The ℓ-test: leveraging sparsity in the Gaussian linear model for improved inference

Souhardya Sengupta and Lucas Janson

Department of Statistics, Harvard University

Abstract

We develop novel LASSO-based methods for coefficient testing and confidence interval construction in the Gaussian linear model with $n \geq d$. Our methods' finite-sample guarantees are identical to those of their ubiquitous ordinary-least-squares-t-test-based analogues, yet have substantially higher power when the true coefficient vector is sparse. In particular, our coefficient test, which we call the ℓ-test, performs like the *one-sided* t-test (despite not being given any information about the sign) under sparsity, and the corresponding confidence intervals are more than 10% shorter than the standard t-test based intervals. The nature of the ℓ-test directly provides a novel exact adjustment conditional on LASSO selection for post-selection inference, allowing for the construction of post-selection p-values and confidence intervals. None of our methods require resampling or Monte Carlo estimation. We perform a variety of simulations and a real data analysis on an HIV drug resistance data set to demonstrate the benefits of the ℓ-test. We end with a discussion of how the ℓ-test may asymptotically apply to a much more general class of parametric models.

1 Introduction

1.1 Motivation

Assume we have data $(\boldsymbol{y}, \boldsymbol{X})$ from a (homoskedastic Gaussian) linear model:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}), \tag{1.1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is full column-rank (and in particular, assume $n \geq d$) and treated as non-random, and $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\sigma^2 \in \mathbb{R}_{>0}$ are unknown. For a given covariate of interest X_j this paper will consider testing $H_j : \beta_j = 0$ and the related problem of constructing a confidence interval for β_j. It will leverage the LASSO to do so and our method's construction will also make it easy to construct conditionally valid versions, conditioned on LASSO selection.

The go-to solution for this type of single covariate inference is based on the linear regression t-test for H_j, which can be efficiently inverted to obtain a t-test-based confidence interval for β_j. The linear regression t-test dates back over a century (Fisher, 1922) and is

ubiquitous in introductory statistics courses and methods courses in nearly every domain of science and engineering. As a result, it is hard to overstate how universally widely used it is in practice. And it is easy to see why: the t-test is intuitive, easy to compute, and comes with strong theoretical guarantees.

The goal of this paper is to allow an analyst to leverage a belief in *sparsity* (of $\boldsymbol{\beta}$) to conduct more informative inference (when sparsity holds) without sacrificing the statistical guarantees of the t-test (even when sparsity does not hold). Sparsity is a widely held belief throughout applications in science and engineering (indeed, this belief is so ubiquitous that it has a name: the principle of parsimony), and while leveraging sparsity in regression is a heavily studied subject (we review existing approaches in Section 1.3), methods developed to leverage it often rely on sparsity for *both* validity and increased power, while we explicitly seek to rely on it *only* for increased power.

1.2 Summary of our contributions

We develop a hypothesis test for $H_j : \beta_j = 0$, which we call the ℓ-test, that uses the absolute value of the fitted LASSO coefficient as its test statistic. Using novel analysis of the conditional distribution of the LASSO estimator given the sufficient statistic of the linear model, we derive the test statistic's exact null distribution, allowing us to efficiently compute p-values without resampling or Monte Carlo. We argue that the ℓ-test will often achieve nearly the power of the one-sided t-test when $\boldsymbol{\beta}$ is sparse, despite not knowing the sign of β_j and remaining exactly valid under identical assumptions as the two-sided t-test. We show the ℓ-test can be efficiently inverted to produce exact confidence intervals, which due to the power improvement of the ℓ-test are typically more than about 10% shorter than t-test-based confidence intervals when $\boldsymbol{\beta}$ is sparse. For both the ℓ-test and its corresponding confidence interval, we show that a cross-validation procedure can be used to select the penalty parameter λ in the LASSO from the data without impacting our validity guarantees, making our proposed procedures tuning-parameter-free. The nature of the ℓ-test and our formula for its null distribution make it straightforward to derive and compute (novel) post-selection ℓ-test p-values such that the post-selection ℓ-test p-value for H_j is exactly (and non-conservatively) valid *conditional* on the LASSO estimate of β_j being nonzero; this conditional test can also be inverted to construct a confidence interval that is valid conditional on selection. A wide range of simulations and an application to HIV drug resistance demonstrate our methods to be powerful, efficient, and robust. In our discussion, we point out that the ℓ-test can be directly generalized to any normal means problem with known covariance, and hence in particular should be applicable to any parametric model's maximum likelihood estimator in its asymptotic Gaussian limit.

1.3 Related work

The standard choice for testing $H_j : \beta_j = 0$ (and, via inversion, constructing confidence intervals) in the linear model is the (two-sided) t-test, or the one-sided t-test when the sign is known. Given the age and ubiquity of the linear model in statistics, we do not attempt to cover all related literature (though Lei and Bickel (2020), particularly their Appendix B, provides an excellent and detailed review), but just note that many such works focus

on developing methods with some form of guarantees under weaker assumptions than the standard (homoskedastic) Gaussian linear model assumed in this paper (Friedman, 1937; Pitman, 1937, 1938; Kruskal and Wallis, 1952; Tukey, 1958; Hajek, 1962; Adichie, 1967; Jaeckel, 1972; Efron, 1979; Freedman, 1981; Gutenbrunner et al., 1993; Lei and Bickel, 2020). As their goal is robustness, these methods generally do not outperform the t-test in the (homoskedastic) Gaussian linear model, whereas this is exactly the goal of the current paper: maintain the same guarantees as the t-test while improving its power when $\boldsymbol{\beta}$ is sparse. To our knowledge, the only other work that leverages a linear model's sparsity for testing an individual coefficient is the de-biased LASSO (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014), but, when $n \geq d$, it either reduces exactly to the t-test or is only asymptotically valid under strong sparsity assumptions on $\boldsymbol{\beta}$; in contrast, the validity of the ℓ-test holds regardless of the sparsity of $\boldsymbol{\beta}$. There are a number of excellent works that aim to leverage sparsity for more powerful inference in the Gaussian linear model, including (fixed-X) knockoffs (Barber and Candès, 2015) and subsequent follow-ups that improve its performance (e.g., Spector and Janson (2022); Luo et al. (2022); Ren and Barber (2023); Lee and Ren (2024)), as well as methods based on mirror statistics (Xing et al., 2023; Dai et al., 2022), but all of these methods can only be used for variable selection and do not provide single-variable inference.

Two approaches share a similar goal as ours in seeking to improve the t-test's power without making further assumptions. The first approach is Habiger and Peña (2014), which proposes to split the observations into two disjoint (thus independent) parts and uses the first part to estimate the sign of β_j and then leverages the estimated sign in a test using the second part of the data. As we will discuss in Section 2.2, the ℓ-test also leverages an estimated sign of β_j, but since it does not involve data splitting, it does not suffer from the associated loss in sample size and power. The second approach is called Frequentist, assisted by Bayes (FAB), which, given a prior on $\boldsymbol{\beta}$ and σ^2, produces Bayes-optimal power (or confidence interval width) subject to maintaining the same frequentist validity as the t-test (Hoff and Yu, 2019). But this paper only considers (dense) Gaussian priors for $\boldsymbol{\beta}$ and hence, unlike the ℓ-test, does not leverage sparsity.[1]

While the goal of the ℓ-test is most similar to the works mentioned so far, its approach is most closely related to the idea of conditioning on a sufficient statistic under the null hypothesis. This approach is perhaps most prominently used in constructing uniformly most powerful unbiased tests (Lehmann and Scheffé, 1955), including the t-test. But this idea is also fundamental to co-sufficient sampling (Bartlett, 1937; Stephens, 2012), which is used for testing in a wide variety of contexts; see Barber and Janson (2022) for a recent review of such tests. However, the ℓ-test is not sampling-based, and besides, to the best of our knowledge, the only work applying co-sufficient sampling to testing in the linear model is Huang and Janson (2020), but there it is applied to knockoffs (Barber and Candès, 2015; Candès et al., 2018) which is a method for variable selection and cannot perform single-coefficient inference.

Related to this paper's post-selection inference methods, there is a rich literature on obtaining p-values that are valid post-selection (Cox, 1975; Berk et al., 2013; Fithian et al., 2014; Rinaldo et al., 2016; Bachoc et al., 2016; Tian and Taylor, 2018) and in particular in

[1]The discussion section of Hoff and Yu (2019) mentions the possibility of using FAB with spike-and-slab priors to incorporate sparsity but does not pursue it.

the linear model conditioned on LASSO selection (Lee et al., 2016; Tibshirani et al., 2016; Liu et al., 2018; Panigrahi et al., 2021). Closest to our work is Liu et al. (2018), which can be thought of as adjusting the standard t-test (or, more accurately, the z-test, since they assume σ^2 known) for a coefficient to make it valid conditional on that coefficient being selected by the LASSO. The conditional ℓ-test can be thought of as making an analogous adjustment to the ℓ-test rather than the t-test, resulting in similar gains under sparsity over Liu et al. (2018)'s method as the unconditional ℓ-test achieves over the standard t-test; see Sections 4, 5.4, and Appendix F.4 for further comparison and discussion.

1.4 Notation

Throughout this paper, we will use boldfaced symbols to denote matrices and vectors. For a matrix $\boldsymbol{A}$, unless otherwise stated, $\boldsymbol{A}_j$ denotes its j^{th} column, $\boldsymbol{A}_{-j}$ denotes its sub-matrix with the j^{th} column dropped while A_{ij} (note the symbol is no longer boldfaced) denotes the entry in the i^{th} row and the j^{th} column. Similarly for a vector $\boldsymbol{a} \in \mathbb{R}^d$, unless otherwise stated, a_j denotes its j^{th} entry while $\boldsymbol{a}_{-j}$ denotes the sub-vector of $\boldsymbol{a}$ without the j^{th} entry. We will also use $\mathbb{I}(\cdot)$ as the *indicator function* that takes the value 1 if the condition within the parantheses is true, and 0 otherwise.

1.5 Software

Functions in **R** (R Core Team, 2024) for running all methods in this paper are available at `github.com/SSouhardya/l-test`.

2 The ℓ-test

Our proposed test for $H_j : \beta_j = 0$ is primarily based on the simple idea of conditioning on sufficient statistics. The minimal sufficient statistic for the linear model under H_j is $\boldsymbol{S}^{(j)} := (\boldsymbol{X}_{-j}^T \boldsymbol{y}, \boldsymbol{y}^T \boldsymbol{y})$. So, by sufficiency, the conditional distribution of $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$ does not depend on any unknown parameters under H_j, and the same holds true for the conditional distribution of any fixed function of $\boldsymbol{y}$, including the LASSO coefficient estimate. Let

$$\hat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_k |\beta_k| \right)$$

denote the LASSO estimator of the entire coefficient vector with penalty parameter λ. Denote the cumulative distribution function (CDF) for the conditional distribution of $\hat{\beta}_j^\lambda \mid \boldsymbol{S}^{(j)}$ under H_j, which we call the ℓ-distribution, evaluated at some $b \in \mathbb{R}$, by

$$F_\ell^\lambda(b \mid \boldsymbol{S}^{(j)}) := \mathbb{P}_{H_j}(\hat{\beta}_j^\lambda \leq b \mid \boldsymbol{S}^{(j)}).$$

By sufficiency, the ℓ-distribution does not depend on any unknown parameters under H_j, and hence we can in principle compute a valid p-value for H_j that rejects for large $|\hat{\beta}_j^\lambda|$:

$$\bar{F}_{|\ell|}^\lambda(|\hat{\beta}_j^\lambda| \mid \boldsymbol{S}^{(j)}),$$

where $\bar{F}^{\lambda}_{|\ell|}(b \mid \boldsymbol{S}^{(j)}) := \mathbb{P}_{H_j}(|\hat{\beta}^{\lambda}_j| \geq b \mid \boldsymbol{S}^{(j)}) = 1 - \lim_{b' \to b^-} F^{\lambda}_{\ell}(b' \mid \boldsymbol{S}^{(j)}) + F^{\lambda}_{\ell}(-b \mid \boldsymbol{S}^{(j)})$ denotes the tail probability of $|\hat{\beta}^{\lambda}_j|$ and is entirely determined by the ℓ-distribution F^{λ}_{ℓ}. We call the test that uses the above p-value the ℓ-test, though our recommended usage of it involves two modifications, one about breaking ties in the p-value when $\hat{\beta}^{\lambda}_j = 0$ (which currently cause a point mass at 1 in the p-value distribution) and the other about the choice of λ. We defer these two choices to Sections 2.3 and 2.4, respectively, and first provide a characterization of the ℓ-distribution which allows us to efficiently compute the ℓ-test p-value and helps explain why, when, and by how much the ℓ-test increases power over the t-test.

2.1 The ℓ-distribution

As the first step towards characterizing the ℓ-distribution $F^{\lambda}_{\ell}(\cdot \mid \boldsymbol{S}^{(j)})$, we restate (Luo et al., 2022, Proposition E.1) that exactly characterizes the distribution of $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$ under H_j. Let $\boldsymbol{P}_{-j} = \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^T\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^T$ denote the projection matrix onto the column space of $\boldsymbol{X}_{-j}$.

Lemma 2.1 (Luo et al. (2022))**.** *For the Gaussian linear model (1.1), define $\hat{\boldsymbol{y}}_j = \boldsymbol{P}_{-j}\boldsymbol{y}$ and $\hat{\sigma}_j^2 = \|\boldsymbol{y} - \hat{\boldsymbol{y}}_j\|^2$, and let $\boldsymbol{V} \in \mathbb{R}^{n \times (n-d+1)}$ denote an orthonormal matrix orthogonal to the column space of $\boldsymbol{X}_{-j}$ with first column given by $\boldsymbol{V}_1 = \frac{(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}$. Then, there exists a unique vector $\boldsymbol{u} \in \mathbb{R}^{n-d+1}$, such that $\|\boldsymbol{u}\| = 1$ and the following relation holds:*

$$\boldsymbol{y} = \hat{\boldsymbol{y}}_j + \hat{\sigma}_j \boldsymbol{V}\boldsymbol{u}. \tag{2.1}$$

Furthermore, under H_j,

$$\boldsymbol{u} \mid \boldsymbol{S}^{(j)} \sim \text{Unif}\left(\mathbb{S}^{n-d}\right), \tag{2.2}$$

where $\mathbb{S}^{n-d}$ denotes the unit sphere of dimension $n-d$.

For completeness, a proof of the lemma is provided in Section C.1 of the Appendix. Next our main theoretical result, Theorem 2.1, establishes a mapping between $\hat{\beta}^{\lambda}_j$ and just the first element of $\boldsymbol{u}$, u_1 (there is nothing special about index 1 here except that we defined $\boldsymbol{V}$ to have only its first column non-orthogonal to $\boldsymbol{X}_j$), which will give an immediate characterization of F^{λ}_{ℓ} via the known distribution of u_1 from Equation (2.2).

Theorem 2.1 (Characterization of the ℓ-distribution)**.** *Consider the unique decomposition (2.1) from Lemma 2.1 and for any $b \in \mathbb{R}$ and $\epsilon \in \{-1, 1\}$, define the functions*

$$\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b) := \underset{\boldsymbol{\beta}_{-j} \in \mathbb{R}^{d-1}}{\arg\min} \left(\frac{1}{2n}\|\boldsymbol{y} - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\|^2 + \lambda\|\boldsymbol{\beta}_{-j}\|_1 \right), \tag{2.3}$$

$$\Lambda_j(b, \epsilon) = \frac{-\boldsymbol{X}_j^T\left(\hat{\boldsymbol{y}}_j - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}(b)\right) + n\lambda\epsilon}{\hat{\sigma}_j\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|}.$$

Then the function $f_{\boldsymbol{S}^{(j)}} : \mathbb{R} \mapsto \mathbb{R}$, defined via its inverse as

$$f^{-1}_{\boldsymbol{S}^{(j)}}(b) = \begin{cases} \Lambda_j(b, \text{sign}(b)), & \text{if } b \neq 0 \\ [\Lambda_j(0, -1), \Lambda_j(0, 1)], & \text{if } b = 0 \end{cases}, \tag{2.4}$$

satisfies $\hat{\beta}^{\lambda}_j = f_{\boldsymbol{S}^{(j)}}(u_1)$ and is continuous and non-decreasing in its domain, and strictly increasing on the set $\{u : f_{\boldsymbol{S}^{(j)}}(u) \neq 0\}$.

Theorem 2.1 exactly characterizes the ℓ-distribution: for $b \neq 0$, we have

$$\{\hat{\beta}_j^\lambda \leq b\} = \left\{ f_{\boldsymbol{S}^{(j)}}^{-1}\left(\hat{\beta}_j^\lambda\right) \leq f_{\boldsymbol{S}^{(j)}}^{-1}(b) \right\} = \{u_1 \leq \Lambda_j(b, \text{sign}(b))\}\,. \tag{2.5}$$

Thus, in particular, if we let F_u denote the CDF of u_1 under H_j, then $F_\ell^\lambda(b) = \mathbb{P}_{H_j}(\hat{\beta}_j^\lambda \leq b \mid \boldsymbol{S}^{(j)}) = F_u(\Lambda_j(b, \text{sign}(b)))$ because $\Lambda_j(b, \text{sign}(b))$ is a function of $\boldsymbol{S}^{(j)}$ for any fixed b. Similarly, for $b = 0$, it follows from Theorem 2.1 that $F_\ell^\lambda(0) = F_u\left(\Lambda_j(0,1)\right)$. Note that F_u is easily evaluated via a one-to-one mapping to a t-distribution, namely, $\frac{\sqrt{n-d}u_1}{\sqrt{1-u_1^2}} \sim t_{n-d}$ (see Appendix C.5 for a proof), which, along with the relations above, can be used to explicitly calculate quantiles of the ℓ-distribution. The proof of Theorem 2.1 in Appendix C.3 hinges on two main ideas—first, we use blockwise coordinate descent to characterize the event $\{\hat{\beta}_j^\lambda = b\}$ in terms of u_1, and second, we characterize $f_{\boldsymbol{S}^{(j)}}$ by obtaining an exact expression for $\frac{\partial}{\partial u_1} f_{\boldsymbol{S}^{(j)}}$, which turns out to be non-negative throughout, thereby showing $f_{\boldsymbol{S}^{(j)}}$ is non-decreasing. We will see next that Theorem 2.1 also provides critical insights into the power of the ℓ-test.

2.2 The power of the ℓ-test

Our simulations in Section 5.1 show that not only does the ℓ-test consistently beat the usual two-sided t-test when $\boldsymbol{\beta}$ is sparse, it achieves power close to the *one-sided* t-test (in the correct direction), being nearly identical in some cases, without any knowledge about the true sign of β_j.

To explain this behavior, we first characterize the relationship between the t-test statistic and u_1 in Lemma 2.2. It turns out that u_1 is a scalar multiple of $\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{y}$ (we argue this in Equations (C.1) through (C.4) of the Appendix and is a major component in the proof of the lemma), and hence, u_1 is a measure of the association between $\boldsymbol{X}_j$ and the component of $\boldsymbol{y}$ that cannot be explained by the rest of the columns.

Lemma 2.2. *Let T_j denote the t-test statistic for testing $H_j : \beta_j = 0$. Then, there exists a continuous, strictly increasing, anti-symmetric function $g_{\boldsymbol{S}^{(j)}}$ that is a functional of the sufficient statistic $\boldsymbol{S}^{(j)}$, such that $T_j = g_{\boldsymbol{S}^{(j)}}(u_1)$.*

We prove this result in Section C.2 of the appendix. Lemma 2.2 and Theorem 2.1 together show that a (one-sided) conditional-on-$\boldsymbol{S}^{(j)}$ test based on any of the test-statistics—u_1, T_j and $\hat{\beta}_j^\lambda$, yield exactly the same p-values as long as the observed LASSO estimate is non-zero, as in this case all the three test statistics are strictly increasing in each other. We also know that T_j is independent of $\boldsymbol{S}^{(j)}$, which follows from standard theory on ancillary statistics, however we supply a separate proof for this in Section C.4 of the Appendix. This implies that a one-sided conditional test based on $T_j \mid \boldsymbol{S}^{(j)}$ yields exactly the one-sided t-test p-value, which by the above argument is exactly equal to the one-sided p-value of the conditional test based on $\hat{\beta}_j^\lambda \mid \boldsymbol{S}^{(j)}$ when the observed LASSO estimate is non-zero.

Even though the above paragraph establishes that conditional one-sided testing based on $\hat{\beta}_j^\lambda$ can do only as well as the corresponding one-sided t-test, we will now argue that the former test can gain considerable power over the t-test in the *two*-sided testing regime. As a first step, note that we can use Theorem 2.1 to characterize the set $\mathcal{R} := \{u_1 : \hat{\beta}_j^\lambda \neq 0\}$ as a disjoint union of two intervals: $\mathcal{R} = (-\infty, v_-) \cup (v_+, \infty)$, where, $v_\pm = \Lambda_j(0, \pm 1)$.

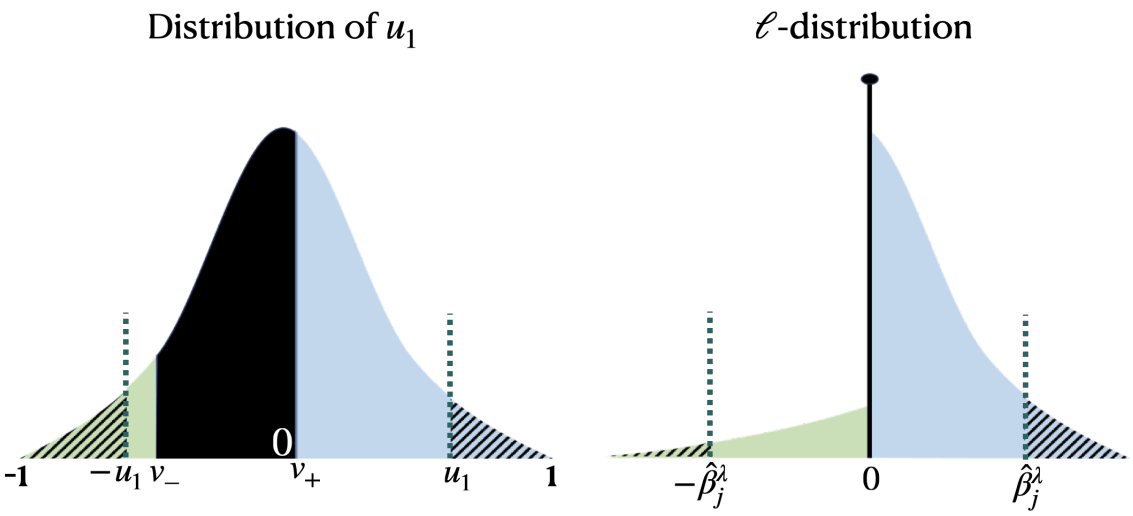

Figure 1: Connection between the distribution of u_1 (left) and the ℓ-distribution (right). Regions of matching color between the two plots represent a one-to-one correspondence between the random variable values within them, as proved in Theorem 2.1.

We will argue that when $\beta_j \neq 0$, the test based on $|\hat{\beta}_j^\lambda|$ (i.e., the ℓ-test) leverages the asymmetry of the interval $[v_-, v_+]$ about 0 to gain power. Observe that from Theorem 2.1, $\hat{\beta}_j^\lambda$ is negative when $u_1 \leq v_-$, while it is positive when $u_1 \geq v_+$. Without loss of generality, we will assume that $\beta_j > 0$ for the proceeding discussion, and assume the center of the interval $[v_-, v_+]$ is negative (we will justify this latter assumption in a little bit). Consider Figure 1 for a visual representation of this, where under H_j, the left and the right figures show the conditional distributions of u_1 (i.e., F_u) and $\hat{\beta}_j^\lambda$ (i.e., the ℓ-distribution $F_\ell^\lambda(\cdot \mid \boldsymbol{S}^{(j)})$), respectively. The correspondence between the two distributions is shown by matching colors—for example, as is evident from Theorem 2.1, the mass that the distribution of u_1 puts to the left of v_- is exactly the mass the ℓ-distribution puts on the negative half, and hence both these regions are colored green. As can be seen from the figure, the asymmetry in $[v_-, v_+]$ (and the symmetry of F_u) directly implies asymmetry in the ℓ-distribution.

Since $\beta_j > 0$, we expect that $\hat{\beta}_j^\lambda > 0$ as well, as reflected in the righthand plot, with corresponding positive u_1 also marked in the lefthand plot. Due to the symmetry of u_1's distribution, the p-value of the two-sided test using $|u_1|$ is just twice the mass to the right of u_1, and by Lemma 2.2, this is also the p-value of the (two-sided) t-test. To understand the ℓ-test p-value for comparison, note that by Theorem 2.1, the mass to the right of u_1 in the left plot is exactly the mass to the right of $\hat{\beta}_j^\lambda$ in the right plot, but, critically, the mass to the *left* of $-\hat{\beta}_j^\lambda$ in the right plot is *far less* than the mass to the left of $-u_1$ in the left plot. Thus, the ℓ-test's p-value, which is exactly the mass of the shaded regions in the right plot, is dominated by the right-most shaded region, whose mass is exactly the value of the *one*-sided t-test.

Next, we argue why we expect the interval $[v_-, v_+]$ to lean opposite to the true sign of β_j,

that is, towards the negative side in this case. Note that the interval $[v_-, v_+]$ has mid-point

$$\hat{m}_j := \frac{v_- + v_+}{2} = \frac{\Lambda_j(0,-1) + \Lambda_j(0,+1)}{2} = \frac{-\boldsymbol{X}_j^T \left(\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}^\lambda_{-j}(0)\right)}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|\hat{\sigma}_j}. \tag{2.6}$$

We can think of $\hat{m}_j$ as an estimator of m_j, where m_j has the exact same expression with $\hat{\boldsymbol{\beta}}^\lambda_{-j}(0)$ replaced by its estimand, $\boldsymbol{\beta}_{-j}$. Defining $\tau_j^2 := \boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \geq 0$, it can be seen that the numerator of m_j satisfies

$$-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}) \sim \mathcal{N}\left(-\beta_j\tau_j^2, \sigma^2\tau_j^2\right).$$

Thus under the alternative, m_j's distribution is shifted towards the opposite sign of β_j as long as $\boldsymbol{X}_j$ is not exactly orthogonal to $\boldsymbol{X}_{-j}$. And when $\boldsymbol{\beta}$ is sparse, we expect the lasso estimator $\hat{\boldsymbol{\beta}}^\lambda_{-j}(0)$ of $\boldsymbol{\beta}_{-j}$ to be a good one, and hence that $\hat{m}_j$'s distribution will also be shifted towards the opposite sign of β_j. In particular, when $\beta_j > 0$, this means we expect $[v_-, v_+]$ to be shifted in the negative direction, as we assumed it would be earlier in this subsection.

In sum, when $\boldsymbol{\beta}$ is sparse, the LASSO leverages information in $\boldsymbol{S}^{(j)}$ (via $\hat{\boldsymbol{\beta}}^\lambda_{-j}(0)$) to guess the sign of β_j, and the point mass in the ℓ-distribution uses that guess (via $[v_-, v_+]$) to reduce the "wrong" tail of the ℓ-test, resulting in a test with power approximating the one-sided t-test; see Appendix B for further discussion. Note that although the intuition for the power gain of the ℓ-test over the t-test relied on sparsity, we emphasize that the *validity* guarantees of the ℓ-test remain identical to those of the t-test, and in particular do not require sparsity.

2.3 Breaking ties when $\hat{\beta}^\lambda_j = 0$

As alluded to earlier in this section, the ℓ-test p-value $\bar{F}^\lambda_{|\ell|}(|\hat{\beta}^\lambda_j| \mid \boldsymbol{S}^{(j)})$ is not $\text{Unif}(0,1)$ under H_j. Instead, its conditional distribution given $\boldsymbol{S}^{(j)}$ is a mixture of $\text{Unif}(0, \mathbb{P}_{H_j}(\hat{\beta}^\lambda_j \neq 0 \mid \boldsymbol{S}^{(j)}))$ and a point mass at 1 of weight $\mathbb{P}_{H_j}(\hat{\beta}^\lambda_j = 0 \mid \boldsymbol{S}^{(j)})$ because $|\hat{\beta}^\lambda_j| = 0$ is the "least significant" value of the test statistic and occurs with positive probability. It is preferable not to have such a point mass, since it makes the ℓ-test somewhat conservative and because both users and many procedures which take p-values as inputs generally assume uniform null p-values. To remedy this, we need a way to break ties among data values that give $\hat{\beta}^\lambda_j = 0$, since the test statistic $|\hat{\beta}^\lambda_j|$ does not distinguish between them. The strong connection between $\hat{\beta}^\lambda_j$ and u_1 established in Theorem 2.1 and visualized in Figure 1 suggests that u_1, whose distribution is continuous on $[v_-, v_+]$ given the event $\{\hat{\beta}^\lambda_j = 0\}$, provides a way forward. In particular, since $\hat{\beta}^\lambda_j$ approaches zero as u_1 approaches v_- from the left or v_+ from the right, it is natural (and continuous in the data) to set v_- and v_+ as tied for the "most significant" values on the interval $[v_-, v_+]$, and then have the significance decrease as u_1 moves inward from those endpoints. This corresponds to breaking ties according to $|u_1 - \hat{m}_j|$ when $\hat{\beta}^\lambda_j = 0$, and can equivalently be thought of as using $|\hat{\beta}^\lambda_j| + \min\{|u_1 - \hat{m}_j| - (v_+ - \hat{m}_j), 0\}$ as the test statistic. Recalling that F_u denotes u_1's CDF and defining $\bar{F}_{|u|}(u' \mid \boldsymbol{S}^{(j)}) := \mathbb{P}_{H_j}(|u_1 - \hat{m}_j| \geq u' \mid \boldsymbol{S}^{(j)}) = 1 - F_u(\hat{m}_j + u') + F_u(\hat{m}_j - u')$ as the tail probability of $|u_1 - \hat{m}_j|$, we can express

this p-value as

$$p_j^\lambda := \begin{cases} \bar{F}^\lambda_{|\ell|}(|\hat{\beta}^\lambda_j| \mid \boldsymbol{S}^{(j)}), & \text{if } \hat{\beta}^\lambda_j \neq 0 \\ \bar{F}_{|u|}\left(|u_1 - \hat{m}_j| \mid \boldsymbol{S}^{(j)}\right), & \text{if } \hat{\beta}^\lambda_j = 0 \end{cases}, \tag{2.7}$$

which is exactly $\mathrm{Unif}(0,1)$ under H_j and never larger than $\bar{F}^\lambda_{|\ell|}(|\hat{\beta}^\lambda_j| \mid \boldsymbol{S}^{(j)})$, the p-value proposed in Section 2.1.

2.4 The choice of λ

Thus far, we have treated λ as a fixed tuning parameter, but in practice it is preferable to have an automated, data-dependent way to choose it. Standard practice for the LASSO is to choose λ via cross-validation (on the full data $(\boldsymbol{y}, \boldsymbol{X})$), but while this choice invalidates the theoretical guarantees of the ℓ-test, just a slight modification of it is sufficient to retain those guarantees. Let $\tilde{\boldsymbol{u}} \sim \mathrm{Unif}(\mathbb{S}^{n-d})$ be drawn independently of $\boldsymbol{u}$, and plug it into Equation (2.1) and call the resulting lefthand side $\tilde{\boldsymbol{y}}$, so that $\tilde{\boldsymbol{y}}$ is conditionally independent of $\boldsymbol{y}$ given $\boldsymbol{S}^{(j)}$. Then it is easy to see that cross-validation on $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$ produces a λ, which we denote by $\hat{\lambda}_j$, that is exactly valid to use in the ℓ-test, since conditioning on $\hat{\lambda}_j$ does not change the ℓ-distribution.[2] And empirically, $\hat{\lambda}_j$ (our implementation uses 10-fold cross-validation) seems to be just as powerful as the (technically invalid) choice via cross-validation on $(\boldsymbol{y}, \boldsymbol{X})$; see Appendix E.2. Although $\hat{\lambda}_j$ is randomized through $\tilde{\boldsymbol{u}}$ and the random 10-fold partition of the data used by cross-validation, we find this exogenous randomness barely makes a difference: in our simulations it empirically accounts for at most about 0.8% of the variability in the ℓ-test p-value $p_j^{\hat{\lambda}_j}$; see Appendix E.2. Furthermore, if desired, this randomness could be arbitrarily reduced by computing many conditionally independent $\hat{\lambda}_j$'s and using their mean or median for the ℓ-test.

2.5 Putting it all together: the ℓ-test

We can now state our recommended implementation of the ℓ-test: the p-value $p_j^{\hat{\lambda}_j}$ which combines the main ℓ-test idea with the tie-breaking of Sections 2.3 and the λ choice of Section 2.4. Computing $p_j^{\hat{\lambda}_j}$ requires, aside from a cross-validated LASSO to compute $\hat{\lambda}_j$, one (non-cross-validated) LASSO to compute $\hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}$ and, if $\hat{\beta}_j^{\hat{\lambda}_j} \neq 0$, then a second (non-cross-validated) LASSO to compute $\hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}(-\hat{\beta}_j^{\hat{\lambda}_j})$. When $\hat{\beta}_j^{\hat{\lambda}_j} \neq 0$, these two LASSO's allow us to compute the two tails for the ℓ-test p-value via Theorem 2.1, since $\hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}(\hat{\beta}_j^{\hat{\lambda}_j}) = \hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}$. And when $\hat{\beta}_j^{\hat{\lambda}_j} = 0$, by Equation (2.7), the only LASSO quantity needed is $\hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}(0)$ to compute $\hat{m}_j$, but since $\hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}(0) = \hat{\boldsymbol{\beta}}^{\hat{\lambda}_j}_{-j}$ in this case, no additional LASSO run is needed beyond the first. Thus computation for the ℓ-test requires just a very small constant number of LASSO runs.

[2] Cross-validation on $(\tilde{\boldsymbol{y}}, \boldsymbol{X})$ would also be valid and natural, but we prefer $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$ for computational reasons; see Appendices E.1 and E.2 for details on this and other ways to choose λ that we considered, all of which we found to be empirically dominated by $\hat{\lambda}_j$.

It is an immediate consequence of our construction that $p_j^{\hat{\lambda}_j}$ is valid and non-conservative under no further assumptions than the (homoskedastic Gaussian) linear model, which we formally state here as a corollary.

Corollary 2.1 (Validity of the ℓ-test)**.** *For model* (1.1)*, for all* $\alpha \in [0,1]$, $\mathbb{P}_{H_j}(p_j^{\hat{\lambda}_j} \leq \alpha) = \alpha$.

3 ℓ-test confidence intervals

If we can use the ℓ-test to test $H_j(\gamma) : \beta_j = \gamma$ for any $\gamma \in \mathbb{R}$, then this family of tests can be inverted to obtain a valid confidence region for β_j. But extending the ℓ-test to $H_j(\gamma)$ is straightforward, since $\boldsymbol{y}$ satisfying $H_j(\gamma)$ is equivalent to $\boldsymbol{y} - \gamma \boldsymbol{X}_j$ satisfying $H_j(0) = H_j$, so we can simply apply the regular ℓ-test (exactly as detailed in Section 2) to the data $(\boldsymbol{y} - \gamma \boldsymbol{X}_j, \boldsymbol{X})$. Defining $p_j^{\hat{\lambda}_j(\gamma)}(\gamma)$ as the ℓ-test p-value for $H_j(\gamma)$, the $100(1-\alpha)\%$ ℓ-test confidence region is given by $\{\gamma \in \mathbb{R} : p_j^{\hat{\lambda}_j(\gamma)}(\gamma) > \alpha\}$, and for interpretability purposes, we take its convex hull (i.e., the smallest interval containing it) as our ℓ-test confidence interval:

$$\hat{C}_j := \mathrm{conv}(\{\gamma \in \mathbb{R} : p_j^{\hat{\lambda}_j(\gamma)}(\gamma) > \alpha\}).$$

The validity of $\hat{C}_j$ follows directly from that of the ℓ-test p-values $p_j^{\hat{\lambda}_j(\gamma)}(\gamma)$ and the fact that taking the convex hull can only make a set bigger and hence only increase its coverage. We recommend using the same $\tilde{\boldsymbol{u}}$ and cross-validation partition for all γ when constructing $\hat{C}_j$, so that the slight randomness in the ℓ-test p-values is consistent across γ.

Computationally, one may be concerned that computing $\hat{C}_j$ requires many LASSO runs for a fine grid of γ values. It is known (Efron et al., 2004; Rosset and Zhu, 2007) that the LASSO solution is piecewise linear in λ and that these paths can be generated by efficient algorithms, but in Appendix D, we show that the LASSO solution $(\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(\gamma))$ is also piecewise linear in γ (for fixed λ) and we provide an algorithm to efficiently generate these paths as well. Combining these two path-generating algorithms (in λ and γ) provides an efficient way to share computation to efficiently compute all the LASSO solutions needed for $\hat{C}_j$.

4 ℓ-test inference conditional on LASSO selection

The ℓ-test p-value's distributional form (2.7) makes it extremely straightforward (both conceptually and computationally) to adjust it to be conditionally valid given $\hat{\beta}_j^{\lambda} \neq 0$: simply divide the ℓ-test p-value p_j^{λ} by $r_j^{\lambda} := \mathbb{P}_{H_j}(\hat{\beta}_j^{\lambda} \neq 0 \mid \boldsymbol{S}^{(j)}) = \mathbb{P}_{H_j}(u_1 \notin [v_-, v_+] \mid \boldsymbol{S}^{(j)}) = 1 - F_u(v_+) + F_u(v_-)$ to get $\underline{p}_j^{\lambda} := p_j^{\lambda}/r_j^{\lambda}$. The fact that, under H_j $\underline{p}_j^{\lambda}$ has a $\mathrm{Unif}(0,1)$ distribution conditional on $\boldsymbol{S}^{(j)}$ and $\hat{\beta}_j^{\lambda} \neq 0$ follows because r_j^{λ} is exactly the supremum value p_j^{λ} can take as long as $\hat{\beta}_j^{\lambda} \neq 0$, and the density of p_j^{λ} between 0 and r_j^{λ} is uniformly distributed (see Section 2.3); it follows further that $\underline{p}_j^{\lambda}$'s null distribution conditional only on $\hat{\beta}_j^{\lambda} \neq 0$ is also $\mathrm{Unif}(0,1)$.

Corollary 4.1 (Validity of the conditional ℓ-test)**.** *For model* (1.1)*, for all* $\alpha \in [0,1]$, $\mathbb{P}_{H_j}(\underline{p}_j^{\lambda} \leq \alpha \mid \hat{\beta}_j^{\lambda} \neq 0) = \alpha$.

Computationally, only one extra LASSO (for $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(0)$, which goes into r^{λ}_j) needs to be run to compute $\underline{p}^{\lambda}_j$ for an index j with $\hat{\beta}^{\lambda}_j \neq 0$. And since everything above is conditional on $\boldsymbol{S}^{(j)}$, the same result holds true when using $\hat{\lambda}_j$ from Section 2.4 (i.e., $\underline{p}^{\hat{\lambda}_j}_j$ is conditionally valid given $\hat{\beta}^{\hat{\lambda}_j}_j \neq 0$), since $\hat{\lambda}_j$ is conditionally independent of the data given $\boldsymbol{S}^{(j)}$.

Now for obtaining post-LASSO-selection valid confidence interval for β_j, we need to invert a conditionally valid test for $H_j(\gamma), \gamma \in \mathbb{R}$. For testing $H_j(\gamma)$, following the suggestion in Section 3, the test statistic should be based on the LASSO estimate on $(\boldsymbol{y}-\gamma\boldsymbol{X}_j, \boldsymbol{X})$, that is $|\hat{\beta}^{\lambda}_j(\gamma)|$, whereas the model selection event is still based on the original, un-centered LASSO estimate, $\hat{\beta}^{\lambda}_j$. Furthermore, one can choose to use a different penalty parameter for the test statistic than the one used for the selection event. This prompts us to understand tests for $H_j(\gamma)$ based on $|\hat{\beta}^{\lambda_\ell}_j(\gamma)|$, valid conditionally on $\{\hat{\beta}^{\lambda_s}_j \neq 0\}$, where λ_ℓ and λ_s need not be the same. Because now the test-statistic and the selection event are based on different LASSO estimates, a conditional p-value would not have such a simple form as for $\underline{p}^{\lambda}_j$, but we can still obtain valid p-value using CDF transforms if we can characterize the distribution of $|\hat{\beta}^{\lambda_\ell}_j(\gamma)| \mid \{\hat{\beta}^{\lambda_s}_j \neq 0\}$, under $H_j(\gamma)$.

First note that, as discussed in Section 3, Lemma 2.1 can be applied to $(\boldsymbol{y} - \gamma\boldsymbol{X}_j, \boldsymbol{X})$ under $H_j(\gamma)$ to show that $\boldsymbol{S}^{(j)}(\gamma) = (\boldsymbol{X}_j^T\boldsymbol{y}, \hat{\sigma}_j(\gamma))$ is sufficient under $H_j(\gamma)$, where, $\hat{\sigma}_j(\gamma) = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{y} - \gamma\boldsymbol{X}_j)\|$, and that $\boldsymbol{y} - \gamma\boldsymbol{X}_j$ can be written as $\boldsymbol{P}_{-j}(\boldsymbol{y} - \gamma\boldsymbol{X}_j) + \hat{\sigma}_j(\gamma)\boldsymbol{V}\boldsymbol{u}^{\gamma}$, with $\boldsymbol{u}^{\gamma} \mid \boldsymbol{S}^{(j)}(\gamma) \overset{H_j(\gamma)}{\sim} \text{Unif}(\mathbb{S}^{n-d})$. In light of this result, one can now apply Theorem 2.1 to $(\boldsymbol{y} - \gamma\boldsymbol{X}_j, \boldsymbol{X})$ to conclude that $\hat{\beta}^{\lambda_\ell}_j(\gamma) \leq b$ if and only if $u^{\gamma}_1 \leq \Lambda^*_j(b, \text{sign}(b); \gamma)$, where $\Lambda^*_j(b, \epsilon; \gamma)$ is exactly equal to $\Lambda(b, \epsilon)$ but with $\boldsymbol{y}$ replaced with $\boldsymbol{y} - \gamma\boldsymbol{X}_j$ and $\text{sign}(0) := 1$. In fact Theorem A.1 (that characterizes the distribution of $\hat{\beta}^{\lambda}_j \mid \boldsymbol{S}^{(j)}(\gamma)$ under $H_j(\gamma)$) in Appendix A shows that this same u^{γ}_1 can be used to characterize the event $\{\hat{\beta}^{\lambda_s}_j = 0\}$, stating it is equivalent to $\{u^{\gamma}_1 \in [\Lambda_j(0, \pm 1; \gamma)]\}$, where,

$$\Lambda_j(0, \pm 1; \gamma) = \frac{-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j + \gamma(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}^{\lambda_s}_{-j}(0)) \pm n\lambda_s}{\hat{\sigma}_j(\gamma)\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|}.$$

Because $\Lambda_j(0, \pm 1; \gamma)$ and $\Lambda^*_j(b, \text{sign}(b); \gamma)$ are all functions of $\boldsymbol{S}^{(j)}(\gamma)$, we have that

$$\begin{aligned}\underline{F}^{\lambda_\ell, \lambda_s}_{\ell}(b \mid \boldsymbol{S}^{(j)}(\gamma); \gamma) &:= \mathbb{P}_{H_j(\gamma)}(\hat{\beta}^{\lambda_\ell}_j(\gamma) \leq b \mid \hat{\beta}^{\lambda_s}_j \neq 0, \boldsymbol{S}^{(j)}(\gamma)) \\ &= \mathbb{P}_{H_j(\gamma)}(u^{\gamma}_1 \leq \Lambda^*_j(b, \text{sign}(b); \gamma) \mid u^{\gamma}_1 \notin [\Lambda_j(0, \pm 1; \gamma)], \boldsymbol{S}^{(j)}(\gamma)),\end{aligned}$$

where the last expression can exactly be evaluated using the known quantiles of F_u defined in Section 2.1. This gives us the adjusted p-value for testing $H_j(\gamma)$:

$$\underline{p}^{\lambda_\ell, \lambda_s}_j(\gamma) := 1 - \underline{F}^{\lambda_\ell, \lambda_s}_{\ell}(|\hat{\beta}^{\lambda_\ell}_j(\gamma)| \mid \boldsymbol{S}^{(j)}(\gamma); \gamma) + \lim_{b \to |\hat{\beta}^{\lambda_\ell}_j(\gamma)|^-} \underline{F}^{\lambda_\ell, \lambda_s}_{\ell}(b \mid \boldsymbol{S}^{(j)}(\gamma); \gamma),$$

where we can also use the strategy in Section 2.3 to break ties when $\hat{\beta}^{\lambda_\ell}_j = 0$. As one would expect, our original conditional ℓ-test p-value $\underline{p}^{\lambda}_j$ is a special case of the above, with $\underline{p}^{\lambda}_j = \underline{p}^{\lambda, \lambda}_j(0)$. Finally, for $\lambda_s = \lambda$, the above test can be inverted to obtain a confidence

interval for β_j valid conditionally on $\{\hat{\beta}_j^\lambda \neq 0\}$. Two particularly interesting choices for such a $100(1-\alpha)\%$ confidence interval are

$$\underline{\hat{C}}_j^\lambda = \text{conv}\left(\left\{\gamma : \underline{p}_j^{\lambda,\lambda}(\gamma) > \alpha\right\}\right) \text{ and } \underline{\hat{C}}_j^{*\lambda} = \text{conv}\left(\left\{\gamma : \underline{p}_j^{\hat{\lambda}_\ell(\gamma),\lambda}(\gamma) > \alpha\right\}\right),$$

which use $\lambda_\ell = \lambda_s = \lambda$ and a cross-validated choice for λ_ℓ (see Section 2.4), respectively. Note that our cross-validation strategy does not allow for a data-adaptive choice for the selection λ (i.e., λ_s), as for the individual ℓ-test p-values $\underline{p}_j^{\lambda_\ell,\lambda_s}(\gamma)$ to be valid, λ_s needs to be a function of $\boldsymbol{S}^{(j)}(\gamma)$ (and maybe some external, conditionally independent, sources of randomness), and this needs to hold for all $\gamma \in \mathbb{R}$.

Like Liu et al. (2018) but unlike, e.g., Lee et al. (2016), our conditional inferences do not condition on anything about the LASSO's selection except that it selects the j^{th} coefficient. As mentioned in Section 1.3, the key difference between our conditional inference and Liu et al. (2018)'s is essentially the same as the difference between the (unconditional) ℓ-test and the t-test, and indeed in Section 5.4 we find that our conditional confidence intervals improve over those of Liu et al. (2018) similarly to how the unconditional ℓ-test confidence intervals from Section 3 improve over standard t-test confidence intervals.

5 Experiments

In this section, we perform experiments to evaluate the performance of the ℓ-test and its corresponding confidence intervals and post-selection procedures. For all simulations except in Section 5.2 (where we study the robustness of the ℓ-test to deviations from model (1.1)), we use a linear model (1.1) with k out of the d elements of $\boldsymbol{\beta}$ chosen uniformly without replacement and set to A or $-A$ with equal probability, and all the other remaining entries set to 0. We perform inference on one randomly chosen signal coefficient, β_j. The rows of $\boldsymbol{X}$ are drawn i.i.d. from $\mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$ and then the columns are normalized. We will specify the values of $n, d, k, A, \boldsymbol{\Sigma}$, and σ for each of the simulation settings we consider.

5.1 Power of the ℓ-test

We compare the power of the following three tests: The ℓ-test, the two-sided t-test, and the one-sided t-test in the direction of the true sign of β_j, under simulation settings studying the effect of varying the amplitude of the signal variables, the number of signal variables, and the inter-variable correlation. Note there is no need to compare Type I error rates, since all three methods have guaranteed exactly nominal Type I error (as long as model (1.1) is well-specified, which it is in this subsection). The results are reported in Figure 2 (see the caption for further details of the simulations). A simulation considering the case where d is closer to n is provided in Appendix F.1; the agreement between the ℓ-test and the one-sided t-test becomes even stronger in this case.

The ℓ-test significantly outperforms the t-test in sparse settings for any signal amplitude, achieving essentially one-sided t-test power for moderate-to-high signal amplitudes. The ℓ-test's power remains close to that of the one-sided t-test when as many as 30% of the coefficients are signals, and this phenomenon seems to be similar across covariate correlation

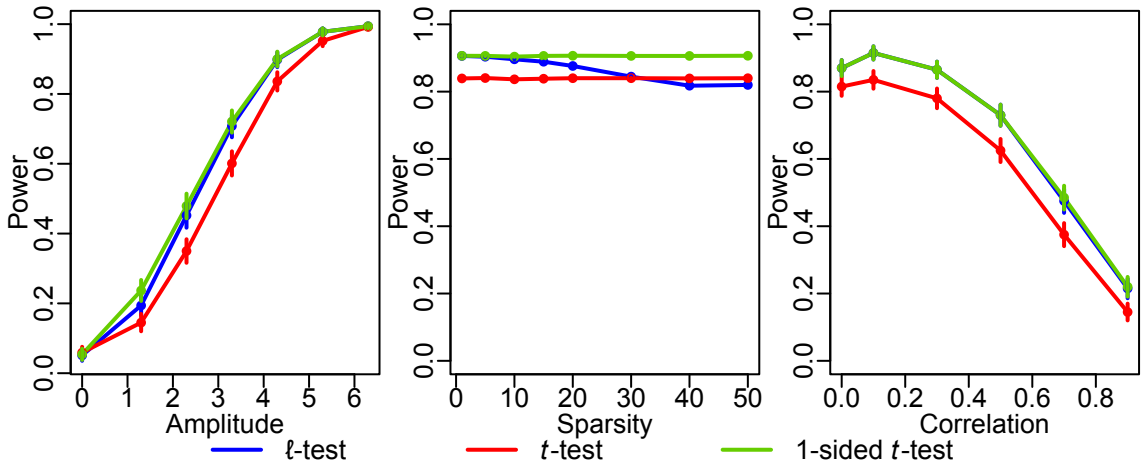

Figure 2: Power comparison of size-5% tests for H_j. For all the settings, we fix $n = 100, d = 50, \sigma = 1$. For left, we fix $k = 5, \mathbf{\Sigma} = \boldsymbol{I}$ and vary the amplitude A. For center, we fix $A = 4.3, \mathbf{\Sigma} = \boldsymbol{I}$ and vary the sparsity level k. For right, we fix $A = 4.3, k = 5, \Sigma_{ij} = \rho^{|i-j|}$ and vary the inter-variable correlation ρ. The error bars represent plus or minus two standard errors.

levels. Furthermore, the ℓ-test only starts to underperform the t-test after more than 60% of the coefficients are non-zero (as one might expect, given the ℓ-test is designed to leverage sparsity), and even when the signal is fully dense (100% nonzero entries, all with equal magnitude), the ℓ-test's power loss is still only a fraction of its power gain in sparse settings.

5.2 Robustness of the ℓ-test

One thing that makes the t-test remarkable and so useful in practice is its robustness, even in relatively small samples, to violations of model (1.1). To evaluate the ℓ-test's robustness, we fix $d = n/2, k = 1, \mathbf{\Sigma} = \boldsymbol{I}$, $A = 3.3$, and test on a null index $\beta_j = 0$. Figure 3 shows Type I error results for the ℓ-test and t-test for four types of model violation: heavy-tailed errors, skewed errors, heteroskedastic errors, and model non-linearity (the figure caption gives exact specifications of each of the model violations) for a range of small sample sizes. In Appendix F.2 we present further simulations of these same four types of model violations, but Figure 3 shows the most extreme example of each of the four. Despite substantial deviations from (1.1), the ℓ-test remains quite robust: like the t-test, it only makes even noticeable Type I error violations in the case of (very subtantial) heteroskedasticity, but even in this case (which is a well-known source of non-robustness for the t-test), the Type I error violation of the ℓ-test remains close to that of the t-test and both are just a few percentage points above nominal.

5.3 Confidence Intervals for β_j

We now perform simulations to compare the ℓ-test confidence interval with the usual (two-sided) t-test confidence interval, as well as the interval obtained by inverting the one-sided

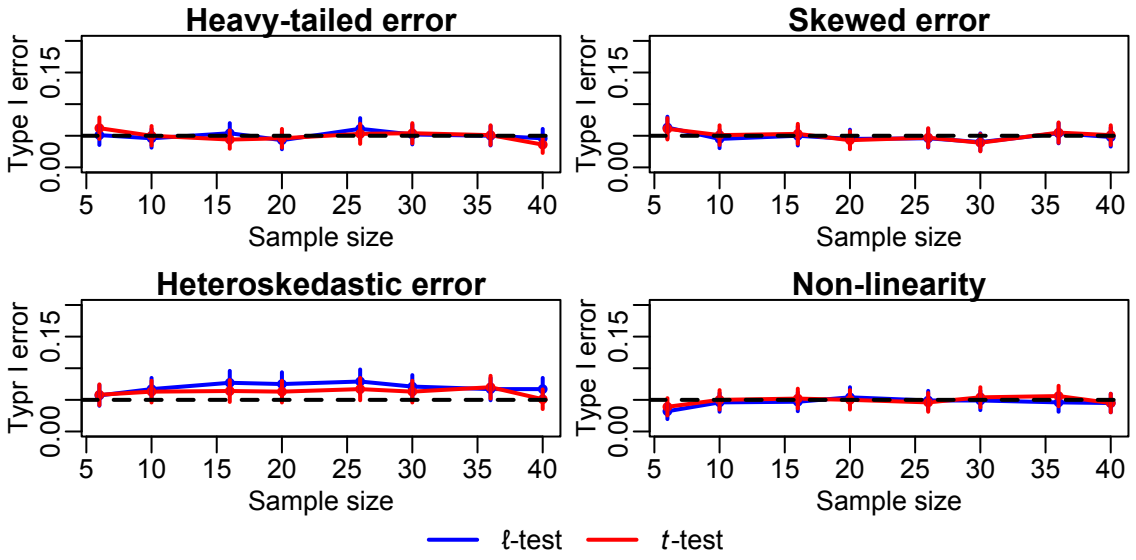

Figure 3: We vary the sample size (n) on the x-axis and set $d = n/2$. For the top-left, we draw $\epsilon_i \overset{i.i.d.}{\sim} t_2$ (which does not have a finite second moment), for the top-right, $\epsilon_i \overset{i.i.d.}{\sim}$ Exp(1), standardized with its theoretical mean and standard deviation, and for the bottom-left, $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$ if mean of the i^{th} row of $\boldsymbol{X}$ is less than the median of the row-means, while, $\epsilon_i \overset{i.i.d.}{\sim} N(0,8)$, otherwise. For bottom-right, we generate $y_i \sim \mathcal{N}\left((\boldsymbol{X}_i^4)^T\boldsymbol{\beta}, \boldsymbol{I}\right)$, where, $(\boldsymbol{X}_i^4)_j = X_{ij}^4$. All these settings use the nominal size of 5% and the error bars represent plus or minus two units of standard error.

t-test in the direction of the true sign of the alternative for every alternative $\beta_j = \gamma$. In Appendix G, we discuss an explicit characterization of this last interval, however a point we highlight here is that this is an oracle procedure (even more so than the one-sided t-test we compare to in Section 5.1) and can only be constructed if we know the *exact* true value of the coefficient, not just its sign. To obtain the ℓ-test confidence interval, we choose a grid of candidate values for β_j and report the coverage and length of the smallest interval that strictly encloses the rejected values of β_j from both the ends. Note that this interval will always have length and coverage at least as large as the true ℓ-test confidence interval (which we can never obtain exactly from a finite grid of β_j values). We use brute force calculations using the highly optimized functions available in the package `glmnet` (Friedman et al., 2010) in **R** instead of our theoretically efficient algorithm in Appendix D for ℓ-test inversion (and defer the task of designing an optimized implementation for it to future research).

For our simulations we consider the exact settings as in Figure 2 and summarize the results in Figure 4. We only report the lengths of all the intervals for a more compact presentation whereas the full results with empirical converges are reported in Appendix F.3 (the coverage is always extremely close to the nominal 95%). As expected, we see similar trends as in Figure 2, with the ℓ-test confidence intervals being close to the oracle one-sided t-test intervals, consistently across amplitudes, in sparse settings. In the left and right plots, the ℓ-test confidence intervals are consistently about 12% shorter than their t-test based counterparts. Perhaps surprisingly, the center plot shows that the ℓ-test's benefit

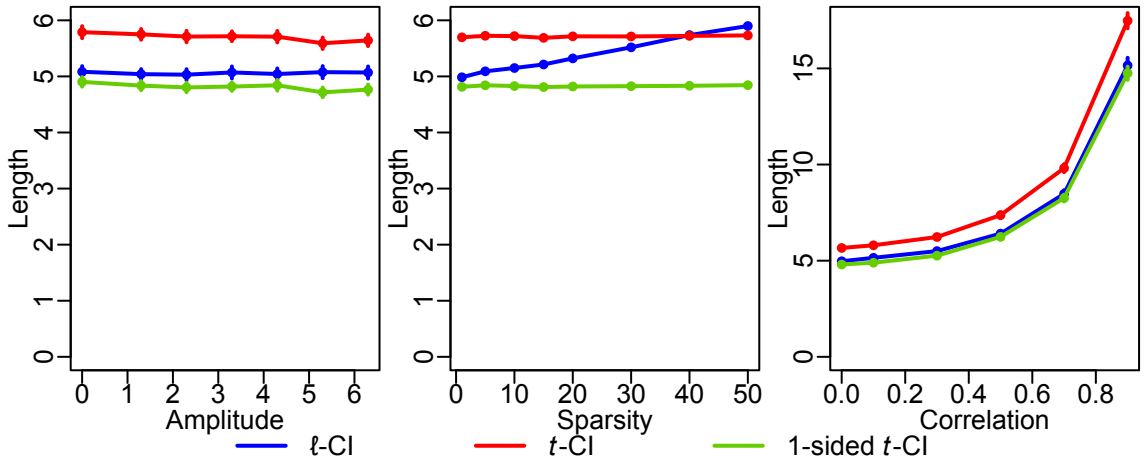

Figure 4: Coverage and lengths of the 95% confidence intervals. Figures on the left, center and right are under the exact same experimental settings as are the respective figures in Figure 2. The error bars represent plus or minus two standard errors.

in confidence interval width over the t-test remains up until about 80% nonzero entries in the coefficient vector, which is a larger outperformance range of sparsity than in Figure 2. Similar to the power results, we see only a small detriment to using the ℓ-test confidence interval in the densest setting, relative to its benefit in the sparsest setting. As with the power simulations, we also consider a setting with d closer to n in Appendix F.3, and again find this further narrows the gap between the ℓ-test and one-sided t-test procedures.

5.4 Post-selection ℓ-test inference

In this section, we perform simulations to empirically evaluate the performance of the adjusted ℓ-test confidence intervals in Section 4 for post-selection inference in the linear model under LASSO selection. In particular, we will compare $\underline{\hat{C}}_j^{\lambda}$ and $\underline{\hat{C}}_j^{*\lambda}$, where for both of these methods we invert the respective adjusted ℓ-tests on a grid of values using the same strategy as in Section 5.3, along with the conditional confidence interval procedure of Liu et al. (2018). Figure 5 shows the effect of varying coefficient amplitude on the length and coverage of the intervals (conditional on the LASSO selecting the coefficient) for the same setting as the left panels of Figures 2. We see there is practically no difference between the performance of $\underline{\hat{C}}_j^{\lambda}$ and $\underline{\hat{C}}_j^{*\lambda}$ (see figure caption for how λ was chosen) and both are consistently shorter than the method in Liu et al. (2018), reflecting again the benefits of the ℓ-test under sparsity but now conditional on LASSO selection. In Appendix F.4, we show results for an additional setting where d is closer to n.

5.5 Analysis of the HIV drug resistance data

The HIV drug resistant data (Rhee et al., 2006) consists of 16 different regressions, each containing data on a set of genetic mutations (the covariates) and a score measuring resistance to an HIV drug (the response). We follow the same pre-processing step suggested in

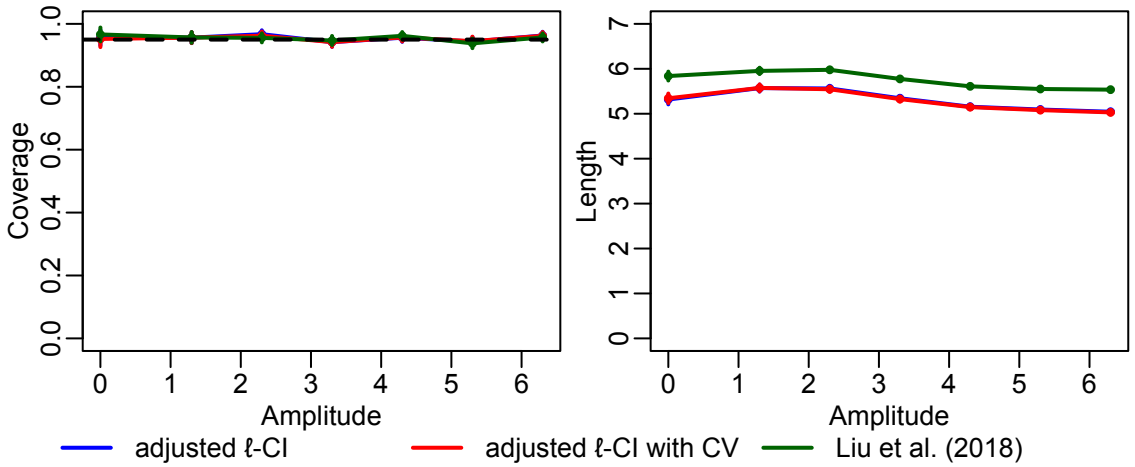

Figure 5: Length and coverage of various post-selection 95% confidence intervals. We are under exactly the same setting as the left panel of Figure 4. Here 'adjusted ℓ-CI' and 'adjusted ℓ-CI with CV' refer to $\hat{\underline{C}}_j^{\lambda}$ and $\hat{\underline{C}}_j^{*\lambda}$, respectively, and we have used $\lambda = 0.01$ (which approximately matched the average cross-validated value found on independent data sets, for the entire range of amplitudes). The error bars represent plus or minus two standard errors.

Barber and Candès (2015), resulting in 16 regressions with n ranging between 328 and 842 and d ranging between 147 and 313. Running the t-test on all covariates across all data sets results in an average power of 16.7%, while for the ℓ-test it is 18.6% (this represents an 11% improvement), showing that the ℓ-test's theoretical benefits under sparsity provide genuine power gains in real data sets. Similarly, the t-test confidence intervals have an average width of 3.85 while the ℓ-test confidence intervals' is 3.54, representing about an 8% improvement. We also note that previous works studying the same data have aggregated discovered mutations to the gene level, and if we do this, the comparison remains similar: the t-test's average power is 31.8% discovered genes while the ℓ-test's is 36.4% (a 14% improvement).

6 Discussion

The ℓ-test leverages sparsity by using the LASSO coefficient estimate $|\hat{\beta}_j^{\lambda}|$ as its test statistic and can achieve power close to that of the one-sided t-test without any knowledge about the true sign of the coefficient. The ℓ-test can be inverted to confidence intervals that are over 10% shorter than t-test intervals under sparsity, and the ℓ-test and confidence intervals can also be adjusted for LASSO selection. A number of questions remain for future work:

1. *A recentered u_1-based test.* Section 2.2 argued that the ℓ-test's power gains under sparsity are derived from $\hat{m}_j$ tending to take the opposite sign of β_j, as this results in the ℓ-test p-value (which uses the test statistic $|\hat{\beta}_j^{\lambda}|$) putting most of its weight on the "correct" tail of $\hat{\beta}_j^{\lambda}$. In fact, due to the increasing relationship between u_1 and $\hat{\beta}_j^{\lambda}$ established in Theorem 2.1, it is easy to see that a similar phenomenon occurs if we use $|u_1 - \hat{m}_j|$ as the test statistic (note this is the same test statistic used to break

ties in the ℓ-test in Section 2.3), and indeed, although doing so produces a test that is numerically distinct from the ℓ-test, the two tests perform very similarly. We presented the ℓ-test in its current form because we expect it to be more interpretable for most (especially non-statistician) users to have its test statistic be a natural estimator of β_j under sparsity, as well as because of the easy extension to post-selection inference, but recognizing that the test statistic $|u_1 - \hat{m}_j|$ produces a very similar test may be helpful in generalizing the ℓ-test idea or for theoretical study such as power analysis.

2. *Extension to Gaussian means and, asymptotically, to general parametric models.* A Gaussian means problem, wherein $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is observed for some known positive-definite $\boldsymbol{\Sigma}$ (but $\boldsymbol{\beta}$ unknown), can always be converted to linear regression by taking $\boldsymbol{X} = \boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{y} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}$, so that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n)$. The key point is that the ℓ-test[3] can be directly applied to any Gaussian means problem with known covariance matrix, with the analogous expectation that, when $\boldsymbol{\beta}$ is sparse, the ℓ-test will outperform the standard z-test as long as $\boldsymbol{\Sigma}$ has sufficient off-diagonal entries (recall from Section 2.2 that the ℓ-test's power gain under sparsity derives from the sign-guessing ability of $\hat{m}_j$, which in turn relies on non-orthogonality of the columns of $\boldsymbol{X}$). This opens up many possible further applications of the ℓ-test, including to *any* multivariate parametric model that admits an asymptotically multivariate Gaussian estimator and a consistent estimator for its covariance matrix, which includes both classical low-dimensional maximum likelihood or M-estimator asymptotics (see (Li and Fithian, 2021, Theorem 4) for an asymptotic result like this for knockoffs) as well as certain proportional asymptotic regimes with $n \propto d$ (Sur and Candès, 2019; Zhao et al., 2022).

3. *Leveraging structure other than sparsity.* The ℓ-test leverages sparsity to improve upon the t-test, but it would be helpful to have analogous methods to use in settings that may not be sparse, but are believed to satisfy other forms of structure. For instance, if $\boldsymbol{\beta}$ is dense but smooth (i.e., it has small total variation), a test based on the fused LASSO (Tibshirani et al., 2005) may be more appropriate. The FAB method (Hoff, 2022) mentioned in Section 1.3 achieves a similar goal as the ℓ-test for other forms of prior information and it may also be interesting to understand the relationship between the two.

Acknowledgements

The authors would like to thank Danielle Paulson and Jonathan Taylor for valuable discussions and comments. LJ and SS were partially supported by DMS-2045981.

[3]When σ^2 is known, the ℓ-test can easily use this information for a slight improvement by shrinking $\boldsymbol{S}^{(j)}$ to just its first element $\tilde{\boldsymbol{S}}^{(j)} := \boldsymbol{X}_{-j}^T\boldsymbol{y}$ and replacing Equation (2.2) with the statement that $\hat{\sigma}_j\boldsymbol{u} \mid \tilde{\boldsymbol{S}}^{(j)} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n-d+1})$. Theorem 2.1 remains unchanged, and in particular Equation (2.5) equates the event $\{\hat{\beta}_j^\lambda \le b\}$ with $\{\hat{\sigma}_j u_1 \le \hat{\sigma}_j \Lambda_j(b, \text{sign}(b))\}$, where $\hat{\sigma}_j\Lambda_j(b, \text{sign}(b))$ is easily seen to be a function only of $\tilde{\boldsymbol{S}}^{(j)}$ as required for straightforward evaluation of the probability of this event via the Gaussian CDF.

References

J. N. Adichie. Asymptotic Efficiency of a Class of Non-Parametric Tests for Regression Parameters. *The Annals of Mathematical Statistics*, 38(3):884 – 893, 1967. doi: 10.1214/aoms/1177698882. URL https://doi.org/10.1214/aoms/1177698882.

F. Bachoc, D. Preinerstorfer, and L. Steinberger. Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48, 11 2016. doi: 10.1214/19-AOS1815.

R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055 – 2085, 2015. doi: 10.1214/15-AOS1337. URL https://doi.org/10.1214/15-AOS1337.

R. F. Barber and L. Janson. Testing goodness-of-fit and conditional independence with approximate co-sufficient sampling. *The Annals of Statistics*, 50(5):2514 – 2544, 2022. doi: 10.1214/22-AOS2187. URL https://doi.org/10.1214/22-AOS2187.

M. S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282, 1937.

C. Berge. Topological spaces including a treatment of multi-valued functions, vector spaces and convexity. Edinburgh-London: Oliver & Boyd, Ltd. XIII, 270 p. (1963)., 1963.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802 – 837, 2013. doi: 10.1214/12-AOS1077. URL https://doi.org/10.1214/12-AOS1077.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018.

D. R. Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62: 441–444, 1975. URL https://api.semanticscholar.org/CorpusID:119955345.

C. Dai, B. Lin, X. Xing, and J. S. Liu. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 0(0):1–18, 2022. doi: 10.1080/01621459.2022.2060113. URL https://doi.org/10.1080/01621459.2022.2060113.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL https://doi.org/10.1214/009053604000000067.

R. A. Fisher. The goodness of fit of regression formulae, and the distribution of regression coefficients. *Journal of the Royal Statistical Society*, 85(4):597–612, 1922.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. 10 2014.

D. A. Freedman. Bootstrapping Regression Models. *The Annals of Statistics*, 9(6):1218 – 1228, 1981. doi: 10.1214/aos/1176345638. URL https://doi.org/10.1214/aos/1176345638.

J. Friedman, R. Tibshirani, and T. Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937. doi: 10.1080/01621459.1937.10503522. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

C. Gutenbrunner, J. JurečKová, R. Koenker, and S. Portnoy. Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, 2(4):307–331, Jan. 1993. ISSN 1048-5252. doi: 10.1080/10485259308832561.

J. D. Habiger and E. A. Peña. Compound p-value statistics for multiple testing procedures. *Journal of Multivariate Analysis*, 126:153–166, 2014. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2014.01.007. URL https://www.sciencedirect.com/science/article/pii/S0047259X14000153.

J. Hajek. Asymptotically Most Powerful Rank-Order Tests. *The Annals of Mathematical Statistics*, 33(3):1124 – 1147, 1962. doi: 10.1214/aoms/1177704476. URL https://doi.org/10.1214/aoms/1177704476.

P. Hoff. Smaller p-values via indirect information. *Journal of the American Statistical Association*, 117(539):1254–1269, 2022. doi: 10.1080/01621459.2020.1844720. URL https://doi.org/10.1080/01621459.2020.1844720.

P. Hoff and C. Yu. Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13(1):94 – 119, 2019. doi: 10.1214/18-EJS1517. URL https://doi.org/10.1214/18-EJS1517.

D. Huang and L. Janson. Relaxing the assumptions of knockoffs by conditioning. *Ann. Statist.*, 48(5):3021–3042, 10 2020. doi: 10.1214/19-AOS1920. URL https://doi.org/10.1214/19-AOS1920.

L. A. Jaeckel. Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals. *The Annals of Mathematical Statistics*, 43(5):1449 – 1458, 1972. doi: 10.1214/aoms/1177692377. URL https://doi.org/10.1214/aoms/1177692377.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014. URL http://jmlr.org/papers/v15/javanmard14a.html.

W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280779.

J. Lee and Z. Ren. Boosting e-bh via conditional calibration, 2024.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

E. L. Lehmann and H. Scheffé. Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236, 1955. ISSN 00364452. URL http://www.jstor.org/stable/25048243.

L. Lei and P. J. Bickel. An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412, 09 2020. ISSN 0006-3444. doi: 10.1093/biomet/asaa079. URL https://doi.org/10.1093/biomet/asaa079.

X. Li and W. Fithian. Whiteout: when do fixed-x knockoffs fail?, 2021.

K. Liu, J. Markovic, and R. Tibshirani. More powerful post-selection inference, with application to the lasso, 2018.

M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 2021+. To Appear.

Y. Luo, W. Fithian, and L. Lei. Improving knockoffs with conditional calibration, 2022.

S. Panigrahi, J. Taylor, and A. Weinstein. Integrative methods for post-selection inference under convex constraints. *The Annals of Statistics*, 49(5):2803–2824, 2021.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2):225–232, 1937. ISSN 14666162. URL http://www.jstor.org/stable/2983647.

E. J. G. Pitman. Significance Tests which May be Applied to Samples from any Populations: III. The Analysis of Variance Test. *Biometrika*, 29(3-4):322–335, 02 1938. ISSN 0006-3444. doi: 10.1093/biomet/29.3-4.322. URL https://doi.org/10.1093/biomet/29.3-4.322.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

Z. Ren and R. F. Barber. Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad085, 09 2023. ISSN 1369-7412. doi: 10.1093/jrsssb/qkad085. URL https://doi.org/10.1093/jrsssb/qkad085.

S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006. doi: 10.1073/pnas.0607274103. URL https://www.pnas.org/doi/abs/10.1073/pnas.0607274103.

A. Rinaldo, L. Wasserman, M. G'Sell, J. Lei, and R. Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *The Annals of Statistics*, 47, 11 2016. doi: 10.1214/18-AOS1784.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012 – 1030, 2007. doi: 10.1214/009053606000001370. URL https://doi.org/10.1214/009053606000001370.

A. Spector and L. Janson. Powerful knockoffs via minimizing reconstructability. *Annals of Statistics*, 50(1):252–276, 2022.

M. Stephens. Goodness-of-fit and sufficiency: Exact and approximate tests. *Methodology and Computing in Applied Probability*, 14, 09 2012. doi: 10.1007/s11009-011-9267-2.

P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

X. Tian and J. Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108, 2005.

R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111: 600–620, 04 2016. doi: 10.1080/01621459.2015.1108848.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, jun 2001. ISSN 0022-3239. doi: 10.1023/A:1017501703105. URL https://doi.org/10.1023/A:1017501703105.

J. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614, 01 1958. doi: 10.1214/aoms/1177706647.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166 – 1202, 2014. doi: 10.1214/14-AOS1221. URL https://doi.org/10.1214/14-AOS1221.

X. Xing, Z. Zhao, and J. S. Liu. Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241, 2023. doi: 10.1080/01621459.2021.1923510. URL https://doi.org/10.1080/01621459.2021.1923510.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/24772752.

Q. Zhao, P. Sur, and E. J. Candes. The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli*, 28(3):1835–1861, 2022.

A Characterization of the ℓ-distribution under $H_j(\gamma) : \beta_j = \gamma$

In Section 2.1, we characterized the conditional distribution, $\hat{\beta}_j^\lambda \mid \boldsymbol{S}^{(j)}$ under H_j based on the quantiles of u_1. In this section, we extend the result to provide a similar characterization of the ℓ-distribution under $H_j(\gamma) : \beta_j = \gamma$.

Theorem A.1. *Consider the linear model defined in Theorem 2.1, fix $\gamma \in \mathbb{R}$ and define $\hat{\sigma}_j(\gamma) = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{y} - \gamma \boldsymbol{X}_j)\|$. Then,*

1. *$\boldsymbol{S}^{(j)}(\gamma) = (\boldsymbol{X}_{-j}^T \boldsymbol{y}, \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{y} - \gamma \boldsymbol{X}_j)\|)$ is sufficient under $H_j(\gamma)$. Furthermore, fixing an orthogonal matrix $\boldsymbol{V}$ for the column-space of $\boldsymbol{X}_{-j}$ as in Lemma 2.1, there exists a unique vector $\boldsymbol{u}^\gamma \in \mathbb{S}^{n-d}$ such that $\boldsymbol{y} = \hat{\boldsymbol{y}}_j + \gamma(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j + \hat{\sigma}_j(\gamma)\boldsymbol{V}\boldsymbol{u}^\gamma$ and*

$$\boldsymbol{u}^\gamma \mid \boldsymbol{S}^{(j)}(\gamma) \overset{H_j(\gamma)}{\sim} \text{Unif}\left(\mathbb{S}^{n-d}\right).$$

2. *Furthermore, analogously define $\Lambda_j(\cdot, \cdot; \gamma) : \mathbb{R}^2 \mapsto \mathbb{R}$ by*

$$\Lambda_j(b, \epsilon; \gamma) = \frac{-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j + \gamma(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}^\lambda(b)) + n\lambda\epsilon}{\hat{\sigma}_j(\gamma)\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|}.$$

Then the function $f^\gamma_{\boldsymbol{S}^{(j)}(\gamma)} : \mathbb{R} \mapsto \mathbb{R}$, whose inverse is defined as

$$\left(f^\gamma_{\boldsymbol{S}^{(j)}(\gamma)}\right)^{-1}(b) = \begin{cases} \Lambda_j(b, \text{sign}(b); \gamma), & b \neq 0 \\ [\Lambda_j(0, -1; \gamma), \Lambda_j(0, 1; \gamma)], & b = 0 \end{cases},$$

is continuous and increasing in $\mathbb{R}$ and strictly increasing in $\{u : f^\gamma_{\boldsymbol{S}^{(j)}(\gamma)}(u) \neq 0\}$ and satisfies $\hat{\beta}_j^\lambda = f^\gamma_{\boldsymbol{S}^{(j)}}(u_j^\gamma)$.

Note that, as discussed in Section 4, the proof of item 1 of the above theorem directly follows from Lemma 2.1 applied to $(\boldsymbol{y} - \gamma \boldsymbol{X}_j, \boldsymbol{X})$. Analogous to the proof of Theorem 2.1 using Lemma 2.1, one can derive item 2 of Theorem A.1 from its item 1, by defining $\boldsymbol{z} := \boldsymbol{y} - \gamma \boldsymbol{X}_j$ and $\boldsymbol{\delta}$ by $\boldsymbol{\delta}_{-j} = \boldsymbol{\beta}_{-j}$ and $\delta_j = \beta_j - \gamma$ and copying exactly the same proof as in Appendix C.3 but by replacing $(\boldsymbol{y}, \boldsymbol{\beta}, \hat{\sigma}_j)$ with $(\boldsymbol{z}, \boldsymbol{\delta}, \hat{\sigma}_j(\gamma))$ and then substituting back for $\boldsymbol{\delta}$ and $\boldsymbol{z}$ at the end. We thus skip an explicit proof of Theorem A.1.

Here we would also like to draw attention to the fact that the function $f^{\gamma}_{\boldsymbol{S}^{(j)}(\gamma)}$, for any $\gamma \in \mathbb{R}$, is defined for *any* real number u and not necessarily restricted to $[0,1]$ and also that $\Lambda_j(b, \text{sign}(b))$ can take values outside the $[-1,1]$ range. In fact if the values exceed this range, we can often draw conclusive insights about the behavior of the LASSO estimate $\hat{\beta}^{\lambda}_j$. For example, $\Lambda(0,\pm 1) < -1$ implies that $u^{\gamma}_1 > \Lambda(0,1)$ and hence that for any $\boldsymbol{y}$ generated using the condition in Theorem A.1 for a $\boldsymbol{u}^{\gamma} \in \mathbb{S}^{n-d}$, the resultant LASSO estimator of the j^{th} coefficient, $\hat{\beta}^{\lambda}_j$, will always be positive. This can happen in situations as we next described in Section B and can in-fact, result in the ℓ-test producing exactly the one-sided t-test p-value.

B Achieving the power of a one-sided t-test

The conclusions of the previous section show that if $\beta_j > 0$ (which, without any loss of generality, we will assume throughout this section), the ℓ-test would produce the exact p-value of a one-sided t-test if $\Lambda(0,\pm 1) = v_{\pm} < -1$ (as in this case $\hat{\beta}^{\lambda}_j > 0$ under the null conditional distribution of $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$, and hence, the ℓ-distribution puts all its mass on the positive half and we saw in Section 2 that the contribution from this part to the ℓ-test p-value is exactly the one-sided t-test p-value). In this section, we will try to take a closer look at when this can be the case. Note that we introduced $\hat{m}_j$ in Section 2.2, defined by,

$$\hat{m}_j = \frac{-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(0))}{\hat{\sigma}_j \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|},$$

which is the mid-point of the interval $[v_-, v_+]$, and argued that its numerator is a proxy to a quantity given by,

$$-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}) \sim \mathcal{N}(-\beta_j \boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j, \sigma^2 \boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j).$$

Thus roughly speaking, one can observe $v_-, v_+ \leq -1$ if $q_j = \frac{\beta_j \boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}$ is highly positive. In particular, the magnitude of q_j intuitively quantifies the reliability of the sign-guess. Notably, larger values of $|q_j|$ indicates higher differences between the mass the ℓ-distribution assigns in the two halves of the real line (and hence is more asymmetric), thereby implying that the ℓ-test p-value is much different from its two-sided-t-test counterpart. Now that we have an understanding of the role that q_j plays, we next describe two situations in which one can observe $v_-, v_+ \leq -1$:

- **Strong signal size:** For any fixed design matrix $\boldsymbol{X}$, if the signal size β_j is strong enough to make the quantity q_j highly positive, one can expect that $\hat{m}_j$ would be sufficiently negative and hence, we can expect to see the p-value of a one-sided t-test. This result is intuitive as with increase in the signal size the variable gets more and more distinguishable. Thus the power of the one-sided t-test, the two-sided test and the ℓ-test, all increase with the power of the latter getting closer to the power of the one-sided t-test.

- **High feature correlation:** Note that except for β_j, the other factor in q_j,

$$\frac{\boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} = \frac{\boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j}{\sqrt{\|\boldsymbol{X}_j\|^2 - \boldsymbol{X}_j^T \boldsymbol{P}_{-j}\boldsymbol{X}_j}},$$

depends on the design. This shows that as $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$ increases and gets closer to $\boldsymbol{X}_j^T \boldsymbol{X}_j$, this factor starts blowing up and makes q_j more positive highlighting another case when the power of the ℓ-test can get closer to the power of the one-sided t-test. This, on the first glance, might seem non-intuitive because the increase in $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$ actually implies that the j^{th} variable gets more correlated with the rest of the variables and hence it should become harder to distinguish its effect. Note that unlike the previous case, in this case the power does not increase and as expected, the larger the quantity $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$ gets, the more the performance of all the three tests deteriorate. However with increase in $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$, the quantity q_j increases in magnitude suggesting an increase in the belief about the validity of the sign guess. Put another way, with increase in $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$, it becomes possible to obtain a more reliable sign-guess as a function of the sufficient statistic, $\boldsymbol{S}^{(j)}$. Thus, though with this increasing correlation the tests loose power, the performances of the ℓ-test and the one-sided t-test gets closer because of the improved sign-guessing ability of the former.

Note that for design matrices, $\boldsymbol{X}$, of dimension $n \times d$ with i.i.d. drawn Gaussian columns (as is the case with most of the simulations in this paper), it indeed holds that with d getting closer to n, the component $\boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j$ increases in magnitude. Thus, in this case we would expect that the power curves of ℓ-test and the one-sided t-test come closer as d gets closer to n (that is, as we move closer to un-identifiability).

Finally, note that it follows as a direct consequence of the discussions in this section and Section 2.2 that the ℓ-test gains no power over the t-test if $\boldsymbol{X}_j$ is orthogonal with the rest of the columns (and in particular, for orthogonal designs). In this case, $\hat{m}_j = 0$, so that we have no estimate of the sign of β_j and hence, the ℓ-distribution is symmetric about 0. With smoothing out of the p-value at 1, we would expect the ℓ-test and the t-test, as well as the respective confidence intervals, to perform similarly.

C Proofs

C.1 Proof of Lemma 2.1

Proof. Because $\boldsymbol{V}$ is a full column-rank matrix, there exists a unique $\boldsymbol{u}$, such that,

$$\boldsymbol{V}\boldsymbol{u} = \frac{\boldsymbol{y} - \hat{\boldsymbol{y}}_j}{\|\boldsymbol{y} - \hat{\boldsymbol{y}}_j\|} = \frac{\boldsymbol{y} - \hat{\boldsymbol{y}}_j}{\hat{\sigma}_j}.$$

Clearly,

$$\|\boldsymbol{V}\boldsymbol{u}\| = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}_j\|}{\|\boldsymbol{y} - \hat{\boldsymbol{y}}_j\|} = 1.$$

Because orthogonal transformations preserve norm, we must have that this unique $\boldsymbol{u}$ satisfy,

$$\|\boldsymbol{u}\| = \|\boldsymbol{V}\boldsymbol{u}\| = 1,$$

which proves (2.1) in the statement of Lemma 2.1. The proof (2.2) follows directly from (Luo et al., 2022, Proposition E.1). □

C.2 Proof of Lemma 2.2

Proof. Note that we can write the OLS coefficient as (see (Fithian et al., 2014, Section 4)),

$$\hat{\beta}_{j,\text{OLS}} = \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{y}}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}. \tag{C.1}$$

We first start by showing that the OLS estimate, $\hat{\beta}_{j,\text{OLS}}$, is a constant multiple of the statistic, u_1, where the constant is a deterministic function of the sufficient statistic, $\boldsymbol{S}^{(j)}$. Based on Lemma 2.1, we have the decomposition,

$$\boldsymbol{y} = \hat{\boldsymbol{y}}_j + \underbrace{\hat{\sigma}_j \boldsymbol{V}\boldsymbol{u}}_{\hat{\boldsymbol{e}}_j}.$$

This implies,

$$\begin{aligned}\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{y} &= \boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{P}_{-j}\boldsymbol{y} + \hat{\sigma}_j \boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{V}\boldsymbol{u} \\ &= \hat{\sigma}_j \boldsymbol{X}_j^T \boldsymbol{V}\boldsymbol{u}. \text{ [since the columns of } \boldsymbol{V} \text{ are orthogonal to the columns of } \boldsymbol{X}_{-j}]\end{aligned} \tag{C.2}$$

Note that from the choice of $\boldsymbol{V}$, we have that $\boldsymbol{V}_1 = \frac{(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}$, and hence one can find other orthogonal vectors $\boldsymbol{v}_i, i \in \{2, \ldots, n-d+1\}$, such that, $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-d+1}]$. Then it follows that $\boldsymbol{X}_j^T \boldsymbol{v}_i = 0, i > 1$ and hence for $\boldsymbol{u} = (u_1, \ldots, u_{n-d+1})^T$,

$$\boldsymbol{X}_j^T \boldsymbol{V}\boldsymbol{u} = \boldsymbol{X}_j^T \sum_{i=1}^{n-d+1} \boldsymbol{v}_i u_i = \boldsymbol{X}_j^T \boldsymbol{V}_1 u_1 = \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} u_1 = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\| u_1, \tag{C.3}$$

thereby implying that,

$$\hat{\beta}_{j,\text{OLS}} = \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\hat{\boldsymbol{e}}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}. = \frac{\hat{\sigma}_j u_1}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} \tag{C.4}$$

Note that the factor pre-multiplying u_1 in the above equation is a function of the sufficient statistic, $\boldsymbol{S}^{(j)}$, and the design matrix, $\boldsymbol{X}$. Thus for samples from $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$, under H_j, showing that a statistic is increasing in u_1 or $\hat{\beta}_{j,\text{OLS}}$ are equivalent.

Next, we show that the t-test statistic is an increasing function of $\hat{\beta}_{j,\text{OLS}}$. For that, we first decompose the error term $\hat{\boldsymbol{e}}_j$ as $\hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_\parallel + \hat{\boldsymbol{e}}_\perp$, where,

$$\begin{aligned}\hat{\boldsymbol{e}}_\parallel &= \frac{(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\hat{\boldsymbol{e}}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} = (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\hat{\beta}_{j,\text{OLS}}, \text{ and} \\ \hat{\boldsymbol{e}}_\perp &= \left(\boldsymbol{I} - \frac{(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}\right)\hat{\boldsymbol{e}}_j.\end{aligned} \tag{C.5}$$

Note that $\hat{\boldsymbol{e}}_\parallel$ is the component of $\hat{\boldsymbol{e}}_j$ along the component of $\boldsymbol{X}_j$ orthogonal to $\boldsymbol{X}_{-j}$, while the component $\hat{\boldsymbol{e}}_\perp$ is perpendicular to the columnspace of the entire $\boldsymbol{X}$. Furthermore, the

components, $\hat{\boldsymbol{e}}_\perp$ and $\hat{\boldsymbol{e}}_\parallel$ are themselves orthogonal to each other. Using the relation $\hat{\boldsymbol{e}}_j = \hat{\sigma}_j \boldsymbol{V}\boldsymbol{u}$ and the fact that the matrix pre-multiplying $\hat{\boldsymbol{e}}_j$ to obtain $\hat{\boldsymbol{e}}_\perp$ is idempotent, one can write,

$$\begin{aligned}
\|\hat{\boldsymbol{e}}_\perp\|^2 &= \hat{\sigma}_j^2 \boldsymbol{u}^T \boldsymbol{V}^T \left(\boldsymbol{I} - \frac{(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j \boldsymbol{X}_j^T (\boldsymbol{I}-\boldsymbol{P}_{-j})}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} \right) \boldsymbol{V}\boldsymbol{u} \\
&= \hat{\sigma}_j^2 \boldsymbol{u}^T \boldsymbol{V}^T \boldsymbol{V}\boldsymbol{u} - \frac{\left(\boldsymbol{X}_j^T \hat{\boldsymbol{e}}_j\right)^2}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} \\
&= \hat{\sigma}_j^2 - \frac{\left(\boldsymbol{X}_j^T \hat{\boldsymbol{e}}_j\right)^2}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} \quad \left[\because \boldsymbol{V} \text{ is an orthogonal matrix and } \boldsymbol{u}^T\boldsymbol{u} = 1\right] \\
&= \hat{\sigma}_j^2 - \left(\hat{\beta}_{j,\mathrm{OLS}}\right)^2 \|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2
\end{aligned}$$

Because $\hat{\boldsymbol{e}}_\perp$ is in the orthogonal complement of the columnspace of $\boldsymbol{X}$, we can write,

$$(\boldsymbol{I}-\boldsymbol{P})\hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_\perp + (\boldsymbol{I}-\boldsymbol{P})\hat{\boldsymbol{e}}_\parallel.$$

Also,

$$\begin{aligned}
\|(\boldsymbol{I}-\boldsymbol{P})\hat{\boldsymbol{e}}_j\|^2 &= \|\hat{\boldsymbol{e}}_\perp\|^2 + \|(\boldsymbol{I}-\boldsymbol{P})\hat{\boldsymbol{e}}_\parallel\|^2 \\
&= \hat{\sigma}_j^2 - \left(\hat{\beta}_{j,\mathrm{OLS}}\right)^2 \|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 + \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 (\hat{\beta}_{j,\mathrm{OLS}})^2 \\
&= \hat{\sigma}_j^2 - (\hat{\beta}_{j,\mathrm{OLS}})^2 \left(\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 - \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 \right) \\
&= \hat{\sigma}_j^2 - (\hat{\beta}_{j,\mathrm{OLS}})^2 \underbrace{\|\boldsymbol{P}(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}_{=:\kappa}.
\end{aligned}$$

With these expressions, we can write,

$$T_j = \frac{\hat{\beta}_{j,OLS}}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}}} = C \frac{\hat{\beta}_{j,\mathrm{OLS}}}{\sqrt{\hat{\sigma}_j^2 - \left(\hat{\beta}_{j,\mathrm{OLS}}\right)^2 \kappa}} = C \frac{\frac{\hat{\sigma}_j u_1}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}}{\sqrt{\hat{\sigma}_j^2 - u_1^2 \left(\frac{\hat{\sigma}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}\right)^2 \kappa}},$$

where, $C = \sqrt{\frac{n-d}{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}}}$ and κ are both positive. Defining, $C' = C \cdot \frac{\hat{\sigma}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}$, $\kappa' = \kappa \cdot \left(\frac{\hat{\sigma}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}\right)^2$ and $g_{\boldsymbol{S}^{(j)}}(u) = \frac{C'u}{\sqrt{\hat{\sigma}_j^2 - u^2\kappa'}}$, the above equation shows $T_j = g_{\boldsymbol{S}^{(j)}}(u_1)$, thereby establishing that, $g_{\boldsymbol{S}^{(j)}}$ is a functional of $\boldsymbol{S}^{(j)}$ and itself is continuous, strictly increasing and anti-symmetric. □

C.3 Proof of Theorem 2.1

We will first prove that $f_{\boldsymbol{S}^{(j)}}$ is continuous and strictly increasing in $\{u : f_{\boldsymbol{S}^{(j)}}(u) \neq 0\}$ (in C.3.1), followed by a proof of the characterization of its inverse (in C.3.2).

C.3.1 Proof of continuity and increasing properties of $f_{\mathbf{S}_j}$

In this section, we will prove that $f_{\boldsymbol{S}^{(j)}}$ is continuous and strictly increasing in $\{u : f_{\boldsymbol{S}^{(j)}}(u) \neq 0\}$.

Proof. We again start with the decomposition,

$$\boldsymbol{y} = \hat{\boldsymbol{y}}_j + \underbrace{\hat{\sigma}_j \boldsymbol{V}\boldsymbol{u}}_{\hat{\boldsymbol{e}}_j},$$

and as we did in the proof in Section C.2, we again decompose,

$$\hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_{\|} + \hat{\boldsymbol{e}}_{\perp}.$$

Next using Equation (C.1) and (C.4),

$$u_1 = \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{y}}{\hat{\sigma}_j \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} = \frac{\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\hat{\boldsymbol{e}}_j}{\hat{\sigma}_j \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|}.$$

Also from Equation (C.5),

$$\hat{\boldsymbol{e}}_{\|} = \frac{(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\boldsymbol{X}_j^T(\boldsymbol{I} - \boldsymbol{P}_{-j})\hat{\boldsymbol{e}}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} = \frac{\hat{\sigma}_j(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j u_1}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|}.$$

Based on these relations, we have,

$$\begin{aligned}
&\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \\
&= \|\hat{\boldsymbol{y}}_j + \hat{\boldsymbol{e}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{X}_j\beta_j\|^2 + 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \\
&= \|\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j + \hat{\boldsymbol{e}}_j - (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j\|^2 + 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \\
&= \|\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j + \hat{\boldsymbol{e}}_{\perp} + \hat{\boldsymbol{e}}_{\|} - (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j\|^2 \\
&+ 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \ [\hat{\boldsymbol{e}}_{\perp}, \hat{\boldsymbol{e}}_{\|} \text{ are defined above}] \\
&= \|\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j\|^2 + \|\hat{\boldsymbol{e}}_{\perp}\|^2 + \|\hat{\boldsymbol{e}}_{\|} - (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j\|^2 + 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \\
&= \|\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j\|^2 + \|\hat{\boldsymbol{e}}_{\perp}\|^2 + \left(\frac{u_1\hat{\sigma}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} - \beta_j\right)^2 \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 \\
&+ 2n\lambda|\beta_j| + 2n\lambda\sum_{i\neq j}|\beta_i| \\
&= \tilde{f}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X}) + \left(\frac{u_1\hat{\sigma}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} - \beta_j\right)^2 \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2,
\end{aligned} \tag{C.6}$$

$\tilde{f}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})$ denotes the expression it is replacing, and whenever the context is clear, we will use $\tilde{f}$ to denote $\tilde{f}(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})$. Note that $\tilde{f}$ also depends on $\lambda > 0$, but for compactness, we have suppressed this in the notation as in the following lines we will not be interested in the behavior of $\tilde{f}$ as a function of λ. In fact, we will only analyze $\tilde{f}$ as a function of the argument $\boldsymbol{\beta}$. Thus we have,

$$\begin{aligned}
&\arg\min_{\boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2n\lambda|\beta_j| + 2n\lambda \sum_{i \neq j} |\beta_i| \right) \\
&= \arg\min_{\boldsymbol{\beta}} \left(\tilde{f} + \left(\frac{\hat{\sigma}_j u_1}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|} - \beta_j \right)^2 \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2 \right).
\end{aligned}$$

Define

$$\hat{\boldsymbol{\beta}}(a) := \arg\min_{\boldsymbol{\beta}} \left(\underbrace{\tilde{f} + (a - \beta_j)^2 \|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}_{=:U(a, \boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{X})} \right). \tag{C.7}$$

Thus, $\hat{\beta}_j \left(\frac{\hat{\beta}_{j,\text{OLS}}}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2} \right)$ is just the LASSO estimate of β_j, $\hat{\beta}_j^\lambda$ and hence, it suffices to show that $\hat{\beta}_j^\lambda(a)$ is a non-decreasing function of a. As is the case with the function, $\tilde{f}$, we will also be primarily be interested in the behavior of U as a function of the arguments, $(a, \boldsymbol{\beta})$. We start with listing some properties of the function, U. First, note that U is a continuous, convex function in its arguments and is strictly convex in a for any fixed value of $\boldsymbol{\beta}$. This implies that $\hat{\boldsymbol{\beta}}(a)$ is continuous in a. To see this, first note that as $\|\boldsymbol{\beta}\| \to \infty$, the function, U diverges, so that for the minimization problem to obtain $\hat{\boldsymbol{\beta}}(a)$, we can constraint $\boldsymbol{\beta}$ within some compact set, $C(a) \subset \mathbb{R}^d$. One can then apply Berge's Maximum Theorem (Berge, 1963), to conclude that $\hat{\boldsymbol{\beta}}(a)$ is continuous in a.

Next, note that the only non-differentiable component in the expression of U is the ℓ_1-penalty of $\boldsymbol{\beta}$, which implies that the U has a partial derivative in β_j at all non-zero values of β_j. Consider a value of a such that $\hat{\beta}_j(a) \neq 0$ and let $\mathcal{A}(a) = \{i \in [1:k] \setminus \{j\} : \hat{\beta}_i(a) \neq 0\}$ denote the active set among the remaining variables.

Note that because $\hat{\boldsymbol{\beta}}(a)$ maximizes U, and $\forall i \notin \mathcal{A}(a) \cup \{j\}$, $\hat{\boldsymbol{\beta}}_j(a) = 0$, it holds that $\hat{\boldsymbol{\beta}}_{\mathcal{A}(a)\cup\{j\}}(a)$ is a minimizer of $U(a, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{X}_{\mathcal{A}(a)\cup\{j\}})$, where now $\boldsymbol{\gamma}$ is a vector of length $|\mathcal{A}(a) \cup \{j\}|$. For a proof, see Lemma 1 of (Liu et al., 2021+, Section 3.2). Because all the entries of $\hat{\boldsymbol{\beta}}_{\mathcal{A}(a)\cup\{j\}}$ are non-zero, $U(a, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{X}_{\mathcal{A}(a)\cup\{j\}})$ is differentiable in $\boldsymbol{\gamma}$ at $\hat{\boldsymbol{\beta}}_{\mathcal{A}(a)\cup\{j\}}$, and because the latter is a minimizer, an appeal to the first-order stationary conditions yield that for any $i \in \mathcal{A}(a)$,

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_i} U(a, \boldsymbol{\gamma}; \boldsymbol{y}, \boldsymbol{X}_{\mathcal{A}(a)\cup\{j\}})\big|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)\cup\{j\}}}(a) = 0 \\
\implies &- \boldsymbol{X}_i^T (\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{\mathcal{A}(a)} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)}(a) - \boldsymbol{P}_{-j}\boldsymbol{X}_j \hat{\beta}_j^\lambda(a)) + 2n\lambda \text{sign}(\hat{\beta}_i(a)) = 0
\end{aligned}$$

Note that from the continuity of $\hat{\boldsymbol{\beta}}(a)$ in the neighbourhood of a where $\mathcal{A}(a)$ does not change, the sign of the active variables also remain constant so that $\text{sign}(\hat{\beta}_i(a))$ is a constant in that

neighborhood, $\forall i \in \mathcal{A}(a) \cup \{j\}$. Thus, differentiating the above equation both the sides with respect to a yields,

$$\begin{aligned}
&\boldsymbol{X}_i^T \boldsymbol{X}_{\mathcal{A}(a)} \frac{\partial}{\partial a} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)} + \boldsymbol{X}_i^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a) = 0, \forall i \in \mathcal{A}(a) \\
\implies &\boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{X}_{\mathcal{A}(a)} \frac{\partial}{\partial a} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)} + \boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a) = \boldsymbol{0} \\
\implies &\frac{\partial}{\partial a} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)} = -\left(\boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{X}_{\mathcal{A}(a)}\right)^{-1} \boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a)
\end{aligned}$$

Similarly using the first-order stationary condition on the index j yields,

$$\begin{aligned}
&-\boldsymbol{X}_j^T \boldsymbol{P}_{-j}(\hat{\boldsymbol{y}}_j - \boldsymbol{X}_{\mathcal{A}(a)} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)}(a) - \boldsymbol{P}_{-j} \boldsymbol{X}_j \hat{\beta}_j^\lambda(a)) - (a - \hat{\beta}_j^\lambda(a)) \|(\boldsymbol{I} - \boldsymbol{P}_{-j}) \boldsymbol{X}_j\|^2 = 0 \\
\implies &\boldsymbol{X}_j^T \boldsymbol{X}_{\mathcal{A}(a)} \frac{\partial}{\partial a} \hat{\boldsymbol{\beta}}_{\mathcal{A}(a)}(a) + \boldsymbol{X}_j^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a) - \left(1 - \frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a)\right) \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2 = 0 \\
\implies &-\boldsymbol{X}_j^T \boldsymbol{X}_{\mathcal{A}(a)} \left(\boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{X}_{\mathcal{A}(a)}\right)^{-1} \boldsymbol{X}_{\mathcal{A}(a)}^T \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a) + \boldsymbol{X}_j^T \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a) = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2 \\
\implies &-\boldsymbol{X}_j^T \boldsymbol{P}_{\mathcal{A}(a)} \boldsymbol{P}_{-j} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a) + \|\boldsymbol{X}_j\|^2 \frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a) = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2 \\
\implies &-\boldsymbol{X}_j^T \boldsymbol{P}_{\mathcal{A}(a)} \boldsymbol{X}_j \frac{\partial}{\partial a} \hat{\beta}_j(a) + \|\boldsymbol{X}_j\|^2 \frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a) = \|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2 \\
\implies &\frac{\partial}{\partial a} \hat{\beta}_j^\lambda(a) = \frac{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2}{\|\boldsymbol{X}_j\|^2 - \|\boldsymbol{P}_{\mathcal{A}(a)} \boldsymbol{X}_j\|^2} = \frac{\|(\boldsymbol{I} - \boldsymbol{P}_{-j})(\boldsymbol{X}_j)\|^2}{\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{A}(a)})(\boldsymbol{X}_j)\|^2},
\end{aligned}$$

which is positive, whenever $\hat{\beta}_j^\lambda(a) \neq 0$. Hence, for an $a \in \mathbb{R}$, either $\hat{\beta}_j^\lambda(a) = 0$ or $\frac{\partial \hat{\beta}_j(a)}{\partial a} > 0$, showing that $\hat{\beta}_j^\lambda$ is locally increasing around a in the latter case. Now define,

$$f_{\boldsymbol{S}^{(j)}}(u) = \hat{\beta}_j^\lambda \left(u \cdot \frac{\hat{\sigma}_j}{\|(\boldsymbol{I} - \boldsymbol{P}_{-j}) \boldsymbol{X}_j\|}\right),$$

and note that, $\hat{\beta}_j^\lambda = f_{\boldsymbol{S}^{(j)}}(u_1)$. Furthermore, note that $f_{\boldsymbol{S}^{(j)}}$ is a functional of the sufficient statistic, $\boldsymbol{S}^{(j)}$, and using the arguments above, is a continuous, piece-wise linear function that is increasing in the region, $\{u : f_{\boldsymbol{S}^{(j)}}(u) \neq 0\}$. This establishes all the claims about $f_{\boldsymbol{S}^{(j)}}$ in the statement of Theorem 2.1 except for (2.4), which we turn to next.

□

C.3.2 Proof of Equation (2.4) in Theorem 2.1

To prove Equation (2.4), we first start with an intermediate result.

Theorem C.1. *For data $(\boldsymbol{y}, \boldsymbol{X})$, with $\boldsymbol{X}$ full column-rank and $\lambda > 0$, define $\hat{\boldsymbol{\beta}}_*^\lambda(b)$ to be b at the j^{th} coordinate and $\hat{\boldsymbol{\beta}}_{-j}^\lambda(b)$ on the rest, where, $\hat{\boldsymbol{\beta}}_{-j}^\lambda(b)$ is defined as in Equation (2.3). Also let*

$$f_\lambda(\boldsymbol{y}; \boldsymbol{\beta}) := \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{C.8}$$

be the LASSO objective function. Then for any $b \in \mathbb{R}$, the following are equivalent

(a) $0 \in \partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^\lambda_*(b)}$

(b) $\hat{\boldsymbol{\beta}}^\lambda = \hat{\boldsymbol{\beta}}^\lambda_*(b)$

(c) $\hat{\beta}^\lambda_j = b$

Proof of Theorem C.1. We first show the equivalence of (a) and (b). Assume that, $0 \in \partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^\lambda_*(b)}$. The convexity of $f_\lambda(\boldsymbol{y};(\boldsymbol{\beta}_{-j} = \hat{\boldsymbol{\beta}}^\lambda_{-j}(b), \beta_j))$ in β_j shows that the j^{th} entry of $\hat{\boldsymbol{\beta}}^\lambda_*(b)$ is the minimizer of $f_\lambda(\boldsymbol{y};(\boldsymbol{\beta}_{-j} = \hat{\boldsymbol{\beta}}^\lambda_{-j}(b), \beta_j))$ in β_j. That is,

$$\arg\min_{\beta_j} f_\lambda(\boldsymbol{y};(\boldsymbol{\beta}_{-j} = \hat{\boldsymbol{\beta}}^\lambda_{-j}(b), \beta_j)) = b.$$

But by the definition of $\hat{\boldsymbol{\beta}}^\lambda_{-j}(b)$, we know that

$$\arg\min_{\boldsymbol{\beta}_{-j}} f_\lambda(\boldsymbol{y};(\boldsymbol{\beta}_{-j}, \beta_j = b)) = \hat{\boldsymbol{\beta}}^\lambda_{-j}(b).$$

Thus, if one runs a blockwise coordinate descent with blocks $\{j\}$ and $[1:p]\setminus\{j\}$ starting at $\hat{\boldsymbol{\beta}}^\lambda_*(b)$, we see that the iterates will be constant at $\hat{\boldsymbol{\beta}}^\lambda_*(b)$, thereby implying that this is a limit point of the iterates. One can now invoke (Tseng, 2001, Proposition 5.1) (the conditions for applying this proposition follow directly as f_λ can be separated into the squared error loss and the non-differentiable ℓ_1-penalty) to conclude that,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} f_\lambda(\boldsymbol{y};\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}^\lambda_*(b),$$

which establishes the implication of (a) to (b). The reverse implication follows directly from the fact that $\hat{\boldsymbol{\beta}}$ is the optimizer of the LASSO objective, so that each coordinate, and in particular the j^{th} coordinate of the sub-gradient, contains 0, that is, $0 \in \partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. The fact that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^\lambda_*(b)$ completes the argument. It is also straightforward to see that (b) implies (c). To show that (c) implies (b), note that one can use the blockwise coordinate descent argument used above to conclude that $\hat{\boldsymbol{\beta}}^\lambda = \hat{\boldsymbol{\beta}}^\lambda(\hat{\beta}^\lambda_j)$, and then use the hypothesis of (c) (that is, $\hat{\beta}^\lambda_j = b$) to conclude (b). □

Hence, Theorem C.1 now implies that $\hat{\beta}^\lambda_j = b$ if and only if $0 \in \partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^\lambda_*(b)}$. We will now prove item 2 of Theorem 2.1 by evaluating this sub-gradient.

Proof of item 2 of Theorem 2.1. Note that we have the following decomposition,

$$\begin{aligned}
&f_\lambda(\boldsymbol{y};\boldsymbol{\beta}) \\
&= \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \frac{1}{2n}\|\hat{\boldsymbol{y}}_j + \hat{\sigma}_j \boldsymbol{V}\boldsymbol{u} - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j - (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \frac{1}{2n}\|\hat{\boldsymbol{y}}_j - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\|^2 + \frac{1}{2n}\|\hat{\sigma}_j\boldsymbol{V}\boldsymbol{u} - (\boldsymbol{I} - \boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j\| + \lambda|\beta_j| + \lambda\sum_{i\neq j}|\beta_i|
\end{aligned}$$

Now define,

$$s(\beta_j) = \begin{cases} [-1,1], & \beta_j = 0 \\ \text{sign}(\beta_j), & \beta_j \neq 0 \end{cases}.$$

Then we have,

$$\begin{aligned}
&\partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta}) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{P}_{-j}(\hat{\boldsymbol{y}}_j - \boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}) - \frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{P}_{-j})\left(\hat{\sigma}_j\boldsymbol{V}\boldsymbol{u} - (\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j\right) + \lambda s(\beta_j) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T\hat{\boldsymbol{y}}_j - \frac{\hat{\sigma}_j}{n}\boldsymbol{X}_j^T\boldsymbol{V}\boldsymbol{u} + \frac{\boldsymbol{X}_j^T\boldsymbol{P}_{-j}\boldsymbol{X}_j\beta_j + \boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\beta_j + \boldsymbol{X}_j^T\boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}}{n} + \lambda s(\beta_j) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T\hat{\boldsymbol{y}}_j - \frac{\hat{\sigma}_j}{n}\boldsymbol{X}_j^T\boldsymbol{V}\boldsymbol{u} + \frac{\boldsymbol{X}_j^T\boldsymbol{X}_j\beta_j + \boldsymbol{X}_j^T\boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}}{n} + \lambda s(\beta_j) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T\hat{\boldsymbol{y}}_j - \frac{\hat{\sigma}_j}{n}\boldsymbol{X}_j^T\boldsymbol{V}\boldsymbol{u} + \frac{\boldsymbol{X}_j^T\boldsymbol{X}\boldsymbol{\beta}}{n} + \lambda s(\beta_j) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - \boldsymbol{X}\boldsymbol{\beta}) - \frac{\hat{\sigma}_j}{n}\boldsymbol{X}_j^T\boldsymbol{V}\boldsymbol{u} + \lambda s(\beta_j).
\end{aligned}$$

From the calculations in Section C.2,

$$\boldsymbol{X}_j^T\boldsymbol{V}\boldsymbol{u} = \boldsymbol{X}_j^T\sum_{i=1}^{n-d+1}\boldsymbol{v}_i u_i = \boldsymbol{X}_j^T\boldsymbol{v}_j u_1 = \frac{\boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}u_1 = \|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|u_1,$$

and thus we have,

$$\begin{aligned}
&\partial_{\beta_j} f_\lambda(\boldsymbol{y};\boldsymbol{\beta}) \\
&= -\frac{1}{n}\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - \boldsymbol{X}\boldsymbol{\beta}) - \frac{\hat{\sigma}_j\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}{n}u_1 + \lambda s(\beta_j).
\end{aligned}$$

Setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^*(b)$ for a b, the above equation along with Theorem C.1 establishes that for $b \neq 0$, $\hat{\beta}_j^\lambda = b$ if and only if,

$$u_1 = \frac{-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}(b)) + n\lambda\text{sign}(b)}{\hat{\sigma}_j\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|} = \Lambda_j(b, \text{sign}(b)),$$

while for $b \neq 0$, $\hat{\beta}_j^\lambda = 0$ if and only if,

$$u_1 \in \left[\frac{-\boldsymbol{X}_j^T(\hat{\boldsymbol{y}}_j - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}(b)) \pm n\lambda}{\hat{\sigma}_j\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|}\right] = [\Lambda_j(0,-1), \Lambda_j(0,1)].$$

Noting that $\hat{\beta}_j^\lambda = f_{\boldsymbol{S}^{(j)}}(u_1)$ completes the proof of item 2. □

C.4 Proof of the fact that T_j is independent of the sufficient statistic, $\mathbf{S}_j$ under $H_j : \beta_j = 0$

Proof. In Section C.2, we showed that,

$$\begin{aligned} T_j &\propto \frac{\hat{\beta}_{j,\text{OLS}}}{\sqrt{\hat{\sigma}_j^2 - \hat{\beta}_{j,\text{OLS}}^2}} \\ &= \frac{\frac{\hat{\beta}_{j,\text{OLS}}}{\hat{\sigma}_j}}{\sqrt{1 - \left(\frac{\hat{\beta}_{j,\text{OLS}}}{\hat{\sigma}_j}\right)^2}}, \end{aligned}$$

where the proportionality constant consists of terms that entirely depend on the design matrix. Thus, the only stochastic component in the expression for T_j is $\frac{\hat{\beta}_{j,\text{OLS}}}{\hat{\sigma}_j} = \frac{\boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{y}}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2\hat{\sigma}_j} =: \frac{\boldsymbol{X}_j^T}{\|(\boldsymbol{I}-\boldsymbol{P}_{-j})\boldsymbol{X}_j\|^2}\boldsymbol{L}_j$, where $\boldsymbol{L}_j$ equals the term its replacing. Thus, it suffices to show that under H_j, the unconditional distribution of $\boldsymbol{L}_j$ is the same as its conditional distribution, $\boldsymbol{L}_j \mid \boldsymbol{S}^{(j)}$.

Note that from Equation (2.1), we have that under H_j

$$\boldsymbol{L}_j \mid \boldsymbol{S}^{(j)} \sim \boldsymbol{V}\boldsymbol{u},$$

where, $\boldsymbol{u}$ is uniformly distributed over $\mathbb{S}^{n-d}$. Now, let us evaluate the unconditional distribution. Under H_j, we can write,

$$\boldsymbol{y} \sim \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j} + \boldsymbol{\epsilon},$$

for some, $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$. Then, because $\boldsymbol{V}$ denotes a matrix with columns forming an orthonormal basis for the complement of the columnspace of $\boldsymbol{X}_{-j}$, we have, $\boldsymbol{I}-\boldsymbol{P}_{-j} = \boldsymbol{V}\boldsymbol{V}^T$. Thus, we have under H_j,

$$\begin{aligned} \boldsymbol{L}_j &= \frac{\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{y}}{\|\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{y}\|} \\ &= \frac{\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{\epsilon}}{\|\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{\epsilon}\|} \\ &= \boldsymbol{V}\frac{\boldsymbol{V}^T\boldsymbol{\epsilon}}{\|\boldsymbol{V}^T\boldsymbol{\epsilon}\|}. \text{ [since orthogonal transformations do not change the norm].} \end{aligned}$$

Now, since $\boldsymbol{V}$ is orthogonal, we have, $\boldsymbol{v} = \boldsymbol{V}^T\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{rank(\boldsymbol{V})}) = \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n-d+1})$. Thus we have under H_j,

$$\boldsymbol{L}_j = \boldsymbol{V}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \boldsymbol{V}\boldsymbol{u}^*,$$

where, $\boldsymbol{u}^* = \boldsymbol{v}/\|\boldsymbol{v}\|$ is uniformly distributed over $\mathbb{S}^{n-d}$. This completes the proof. □

C.5 Proof of the map of U_j to the t-distribution

In this section, we will prove that $\boldsymbol{U} \sim \text{Unif}(\mathbb{S}^m)$ implies that

$$\frac{\sqrt{m} \cdot U_j}{\sqrt{1 - U_j^2}} \sim t_m.$$

The proof follows from the following representation of $\boldsymbol{U}$: Let $\boldsymbol{X} \sim N_{m+1}(\boldsymbol{0}, \boldsymbol{I}_{m+1})$, then we have that,

$$U_j \overset{d}{=} \frac{X_j}{\sqrt{X_j^2 + \sum_{i \neq j} X_i^2}} \implies \sqrt{m} \cdot \frac{U}{\sqrt{1 - U^2}} \overset{d}{=} \frac{\sqrt{m} X_j}{\sqrt{\sum_{i \neq j} X_i^2}}.$$

The proof follows from the fact that $X_j \sim N(0, 1)$ and $\sum_{i \neq j} X_i^2 \sim \chi^2_m$, independent of X_j.

D Characterization of $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$ as a function of b

Algorithm 1: Piecewise linear characterization of $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ (as defined in (D.1))

1 **Input:** Data: $(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{X})$, Regularization parameter: λ.
2 **Output:** Piecewise linear characterization of $\hat{\boldsymbol{\beta}}^{\lambda}(b)$.
3 Find a point a point x_0 such that $\hat{\boldsymbol{\beta}}^{\lambda}$ is differentiable at x_0 and set $b = x_0$
4 Calculate $\boldsymbol{\gamma}(x_0)$ and set $\boldsymbol{\gamma} = \boldsymbol{\gamma}(x_0)$.
5 **while** *Not stopped* **do**
6 Find $d_1 := \min\left\{d > 0 : \left|\boldsymbol{Z}_i^T\left(\boldsymbol{y} - \boldsymbol{v}b - \boldsymbol{Z}(\hat{\boldsymbol{\beta}}^{\lambda}(b) + d\boldsymbol{\gamma})\right)\right| = n\lambda\right\}$.
7 Find $d_2 := \min\left\{d > 0 : \text{ at least one coordinate of } \hat{\boldsymbol{\beta}}^{\lambda}_{S(b)} + d\boldsymbol{\gamma}_{S(b)} \text{ is } 0\right\}$.
8 **if** *the minima could not be found in the above two steps* **then**
9 Label b as a terminal knot
10 Record $\boldsymbol{\gamma}(b)$ and tag it with b.
11 Break the While loop.
12 **end**
13 Set $d \leftarrow \min\{d_1, d_2\}$.
14 Set $\hat{\boldsymbol{\beta}}^{\lambda}(b + d) = \hat{\boldsymbol{\beta}}^{\lambda}(b) + d\boldsymbol{\gamma}$.
15 Set $b = b + d$
16 Set $\boldsymbol{\gamma} = \boldsymbol{\gamma}(b)$ Label b a non-terminal knot.
17 Record $(b, \hat{\boldsymbol{\beta}}^{\lambda}(b))$.
18 **end**
19 Repeat from Step 4, but with $\boldsymbol{\gamma} = -\boldsymbol{\gamma}(x_0)$ instead.
20 Generate piecewise linear paths by linearly interpolating $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ between two non-terminal knots or a terminal and non-terminal knot. For a terminal knot, b_t, towards the side where there is no other knot, generate the linear path with slope $\boldsymbol{\gamma}(b_t)$ originating at $(b_t, \hat{\boldsymbol{\beta}}^{\lambda}(b_t))$.

In Sections 3 and 4, we saw that for obtaining the ℓ-test confidence intervals for β_j, one needs to evaluate the function $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$ for different values of b. Recall from (2.3) that $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$ is defined as

$$\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b) := \underset{\boldsymbol{\beta}_{-j}\in\mathbb{R}^{d-1}}{\arg\min} \left(\frac{1}{2n}\|\boldsymbol{y} - b\boldsymbol{X}_j - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}\|^2 + \lambda\|\boldsymbol{\beta}_{-j}\|_1 \right).$$

Certainly, evaluating $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$ for different values of b is computationally expensive as each evaluation requires a LASSO run. In this section, we show that this computation burden can be relieved significantly by providing a characterization of $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$.

We will ease notations and for this section we will assume that we have data of the form $(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{Z})$ and define,

$$\hat{\boldsymbol{\beta}}^{\lambda}(b) := \underset{\boldsymbol{\beta}}{\arg\min} \left(\frac{1}{2n}\|\boldsymbol{y} - b\boldsymbol{v} - \boldsymbol{Z}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \right). \tag{D.1}$$

The above notation is valid in this section only and any mention of $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ would always imply the above. We will characterize $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ as a function of $b \in \mathbb{R}$. Note that one can obtain $\hat{\boldsymbol{\beta}}^{\lambda}_{-j}(b)$ (as defined in (2.3)) by substituting $\boldsymbol{Z} = \boldsymbol{X}_{-j}$ and $\boldsymbol{v} = \boldsymbol{X}_j$.

In the following proposition, we first show that $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ is a piecewise linear function of b, in which we take heavy inspiration from Rosset and Zhu (2007) where the authors establish that the optimizers of ℓ_1-penalized, twice-differentiable likelihoods are piecewise linear in the regularization parameter λ and provide algorithm for generating these paths.

Proposition D.1. *Assume that $\boldsymbol{y}, \boldsymbol{v} \in \mathbb{R}^n$ and $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$, for $n > d$. Then, for $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ defined as in (D.1), we have that $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ is a continuous, piecewise linear function of b.*

Proof sketch. First note that due to Berge's Maximum Theorem (Berge, 1963), $b \mapsto \hat{\boldsymbol{\beta}}^{\lambda}(b)$ is a continuous transformation. Now for any $b \in \mathbb{R}$, define,

$$S(b) = \left\{ j \in [1:d] : \left(\hat{\boldsymbol{\beta}}^{\lambda}(b)\right)_j \neq 0 \right\}. \tag{D.2}$$

Note that $S(b)$ is constant in a neighborhood $b \in \mathcal{N}$, if and only if $\hat{\boldsymbol{\beta}}^{\lambda}_{S(b)}(b)$ is differentiable in b. Pick such a b, then it follows that

$$\frac{\partial}{\partial b}\hat{\boldsymbol{\beta}}^{\lambda}_{S(b)}(b) = -\left(\boldsymbol{Z}^T_{S(b)}\boldsymbol{Z}_{S(b)}\right)^{-1}\boldsymbol{Z}^T_{S(b)}\boldsymbol{v} \tag{D.3}$$

Thus this shows that $\hat{\boldsymbol{\beta}}^{\lambda}_{S(b)}(b)$, and hence, $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ is linear in that neighborhood. And hence, $\hat{\boldsymbol{\beta}}^{\lambda}(b)$ is piecewise linear for $b \in \mathbb{R}$. □

We now turn our attention to methods of exactly generating these piecewise linear paths, following an approach analogous to Rosset and Zhu (2007). To begin with, denote,

$$L(\boldsymbol{\beta}) := \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{v}b - \boldsymbol{Z}\boldsymbol{\beta}\|^2$$

Then, note that we can decompose $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$ to write our LASSO minimization problem as,

$$\begin{aligned} &\text{Minimize: } L(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + \lambda \sum (\beta_i^+ - \beta_i^-) \\ &\text{Subject to: } \beta_i^+, \beta_i^- \geq 0, \forall i. \end{aligned}$$

We introduce Lagarange multipliers for each of these $2p$ elements to get the Lagarangian

$$\mathcal{L} = L(\boldsymbol{\beta}^+ - \boldsymbol{\beta}^-) + \lambda \sum (\beta_i^+ + \beta_i^-) - \sum \lambda_i^+ \beta_i^+ - \sum \lambda_i^- \beta_i^-.$$

KKT conditions on the primal show that the following relations are suggested at the optimum for all $1 \leq i \leq d$:

$$\begin{aligned} (\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}))_i + \lambda - \lambda_i^+ &= 0 \\ -(\partial_{\boldsymbol{\beta}} L(\boldsymbol{\beta}))_i + \lambda - \lambda_i^- &= 0 \\ \lambda_i^+ \beta_i^+ &= 0 \\ \lambda_i^- \beta_i^- &= 0 \end{aligned}$$

The above set of equations suggest that,

$$\hat{\beta}_i^\lambda(b) \neq 0 \Leftrightarrow \left| \left(\partial_{\boldsymbol{\beta}} L \big|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^\lambda(b)} \right)_i \right| = \lambda \Leftrightarrow \left| \boldsymbol{Z}_i^T \left(\boldsymbol{y} - \boldsymbol{v} b - \boldsymbol{Z} \hat{\boldsymbol{\beta}}^\lambda(b) \right) \right| = n\lambda. \tag{D.4}$$

We use Equation (D.3), (D.4), along with Proposition D.1 and the fact that the set $S(b)$ (from (D.2)) changes only at non-differentiability points of $\hat{\boldsymbol{\beta}}(b)$ to devise an algorithm to exactly generate the paths of $\hat{\boldsymbol{\beta}}(b)$ as a function of b in Algorithm 1. The algorithm uses a notation defined below for the 'derivative' of $\hat{\boldsymbol{\beta}}^\lambda(b)$:

$$\begin{aligned} \boldsymbol{\gamma}(b) := &\begin{cases} \boldsymbol{\gamma}_{S(b)}(b) = -\left(\boldsymbol{Z}_{S(b)}^T \boldsymbol{Z}_{S(b)} \right)^{-1} \boldsymbol{Z}_{S(b)}^T \boldsymbol{v} \\ \boldsymbol{\gamma}_{S(b)^c}(b) = \boldsymbol{0} \end{cases}, \text{ if } S(b) \neq \phi \\ :=& \boldsymbol{0}, \text{ otherwise.} \end{aligned}$$

E Choice of the tuning parameter, λ

E.1 The min rule vs. The 1se rule

In Section 2.4, we justified that a reasonable way of choosing λ for testing H_j so that it does not invalidate our theory surrounding the ℓ-test can be to cross validate on $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$, where $\tilde{\boldsymbol{y}}$ is drawn from the conditional distribution of $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$, under H_j. For computational convenience, we have used cross-validation on $(\boldsymbol{X}_{-j}, \tilde{\boldsymbol{y}})$ instead of $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$, as the two choices result in almost identical performances of the resulting methods and because the LASSO estimates obtained using $(\boldsymbol{X}_{-j}, \boldsymbol{y})$ is the same as that using $(\boldsymbol{X}_{-j}, \tilde{\boldsymbol{y}})$, this enables us to recycle some common information from the latter dataset, thereby saving computation time while obtaining the conditional distribution of $\hat{\beta}_j \mid \boldsymbol{S}^{(j)}$, under H_j.

Throughout this paper, we recommend using the min rule for choosing λ using cross-validation—that is, choosing the λ that results in the smallest cross-validated error. However another popular choice with cross-validation can be to pick the largest λ resulting in a cross-validated error within one standard deviation of the minimum cross-validated error, also known as the 1se rule. The latter rule results in stricter selection, but more severe multiplicity correction post-selection, and hence, it is not entirely clear whether the trade-off that the 1se rule presents can be any better than the min rule.

In Section F.3, we provide the empirical coverage and lengths of the resulting confidence intervals when using the 1se rule for selecting λ. To summarize our findings, we observe that the 1se rule and the min rule have similar performances for constructing ℓ-test confidence intervals except for the case when the $\boldsymbol{\beta}$ is very dense, in which case the min rule results in intervals of shorter length. Hence, we recommend using the min rule for choosing λ.

E.2 The randomness in the choice of λ

The rule for the choice of λ we discussed involves sampling, $\tilde{\boldsymbol{y}} \sim \boldsymbol{y} \mid \boldsymbol{S}^{(j)}$, under H_j and then running cross-validation on $(\boldsymbol{X}_{-j}, \tilde{\boldsymbol{y}})$. This suggests towards some inherent sources of variability in the method—the random sampling of $\tilde{\boldsymbol{y}}$ and the random splitting of the dataset to perform cross-validation. In order to understand its effect, we perform $m = 100$ replications of an experiment under the setting of the left panel of Figure 2 where for each replicate we form a linear model with this design matrix and obtain p-value for testing $H_j : \beta_j = 0$ for $m = 100$ samples of $\tilde{\boldsymbol{y}}$. Let p_{ij} denote the p-value after sampling $\tilde{\boldsymbol{y}}$ for the j^{th} time for the dataset $(\boldsymbol{y}^{(i)}, \boldsymbol{X}^{(i)})$ (corresponding to the i^{th} replication of drawing $(\boldsymbol{y}, \boldsymbol{X})$ from a pre-determined distribution). For these $m^2 = 10^4$ p-values, we plot the empirical estimate of the overall standard deviation of the p-values, $\sqrt{\text{Var}(p_{ij})}$ against the standard deviation conditioned on a replicate, $\sqrt{\mathbb{E}\text{Var}\left(p_{ij} \mid (\boldsymbol{y}^{(i)}, \boldsymbol{X}^{(i)})\right)}$ in Figure 6 in the log-scale. These quantities are estimated using $\sqrt{\frac{1}{m^2-1}\left(\sum_{i,j=1}^{m}(p_{ij} - \bar{p}_{\cdot\cdot})^2\right)}$ and $\sqrt{\frac{1}{m^2-m}\left(\sum_{i,j=1}^{m}(p_{ij} - \bar{p}_{i\cdot})^2\right)}$, respectively, where, $\bar{p}_{i\cdot}$ represents the mean of the entries in $\{p_{ij} : 1 \leq j \leq m\}$, while $\bar{p}_{\cdot\cdot}$ represents the mean of all the p-values, $\{p_{ij} : 1 \leq i, j \leq m\}$.

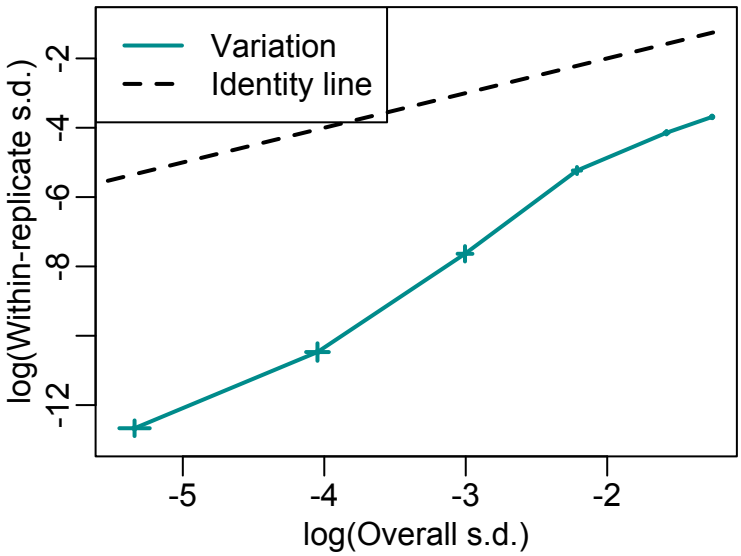

Figure 6: Relative variability in the ℓ-test p-values due to the randomness in $\hat{\lambda}_j$. The error bars represent plus or minus two units of standard errors.

This figure suggests that the variability due to sampling of $\tilde{\boldsymbol{y}}$ is negligible compared to the monte-carlo variability in the p-values due to the replication of the procedure—we see from Figure 6 that sampling of $\tilde{\boldsymbol{y}}$ never accounts for more than about 9% of the overall standard deviation (or about 0.8% of the overall variation). This suggests that even though randomized due to sampling of $\tilde{\boldsymbol{y}}$, the p-values we obtain are stable.

Finally, note that an alternate way of doing cross-validation that does not introduce the randomness in the procedure due to sampling $\tilde{\boldsymbol{y}}$ is by cross-validating on $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$ instead, where $\hat{\boldsymbol{y}}_j$ is the projection of $\boldsymbol{y}$ on the columnspace of $\boldsymbol{X}_{-j}$ (and is the non-zero-mean component of $\tilde{\boldsymbol{y}}$, when sampled from $\boldsymbol{y} \mid \boldsymbol{S}^{(j)}$ under H_j). Note that computation of ℓ-distribution based on either of $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$ or $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$ would exactly be the same. Even though we have established that the sampling of $\tilde{\boldsymbol{y}}$ introduces negligible randomness in the ℓ-test p-value, one might wonder why introduce *any* randomness at all in the first place and not cross-validate on $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$? In Figure 7, we compare three possible datasets we can cross-validate on to choose λ: $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$ (our default choice), $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$ and $(\boldsymbol{y}, \boldsymbol{X})$. Note that, as described in Section 2.4, the last choice is not valid as the chosen λ will not be a function of the sufficient statistic, however as this is cross-validating on the full dataset, we can expect this chosen λ to have the 'optimal performance' and the resulting power curve can be used as a benchmark. Indeed, Figure 7 shows that choosing λ based on $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$ suffers a detriment as compared to our recommended choice (which also performs almost similarly to cross-validating on the full $(\boldsymbol{y}, \boldsymbol{X})$), providing further justification for it.

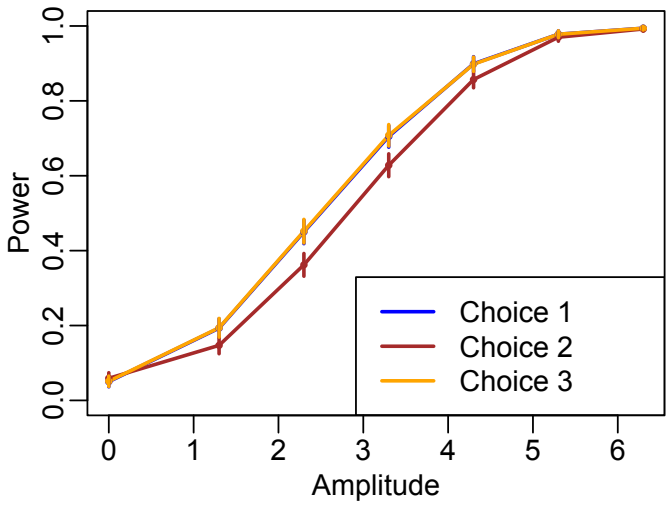

Figure 7: We use exactly the same setting as in the left panel of Figure 2. Choice 1,2 and 3 denote choosing λ by cross-validating on $(\tilde{\boldsymbol{y}}, \boldsymbol{X}_{-j})$, $(\hat{\boldsymbol{y}}_j, \boldsymbol{X}_{-j})$ and $(\boldsymbol{y}, \boldsymbol{X})$, respectively. The curves for Choice 1 and 3 are on top of each other. The error bars represent plus or minus two standard errors.

F Further Experiments

F.1 Performance of the ℓ-test under different settings

To explore further aspects of the performance of the ℓ-test, we test its performance under an additional setting, with the results are summarized in Figure 8. We see that in this case, similar to Figure 2, the power of the ℓ-test increases with increasing amplitude, but almost overlapping with that of the one-sided t-test. Note that, as follows from Appendix B, in this case where our particular choice of the design matrix is closer to un-identifiability, the ℓ-test is more sure about its guess of the sign of the alternate β_j as compared to the $d = 50$ case.

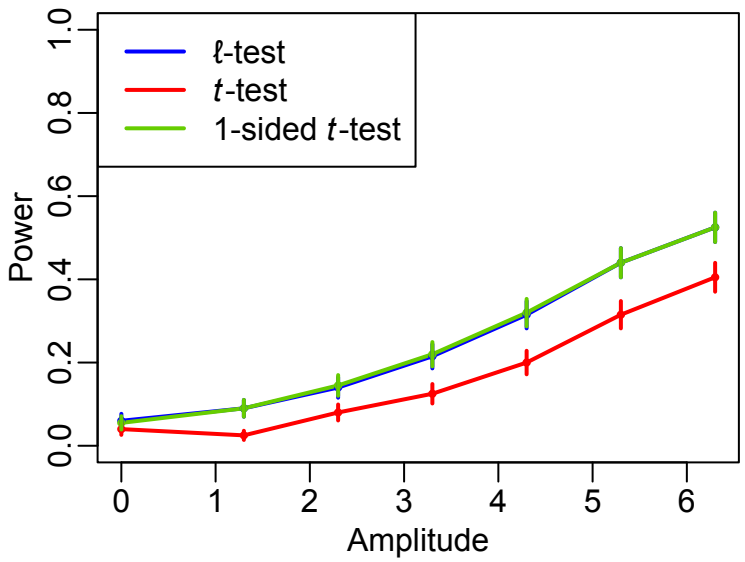

Figure 8: Exactly the same setting as in the left panel of Figure 2 but with $d = 90$. The error bars represent plus or minus two standard errors.

F.2 Robustness of the ℓ-test to violations of the linear model assumptions

In this section, we extend the results in Section 5.2 by performing more extensive experiments to empirically evaluate the robustness of the validity of the ℓ-test to the violations in the assumptions of the Gaussian linear model. We will be under the same exact setup of Section 5.2 and will consider the following settings, each aimed at testing a specific kind of violation.

- **Setting 1 (Violation of the Gaussianity of errors—effect of heavy tails):** For each specific value of (n, d), we draw i.i.d. errors from a t distribution with ν degrees of freedom. We vary ν between 30 and 2. For $\nu > 2$, we also standardize the mean zero errors with the standard deviation of the t_ν distribution. We do not do this for t_2 as it does not have a finite second moment. As ν varies from 30 to 2, the tails of t_ν gets fatter as compared to the normal distribution. We summarize our results in Figure 9.

- **Setting 2 (Violation of the Gaussianity of errors—effect of skewness):** We consider a setup similar to that in Setting 1 but instead consider Gamma distributed error with scale parameter 1 and shape parameter, α. We vary α between 1 and 10 and for each error draw, we standardize the error with the mean and standard deviation of the Gamma$(1, \alpha)$ distribution. This error distribution is asymmetric for smaller values of α and moves towards symmetry as the value of α increases. We summarize our results in Figure 10.

- **Setting 3 (Violation of homoskedasticity of error):** We again consider a similar setup as above, but change the error distribution as follows: For design matrix, $\boldsymbol{X}$, we define $m_{\boldsymbol{X}}$ to be the median of the mean of the rows, that is median of the elements of

$\boldsymbol{X}\mathbf{1}/d$. Let r_i denote mean of the i^{th} row of the design matrix, we generate the error vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, where,

$$\epsilon_i \overset{i.i.d.}{\sim} \begin{cases} \mathcal{N}(0,1), & \text{if } r_i \leq m_{\boldsymbol{X}} \\ \mathcal{N}(0,\eta^2), & \text{if } r_i > m_{\boldsymbol{X}} \end{cases},$$

where $\eta^2 > 0$ is a quantity, specified by us, that controls the heteroskedasticity in the error term. For $\eta = 1$, the distribution is homoskedastic while becomes heteroskedastic for larger and smaller values of η. We vary η^2 in the set $\{0.01, 0.25, 0.5, 1, 4, 8\}$ and compare the performance of the two tests for each of these values. The results are summarized in Figure 11.

- **Setting 4 (Violation of the linearity assumption):** In this setting we test the robustness of the two tests to non-linearity in the model. We consider settings with similar specifications as the above three cases but with i.i.d. homoskedastic, normal errors with variance σ^2. In this case, for design matrix, $\boldsymbol{X}$, and error term, $\boldsymbol{\epsilon}$, we define,

  $$y_i = (\boldsymbol{X}_i^{\delta})^T \boldsymbol{\beta} + \epsilon_i,$$

  where δ is variable we will control, and $\boldsymbol{X}_i^{\delta}$ denotes a vector whose j^{th} entry is the j^{th} entry of $\boldsymbol{X}_i$ raised to the exponent, δ. $\delta = 1$ recovers the usual linear model and larger departure of δ from 1 imparts higher degree of non-linearity to the model. We vary δ in the set, $\{0.3, 0.5, 1, 2, 3, 4\}$ and summarize the results in Figure 12.

The results from all the three simulations suggest that the ℓ-test and the t-test have similar degree of tolerance against violations of the linear model assumptions. The t-test exhibits robustness against the violation of the Gaussianity assumption in the error term, as Figures 9 and 10 indicate and we see a similar behavior for the ℓ-test. Notably, the performance of ℓ-test is almost similar to that of the t-test in the extreme cases, such as when the degrees of freedom for the error t-distribution is 2 in Figure 9 (indicating fat tails of the error distribution) or when the shape parameter of the centered gamma distributed error is 1 (which essentially is a centered exponential distribution with rate 1, indicating a high degree of skeweness in the error term). For Setting 3 with heteroskedastic errors, we see from Figure 11 that the ℓ-test's size does depart from its nominal target of 0.05 as η moves away from 1 (which is the homoskedastic case). However, this departure is of a similar degree as that of the t-test, so that both the tests exhibit similar robustness properties under this setting. From Figure 12 as well, we see that both the t-test and the ℓ-test are robust to the violation of non-linearity.

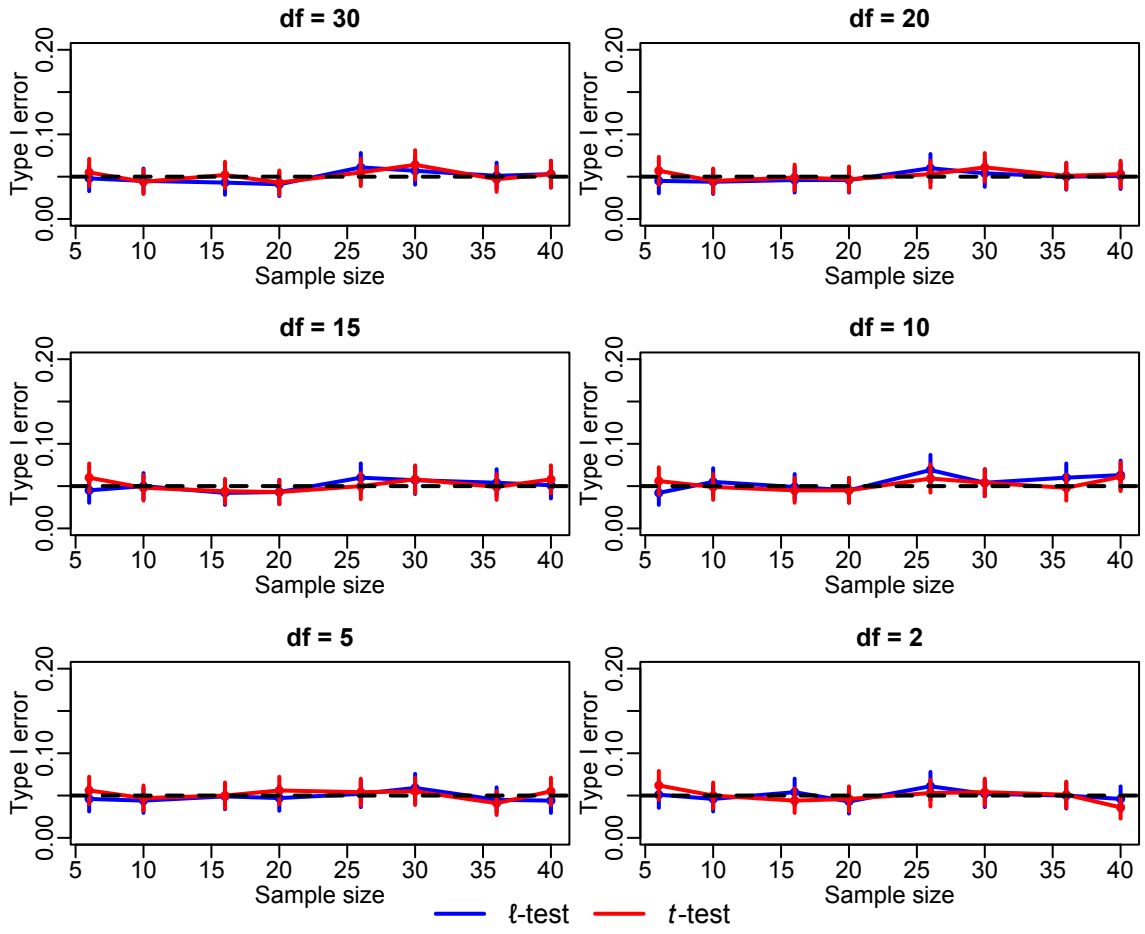

Figure 9: Effect of t-distributed errors on the size of the ℓ-test and the t-test, for different degrees of freedom. The error bars represent plus or minus two standard errors.

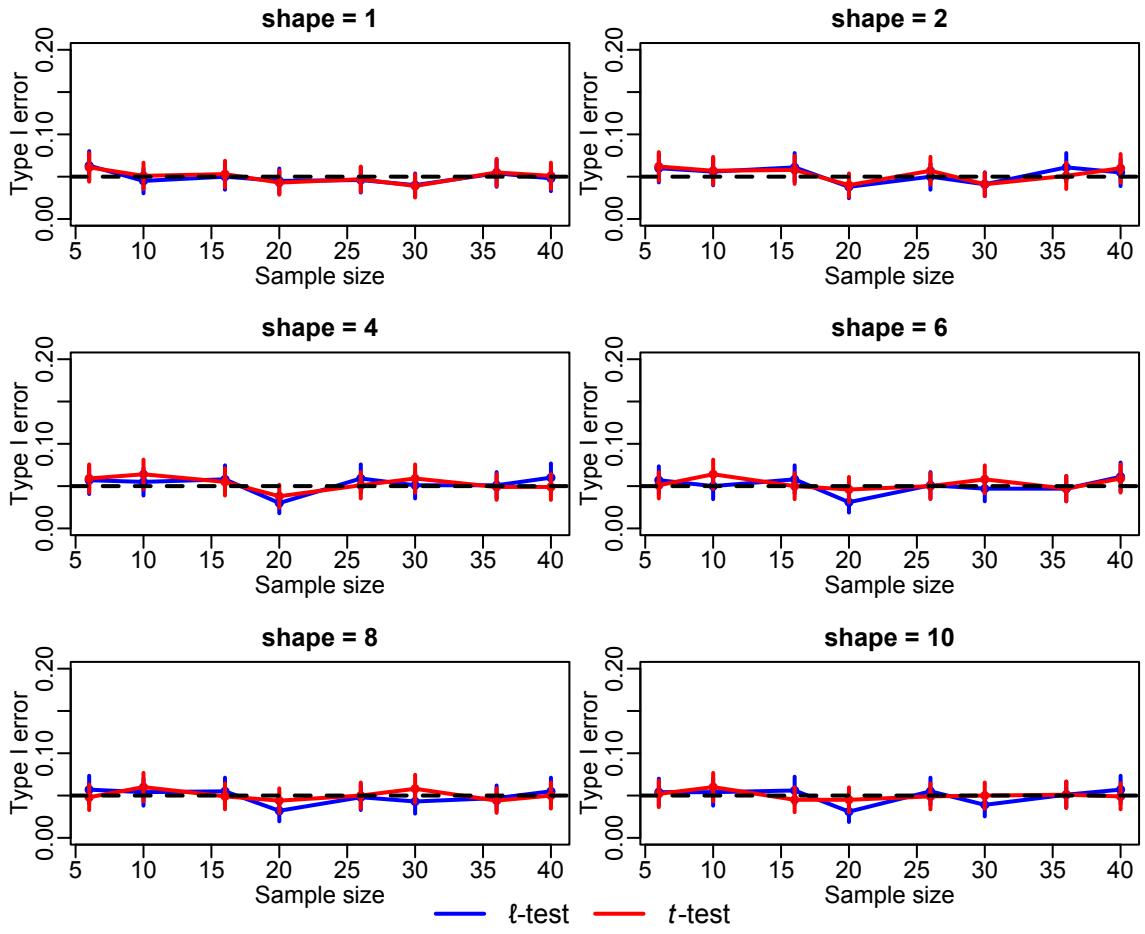

Figure 10: Effect of Gamma distributed errors (with scale parameter 1) on the size of the ℓ-test and the t-test, for different values of the shape parameter. The error bars represent plus or minus two standard errors.

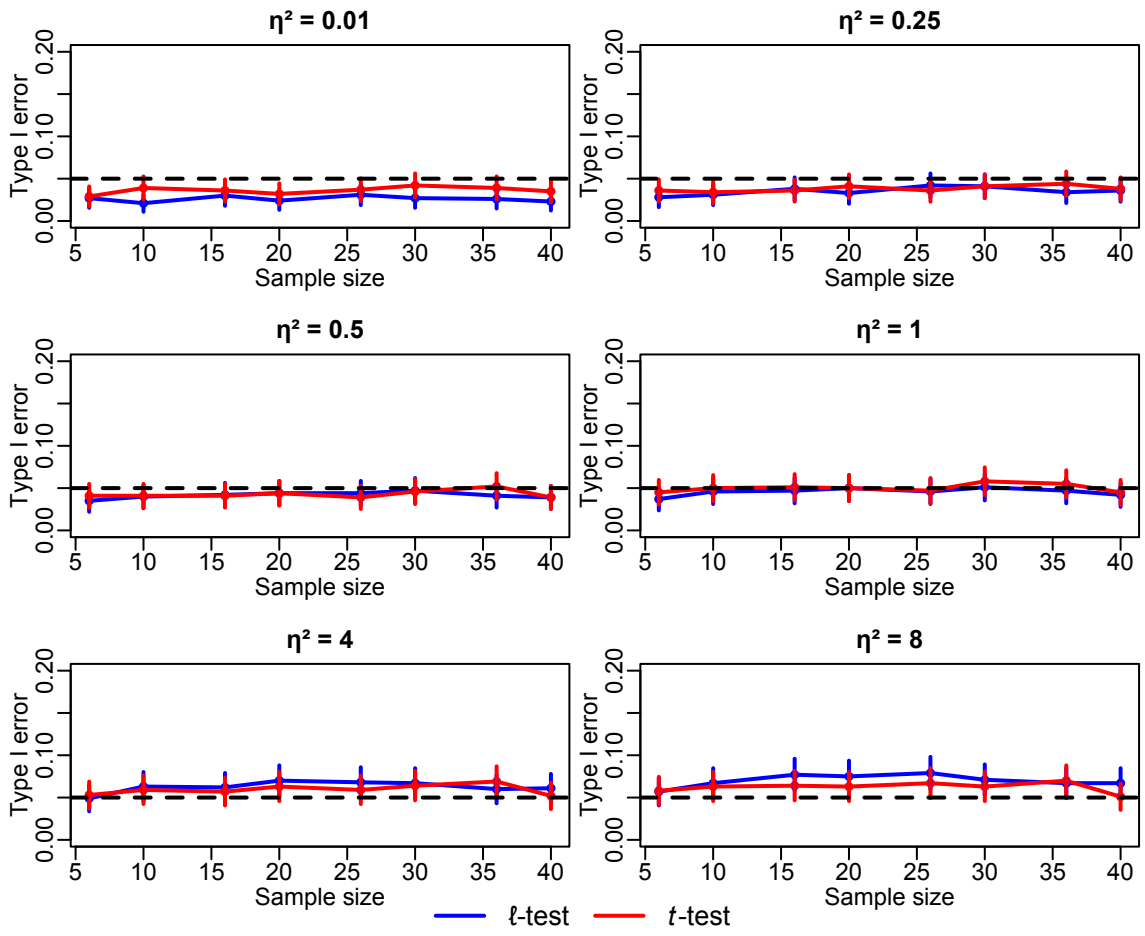

Figure 11: Effect of heteroskedasticity on the size of the ℓ-test and the t-test. Values of η away from 1 indicate higher degree of heteroskedasticity. The error bars represent plus or minus two standard errors.

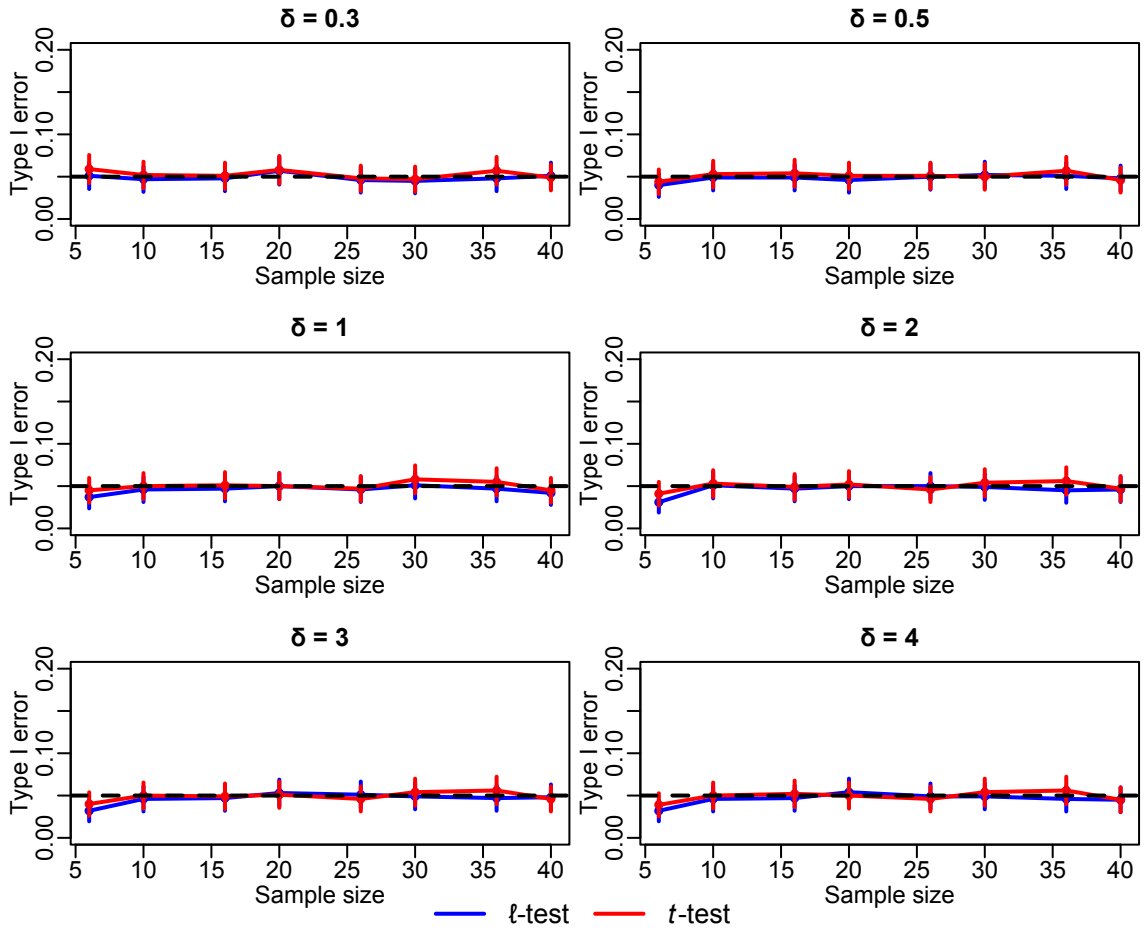

Figure 12: Effect of non-linearity on the size of the ℓ- and the t-test. Departure of the exponent, δ away from 1 indicates higher degree of non-linearity in the model. The error bars represent plus or minus two standard errors.

F.3 Comparison of the various confidence intervals in the linear model

In this section, we summarize the full results of our simulations comparing the various procedures of obtaining confidence intervals for β_j. We consider four different experimental settings in Figure 13 (studying the effect of varying the amplitude with d much smaller than n), Figure 14 (studying the effect of varying the amplitude with d and n close), Figure 15 (studying the effect of varying the sparsity in $\boldsymbol{\beta}$) and Figure 16 (studying the effect of varying the inter-variable correlations). The linear model we consider has exactly the same specifications as in Section 5.3 and the captions of the respective figures contain the further details. For the ℓ-test confidence interval, we consider the performance of the procedure when both the min rule and the 1se rule is used to choose λ. In addition to our observations in Section 5.3 and similar to those in Section F.1, we observe from Figure 14 that the gap between the average length of the confidence interval obtained by inverting the one-sided t-test and the ℓ-test confidence interval is even smaller when $d = 90$, and can again be

justified by our observations in Appendix B. We also see that, except for the cases when the vector $\boldsymbol{\beta}$ is very dense, the performance of the 1se rule and the min rule is almost identical, despite the former being a stricter selection rule. Even though the former chops-off a larger mass from the distribution of u_1, it is less likely that this region significantly overlaps with the corresponding 5% rejection region and hence the corresponding smoothed out statistic (as described in Section 2.3) under both the rules have almost similar distribution in the 5% rejection region. However, as is evident from Figure 15, the 1se rule does perform worse than the min rule when the most of the entries of $\boldsymbol{\beta}$ are signals and can be attributed to the fact that in this case the LASSO is not a good estimator of $\boldsymbol{\beta}$ and most of our intuitions break down.

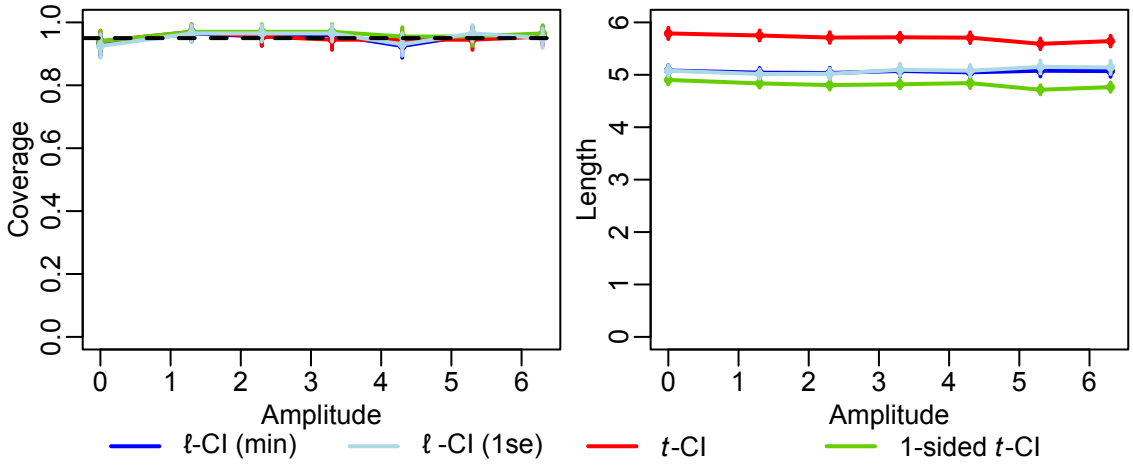

Figure 13: Length and coverage of various 95% confidence intervals. We set $n = 100, d = 50, k = 5, \boldsymbol{\Sigma} = \boldsymbol{I}, \sigma = 1$ and vary A. The error bars represent plus or minus two standard errors.

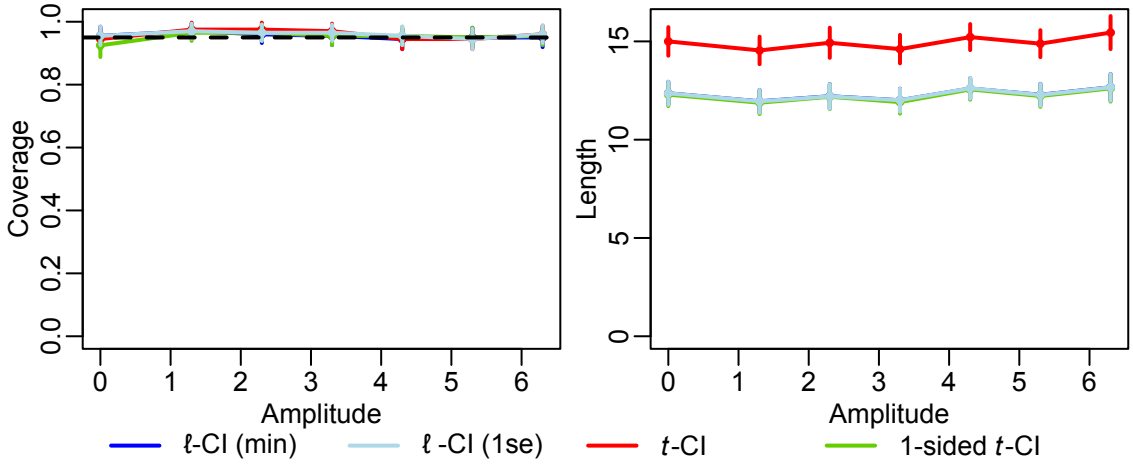

Figure 14: Length and coverage of various 95% confidence intervals. We set $n = 100, d = 90, k = 5, \boldsymbol{\Sigma} = \boldsymbol{I}, \sigma = 1$ and vary A. The error bars represent plus or minus two standard errors.

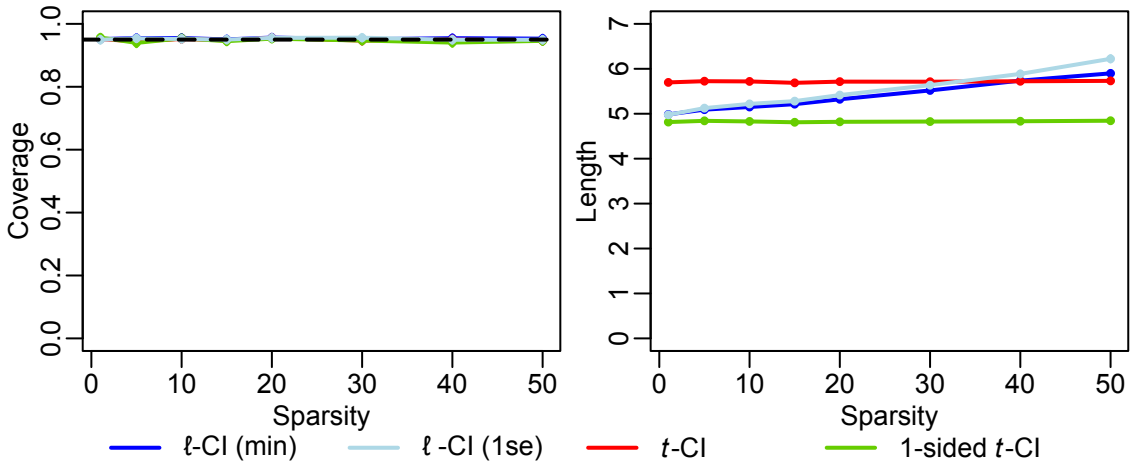

Figure 15: Length and coverage of various 95% confidence intervals while varying the sparsity. We set $n = 100, d = 90, A = 4.3, \mathbf{\Sigma} = \boldsymbol{I}, \sigma = 1$ and vary k. The error bars represent plus or minus two standard errors.

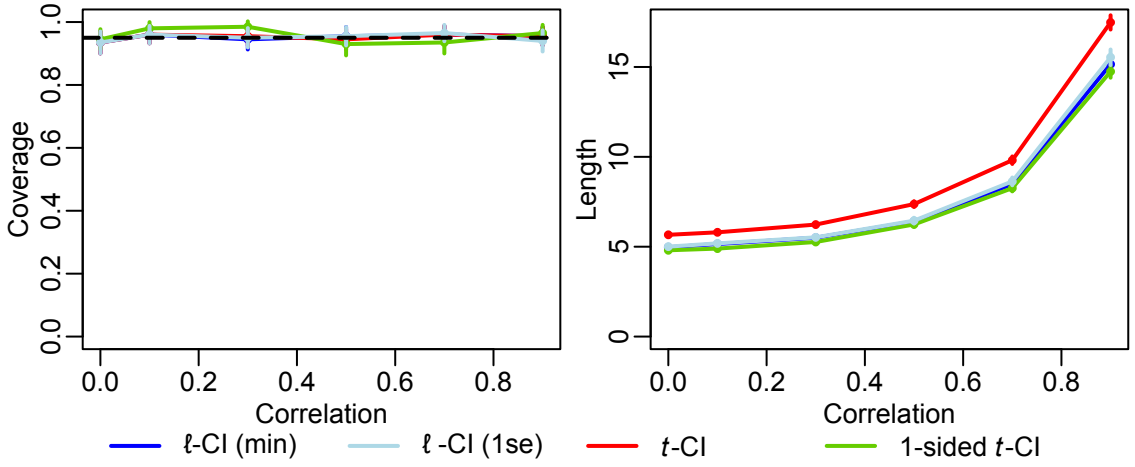

Figure 16: Length and coverage of various 95% confidence intervals while varying the inter-variable correlation. We set $n = 100, d = 90, A = 4.3, k = 4, \mathbf{\Sigma}_{ij} = \rho^{|i-j|}, \sigma = 1$ and vary ρ. The error bars represent plus or minus two standard errors.

F.4 LASSO selection-adjusted confidence sets for linear model coefficients

Expanding on Section 5.4, we compare the various conditional-on-LASSO-selection inference methods under another setting where $d = 90$, with the results plotted in Figure 17. As in Figure 5, there is practically no difference between the performance of $\hat{\underline{C}}_j^{\lambda}$ and $\hat{\underline{C}}_j^{*\lambda}$ but, unlike in Figure 5, Liu et al. (2018) performs quite similarly to them as well. The reason for this is that, as noted in Section 1.3, Liu et al. (2018)'s method assumes known σ^2 (and hence we provide it the true value of σ^2 in both this simulation and that of Figure 5) while the ℓ-test

does not. Unknown σ^2 is what necessitates the second component of the sufficient statistic $\boldsymbol{S}^{(j)}$, namely, $\boldsymbol{y}^T\boldsymbol{y}$, which the ℓ-test conditions on but Liu et al. (2018)'s method does not. This difference becomes information-theoretically more pronounced as d approaches n, as it does in the simulation in Figure 17, because the residual degrees of freedom decreases, resulting in a relative power loss for the ℓ-test-based confidence intervals that approximately offsets the benefit of the ℓ-test. To confirm this explanation, we can easily derive a version of the ℓ-test and corresponding conditional confidence intervals that assumes σ^2 is known and hence does not condition on $\boldsymbol{y}^T\boldsymbol{y}$ (see footnote in Section 6). This is plotted as dashed curves in Figure 17, and as expected these methods again provide consistently shorter confidence intervals than the method in Liu et al. (2018).

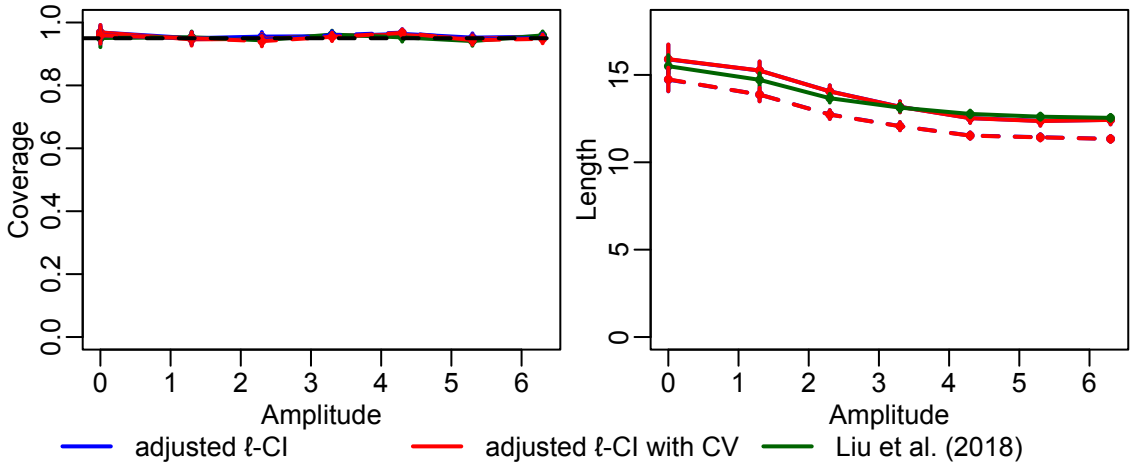

Figure 17: We consider the exact same setting as Figure 5 but with $d = 90$. The dashed curves represent the counterparts of the corresponding ℓ-test procedures with known σ^2. The error bars represent plus or minus two standard errors.

G Inverting the one-sided t-test

In this section, we describe the confidence interval obtained by inverting the one-sided t-test, as mentioned in Section 5.3. For the linear model, $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, the one-sided t-test tests $H_j(\gamma) : \beta_j = \gamma$ by rejecting for large values of the t-test statistic if $\beta_j > \gamma$ and and for small values, if $\beta_j < \gamma$. In case $\beta_j = \gamma$, we, without any loss of generality, fix the convention of rejecting for small values of the t-test statistic, however, the validity of the test is not affected if rejecting for large values as well. Thus, with this convention, the one-sided t-test rejects $H_j(\gamma)$ when $\frac{\hat{\beta}_{j,\mathrm{OLS}}-\gamma}{\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right)} > t_{\alpha;n-d}$, if $\beta_j > \gamma$ and when $\frac{\hat{\beta}_{j,\mathrm{OLS}}-\gamma}{\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right)} < -t_{\alpha;n-d}$, if $\beta_j \leq \gamma$. Here $t_{\alpha;n-d}$ is the quantile of the t_{n-d} distribution putting mass α on its upper tail. Inverting this test gives the following one-sided $100(1-\alpha)\%$ confidence interval,

$$C^t_{1-\mathrm{sided}} = \begin{cases} \left[\hat{\beta}_{j,\mathrm{OLS}} \pm t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right)\right], & \text{if } \beta_j \in \left[\hat{\beta}_{j,\mathrm{OLS}} \pm t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right)\right] \\ \left[\beta_j, \hat{\beta}_{j,\mathrm{OLS}} + t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right)\right), & \text{if } \beta_j < \hat{\beta}_{j,\mathrm{OLS}} - t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right) \\ \left(\hat{\beta}_{j,\mathrm{OLS}} - t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right), \beta_j\right], & \text{if } \beta_j > \hat{\beta}_{j,\mathrm{OLS}} + t_{\alpha;n-d}\hat{\mathrm{SE}}\left(\hat{\beta}_{j,\mathrm{OLS}}\right). \end{cases}$$

Note that the above interval is indeed an oracle interval as one needs to know the exact value of β_j to construct it. Furthermore note that when $\beta_j > \hat{\beta}_{j,\text{OLS}} + t_{\alpha;n-d}\hat{\text{SE}}\left(\hat{\beta}_{j,\text{OLS}}\right)$, the interval surely misses the target parameter, β_j, and this is because of the one-sided nature of the test when we are testing at the true parameter, $\gamma = \beta_j$. In case we were rejecting for large values of the t-test statistic on observing β_j, we would get an open interval in the case when $\beta_j < \hat{\beta}_{j,\text{OLS}} - t_{\alpha;n-d}\hat{\text{SE}}\left(\hat{\beta}_{j,\text{OLS}}\right)$ instead.