The Annals of Statistics
2020, Vol. 48, No. 5, 2791–2823
https://doi.org/10.1214/19-AOS1906

WHICH BRIDGE ESTIMATOR IS THE BEST FOR VARIABLE SELECTION?

BY SHUAIWEN WANG[1,*], HAOLEI WENG[2] AND ARIAN MALEKI[1,†]

[1]*Department of Statistics, Columbia University,* [*]*sw2853@columbia.edu;* [†]*arian@stat.columbia.edu*
[2]*Department of Statistics and Probability, Michigan State University, wenghaol@msu.edu*

We study the problem of variable selection for linear models under the high-dimensional asymptotic setting, where the number of observations n grows at the same rate as the number of predictors p. We consider two-stage variable selection techniques (TVS) in which the first stage uses bridge estimators to obtain an estimate of the regression coefficients, and the second stage simply thresholds this estimate to select the "important" predictors. The asymptotic false discovery proportion (AFDP) and true positive proportion (ATPP) of these TVS are evaluated. We prove that for a fixed ATPP, in order to obtain a smaller AFDP, one should pick a bridge estimator with smaller asymptotic mean square error in the first stage of TVS. Based on such principled discovery, we present a sharp comparison of different TVS, via an in-depth investigation of the estimation properties of bridge estimators. Rather than "orderwise" error bounds with loose constants, our analysis focuses on precise error characterization. Various interesting signal-to-noise ratio and sparsity settings are studied. Our results offer new and thorough insights into high-dimensional variable selection. For instance, we prove that a TVS with Ridge in its first stage outperforms TVS with other bridge estimators in large noise settings; two-stage LASSO becomes inferior when the signal is rare and weak. As a by-product, we show that two-stage methods outperform some standard variable selection techniques, such as LASSO and Sure Independence Screening, under certain conditions.

1. Introduction.

1.1. *Motivation and problem statement.* Although linear models can be traced back to two hundred years ago, they keep shining in the modern statistical research. A problem of major interest in this literature is *variable selection*. Consider the linear regression model

$$y = X\beta + w,$$

with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and $w \in \mathbb{R}^n$. Suppose only a few elements of β are nonzero. The problem of variable selection is to find these nonzero locations of β. Motivated by the concerns about the instability and high computational cost of classical variable selection techniques, such as best subset selection and stepwise selection, Tibshirani proposed LASSO [47] to perform parameter estimation and variable selection simultaneously. The LASSO estimate is given by

$$\hat{\beta}(1,\lambda) := \operatorname*{argmin}_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1.1}$$

where $\lambda \in (0,\infty)$ is the tuning parameter, and $\|\cdot\|_1$ is the ℓ_1 norm. The regularization term $\|\beta\|_1$ stabilizes the variable selection process while the convex formulation of (1.1) reduces the computational cost.

Received March 2019.
MSC2020 subject classifications. 62J05, 62J07.
Key words and phrases. Variable selection, high dimension, bridge regression, two-stage methods, false discovery proportion, true positive proportion, rare signal, large noise, large sample, debiasing.

Compared to LASSO, other convex regularizers such as $\|\beta\|_2^2$ imposes larger penalty to large components of β. Hence, their estimates might be more stable than LASSO. Even though the solutions of many of these regularizers are not sparse (and thus not automatically perform variable selection), we may threshold their estimates to select variables. This observation leads us to the following questions: can such two-stage methods with other regularizers outperform LASSO in variable selection? If so, which regularizer should be used in the first stage? The goal of this paper is to address these questions. In particular, we study the performances of the two-stage variable selection (TVS) techniques mentioned above, with the first stage based on the class of bridge estimators [23]:

$$\hat{\beta}(q,\lambda) := \operatorname*{argmin}_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q, \tag{1.2}$$

where $\|\beta\|_q^q = \sum_i |\beta_i|^q$ with $q \geq 1$. Our variable selection technique takes $\hat{\beta}(q,\lambda)$ and returns the sparse estimate $\bar{\beta}(q,\lambda,s)$ defined as follows:

$$\bar{\beta}(q,\lambda,s) = \eta_0\big(\hat{\beta}(q,\lambda); s^2/2\big),$$

where $\eta_0(u;\chi) = u\mathbb{1}_{\{|u|\geq\sqrt{2\chi}\}}$ denotes the hard threshold function and it operates on a vector in a componentwise manner. The nonzero elements of $\bar{\beta}(q,\lambda,s)$ are used as selected variables. In this paper, we give a thorough investigation of such TVS techniques under the asymptotic setting $n/p \to \delta \in (0,\infty)$. Specifically, the following fundamental questions are addressed:

> *Which value of q offers the best variable selection performance? Does* LASSO *outperform the two-stage methods based on other bridge estimators? What is the impact of the signal-to-noise ratio (SNR) and the sparisty level on the optimal choice of q?*

1.2. *Our contribution.* Different from most of the previous works, our study adopts a high-dimensional regime in which variable selection consistency is unattainable. Under our asymptotic framework, we are able to obtain a sharp characterization of the variable selection "error" (we will clarify our definition of this error in Section 2). The *asymptotically exact expressions* we derive for the error open a new way for comparing the aforementioned variable selection techniques accurately.

It turns out that the variable selection performance of TVS is closely connected with the estimation quality of the bridge estimator in the first stage; a bridge estimator with a smaller asymptotic mean square error (AMSE) in the first stage offers a better variable selection performance in the TVS. This novel observation enables us to connect and translate the study of TVS to the comparison of the estimation accuracy of different bridge estimators.

Due to the nature of different ℓ_q regularizers, each bridge estimator has its own strength under different model settings. To clarify the strength and weakness of different bridge estimators, we study and compare their AMSE under the following important scenarios: (i) rare signal scenario; (ii) large noise scenario; (iii) large sample scenario. For the first two, new phenomena are discovered: the Ridge estimator is optimal among all the bridge estimators in large noise settings; in the setting of rare signals, LASSO achieves the best performance when the signal strength exceeds a certain level. However, for signals below that level, other bridge estimators may outperform LASSO. In the large sample scenario, we connect our analyses with the fruits of the classical low-dimensional asymptotic studies. We will provide new comparison results not available in classical asymptotic analyses of bridge estimators.

In summary, our studies reveal the intricate impact of the combination of SNR and sparsity level on the estimation of the coefficients. New insights into high-dimensional variable selection are discovered. We present our contributions more formally in Section 3.

1.3. *Related work.* The literature on variable selection is very rich. Hence, the related works we choose to discuss can only be illustrative rather than exhaustive.

Traditional methods of variable selection include best subset selection and stepwise procedures. Best subset selection suffers from high computational complexity and high variance. The greedy nature of stepwise procedures reduces the computational complexity, but limits the number of models that are checked by such procedures; see [37] for a comprehensive treatment of classical subset selection. To overcome these limitations, [47] proposed the LASSO that aims to perform variable selection and parameter estimation simultaneously. Both the variable selection and estimation performance of LASSO have been studied extensively in the past decade. It has been justified in the works of [35, 57, 59] that a type of "irrepresentable condition" is almost sufficient and necessary to guarantee sign consistency for the LASSO. Later [48] established sharp conditions under which LASSO can perform a consistent variable selection. One implication of [48] that is relevant to our paper is that, consistent variable selection is impossible under the linear asymptotic regime[1] that we consider in this paper. This result is consistent with that of [44] and our paper. Hence, we should expect that both the true positive proportion (TPP) and false discovery proportion (FDP) play a major role in our analyses and comparisons. It is worth mentioning that the rate of convergence for variable selection under Hamming loss has been studied in a sequel of works [7, 24, 29, 30].

Since LASSO requires strong conditions for variable selection consistency, several authors have considered a few variants, such as adaptive LASSO [62] and thresholded LASSO [36]. Thresholded LASSO is an instance of two-stage variable selection schemes we study in this paper. Meinshausen and Yu [36] proved that thresholded LASSO offers a variable selection consistency under weaker conditions than the irrepresentable condition required by LASSO. As we will see later, even the thresholded LASSO does not obtain variable selection consistency under the asymptotic framework of this paper. However, we will show that it outperforms the LASSO in variable selection. Other authors have also studied two-step or even multistep variable selection schemes in the hope of weakening the required conditions [33, 54, 58, 61]. Note that none of these methods provide consistent variable selection under the linear asymptotic setting we consider in this paper. Study and comparison of these other schemes under our asymptotic setting is an interesting open problem for future research.

A more delicate study of the LASSO estimator and more generally the bridge estimators is necessary for an accurate analysis of two-stage methods under the linear asymptotic regime. Our analysis relies on the recent results in the study of bridge estimators [4, 5, 15, 16, 34, 44, 55]. These papers use the platform offered by approximate message passing (AMP) to characterize sharp asymptotic properties. In particular, the most relevant work to our paper is [44] which studies the solution path of LASSO through the trade-off diagram of the asymptotic FDP and TPP. The present paper makes further steps in the analysis of bridge estimator based two-stage methods under various interesting signal-to-noise ratio settings that have not been considered in [44].

Another line of two-stage methods is the idea of screening [9, 21, 28, 53]. For instance, in [21] a preliminary estimate of the jth regression coefficient is obtained by regressing y on only the jth predictor. Then a hard threshold function is applied to all the estimates to infer the location of the nonzero coefficients. As we will discuss in Section 4.2, this approach is a special form of our TVS with a debiasing performed in the first stage, and hence our variable selection technique under appropriate tuning outperforms Sure Independence Screening

[1]Throughout the paper, the linear asymptotic is referred to the asymptotic setting with (a) and (b) in Definition 2.1 satisfied. Typically, in this case, we have n, p and the number of nonzero coefficients k go to infinity proportionally.

of [21]. Compared to Sure Independence Screening, the work of [53] uses more complicated estimators in the first stage, which is more aligned to our approach. However, [53] requires data splitting. While this data splitting achieves certain theoretical improvement, in practice (especially in high dimensions) this may degrade the performance of a variable selection technique. In this paper, we avoid data splitting. We should also mention that two-stage or multistage methods (that have a thresholding step) are also popular for estimation purposes; see, for instance, [56]. Due to limited space, the current paper will be focused on variable selection and not discuss the estimation performance of TVS. However, an accurate analysis of multistage estimation techniques is an interesting problem to study.

Finally, there exists one stream of research with emphasis on the derivation of sufficient and necessary conditions for variable selection consistency under different types of restrictions on the model parameters [1, 10, 22, 39, 41, 49, 52]. These works typically assume that all the entries of the design matrix X and error vector w are independent zero-mean Gaussian, with which they are able to obtain accurate information theoretical thresholds and phase transition for exact support recovery of the coefficients β. We refer to [39] for a detailed discussion of such results. As will be shown shortly in Section 2, we make the same assumption on the design X, but allow much weaker conditions on the error term w. More importantly, we push the analysis one step further by analyzing a class of TVS when exact recovery is impossible information theoretically.

2. Our asymptotic framework and some preliminaries.

2.1. *Asymptotic framework.* In this section, we review the asymptotic framework under which our studies are performed. We start with the definition of a converging sequence adapted from [5].

DEFINITION 2.1. The sequence of instances $\{\beta(p), w(p), X(p)\}_{p\in\mathbb{N}}$, indexed by p, is said to be a standard converging sequence if:

(a) $n = n(p)$ such that $\frac{n}{p} \to \delta \in (0, \infty)$.

(b) The empirical distribution of the entries of $\beta(p)$ converges weakly to a probability measure p_B on $\mathbb{R}$ with finite second moment. Further, $\frac{1}{p}\sum_{i=1}^{p} \beta_i(p)^2$ converges to the second moment of p_B; and $\frac{1}{p}\sum_{i=1}^{p} \mathbb{I}(\beta_i(p) = 0) \to p_B(\{0\})$.

(c) The empirical distribution of the entries of $w(p)$ converges weakly to a zero-mean distribution with variance σ^2. Furthermore, $\frac{1}{n}\sum_{i=1}^{n} w_i(p)^2 \to \sigma^2$.

(d) $X_{ij}(p) \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$.

The asymptotic scaling $n/p \to \delta$ specified in Condition (a) was proposed by Huber in 1973 [27], and has become one of the most popular asymptotic settings especially for studying problems with moderately large dimensions [11, 13, 18, 19, 45, 46]. Regarding Condition (b), suppose the entries of $\beta(p)$ form a stationary ergodic sequence with marginal distribution determined by some probability measure p_B. According to Birkhoff's ergodic theorem, it is clear that Condition (b) will hold almost surely. Thus Condition (b) can be considered as a weaker notion of this Bayesian set-up. Similar interpretation works for Condition (c). Regarding Condition (d), as discussed in Section 1.3, many related works assume it as well. Moreover, we would like to point out that there are a lot of empirical and a few theoretical studies revealing the universal behavior of i.i.d. Gaussian design matrices over a wider class of distributions; see [3] and references therein. Hence, the Gaussianity of the design does not play a critical role in our final results. The numerical studies presented in Section 5.7 confirm this claim. The independence assumption of the design entries is critical for our

analysis. Given that our analyses for i.i.d. matrices are already complicated, and the obtained results are highly nontrivial (as will be seen in Section 3), we leave the study of general design matrices for a future research. However, the numerical studies performed in Section 5.7 imply that the main conclusions of our paper are valid even when the design matrix is correlated.

In the rest of the paper, we assume the vector of regression coefficients β is sparse. More specifically, we assume $p_B = (1-\epsilon)\delta_0 + \epsilon p_G$, where δ_0 denotes a point mass at 0 and p_G is a probability measure without any point mass at 0. Accordingly, the mixture proportion ϵ represents the sparsity level of $\beta(p)$ in the converging sequence. Throughout the paper, B and G will be used as random variables with distribution specified by p_B and p_G, respectively. Z represents a standard normal random variable. Subscripts like i attached to a vector are used to denote its ith component. The *asymptotic mean square error* (AMSE) of the bridge estimator $\hat{\beta}(q,\lambda)$ is defined as the almost sure limit

$$\mathrm{AMSE}(q,\lambda) \triangleq \lim_{p\to\infty} \frac{1}{p}\|\hat{\beta}(q,\lambda)-\beta\|_2^2. \tag{2.1}$$

According to [4, 55], $\mathrm{AMSE}(q,\lambda)$ is well defined for $q \in [1,\infty)$ and $\lambda > 0$. In this paper, one of our focuses will be on bridge estimators with optimal tuning λ_q^* defined as

$$\lambda_q^* \triangleq \operatorname*{argmin}_{\lambda>0} \mathrm{AMSE}(q,\lambda).$$

Further, we denote the thresholded estimators as

$$\bar{\beta}(q,\lambda,s) = \eta_0\big(\hat{\beta}(q,\lambda); s^2/2\big) = \hat{\beta}(q,\lambda)\mathbb{1}_{\{|\hat{\beta}(q,\lambda)|\geq s\}}.$$

Since under our asymptotic setting the exact recovery of the nonzero locations of β is impossible [43, 48], we expect to observe both false positives and false negatives. Hence, for a given sparse estimator $\hat{\beta}$, we follow [44] and measure its variable selection performance by the false discovery proportion (FDP) and true positive proportion (TPP), defined as

$$\mathrm{FDP}(\hat{\beta}) = \frac{\#\{i:\hat{\beta}_i\neq 0, \beta_i = 0\}}{\#\{i:\hat{\beta}_i\neq 0\}}, \qquad \mathrm{TPP}(\hat{\beta}) = \frac{\#\{i:\hat{\beta}_i\neq 0, \beta_i \neq 0\}}{\#\{i:\beta_i\neq 0\}}.$$

In particular, our study will focus on the asymptotic version of FDP and TPP for the LASSO estimate $\hat{\beta}(1,\lambda)$ and thresholded estimators $\bar{\beta}(q,\lambda,s)$. We define (the limits are in almost surely senses)

$$\mathrm{AFDP}(1,\lambda) = \lim_{p\to\infty} \mathrm{FDP}\big(\hat{\beta}(1,\lambda)\big), \qquad \mathrm{AFDP}(q,\lambda,s) = \lim_{p\to\infty} \mathrm{FDP}\big(\bar{\beta}(q,\lambda,s)\big).$$

Similar definitions are used for $\mathrm{ATPP}(1,\lambda)$ and $\mathrm{ATPP}(q,\lambda,s)$. The following result adapted from [6] characterizes the AFDP and ATPP for LASSO.

LEMMA 2.1. *For any given $\lambda > 0$, almost surely*

$$\begin{aligned}\mathrm{AFDP}(1,\lambda) &= \frac{(1-\epsilon)\mathbb{P}(|Z|>\alpha)}{(1-\epsilon)\mathbb{P}(|Z|>\alpha)+\epsilon\mathbb{P}(|G+\tau Z|>\alpha\tau)},\\ \mathrm{ATPP}(1,\lambda) &= \mathbb{P}\big(|G+\tau Z|>\alpha\tau\big),\end{aligned} \tag{2.2}$$

where (α,τ) is the unique solution to the following equations with $q=1$:

$$\tau^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\big(\eta_q\big(B+\tau Z;\alpha\tau^{2-q}\big)-B\big)^2, \tag{2.3}$$

$$\lambda = \alpha\tau^{2-q}\bigg(1-\frac{1}{\delta}\mathbb{E}\eta_q'\big(B+\tau Z;\alpha\tau^{2-q}\big)\bigg), \tag{2.4}$$

with $\eta_q(\cdot;\cdot)$ *being the proximal operator defined as*

$$\eta_q(u;\chi) = \operatorname*{argmin}_z \frac{1}{2}(u-z)^2 + \chi|z|^q,$$

and $\eta_q'(\cdot;\cdot)$ *being the derivative of* η_q *with respect to its first argument.*

The formulas in this lemma have been derived in terms of convergence in probability in [6]. The extension to almost sure convergence is straightforward and is hence skipped; see Appendix C.1 of [50] for more information. One of the main goals of this paper is to compare the performance of two-stage variable selection techniques with LASSO. In the next lemma, we derive the AFDP and ATPP of the thresholded estimate $\bar{\beta}(q,\lambda,s)$.

LEMMA 2.2. *For any given* $q \in [1,\infty)$, $\lambda > 0$, $s > 0$, *almost surely*

$$\text{AFDP}(q,\lambda,s) = \frac{(1-\epsilon)\mathbb{P}(\eta_q(|Z|;\alpha) > \frac{s}{\tau})}{(1-\epsilon)\mathbb{P}(\eta_q(|Z|;\alpha) > \frac{s}{\tau}) + \epsilon\mathbb{P}(|\eta_q(G+\tau Z;\alpha\tau^{2-q})| > s)},$$

$$\text{ATPP}(q,\lambda,s) = \mathbb{P}\big(\big|\eta_q\big(G+\tau Z;\alpha\tau^{2-q}\big)\big| > s\big),$$

(2.5)

where (α,τ) *is the unique solution of* (2.3) *and* (2.4).

The proof of this lemma is presented in Section C of the Supplementary Material [51].

3. Our main contribution.

3.1. *How to compare two variable selection schemes.* The main objective of this paper is to compare the performance of the TVS techniques under the asymptotic setting of Section 2. A natural way for performing this comparison is to set ATPP to a fixed value $\zeta \in [0,1]$ for different variable selection schemes and then compare their AFDPs.

The first challenge we face in such a comparison is that the TVS may have many different ways for setting ATPP to ζ. If $q > 1$, Lemma 2.2 shows that for every given value of the regularization parameter λ, we can set s (the threshold parameter) in a way that it returns the right level of ATPP. Which of these parameter choices should be used when we compare a TVS with another variable selection technique, such as LASSO? Despite the fact that different choices of (λ,s) achieve the same ATPP level ζ, they may result in different values of AFDP. Thus for fair comparison we pick the one that minimizes AFDP. The next theorem explains how this optimal pair can be found.

THEOREM 3.1. *Consider* $q \in (1,\infty)$. *Given an* ATPP *level* $\zeta \in [0,1]$, *for every value of* $\lambda > 0$ *there exists* $s = s(\lambda,\zeta)$ *such that* $\text{ATPP}(q,\lambda,s) = \zeta$. *Furthermore, the value of* λ *that minimizes* $\text{AFDP}(q,\lambda,s(\lambda,\zeta))$ *also minimizes* $\text{AMSE}(q,\lambda)$.

The proof of this theorem can be found in Section D.1 of the Supplementary Material [51]. Before discussing the implications of this theorem, we state a similar result for LASSO.

THEOREM 3.2. *For any* $\zeta \in [0, \text{ATPP}(1,\lambda_1^*)]$, *there exists at least one* λ *s.t.* $\text{ATPP}(1,\lambda) = \zeta$. *Further, there exists a unique* $s = s_\zeta$ *such that* $\text{ATPP}(1,\lambda_1^*,s) = \zeta$. *There may also exist other* (λ,s) *s.t.* $\text{ATPP}(1,\lambda,s) = \zeta$. *Among all these estimators, the one that offers the minimal* AFDP *is* $\bar{\beta}(1,\lambda_1^*,s_\zeta)$, *that is, the two-stage* LASSO *with the optimal tuning value* $\lambda = \lambda_1^*$.

The proof of this theorem is shown in Section D.2 of the Supplementary Material [51]. There are a couple of points we would like to emphasize here:

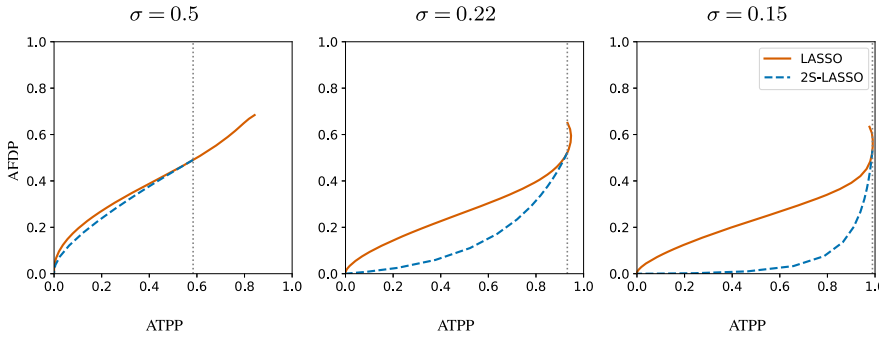

FIG. 1. *Comparison of* AFDP–ATPP *curve between* LASSO *and two-stage* LASSO. *Here, we pick the setting* $\delta = 0.8$, $\epsilon = 0.3$, $\sigma \in \{0.5, 0.22, 0.15\}$, $p_G = \delta_1$. *For two-stage LASSO, we use optimal tuning* λ_1^* *in the first stage. All the curves are calculated based on equations* (2.2) *and* (2.5). *The gray dotted line is the upper bound of* ATPP *that the two-stage* LASSO *can reach. Notice that even for* LASSO, *there is an upper bound which it cannot exceed.*

(i) Consider a TVS technique. According to Theorems 3.1 and 3.2, for $q \in (1, \infty)$, the optimal choice of λ does not depend on the ATPP level ζ we are interested in. Even for $q = 1$, the optimal choice of λ is independent of ζ in a large range of ATPPs. It is the optimal tuning λ_q^* for AMSE.

(ii) An implication of Theorem 3.2 is that, for a wide range of ζ, a second thresholding step helps with the variable selection of LASSO. Figure 1 compares the AFDP–ATPP curve of LASSO with that of the two-stage LASSO. As is clear in this figure, when SNR is higher, the gap between the performance of two-stage LASSO and LASSO becomes larger. We should emphasize that the ATPP level of the two-stage LASSO (with optimal tuning) cannot exceed that of $\hat{\beta}(1, \lambda_1^*)$. We discuss *debiasing* to resolve this issue in Section 4.

(iii) Theorems 3.1 and 3.2 do not explain how λ_q^* can be estimated in practice. This issue will be discussed in Section 5. But in a nutshell, any approach that optimizes λ for minimizing the out-of-sample prediction error works well.

REMARK 3.1. Theorems 3.1 and 3.2 prove that the optimal way to use two-stage variable selection is to set $\lambda = \lambda_q^*$ for the regularization parameter in the first stage. It is important to point out that λ_q^* minimizes $\mathrm{AMSE}(q, \lambda)$, and thus is the optimal tuning for parameter estimation. Therefore, the optimal tuning of the regularization parameter in bridge regression is the same for estimation and variable selection.

In the rest of the paper, we will use the notation $s_q^*(\zeta)$ for the value of threshold that satisfies $\mathrm{ATPP}(q, \lambda_q^*, s_q^*(\zeta)) = \zeta$.

3.2. *The best bridge estimator for variable selection.*

3.2.1. *Summary.* The two theorems we presented in the last section pave our way in addressing the question we raised in Section 1.1, that is, finding the best bridge estimator based TVS technique. Consider $q_1, q_2 \in [1, \infty)$. We would like to compare $\mathrm{AFDP}(q_1, \lambda_{q_1}^*, s_{q_1}^*(\zeta))$ and $\mathrm{AFDP}(q_2, \lambda_{q_2}^*, s_{q_2}^*(\zeta))$. The following corollary of Theorems 3.1 and 3.2 shows the equivalence of the variable selection and estimation performance of bridge estimators.

COROLLARY 3.1. *Let* $q_1, q_2 \geq 1$. *If* $\mathrm{AMSE}(q_1, \lambda_{q_1}^*) < \mathrm{AMSE}(q_2, \lambda_{q_2}^*)$, *then for every* $\zeta \in [0, 1]$

$$\mathrm{AFDP}\big(q_1, \lambda_{q_1}^*, s_{q_1}^*(\zeta)\big) \leq \mathrm{AFDP}\big(q_2, \lambda_{q_2}^*, s_{q_2}^*(\zeta)\big).$$

The proof of this result is presented in Section D.3 of the Supplementary Material [51]. According to Corollary 3.1, in order to see which two-stage method is better, we can compare their AMSE under optimal tuning λ_q^*. Such AMSE is given by (see Theorem B.1 and Lemma B.1 in the Supplementary Material [51])

$$\mathrm{AMSE}(q, \lambda_q^*) = \mathbb{E}\big(\eta_q\big(B + \tau_* Z; \alpha_* \tau_*^{2-q}\big) - B\big)^2,$$

where τ_* and α_* satisfy (2.3) and (2.4) with $\lambda = \lambda_q^*$.

The stage is finally set for comparing different two-stage variable selection techniques. Note that in the calculation of $\mathrm{AMSE}(q, \lambda_q^*)$, the values of α_* and τ_* are required and can only be calculated through the fixed-point equations (2.3) and (2.4). Therefore, we have no access to an explicit formula for $\mathrm{AMSE}(q, \lambda_q^*)$. Furthermore, AMSE depends on many factors including δ, σ and p_B. This poses an extra challenge to completely evaluate and compare AMSE for different values of q. To address these issues, we focus on a few regimes that researchers have found useful in applications, and develop techniques to obtain explicit and accurate expressions for $\mathrm{AMSE}(q, \lambda_q^*)$. These sharp results enable an accurate comparison among different TVS methods in each setting. The regimes we will consider are the following:

(i) Nearly black objects or rare signals: In this regime, ϵ is assumed to be small. In other words, there are very few nonzero coefficients that need to be detected. This model is called nearly black objects [14] or rare signals [12]. Intuitively speaking, it is also equivalent to the models considered in many other papers in which the sparsity level is assumed to be much smaller than the number of features; see, for instance, [35, 57, 59] and the references therein. We will allow the signal strength to vary with respect to ϵ. It turns out that the rate of signal strength affects the choice of optimal bridge estimator.

(ii) Low SNR: In this model, σ is considered to be large. This assumption is accurate in many social and medical studies. For more information, the reader may refer to [25]. To explain the effect of SNR on the best choice of q, we will also mention a result for high SNR. Such assumption is also standard in the engineering applications, where the quality of measurements is carefully controlled. The analysis that is performed under the low noise setting is often called phase transition analysis, noise sensitivity analysis, or nearly exact recovery; see, for instance, [16, 17, 40].

(iii) Large sample regime: In this regime, the per-feature sample size δ is large. This regime, as will be seen later, is closely related to the classical asymptotic regime $n/p \to \infty$, and is appropriate for traditional applied statistical problems; see, for instance, [31] for the asymptotic analysis of bridge estimators.

3.2.2. *Analysis of* AMSE *for nearly black objects.* As discussed in the preceding section, the formulas of AMSE are implicit and depend on δ, σ and p_B in a complicated way. The goal of this section is to obtain explicit and accurate expressions for $\mathrm{AMSE}(q, \lambda_q^*)$ when ϵ is small (i.e., the signal is very sparse). Toward this goal, a critical issue as made in, for example, [14] for the case of orthogonal design, is that the strength of the signal affects the performance of each estimator. Hence, in our analysis we let the strength of the signal vary with ϵ. This generalization requires an extra notation we introduce here. Recall G is the random variable with probability measure p_G, which determines the values of the nonzero entries of β. Define

$$b_\epsilon = \sqrt{\mathbb{E}G^2}, \quad \tilde{G} = G/b_\epsilon.$$

Under this parameterization, $\mathbb{E}\tilde{G}^2 = 1$ and b_ϵ represents the (average) magnitude of each nonzero coefficient. We refer to b_ϵ as the signal strength and will allow it to change with the sparsity level ϵ. Our first theorem characterizes the behavior of bridge estimators for $q > 1$ and small values of ϵ.

THEOREM 3.3. *Suppose that* $b_\epsilon \to \infty$ *and* $b_\epsilon = O(1/\sqrt{\epsilon})$.[2] *For* $q > 1$, *we have*:

- *If* $b_\epsilon = \omega(\epsilon^{\frac{1-q}{2}})$, *then*
$$\lim_{\epsilon\to 0} \epsilon^{-\frac{1}{q}} b_\epsilon^{-\frac{2(q-1)}{q}} \operatorname{AMSE}(q, \lambda_q^*) = q(q-1)^{\frac{1}{q}-1}\sigma^{\frac{2}{q}}\big[\mathbb{E}|Z|^{\frac{2}{q-1}}\big]^{\frac{q-1}{q}}\big[\mathbb{E}|\tilde{G}|^{2q-2}\big]^{\frac{1}{q}}.$$
- *If* $b_\epsilon = o(\epsilon^{\frac{1-q}{2}})$, *then* $\lim_{\epsilon\to 0} \epsilon^{-1} b_\epsilon^{-2} \operatorname{AMSE}(q, \lambda_q^*) = 1$.
- *If* $\lim_{\epsilon\to 0} b_\epsilon \epsilon^{\frac{q-1}{2}} = c_r \in (0, \infty)$, *then*
$$\lim_{\epsilon\to 0} \epsilon^{-\frac{1}{q}} b_\epsilon^{-\frac{2(q-1)}{q}} \operatorname{AMSE}(q, \lambda_q^*) = \min_C h(C),$$
where $h : \mathbb{R}^+ \to \mathbb{R}$ *and* $h(C) \triangleq (Cq)^{-\frac{2}{q-1}}\sigma^2\mathbb{E}|Z|^{\frac{2}{q-1}} + \mathbb{E}(\eta_q(c_r\tilde{G}; C\sigma^{2-q}) - c_r\tilde{G})^2$. *Furthermore, the minimizer of* $h(C)$ *is finite.*

We note that when $q > 2$, $b_\epsilon = o(\epsilon^{\frac{1-q}{2}})$ *always holds, hence only the second item applies. When* $q = 2$, *only the second and the third items apply.*

This theorem is proved in Section E of the Supplementary Material [51]. Before we interpret this result, we characterize $\operatorname{AMSE}(1, \lambda_1^*)$ in Theorem 3.4.

THEOREM 3.4. *Suppose that* $b_\epsilon \to \infty$ *and* $b_\epsilon = O(1/\sqrt{\epsilon})$. *We have*:

- *If* $b_\epsilon = \omega(\sqrt{\log \epsilon^{-1}})$, *then* $\lim_{\epsilon\to 0} \frac{\operatorname{AMSE}(1,\lambda_1^*)}{\epsilon \log \epsilon^{-1}} = 2\sigma^2$.
- *If* $b_\epsilon = o(\sqrt{\log \epsilon^{-1}})$, *then* $\lim_{\epsilon\to 0} \frac{\operatorname{AMSE}(1,\lambda_1^*)}{\epsilon b_\epsilon^2} = 1$.
- *If* $\frac{b_\epsilon}{\sqrt{2\log \epsilon^{-1}}} \to c \in (0, \infty)$, *then* $\lim_{\epsilon\to 0} \frac{\operatorname{AMSE}(1,\lambda_1^*)}{\epsilon \log \epsilon^{-1}} = \mathbb{E}(\eta_1(c\tilde{G}; \sigma) - c\tilde{G})^2$.

This theorem will be proved in Section F of the Supplementary Material [51]. There are a few points that we should emphasize about Theorems 3.3 and 3.4.

REMARK 3.2. First, let us discuss the assumptions of these two theorems. It is straightforward to show that with $b_\epsilon = \omega(1/\sqrt{\epsilon})$, the SNR per measurement goes to infinity. Such scenarios seem uncommon in applications, and for the sake of brevity we have only considered $b_\epsilon = O(1/\sqrt{\epsilon})$. Otherwise, the techniques we developed can be applied to higher SNR as well. Furthermore, we postpone the discussion about the case $b_\epsilon = O(1)$ to Theorem 3.5.

REMARK 3.3. The work of [14] has studied the problem of estimating an extremely sparse signal under the orthogonal design. The main goal of [14] is to obtain the minimax risk for the class of ϵ-sparse signals (similar to our model) without any constraint on the signals' power. They have shown that the approximately least favorable distribution has a point mass at $\Theta(\sqrt{\log(\epsilon^{-1})})$, and that LASSO achieves the minimax risk. Note that there are two major differences between Theorem 3.4 and the work of [14]: (i) our result is for nonorthogonal design, and (ii) we are not concerned with the minimax performance. In fact, we fix the power of the signal and obtain the asymptotic mean square error. This platform enables us to observe several delicate phenomena that are not observed in minimax settings. For instance, as is clear from Theorem 3.4, the rate of $\operatorname{AMSE}(1, \lambda_1^*)$ undergoes a transition at the signal

[2] O notation used here is the standard big-O notation. We will also use other standard asymptotic notation. If the reader is not familiar with these notation, he/she may refer to Section B.1 in the Supplementary Material [51].

strength level $\Theta(\sqrt{\log(\epsilon^{-1})})$. As we will discuss later, below this threshold, LASSO is not necessarily optimal. However, since the risk of the Bayes estimator and LASSO is maximized for $b_\epsilon = \Theta(\sqrt{\log(\epsilon^{-1})})$, this important information is missed in minimax analysis.

REMARK 3.4. Compared to other bridge estimators, the performance of LASSO is much less sensitive to the strength of the signal: $\mathrm{AMSE}(1, \lambda_1^*) \sim \epsilon \log \epsilon^{-1}$ as long as $b_\epsilon = \Omega(\sqrt{\log \epsilon^{-1}})$, while the order of $\mathrm{AMSE}(q, \lambda_q^*)$ continuously changes as b_ϵ varies.

Theorems 3.3 and 3.4 can be used for comparing different bridge estimators, as clarified in our next corollary.

COROLLARY 3.2. *Suppose that $b_\epsilon = \epsilon^{-\gamma}$ for $\gamma \in (0, 1/2]$. We have*:

- *If $q > 2\gamma + 1$, then* $\mathrm{AMSE}(q, \lambda_q^*) \sim \epsilon^{1-2\gamma}$.
- *If $1 < q \le 2\gamma + 1$, then* $\mathrm{AMSE}(q, \lambda_q^*) \sim \epsilon^{\frac{1-2\gamma(q-1)}{q}}$.
- *If $q = 1$, then* $\mathrm{AMSE}(q, \lambda_q^*) \sim \epsilon \log(\epsilon^{-1})$.

The above result implies that in a wide range of signal strength, $q = 1$ offers the smallest AMSE when the value of ϵ is very small. Consequently, according to Corollary 3.1, the two-stage LASSO provides the best variable selection performance. One can further confirm that the same conclusion continues to hold as long as $b_\epsilon = \omega(\sqrt{\log \epsilon^{-1}})$.

So far, we have seen that if the signal is reasonably strong, that is, $b_\epsilon = \omega(\sqrt{\log \epsilon^{-1}})$, then two-stage LASSO outperforms all the other variable selection techniques. However, once $b_\epsilon = O(\sqrt{\log \epsilon^{-1}})$, we can see that $\mathrm{AMSE}(q, \lambda_q^*) \sim \epsilon b_\epsilon^2$ for all $q \ge 1$. Hence, in order to provide a fair comparison, one should perform finer analyses and obtain a more accurate expression for AMSE. Our next result shows how this can be done.

THEOREM 3.5. *Consider $b_\epsilon = 1$, and hence $\tilde{G} = G$. Assume G is bounded from above. Then we have*

$$\text{(3.1)} \quad \textit{For } q = 1: \quad \mathrm{AMSE}(1, \lambda_1^*) = \epsilon \mathbb{E} G^2 + o(\epsilon^k) \quad \forall k \in \mathbb{N};$$

$$\text{(3.2)} \quad \textit{For } q > 1: \quad \mathrm{AMSE}(q, \lambda_q^*) = \epsilon \mathbb{E} G^2 - \epsilon^2 \frac{\mathbb{E}^2(|\frac{G}{\sigma} + Z|^{\frac{1}{q-1}} \operatorname{sgn}(\frac{G}{\sigma} + Z) G)}{\mathbb{E}|Z|^{\frac{2}{q-1}}} + o(\epsilon^2),$$

where $\operatorname{sgn}(\cdot)$ *denotes the sign of a random variable.*

The proof of this theorem is presented in Section G of the Supplementary Material [51]. The first interesting observation about this theorem is that the first dominant term of AMSE is the same for all bridge estimators. The second dominant term, on the other hand, is much smaller for $q = 1$ compared to the other values of q. Hence, LASSO is *suboptimal* in this setting. Accordingly, two-stage LASSO is outperformed by other TVS methods. However, as is clear from Theorem 3.5, we should not expect the bridge estimator with $q > 1$ to outperform LASSO by a large margin when ϵ is too small. In fact, the second dominant term is proportional to ϵ^2 (for $q > 1$), while the first dominant term is proportional to ϵ. Hence, the second dominant term is expected to become important for moderately small values of ϵ. In such cases, we expect $q > 1$ to offer more significant improvements. Regarding the optimal choice of q, it is determined by the constant of the second-order term in (3.2). As is shown in Figure 2, while the optimal value of q is case-dependent, it gets closer to 1 as the signal strength increases. This observation is consistent with the message delivered by Theorems 3.3 and 3.4.

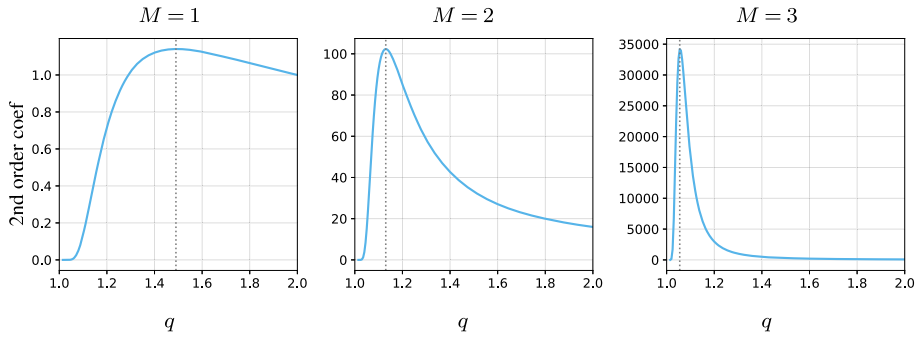

FIG. 2. *The constant coefficient of the second order term in* (3.2). *We set* $G = M$ *with* $M = 1, 2, 3$, *respectively, and* $\sigma = 1$. *As the signal strength* M *increases, the optimal choice of* q *shifts toward* 1.

3.2.3. *Analysis of* AMSE *in large noise scenario*. This section aims to obtain explicit formulas for the optimal AMSE of bridge estimators in low SNR. This regime is particularly important, since in many social and medical studies, variable selection plays a key role and the SNR is low. The following theorem summarizes the main result of this section.

THEOREM 3.6. *As* $\sigma \to \infty$, *we have the following expansions of* $\mathrm{AMSE}(q, \lambda_q^*)$:

(i) *For* $q = 1$, *when* G *has a sub-Gaussian tail, we have*

$$\mathrm{AMSE}(1, \lambda_1^*) = \epsilon \mathbb{E}|G|^2 + o\big(e^{-\frac{C^2\sigma^2}{2}}\big), \tag{3.3}$$

where C *can be any positive number smaller than* C_0, *and* $C_0 > 0$ *is a constant only depending on* ϵ *and* G. *The explicit definition of* C_0 *can be found in the proof.*

(ii) *For* $1 < q \le 2$, *if all the moments of* G *are finite, then*

$$\mathrm{AMSE}(q, \lambda_q^*) = \epsilon \mathbb{E}|G|^2 - \frac{\epsilon^2 (\mathbb{E}|G|^2)^2 c_q}{\sigma^2} + o(\sigma^{-2}), \tag{3.4}$$

with $c_q = \dfrac{(\mathbb{E}|Z|^{\frac{2-q}{q-1}})^2}{(q-1)^2 \mathbb{E}|Z|^{\frac{2}{q-1}}}$.

(iii) *For* $q > 2$, *if* G *has sub-Gaussian tail, then* (3.4) *holds.*

We present our proofs in Section H of the Supplementary Material [51]. Figure 3 compares the accuracy of the first-order approximation and second-order approximation for moderate values of σ. As is clear, for $q \in (1, \infty)$, the second-order approximation provides an accurate approximation of $\mathrm{AMSE}(q, \lambda_q^*)$ for a wide range of σ. Moreover, the first-order approximation for $\mathrm{AMSE}(1, \lambda_1^*)$ is already accurate as can be justified by its exponentially small second-order term in (3.3).

According to this theorem, we can conclude that for sufficiently large σ, two-stage method with any $q > 1$ can outperform the two-stage LASSO. This is because while the first dominant term is the same for all the bridge estimators with $q \in [1, \infty)$, the second-order term for LASSO is exponentially smaller (in magnitude) than that of the other estimators. More interestingly, the following lemma shows that in fact $q = 2$ leads to the smallest AMSE in the large noise regime.

LEMMA 3.1. *The maximum of* c_q, *defined in Theorem* 3.6, *is achieved at* $q = 2$.

See Figure 4 for the plot of c_q.

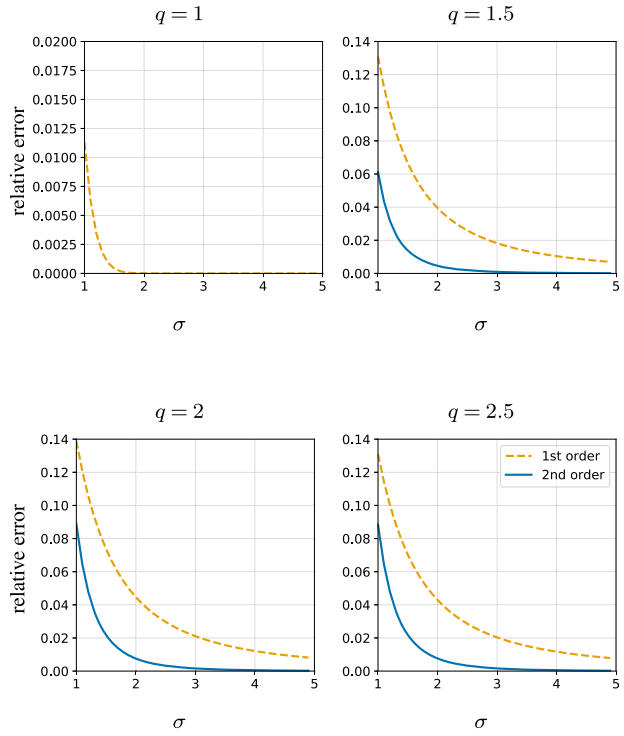

FIG. 3. *Absolute relative error of first-order and second-order approximations of* AMSE *under large noise scenario. In these four figures,* $p_B = (1-\epsilon)\delta_0 + \epsilon\delta_1$, $\delta = 0.4$, $\epsilon = 0.2$.

PROOF. A simple integration by part yields

$$\mathbb{E}|Z|^{\frac{2-q}{q-1}} = 2(q-1)\int_0^\infty z^{\frac{q}{q-1}}\phi(z)\,dz = (q-1)\mathbb{E}|Z|^{\frac{q}{q-1}}.$$

We can then apply Hölders's inequality to obtain

$$c_q = \frac{(\mathbb{E}|Z|^{\frac{q}{q-1}})^2}{\mathbb{E}|Z|^{\frac{2}{q-1}}} \le \frac{\mathbb{E}|Z|^{\frac{2}{q-1}}\mathbb{E}Z^2}{\mathbb{E}|Z|^{\frac{2}{q-1}}} = 1 = c_2. \qquad \square$$

Therefore, while the AMSE of all bridge estimators share the same first dominant term, Ridge offers the largest second dominant term (in magnitude), and hence the lowest AMSE. If we combine this result with Corollary 3.1, we conclude that in low SNR regime, two-stage

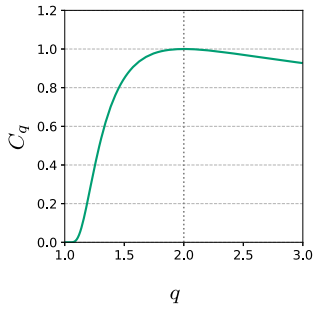

FIG. 4. *The constant* c_q *in Theorem* 3.6 *part* (*ii*). *The maximum is achieved at* $q = 2$.

Ridge obtains the best variable selection performance among TVS schemes with their first stage picked from the class of bridge estimators.

A comparison of this result with that for the high SNR derived in [55] clarifies the impact of SNR on the best choice of q.

THEOREM 3.7. *Assume* $\epsilon \in (0,1)$. *As* $\sigma \to 0$, *we have the following expansions of* $\mathrm{AMSE}(q, \lambda_q^*)$ *in terms of* σ:

(i) *For* $q = 1$, *if* $\mathbb{P}(|G| \geq \mu) = 1$ *for some* $\mu > 0$, $\delta > M_1(\epsilon)$, *and* $\mathbb{E}|G|^2 < \infty$, *then*

$$\mathrm{AMSE}(1, \lambda_1^*) = \frac{\delta M_1(\epsilon)}{\delta - M_1(\epsilon)}\sigma^2 + o\big(e^{\frac{(M_1(\epsilon)-\delta)\tilde{\mu}^2}{2\delta\sigma^2}}\big), \tag{3.5}$$

where $M_1(\epsilon) = \min_\chi (1-\epsilon)\mathbb{E}\eta_1^2(Z;\chi) + \epsilon(1+\chi^2)$, *and* $\tilde{\mu}$ *can be any positive number smaller than* μ.

(ii) *For* $1 < q < 2$, *if* $\mathbb{P}(|G| \leq x) = O(x)$ *(as* $x \to 0$*),* $\delta > 1$, *and* $\mathbb{E}|G|^2 < \infty$ *then*

$$\mathrm{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{1-1/\delta} - \sigma^{2q}\frac{\delta^{q+1}(1-\epsilon)^2(\mathbb{E}|Z|^q)^2}{(\delta-1)^{q+1}\epsilon\mathbb{E}|G|^{2q-2}} + o\big(\sigma^{2q}\big). \tag{3.6}$$

(iii) *For* $q = 2$, *if* $\delta > 1$ *and* $\mathbb{E}|G|^2 < \infty$, *we have*

$$\mathrm{AMSE}(2, \lambda_2^*) = \frac{\sigma^2}{1-1/\delta} - \sigma^4\frac{\delta^3}{(\delta-1)^3\epsilon\mathbb{E}|G|^2} + o\big(\sigma^4\big). \tag{3.7}$$

(iv) *For* $q > 2$, *if* $\delta > 1$ *and* $\mathbb{E}|G|^{2q-2} < \infty$, *then*

$$\mathrm{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{1-1/\delta} - \sigma^4\frac{\delta^3\epsilon(q-1)^2(\mathbb{E}|G|^{q-2})^2}{(\delta-1)^3\mathbb{E}|G|^{2q-2}} + o\big(\sigma^4\big). \tag{3.8}$$

The results for $q \in [1,2]$ are taken from [55]. The proof for the case $q > 2$ can be found in Appendix I of [50]. It is straightforward to see that $M_1(\epsilon)$ is an increasing function of $\epsilon \in [0,1]$ and $M_1(1) = 1$. This implies that $\mathrm{AMSE}(1, \lambda_1^*)$ is the smallest among all $\mathrm{AMSE}(q, \lambda_q^*)$ with $q \in [1, \infty)$. As is clear, the first-order terms in the expansion of $\mathrm{AMSE}(q, \lambda_q^*)$ are the same for all $q \in (1, \infty)$. However, the second dominant term shows that the smaller values of q are preferable (note the strict monotonicity only occurs in the range $(1,2]$).

Combining the above results with Corollary 3.1 implies that in the high SNR setting, two-stage LASSO offers the best variable selection performance. We should also emphasize that as depicted in Figure 1, in this regime two-stage LASSO offers a much better variable selection performance than LASSO.

REMARK 3.5. Theorems 3.6 and 3.7 together give a full and sharp evaluation of the noise-sensitivity of bridge estimators. Among all the bridge estimators with $q \in [1, \infty)$, LASSO and Ridge are optimal for parameter estimation and variable selection, in the low and large noise settings, respectively. This result delivers an intriguing message: sparsity inducing regularization is not necessarily preferable even in sparse models. Such phenomenon might be well explained by the bias-variance tradeoff: variance is the major factor in very noisy settings, thus a regularization that produces more stable estimator is preferred, when the noise is large.

3.2.4. *Analysis of* AMSE *in large sample scenario.* Our analysis in this section is concerned with the large δ regime. Since $n/p \to \delta$ in our asymptotic setting, large δ means large sample size (relative to the dimension p). Intuitively speaking, this is similar to the classical asymptotic setting where $n \to \infty$ and p is fixed (especially if we assume the fixed number p is large). We will later connect the results we derive in the large δ regime to those obtained in classical asymptotic regime, and provide new insights.

In our original set-up, the elements of the design matrix are $X_{ij} \overset{i.i.d.}{\sim} N(0, \frac{1}{n})$. This means the SNR $\operatorname{var}(\sum_j X_{ij}\beta_j)/\operatorname{var}(w_i) \to \frac{\mathbb{E}|B|^2}{\delta\sigma^2}$ as $n \to \infty$. Therefore, if we let $\delta \to \infty$, the SNR will decrease to zero, which is not consistent with the classical asymptotics in which the SNR is assumed to be fixed. To resolve this discrepancy, we scale the noise term by $\sqrt{\delta}$ and use the model:

$$y = X\beta + \frac{1}{\sqrt{\delta}}w, \tag{3.9}$$

where $\{\beta, w, X\}$ is the converging sequence in Definition 2.1. Under this model, we compare the AMSE of different bridge estimators. The next theorem summarizes the main result.

THEOREM 3.8. *Consider the model in* (3.9) *and* $\epsilon \in (0, 1)$. *As* $\delta \to \infty$, *we have*:

(i) *For* $q = 1$, *if* $\mathbb{P}(|G| \geq \mu) = 1$ *for some* $\mu > 0$ *and* $\mathbb{E}|G|^2 < \infty$, *then*

$$\operatorname{AMSE}(1, \lambda_1^*) = \frac{M_1(\epsilon)\sigma^2}{\delta} + o(\delta^{-1}), \tag{3.10}$$

where $M_1(\epsilon)$ *has the same definition as in Theorem* 3.7(*i*).

(ii) *For* $1 < q < 2$, *if* $\mathbb{P}(|G| \leq x) = O(x)$ *(as* $x \to 0$*) and* $\mathbb{E}|G|^2 < \infty$, *then*

$$\operatorname{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{\delta} - \frac{\sigma^{2q}}{\delta^q}\frac{(1-\epsilon)^2(\mathbb{E}|Z|^q)^2}{\epsilon\mathbb{E}|G|^{2q-2}} + o(\delta^{-q}). \tag{3.11}$$

(iii) *For* $q = 2$, *if* $\mathbb{E}|G|^2 < \infty$, *then we have*

$$\operatorname{AMSE}(2, \lambda_2^*) = \frac{\sigma^2}{\delta} + \frac{\sigma^2}{\delta^2}\left[1 - \frac{\sigma^2}{\epsilon\mathbb{E}G^2}\right] + o(\delta^{-2}). \tag{3.12}$$

(iv) *For* $q > 2$, *if* $\mathbb{E}|G|^{2q-2} < \infty$, *then*

$$\operatorname{AMSE}(q, \lambda_q^*) = \frac{\sigma^2}{\delta} + \frac{\sigma^2}{\delta^2}\left[1 - \frac{\epsilon(q-1)^2\sigma^2(\mathbb{E}|G|^{q-2})^2}{\mathbb{E}|G|^{2q-2}}\right] + o(\delta^{-2}). \tag{3.13}$$

The proof of Theorem 3.8 can be found in Section I of the Supplementary Material [51]. Figure 5 compares the accuracy of the first- and second-order expansions in large range of δ. As is clear from this figure, the second-order term often offers an accurate approximation over a wide range of δ.

REMARK 3.6. As mentioned in Section 3.2.3, $M_1(\epsilon)$ is an increasing function of $\epsilon \in [0, 1]$ and $M_1(1) = 1$. This implies that $\operatorname{AMSE}(1, \lambda_1^*)$ is the smallest among all $\operatorname{AMSE}(q, \lambda_q^*)$ with $q \in [1, \infty)$. Therefore, in this regime LASSO gives the smallest estimation error, and thus two-stage LASSO offers the best variable selection performance.

REMARK 3.7. The $\operatorname{AMSE}(q, \lambda_q^*)$ with $q > 1$ share the same first dominant term, but have different second-order terms. Furthermore, for $q \in (1, 2]$, the smaller q is, the better its performance will be. Such monotonicity does not hold beyond $q = 2$.

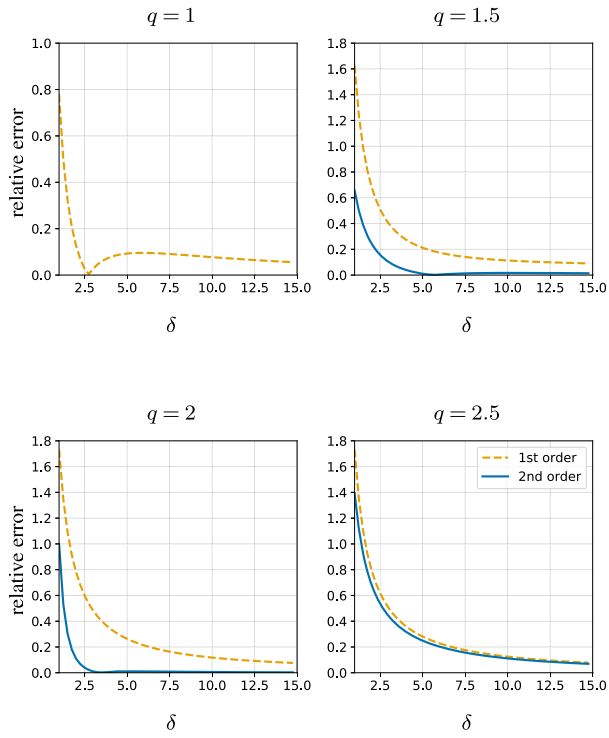

FIG. 5. *Absolute relative error of first-order and second-order approximations of AMSE under large sample scenario. In these four figures,* $p_B = (1-\epsilon)\delta_0 + \epsilon\delta_1$, $\epsilon = 0.5$, $\sigma = 1$.

We now connect our results in this large δ regime to those obtained in classical asymptotic setting. The classical asymptotics (p fixed) of bridge estimators for all the values of $q \in [0, \infty)$ is studied in [31]. We explain LASSO first. According to [31], if $\frac{\lambda}{\sqrt{n}} \to \lambda_0 \geq 0$ and $\frac{1}{n}X^T X \to C$, then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \arg\min_u V(u), \tag{3.14}$$

where $V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j)\mathbb{1}_{\{\beta_j \neq 0\}} + |u_j|\mathbb{1}_{\{\beta_j = 0\}}]$ with $W \sim \mathcal{N}(0, \sigma^2 C)$. We will do the following calculations to explore the connections. Since $X_{ij} \sim N(0, 1/n)$ in our paper, we first make the following changes to LASSO to make our set-up consistent with that of [31]:

$$\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 = \frac{1}{2}\left(\left\|y - \sqrt{n}X\frac{\beta}{\sqrt{n}}\right\|_2^2 + 2\sqrt{n}\lambda\left\|\frac{\beta}{\sqrt{n}}\right\|_1\right).$$

We thus have $C = \frac{1}{n}(\sqrt{n}X)^T(\sqrt{n}X) \to I$ and $\lambda_0 = 2\lambda$. Now suppose the result (3.14) works for $\hat{\beta}(1, \lambda)$. Then we have

$$\hat{\beta}(1, \lambda) - \beta \xrightarrow{d} \arg\min_u V(u), \tag{3.15}$$

where $V(u) = -2u^T W + u^T u + 2\lambda \sum_{j=1}^p [u_j \operatorname{sgn}(\beta_j)\mathbb{1}_{\{\beta_j \neq 0\}} + |u_j|\mathbb{1}_{\{\beta_j = 0\}}]$ with $W \sim \mathcal{N}(0, \frac{\sigma^2}{\delta} I)$. It is straightforward to see that the optimal choice of u in (3.15) has the following form:

$$\hat{u}_j = \begin{cases} W_j - \lambda \operatorname{sgn}(\beta_j) & \text{when } \beta_j \neq 0, \\ W_j - \lambda s(\hat{u}_j) & \text{when } \beta_j = 0, \end{cases}$$

where $s(u_j) = \operatorname{sgn}(u_j)$ when $u_j \neq 0$ and $|s(u_j)| \leq 1$ when $u_j = 0$. Furthermore, for the case of $\beta_j = 0$, $\hat{u}_j = 0$ is equivalent to $|W_j| \leq \lambda$ and $\operatorname{sgn}(W_j) = \operatorname{sgn}(\hat{u}_j)$ when $\hat{u}_j \neq 0$. Based on this result, we do the following heuristic calculation to connect our results with those of [31]:

$$\begin{aligned}
&\frac{1}{p}\|\hat{\beta}(1,\lambda) - \beta\|_2^2 \\
&\quad\approx \frac{1}{p}\mathbb{E}\Big[\sum_{j:\beta_j\neq 0}\big[W_j^2 - 2\lambda\operatorname{sgn}(\beta_j)W_j + \lambda^2\big] + \sum_{j:\beta_j=0,\hat{u}_j\neq 0}\big[W_j^2 - 2\lambda W_j\operatorname{sgn}(\hat{u}_j) + \lambda^2\big]\Big] \\
&\quad\approx \frac{1}{p}\Big[\sum_{j:\beta_j\neq 0}\Big(\frac{\sigma^2}{\delta} + \lambda^2\Big) + \sum_{j:\beta_j=0}\mathbb{E}\eta_1^2(W_j;\lambda)\Big] = \frac{k}{p}\Big(\frac{\sigma^2}{\delta} + \lambda^2\Big) + \frac{p-k}{p}\mathbb{E}\eta_1^2(W_j;\lambda) \\
&\quad= \frac{\sigma^2}{\delta}\Big[\frac{p-k}{p}\mathbb{E}\eta_1^2(Z;\sqrt{\delta}\lambda/\sigma) + \frac{k}{p}\big(1 + (\sqrt{\delta}\lambda/\sigma)^2\big)\Big],
\end{aligned}$$

where k is the number of nonzero elements of β and $Z \sim N(0,1)$. Note that in our asymptotic setting $k/p \to \epsilon$ and we consider the optimal tuning λ_1^*. Therefore, following the above calculations, we obtain

$$\min_{\lambda}\frac{1}{p}\|\hat{\beta}(1,\lambda) - \beta\|_2^2 \approx \frac{\sigma^2}{\delta}\min_{\chi}(1-\epsilon)\mathbb{E}\eta_1^2(Z;\chi) + \epsilon\big(1+\chi^2\big) = \frac{M_1(\epsilon)\sigma^2}{\delta}.$$

This is consistent with (3.10) in our asymptotic analysis. We can do similar calculations to show that the asumptotic analysis of [31] leads to the first-order expansion of AMSE in Theorem 3.8 for the case $q > 1$.

Based on this heuristic argument, we may conclude that the information provided by the classical asymptotic analysis is reflected in the first-order term of $\mathrm{AMSE}(q, \lambda_q^*)$. Moreover, our large sample analysis is able to derive the second dominant term for $q > 1$. This term enables us to compare the performance of different values of $q > 1$ more accurately (note they all have the same first-order term). Such comparisons cannot be performed in [31].

4. Debiasing.

4.1. *Implications of debiasing for* LASSO. As is clear from Theorem 3.2, since LASSO produces a sparse solution, it is not possible for a LASSO based two-stage method to achieve ATPP values beyond what is already reached by the first stage. This problem can be resolved by *debiasing*. In this approach, instead of thresholding the LASSO estimate (or in general a bridge estimate), we threshold its debiased version. Below we will add a dagger † to aforementioned notation to denote their corresponding debiased version. Recall $\hat{\beta}(q,\lambda)$ denotes the solution of bridge regression for any $q \geq 1$. Define the debiased estimates as:

(i) For $q = 1$,

$$\hat{\beta}^{\dagger}(1,\lambda) \triangleq \hat{\beta}(1,\lambda) + X^T\frac{y - X\hat{\beta}(1,\lambda)}{1 - \|\hat{\beta}(1,\lambda)\|_0/n},$$

where $\|\cdot\|_0$ counts the number of nonzero elements in a vector.

(ii) For $q > 1$,

$$\hat{\beta}^{\dagger}(q,\lambda) \triangleq \hat{\beta}(q,\lambda) + X^T\frac{y - X\hat{\beta}(q,\lambda)}{1 - f(\hat{\beta}(q,\lambda),\hat{\gamma}_\lambda)/n}, \tag{4.1}$$

where $f(v,w)=\sum_{i=1}^{p}\frac{1}{1+wq(q-1)|v_i|^{q-2}}$ and $\gamma=\hat{\gamma}_\lambda$ is the unique solution of the following equation:

$$\frac{\lambda}{\gamma}=1-\frac{1}{n}f\big(\hat{\beta}(q,\lambda),\gamma\big). \tag{4.2}$$

We have the following theorem to confirm the validity of the debiasing estimator $\hat{\beta}^\dagger(q,\lambda)$.

THEOREM 4.1. *For any given $q\in[1,\infty)$, with probability one, the empirical distribution of the components of $\hat{\beta}^\dagger(q,\lambda)-\beta$ converges weakly to $N(0,\tau^2)$, where τ is the solution of* (2.3) *and* (2.4).

See Section J in the Supplementary Material [51] for the proof. In order to perform variable selection, one may apply the hard thresholding function to these debiased estimates, that is,

$$\bar{\beta}^\dagger(q,\lambda,s)=\eta_0\big(\hat{\beta}^\dagger(q,\lambda);s^2/2\big)=\hat{\beta}^\dagger(q,\lambda)\mathbb{1}_{\{|\hat{\beta}^\dagger(q,\lambda)|\geq s\}}.$$

We use the notation $\mathrm{ATPP}^\dagger(q,\lambda,s)$ and $\mathrm{AFDP}^\dagger(q,\lambda,s)$ to denote the ATPP and AFDP of $\bar{\beta}^\dagger(q,\lambda,s)$, respectively. In the case of LASSO, note that unlike $\hat{\beta}(1,\lambda)$ the debiased estimator $\hat{\beta}^\dagger(1,\lambda)$ is dense. Hence we expect the two-stage variable selection estimate $\bar{\beta}^\dagger(1,\lambda,s)$ to be able to reach any value of ATPP between $[0,1]$. The following theorem confirms this claim.

THEOREM 4.2. *Given the* ATPP *level $\zeta\in[0,1]$, for every value of $\lambda>0$, there exists $s(\lambda,\zeta)$ such that* $\mathrm{ATPP}^\dagger(1,\lambda,s(\lambda,\zeta))=\zeta$. *Furthermore, whenever $\bar{\beta}^\dagger(1,\lambda,s)$ and $\bar{\beta}(1,\lambda,\tilde{s})$ reach the same level of* ATPP, *they have the same* AFDP. *The value of λ that minimizes* $\mathrm{AFDP}^\dagger(1,\lambda,s(\lambda,\zeta))$ *also minimizes* $\mathrm{AMSE}(1,\lambda)$.

As expected since the solution of bridge regression for $q>1$ is dense, the debiasing step does not help variable selection for $q>1$. Our next theorem confirms this claim.

THEOREM 4.3. *Consider $q>1$. Given the* ATPP *level $\zeta\in[0,1]$, for every value of $\lambda>0$, there exists $s(\lambda,\zeta)$ such that* $\mathrm{ATPP}^\dagger(q,\lambda,s(\lambda,\zeta))=\zeta$. *Furthermore, whenever $\bar{\beta}^\dagger(q,\lambda,s)$ and $\bar{\beta}(q,\lambda,\tilde{s})$ reach the same level of* ATPP, *they have the same* AFDP. *Also, the value of λ that minimizes* $\mathrm{AFDP}^\dagger(q,\lambda,s(\lambda,\zeta))$ *also minimizes* $\mathrm{AMSE}(q,\lambda)$. *As a result, the optimal value of* $\mathrm{AFDP}^\dagger(q,\lambda,s(\lambda,\zeta))$ *is the same as* $\mathrm{AFDP}(q,\lambda_q^*,s_q^*(\zeta))$.

For the proof of Theorems 4.2 and 4.3, please refer to Section J in the Supplementary Material [51].

REMARK 4.1. Comparing Theorem 4.2 with Theorem 3.2, we see that replacing LASSO in the first stage with the debiased version enables to achieve wider range of ATPP level. On the other hand, given the value of λ, if $\bar{\beta}^\dagger(1,\lambda,s)$ and $\bar{\beta}(1,\lambda,\tilde{s})$ reach the same level of ATPP, their AFDP are equal as well. Therefore, the debiasing for LASSO expands the range of AFDP–ATPP curve without changing the original one. Figure 6 compares the variable selection performance of LASSO with that of the two-stage scheme having the debiased LASSO estimate in the first stage. Compare this figure with Figure 1 to see the difference between the two-stage LASSO and two-stage debiased LASSO.

REMARK 4.2. The debiasing does not present any extra gain to the two-stage variable selection technique based on bridge estimators with $q>1$. In other words, debiasing does not change the AFDP–ATPP curve for $q>1$.

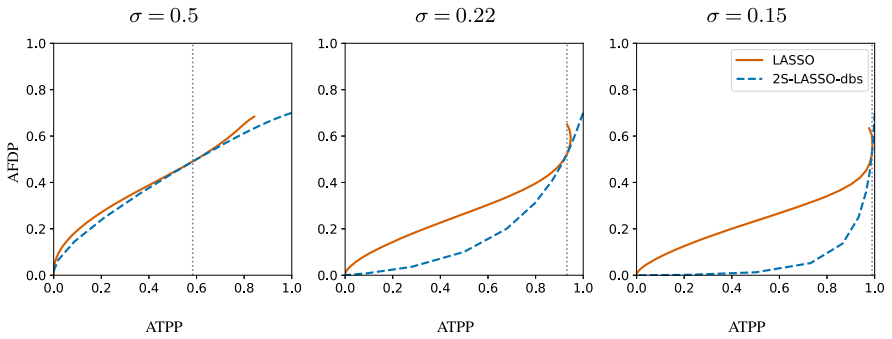

FIG. 6. *Comparison of* AFDP–ATPP *curve between* LASSO *and two-stage debiased* LASSO. *Here, we pick the setting* $\delta = 0.8$, $\epsilon = 0.3$, $\sigma \in \{0.5, 0.22, 0.15\}$, $p_G = \delta_1$. *For the two-stage debiased* LASSO, *we use optimal tuning* λ_1^* *in the first stage. The gray dotted line is the upper bound for the two-stage* LASSO *without debiasing can reach.*

4.2. *Debiasing and sure independence screening.* Sure Independence Screening (SIS) is a variable selection scheme proposed for ultra-high dimensional settings [21]. Our asymptotic setting is not considered an ultrahigh dimensional asymptotic. We are also aware that SIS is typically used for screening out irrelevant variables and other variable selection methods, such as LASSO, will be applied afterwards. Nevertheless, we present a connection and comparison between our two-stage methods and SIS in the linear asymptotic regime. Such comparisons shed more light on the performance of SIS. It is straightforward to confirm that sure independence screening is equivalent to

$$\bar{\beta}^\dagger(q, \infty, s) = \eta_0\big(\hat{\beta}^\dagger(q, \infty); s^2/2\big) = \eta_0\big(X^T y; s^2/2\big).$$

Therefore, the main difference between the approach we propose in this paper and SIS, is that SIS sets λ to ∞, while we select the value of λ that minimizes AMSE.[3] This simple difference may give a major boost to the variable selection performance. The following lemma confirms this claim.

LEMMA 4.1. *Consider* $q \geq 1$. *Given any* ATPP *level* $\zeta \in [0, 1]$, *let* $\mathrm{AFDP}_{\mathrm{sis}}(\zeta)$ *and* $\mathrm{AFDP}^\dagger(q, \lambda_q^*, s(\lambda_q^*, \zeta))$ *denote the asymptotic* FDP *of SIS and two-stage debiased bridge estimator respectively, when their* ATPP *is equal to* ζ. *Then* $\mathrm{AFDP}^\dagger(q, \lambda_q^*, s(\lambda_q^*, \zeta)) \leq \mathrm{AFDP}_{\mathrm{sis}}(\zeta)$.

Refer to Section J of the Supplementary Material [51] for the proof. Note that when the noise σ is large, we expect the optimally tuned λ to be large, and hence the performance of SIS gets closer to the TVS. However, as σ decreases, the gain obtained from using a better estimator in the first stage improves. Figure 7 compares the performance of SIS and TVS under different noise settings.

5. Numerical experiments.

5.1. *Objective and simulation set-up.* This section aims to investigate the finite sample performances of various two-stage variable selection estimators under the three different regimes analyzed in Section 3.2. In particular, we will study to what extent our theory works for more realistic situations, where model parameters σ, ϵ, δ are of moderate magnitudes or the i.i.d. Gaussian design assumption is violated. For brevity, we will use bridge estimator

[3] Our approach is more aligned with the approach proposed in [53]. However, [53] uses data splitting to select λ.

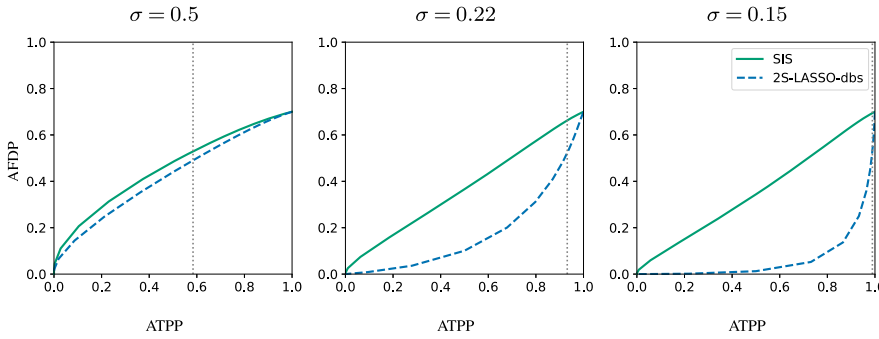

FIG. 7. *Comparison of* AFDP–ATPP *curve between SIS and the two-stage debiased* LASSO. *Here, we pick the setting* $\delta = 0.8$, $\epsilon = 0.3$, $\sigma \in \{0.5, 0.22, 0.15\}$, $p_G = \delta_1$. *For the two-stage debiased* LASSO, *we use optimal tuning* λ_1^* *in the first stage. The gray dotted line is the upper bound that the two-stage* LASSO *without debiasing can reach.*

to refer to the corresponding two-stage method whenever it does not cause any confusion. More specifically, in all the figures, ℓ_q will be used to denote the TVS that uses the bridge estimator with q in the first stage, and ℓ_1-db denotes the two-stage debiased LASSO. The performances of different methods will be compared via the AFDP–ATPP curves.[4]

The organization of this section is as follows. In Sections 5.2–5.6, we focus on experiments under i.i.d. Gaussian design as assumed in our theories. In Section 5.7, we present numerical results for non-i.i.d. or non-Gaussian designs to evaluate the accuracy of our results, when i.i.d. Gaussian assumption on X is violated.

We adopt the following settings for i.i.d. Gaussian design. The settings for general design are described in Section 5.7.

1. Number of variables is fixed at $p = 5000$. Sample size $n = p\delta$ is then decided by δ.

2. Given the values of δ, ϵ, σ, we sample $X \in \mathbb{R}^{n\times p}$ with $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n})$. We pick the probability measure p_G as a point mass at M where M will be specified in each scenario. We generate $\beta \in \mathbb{R}^p$ with $\beta_i \overset{i.i.d.}{\sim} p_B = (1-\epsilon)\delta_0 + \epsilon p_G$, and $w \in \mathbb{R}^n$ with $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ or $\mathcal{N}(0, \frac{\sigma^2}{\delta})$.[5] Construct y according to $y = X\beta + w$.

3. For each data set (y, X), AFDP–ATPP curves will be generated for different variable selection methods. In each setting of parameters, 80 samples are drawn and the average AFDP–ATPP curves are calculated. The associated one standard deviation confidence interval will be presented.

We compute bridge estimators via coordinate descent algorithm, with the proximal operator $\eta_q(x;\tau)$ calculated through a properly implemented Newton's method.

We discuss how to pick optimal tuning under i.i.d. Gaussian design in Section 5.2. Section 5.3 presents the large/small noise scenario. Section 5.4 is devoted to the large sample regime. Section 5.5 covers the nearly black object scenario. In Section 5.6, we compare the performance of LASSO and two-stage LASSO to shed more lights on our two-stage methods.

5.2. *Estimating the optimal tuning* λ_q^*. For two-stage variable selection procedures, it is critical to have a good estimator in the first step. One challenge here is to search for the

[4]Since the simulations are in finite samples, the curve we calculate is actually FDP–TPP instead of the asymptotic version. With a little abuse of notation, we will call it AFDP–ATPP curve throughout the section.

[5]The setting $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{\sigma^2}{\delta})$ will be used in the large sample scenario, since we have scaled the error term by $\sqrt{\delta}$ in our asymptotic analysis in Section 3.2.4.

optimal tuning that minimizes AMSE of $\hat{\beta}(q,\lambda)$. According to the result of Theorem B.1 of the Supplementary Material [51] and the definition of AMSE in (2.1), it is straightforward to see that $\tau^2 = \sigma^2 + \frac{1}{\delta}$ AMSE. Hence, one can minimize τ^2 to achieve the same optimal tuning. Motivated by [38], we can obtain a consistent estimator of τ^2:

$$q=1: \quad \hat{\tau}^2 = \frac{\|y - X\hat{\beta}(1,\lambda)\|_2^2}{n(1-\|\hat{\beta}(1,\lambda)\|_0/n)^2}, \qquad q>1: \quad \hat{\tau}^2 = \frac{\|y - X\hat{\beta}(q,\lambda)\|_2^2}{n(1-f(\hat{\beta}(q,\lambda),\hat{\gamma}_\lambda)/n)^2},$$

where $f(\cdot,\cdot)$, $\hat{\gamma}_\lambda$ are the same as the ones in (4.1) and (4.2). The consistency $\hat{\tau} \overset{a.s.}{\to} \tau$ can be easily seen from the proof of Theorem 4.1. We thus do not repeat it. As a result, we approximate λ_q^* by searching for the λ that minimizes $\hat{\tau}^2$. Notice that this problem has been studied for LASSO in [38] and a generalization is straightforward for other bridge estimators. We use the following grid search strategy:

- Initialization: An initial search region $[a,b]$, a window size Δ and a grid size m.
- Searching: A grid with size m is built over $[a,b]$, upon which we search in descending order for λ that minimizes $\hat{\tau}^2$ with warm initialization.
  - If the minimal point $\hat{\lambda} \in (a,b)$, stop searching and return $\hat{\lambda}$.
  - If $\hat{\lambda} = a$ or b, update the search region with $[\frac{a}{10}, a]$ or $[b, b+\Delta]$ and do the next round of searching.
- Stability: If the optimal $\hat{\lambda}$ obtained from two consecutive search regions are smaller than a threshold ϵ_0, we stop and return the previous optimal $\hat{\lambda}$; If the number of nonzero locations of a LASSO estimator is larger than n (which may happen numerically for very small tuning), we set its $\hat{\tau}^2$ to ∞.

For our experiments, we pick the initial $[a,b] = [0.1, \frac{1}{2}\|X^T y\|_\infty]$, $\Delta = \frac{1}{2}\|X^T y\|_\infty$ and $m = 15$.

5.3. *From large noise to small noise.* Theorems 3.6 and 3.7 showed that in low and high SNR situations, ridge and LASSO offer the best performances, respectively. These results are obtained for limiting cases $\sigma \to \infty$ and $\sigma \to 0$. In this section, we run a few simulations to clarify the scope of applicability of our analysis. Toward this goal, we fix the probability measure $p_G = \delta_M$ with $M = 8$ and run TVS for $q \in \{1, 1.2, 2, 4\}$ and debiased LASSO[6] under four settings:

1. $\delta = 0.8$, $\epsilon = 0.2$: The results are shown in Figure 8. Here, we pick $\sigma \in \{1.5, 3, 5\}$. As expected from our theoretical results, for small values of noise LASSO offers the best performance. As we increase the noise, eventually ridge outperforms LASSO and the other bridge estimators. Note that under this setting, the outperformance occurs at a high noise level so that all estimators have large errors. In this example, we make $1 > \delta > M_1(\epsilon)$. Refer to Theorem 3.7 for the importance of this condition.
2. $\delta = 2$, $\epsilon = 0.4$: The results are included in Figure 8. Here, we pick $\sigma \in \{2, 4, 8\}$. Similar phenomena are observed. However, for all choices of σ, the AFDP–ATPP curves of different methods are quite close to each other.
3. $\delta = 0.6$, $\epsilon = 0.4$: Figure 9 contains the results for this part. Here, we have $\sigma \in \{0.25, 0.75, 2\}$. An important feature of this simulation is that $\delta < M_1(\epsilon)$, which does not satisfy the condition of Theorem 3.7. It is interesting to observe that in this case, ridge outperforms LASSO even for small values of the noise. We thus see that the superiority of LASSO

[6]We include the results for two-stage debiased LASSO in Sections 5.3–5.5 to validate the effect of debiasing stated in Theorem 4.2 and Remark 4.1.

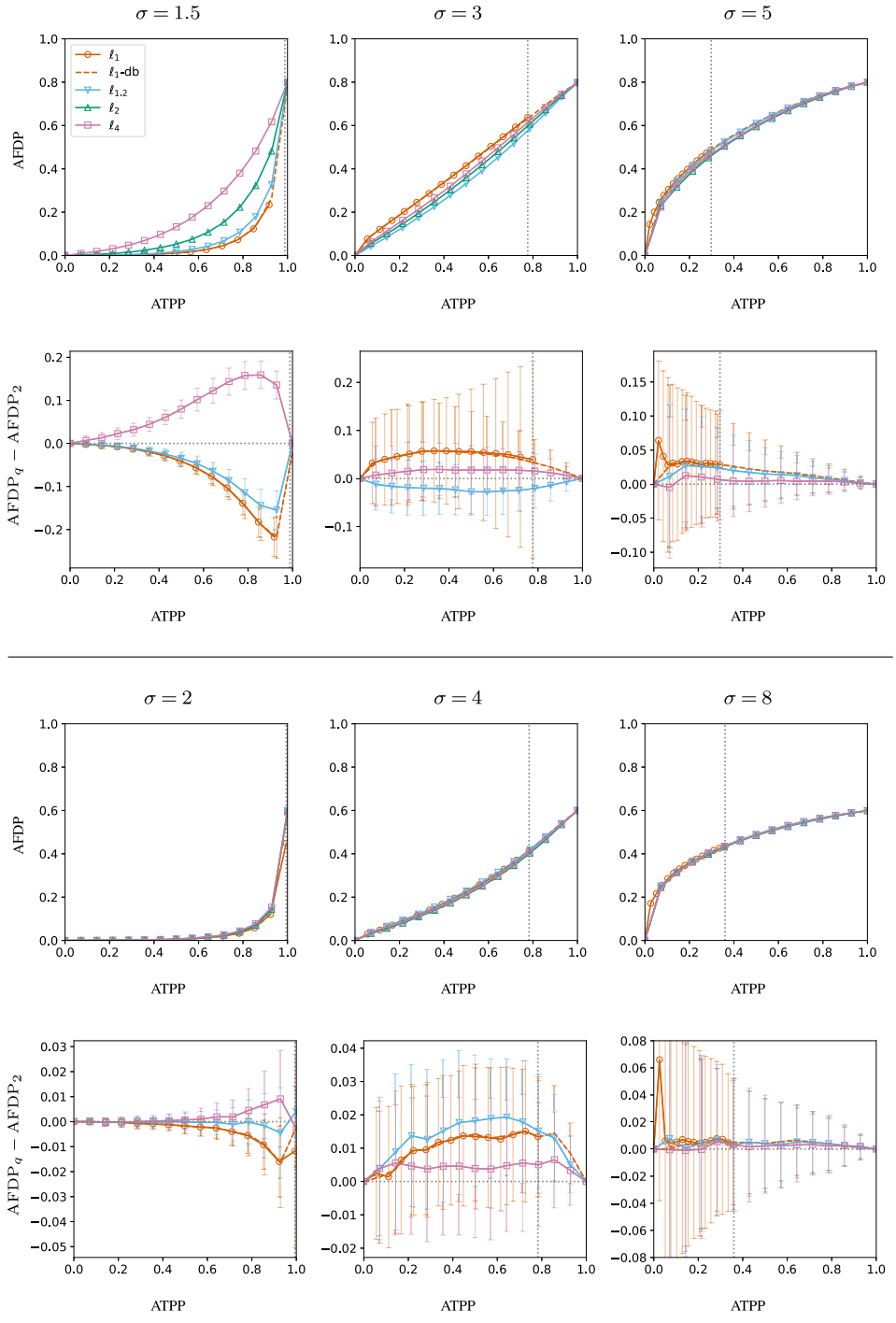

FIG. 8. *Top row*: AFDP–ATPP *curve under the setting* $\delta = 0.8$, $\epsilon = 0.2$, $\sigma \in \{1.5, 3, 5\}$. *Second row*: *Y-axis is the difference of* AFDP *between the other bridge estimators and ridge. One standard deviation of the difference is added. Third and fourth rows*: *the same type of plots as in the first two rows, under the setting* $\delta = 2$, $\epsilon = 0.4$, $\sigma \in \{2, 4, 8\}$.

in small noise characterized by Theorem 3.7 may not hold when the conditions of the theorem are violated. In fact, Theorem 3.7 is restricted to the regime below the phase transition (i.e., when the signal can be fully recovered without noise). However, in the current setting, the optimal AMSE for $q = 1, 1.2, 2, 4$ at $\sigma = 0$ are 14.9, 12.2, 10.2, 11.6, respectively.

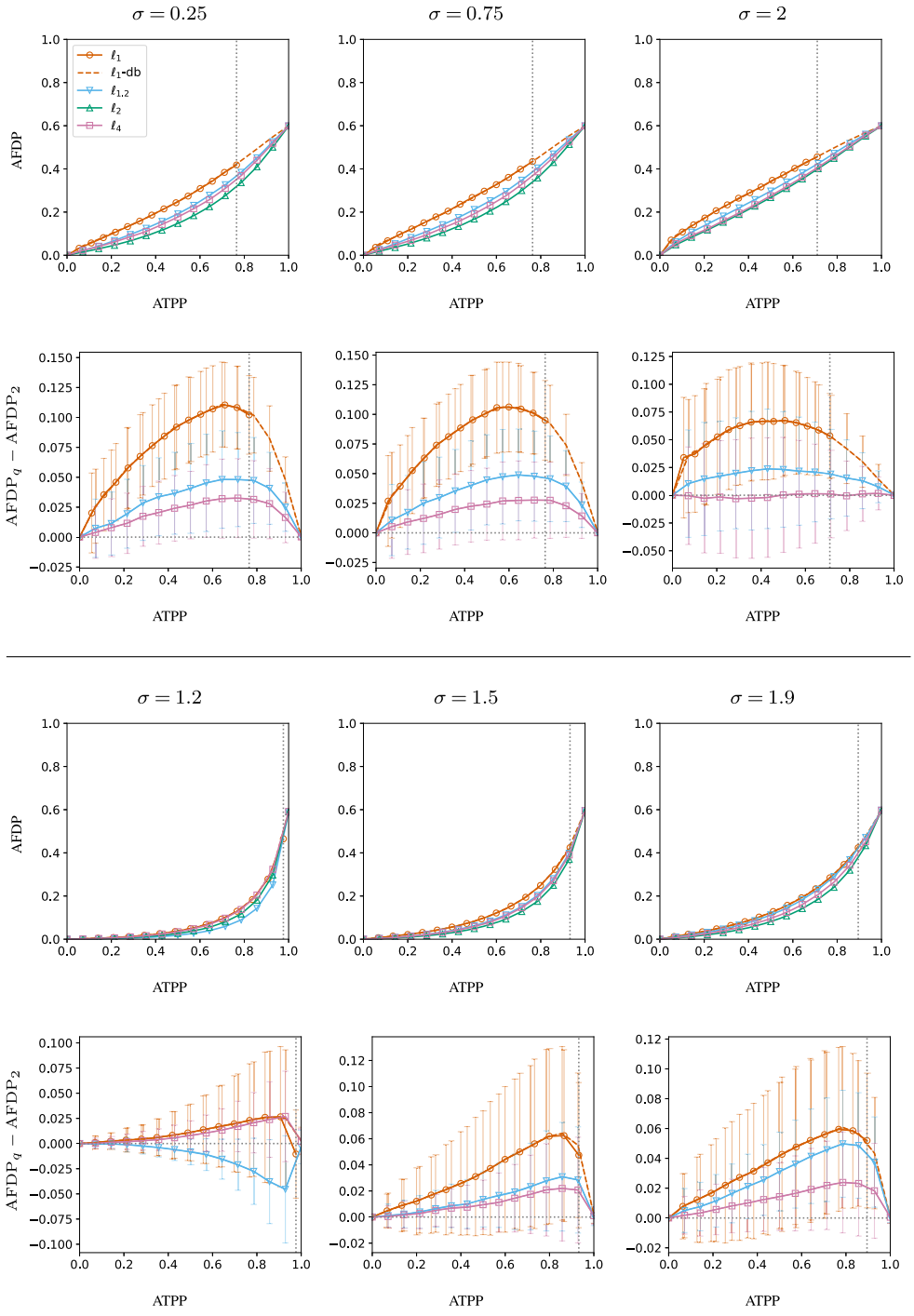

FIG. 9. *Top row*: AFDP–ATPP *curve under the setting* $\delta = 0.6$, $\epsilon = 0.4$, $\sigma \in \{0.25, 0.75, 2\}$. *Second row*: *Y-axis is the difference of* AFDP *between the other bridge estimators and ridge. One standard deviation of the difference is added. Third and fourth rows*: *the same type of plots as in the first two rows*, *under the setting* $\delta = 0.9$, $\epsilon = 0.4$, $\sigma \in \{1.2, 1.5, 1.9\}$.

4. $\delta = 0.9$, $\epsilon = 0.4$: The results are shown in Figure 9. Here, we have $\sigma \in \{1.2, 1.5, 1.9\}$. This group of figures provide us with examples where ridge based TVS outperforms the other two-stage methods, and at the same time reaches a quite satisfactory AFDP–ATPP trade-off.

For instance, when $\sigma = 1.5$ and AFDP ≈ 0.2, for ridge we have ATPP ≈ 0.8 while that for LASSO is around 0.7. Note that here $M_1(\epsilon) < \delta < 1$.

5.4. *Large sample regime.* We will validate the results in Theorem 3.8, which are obtained under the limiting case $\delta \to \infty$. We fix the probability measure $p_G = \delta_M$ with $M = 1$ and consider the following settings for $q \in \{1, 1.5, 2, 4\}$ and debiased LASSO:

1. $\epsilon = 0.1$, $\sigma = 0.4$: The results for this setting are shown in Figure 10. We vary $\delta \in \{2, 3, 4\}$. As is clear, LASSO starts to outperform the others even when $\delta = 2$. As δ increases, LASSO remains the best, but all the methods are becoming better and the AFDP–ATPP curves get closer to each other.
2. $\epsilon = 0.3$, $\sigma = 0.4$: The results can be found in Figure 10. Again $\delta \in \{2, 3, 4\}$. Similar phenomena are observed. Compared to the previous setting, a larger ϵ leads to a higher SNR and all the methods have improved performances.
3. $\epsilon = 0.4$, $\sigma = 0.22$: The results are shown in Figure 11. We set $\delta \in \{0.7, 0.8, 1.2\}$. When δ is 0.7 or 0.8, ridge significantly outperforms the others. As δ is increased to 1.2, LASSO starts to lead the performances.

5.5. *Nearly black object.* In this section, we verify our theoretical results which are presented in Section 3.2.2 for the nearly black object setting. Recall $b_\epsilon = \sqrt{\mathbb{E}G^2}$ and $\tilde{G} = G/b_\epsilon$. We consider the following setting: $\delta = 0.8$, $\sigma \in \{3, 5\}$, $b_\epsilon = 4/\sqrt{\epsilon}$, $\tilde{G} = 1$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. The simulation results are displayed in Figure 12. We observe that under both noise levels $\sigma = 3, 5$, LASSO is suboptimal at sparsity level $\epsilon = 0.25$. As ϵ decreases, LASSO becomes better. When ϵ is reduced to 0.04, LASSO outperforms the other bridge estimators by a large margin. Note that in this simulation, the signal strength b_ϵ scales with ϵ at the rate $\epsilon^{-1/2}$. This is the regime where LASSO is proved to be optimal in Section 3.2.2.

5.6. *LASSO versus two-stage LASSO.* In Theorem 3.2 we proved that two-stage LASSO with its first stage optimally tuned outperforms LASSO on variable selection. We now provide a brief simulation to verify this result. We choose $p_G = \delta_M$ with $M = 8$ and set $\delta = 0.8$, $\epsilon = 0.2$, $\sigma \in \{1, 3, 5\}$. As shown in Figure 13, two-stage LASSO improves over LASSO. When the noise is small ($\sigma = 1$), the improvement is the most significant. As the noise level increases, the difference between the two approaches becomes smaller. When the noise is large ($\sigma = 5$), both have large errors.

5.7. *General design.* In this section, we extend our simulations to general design matrices. Given that our theoretical results in Section 3 are derived under the i.i.d. Gaussian assumption on X, the aim of this section is to numerically study the validity scope of our main conclusions when such an assumption does not hold. In particular, we consider the following correlated designs and i.i.d. non-Gaussian designs:

- Correlated design: We consider the model $y = X\Sigma^{\frac{1}{2}}\beta + w$, where $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n})$ and Σ is a Toeplitz matrix with $\Sigma_{ij} = \rho^{|i-j|}$. Here, $\rho \in (0, 1)$ controls the correlation strength.
- i.i.d. non-Gaussian design: We generate X with i.i.d. components $X_{ij} \sim \sqrt{\frac{\nu-2}{n\nu}} t_\nu$ where t_ν is the t-distribution with degrees of freedom ν. The scaling $\sqrt{\frac{\nu-2}{n\nu}}$ ensures $\mathrm{var}(X_{ij}) = \frac{1}{n}$ as in the i.i.d. Gaussian case.

Throughout this section, we choose $p = 2500$, $p_G = \delta_M$, $n = \delta p$, $w_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

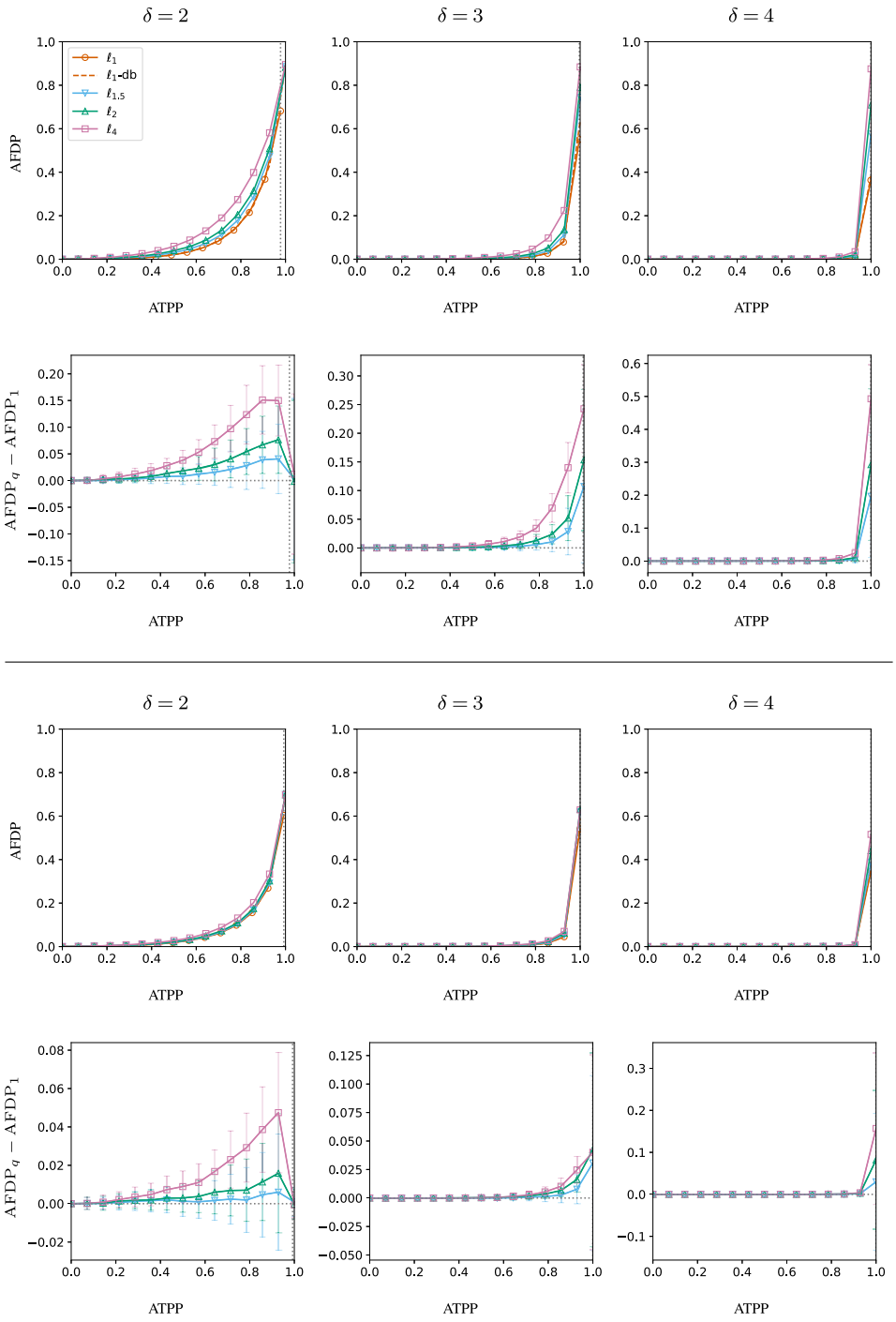

FIG. 10. *Top row*: AFDP–ATPP *curve under the setting* $\epsilon = 0.1$, $\sigma = 0.4$, $\delta \in \{2, 3, 4\}$. *Second row*: *Y-axis is the difference of* AFDP *between the other bridge estimators and LASSO. One standard deviation of the difference is added. Third and fourth rows*: *the same type of plots as in the first two rows, under the setting* $\epsilon = 0.3$, $\sigma = 0.4$, $\delta \in \{2, 3, 4\}$.

Large/small noise. We set $M = 8$, $\delta = 0.9$, $\epsilon = 0.4$. For correlated design, we vary $\rho \in \{0.1, 0.5, 0.9\}$ to allow for different levels of correlations among the predictors. Figure 14 shows the simulation results. There are a few important observations:

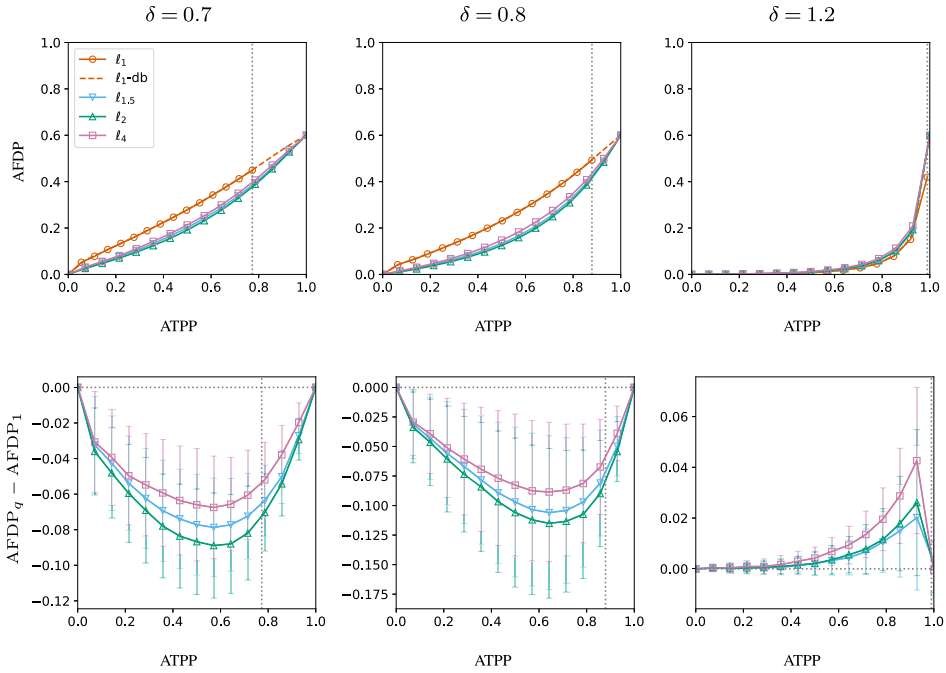

FIG. 11. *Top row*: AFDP–ATPP *curve under the setting* $\epsilon = 0.4$, $\sigma = 0.22$, $\delta \in \{0.7, 0.8, 1.2\}$. *Second row*: *Y-axis is the difference of* AFDP *between the other bridge estimators and LASSO. One standard deviation of the difference is added.*

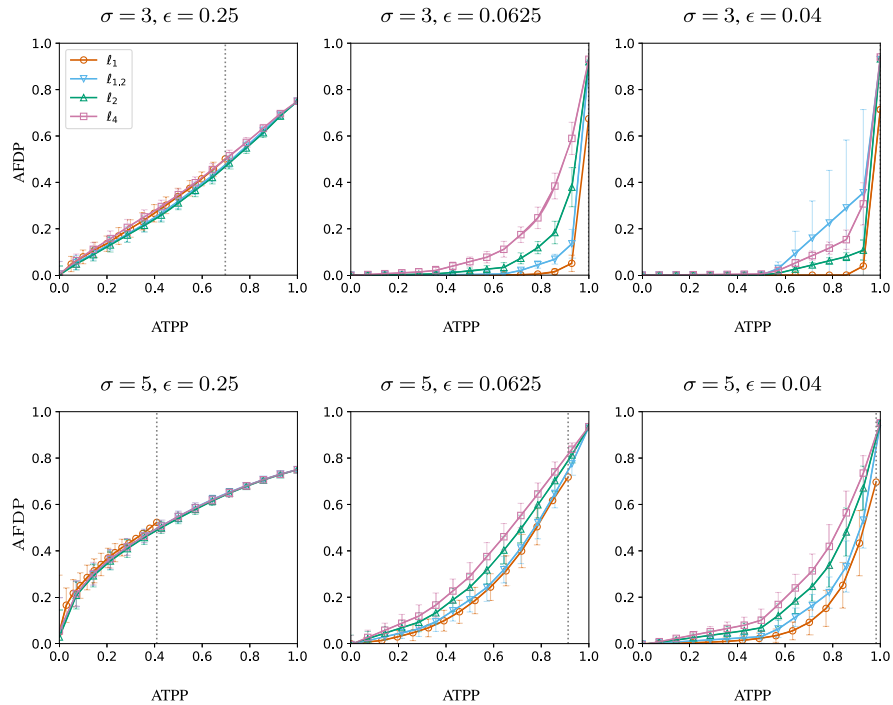

FIG. 12. *Top row*: AFDP–ATPP *curve under the setting* $b_\epsilon = 4/\sqrt{\epsilon}$, $\sigma = 3$, $\delta = 0.8$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. *Second row*: AFDP–ATPP *curve under the setting* $b_\epsilon = 4/\sqrt{\epsilon}$, $\sigma = 5$, $\delta = 0.8$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. *One standard deviation is added.*

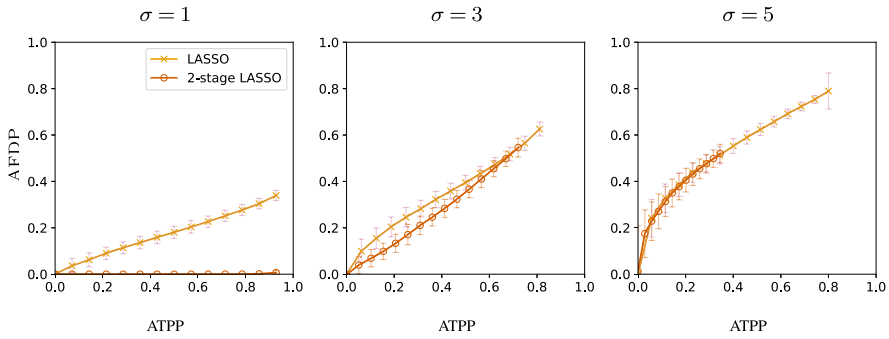

FIG. 13. *LASSO versus two-stage LASSO. Here,* $\delta = 0.8$, $\epsilon = 0.2$, $M = 8$, $\sigma \in \{1, 3, 5\}$. *The outperformance of two-stage LASSO is the most significant when the noise level is low. When noise gets higher, the gap becomes smaller and smaller.*

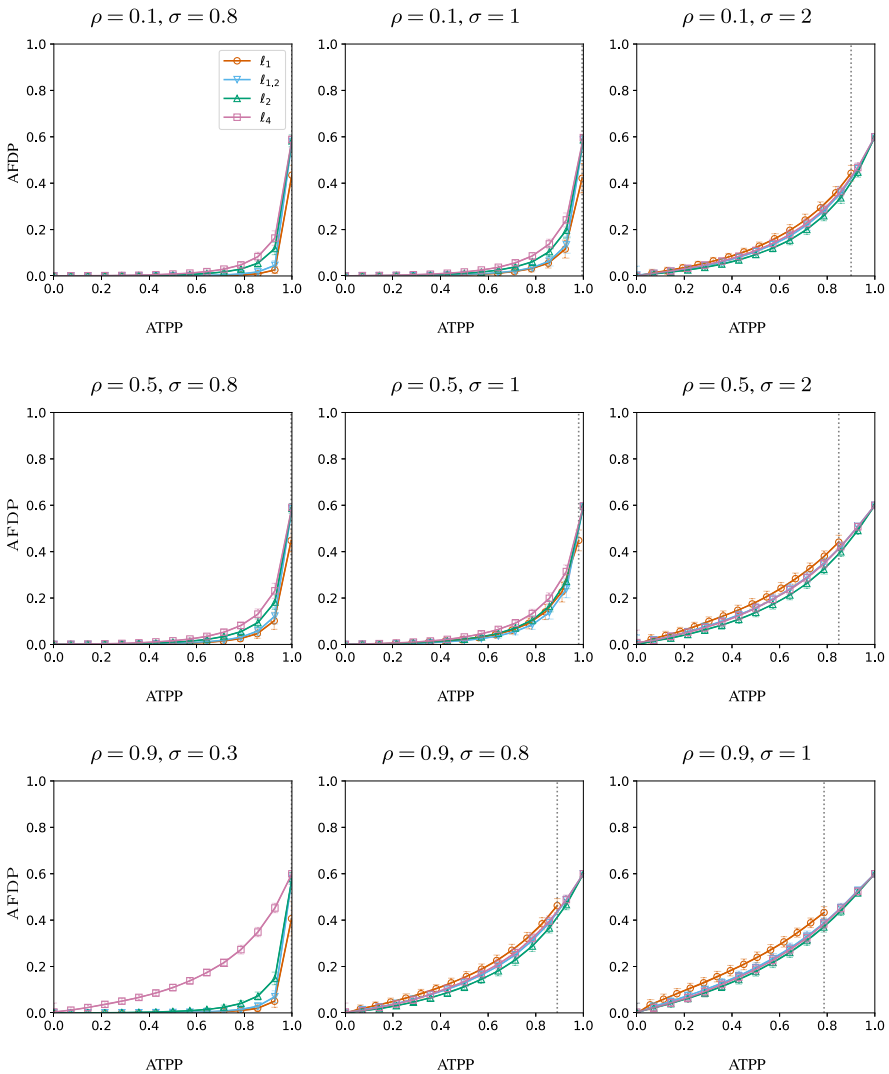

FIG. 14. *Large/small noise scenario under correlated design.*

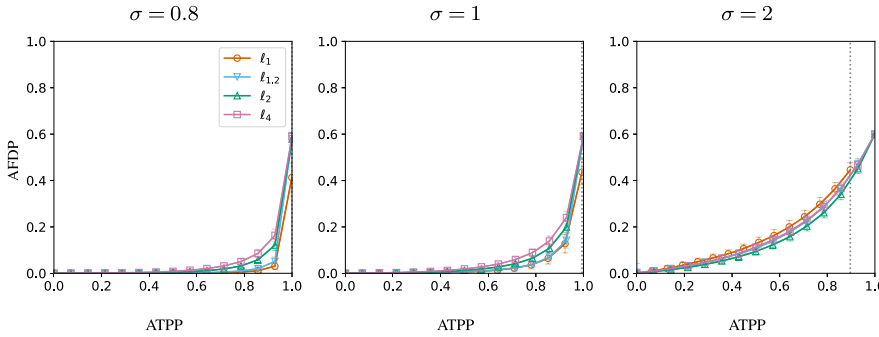

FIG. 15. *Large/small noise scenario under i.i.d. non-Gaussian design. We set* $\delta = 0.9$, $\epsilon = 0.4$, $M = 8$, $\sigma \in \{0.8, 1, 2\}$. *The degrees of freedom of the t-distribution is* $\nu = 3$.

(i) For a given $\rho \in \{0.1, 0.5, 0.9\}$, the comparison of bridge estimators under different noise levels is similar to what we observe for i.i.d. Gaussian designs: LASSO performs best in low noise case, and ridge becomes optimal when the noise is large.

(ii) Given the noise level $\sigma = 0.8$, as the design correlation ρ varies in $\{0.1, 0.5, 0.9\}$, it is interesting to observe that, LASSO outperforms the other estimators when the correlation is not high ($\rho = 0.1, 0.5$), while ridge becomes the optimal one when the correlation is increased to 0.9. Similar phenomenon happens at the noise level $\sigma = 1$. It seems that in terms of variable selection performance comparison of TVS, adding dependency among the predictors is like increasing the noise level in the system. We leave a theoretical analysis of the impact of correlation on our results as an interesting future research.

Regarding i.i.d. non-Gaussian design, we choose the t-distribution t_ν with $\nu = 3$. Note that among all the t-distributions $\{t_\nu, \nu \in \mathbb{N}\}$ with finite variance, t_3 has the heaviest tail. The results are shown in Figure 15. We again observe the comparison predicted by our theory: LASSO outperforms the other bridge estimators when the noise level is low ($\sigma = 0.8$), and ridge performs best as the noise level increases to $\sigma = 2$.

Nearly black object. For nearly black objects, we consider $\delta = 0.8$, $\sigma = 3$, $b_\epsilon = \frac{4}{\sqrt{\epsilon}}$, $\tilde{G} = 1$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. We construct the design matrix in the following ways:

(i) Set a correlated Gaussian design with correlation levels $\rho = 0.5, 0.9$.
(ii) Set an i.i.d. non-Gaussian design with t_3.

Figures 16 and 17 contain the results for the correlated design and i.i.d. non-Gaussian design, respectively. We can see that as the model becomes sparser, LASSO starts to outperform other choices of bridge estimator and eventually becomes optimal. This is consistent with the main conclusion we have proved for the i.i.d. Gaussian designs.

LASSO versus two-stage LASSO. We compare LASSO and two-stage LASSO under more general designs. As in Section 5.6 for i.i.d. Gaussian design, we set $\delta = 0.8$, $\epsilon = 0.2$, $M = 8$ and $\sigma = 1, 3, 5$. For correlated designs, we pick $\rho = 0.5, 0.9$. For i.i.d. non-Gaussian design, we choose $\nu = 3$. As is seen in Figure 18, the same phenomenon observed in i.i.d. Gaussian design also occurs under general designs: two-stage LASSO outperforms LASSO by a large margin when the noise is small, and the outperformance becomes marginal in large noise.

6. Discussion.

6.1. *Nonconvex bridge estimators.* In this paper, our discussion has been focused on the bridge estimators with $q \in [1, \infty)$. When q falls in $[0, 1)$, the corresponding bridge regression becomes a nonconvex problem. Given that certain nonconvex regularizations have been

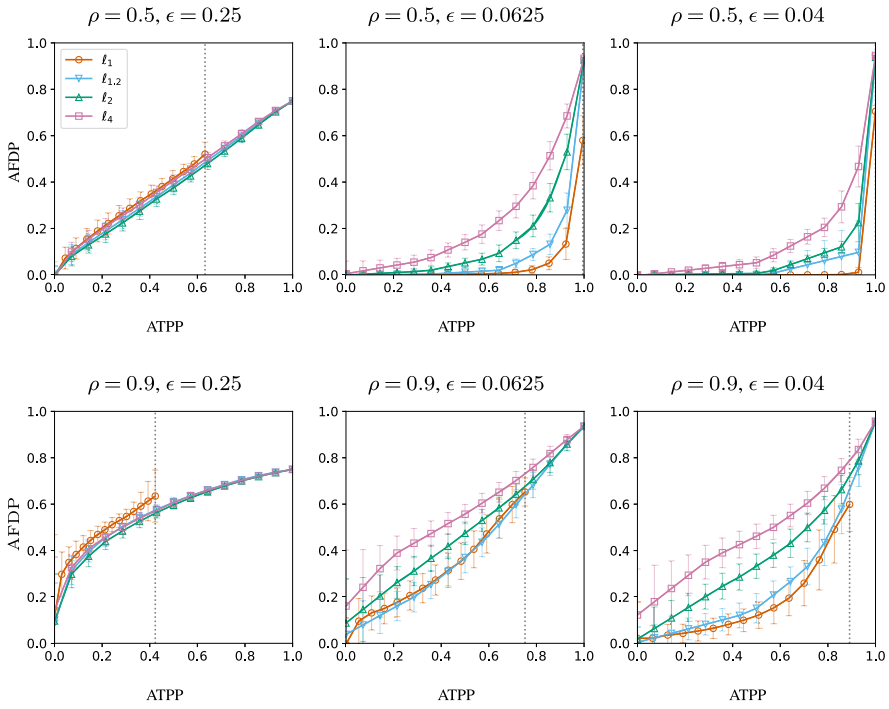

FIG. 16. *Nearly black object with correlated design. We fix* $\delta = 0.8$, $\sigma = 3$ *and* $b_\epsilon = 4/\sqrt{\epsilon}$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. *The correlation* ρ *is set to* 0.5 *and* 0.9 *in the two rows.*

shown to achieve variable selection consistency under weaker conditions than LASSO [32], it is of great interest to analyze the variable selection performance of nonconvex bridge estimators. An early work [26] has showed that bridge estimators for $q \in (0, 1)$ enjoy an oracle property in the sense of [20] under appropriate conditions. However, the asymptotic regime considered in [26] is fundamentally different from the linear asymptotic in the current paper. A more relevant work is [60] which studied the estimation property of bridge regression when q belongs to $[0, 1]$ under a similar asymptotic framework to ours. Nevertheless, the main focus of [60] is on the estimators returned by an iterative local algorithm. The analysis of the global minimizer in [60] relies on the replica method [42] from statistical physics, which has not been fully rigorous yet. To the best of our knowledge, under the linear asymptotic setting, no existing works have provided a fully rigorous analysis of the global solution from non-

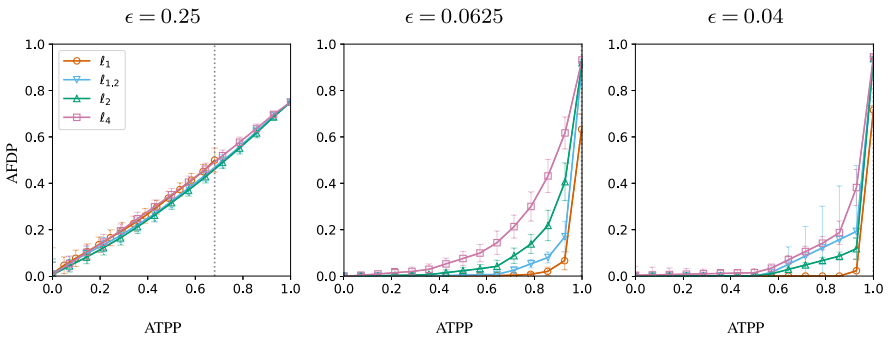

FIG. 17. *Nearly black object with i.i.d. non-Gaussian design. We fix* $\delta = 0.8$, $\sigma = 3$ *and* $b_\epsilon = 4/\sqrt{\epsilon}$, $\epsilon \in \{0.25, 0.0625, 0.04\}$. *The degrees of freedom for the t-distribution design is* $\nu = 3$.

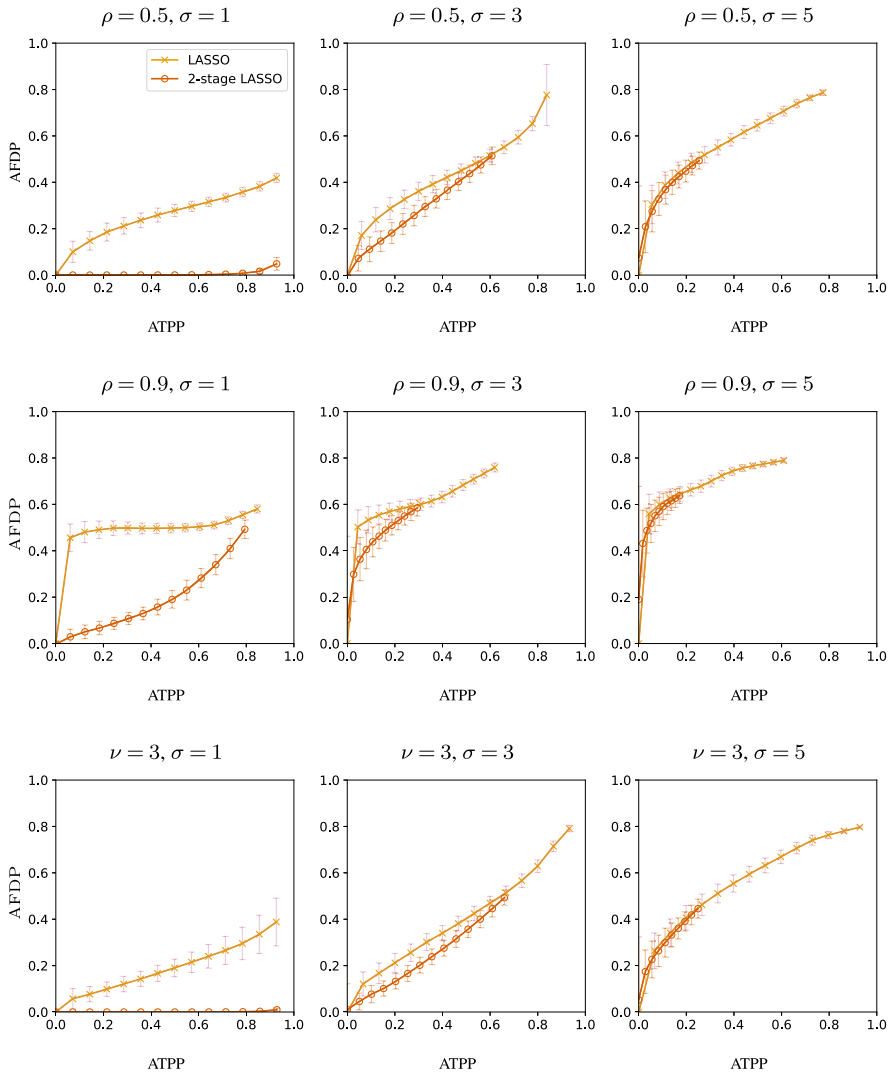

FIG. 18. *LASSO versus two-stage LASSO under general designs. Here,* $\delta = 0.8$, $\epsilon = 0.2$, $M = 8$, $\sigma \in \{1, 3, 5\}$. *The first two rows are for* $\rho = 0.5, 0.9$ *in correlated design. The last row is for* $\nu = 3$ *in i.i.d. non-Gaussian design.*

convex regularization in linear regression models. We leave this important and challenging problem as a future research.

6.2. *Tuning parameter selection for a two-stage variable selection scheme.* Two-stage variable selection techniques discussed in this paper have two tuning parameters: the regularization parameter λ in the first stage and the threshold s from the second stage. Furthermore, given that TVS using different bridge estimators offer the best performance in different regimes, we may see q as another tuning parameter. How can these parameters be optimally tuned in practice? As proved in Section 3, the TVS with an estimator of smaller AMSE in the first stage provides a better variable selection. Hence, the parameter λ can be set by minimizing the estimated risk of the bridge estimator. Similarly, one can estimate the risk for different values of q and choose the one that offers the smallest estimated risk. Section 5.2 has showed how this can be done.

It remains to determine the parameter s. As presented in our results, the threshold s controls the trade-off between AFDP and ATPP. By increasing s, we decrease the number of

false discoveries, but at the same time, we decrease the number of correct discoveries. Therefore, the choice of s depends on the accepted level of false discoveries (or similar quantities). For instance, one can control the false discovery rate by combining the two-stage approach with the knock-off framework [2]. Specifically, if we would like to control FDP at a rate of $\rho \in (0, 1)$, we can go through the following procedure:

1. Construct the knock-off features $\tilde{X} \in \mathbb{R}^{n\times p}$ as stated in [2];
2. Run bridge regression on the joint design $[X, \tilde{X}]$ and obtain the corresponding estimator $\left[\begin{smallmatrix}\hat{\beta}\\ \tilde{\beta}\end{smallmatrix}\right]$. Let $W_j = \max(|\hat{\beta}_j|, |\tilde{\beta}_j|)\operatorname{sign}(|\hat{\beta}_j| - |\tilde{\beta}_j|)$, $j = 1, 2, \ldots, p$. Define the threshold s as $s = \min\{t > 0 : \frac{1+\#\{j:W_j\le -t\}}{\#\{j:W_j\ge t\}\vee 1} \le \rho\}$.
3. Select all the predictors with $\{j : W_j \ge s\}$.

The above procedure only works for $n \ge p$. We may adapt the new knockoff approach in [8] when $n < p$.

7. Conclusion. We studied two-stage variable selection schemes for linear models under the high-dimensional asymptotic setting, where the number of observations n grows at the same rate as the number of predictors p. Our TVS has a bridge estimator in the first stage and a simple threshold function in the second stage. For such schemes, we proved that for a fixed ATPP, in order to obtain the smallest AFDP one should pick an estimator that minimizes the asymptotic mean square error in the first stage of TVS. This connection between parameter estimation and variable selection further led us to a thorough investigation of the AMSE under different regimes including rare and weak signals, small/large noise and large sample. Our analyses revealed several interesting phenomena and provided new insights into variable selection. For instance, the variable selection of LASSO can be improved by debiasing and thresholding; a TVS with ridge in its first stage outperforms TVS with other bridge estimators for large values of noise; the optimality of two-stage LASSO among two-stage bridge estimators holds for very sparse signals until the signal strength is below some threshold. We conducted extensive numerical experiments to support our theoretical findings and validate the scope of our main conclusions for general design matrices.

Acknowledgments. The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments and suggestions that improved the quality of this paper.

SUPPLEMENTARY MATERIAL

Supplement to "Which bridge estimator is the best for variable selection?" (DOI: 10.1214/19-AOS1906SUPP; .pdf). Due to space constraints, additional technical proofs are relegated a supplementary document in [51], which contains Sections A–J.

REFERENCES

[1] AERON, S., SALIGRAMA, V. and ZHAO, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory* **56** 5111–5130. MR2808668 https://doi.org/10.1109/TIT.2010.2059891

[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

[3] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[4] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[5] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612
[6] BOGDAN, M., VAN DEN BERG, E., SU, W. and CANDES, E. J. (2013). Supplementary materials for "Statistical estimation and testing via the sorted l1 norm". *Ann. Statist.*
[7] BUTUCEA, C., NDAOUD, M., STEPANOVA, N. A. and TSYBAKOV, A. B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572
[8] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265
[9] CHO, H. and FRYZLEWICZ, P. (2012). High dimensional variable selection via tilting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 593–622. MR2925375 https://doi.org/10.1111/j.1467-9868.2011.01023.x
[10] DAVID, G. and ILIAS, Z. (2017). High dimensional regression with binary coefficients. Estimating squared error and a phase transition. In *Conference on Learning Theory* 948–953.
[11] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279. MR3766952 https://doi.org/10.1214/17-AOS1549
[12] DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506
[13] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[14] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714
[15] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.
[16] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823
[17] DONOHO, D. L. and TANNER, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102** 9446–9451. MR2168715 https://doi.org/10.1073/pnas.0502269102
[18] EL KAROUI, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *Ann. Statist.* **38** 3487–3566. MR2766860 https://doi.org/10.1214/10-AOS795
[19] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.
[20] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273
[21] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x
[22] FLETCHER, A. K., RANGAN, S. and GOYAL, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory* **55** 5758–5772. MR2597192 https://doi.org/10.1109/TIT.2009.2032726
[23] FRANK, L. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.
[24] GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354
[25] HASTIE, T., TIBSHIRANI, R. and TIBSHIRANI, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso. arXiv preprint arXiv:1707.08692.
[26] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808 https://doi.org/10.1214/009053607000000875
[27] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373
[28] JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180 https://doi.org/10.1214/11-AOS947
[29] JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749 https://doi.org/10.1631/jzus.a1400233

[30] KE, Z. T., JIN, J. and FAN, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202–2242. MR3269978 https://doi.org/10.1214/14-AOS1243
[31] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787 https://doi.org/10.1214/aos/1015957397
[32] LOH, P.-L. and WAINWRIGHT, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.* **45** 2455–2482. MR3737898 https://doi.org/10.1214/16-AOS1530
[33] LUO, S. and CHEN, Z. (2014). Sequential lasso cum EBIC for feature selection with ultra-high dimensional feature space. *J. Amer. Statist. Assoc.* **109** 1229–1240. MR3265693 https://doi.org/10.1080/01621459.2013.877275
[34] MALEKI, A., ANITORI, L., YANG, Z. and BARANIUK, R. G. (2013). Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Trans. Inf. Theory* **59** 4290–4308. MR3071330 https://doi.org/10.1109/TIT.2013.2252232
[35] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281
[36] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351 https://doi.org/10.1214/07-AOS582
[37] MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. *Monographs on Statistics and Applied Probability* **95**. CRC Press/CRC, Boca Raton, FL. MR2001193 https://doi.org/10.1201/9781420035933
[38] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2018). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **46** 119–148. MR3766948 https://doi.org/10.1214/17-AOS1544
[39] NDAOUD, M. and TSYBAKOV, A. B. (2018). Optimal variable selection and adaptive noisy compressed sensing. arXiv preprint arXiv:1809.03145.
[40] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131 https://doi.org/10.1007/s10208-015-9278-4
[41] RAD, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inf. Theory* **57** 4672–4679. MR2840483 https://doi.org/10.1109/TIT.2011.2145670
[42] RANGAN, S., GOYAL, V. and FLETCHER, A. K. (2009). Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems* 1545–1553.
[43] REEVES, G. and GASTPAR, M. C. (2013). Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inf. Theory* **59** 3451–3465. MR3061258 https://doi.org/10.1109/TIT.2013.2253852
[44] SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. MR3718164 https://doi.org/10.1214/16-AOS1521
[45] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116
[46] SUR, P., CHEN, Y. and CANDÈS, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a *rescaled* chi-square. *Probab. Theory Related Fields* **175** 487–558. MR4009715 https://doi.org/10.1007/s00440-018-00896-9
[47] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
[48] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018
[49] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816
[50] WANG, S., WENG, H. and MALEKI, A. (2020). Which bridge estimator is optimal for variable selection? arXiv preprint arXiv:1705.08617.
[51] WANG, S., WENG, H. and MALEKI, A. (2020). Supplement to "Which bridge estimator is the best for variable selection?" https://doi.org/10.1214/19-AOS1906SUPP
[52] WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. MR2683451 https://doi.org/10.1109/TIT.2010.2046199
[53] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646
[54] WENG, H., FENG, Y. and QIAO, X. (2019). Regularization after retention in ultrahigh dimensional linear regression models. *Statist. Sinica* **29** 387–407. MR3889373

[55] WENG, H., MALEKI, A. and ZHENG, L. (2018). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Ann. Statist.* **46** 3099–3129. MR3851766 https://doi.org/10.1214/17-AOS1651

[56] YANG, E., LOZANO, A. and RAVIKUMAR, P. (2014). Elementary estimators for high-dimensional linear regression. In *International Conference on Machine Learning* 388–396.

[57] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

[58] ZHANG, T. (2009). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109–2144. MR2543687 https://doi.org/10.1214/08-AOS659

[59] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[60] ZHENG, L., MALEKI, A., WENG, H., WANG, X. and LONG, T. (2017). Does ℓ_p-minimization outperform ℓ_1-minimization? *IEEE Trans. Inf. Theory* **63** 6896–6935. MR3724407 https://doi.org/10.1109/TIT.2017.2717585

[61] ZHOU, S. (2009). Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing Systems* 2304–2312.

[62] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735