Uniform Asymptotic Inference and the Bootstrap After Model Selection

Ryan Tibshirani, Alessandro Rinaldo, Rob Tibshirani, and Larry Wasserman

Carnegie Mellon University and Stanford University

Abstract

Recently, Tibshirani et al. (2016) proposed a method for making inferences about parameters defined by model selection, in a typical regression setting with normally distributed errors. Here, we study the large sample properties of this method, without assuming normality. We prove that the test statistic of Tibshirani et al. (2016) is asymptotically valid, as the number of samples n grows and the dimension d of the regression problem stays fixed. Our asymptotic result holds uniformly over a wide class of nonnormal error distributions. We also propose an efficient bootstrap version of this test that is provably (asymptotically) conservative, and in practice, often delivers shorter intervals than those from the original normality-based approach. Finally, we prove that the test statistic of Tibshirani et al. (2016) does not enjoy uniform validity in a high-dimensional setting, when the dimension d is allowed grow.

1 Introduction

There has been a recent surge of work on conducting formally valid inference in a regression setting after a model selection event has occurred, see Berk et al. (2013), Lockhart et al. (2014), Tibshirani et al. (2016), Lee et al. (2016), Fithian et al. (2014), Bachoc et al. (2014), just to name a few. Our interest in this paper stems in particular from the work of Tibshirani et al. (2016), who presented a method to produce valid p-values and confidence intervals for adaptively fitted coefficients from any given step of a sequential regression procedure like forward stepwise regression (FS), least angle regression (LAR), or the lasso (the lasso is meant to be thought of as tracing out a sequence of models along its solution path, as the penalty parameter descends from $\lambda = \infty$ to $\lambda = 0$). These authors use a statistic that is carefully crafted to be pivotal after conditioning on the model selection event. This idea is not specific to the sequential regression setting, and is an example of a broader framework that we might call *selective pivotal inference*, applicable to many other settings, e.g., developed in Taylor et al. (2016), Lee et al. (2016), Lee & Taylor (2014), Loftus & Taylor (2014), Reid et al. (2014), Choi et al. (2014), Fithian et al. (2014), Hyun et al. (2016).

A key to the proposal in Tibshirani et al. (2016) (and much of the work in selective pivotal inference) is to assume normality of the errors. To fix notation, consider the regression of a response $Y \in \mathbb{R}^n$ on predictor variables $X_1, \ldots, X_d \in \mathbb{R}^n$, stacked together as columns of a matrix $X \in \mathbb{R}^{n \times d}$. We will treat the predictors X are fixed (nonrandom), and assume that the response follows the model

$$Y_i = \theta_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\theta \in \mathbb{R}^n$ is an unknown mean parameter of interest. Tibshirani et al. (2016) assume that the errors $\epsilon_1, \ldots \epsilon_n$ are i.i.d. $N(0, \sigma^2)$, with the error variance $\sigma^2 > 0$ known. An advantage of their approach is that it does not require θ to be an exact linear combination of the predictors $X_1, \ldots, X_d$, and makes no assumptions about the correlations among these predictors. But as far as the finite-sample guarantees are concerned, normality of the errors is crucial. In this work, we examine the properties of the test statistic proposed in Tibshirani et al. (2016)—hereafter, the *truncated Gaussian* (TG) statistic—without using an assumption about normal errors. We only assume that $\epsilon_1, \ldots \epsilon_n$ are i.i.d. from a distribution with mean zero and essentially no other restrictions.

A high-level description of the selective pivotal inference framework for sequential regression is as follows (details are provided in Section 2). FS, LAR, or the lasso is run for some number of steps k, and a model is selected, call it M. For FS and LAR, this model will always have k active variables, and for the lasso, it will have at most k, as variables can be added to or deleted from the active set at each step. We specify a linear contrast of the mean $v^T\theta$ of interest, e.g., one giving the coefficient of a variable of interest in the model M at step k, in the regression of θ onto the active variables. By assuming normal errors in (1), and examining the distribution of $v^T Y$ conditional on having selected model M, which we denote by $\widehat{M}(Y) = M$, we can construct a confidence interval C_α satisfying

$$\mathbb{P}\Big(v^T\theta \in C_\alpha \,\Big|\, \widehat{M}(Y) = M\Big) = 1 - \alpha,$$

for a given $\alpha \in [0,1]$. The interpretation: if we were to repeatedly draw Y from (1) and run FS, LAR, or the lasso for k steps, and only pay attention to cases in which we selected model M, then among these cases, the constructed intervals $C_\alpha = C_\alpha(Y; M)$ contain $v^T\theta$ with frequency tending to $1-\alpha$.

The above is a *conditional* perspective of the selective pivotal inference approach for FS, LAR, and lasso. Asymptotic analysis turns out to be more tractable from an *unconditional* point of view, which we now describe. For each possible selected model M, a constrast vector v_M is specified, and the contrast $v_M^T\theta$ is considered when model M is selected, $\widehat{M}(Y) = M$. To be concrete, we can again think of a setup such that $v_M^T\theta$ gives the coefficient of a variable of interest in the model M at step k, in the projection of θ onto the active set. Confidence intervals are constructed in exactly the same manner as above (without change), and conditional coverage over all models M implies the following unconditional property for C_α,

$$\mathbb{P}\Big(v_{\widehat{M}(Y)}^T\theta \in C_\alpha\Big) = 1 - \alpha.$$

The interpretation is different: if we were to repeatedly draw Y from (1) and run FS, LAR, or lasso for k steps, and construct confidence intervals $C_\alpha = C_\alpha(Y; \widehat{M}(Y))$, then these intervals contain their respective targets $v_{\widehat{M}(Y)}^T\theta$ with frequency approaching $1-\alpha$. Note that, by construction, the target itself may change each time we draw Y, though it is the same for all Y that give rise to the same selected model. In terms of the setting for regression contrasts described above, each time we draw Y and carry out the inferential procedure, the interval C_α covers the coefficient of a possibly different variable in the active model, in the projection of θ onto the active variables. Figure A.1, deferred to Appendix A.1 of the online supplement, demonstrates this point.

1.1 Uniform convergence

When making asymptotic inferential guarantees, as we do in this paper, it is important to be clear about the type of guarantee. Here we review the concepts of uniform convergence and validity. Let $\xi_1, \ldots, \xi_n \in \mathbb{R}^s$ be random vectors with joint distribution $(\xi_1, \ldots, \xi_n) \sim F_n$, where $F_n \in \mathcal{P}_n$, and $\mathcal{P}_n$ is a class of distributions. For example, we could have $\xi_1, \ldots, \xi_n \in \mathbb{R}^s$ i.i.d. from F, and the class $\mathcal{P}_n$ could

contain product distributions of the form $F_n = F \times \ldots \times F$ (n times); our notation allows for a more general setup than this one. Let $W_n = T_n(\xi_1, \ldots, \xi_n)$ for a statistic T_n, and $W \sim G$, where $W_n, W \in \mathbb{R}^q$. We will say that W_n, converges *uniformly in distribution* to W, over $\mathcal{P}_n$, provided that

$$\lim_{n\to\infty} \sup_{F_n \in \mathcal{P}_n} \sup_{x \in \mathbb{R}^q} \left| \mathbb{P}_{F_n}(W_n \leq x) - \mathbb{P}(W \leq x) \right| = 0. \tag{2}$$

(The above inequalities, as in $W_n \leq x$ and $W \leq x$, are meant to be interpreted componentwise; we are also implicitly assuming that the limiting distribution G is continuous, otherwise the above inner supremum should be restricted to continuity points x of G.) This is much stronger than the notion of *pointwise* convergence in distribution, which only requires that

$$\lim_{n\to\infty} \sup_{x \in \mathbb{R}^q} \left| \mathbb{P}_{F_n}(W_n \leq x) - \mathbb{P}(W \leq x) \right| = 0, \tag{3}$$

for a particular sequence of distributions F_n, $n = 1, 2, 3, \ldots$.

A recent article by Kasy (2015) emphasizes the importance of uniformity in asymptotic approximations. This authors points out that uniform versions of the continuous mapping theorem and the central limit theorem for triangular arrays follow from standard proofs of these results (e.g., following their proofs in van der Vaart (1998)). For convenience, these basic uniform convergence results are transcribed in Appendix A.2.

In our work, a motivating reason for the study of uniform convergence is the associated property of *uniform validity* of asymptotic confidence intervals. That is, if $W_n = W_n(\mu)$ depends on a parameter $\mu = \mu(F_n)$ of the distribution F_n, but W does not, then we can consider any $(1-\alpha)$ confidence set $C_{n,\alpha}$ built from a $(1-\alpha)$ probability rectangle R_α of W,

$$C_{n,\alpha} = \{\mu : W_n(\mu) \in R_\alpha\},$$

and the uniform convergence of W_n to W, really just by rearranging its definition in (2), implies

$$\lim_{n\to\infty} \sup_{F_n \in \mathcal{P}_n} \sup_{\alpha \in [0,1]} \left| \mathbb{P}_{F_n}\Big(\mu(F_n) \in C_{n,\alpha}\Big) - (1-\alpha) \right| = 0. \tag{4}$$

Meanwhile, pointwise convergence as in (3) only implies

$$\lim_{n\to\infty} \sup_{\alpha \in [0,1]} \left| \mathbb{P}_{F_n}\Big(\mu(F_n) \in C_{n,\alpha}\Big) - (1-\alpha) \right| = 0, \tag{5}$$

for a particular sequence F_n, $n = 1, 2, 3, \ldots$. For a confidence set satisfying (4), and a given tolerance $\epsilon > 0$, there exists a sample size $n(\epsilon)$ such that the coverage is guaranteed to be at least $1 - \alpha - \epsilon$, for $n \geq n(\epsilon)$, no matter the underlying distribution (over the class of distributions in question). Note that this is not necessarily true for a pointwise confidence set as in (5), as the required sample size here could depend on the particular distribution under consideration.

1.2 Summary of main results

An overview of our main contributions is as follows.

1. We describe a new unconditional perspective for selective pivotal inference (Section 3), and we establish that the TG statistic and the FS, LAR, and lasso paths only depend on the data (X, Y) through $\frac{1}{n}X^TX$ and $\frac{1}{\sqrt{n}}X^TY$ (Lemmas 1, 2, and 3), which is important because the latter two quantities have asymptotic limits.

2. Placing mild constraints on the mean and error distribution in (1), and treating the dimension d as fixed, we prove that the TG test statistic is asymptotically pivotal, converging to $U(0,1)$ (the standard uniform distribution), when evaluated at the true population value for its pivot argument. We show that this holds uniformly over a wide class of distributions for the errors, without any real restrictions on the predictors X (first part of Theorem 5).

3. The resulting confidence intervals are therefore asymptotically uniformly valid, over the same class of distributions (second part of Theorem 5).

4. The above asymptotic results assume that the error variance σ^2 is known, so for σ^2 unknown, we propose a plug-in approach that replaces σ^2 in the TG statistic with a simple estimate, and alternatively, an efficient bootstrap approach. Both allow for conservative asymptotic inference (Theorem 9).

5. We present detailed numerical experiments that support the asymptotic validity of the TG p-values and confidence intervals for inference in low-dimensional regression problems that have nonnormal errors (Section 6). Our experiments reveal that the plug-in and bootstrap versions also show good performance, and the bootstrap method can often deliver substantially shorter intervals than those based directly on the TG statistic.

6. Our experiments also also suggest that the TG test statistic (and plug-in, bootstrap variants) may be asymptotically valid in even broader settings not covered by our theory, e.g., problems with heteroskedastic errors and (some) high-dimensional problems.

7. We prove that TG statistic does not exhibit a general uniform convergence to $U(0,1)$ when the dimension d is allowed to increase (Theorem 10).

1.3 Related work

A recent paper by Tian & Taylor (2015) is very related to our work here. These authors examine the asymptotic distribution of the TG statistic under nonnormal errors. Their main result proves that the TG statistic is asymptotically pivotal, under some restrictions on the model selection events in question. We view their work as providing a complementary perspective to our own: they consider a setting where the dimension d grows, but place strong regularity conditions on the selected models; we adopt a more basic setting with d fixed, and prove more broad uniformly valid convergence results for the TG pivot, free of regularity conditions.

In a sequence of papers, Leeb & Potscher (2003, 2006, 2008) prove that in a classical regression setting, it is impossible to find the distribution of a post-selection estimator of the underlying coefficients, even asymptotically. Specifically, they prove for an estimate $\widehat{\beta}$ of some underlying coefficient vector β_0, the usual pivot $Q_n = \sqrt{n}(\widehat{\beta} - \beta_0)$ cannot be used for inference after model selection. Though Q_n, once appropriately scaled, is pivotal (or at least asymptotically pivotal), this is no longer true in the presence of selection, even if the dimension d is fixed and the sample size n approaches ∞. Furthermore, they show that there is no uniformly consistent estimate of the distribution of Q_n (either conditionally or unconditionally), which makes Q_n unsuitable for inference. This fact is essentially a manifestation of the well-known Hodges phenomenon. The selective pivotal framework, and hence the perspective of our paper, avoids this problem for the following reason: this method is does not claim (nor attempt) to estimate the distribution of Q_n whatsoever, and makes inferences based on an entirely different pivotal quantity that is constructed via a clever conditioning scheme.

1.4 Notation

As our paper considers an asymptotic regime, with the number of samples n growing, we will often use a subscript n to mark the dependence of various quantities on the sample size. An exception is our notation for the predictors, response, and mean, which we will always denote by X, Y, θ, respectively. Though these quantities will (of course) vary with n, our notation hides this dependence for simplicity.

When it comes to probability statements involving Y, drawn from (1), we will write $\mathbb{P}_{f(\theta)=\mu}(\cdot)$ to denote the probability operator under a mean vector θ such that $f(\theta) = \mu$. With a subscript omitted, as in $\mathbb{P}(\cdot)$, it is implicit that the probability is taken under θ. Also, we will generally write y (lowercase) for an arbitrary response vector, and Y (uppercase) for a random response vector drawn from (1). This is intended to distinguish statements that hold for an arbitrary y, and statements that hold for a random Y with a certain distribution. Lastly, we will denote $\widehat{M}$ the model selection procedure associated with the regression algorithm under consideration (FS, LAR, or lasso), and we will treat this as a mapping from $\mathbb{R}^n$ to the space of models, so that $\widehat{M}(y)$ is a fixed quantity, representing the model selected when the response is the fixed vector y, and $\widehat{M}(Y)$ is a random variable, representing the model selected when the response is the random vector Y. Similar notation will be used for related quantities.

2 Conditional inference

In this section, we describe the selective pivotal inference framework for sequential regression procedures. We present this framework from a conditional point of view; in a sense, this is the simplest way to portray the ideas of inference after model selection.

2.1 Model selection

Consider forward stepwise regression (FS), least angle regression (LAR), or the lasso, run for a number of steps k, where k is arbitrary (but treated as fixed throughout this paper). Such a procedure defines a *partition* of the sample space, $\mathbb{R}^n = \bigcup_{M \in \mathcal{M}} \Pi_M$, with elements

$$\Pi_M = \{y : \widehat{M}(y) = M\}, \quad M \in \mathcal{M}. \tag{6}$$

Here $\widehat{M}(y)$ denotes the *selected model* from the given k-step procedure, run on y, and $\mathcal{M}$ is the space of possible models. Calling $\widehat{M}(Y)$ a selected model may be bit of an abuse of common nomenclature, because, as we will see, $\widehat{M}(y)$ will describe more than just a set of selected variables at the point y. In fact, one can think of $\widehat{M}(y)$ as a representation of the decisions made by the algorithm across its k steps. For FS, we define $\widehat{M}(y) = \{(\widehat{A}_\ell(y), \widehat{s}_\ell(y)) : \ell = 1, \ldots, k\}$, comprised of two things:

1. a sequence of active sets $\widehat{A}_\ell(y)$, $\ell = 1, \ldots, k$, denoting the variables that are given nonzero coefficients, at each of the k steps;

2. a sequence of sign vectors $\widehat{s}_\ell(y)$, $\ell = 1, \ldots, k$, denoting the signs of nonzero coefficients, at each of the k steps.

The active sets are nested across steps, $\widehat{A}_1(y) \subseteq \widehat{A}_2(y) \subseteq \widehat{A}_3(y) \subseteq \ldots$, as FS selects one variable to add to the active set at each step. However, the sign vectors $\widehat{s}_1(y), \widehat{s}_2(y), \widehat{s}_3(y), \ldots$ are not, since these are

determined by least squares on the active variables at each step. Hence, as defined, the number of possible models $\widehat{M}(y)$ after k steps of FS is

$$|\mathcal{M}| = d \cdot (d-1) \cdots (d-k+1) \cdot 2 \cdot 2^2 \cdots 2^k = O(d^k 2^{k^2}).$$

Moreover, the corresponding partition elements Π_M, $M \in \mathcal{M}$ in (6) are all convex cones. The proof of this fact is not difficult, and requires only a slight modification of the arguments in Tibshirani et al. (2016), given in Appendix A.3 for completeness. The result is easily seen for $k = 1$: after one step of FS, assuming without a loss of generality that $X_1, \ldots, X_d$ have unit norm, we can express, e.g.,

$$\begin{aligned} \big\{y : \big(\widehat{A}_1(y), \widehat{s}_1(y)\big) = (1,1)\big\} &= \big\{y : X_1^T y \geq \pm X_j^T y,\, j = 2, \ldots, d\big\} \\ &= \bigcap_{j=2}^{d} \big\{y : (X_1 - X_j)^T y \geq 0\big\} \cap \big\{y : (X_1 + X_j)^T y \geq 0\big\}, \end{aligned}$$

the right-hand side above being an intersection of half-spaces passing through zero, and therefore a convex cone. As we enumerate the possible choices for $(\widehat{A}_1(y), \widehat{s}_1(y))$, these cones form a partition of $\mathbb{R}^n$. Figure 1 shows an illustration.

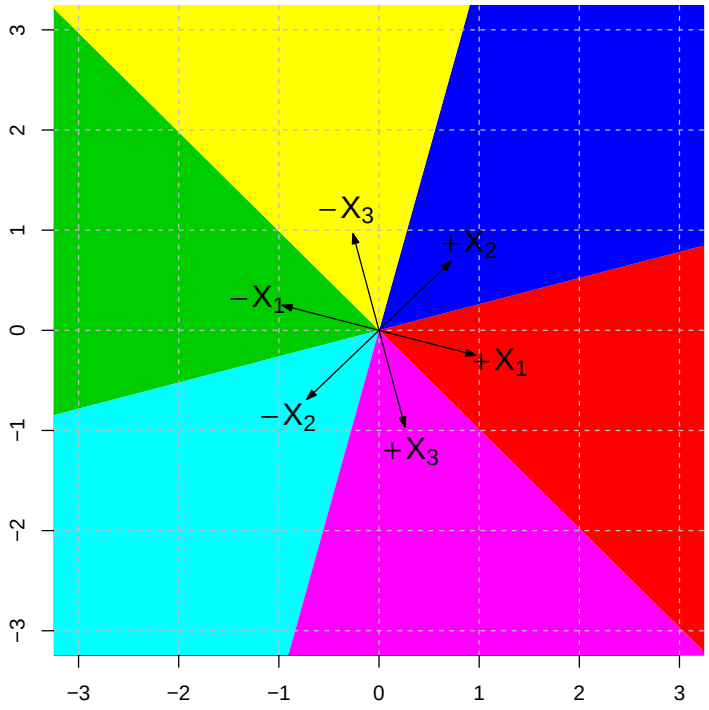


Figure 1: *An example of the model selection partition from one step of FS (the variables are normalized, and this is equivalent to one step of LAR, or lasso). Here $n = 2$ and $d = 3$. The colors indicate the regions of the sample space $\mathbb{R}^2$ for which different models—pairs of active variables and signs—are selected, so that, e.g., the red region contains points in $\mathbb{R}^2$ that are maximally aligned with X_1.*

For LAR and the lasso, we need to modify the definition of the selected model $\widehat{M}(y)$ in order for the resulting partition elements in (6) to be convex cones. We add an "extra" bit of model information

and define $\widehat{M}(y) = \{(\widehat{A}(y), \widehat{s}(y), \widehat{I}_\ell(y)) : \ell = 1, \ldots, k\}$, where $\widehat{I}_\ell(y)$ is a list of variables that play a special role in the construction of the LAR or lasso active set at the ℓth step, but that a user would not typically pay attention to. In truth, the latter quantity is only a detail that is included so that Π_M, $M \in \mathcal{M}$ are convex cones (without it, the partition elements would each be a union of cones), and so we do not describe it here. Furthermore, it does not affect our treatment of inference in what follows, and for this reason, we will largely ignore the minor differences in model selection events between FS, LAR, and lasso hereafter.

The description of $\widehat{I}_\ell(y)$, $\ell = 1, \ldots, k$, and the proof that the partition elements Π_M, $M \in \mathcal{M}$ are cones for LAR and lasso, mirrors that in Tibshirani et al. (2016), and is again given in Appendix A.3. Like FS, the active sets from LAR are nested, $\widehat{A}_1(y) \subseteq \widehat{A}_2(y) \subseteq \widehat{A}_3(y) \subseteq \ldots$, since one variable is added to the active set at each step. But for the lasso, this is not necessarily true, as in this case variables can be either added or deleted at each step.

2.2 Testing after selection

We review the selective pivotal inference approach for hypothesis testing after model selection with FS, LAR, or the lasso. The technical details of the TG statistic are deferred to the next two subsections, as they are not needed to understand how the method is used. The null hypotheses we consider are of the form $H_0 : v^T\theta = 0$. An important special case occurs when the linear contrast $v^T\theta$ gives a normalized coefficient in the regression of θ onto a subset of the variables in X. To be specific, in this case $v = X_A(X_A^TX_A)^{-1}e_j/(e_j^T(X_A^TX_A)^{-1}e_j)^{1/2}$, for a subset $A \subseteq \{1, \ldots, d\}$, where we let $X_A \in \mathbb{R}^{n\times|A|}$ denote the submatrix of X whose columns correspond to elements of A (with $X_A^TX_A$ assumed to be invertible for the chosen subset), and we write e_j for the jth standard basis vector. This gives

$$v^T\theta = \frac{e_j^T(X_A^TX_A)^{-1}X_A^T\theta}{\sqrt{e_j^T(X_A^TX_A)^{-1}e_j}} := \beta_j(A), \tag{7}$$

and therefore $H_0 : v^T\theta = 0$ is a test for the significance of the jth normalized coefficient in the linear projection of θ onto X_A, written as $\beta_j(A)$ for short. (Though the normalization in the denominator is irrelevant for this significance test, it acts as a key scaling factor for the asymptotics in Section 4.) The idea of using a projection parameter for inference, $\beta_j(A)$, has also appeared in, e.g., Berk et al. (2013), Wasserman (2014), Lee et al. (2016). Here is now a summary of the testing framework.

- For each possible model $M \in \mathcal{M}$, and any $v \in \mathbb{R}^n$ and $\mu \in \mathbb{R}$, a TG statistic $T(\cdot; M, v, \mu)$ is defined (see (10), in the next subsection), whose domain is the partition element Π_M. This can be used as follows: if Y is drawn from (1), and lands in the partition element Π_M for model M, then the statistic $T(Y; M, v, \mu)$ provides us with a test for the hypothesis $H_0 : v^T\theta = \mu$.

- A concrete case to keep in mind, denoting $M = \{(A_\ell, s_\ell) : \ell = 1, \ldots, k\}$, is a choice of v such that $v^T\theta = \beta_j(A_\ell)$, in the notation of (7). This is the jth normalized coefficient in the regression of θ onto the active variables X_{A_ℓ}, for an active set A_ℓ at some step $\ell = 1, \ldots, k$.

- Assume i.i.d. $N(0, \sigma^2)$ errors in (1). Under the null hypothesis, the TG statistic has a standard uniform distribution, over draws of Y that land in Π_M. Mathematically, this is the property

$$\mathbb{P}_{v^T\theta=\mu}\Big(T(Y; M, v, \mu) \leq t \,\Big|\, \widehat{M}(Y) = M\Big) = t, \tag{8}$$

for all $t \in [0,1]$. The probability above is taken over an arbitrary mean parameter θ for which $v^T\theta = \mu$ (in fact, the TG statistic is constructed so that the law of $T(Y;M,v,\mu)\,|\,\widehat{M}(Y) = M$ only depends on θ through $v^T\theta$, so this is unambiguous). In order for (8) to hold, of course, v and μ cannot be random, i.e., they cannot depend on Y, though they can be functions of M.

- Thus $T(Y;M,v,\mu)$ serves as a valid p-value (with exact finite sample size) for testing the null hypothesis $H_0 : v^T\theta = \mu$, conditional on $\widehat{M}(Y) = M$.

- A confidence interval is obtained by inverting the test in (8). Given a desired confidence level $1-\alpha$, we define C_α to be the set of all values μ such that $\alpha/2 \le T(Y;M,v,\mu) \le 1-\alpha/2$. Then, by construction, the property in (8) translates into

$$\mathbb{P}\Big(v^T\theta \in C_\alpha \,\Big|\, \widehat{M}(Y) = M\Big) = 1-\alpha. \tag{9}$$

  The interpretation of the above statement is straightforward: the random interval C_α contains the fixed parameter $v^T\theta$ with probability $1-\alpha$, conditional on $\widehat{M}(Y) = M$.

We reiterate that the properties (8), (9) assume i.i.d. $N(0,\sigma^2)$ errors in (1), and our goal in this paper is to establish analogous asymptotic properties without a normal error model. The conditional perspective described here, however, where each inferential statement is conditioned in the event $\widehat{M}(Y) = M$, turns out to be harder to study asymptotically than an unconditional version. Therefore, later in Section 3, we cast the testing framework for sequential regression in an unconditional light.

2.3 The truncated Gaussian pivot

We now review the truncated Gaussian (TG) pivotal quantity. As defined in Section 2.1, if we write $\widehat{M}(y)$ for the selected model from the given algorithm (FS, LAR, or lasso), run for k steps on y, then $\Pi_M = \{y : \widehat{M}(y) = M\}$ is a convex cone, for any fixed achieveable model M. Hence

$$\Pi_M = \{y : \widehat{M}(y) = M\} = \{y : Q_M\, y \ge 0\},$$

for some fixed matrix Q_M (here the inequality is meant to be interpreted componentwise). The pivot $T(\cdot;M,v,\mu)$ for testing $H_0 : v^T\theta = \mu$ can be defined by

$$T(y;M,v,\mu) = \frac{\Phi\Big(\dfrac{b(y;M,v)-\mu}{\sigma\|v\|_2}\Big) - \Phi\Big(\dfrac{v^Ty-\mu}{\sigma\|v\|_2}\Big)}{\Phi\Big(\dfrac{b(y;M,v)-\mu}{\sigma\|v\|_2}\Big) - \Phi\Big(\dfrac{a(y;M,v)-\mu}{\sigma\|v\|_2}\Big)}. \tag{10}$$

which is the evaluation of the truncated Gaussian survival function at v^Ty, for specific truncation limits $a(y;M,v), b(y;M,v)$ defined in Appendix A.4. This pivot has the following property, as stated in (8): when Y is drawn from (1) with i.i.d. $N(0,\sigma^2)$ errors, and $v^T\theta = \mu$, the pivot $T(Y;M,v,\mu)$ is uniformly distributed conditional on $\widehat{M}(Y) = M$. See Lemmas 1 and 2 in Tibshirani et al. (2016) for a proof of this result.

P-values and confidence intervals follow directly from the construction of the pivot above. For the null hypothesis $H_0 : v^T\theta = 0$, we prefer the one-sided p-value $T(Y;M,v,0)$, as in (8). For confidence intervals, we prefer the two-sided interval given by inverting the two-sided pivot $2\min\{T(Y;M,v,\mu),\, 1-T(Y;M,v,\mu)\}$, as in (9). Further discussion in deferred until Appendx A.5.

3 Unconditional inference

We have portrayed selective pivotal inference, in sequential regression procedures, as a method for producing conditional p-values and intervals. But an unconditional interpretation is also possible, and in fact, this view ends up being crucial for the development of uniform asymptotic results.

3.1 Testing after selection, revisited

We describe an unconditional picture for our testing approach with sequential regression procedures.

- For each model $M \in \mathcal{M}$, a contrast vector $v_M \in \mathbb{R}^n$ and pivot value $\mu_M \in \mathbb{R}$ are identified, so that the hypothesis $H_{0,M} : v_M^T \theta = \mu_M$ is to be tested whenever $y \in \Pi_M$, i.e., whenever $\widehat{M}(y) = M$. A TG statistic $\mathcal{T}(\cdot; V, U)$ is then defined, whose domain is the entire sample space $\mathbb{R}^n$. Here we write $V = \{v_M : M \in \mathcal{M}\}$ and $U = \{\mu_M : M \in \mathcal{M}\}$ to denote the collection of contrast vectors and pivot values, respectively, across partition elements—we will also refer to these as *catalogs*. This unconditional TG statistic is defined by

$$\mathcal{T}(\cdot; V, U) = \sum_{M \in \mathcal{M}} T(\cdot; M, v_M, \mu_M) 1_{\Pi_M}(\cdot),$$

  where $1_{\Pi_M}(\cdot)$ denotes the indicator function for the partition element Π_M (and $T(\cdot; M, v_M, \mu_M)$ is as before, defined in (10)). The unconditional statistic can be used as follows: if a response Y is drawn from (1), then we can form $\mathcal{T}(Y; V, U) = T(Y; \widehat{M}(Y), v_{\widehat{M}(y)}, \mu_{\widehat{M}(y)})$ to test the hypothesis $H_0 : v_{\widehat{M}(Y)}^T \theta = \mu_{\widehat{M}(Y)}$.

- A concrete case to keep in mind is when V assigns a contrast vector v_M to each model M, such that $v_M^T \theta = \beta_{j_M}(A_{\ell_M})$, in the notation of (7), where $M = \{(A_\ell, s_\ell) : \ell = 1, \ldots, k\}$ as usual. This is the j_Mth normalized coefficient from projecting θ onto $X_{A_{\ell_M}}$, the active variables at step ℓ_M.

- Assume that the errors in (1) are i.i.d. $N(0, \sigma^2)$. Then under the proper hypothesis, by summing up the conditional property in (8) across partition elements, we have

$$\mathbb{P}_{V^T\theta = U}\Big(\mathcal{T}(Y; V, U) \le t\Big) = t, \tag{11}$$

  for all $t \in [0, 1]$. The assertion above holds for a parameter θ such that $V^T \theta = U$, which we use as shorthand for $v_M^T \theta = \mu_M$ for all $M \in \mathcal{M}$. Note that this full specification, across all $M \in \mathcal{M}$, is critical in order to apply the relevant null probability within each partition element (giving rise to the equality in (11)).

- Therefore $\mathcal{T}(Y; V, U)$ serves as a valid p-value (with exact finite sample size)—but for testing what null hypothesis? Formally, it is attached to $H_0 : V^T \theta = U$, an exhaustive specification of $v_M^T \theta = \mu_M$, over all $M \in \mathcal{M}$, but in truth, $\mathcal{T}(Y; V, U)$ carries no information about models other than the selected one, $\widehat{M}(Y)$. For this reason, we actually consider $\mathcal{T}(Y; V, U)$ to be a p-value for the *random* null hypothesis $H_0 : v_{\widehat{M}(Y)}^T \theta = \mu_{\widehat{M}(Y)}$. This is made more precise through confidence intervals.

- A confidence interval is obtained by inverting the test in (11). But the TG statistic at Y,

$$\mathcal{T}(Y; V, U) = \sum_{M \in \mathcal{M}} T(Y; M, v_M, \mu_M) 1_{\Pi_M}(Y) = T\big(Y; \widehat{M}(Y), v_{\widehat{M}(Y)}, \mu_{\widehat{M}(Y)}\big),$$

only depends on U through $\mu_{\widehat{M}(Y)}$. Thus, given a desired confidence level $1-\alpha$, let us define D_α to be the set of U such that $\alpha/2 \le \mathcal{T}(Y;V,U) \le 1-\alpha/2$, and C_α to be the set of $\mu_{\widehat{M}(Y)}$ such that $\alpha/2 \le T(Y;\widehat{M}(Y),v_{\widehat{M}(Y)},\mu_{\widehat{M}(Y)}) \le 1-\alpha/2$. Then we can see that

$$U \in D_\alpha \iff \mu_{\widehat{M}(Y)} \in C_\alpha,$$

so inverting the test in (11) yields

$$\mathbb{P}\Big(v_{\widehat{M}(Y)}^T \theta \in C_\alpha\Big) = 1-\alpha. \tag{12}$$

The above expression says that the random interval C_α traps the random parameter $v_{\widehat{M}(Y)}^T \theta$ with probability $1-\alpha$. This supports the interpretation of $H_0 : v_{\widehat{M}(Y)}^T \theta = \mu_{\widehat{M}(Y)}$ as the null hypothesis underlying the unconditional TG statistic.

Remark 1. The pivotal property in (11) is derived under the distributional assumption that $V^T\theta = U$, i.e., $v_M^T\theta = \mu_M$ for all $M \in \mathcal{M}$, which may seem unnatural. After all, the catalog U of pivot values is large, $|\mathcal{M}|$-dimensional (where recall $|\mathcal{M}|$ is on the order of d^k after k steps of FS), so this is a condition on (typically) many of contrasts of θ. Our use of U in the above and in what follows is just a notational formality used to derive the unconditional confidence interval property in (12), and the reader *should not view it as a requirement to fully specify U in order to use this inference framework*. Here are two alternative interpretations for (11). First, note that we can rewrite (11) as

$$\mathbb{P}\Big(\mathcal{T}(Y;V,V^T\theta) \le t\Big) = t, \tag{13}$$

for all $t \in [0,1]$. This says that the TG statistic has the "right" distribution (uniform) when we plug in the *true population parameter* $V^T\theta$ as its pivot argument. This is indeed the defining property of any pivotal statistic. The statement (13) may stand out because the population parameter $V^T\theta$ is multidimensional, and its domain could be a strict subset of $\mathbb{R}^{|\mathcal{M}|}$ (meaning some catalogs U are not compatible with V), but this is still well within the realm of the classic notion of a pivotal statistic in hypothesis testing. Second, note that we can rewrite (11) as

$$\sum_{M \in \mathcal{M}} \mathbb{P}_{v_M^T\theta = \mu_M}\Big(T(Y;M,v_M,\mu_M) \le t \,\Big|\, \widehat{M}(Y) = M\Big)\mathbb{P}\big(\widehat{M}(Y) = M\big) = t, \tag{14}$$

for all $t \in [0,1]$. The interpretation is that the TG tests of $H_{0,M} : v_M^T\theta_0 = \mu_M$, $M \in \mathcal{M}$ have the correct conditional size in a suitable *weighted average sense*, where the weights $w_M = \mathbb{P}(\widehat{M}(Y) = M)$, $M \in \mathcal{M}$ are given by the model selection probabilities.

Finally, it is worth emphasizing once again that the testing property in (11) is written in such a way that it is easy to establish the confidence interval property in (12). Thus, a third way to address any concerns about interpreting (11) (or even (13) or (14)) is to switch the focus from unconditional hypothesis testing to unconditional confidence intervals; in many ways, we find the latter the more natural of the two perspectives, from an unconditional point of view.

3.2 The master statistic

Given a response y and predictors X, our description thus far of the selected model $\widehat{M}(y)$, statistics $T(y;M,v,\mu)$ and $\mathcal{T}(y;V,U)$, etc., has ignored the role of X. This was done for simplicity. The theory to come in Section 4 will consider X to be nonrandom, but asymptotically X must (of course) grow

with n, and so it will help to be precise about the dependence of the selected model and statistics on X. We will denote these quantities by $\widehat{M}(X,y)$, $T(X,y;M,v,\mu)$, and $\mathcal{T}(X,y;V,U)$ to emphasize this dependence. We define

$$\Omega_n = \Big(\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^Ty\Big),$$

a $d(d+3)/2$-dimensional quantity that we will call the *master statistic*. As its name might suggest, this plays an important role: all normalized coefficients from regressing y onto subsets of the variables X can be written in terms of Ω_n. That is, for an arbitrary set $A \subseteq \{1,\ldots,p\}$, the jth normalized coefficient from the regression of y onto X_A is

$$\frac{(e_j^T X_A^T X_A)^{-1} X_A^T y}{\sqrt{e_j^T (X_A^T X_A)^{-1} e_j}} = \frac{e_j^T n (X_A^T X_A)^{-1} \frac{1}{\sqrt{n}} X_A^T y}{\sqrt{e_j^T n (X_A^T X_A)^{-1} e_j}},$$

which only depends on (X,y) through Ω_n. The same dependence is true, it turns out, for the selected models from FS, LAR, and the lasso. We defer the proof of the next lemma, as with all proofs in this paper, until the supplement.

Lemma 1. *For each the FS, LAR, and lasso procedures, run for k steps on data (X,y), the selected model $\widehat{M}(X,y)$ only depends on (X,y) through $\Omega_n = (\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^Ty)$, the master statistic.*

In more detail, for any fixed $M \in \mathcal{M}$, the matrix $Q_M(X)$ such that $\widehat{M}(X,y) = M \iff Q_M(X)y \geq 0$ can be written as $Q_M(X) = P_M(\frac{1}{n}X^TX)\frac{1}{\sqrt{n}}X^T$, where P_M depends only on $\frac{1}{n}X^TX$. Hence

$$\widehat{M}(X,y) = M \iff P_M\Big(\frac{1}{n}X^TX\Big)\frac{1}{\sqrt{n}}X^Ty \geq 0.$$

This lemma asserts that the master statistic governs model selection, as performed by FS, LAR, and the lasso. It is also central to TG pivot for these procedures. Denoting $M = \widehat{M}(X,y)$, the statistic $T(X,y;M,v,\mu)$ in (10) only depends on (X,y) through three quantities:

$$\frac{v^Ty}{\|v\|_2}, \quad \frac{Q_M(X)v}{\|v\|_2}, \quad \text{and} \quad Q_M(X)y.$$

The third quantity is always a function of Ω_n, by Lemma 1. When v is chosen so that v^Ty is a normalized coefficient in the regression of y onto a subset of the variables in X, the first two quantities are also functions of Ω_n. Thus, in this case, the TG pivot only depends on (X,y) through the master statistic Ω_n; in fact, for fixed $\frac{1}{n}X^TX$, it is a smooth function of $\frac{1}{\sqrt{n}}X^Ty$.

Lemma 2. *Fix any model $M \in \mathcal{M}$, and suppose that v is chosen so that v^Ty is a normalized coefficient from projecting y onto a subset of the variables in X. Then the TG statistic only depends on (X,y) by means of Ω_n, so that we may write*

$$T(X,y;M,v,\mu) = \psi_M\Big(\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^Ty\Big).$$

The above function ψ_M, with its first argument fixed at (any arbitrary value) $\frac{1}{n}X^TX$, is continuous in its second argument, on the interior of the cone $\{\eta : P_M(\frac{1}{n}X^TX)\eta \geq 0\} \subseteq \mathbb{R}^d$.

Consider now the behavior of the unconditional TG statistic $\mathcal{T}(X,y;V,U)$ over $y \in \mathbb{R}^n$, the entire sample space. Given a catalog $V = \{v_M : M \in \mathcal{M}\}$ such that $v_M^T y$ is a normalized regression coefficient from projecting y onto some subset of the variables, for each $M \in \mathcal{M}$, Lemmas 1 and 2 combine to imply that the unconditional TG statistic is still only a function of (X,y) through the master statistic Ω_n. When $\frac{1}{n}X^TX$ is fixed, the discontinuities of this function over $\frac{1}{\sqrt{n}}X^Ty$ lie on the boundaries of the partition elements, which have measure zero.

Lemma 3. *Suppose that the catalog $V = \{v_M : M \in \mathcal{M}\}$ is chosen so that each $v_M^T y$ gives a normalized coefficient from regressing y onto a subset of the variables in X, for $M \in \mathcal{M}$. Then the unconditional TG statistic only depends on (X,y) by means of Ω_n, so that we may write*

$$\mathcal{T}(X,y;V,U) = \psi\Big(\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^Ty\Big).$$

The above function ψ, with its first argument fixed at (an arbitrary value of) $\frac{1}{n}X^TX$, is continuous in its second argument, on a set $D \subseteq \mathbb{R}^d$ with full Lebesgue measure (i.e., $\mathbb{R}^d \setminus D$ has measure zero).

In addition, if Y is drawn from (1)*, and we construct the master statistic $\Omega_n = (\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^TY)$, then there is a function g such that*

$$V^T\theta = g(\mathbb{E}(\Omega_n)).$$

Therefore, if the errors in (1) *are i.i.d. $N(0,\sigma^2)$, then the pivotal property* (11) *of the TG statistic can be reexpressed as*

$$\mathbb{P}_{g(\mathbb{E}(\Omega_n))=U}\Big(\psi(\Omega_n) \le t\Big) = t,$$

for all $t \in [0,1]$.

Equipped with this last lemma, asymptotic results about the TG statistic, for fixed d, are not far off. Under weak conditions on the generative model in (1), the central limit theorem tells us that $\frac{1}{\sqrt{n}}X^TY$ converges weakly to a normal random variable. With $\frac{1}{n}X^TX$ converging to a deterministic matrix, the continuous mapping theorem will then establish the appropriate asymptotic limit for the statistic $\mathcal{T}(X,y;V,U) = \psi(\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^TY)$. This is made more precise next.

4 Asymptotic theory

Here we treat the dimension d as fixed, and consider the limiting distribution of the TG statistic as $n \to \infty$. (See Section 7 for the case when d grows.) Throughout, the matrix $X \in \mathbb{R}^{n\times d}$ will be treated as nonrandom, and we consider a sequence of predictor matrices satisfying two conditions:

$$\lim_{n\to\infty} \frac{1}{n}X^TX = \Sigma, \text{ and } \lim_{n\to\infty} \max_{i=1,\ldots,n} \frac{\|x_i\|_2}{\sqrt{n}} = 0, \tag{15}$$

for a nonsingular matrix $\Sigma \in \mathbb{R}^{d\times d}$, and for $x_i \in \mathbb{R}^d$, $i = 1,\ldots,n$ denoting the rows of X. These are not strong conditions.

4.1 A nonparametric family of distributions

We specify the class of distributions that we will be working with for Y in (1). Let $\sigma^2 > 0$ be a fixed, known constant. First we define a set of error distributions

$$\mathcal{E} = \Big\{F : \int x\,dF(x) = 0, \int x^2\,dF(x) = \sigma^2\Big\}.$$

The first moment condition in the above definition is needed to make the model identifiable, and the second condition is used for simplicity. Aside from these moment conditions, the class $\mathcal{E}$ contains a small neighborhood (say, as measured in the total variation metric) around essentially every element. Thus, modulo the moment assumptions, $\mathcal{E}$ is strongly nonparametric in the sense of Donoho (1988). Given $\mu \in \mathbb{R}$, let F_μ denote the distribution of $\mu + \delta$, where $\delta \sim F$, and given $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$, let $F_n(\theta) = F_{\theta_1} \times \ldots, \times F_{\theta_n}$. Now we define a class of distributions

$$\mathcal{P}_n(\theta) = \Big\{ F_n(\theta) = F_{\theta_1} \times \ldots \times F_{\theta_n} \,:\, F \in \mathcal{E} \Big\}. \tag{16}$$

In words, assigning a distribution $Y \sim F_n(\theta)$ means that Y in drawn from the model (1), with mean $\theta \in \mathbb{R}^n$, and errors $\epsilon_1, \ldots, \epsilon_n$ i.i.d. from an arbitrary centered distribution F with variance σ^2.

As n grows, we allow the underlying mean θ to change, but we place a restriction on this parameter so that it has an appropriate asymptotic limit. Specifically, we consider a class Θ of sequences of mean parameters such that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} X^T \theta \text{ exists.} \tag{17}$$

To be clear, we emphasize that $\theta \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ will both vary with n, i.e., we can think of θ and the columns of X as triangular arrays, though our notation suppresses this dependence for simplicity. Furthermore, we will sometimes write, in a slight abuse of notation, $\theta \in \Theta$ to represent a sequence of mean parameters that come from the class defined above.

4.2 Uniform convergence results

We begin with a result on the uniform convergence of (the random part of) the master statistic to a normal distribution.

Lemma 4. *Assume that X has asymptotic covariance matrix Σ, and satisfies the normalization condition, as in* (15)*. Let $Y \sim F_n(\theta) \in \mathcal{P}_n(\theta)$, this class as defined in* (16)*, for an arbitrary mean θ. Then $S_n = \frac{1}{\sqrt{n}}(X^T Y - X^T \theta)$ converges in distribution to $Z \sim N(0, \sigma^2 \Sigma)$, uniformly over $\mathcal{P}_n(\theta)$, and uniformly over all θ. That is,*

$$\lim_{n\to\infty} \sup_{\theta \in \mathbb{R}^n} \sup_{F_n(\theta) \in \mathcal{P}_n(\theta)} \sup_{x \in \mathbb{R}^d} \Big| \mathbb{P}(S_n \le x) - \mathbb{P}(Z \le x) \Big| = 0.$$

This leads us to a uniform asymptotic result about the unconditional TG statistic. We remind the reader that k, the number of steps, is to be considered fixed in the next result (as it is throughout the paper).

Theorem 5. *Assume the conditions of Lemma 4, and moreover, let us restrict attention to $\theta \in \Theta$, i.e., consider a sequence of mean parameters sastisfying* (17)*. Suppose FS, LAR, or the lasso is run for k steps on (X, Y). Let $V = \{v_M : M \in \mathcal{M}\}$ be a catalog of vectors such that each $v_M^T \theta$ yields a normalized coefficient in the projection of θ onto a subset of the variables in X, for $M \in \mathcal{M}$.*

Let $U = \{\mu_M : M \in \mathcal{M}\}$ be any fixed catalog of pivot values. Then under $V^T \theta = U$, the TG statistic $\mathcal{T}(X, Y; V, U)$ converges in distribution to $W \sim U(0,1)$, uniformly over $\mathcal{P}_n(\theta)$, and over $\theta \in \Theta$. That is,

$$\lim_{n\to\infty} \sup_{\theta \in \Theta} \sup_{F_n(\theta) \in \mathcal{P}_n(\theta)} \sup_{t \in [0,1]} \Big| \mathbb{P}_{V^T\theta = U}\Big(\mathcal{T}(X, Y; V, U) \le t \Big) - t \Big| = 0.$$

Furthermore, if we define C_α to be the set of μ such that $\alpha/2 \le T(X,Y;\widehat{M}(Y),v_{\widehat{M}(Y)},\mu) \le 1-\alpha$, then C_α is an asymptotically uniformly valid confidence interval for $v_{\widehat{M}(Y)}^T\theta$. That is,

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \sup_{F_n(\theta)\in\mathcal{P}_n(\theta)} \sup_{\alpha\in[0,1]} \left|\mathbb{P}\left(v_{\widehat{M}(Y)}^T\theta \in C_\alpha\right) - (1-\alpha)\right| = 0.$$

Remark 2. We could broaden the scope of allowed catalogs $V = \{v_M : M \in \mathcal{M}\}$ in Theorem 5 to include arbitrary catalogs such that

$$\frac{v_M^T y}{\|v_M\|_2^2} \text{ and } \frac{Q_M(X)v_M}{\|v_M\|_2}$$

are continuous functions of the master statistic $\Omega_n = (\frac{1}{n}X^TX, \frac{1}{\sqrt{n}}X^Ty)$, for each $M \in \mathcal{M}$, since this is all that is needed in order for ψ to be a function of the master statistic (and, continuous in its second argument a.e., whenever its first argument is fixed). Catalogs for which each v_M^Ty is a normalized coefficient from regressing y onto a subset of the variables in X, over $M \in \mathcal{M}$, are the most natural type that meet the requisite conditions, so the results are expressed in terms of these catalogs, for simplicity.

5 Unknown σ^2 and the bootstrap

The results of the previous section assumed that the error variance σ^2 in the model (1) was known. Here we consider two strategies when σ^2 is unknown. The first plugs a (rather naive) estimate of σ^2 into the usual TG statistic. The second is a computationally efficient bootstrap method. Both, as we will show, yield asymptotically conservative p-values. (In practice, the bootstrap often gives shorter confidence intervals than those based on the TG pivot; see Section 6.)

5.1 A simple plug-in approach

Consider the TG statistic $\mathcal{T}(X,Y;V,U)$, with catalogs $V = \{v_M : M \in \mathcal{M}\}$, $U = \{\mu_M : M \in \mathcal{M}\}$. Let us abbreviate the pivot value for the model $\widehat{M}(Y)$ by $\mu = \mu_{\widehat{M}(Y)}$, and also

$$\widehat{v} = v_{\widehat{M}(Y)}, \;\; \widehat{a} = a(Y;\widehat{M}(Y),v_{\widehat{M}(Y)}), \text{ and } \widehat{b} = b(Y;\widehat{M}(Y),v_{\widehat{M}(Y)}),$$

where recall, the latter two functions define the truncation limits for the TG statistic, and are defined precisely in Appendix A.4. In this notation, we can succintly write the TG statistic as

$$\mathcal{T}(X,Y;V,U) = \frac{\Phi\left(\dfrac{\widehat{b}-\mu}{\sigma\|\widehat{v}\|_2}\right) - \Phi\left(\dfrac{\widehat{v}^TY-\mu}{\sigma\|\widehat{v}\|_2}\right)}{\Phi\left(\dfrac{\widehat{b}-\mu}{\sigma\|\widehat{v}\|_2}\right) - \Phi\left(\dfrac{\widehat{a}-\mu}{\sigma\|\widehat{v}\|_2}\right)}. \tag{18}$$

When σ^2 is unknown, we propose a simple plug-in approach that replaces σ with a small multiple of the sample variance, cs_Y, where $s_Y^2 = \frac{1}{n}\sum_{i=1}^n |Y_i - \overline{Y}|^2$ (here $\overline{Y} = \sum_{i=1}^n Y_i/n$ denotes the sample mean), and $c > 1$ is a fixed constant. To be explicit, we consider the modified TG statistic

$$\widetilde{\mathcal{T}}(X,Y;V,U) = \frac{\Phi\left(\dfrac{\widehat{b}-\mu}{cs_Y\|\widehat{v}\|_2}\right) - \Phi\left(\dfrac{\widehat{v}^TY-\mu}{cs_Y\|\widehat{v}\|_2}\right)}{\Phi\left(\dfrac{\widehat{b}-\mu}{cs_Y\|\widehat{v}\|_2}\right) - \Phi\left(\dfrac{\widehat{a}-\mu}{cs_Y\|\widehat{v}\|_2}\right)}. \tag{19}$$

The scaling factor c facilitates our theoretical study of the above plug-in statistic, and practically, we have found that ignoring it (i.e., setting $c = 1$) works perfectly well, though a choice of, say, $c = 1.0001$ seems to have a minor effect anyway.

When the mean θ of Y is nonzero, the sample variance s_Y^2 is generally too large as an estimate of σ^2. As we will show, the modified statistic in (19) thus yields asymptotically conservative p-values. Residual based estimates of σ^2 are not as useful in our setting because they depend more heavily on the linearity of the underlying regression model, and they suffer practically when d is close to n (see also the discussion at the start of Section 6).

5.2 An efficient bootstrap approach

As an alternative to the plug-in method of the last subsection, we investigate a highly efficient bootstrap scheme that does not rely on knowledge of σ^2. Our general framework so far treats X as fixed, and for our bootstrap strategy to respect this assumption, we cannot use, say, the pairs bootstrap, and must perform sampling with respect to Y only. The residual bootstrap is ruled out since we do not assume that the mean θ follows a linear model in X. This leaves us to consider simple bootstrap sampling of the components of Y. This is somewhat nonstandard, as the components of Y in (1) are not i.i.d., but it provides a mechanism for provably conservative asymptotic inference, and it is what makes our approach so computationally efficient.

Given $Y = (Y_1, \dots, Y_n)$ drawn from the model in (1), let $Y^* = (Y_1^*, \dots, Y_n^*)$ denote a bootstrap sample of Y. We will denote by $\mathbb{P}_*$ the conditional distribution of Y^* on Y, and $\mathbb{E}_*$ the associated expectation operator. That is, $\mathbb{P}_*(Y^* \in A)$ is shorthand for $\mathbb{P}(Y^* \in A|Y)$, and similarly for $\mathbb{E}_*$. Using the notation of the last subsection (notation for $\mu, \widehat{v}, \widehat{a}, \widehat{b}$), and assuming without a loss of generality that $\|\widehat{v}\|_2 = 1$, let us motivate our bootstrap proposal by expressing the TG statistic as

$$\mathcal{T}(X,Y;V,U) = \mathbb{P}\Big(Z_{\mu,\sigma^2} \geq \widehat{v}^T Y \,\Big|\, \widehat{a} \leq Z_{\mu,\sigma^2} \leq \widehat{b}, Y\Big),$$

where the probability on the right-hand side is taken with Y (and hence $\widehat{v}, \widehat{a}, \widehat{b}$) treated as fixed, and with Z_{μ,σ^2} denoting a $N(\mu,\sigma^2)$ random variable. The main idea is now to approximate the truncated normal distribution underlying the TG statistic with an appropriate one from bootstrap samples, as in

$$\mathbb{P}\Big(Z_{\mu,\sigma^2} \geq \widehat{v}^T Y \,\Big|\, \widehat{a} \leq Z_{\mu,\sigma^2} \leq \widehat{b}, Y\Big) \approx \mathbb{P}_*\Big(\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) + \mu \geq \widehat{v}^T Y \,\Big|\, \widehat{a} \leq \widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) + \mu \leq \widehat{b}\Big).$$

Recall $\overline{Y} = \sum_{i=1}^n Y_i/n$ is the sample mean of Y, so $\mathbb{E}_*(\widehat{v}^T Y^*) = \widehat{v}^T(\overline{Y}\mathbb{1})$ (with $\mathbb{1} \in \mathbb{R}^n$ denoting the vector of all 1s), and we have shifted $\widehat{v}^T Y^*$ so that the resulting quantity $\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) + \mu$ mimics a normal variable with mean μ. The right-hand side above very nearly defines our bootstrap version of the TG statistic, except that for technical reasons, we must make two small modifications. In particular, we define the bootstrap TG statistic as

$$\mathcal{T}^*(X,Y;V,U) = \frac{\mathbb{P}_*\big(\widehat{v}^T Y \leq c\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) + \mu \leq \widehat{b}\big) + \delta_n}{\mathbb{P}_*\big(\widehat{a} \leq c\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) + \mu \leq \widehat{b}\big) + \delta_n}, \tag{20}$$

where $c > 1$ is a constant as before, and $\delta_n = \gamma n^{-1/4}$ for a small constant $\gamma > 0$. Again, we have found that ignoring the scaling factor c (i.e., setting $c = 1$) works just fine in practice, though a choice like $c = 1.0001$ does not cause major differences anyway. On the contrary, a nonzero choice of the padding factor like $\delta_n = 10^{-4} n^{-1/4}$ does play an important practical role, since the bootstrap probabilities in the numerator and denominator in (20) can sometimes be zero.

Lastly, it is worth emphasizing that practical estimation of the bootstrap probabilities appearing in (20) is quite an easy computational task, because the regression procedure in question, be it FS, LAR, or the lasso, need not be rerun beyond its initial run on the observed Y. After this initial run, we can save the quantities $\widehat{v}, \widehat{a}, \widehat{b}$, and then draw, say, $B = 1000$ bootstrap samples Y^* in order to estimate the probabilities in (20). This is not at all computationally expensive. Moreover, to estimate (20) over multiple trial values of μ (so that we can invert these bootstrap p-values for a bootstrap confidence interval), only a single common set of bootstrap samples is needed, since we can just shift $\widehat{v}^T Y^*$ appropriately for each bootstrap sample Y^*.

5.3 Asymptotic theory for unknown σ^2

Treating the dimension d as fixed, we will assume the previous limiting conditions (15) on the matrix X, and additionally, that

$$\frac{1}{n} \sum_{i=1}^n \|x_i\|_2^3 = O(1). \tag{21}$$

Note that the first condition in (15) already implies $\frac{1}{n} \sum_{i=1}^n \|x_i\|_2^2 \to \mathrm{tr}(\Sigma)$, and the above condition is a little stronger, though it is still not a strong condition by any means. For example, it is satisfied when $\max_{i=1,\ldots,n} \|x_i\|_2 = O(1)$. These conditions on X imply important scaling properties for our usual choices of contrast vectors.

Lemma 6. *Assume that X satisfies* (15)*,* (21)*, and let $V = \{v_M : M \in \mathcal{M}\}$ be a catalog such that each $v_M^T \theta$ gives a normalized regression coefficient from projecting θ onto some subset of the variables in X, for $M \in \mathcal{M}$. Then*

$$\max_{M \in \mathcal{M}} \|v_M\|_3^3 = O\Big(\frac{1}{\sqrt{n}}\Big).$$

We specify assumptions on the distribution of Y in (1) that are similar to (but slightly stronger than) those in Section 4.1. For constants $\sigma^2, \tau, \kappa > 0$, we define a set of error distributions

$$\mathcal{E}' = \Big\{ F : \int x dF(x) = 0, \int x^2 dF(x) = \sigma^2, \int x^3 dF(x) \le \tau, \int x^4 dF(x) \le \kappa \Big\}.$$

We also define a class of distributions

$$\mathcal{P}_n'(\theta) = \Big\{ F_n(\theta) = F_{\theta_1} \times \ldots \times F_{\theta_n} : F \in \mathcal{E}' \Big\}. \tag{22}$$

where as before, F_μ denotes the distribution of $\mu + \delta$, with $\delta \sim F$. For constants $S, R > 0$, we further define a domain $\mathcal{B}$ for the mean parameter θ,

$$\mathcal{B} = \Big\{ \theta \in \mathbb{R}^n : s_\theta^2 = \frac{1}{n} \sum_{i=1}^n |\theta_i - \overline{\theta}|^2 \le S, \;\; r_\theta^3 = \frac{1}{n} \sum_{i=1}^n |\theta_i - \overline{\theta}|^3 \le R \Big\}, \tag{23}$$

where $\overline{\theta} = \sum_{i=1}^n \theta_i / n$. Note that assuming $Y \sim F_n(\theta)$ with $\theta \in \mathcal{B}$ is not a strong assumption; we require the existence of two more moments compared to our distributional assumptions in Section 4.1, and place a very weak condition on the growth of (the components of) θ. These conditions are sufficient to prove the following helpful lemma.

Lemma 7. *Let $Y \sim F_n(\theta) \in \mathcal{P}'_n(\theta)$, where this class is as in* (22)*, and let $\theta \in \mathcal{B}$, where this set is as in* (23)*. Then, defining $c = 1+\rho$ for any fixed $\rho > 0$, we have*

$$\lim_{n\to\infty} \sup_{\theta\in\mathcal{B}} \sup_{F_n(\theta)\in\mathcal{P}'_n(\theta)} \mathbb{P}\big(cs_Y \geq \sigma\big) = 1.$$

In words, the event $\{cs_Y \geq \sigma\}$ has probability approaching 1, uniformly over $\mathcal{P}'_n(\theta)$, and over $\theta \in \mathcal{B}$. Furthermore, denoting the sample third moment of Y as

$$r_Y^3 = \frac{1}{n}\sum_{i=1}^n |Y_i - \overline{Y}|^3,$$

we have that for any $\delta > 0$, there exists $C > 0$ such that

$$\sup_{\theta\in\mathcal{B}} \sup_{F_n(\theta)\in\mathcal{P}'_n(\theta)} \mathbb{P}\bigg(\frac{r_Y^3}{s_Y^3} \geq C\bigg) \leq \delta,$$

for n sufficiently large. In words, $r_Y^3/s_Y^3 = O_{\mathbb{P}}(1)$ uniformly over $\mathcal{P}'_n(\theta)$, and over $\theta \in \mathcal{B}$.

The last two lemmas allow us to tie the distribution function of our bootstrap contrast to that of a normal random variable.

Lemma 8. *Assume that X satisfies* (15)*,* (21)*. Let $V = \{v_M : M \in \mathcal{M}\}$ be a catalog such that each $v_M^T\theta$ gives a normalized regression coefficient from projecting θ onto some subset of the variables in X, for $M \in \mathcal{M}$. Let $Y \sim F_n(\theta) \in \mathcal{P}'_n(\theta)$, as defined in* (22)*, and let $\theta \in \mathcal{B}$, as defined in* (23)*. Then for any $\delta > 0$, there exists $C > 0$ such that*

$$\sup_{\theta\in\mathcal{B}} \sup_{F_n(\theta)\in\mathcal{P}'_n(\theta)} \mathbb{P}\bigg(\sup_{t\in\mathbb{R}} \Big|\mathbb{P}_*\big(\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) \leq t\big) - \mathbb{P}\big(s_Y Z \leq t \,\big|\, Y\big)\Big| \geq \frac{C}{\sqrt{n}}\bigg) \leq \delta,$$

for all n sufficiently large, where we use $Z \sim N(0,1)$ to denote a standard normal random variable. In other words, $\sup_{t\in\mathbb{R}} |\mathbb{P}_*(\widehat{v}^T(Y^* - \overline{Y}\mathbb{1}) \leq t) - \mathbb{P}(s_Y Z \leq t \,|\, Y)| = O_{\mathbb{P}}(1/\sqrt{n})$, uniformly over $\mathcal{P}'_n(\theta)$, and over $\theta \in \mathcal{B}$.

We are now ready to present uniform asymptotic results for both the bootstrap TG statistic, as well as the modified plug-in version of the TG statistic. We remind the reader the number of steps k is treated as fixed in the result below (just as it is throughout this paper).

Theorem 9. *Assume the conditions of Lemma 8, and further, let us restrict attention to $\theta \in \Theta$, i.e., consider a sequence of mean parameters satisfying* (17)*. Suppose FS, LAR, or the lasso is run for k steps on (X,Y). Then under $V^T\theta = 0$, both the plug-in TG statistic $\widetilde{\mathcal{T}}(X,Y;V,0)$ and the bootstrap TG statistic $\mathcal{T}^*(X,Y;V,0)$ are asymptotically larger than a $U(0,1)$ distribution, uniformly so over $\mathcal{P}'_n(\theta)$, and over $\theta \in \Theta \cap \mathcal{B}$. That is,*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta\cap\mathcal{B}} \sup_{F_n(\theta)\in\mathcal{P}'_n(\theta)} \sup_{t\in[0,1]} \Big[\mathbb{P}_{V^T\theta=0}\Big(\widetilde{\mathcal{T}}(X,Y;V,0) \leq t\Big) - t\Big]_+ = 0,$$

and

$$\lim_{n\to\infty} \sup_{\theta\in\Theta\cap\mathcal{B}} \sup_{F_n(\theta)\in\mathcal{P}'_n(\theta)} \sup_{t\in[0,1]} \Big[\mathbb{P}_{V^T\theta=0}\Big(\mathcal{T}^*(X,Y;V,0) \leq t\Big) - t\Big]_+ = 0,$$

where $x_+ = \max\{x,0\}$ denotes the positive part of x. (Also, the notation $\theta \in \Theta \cap \mathcal{B}$ refers to a sequence of mean parameters that satisfies (17)*, and is contained in the set $\mathcal{B}$ in* (23) *for each n.)*

Remark 3. For simplicity, we analyzed the plug-in and bootstrap statistics simultaneously. Consequently, the conditions assumed to prove asymptotic properties of the plug-in approach are stronger than what we would need if we were to study this method on its own, but there are not major differences in these conditions.

Theorem 9 establishes that the plug-in and bootstrap versions of the TG statistic are asymptotically conservative when viewed as p-values under $V^T\theta = 0$, which, recall, we can think of as p-values for the random null hypothesis $H_0 : \widehat{v}^T\theta = 0$. If we look more broadly at the distribution of these test statistics under $V^T\theta = U$, for an arbitrary catalog $U = \{\mu_M : M \in \mathcal{M}\}$ of pivot values, then a technical barrier arises. For each statistic, our proof of its asymptotic conservativeness leveraged the fact that the truncated Gaussian survival function decreases (in a pointwise sense), as its underlying variance parameter decreases. To extend these results to the case of an arbitrary catalog U, we would need the analogous fact to hold when we replace the survival function of the Gaussian variate $cs_Y Z + \mu$ truncated to $[\widehat{a}, \widehat{b}]$, with that of $\sigma Z + \mu$ tuncated to $[\widehat{a}, \widehat{b}]$, on the event $\{cs_Y \geq \sigma\}$. Yet, without the guarantee that $\widehat{a} \geq \mu$ (which of course cannot always be true, for an arbitrary pivot value μ), it is no longer the case that decreasing the variance from $c^2 s_Y^2$ to σ^2 necessarily decreases the survival functions of these two truncated Gaussians; see Appendix A.15. This means confidence intervals given by directly inverting either of the two statistics do not have provably correct asymptotic coverage properties, under the current analysis.

From the arguments in the proof of Theorem 9, we can construct one-sided confidence intervals with conversative asymptotic coverage, by forcing them to include $\widehat{a}$. We do not pursue the details here, as we have found that these one-sided intervals are practically too wide to be of interest.

Importantly, the plug-in and bootstrap TG statistics often display excellent empirical properties, as we will show in the next section. A more refined analysis is needed to establish asymptotic uniformity for the distribution of these statistics when an arbitrary catalog U is used, such that $V^T\theta = U$. Such asymptotic uniformity, for arbitrary U, would lead to asymptotic coverage guarantees for confidence intervals produced by inverting these statistics, and we leave this extension to future work.

6 Examples

We present empirical examples that support the theory developed in the previous sections, and also suggest that there is much room to refine and expand our current set of results. The first two subsections examine a low-dimensional problem setting that is covered by our theory. The last two look at substantial departures from this theoretical framework, the heteroskedastic and high-dimensional settings, respectively. In all examples, the LAR algorithm is used for variable selection; results with the FS and lasso paths would be roughly similar. Also in all examples, though this is not explicitly stated, the computed p-values are a test of whether the target population value is 0.

It may be worth discussing two potentially common reactions to our experimental setups, especially for the low-dimensional problems described in the next subsections. First, our plug-in statistic uses s_Y^2 as an estimate for σ^2; why not use an estimate from the full least squares model of Y on X, since this would be less conservative? While experiments (not shown) confirm that this works in low-dimensional regression problems, such an estimate becomes anti-conservative as the number of variables grows (particularly, irrelevant ones), and is obviously not applicable in high-dimensional problems. Therefore, we stick with the simple estimate s_Y^2, as this is always applicable and always conservative.

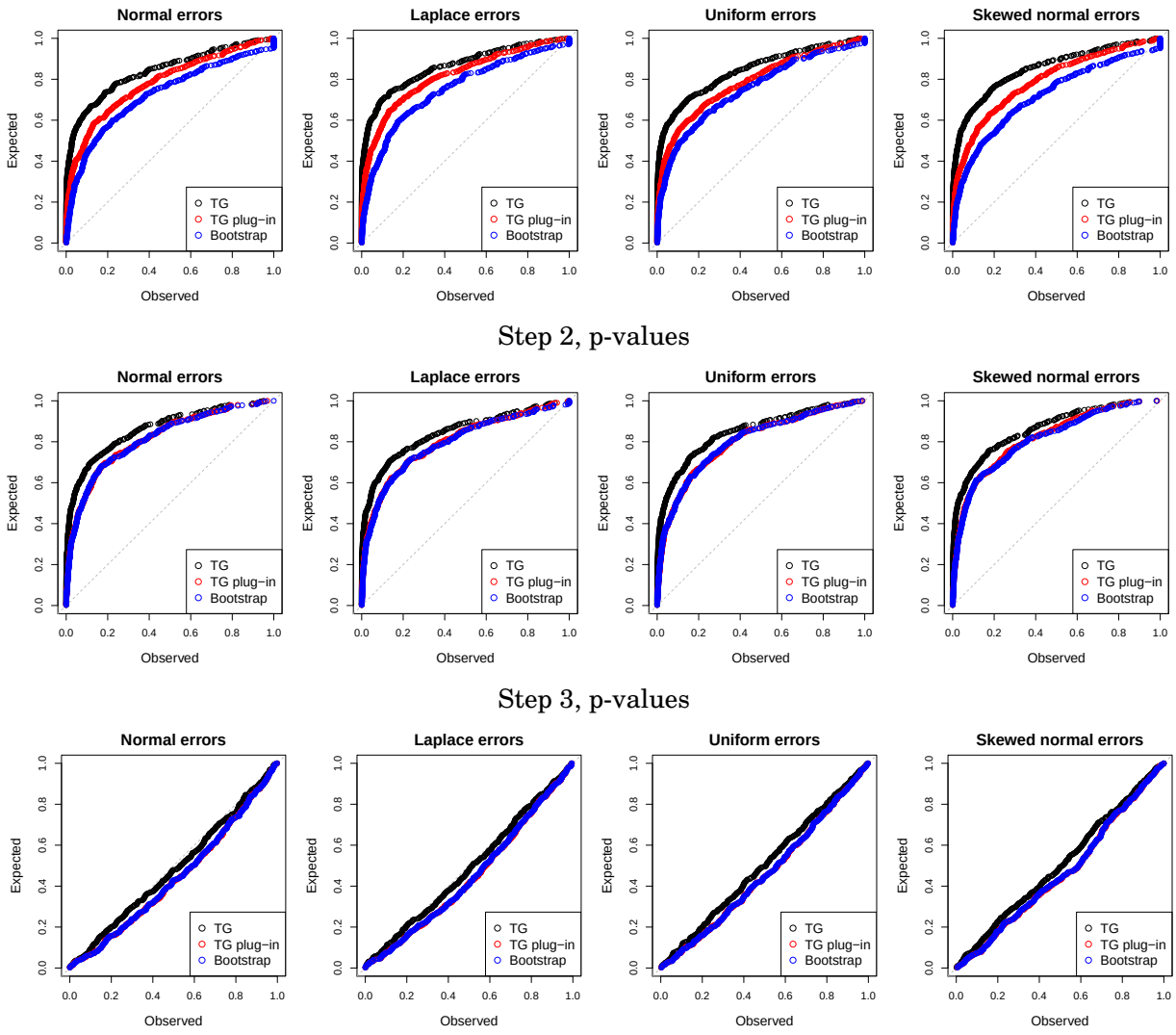
Second, to determine variable significance in a low-dimensional problem, one could of course fit a full regression model and inspect the resulting p-values and confidence intervals. These p-values and intervals could even be Bonferonni-adjusted to account for selection. Of course, this strategy would not be possible for a high-dimensional problem, but if the number of predictors is small enough, then it may work perfectly fine. So when should one use more complex tools for post-selection inference? This is an important question, deserving of study, but it is not the topic of this paper. The examples that follow are intended to portray the robustness of the selective pivotal inference method against nonnormal error distributions; they are not meant to represent the ideal statistical practice in any given scenario.

6.1 P-value examples

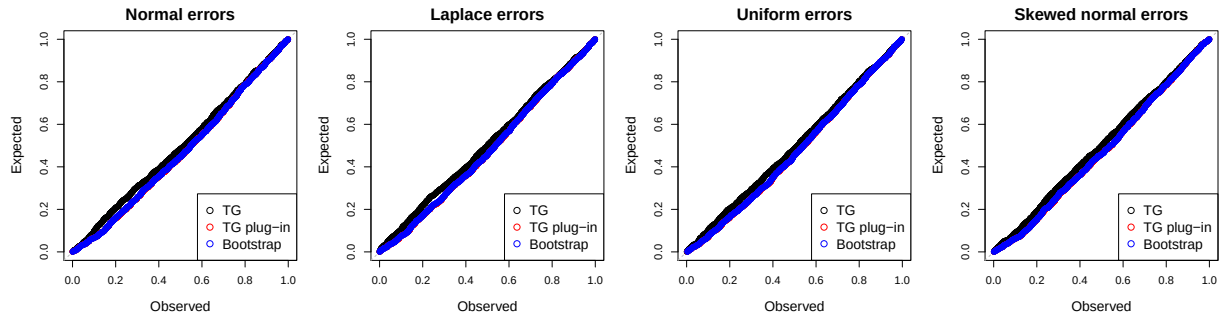
We begin by studying a low-dimensional setting with $n = 50$ and $d = 10$. We defined predictors $X \in \mathbb{R}^{50\times 10}$, by drawing the columns independently according to the following mixture distribution: with equal probability, a column was filled with i.i.d. entries from $N(0,1)$, Bern(0.5), or $SN(0,1,5)$, where $SN(0,1,5)$ denotes the skew normal distribution (O'Hagan & Leonard 1976) with shape parameter equal to 5. We then scaled the columns of X to have unit norm. The underlying mean was defined as $\theta = X\beta_0$, where $\beta_0 \in \mathbb{R}^{10}$ has its first 2 components equal to -4 and 4, and the rest set to 0. Over 500 repetitions, we drew a response $Y \in \mathbb{R}^{50}$ from (1), with i.i.d. errors, and 4 different choices for the error distribution: normal, Laplace, uniform, and skew normal. In each case, we centered the error distribution, and we scaled it to have variance $\sigma^2 = 1$ (for the skew normal distribution, we used a shape parameter 5). Every 10 repetitions, the predictor matrix X was regenerated according to the prescription described above.

Figure 2a displays QQ plots of p-values for testing the significance of the variable entered into the active model, across 3 steps of LAR. (The QQ plots compare the p-values to a standard uniform distribution.) The p-values were computed using the TG statistic with $\sigma^2 = 1$, the plug-in TG statistic with s_Y^2 as its estimate for σ^2, and the bootstrap TG statistic with 50,000 bootstrap samples used to approximate the probabilities in the numerator and denominator of (20), and padding factor $\delta_n = 10^{-4}n^{-1/4}$. (The scaling factor was ignored, i.e., set to $c = 1$, for the plug-in and bootstrap statistics.) In steps 1 and 2, the p-values are restricted to repetitions in which a correct variable selection was made—i.e., variable 1 or 2 was entered into the active LAR model. In step 3, the p-values are from repetitions in which an incorrect variable selection was made—i.e., one of variables 3 through 10 was entered into the active model. Since the underlying signal was fairly strong and the predictors uncorrelated, such selections happened the majority of the time; specifically, the p-values displayed for steps 1, 2, and 3 comprise approximately 95%, 85%, and 87% of the 500 repetitions, respectively. The p-values in steps 1 and 2 show reasonable power, for all 3 statistics (TG, plug-in, and bootstrap types), and all 4 error distributions. Also, the p-values in step 3 are uniform, as desired, again for all statistics and all error distributions. Though the guarantees (for uniform null p-values) are only asymptotic for the Laplace, uniform, and skew normal error distributions, such asymptotic behavior appears to kick in quite early for these distributions, as the sample size here is only $n = 50$. Further, the QQ plots reveal that the p-values for the nonnormal error distributions are not really any farther from uniform than they are in the normal case. This is somewhat remarkable, recalling that the p-values are, by construction, *exactly* uniform under normal errors.

Figure 2b inspects the TG statistic and plug-in and boostrap variants, when the pivot value μ is set to the true population value. That is, we set $\mu = v^T\theta$ in computing the statistics in (18), (19), and (20), in each data instance and each step of LAR. The figure collects the p-values across all 3 steps



(a) *P-values are shown, after each of 3 steps of LAR.*



(b) *Pivotal statistics are shown, aggregated over all 3 steps of LAR.*

Figure 2: *A simulation setup with* $n = 50$ *and* $d = 10$*, and a mean* $\theta = X\beta_0$*, where* β_0 *has 2 nonzero components.*

of LAR, for each of the 4 error distribution types. According to our theory, the distribution of the TG pivotal statistics here should be asymptotically uniform. This is clearly supported by the QQ plots. Interestingly, both plug-in and bootstrap pivotal statistics also appear uniform in the QQ plots, and yet, this is not a case handled by our asymptotic theory: recall, Theorem 9 fixes the pivot value μ to be 0 (as, otherwise, technical difficulties are encountered in its proof). This gives empirical evidence to the idea that a more refined analysis could extend Theorem 9 to the broader setting (of arbitrary pivot values) handled by Theorem 5. Moreover, it suggests that inverting the plug-in and bootstrap TG statistics should yield intervals with proper coverage, which is verified in the next subsection.

Lastly, we repeated all experiments in this subsection with the predictors $X \in \mathbb{R}^{50\times 10}$ generated in such a way to induce a (population) correlation of 0.5 between all pairs of predictor variables. The results are quite similar to those shown in Figure 2, and are hence deferred to Appendix A.16.

6.2 Confidence interval examples

We stay in same setting as the last subsection, so that $n = 50$, $d = 10$, and $\theta = X\beta_0$ for a coefficient vector β_0 with its first 2 components equal to -4 and 4, and the rest equal to 0. We invert the TG, plug-in TG, and bootstrap TG statistics to obtain 90% confidence intervals at each LAR step. See Table 1 for a numerical summary. "Coverage" refers to the average fraction of intervals that contained their respective targets over the 500 repetitions, "power" is the average fraction of intervals that excluded zero, and "width" is the median interval width. These are all recorded in an unconditional sense, i.e., no screening of repetitions was performed based on the variables that were selected across the 3 steps of LAR. From the table, we can see that all 3 methods lead to accurate coverage (around 90%) in all cases. We can further see that the intervals from the bootstrap TG statistic are shorter than those from the plug-in TG statistic in all cases, and considerably shorter than both the plug-in and original TG statistics in steps 2 and 3. The power from the bootstrap TG intervals is generally better than that from the plug-in TG intervals; also, it is on par with the power from the original TG statistic in step 1, but somewhat worse in step 2. We emphasize that the original TG statistic here uses knowledge of the error variance ($\sigma^2 = 1$) but the bootstrap and plug-in variants do not.

		Step 1			Step 2			Step 3		
		Coverage	Power	Width	Coverage	Power	Width	Coverage	Power	Width
N	TG	0.914	0.508	5.622	0.890	0.520	10.309	0.910	0.114	25.155
	Plug-in	0.928	0.378	7.561	0.914	0.404	15.774	0.918	0.100	34.642
	Boot	0.932	0.528	5.477	0.916	0.424	7.856	0.930	0.090	9.141
L	TG	0.904	0.568	5.193	0.926	0.536	11.153	0.912	0.118	26.393
	Plug-in	0.944	0.410	7.271	0.930	0.440	14.859	0.904	0.120	36.206
	Boot	0.944	0.566	5.429	0.944	0.454	7.892	0.924	0.108	9.273
U	TG	0.912	0.538	5.153	0.902	0.504	12.347	0.894	0.128	26.451
	Plug-in	0.928	0.396	7.284	0.910	0.390	17.497	0.886	0.126	39.299
	Boot	0.924	0.540	5.453	0.910	0.422	7.808	0.892	0.118	8.913
S	TG	0.892	0.540	5.346	0.878	0.504	10.876	0.906	0.116	26.592
	Plug-in	0.940	0.402	7.210	0.896	0.380	15.687	0.910	0.106	38.965
	Boot	0.936	0.520	5.477	0.912	0.394	8.060	0.918	0.102	9.057

Table 1: *Summary statistics for 90% confidence intervals constructed in the problem setting of Figure 2. The 4 blocks of rows correspond to the 4 types of noise: normal, Laplace, uniform, and skew normal, respectively. The standard errors are about 0.01, 0.02, and 0.42 for the coverage, power, and width statistics, respectively.*

It is a bit surprising that the bootstrap intervals can be shorter but still have worse power than the original TG intervals. This is easier to understand once the intervals are visualized, as done in Figure 3. The figure shows 100 sample intervals from the first LAR step, under normally distributed errors. Sample intervals from the other error models are in Appendix A.17. We can see that the bootstrap TG intervals are indeed shorter, but compared to the original TG intervals, they are more symmetric around the target population values. The original TG intervals, being more asymmetric, are often shorter on the side (of the target value) facing 0, and this results in better power.

Again, we repeated the experiments here with the predictors $X \in \mathbb{R}^{50\times 10}$ generated to have pairwise correlation 0.5. Comparisons can be drawn between the results in a manner that roughly parallels the discussions following Table 1; however, on an absolute scale, all methods display a decrease in power across the board (as correlated predictors clearly make the problem more difficult). Details are provided in Appendix A.18.

6.3 Heteroskedastic errors

In the same setup as in Sections 6.1 and 6.2, with $n = 50$, $d = 10$, and the predictors X and mean θ generated in the same manner, we consider a heteroskedastic model for Y by drawing ϵ_i', $i = 1,\dots,n$ i.i.d. from the given distribution—normal, Laplace, uniform, or skew normal—and then taking the errors to be $\epsilon_i = \sigma_i \epsilon_i'$, $i = 1,\dots,n$, where $\sigma_i^2 = 10\|x_i\|_2^2$, $i = 1,\dots,n$ (and where $x_i \in \mathbb{R}^d$, $i = 1,\dots,n$ denote the rows of X.) The spread of error variances ended up being fairly substantial, from about 0.3 to 5.5. The original TG statistic was computed with $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ as a surrogate for the common error variance; the plug-in and bootstrap variants were computed as usual. For brevity, we only plot the pivotal statistics, aggregated over 3 steps of LAR, in Figure 4. (This is analogous to what is shown in Figure 2b for the homoskedastic case. P-values at steps 1, 2, and 3, not shown, end up being similar to those in Figure 2a, but the power from all methods is generally lower, due to the heteroskedastic errors.) As we can see, the pivotal statistics in the figure look very close to uniformly distributed, as desired. This is especially encouraging because the current problem setup lies outside of the scope of our asymptotic theory (which assumes a constant error variance), and it suggests that our theory could possibly be extended to accomodate errors with an (unknown) nonconstant variance structure.

6.4 High-dimensional examples

Finally, we consider a high-dimensional regime with $n = 50$ and $d = 1000$ predictors. The matrix $X \in \mathbb{R}^{50\times 1000}$ was generated according to the same recipe as before: each column, with equal probability, was assigned i.i.d. entries from $N(0,1)$, Bern(0.5), or $SN(0,1,5)$, and then scaled to have unit norm. The mean was defined as $\theta = X\beta_0$, where $\beta_0 \in \mathbb{R}^{1000}$ has its first 2 components equal to -4 and 4, and the rest 0. Over 500 repetitions, a response $Y \in \mathbb{R}^{50}$ was generated by adding normal, Laplace, uniform, or skew normal noise to θ, with an error variance of $\sigma^2 = 1$ (and every 10 repetitions, the predictor matrix X was regenerated). Figure 5 plots the pivotal statistics aggregated over the first 3 steps of LAR. (This is as in Figure 2b for the low-dimensional case. P-values from the first 3 LAR steps are omitted for brevity, and are roughly similar to those in Figure 2a, except that they display less power, due to the high-dimensionality.) The pivotal statistics here look quite close to uniform, as desired, and this is again encouraging, especially given that the current high-dimensional case lies outside of the scope of our theory (which assumes that d is fixed). Further work on high-dimensional asymptotic theory should be pursued (see also Tian & Taylor (2015)), though, as we show in the next section, there is no hope for a uniform convergence result in high dimensions that holds as generally

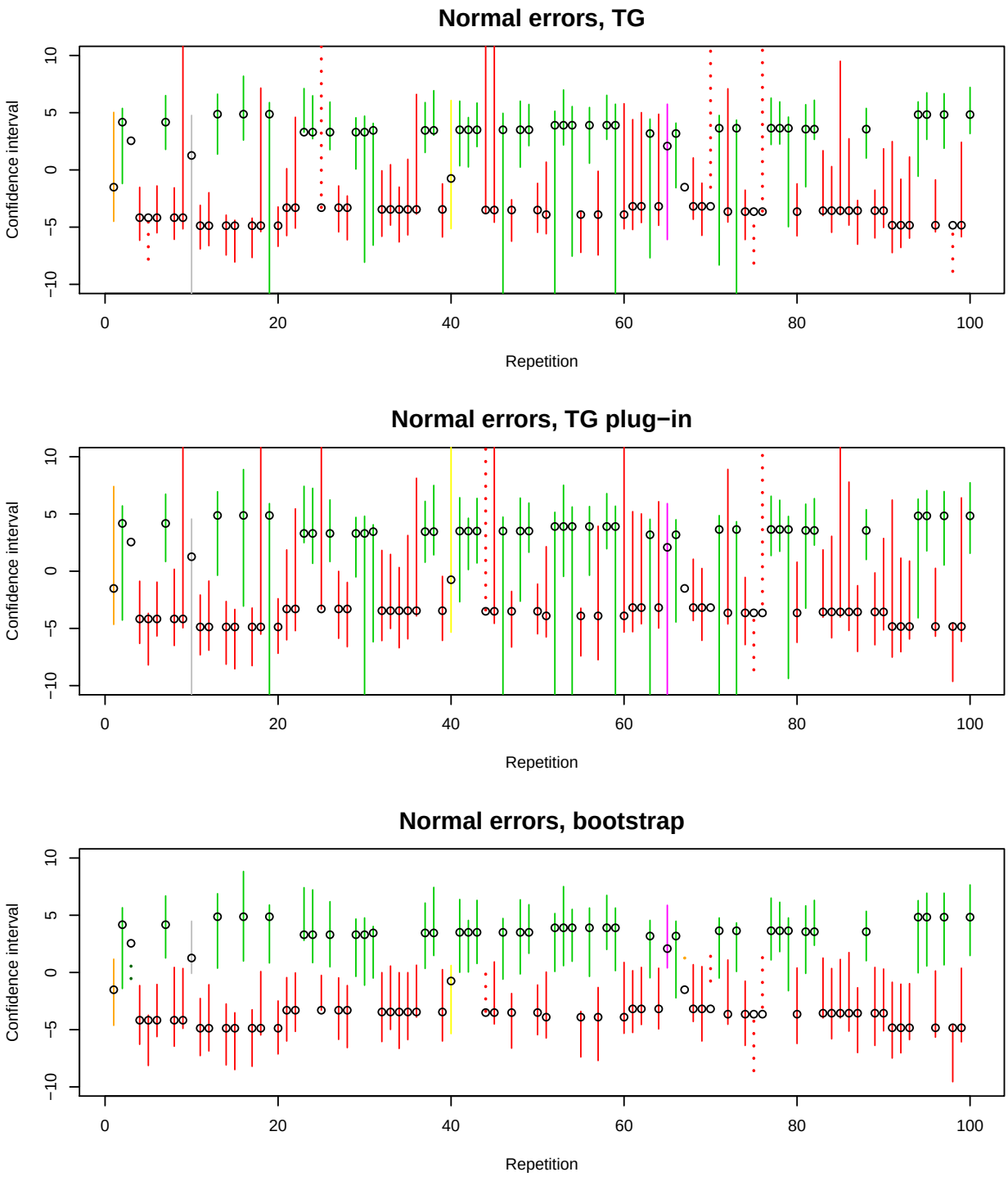


Figure 3: *Confidence intervals from 100 draws of Y from the same model as that in Figure 2. These intervals are constructed from the first step of LAR, under a uniform distribution for noise. The colors are simply a visual aid to mark the selection of different variables at step 1. The open circles denote the true population quantity to be covered (here, the coefficient from projecting θ onto the first selected variable). Intervals that do not contain their targets are drawn as dotted segments.*

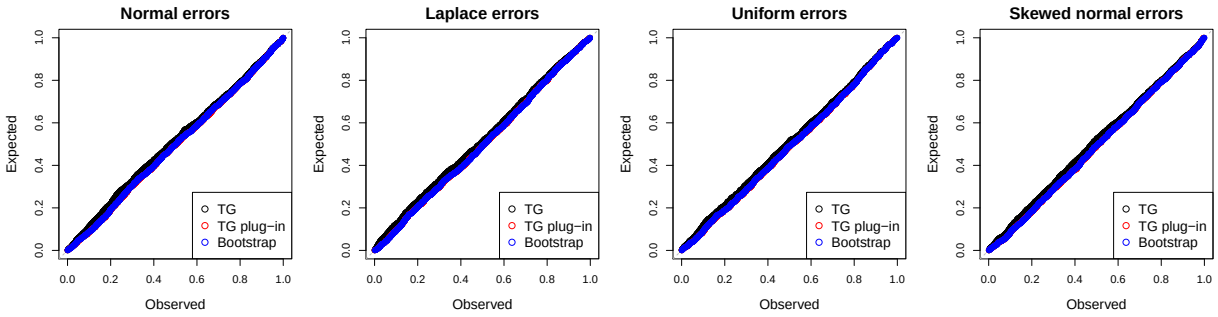


Figure 4: *A simulation setup with $n = 50$ and $d = 10$, but with heteroskedastic errors. Shown are the pivotal statistics aggregated over 3 LAR steps.*

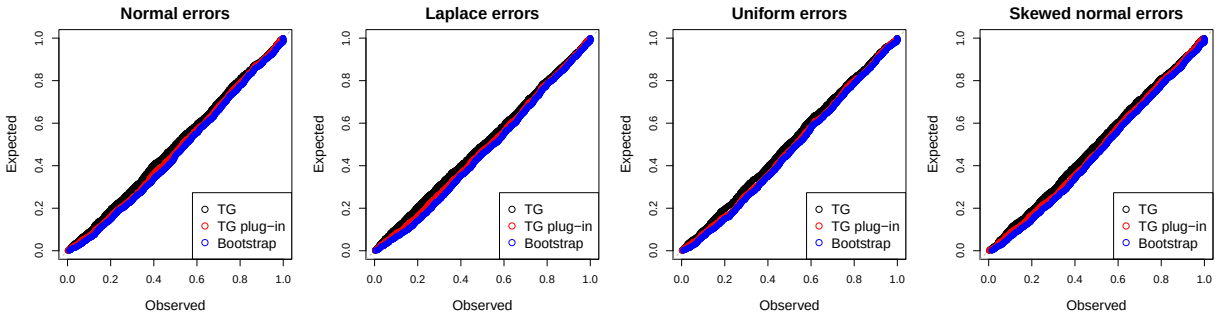


Figure 5: *A simulation setup with $n = 50$ and $d = 1000$. Shown are the pivotal statistics over 3 LAR steps.*

as the one we established in Theorem 5 for low dimensions.

7 A negative result in high dimensions

We prove that the TG statistic fails to converge to a uniform distribution, under the null hypothesis, in a data model that has nonnormal errors and is high-dimensional, but otherwise represents a fairly standard setting: the "many means" setting. We write the observation model as

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1,\dots,m,\ j = 1,\dots,d, \tag{24}$$

where we interpret $i = 1,\dots,m$ as replications, and $j = 1,\dots,d$ as dimensions. In total there are hence $n = md$ observations. Denote

$$\overline{Y}_j = \frac{1}{m}\sum_{i=1}^{m} Y_{ij}, \quad j = 1,\dots,d.$$

We will analyze the TG statistic, when selection is performed based on the largest of $|\overline{Y}_j|$, $j = 1,\dots,d$, and inference is then performed on the corresponding mean parameter. A straightforward change of notation will translate the above into a regression problem, with an orthogonal design $X \in \mathbb{R}^{n\times d}$, but we stick with the many means formulation of the problem for simplicity.

We assume that the errors ϵ_{ij}, $i = 1,\dots,m$, $j = 1,\dots,d$ in (24) are i.i.d. from the following mixture:

$$\pi \cdot N(-B,1) + (1-2\pi)\cdot N(0,1) + \pi \cdot N(B,1). \tag{25}$$

The mixing proportion π and mean shift B will both scale with d. Moreover, they will be chosen so that (for each d) the error variance is

$$\sigma^2 = 1 + 2\pi B^2 = 2.$$

As mentioned, we will consider model selection events of the form

$$\widehat{M}(Y) = (j,s) \iff s\overline{Y}_j \geq \max_{\ell \neq j} |\overline{Y}_\ell|.$$

We note that this is exactly the same selection event as that from the first step of FS, LAR, or lasso paths, when run on the regression version of this problem with orthogonal design X. It is not hard to check that the TG statistic for conditionally testing $\mu_j = 0$, given that $\widehat{M}(Y) = (j,s)$, is

$$T(Y;j,s,0) = \frac{1 - \Phi\left(\frac{\sqrt{m}s\overline{Y}_j}{\sqrt{2}}\right)}{1 - \Phi\left(\frac{\max_{\ell \neq j} \sqrt{m}|\overline{Y}_\ell|}{\sqrt{2}}\right)}. \tag{26}$$

As per the spirit of our paper, we can also view this statistic unconditionally; for this it is helpful to define $W_1 = |\overline{Y}_1|, \ldots, W_d = |\overline{Y}_d|$, and denote by $W_{(1)} \geq \ldots, \geq W_{(d)}$ the order statistics. Then from (26), we can see that the unconditional TG statistic for the selected mean being zero is

$$\mathcal{T}(Y;0) = \frac{1 - \Phi\left(\frac{\sqrt{m}W_{(1)}}{\sqrt{2}}\right)}{1 - \Phi\left(\frac{\sqrt{m}W_{(2)}}{\sqrt{2}}\right)}. \tag{27}$$

The framework underlying the TG statistic tells us that, if $W_{(1)}$ and $W_{(2)}$ are the largest and second largest absolute values of centered normal random variables (each with variance $2/m$), then $\mathcal{T}(Y;0)$ is exactly uniform. But when $W_{(1)}, W_{(2)}$ are large, and come from the order statistics of nonnormal random variates, the statistic $\mathcal{T}(Y;0)$—which in this case is defined by the extreme tail behavior of the normal distribution—could be nonuniform. The next theorem asserts that such nonuniformity does indeed happen asymptotically if we choose the mixture distribution in (25) appropriately.

Theorem 10. *Assume the observation model* (24)*, where the errors are all drawn i.i.d. from* (25)*. Let d and m scale in such a manner that* $(\log d)/m \to \infty$*. Further, let*

$$\pi = \left(\frac{1}{d}\right)^{1/m}, \quad B = \sqrt{\frac{d^{1/m}}{2}},$$

so that the error variance is just $\sigma^2 = 2$, for all d. Then under the global null hypothesis, $\mu = 0$, the TG statistic $\mathcal{T}(Y;0)$ in (27) *does not converge in distribution to $U(0,1)$. In particular, on an event whose limiting probability is at least $1/e$, the statistic $\mathcal{T}(Y;0)$ converges to 0.*

Remark 4. The assumed condition $(\log d)/m \to \infty$ requires the dimension d to diverge to ∞, but not necessarily the number of replications m, though it clearly allows m to diverge at a sufficiently slow rate. On the other hand, if d were fixed and m diverged to ∞, then the result of the theorem would no longer be true, and the limiting distribution of the TG p-value would revert to $U(0,1)$. (To be careful,

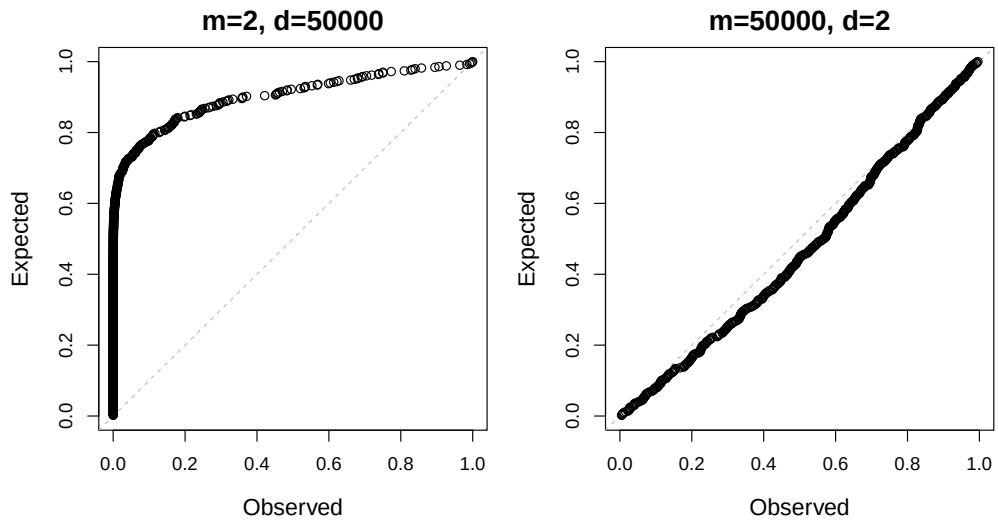


Figure 6: *The left plot shows a QQ plot of TG p-values, computed over 500 repetitions from the many means setup exactly as described in Theorem 10, with* $d = 50,000$ *and* $m = 2$*; we can see that the p-values are clearly nonuniform, and* 34% *of the p-values are 0 (up to computer precision), close to the theoretically predicted proportion of* $1/e$*. The right plot shows p-values from the same model, but having reversed the roles of* d *and* m *(we also had to cap* π *at* $1/2$*); we can see that the p-values are essentially uniform.*

here we would have cap the mixing probability π at 1/2 in order for the mixture to make sense, as the current definition of π diverges with d fixed and m tending to ∞.) In fact, this is ensured by our low-dimensional result in Theorem 5: after rewriting the current many means problem in appropriate regression notation, all of the conditions of Theorem 5 are met by our current setup when d is fixed. This is supported by the simulation in Figure 6.

Remark 5. The precise scaling $(\log d)/m \to \infty$ is chosen since this implies that $\pi = (1/d)^{1/m} \to 0$, i.e., the extreme mixture components $N(-B,1)$ and $N(B,1)$ have probability tending to zero, which seems to be an intuitively reasonable property for the error distribution. This scaling is not important for any other reason, and the proof would still remain correct if $d/m \to \infty$.

Remark 6. In Theorem 3 of Tian & Taylor (2015), the authors show that the TG statistic converges in distribution to a standard uniform random variable, in a high-dimensional problem setting, with some restrictions on the sequences of selection events that are allowed. One might ask what part of our high-dimensional setup here violates their conditions, because both results obviously cannot be true simultaneously. As far as we can tell, the issue lies in the role of δ_n in Assumption 1 of Tian & Taylor (2015). Namely, as we have defined the error distribution in (25), the value of δ_n needed to certify the third condition Assumption 1 of their work is simply too small for the main assumption in their Theorem 3 to hold. Hence Theorem 3 of Tian & Taylor (2015) does not apply to our setup in this section.

8 Discussion

We have studied the selective pivotal inference framework, with a focus on forward stepwise regression (FS), least angle regression (LAR), and the lasso, in regression problems with nonnormal errors.

We have shown that the truncated Gaussian (TG) pivot is asymptotically robust in low-dimensional settings to departures from normality, in that it converges to a $U(0,1)$ distribution (its pivotal distribution under normality), and does so uniformly over a wide class of nonnormal error distributions. When the error variance σ^2 is unknown, we have proposed plug-in and bootstrap versions of the TG statistic, both of which yield provably conservative asymptotic p-values.

Our numerical experiments revealed that the statistics under theoretical investigation generally display excellent finite-sample performance, for highly nonnormal error distributions. These experiments also revealed findings not predicted by our theory: (i) the bootstrap TG statistic often produces shorter confidence intervals than those based on the plug-in TG statistic, and even the TG statistic that relies on the error variance σ^2; and (ii) all three TG statistics show strong empirical properties well-outside of the classic homoskedastic, fixed d regression setting that we presumed theoretically.

However, as we have demonstrated, one should not hope for a convergence result in high dimensions that is as general as the result obtained in low dimensions. In a relatively simple many means problem, we showed the nonconvergence of the TG statistic to $U(0,1)$ as $d \to \infty$, whereas in the same problem but with d fixed, the TG statistic converges to its usual $U(0,1)$ limit.

There is still much left to do in terms of understanding the behavior of selective pivotal inference tools that are constructed to have exact finite-sample guarantees under normality, like the TG statistic of Tibshirani et al. (2016), when applied in high-dimensional regression settings with nonnormal data. When the pivot, the central cog of this framework, is constructed based on assuming (say) a normal distribution for the data (like the TG statistic), this creates robustness issues that are especially worrisome in high dimenisons. Appendix A.20 provides a high-level discussion of some of these issues; a more detailed study will be the subject of future research.

Acknolwedgements

The authors would like to thank Jonathan Taylor for helpful discussions. RJT was supported by NSF grant DMS-1309174; RT was supported by NSF grant DMS-9971405 and NIH grant N01-HV-28183.

References

Bachoc, F., Leeb, H. & Potscher, B. (2014), Valid confidence intervals for post-model-selection predictors. arXiv: 1412.4605.

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013), 'Valid post-selection inference', *Annals of Statistics* **41**(2), 802–837.

Choi, Y., Taylor, J. & Tibshirani, R. (2014), Selecting the number of principal components: estimation of the true rank of a noisy matrix. arXiv: 1410.8260.

Donoho, D. (1988), 'One-sided inference about functionals of a density', *Annals of Statistics* **16**(4), 1390–1420.

Fithian, W., Sun, D. & Taylor, J. (2014), Optimal inference after model selection. arXv: 1410.2597.

Hyun, S., G'Sell, M. & Tibshirani, R. J. (2016), Exact post-selection inference for changepoint detection and other generalized lasso problems. arXv: 1606.03552.

Kasy, M. (2015), Uniformity and the delta method. Unpublished manuscript.

Lee, J., Sun, D., Sun, Y. & Taylor, J. (2016), 'Exact post-selection inference, with application to the lasso', *Annals of Statistics* **44**(3), 907–927.

Lee, J. & Taylor, J. (2014), 'Exact post model selection inference for marginal screening', *Advances in Neural Information Processing Systems* **27**, 136–144.

Leeb, H. & Potscher, B. (2003), 'The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations', *Econometric Theory* **19**(1), 100–142.

Leeb, H. & Potscher, B. (2006), 'Can one estimate the conditional distribution of post-model-selection estimators?', *Annals of Statistics* **34**(5), 2554–2591.

Leeb, H. & Potscher, B. (2008), 'Can one estimate the unconditional distribution of post-model-selection estimators?', *Econometric Theory* **24**(2), 338–376.

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014), 'A significance test for the lasso', *Annals of Statistics* **42**(2), 413–468.

Loftus, J. & Taylor, J. (2014), A significance test for forward stepwise model selection. arXiv: 1405.3920.

O'Hagan, A. & Leonard, T. (1976), 'Bayes estimation subject to uncertainty about parameter constraints', *Biometrika* **63**(1), 201–203.

Reid, S., Taylor, J. & Tibshirani, R. (2014), Post-selection point and interval estimation of signal sizes in Gaussian samples. arXiv: 1405.3340.

Taylor, J., Loftus, J. & Tibshirani, R. J. (2016), 'Inference in adaptive regression via the kac-rice formula', *Annals of Statistics* **44**(2), 743–770.

Tian, X. & Taylor, J. (2015), Asymptotics of selective inference. arXiv: 1501.03588.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* **111**(514), 600–620.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press.

Wasserman, L. (2014), 'Discussion: A significance test for the lasso', *Annals of Statistics* **42**(2), 501–508.