

$$V_{\beta_{n,0}} \geq \dot{g}(\beta_{n,0})^T I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0}) + o(1).$$

Theorem 9 in conjunction with Lemma 3 gives the result summarized in Corollary 2 below.

COROLLARY 2. *Suppose that conditions of Theorem 9 and condition (22) are satisfied and that $\beta_{n,0} + I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0})/\sqrt{n} \in B(\beta_{n,0}, c/\sqrt{n})$. Then*

$$\sqrt{n}\big(T_n - g(\beta_{n,0} + h/\sqrt{n})\big) / V_{\beta_{n,0}}^{1/2} \overset{\beta_{n,0}+h/\sqrt{n}}{\rightsquigarrow} \mathcal{N}(0,1), \tag{23}$$

where

$$V_{\beta_{n,0}} \geq \dot{g}(\beta_{n,0})^T I_{\beta_{n,0}}^{-1} \dot{g}(\beta_{n,0}) + o(1).$$

The corollary implies that asymptotic efficiency is attained by an estimator which is asymptotically linear with an influence function $l_{\beta_{n,0}} = \dot{g}(\beta_{n,0})^T I_{\beta_{n,0}}^{-1} s_{\beta_{n,0}}$, provided that it satisfies condition (22).

We have already shown how these results can be applied to the linear regression setting in Section 8.5. We remark that the result of Theorem 9 is not directly applicable to Gaussian graphical models, where the unknown parameter has overall sparsity ps, where $s = o(\sqrt{n}/\log p)$.

REMARK 3. The sparsity condition $d_n = o(n^{1/3})$ arises when considering Taylor expansion of the log-likelihood for *general* models. Hence, when there is some special structure in the log-likelihood function, weaker sparsity conditions

might be possible. For instance, for linear regression setting, the Hessian of the log-likelihood does not depend on the unknown parameter $\beta_{n,0}$, hence in that case by inspection of the likelihood expansion in the proof of Theorem 9, we see that the condition $d_n = o(\sqrt{n/\log p})$ is sufficient.

11. Conclusions. In this paper, we have proposed a framework for studying asymptotic efficiency in high-dimensional models. We adopted a semiparametric point of view: we concentrated on one-dimensional functions of a high-dimensional parameter for which the lower bounds were derived. The semiparametric efficiency bounds we obtained correspond to the efficiency bounds for parametric models. However, the treatment for high-dimensional models required more elaborate analysis due to the models changing with n and assumed sparsity of the model.

We further considered construction of estimators attaining the lower bounds. We showed that indeed construction of asymptotically efficient estimator is possible: a de-sparsified estimator in linear regression and Gaussian graphical models is asymptotically efficient for estimation of certain simple functionals. Our analysis identified the theoretical conditions on the parameter sparsity and further conditions on the model under which asymptotic efficiency may be shown.

Comparison of the two approaches. The analysis was done in two ways: in the spirit of asymptotic Cramér–Rao bounds and Le Cam's bounds [van der Vaart (1998)]. These are strongly related: both define a restricted set of estimators which are in some sense asymptotically unbiased and claim lower bounds for any estimator in this class.

However, the two lines of work are not directly comparable as they are different results under different assumptions. Le Cam's bounds give a lower bound on *asymptotic variance*, while the Cramér–Rao bounds give a bound on the *variance* of an estimator. We formulated Le Cam's approach for a general sparse model, while the Cramér–Rao bounds were only considered for the linear regression and Gaussian graphical models. Apart from this, the main results arising from the two approaches also present some differences in the assumptions. For the Le Cam's-type results, we assumed a stronger sparsity condition of order $d_n = o(n^{1/3}/\log p)$ because of the Taylor expansion of the likelihood. However, for the linear regression setting, the sparsity condition can be improved to $d_n = o(\sqrt{n}/\log p)$, which is the same as in the Cramér–Rao bounds. For Gaussian graphical models, Le Cam's approach as formulated in this paper cannot be directly used, unlike the approach through the Cramér–Rao bounds.

Extensions. Our results on upper bounds are presented for the case when the parameter of interest is a single entry of the high-dimensional parameter or a linear combination with, for example, bounded ℓ_1-norm. It is interesting to note some

relations to literature on minimax rates. One question is whether asymptotic efficiency can be attained, for example, for estimation of linear functionals in linear regression when the linear combination ξ is sparse. Our results needed that $\|\xi\|_1$ remains bounded. Some recent works on high-dimensional models further consider estimation of more complicated, nonsparse functionals [in linear regression Cai and Guo (2017), for Gaussian sequence models Collier, Comminges and Tsybakov (2015)]. These results are however of a different nature. Consider for instance estimation of $\sum_{i=1}^p \beta_i^0$ in high-dimensional linear regression. In this case, the parametric rate cannot be achieved [Cai and Guo (2017)], and thus it remains unclear what can be said about "asymptotic efficiency".

Furthermore, we have treated the case of a one-dimensional parameter of interest, though the analysis might be extended to settings when the parameter of interest is higher dimensional (of a fixed dimension). Finally, our analysis considered particular examples of de-sparsified estimators, however, other estimators which are in some sense equivalent to these de-sparsified estimators are applicable.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric efficiency bounds for high-dimensional models" (DOI: 10.1214/17-AOS1622SUPP; .pdf). The supplementary material contains proofs.

REFERENCES

BELLEC, P. and TSYBAKOV, A. B. (2016). Bounds on the prediction error of penalized least squares estimators with convex penalty. Available at arXiv:1609.06675.

BICKEL, P. J., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, Berlin.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Ann. Rev. Econ.* **7** 649–688.

COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. Available at arXiv:1502.00665.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9** 432–441.

GAO, C., MA, Z. and ZHOU, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. MR3718162

JANKOVÁ, J. and VAN DE GEER, S. (2014). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. MR3354336

JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality for high-dimensional precision matrix estimation. *TEST* **26** 143–162.

JANKOVÁ, J. and VAN DE GEER, S. (2018). Supplement to "Semiparametric efficiency bounds for high-dimensional models." DOI:10.1214/17-AOS1622SUPP.

JAVANMARD, A. and MONTANARI, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909.

JAVANMARD, A. and MONTANARI, A. (2014b). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

JAVANMARD, A. and MONTANARI, A. (2015). De-biasing the Lasso: Optimal sample size for gaussian designs. Available at arxiv:1508.02757.

KNIGHT, K. and FU, W. (2000). Asymptotics for Lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

VAN DE GEER, S. (2016). *Estimation and Testing under Sparsity*. Springer, Berlin. MR3526202

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

ZHANG, C. H. and ZHANG, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *J. Roy. Statist. Soc. Ser. B* **76** 217–242. MR3153940

SEMINAR FÜR STATISTIK
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: jankova@stat.math.ethz.ch
geer@stat.math.ethz.ch