

The L -test: Increasing the Linear Model F -test’s Power Under Sparsity Without Sacrificing Validity

Danielle Paulson Souhardya Sengupta Lucas Janson

November 2025

Abstract

We introduce a new procedure for testing the significance of a set of regression coefficients in a Gaussian linear model with $n \geq d$. Our method, the L -test, provides the same statistical validity guarantee as the classical F -test, while attaining higher power when the nuisance coefficients are sparse. Although the L -test requires Monte Carlo sampling, each sample’s runtime is dominated by simple matrix–vector multiplications so that the overall test remains computationally efficient. Furthermore, we provide a Monte-Carlo-free variant that can be used for particularly large-scale multiple testing applications. We give intuition for the power of our approach, validate its advantages through extensive simulations, and illustrate its practical utility in both single- and multiple-testing contexts with an application to an HIV drug resistance dataset. In the concluding remarks, we also discuss how our methodology can be applied to a more general class of parametric models that admit asymptotically Gaussian estimators.

1 Introduction

1.1 Motivation

Consider the homoscedastic Gaussian linear model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is full-column rank with $n \geq d$. The design matrix could be random, but we treat it as fixed throughout to avoid having to explicitly condition on it. The parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}_{>0}$ are treated as unknown. To assess the contribution of a specified subset of covariates, we focus without loss of generality on the first k and test the hypothesis

$$H_{1:k} : \boldsymbol{\beta}_{1:k} = \mathbf{0}.$$

The classic approach is the linear regression F -test proposed by Fisher a century ago [13]. While the F -test is widely used, it lacks flexibility in that it does not allow a user to incorporate known structural information. In particular, we are interested in settings where

the coefficient vector is *sparse*, meaning many of the true coefficients are zero (though we find our method also performs well under various forms of *approximate* sparsity as well). Real-world systems in the sciences and engineering often exhibit sparsity because although numerous potential predictors may exist, typically only a small number meaningfully affects the outcome [e.g., 20, 7]. The goal of this work is to allow a practitioner to leverage a belief about sparsity outside the tested subset for increased power while maintaining exact Type I error control under the same assumptions as the F -test. The recent work [35] addresses this goal in the case $k = 1$, but their approach does not extend to the multivariate case due to the greater complexity that quickly arises beyond a single dimension.

1.2 Our contribution

We propose a new hypothesis test for $H_{1:k}$ called the L -test that uses a group LASSO estimate for $\beta_{1:k}$ in its test statistic. Under the same assumptions as the F -test, it retains exact validity while achieving greater power when the nuisance parameters are sparse and, empirically, only suffers at most very minor power loss even when the nuisance parameters are dense. While the L -test is computationally efficient, we show how a slight modification to the L -test statistic yields an even faster companion test valuable in particularly large-scale multiple testing settings. We provide insight into our methodology’s power, demonstrate its power gains and robustness through extensive simulations, and showcase its practical effectiveness on an HIV drug resistance dataset.

1.3 Background and related work

Many works have tried to generalize the F -test to settings where its classical assumptions—particularly normality—may not hold by developing more robust tests [17, 31, 32, 26, 40, 19, 1, 24, 12, 15, 18, 28]. Because these methods are focused on extending the F -test to more general settings, they typically do not surpass the F -test’s performance within the standard Gaussian linear model. In contrast, the present work aims to make the *same* assumptions as the F -test, while leveraging information about sparsity for increased power.

A wide range of methods that leverage sparsity in particular to improve testing in the linear model have been developed. A prominent example is fixed-X knockoffs [4, 9, 23] and its many extensions [e.g., 37, 30, 33, 27] as well as methods based on mirror statistics [10, 42]. All of these methods, however, are designed for multiple testing and cannot be used for testing a single hypothesis, which is the focus of this work.

A related approach to this paper’s is Frequentist, assisted by Bayes (FAB), which incorporates auxiliary information through a prior and uses it to design frequentist procedures—confidence intervals [22] and hypothesis tests [21, 8, 11]—that can outperform classical ones when the prior aligns with the truth. The L -test is similar in spirit, in that it leverages available information about the unknown parameters for improved inference, just not in the form of a Bayesian prior. The L -test is also tailored to exploit sparsity for higher power, and to our knowledge, FAB has not been used to target this structural feature in hypothesis testing.

Closer to our setting, a few works have studied the problem of leveraging sparsity for single multi-parameter hypothesis testing in the linear model. For example, [2] develops procedures

for testing the global null that are asymptotically powerful under alternatives with different sparsity levels. These procedures are calibrated for asymptotic (not finite-sample) validity, and because they are tailored to testing the global null, they rely on sparsity exclusively among the null coefficients. By contrast, our approach is exactly valid in finite samples and leverages sparsity in the nuisance parameters—an advantage when the tested group is small, making it most useful to exploit sparsity outside the null. A complementary line of work uses the asymptotic distribution of the debiased LASSO estimator [41, 44, 25] to construct confidence regions and p-values for low-dimensional subsets of β ; these tests are likewise only asymptotically valid and, moreover, their validity relies on a strict sparsity assumption. In our case, validity does not hinge on sparsity; sparsity is solely used to increase power. The recent work [35] is the one that is closest to ours in the literature, as it achieves our precise objective when $k = 1$. Unfortunately, key elements of that work do not extend beyond $k = 1$, as we will discuss in more detail when we lay out the background for our approach in the next section.

1.4 Notation

Boldfaced symbols represent matrices and vectors. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\text{col}(\mathbf{A})$ refers to the column space of $\mathbf{A}$. We also use $\mathbf{A}_{1:i}$ to represent its first i columns, $\mathbf{A}_{-1:i}$ to represent its submatrix with the first i columns dropped, and A_{ij} to represent the entry in the i -th row and j -th column. Similarly, for a vector $\mathbf{a} \in \mathbb{R}^n$, we use $\mathbf{a}_{1:i}$ to represent its first i elements, $\mathbf{a}_{-1:i}$ to represent its subvector with the first i elements dropped, and a_i to represent its i -th entry. All vectors are assumed to be column vectors. We use the notation $\mathbb{I}[\cdot]$ to refer to the indicator function that takes the value 1 if the condition inside the brackets is true and 0 otherwise. We also use the notation $\mathbb{S}^{n-1}$ to represent the unit sphere in $\mathbb{R}^n$. Finally, we use $\|\cdot\|_2$ to denote the ℓ^2 -norm and $\|\cdot\|_1$ to denote the ℓ^1 -norm. When no subscript is included on a normed expression, the ℓ^2 -norm is assumed.

1.5 Code

Python code for implementing our methodology and reproducing all empirical results in this paper can be accessed at this GitHub repository: <https://github.com/dipaulson/L-test/>.

2 Methodology

2.1 Conceptual framework

We start by following the idea of the ℓ -test [35] and defining an analogous test statistic for general k whose null distribution we obtain by conditioning on a suitable sufficient statistic. The minimal sufficient statistic for the linear model in (1) under $H_{1:k}$ is $\mathbf{S}^{(1:k)} := (\mathbf{X}_{-1:k}^T \mathbf{y}, \mathbf{y}^T \mathbf{y})$. By definition of a sufficient statistic, the distribution of $\mathbf{y} \mid \mathbf{S}^{(1:k)}$ under $H_{1:k}$ is free of unknown parameters, and hence the distribution of any deterministic function of $\mathbf{y}$ conditional on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$ will also be parameter-free. Define the following group LASSO estimate

[3, 45, 43] with penalty λ ¹:

$$\hat{\beta}^\lambda := \arg \min_{\beta \in \mathbb{R}^d} \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda (\|\beta_{1:k}\|_2 + \|\beta_{-1:k}\|_1) \right). \quad (2)$$

Note that while (2) seems natural for our setting, many other LASSO-based estimators have been proposed in the literature [e.g., 39, 47, 46, 36], and we motivate this choice in the context of some of these alternatives in Appendix E.4. Also let $\mathbf{V} \in \mathbb{R}^{n \times (n-d+k)}$ denote a matrix whose columns constitute an orthonormal basis for $\text{col}(\mathbf{X}_{-1:k})^\perp$ with the first k columns given by $\mathbf{V}_i = \frac{(\mathbf{I} - \mathbf{P}_{-1:i})\mathbf{X}_i}{\|(\mathbf{I} - \mathbf{P}_{-1:i})\mathbf{X}_i\|}$ for $i \in \{1, \dots, k\}$ and $\mathbf{P}_{-1:i}$ the projection matrix onto $\text{col}(\mathbf{X}_{-1:i})$ (for an explicit construction, see Appendix E.1). By the previous sufficiency argument, the test statistic

$$T(\mathbf{y}) := \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda \right\| \quad (3)$$

conditional on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$ has a distribution free of unknown parameters. To understand the form of (3), note that the F -test can be shown to be exactly equivalent to rejecting large values of

$$\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\beta}_{1:k, \text{OLS}} \right\| \quad (4)$$

conditional on $\mathbf{S}^{(1:k)}$, where $\hat{\beta}_{\text{OLS}}$ denotes the OLS estimate (see Lemma C.2 for a proof). So, (3) is simply the result of replacing the OLS estimate $\hat{\beta}_{1:k, \text{OLS}}$ with the group LASSO estimate $\hat{\beta}_{1:k}^\lambda$ in the F -test statistic defined in (4). The motivation is that since the group LASSO is a better estimator of $\beta_{1:k}$ than OLS when $\beta_{-1:k}$ is sparse, we expect using $\hat{\beta}_{1:k}^\lambda$ in the test statistic will lead to superior power in such a setting.

One way to compute a p-value for the test that rejects for large values of $T(\mathbf{y})$ conditional on $\mathbf{S}^{(1:k)}$ is via the co-sufficient sampling (CSS) framework [6, 38, 5], using Monte Carlo (MC) draws:

$$\frac{1}{M+1} \left(1 + \sum_{i=1}^M \mathbb{I}[T(\mathbf{y}^{(i)}) \geq T(\mathbf{y})] \right), \quad \mathbf{y}^{(i)} \stackrel{\text{i.i.d.}}{\sim} \mathbf{y} \mid \mathbf{S}^{(1:k)} \text{ under } H_{1:k}. \quad (5)$$

We can sample from $\mathbf{y} \mid \mathbf{S}^{(1:k)}$ under $H_{1:k}$ by letting $\hat{\mathbf{y}}_{1:k} = \mathbf{P}_{-1:k}\mathbf{y}$ and $\hat{\sigma}_{1:k}^2 = \|(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2$ and noting that there exists a unique unit vector $\mathbf{u} \in \mathbb{R}^{n-d+k}$ such that

$$\mathbf{y} = \hat{\mathbf{y}}_{1:k} + \hat{\sigma}_{1:k} \mathbf{V} \mathbf{u}, \quad \mathbf{u} \mid \mathbf{S}^{(1:k)} \stackrel{H_{1:k}}{\sim} \text{Unif}(\mathbb{S}^{n-d+k-1}), \quad (6)$$

a standard result that is derived as a multivariate extension of Proposition E.1 in [30] in Lemma C.1. Because the data $\mathbf{y}$ and its sampled copies $\{\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}\}$ are i.i.d. conditional

¹The penalty in the original definition of the group LASSO by [3] is formulated as a sum of Mahalanobis norms over groups, with each group j associated with a positive semidefinite matrix $\mathbf{K}_j$. Different specifications of $\mathbf{K}_j$ lead to different variants of the group LASSO. We use the common default choice of $\mathbf{K}_j = \mathbf{I}$, suggested in [43].

on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$, the data and its copies are exchangeable under $H_{1:k}$. As a result, the MC p-value in (5) stochastically dominates $\text{Unif}(0, 1)$ under $H_{1:k}$ and is hence valid.

The difficulty with this MC approach, however, is that it requires running the group LASSO on each MC sample—a computationally expensive task, especially when large M is needed for high-resolution p-values, as in multiple testing settings. Everything in this subsection thus far matches the preliminaries of the ℓ -test [35] when $k = 1$, but that paper is then able to avoid MC by deriving the null distribution of $T(\mathbf{y}) \mid \mathbf{S}^{(1:k)}$ (when $k = 1$). Their proof approach relies on insightfully characterizing $\hat{\beta}_1^\lambda$ as a function of u_1 (Theorem 2.1 in [35])—the first component of the unit vector $\mathbf{u}$ defined in (6)—to convert the p-value from $\hat{\beta}_1^\lambda$ -space to u_1 -space, but extending this idea to higher dimensions is intractable. In particular, for $k > 1$, understanding $\hat{\beta}_{1:k}^\lambda$ as a function of $\mathbf{u}_{1:k}$ is more complicated since the matrix derivative $\frac{\partial \hat{\beta}_{1:k}^\lambda}{\partial \mathbf{u}_{1:k}}$ has no simplifying structure, and it is also unclear what analytic properties would even enable a transformation of the p-value into $\mathbf{u}_{1:k}$ -space due to its greater complexity in $\hat{\beta}_{1:k}^\lambda$ -space; see Appendix A.1 for a full discussion. In the remainder of this section, we will introduce a new approach to computationally tractably implement the conceptual idea introduced in this subsection for $k \geq 1$.

2.2 Constructing the L -test

To address the computational bottleneck of solving a group LASSO for every MC draw in (5), we define an affine approximation for $\hat{\beta}_{1:k}^\lambda(\mathbf{y})$ that depends only on $\mathbf{S}^{(1:k)}$, thereby eliminating the need to run a group LASSO on each sample. We call the test that uses this approximation the L -test, which notably only requires two group LASSO runs, a substantial improvement over the original MC test, as Section 4 will show. The experiments in Section 4 also demonstrate that the L -test achieves statistical performance nearly identical to that of the original MC method.

We start by defining the function $f_{\mathbf{S}^{(1:k)}}$ such that $\hat{\beta}_{1:k}^\lambda = f_{\mathbf{S}^{(1:k)}}(\mathbf{u}_{1:k})$. As we discuss in detail in Appendix C.3, $f_{\mathbf{S}^{(1:k)}}$ is challenging to analyze so we instead focus on its inverse $f_{\mathbf{S}^{(1:k)}}^{-1}$. Although evaluating this inverse on an input requires running a LASSO and is therefore computationally demanding (Theorem C.8), the gradient of $f_{\mathbf{S}^{(1:k)}}^{-1}$ constrained to a hypersphere of radius r , $\mathbb{S}^{k-1}(r)$, reveals this inverse has a relatively simple structure. The gradient is shown in Lemma 2.1 below and proved in Appendix C.5.

Lemma 2.1. *Fixing $r, \lambda \in \mathbb{R}_{>0}$, let $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; r) = f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}) \Big|_{\mathbb{S}^{k-1}(r)}$. Then,*

$$\nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; r) = \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \left[\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}(\mathbf{b})}) \mathbf{X}_{1:k} + \frac{n\lambda}{r} \mathbf{I} \right], \quad (7)$$

where

$$\begin{aligned} \mathcal{A}(\mathbf{b}) &= \left\{ i \in \{k+1, \dots, d\} : \hat{\beta}_i^\lambda(\mathbf{b}) \neq 0 \right\} \\ \hat{\beta}_{-1:k}^\lambda(\mathbf{b}) &= \arg \min_{\boldsymbol{\beta}_{-1:k} \in \mathbb{R}^{d-k}} \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}_{1:k} \mathbf{b} - \mathbf{X}_{-1:k} \boldsymbol{\beta}_{-1:k}\|_2^2 + \lambda \|\boldsymbol{\beta}_{-1:k}\|_1 \right). \end{aligned}$$

Since the active set $\mathcal{A}(\mathbf{b})$ is constant in a neighborhood of $\mathbf{b}$, Lemma 2.1 implies that $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; r)$ is piecewise affine. Corollary 2.2 below, proved in Appendix C.6, explicitly specifies the functional form of one of these affine pieces.

Corollary 2.2. *Given a point $\mathbf{b}^* \in \mathbb{R}^k$, for $\mathbf{b}$ in a sufficiently small neighborhood of $\mathbf{b}^*$ such that $\|\mathbf{b}\| = \|\mathbf{b}^*\|$,*

$$\begin{aligned} f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; \|\mathbf{b}^*\|) &= \nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|)\mathbf{b} + \boldsymbol{\nu}(\mathbf{b}^*) \\ \boldsymbol{\nu}(\mathbf{b}^*) &= -\frac{1}{\hat{\sigma}_{1:k}} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \tilde{\boldsymbol{\beta}}_{1:k}^\lambda(\mathbf{b}^*) \\ \tilde{\boldsymbol{\beta}}_{1:k}^\lambda(\mathbf{b}^*) &= \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k}(\mathbf{y} - \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}^*) - P_{\mathcal{A}(\mathbf{b}^*)} \mathbf{X}_{1:k} \mathbf{b}^*) \\ \mathbf{S} &= \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k}. \end{aligned} \tag{8}$$

While an exact characterization of $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; r)$ could be obtained by tracking how the active set $\mathcal{A}(\mathbf{b})$ evolves with its input, it is expensive to compute $\mathcal{A}(\mathbf{b})$ so we instead propose a simpler approach: selecting a single affine piece of this piecewise affine map and using it as a global approximation. This simplification is supported by simulations in which we visualized $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}; r)$ and found it to be approximately ellipsoidal, suggesting that the affine pieces are similar enough for the map to be well-approximated as globally affine; see Appendix E.2 for further details.

Using the affine function corresponding to a point $\mathbf{b}^* \in \mathbb{R}^k$ specified in Corollary 2.2 as an approximation for $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b})$ gives

$$\hat{\boldsymbol{\beta}}_{1:k}^\lambda = f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{u}_{1:k}) \approx (\nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|))^{-1}(\mathbf{u}_{1:k} - \boldsymbol{\nu}(\mathbf{b}^*)),$$

which when plugged into the expression for $T(\mathbf{y})$ in Equation (3) gives

$$T(\mathbf{y}) \approx \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} (\nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|))^{-1}(\mathbf{u}_{1:k} - \boldsymbol{\nu}(\mathbf{b}^*)) \right\|.$$

To use this test statistic, we need to choose a $\mathbf{b}^*$. And, although we have until now treated λ as fixed (and correspondingly suppressed it in our notation), in practice we need a data-driven way to choose it; we will follow the same strategy for choosing both $\mathbf{b}^*$ and λ . The obvious strategy of using cross-validation to choose λ and (the first k elements of) cross-validated group LASSO to choose $\mathbf{b}^*$ invalidates our MC p-value (5) because these choices still depend on the data after conditioning on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$.² Instead, we propose sampling $\tilde{\mathbf{u}} \sim \text{Unif}(\mathbb{S}^{n-d+k-1})$, substituting $\tilde{\mathbf{u}}$ into (6) to obtain $\tilde{\mathbf{y}}$ independent of $\mathbf{y}$ conditional on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$, and then applying this ‘obvious strategy’ to the data $(\tilde{\mathbf{y}}, \mathbf{X})$, denoting the corresponding choices simply (and without risk of confusion, as we fix this λ choice for the rest of the paper) by λ and $\mathbf{b}^* = \tilde{\boldsymbol{\beta}}_{1:k}^\lambda$; Appendix D discusses this approach further.

A last remaining issue is that if $\tilde{\boldsymbol{\beta}}_{1:k}^\lambda = \mathbf{0}$, then Equation (7) for $\nabla f_{\mathbf{S}^{(1:k)}}^{-1}$ diverges to ∞ with $r = 0$ plugged in, resulting in our test statistic in Equation (3) being set to zero for the data and all MC samples. This causes all test statistic comparisons in the MC p-value (5) to evaluate as ties, producing a trivial p-value of 1. We can avoid this pathology by instead

²One could make these choices valid by refitting them to each MC sample, but this sacrifices the computational advantage of our affine approximation.

taking a limit as $r \rightarrow 0$ to give nontrivial comparisons and potentially still a powerful test. Thus, we define our final test statistic for the L -test as

$$L(\mathbf{y}) := \left\| \mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda)(\mathbf{u}_{1:k} - \boldsymbol{\nu}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)) \right\|,$$

where

$$\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda) := \begin{cases} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} (\nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda; \|\tilde{\boldsymbol{\beta}}_{1:k}^\lambda\|))^{-1} & \text{if } \tilde{\boldsymbol{\beta}}_{1:k}^\lambda \neq \mathbf{0} \\ \lim_{r \rightarrow 0} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} (\nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda; r))^{-1} / r = \frac{\sigma_{1:k}}{n\lambda} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} & \text{if } \tilde{\boldsymbol{\beta}}_{1:k}^\lambda = \mathbf{0}. \end{cases}$$

By construction, the MC L -test p-value is valid under the same assumptions as the F -test; we state this formally as a corollary below.

Corollary 2.3 (Validity of the L -test). *For model (1) and all $\alpha \in [0, 1]$,*

$$P_{H_{1:k}} \left(\frac{1}{M+1} \left(1 + \sum_{i=1}^M \mathbb{I}[L(\mathbf{y}^{(i)}) \geq L(\mathbf{y})] \right) \leq \alpha \right) \leq \alpha.$$

2.3 Monte-Carlo-free p-values

Computing the MC p-value with L requires only two group LASSO runs—one to obtain $\tilde{\boldsymbol{\beta}}_{1:k}^\lambda$ and another to obtain $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ —yielding a substantial computational gain over the original MC p-value that requires re-running the group LASSO for each MC sample. However, as can be seen by Equation (5), the MC p-value is lower-bounded by $1/(M+1)$ (recall that M is the number of MC samples). This means that for settings where power requires very small p-values, such as in multiple testing settings with high multiplicities, M may need to be so large that recomputing L for each MC sample, which scales linearly in M , dominates the two group LASSOs' computation time. For such settings that would require an MC p-value to use an astronomically high M , we would prefer a method that does not require MC at all.

To this end, we propose omitting the premultiplying matrix $\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda)$ in L because we can analytically characterize the density of $\|\mathbf{u}_{1:k} - \boldsymbol{\nu}\| \mid \mathbf{S}^{(1:k)}$ under $H_{1:k}$, enabling p-value computation via efficient numerical integration of a two-dimensional integral that is not limited by a p-value lower-bound that scales inversely with computation time. The proof of the following proposition in Appendix C.7 relies on properties of Beta distributions and strategic changes of variables.

Proposition 2.4. *For $\mathbf{u} \sim \text{Unif}(\mathbb{S}^{n-d+k-1})$, $k \geq 2$, and $\mathbf{c} \in \mathbb{R}^k$, the density of the random*

variable $Z = \|\mathbf{u}_{1:k} - \mathbf{c}\|$ is given by

$$\begin{aligned}
f_Z(z) &= \mathbb{I}[\|\mathbf{c}\| \leq 1] \{zg(z)\mathbb{I}[z \in [0, \|\mathbf{c}\| + 1]]\} \\
&\quad + \mathbb{I}[\|\mathbf{c}\| > 1] \{zg(z)\mathbb{I}[z \in [\|\mathbf{c}\| - 1, \|\mathbf{c}\| + 1]]\} \\
g(z) &= \begin{cases} \int_{\|\mathbf{c}\|^2 - 2z\|\mathbf{c}\| + z^2}^{\|\mathbf{c}\|^2 + 2z\|\mathbf{c}\| + z^2} h(z, t) dt & \text{if } 0 \leq z \leq \|\mathbf{c}\| - 1 \\ \int_1^{\|\mathbf{c}\|^2 - 2z\|\mathbf{c}\| + z^2} h(z, t) dt & \text{if } |\|\mathbf{c}\| - 1| < z \leq \|\mathbf{c}\| + 1 \end{cases} \\
h(z, t) &= \frac{D}{\|\mathbf{c}\|} \cdot (1 - t)^{\frac{n-d-2}{2}} \left(t - \left(\frac{t + \|\mathbf{c}\|^2 - z^2}{2\|\mathbf{c}\|} \right)^2 \right)^{\frac{k-3}{2}} \\
D &= \frac{\Gamma(\frac{k}{2})\Gamma(\frac{n-d+k}{2})}{\sqrt{\pi}\Gamma(\frac{k-1}{2})\Gamma(\frac{k}{2})\Gamma(\frac{n-d}{2})}.
\end{aligned}$$

For the density in the $k = 1$ case, see Appendix C.8.

The experiments in Section 4 show that this test often performs similarly to the L -test, but can outperform or underperform in certain situations, and in general less closely matches the original MC test that uses the test statistic (3). Nevertheless, it offers notable gains over the F -test and may be of interest to users who require very precise p-values—such as in multiple testing settings—since they can be obtained efficiently. In Section 5, we show the computational speed up and power gain this test can offer in a real multiple testing setting.

Remark 2.1. This test is similar to the recentered $\mathbf{u}_{1:k}$ -test that is analogous to the recentered test in [35] and discussed in Appendix A.2, except that it uses a different recentering vector and has the key advantage that its p-value can be computed without MC sampling. While both recentering vectors are principled choices and give rise to similar test performance, we prefer the one presented in this work since it fits more neatly with the approach that gives rise to the L -test.

3 Intuition for the L -test's power gain

3.1 An oracle test

To better understand the power gain of the L -test over the F -test under sparsity, it is helpful to compare it to an oracle test that outperforms the F -test. To wit, consider an oracle that knows the *direction* of $\beta_{1:k}$, i.e., $\beta_{1:k}/\|\beta_{1:k}\|$, and uses it to first construct the design matrix $\tilde{\mathbf{X}} := \left(\mathbf{X}_{1:k} \frac{\beta_{1:k}}{\|\beta_{1:k}\|} \quad \mathbf{X}_{-1:k} \right)$ and then perform a 1-sided t -test in the (well-specified) linear model

$$\mathbf{y} \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\gamma}, \sigma^2 \mathbf{I}) \quad (9)$$

of $\tilde{H}_0 : \gamma_1 = 0$ against $\tilde{H}_1 : \gamma_1 > 0$. Since $\gamma_1 = \|\beta_{1:k}\|$, this is a valid test of $H_{1:k}$; note $\tilde{\mathbf{X}}$ is ill-defined under $H_{1:k}$, but as this section is about power, we will only consider the setting

when $H_{1:k}$ is false and hence the oracle is well-defined. When $k = 1$, this oracle is exactly performing the 1-sided t -test on β_1 (in the correct direction), which [35] shows the ℓ -test mimics the performance of. We will argue that the L -test also aims to mimic the performance of the oracle test in the general $k \geq 1$ case, but allowing $k > 1$ creates interesting dynamics not observed in [35]. In particular, the oracle's privileged information, $\beta_{1:k}/\|\beta_{1:k}\|$, is a unit vector in k dimensions, which means that as k grows, the oracle's privileged information gets richer. On the one hand, this makes the oracle's power improvement over the F -test increasing in k . On the other hand, this also makes it increasingly hard for the L -test (or, one would expect, any test) to approach the power of the oracle as k grows. On net, we find in Section 4 that the L -test's power gains over the F -test can be substantially greater for $k > 1$ than the ℓ -test's ($k = 1$) power gain over the t -test.

To compare the oracle with the L -test, it is useful to reframe the oracle procedure as a conditional test given $\mathbf{S}^{(1:k)}$. In fact, it is equivalent to rejecting for large values of

$$O_m := \left\| \mathbf{u}_{1:k} + m \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \frac{\beta_{1:k}}{\|\beta_{1:k}\|} \right\| \quad (10)$$

conditional on $\mathbf{S}^{(1:k)}$, in the limit as $m \rightarrow \infty$. We provide a formal statement and proof of this equivalence in Appendix C.9.

3.2 L -test power

Lemma C.2 shows that the F -test is equivalent to the test that rejects for large $\|\mathbf{u}_{1:k}\|$ conditional on $\mathbf{S}^{(1:k)}$. Thus, the oracle test, F -test, and L -test statistics all take the following form

$$\|\phi(\mathbf{u}_{1:k} - \boldsymbol{\omega})\|,$$

for $\boldsymbol{\omega} \in \mathbb{R}^k$ and positive definite $\phi : \mathbb{R}^k \rightarrow \mathbb{R}^k$ that are conditionally independent of the data given $\mathbf{S}^{(1:k)}$. Letting $\tilde{\mathbf{u}}_{1:k}$ denote the first k elements of a $\text{Unif}(\mathbb{S}^{n-d+k-1})$ drawn independently of the data (and continuing to denote by $\mathbf{u}$ the vector computed from the data via Equation (6)), this test statistic's conditional MC p-value in the limit as the number of MC samples, M , goes to infinity, is

$$\begin{aligned} & P_{H_{1:k}} \left((\tilde{\mathbf{u}}_{1:k} - \boldsymbol{\omega})^T \phi^T \phi (\tilde{\mathbf{u}}_{1:k} - \boldsymbol{\omega}) \geq (\mathbf{u}_{1:k} - \boldsymbol{\omega})^T \phi^T \phi (\mathbf{u}_{1:k} - \boldsymbol{\omega}) \mid \mathbf{S}^{(1:k)} \right) \\ &= P_{H_{1:k}} \left((\tilde{\mathbf{u}}_{1:k} - \boldsymbol{\omega})^T \frac{1}{c} \phi^T \phi (\tilde{\mathbf{u}}_{1:k} - \boldsymbol{\omega}) \geq 1 \mid \mathbf{S}^{(1:k)} \right), \text{ where} \\ & c = (\mathbf{u}_{1:k} - \boldsymbol{\omega})^T \phi^T \phi (\mathbf{u}_{1:k} - \boldsymbol{\omega}). \end{aligned}$$

The set of points $\mathbf{x} \in \mathbb{R}^k$ that satisfy $(\mathbf{x} - \boldsymbol{\omega})^T \frac{1}{c} \phi^T \phi (\mathbf{x} - \boldsymbol{\omega}) = 1$ defines a k -dimensional ellipsoid, so the p-value corresponds to the mass of $\tilde{\mathbf{u}}_{1:k}$ outside this ellipsoid. This ellipsoid's center is determined by the recentering vector $\boldsymbol{\omega}$, and its semi-axes are specified by the eigenvectors and eigenvalues of $\frac{1}{c} \phi^T \phi$. Below we analyze the recentering vector and premultiplier in the context of each of the three tests to shed light on when and how we expect the L -test to gain power over the F -test.

3.2.1 The recentering

The oracle's recentering vector is $\boldsymbol{\omega}^O = m \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \frac{\boldsymbol{\beta}_{1:k}}{\|\boldsymbol{\beta}_{1:k}\|}$ for $m \gg 0$, and the L -test's recentering vector is $\boldsymbol{\omega}^L = \boldsymbol{\nu}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$. The decomposition in (6) implies that $\mathbf{u}_{1:k} = \frac{1}{\hat{\sigma}_{1:k}} \mathbf{V}_{1:k}^T \mathbf{y}$, from which we get that

$$\text{dir}(\mathbf{u}_{1:k}) = \text{dir}(\mathbf{V}_{1:k}^T \mathbf{y}) \approx \text{dir}(\mathbb{E}[\mathbf{V}_{1:k}^T \mathbf{y}]) = \text{dir}(\mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \boldsymbol{\beta}_{1:k}). \quad (11)$$

The oracle essentially recenters with a vector that points with infinite magnitude in the direction exactly opposite (11) because it has the benefit of knowing the exact direction of $\boldsymbol{\beta}_{1:k}$. In a similar manner, the L -test's recentering vector points approximately opposite (11) by making a guess of the direction of $\boldsymbol{\beta}_{1:k}$. We can see this by thinking of $\check{\boldsymbol{\beta}}_{1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ from (8) as an estimator of

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{1:k}^* &:= \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k} (\mathbf{y} - \mathbf{X}_{-1:k} \boldsymbol{\beta}_{-1:k}) \\ &\sim \mathcal{N}(\mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \boldsymbol{\beta}_{1:k}, \sigma^2 \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \mathbf{S}^{-1}), \end{aligned}$$

the same expression that defines $\check{\boldsymbol{\beta}}_{1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ but with $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ replaced with $\boldsymbol{\beta}_{-1:k}$ and $\mathbf{P}_{\mathcal{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)}$ set to zero. Since the group LASSO is exactly designed to handle sparsity, we would expect $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ to estimate $\boldsymbol{\beta}_{-1:k}$ best when $\boldsymbol{\beta}_{-1:k}$ is sparse. Also, when $\boldsymbol{\beta}_{-1:k}$ is sparse, we would expect the group LASSO estimate $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ to be sparse as well, meaning it should have a small active set and hence $\mathbf{P}_{\mathcal{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)} \approx \mathbf{0}$. Taken together, these heuristic arguments suggest that when $\boldsymbol{\beta}_{-1:k}$ is sparse, $\check{\boldsymbol{\beta}}_{1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)$ should approximately be distributed like $\hat{\boldsymbol{\beta}}_{1:k}^*$, which in turn allows us to say that

$$\text{dir}(\boldsymbol{\nu}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)) = -\text{dir}\left(\mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \check{\boldsymbol{\beta}}_{1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)\right) \approx -\text{dir}(\mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k}^*).$$

Further approximating by replacing $\hat{\boldsymbol{\beta}}_{1:k}^*$ by its expectation gives

$$\text{dir}(\boldsymbol{\nu}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda)) \approx -\text{dir}((\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \boldsymbol{\beta}_{1:k}),$$

which forms an obtuse angle with (11). Note that there is no recentering effect when the blocks $\mathbf{X}_{1:k}$ and $\mathbf{X}_{-1:k}$ are orthogonal since $\check{\boldsymbol{\beta}}_{1:k}^\lambda(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda) = \mathbf{0}$, a case we study empirically in Section 4.2.

3.2.2 The premultiplier

The L -test is the only one of the three tests that takes $\boldsymbol{\phi} = \mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda)$ to not be the identity. A detailed analysis of this premultiplier is performed in Appendix B and reveals that its structure depends on how the variance is allocated among the first k predictors. In particular, when the variance is spread evenly across the first k predictors, $\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda) \approx \mathbf{I}$. By contrast, when nearly all the variance is concentrated in a single principal component (PC) and the true $\boldsymbol{\beta}_{1:k}$ is well aligned with that component, then $\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda)$ has one dominant eigenvalue with eigenvector approximately parallel to $\mathbf{u}_{1:k}$, while the remaining $k - 1$ eigenvalues are close to zero. Geometrically, this essentially translates to the ellipse having a single finite principal axis along $\mathbf{u}_{1:k}$ and all other axes orthogonal to it extending infinitely.

3.2.3 Putting the pieces together

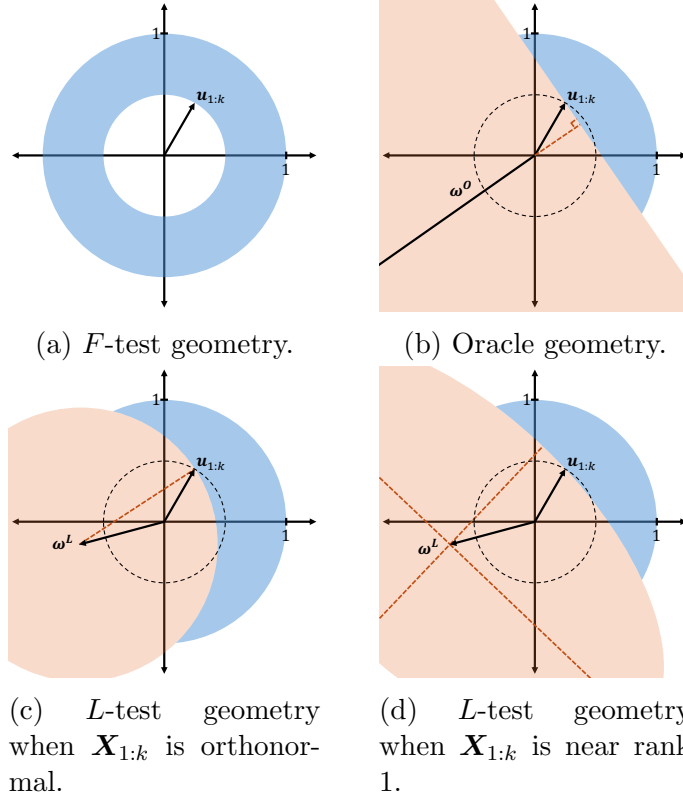


Figure 1: Schematics representative of the geometries of the oracle, F -, and L -tests. In all panels, the blue region corresponds to the region whose mass equals the p-value. Note that the distribution of $\tilde{\mathbf{u}}_{1:k}$ is not uniform within the unit ball in $\mathbb{R}^k$.

Since the F -test performs no recentering and takes $\phi = \mathbf{I}$, its p-value corresponds to the mass in the disk region in Figure 1a. Meanwhile, the oracle p-value corresponds to the mass of $\tilde{\mathbf{u}}_{1:k}$ outside the circle centered at $\omega^O = m\mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \frac{\beta_{1:k}}{\|\beta_{1:k}\|}$ of radius $\left\| \mathbf{u}_{1:k} - m\mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \frac{\beta_{1:k}}{\|\beta_{1:k}\|} \right\|$. As $m \rightarrow \infty$, this is equivalent to the region outside the half-space shown in Figure 1b.

The L -test p-value corresponds to the mass of $\tilde{\mathbf{u}}_{1:k}$ lying outside an ellipse centered at $\omega^L = \nu(\tilde{\beta}_{1:k}^\lambda)$, which will typically be smaller in magnitude and less anti-aligned with $\mathbf{u}_{1:k}$ than ω^O . When variance is distributed relatively uniformly in the span of the first k predictors, the premultiplier essentially reduces to the identity, so the ellipse is simply a circle of radius $\left\| \mathbf{u}_{1:k} - \nu(\tilde{\beta}_{1:k}^\lambda) \right\|$ as Figure 1c shows. This case isolates the role of the recentering vector: it enables the L -test to leverage auxiliary information about sparsity to infer the direction of $\beta_{1:k}$ —and, in turn, the opposite direction of $\mathbf{u}_{1:k}$ —emulating the oracle. At the other extreme, when nearly all variance is captured by a single principal component and the true $\beta_{1:k}$ aligns with it, the ellipse degenerates into one principal axis roughly aligned with $\mathbf{u}_{1:k}$ and $k - 1$ orthogonal axes of very large magnitude, shown in Figure 1d. We can interpret the principal component direction as a second attempt at inferring the direction of $\beta_{1:k}$ since this degenerate geometry only arises if the two are well aligned. Thus, this second

case demonstrates how the premultiplier allows the L -test to sometimes exploit additional, unsupervised information encoded in the geometry of $\mathbf{X}_{1:k}$ to further infer the direction of $\beta_{1:k}$ and come even closer to approximating the power of the oracle.

4 Simulation studies

In this section, we conduct simulations to evaluate the L -test’s performance. In all simulations, we use the linear model (1) and perform inference on the first k coefficients $\beta_{1:k}$. The rows of $\mathbf{X}$ are drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho \in [0, 1)$, and then the columns are standardized. We randomly choose k_1 out of the first k coefficients of β to be $\pm A/\sqrt{k_1}$ with equal probability and sample k_2 out of the remaining $d-k$ independently from a standard Gaussian. The rest of the coefficients are set to zero. All tests are performed at the $\alpha = 0.05$ level. For the implementation of tests that involve Monte Carlo sampling, $M = 200$, unless specified otherwise. Power estimates are computed as averages on 1000 simulations, and error bars represent plus or minus two standard errors. We always take $\sigma = 1$, and in each of the experiments, we specify the parameters n, d, k, A, k_1, k_2 , and ρ .

4.1 Power of the L -test

In Figure 2, we compare the powers of the L -test (“ L -test”), MC-free test from Section 2.3 (“MC-free”), group LASSO MC test from (5) (“gLASSO”), oracle test (“Oracle”), and F -test (“ F -test”) when we vary the signal of the tested coefficients, sparsity among null and non-null variables, and covariate correlations. For an additional benchmark, we also consider a naive extension of the ℓ -test from [35] that applies it separately to each of the tested coefficients with a Bonferroni correction (“Bonf- ℓ ”).

The L -test significantly outperforms the F -test in sparse settings under various signal strengths and covariate correlations. As expected, its performance is insensitive to sparsity among null coefficients (because the group LASSO does not regularize towards sparsity within a group) but weakens when sparsity increases among non-null coefficients. As discussed in Section 3.1, since $k = 10$, the L -test exhibits a large power gap with the oracle but outperforms the F -test by a substantial margin, and that margin is larger than when $k = 1$. For instance, the maximal power gain of the L -test in the top left panel (13.8 percentage points) is about 30% higher than that of the ℓ -test in its paper’s analogous $k = 1$ simulations (10.7 percentage points) [35].

In all settings, the L -test closely tracks the power of the group LASSO MC test (while, of course, being far more computationally efficient). The MC-free test performs similarly to the L -test except when correlations are high, since variance concentrates in a few directions and the premultiplier $A(\tilde{\beta}_{1:k}^\lambda, \lambda)$ exploits this geometry to give the L -test an additional boost; a similar effect occurs when dimensionality increases (see Figure 13 in Appendix E.4). The Bonferroni- ℓ -test generally performs substantially worse than all other tests, approaching our methods’ performance (and outperforming the F -test) only in the special case where the null is 1-sparse. Similar trends are evident in larger models, shown in Figure 12 of Appendix E.4.

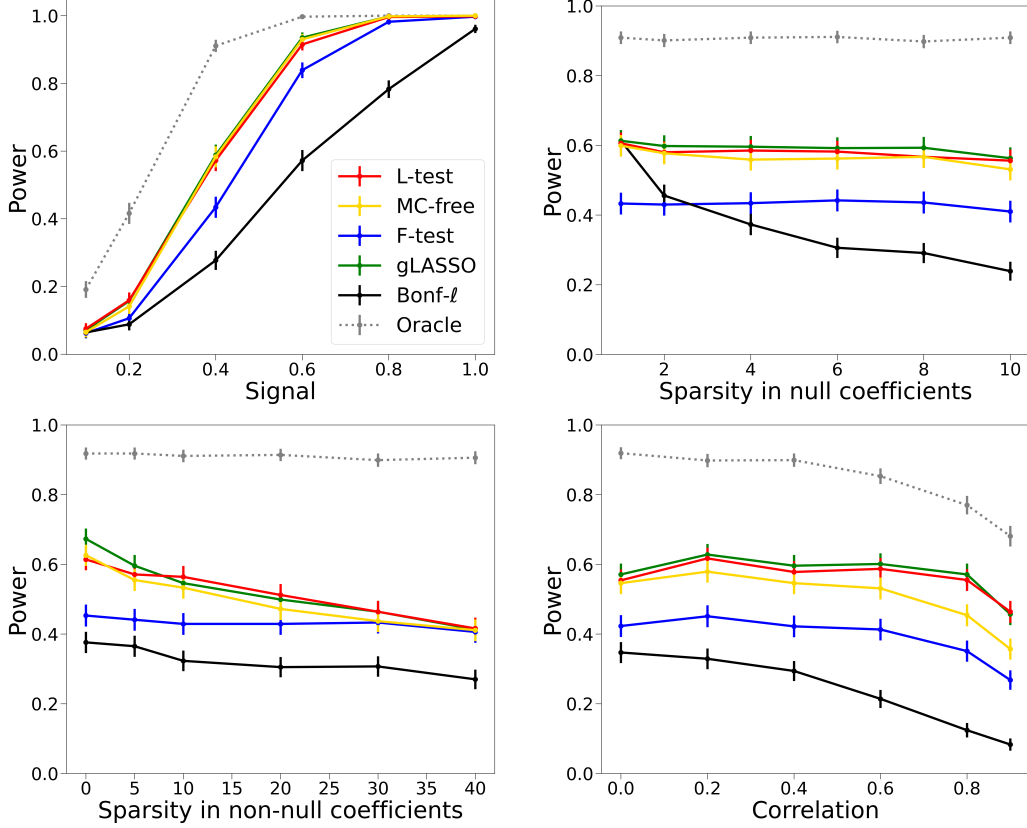


Figure 2: Comparison of test powers. In all panels, $n = 100$, $d = 50$, and $k = 10$. The top left takes $k_1 = k$, $k_2 = 4$, and $\rho = 0$ and varies A ; top right takes $A = 0.4$, $k_2 = 4$, and $\rho = 0$ and varies k_1 ; bottom left takes $A = 0.4$, $k_1 = 4$, and $\rho = 0$ and varies k_2 ; bottom right takes $A = 0.4$, $k_1 = 4$, and $k_2 = 4$ and varies ρ .

4.2 Block-orthogonal design setting

To confirm our understanding of the premultiplier $\mathbf{A}(\tilde{\beta}_{1:k}^\lambda, \lambda)$ in the L -test empirically, we constructed a related procedure, the PC-test, that leverages a similar geometry in $\mathbf{X}_{1:k}$ for power. This test selects the top r PC directions explaining 85% of the variance in $\mathbf{X}_{1:k}$ and then performs an F -test on the retained directions.

To isolate the impact of $\mathbf{A}(\tilde{\beta}_{1:k}^\lambda, \lambda)$, we orthogonalized $\mathbf{X}_{1:k}$ and $\mathbf{X}_{-1:k}$ so the recentering $\nu(\tilde{\beta}_{1:k}^\lambda)$ vanishes, and then assessed the L -test's similarity to the PC-test. As the left panel in Figure 3 shows, both outperform the F -test as ρ increases and variance concentrates in a few directions. Since both rely on alignment between $\beta_{1:k}$ and these PCs, they underperform when $\beta_{1:k}$ is anti-aligned (middle panel) but improve under maximum alignment (left vs. right panels). Further experiments in Appendix E.4 show that the MC-free test is more robust in the anti-aligned regime since it omits the premultiplier.

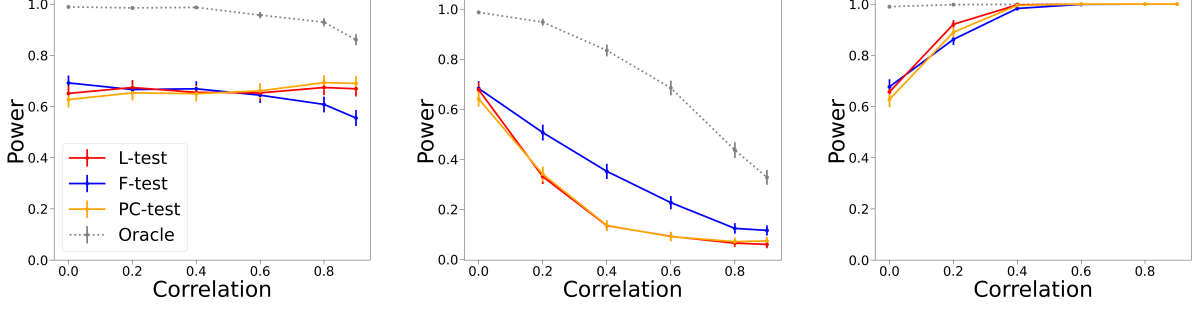


Figure 3: Similarity between L -test and PC-test under block-orthogonal design. In all panels, we orthogonalize $\mathbf{X}_{1:k}$ to $\mathbf{X}_{-1:k}$ by using the design $(\mathbf{X}_{1:k} - (\mathbf{I} - \mathbf{P}_{1:k})\mathbf{X}_{-1:k})$, followed by standardization. The left panel has the same setting as the bottom right panel of Figure 2. The middle panel changes only β to be dense with alternating signs, giving maximum anti-alignment with $\mathbf{1}_k$, the top PC under maximum correlation; the right panel instead uses a dense, nonnegative β , yielding maximum alignment.

4.3 Robustness

In this section, we assess the robustness of the L -test’s validity to linear model violations, similar to the approach taken in [35]. Four different types of violations are considered: heavy-tailed errors, skewed errors, heteroskedasticity, and non-linearity. In all of the experiments, we consider various sample sizes n and take $d = n/2$, $k = 3$, $k_1 = 0$, $k_2 = 1$, and $\rho = 0$ or $\rho = 0.5$. Figure 4 shows that the L -test exhibits similar robustness to the F -test for each model violation considered when $\rho = 0$. Appendix E.5 presents additional settings for each type of model violation—with Figure 4 showing the most extreme setting for each—and also performs the same experiments when $\rho = 0.5$.

5 Application to HIV drug resistance data

To demonstrate the power gains of our methodology on real data, we apply it to the task of discovering genes in the Human Immunodeficiency Virus Type 1 (HIV-1) that are associated with drug resistance. The data set from [34] consists of drug resistance measurements (the response) and genetic mutations (the covariates). It contains 16 data sets, each corresponding to a different drug.

We followed the same pre-processing steps outlined in [4], resulting in 16 regressions with n ranging between 328 and 842 and d ranging between 147 and 313. For each of the 16 regressions, we aggregated mutations to the gene level by grouping them based on their positions. Across the regressions, the group size k ranged from 1 to 11, and the F -test, L -test, and MC-free test were performed for $H_{1:k}$. In the single testing analysis, we also performed the group LASSO MC test in (5) to provide a benchmark for comparison.

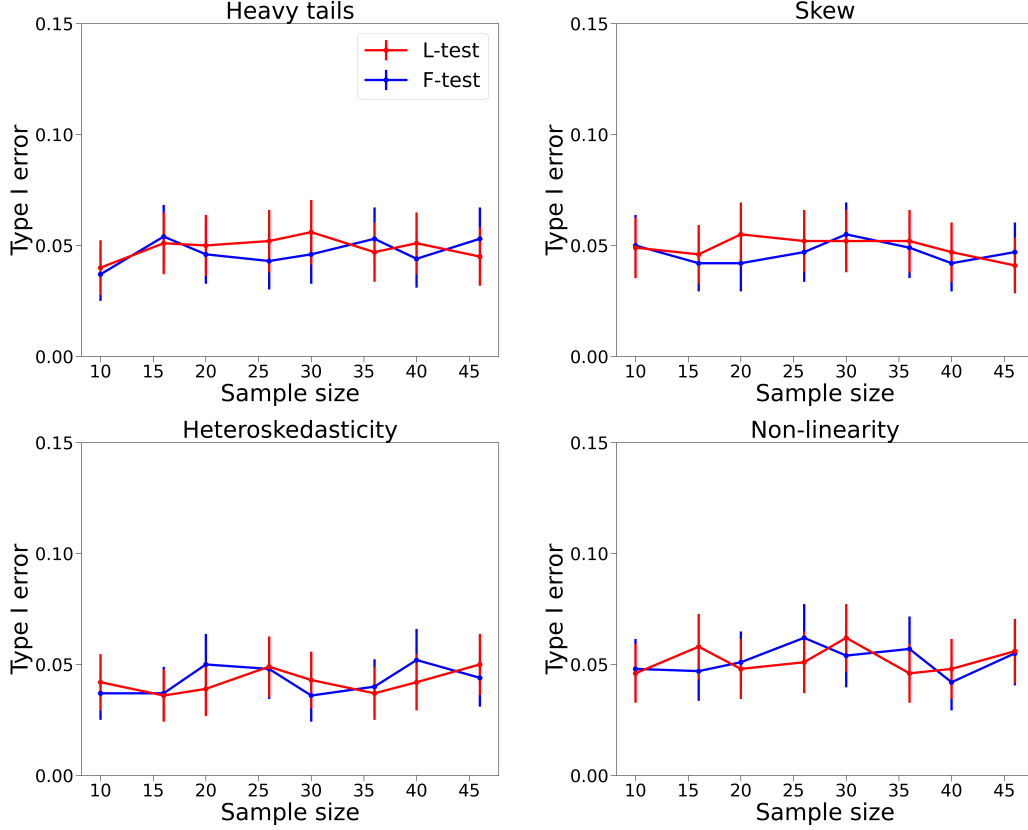


Figure 4: F -test and L -test Type I errors for $H_{1:k}$ under four different violations of model (1) when $\rho = 0$. In the top left panel, $\epsilon_i \stackrel{i.i.d.}{\sim} t_2$. In the top right, $\epsilon_i \stackrel{i.i.d.}{\sim} \text{Expo}(1)$ standardized with its theoretical mean and std. In the bottom left, $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ if the i -th row mean of $\mathbf{X}$ is less than 0 and $\epsilon_i \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 8)$ otherwise. In the bottom right, the design matrix has (i, j) -th entry $\text{sign}(X_{i,j})|X_{i,j}|^4$.

5.1 Single testing

All tests were performed at the $\alpha = 0.05$ level, and the leftmost panel in Figure 5 shows that the F -test and L -test had average powers of 24.30% and 25.88%, respectively, a 6.50% improvement, with the largest power gains occurring on moderate to larger group sizes. Although the group LASSO MC test performed slightly better than the L -test overall with an average power of 26.45%, this is mainly because it performed better on smaller group sizes (overrepresented in the middle panel); the right panel shows their performances were actually quite similar, indicating the L -test is a good approximation even in real data. The MC-free test demonstrated the best performance with an average power of 27.02% (a 11.19% improvement).

5.2 Multiple testing

To demonstrate the efficiency benefit of our MC-free test, we applied two multiple testing procedures: Holm at 5% and Benjamini–Hochberg (BHq) at 10%. As shown in the left panel

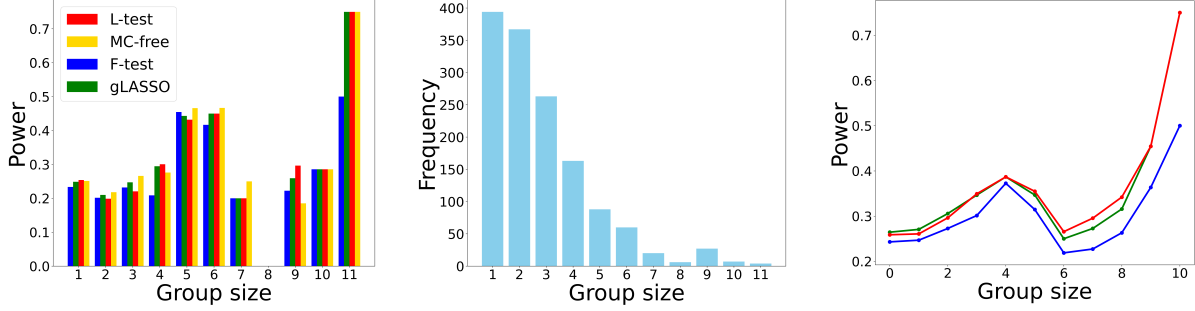


Figure 5: The left panel shows average powers by null group size across all 16 regressions. The middle panel shows the number of tests sorted by null group size across the regressions. The right panel shows average powers across all hypothesis tests that included subsets of coefficients of at least the size indicated on the x -axis.

of Figure 6, achieving a power gain with the L -test over the F -test with the Holm procedure requires more than 10^4 MC samples. Fewer MC samples suffices to obtain more power than the F -test with BHq, but even a similar number of samples is needed for the L -test to reach maximum power. This added precision, however, comes at the cost of substantially higher computation time, as the right panel shows. In contrast, the MC-free test delivers the precision needed for high power at computation times an order of magnitude faster.

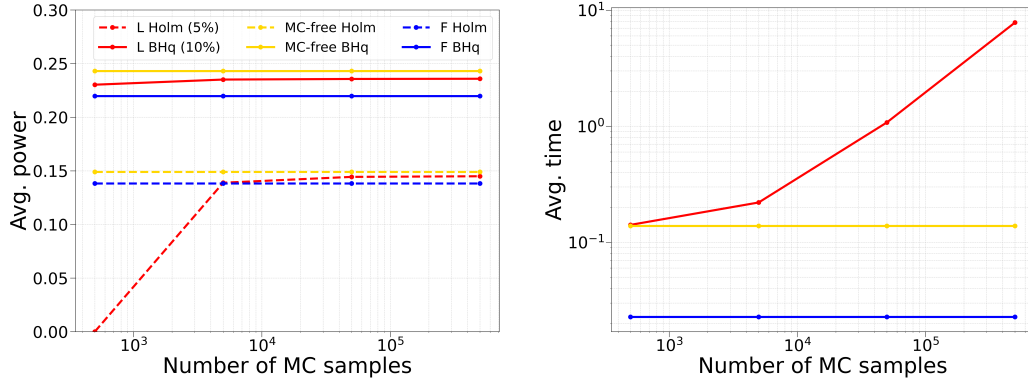


Figure 6: For each procedure (Holm and BHq), power was computed for each regression and then averaged across all 16. In order to be able to attribute the increased L -test power to the increased number of MC samples and not the randomness of the test, we performed this entire data analysis 50 times. The left panel shows the Holm and BHq powers averaged across these 50 iterations for each test. The right panel shows the average time associated with a single application of each test, computed over all tests from the 16 regressions and 50 iterations. Because the standard errors associated with the mean estimates in these plots were negligible, we chose to omit them.

6 Discussion

In closing, we identify some directions for future research:

1. *Generating confidence regions:* Note that we can extend the L -test to the more general hypothesis $H_{1:k}(\mathbf{a}) : \beta_{1:k} = \mathbf{a}$. This is because $\mathbf{y}$ satisfying $H_{1:k}(\mathbf{a})$ is equivalent to $\mathbf{y} - \mathbf{X}_{1:k}\mathbf{a}$ satisfying $H_{1:k}(\mathbf{0}) = H_{1:k}$, so we can just apply the L -test to $(\mathbf{y} - \mathbf{X}_{1:k}\mathbf{a}, \mathbf{X})$. Letting $p^{(1:k)}(\mathbf{a})$ denote the L -test p-value for $H_{1:k}(\mathbf{a})$, this generalization allows us to, in principle, generate a $100(1 - \alpha)\%$ L -test confidence region given by

$$\{\mathbf{a} \in \mathbb{R}^k : p^{(1:k)}(\mathbf{a}) > \alpha\}.$$

First, naively generating the region in this way would be incredibly computationally expensive, so finding an efficient construction is an interesting direction for further work. Second, it would be useful to study whether these regions could be constructed to admit an interpretable geometry, which would be especially useful in higher dimensions.

2. *Exploiting additional structure:* In Section 2.1 and Appendix E.4, we consider replacing the OLS in (4) with alternative estimators when the null group is sparse. It could be interesting to adapt our methodology to exploit additional structure—leveraging sparsity in the null group via the LASSO [39] or sparse group LASSO [36], or encoding known covariance structure within the tested group by using the original Mahalanobis-norm formulation of the group LASSO [3, 43], with $\mathbf{K}_j$ chosen to reflect that structure.
3. *Applying to Gaussian means and, asymptotically, to general parametric models:* In a Gaussian means problem, $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$ is observed for some known positive definite matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ is unknown. This model can always be converted to a Gaussian linear regression setup by taking $\mathbf{X} = \boldsymbol{\Sigma}^{-1/2}$ and $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}\mathbf{w}$ so that $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n)$. The standard method for testing $H_{1:k}$ in this case of known variance is the likelihood ratio test, which uses the test statistic:

$$\frac{1}{\sigma^2} (\|(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2 - \|(\mathbf{I} - \mathbf{P})\mathbf{y}\|^2) \sim \chi_k^2 \text{ under } H_{1:k}.$$

Following similar steps to the proof of Lemma C.4 that rewrites the F -test statistic, it can be shown that the above is equivalent to

$$\frac{1}{\sigma^2} \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2.$$

If we proceed in the same manner we obtained the L -test from the F -test by replacing $\hat{\boldsymbol{\beta}}_{1:k, \text{OLS}}$ with $\hat{\boldsymbol{\beta}}_{1:k}^\lambda$ and conditioning on $\mathbf{X}_{-1:k}^T \mathbf{y}$, we will obtain the conditional test that rejects for large values of the L -test statistic. So, we might analogously expect the L -test to outperform the likelihood ratio test in this setting when the true $\boldsymbol{\beta}_{-1:k}$ is sparse. One promising implication of this is that the L -test may yield more powerful asymptotic inference (under sparsity) for a vast class of multivariate parametric models that admit asymptotically multivariate Gaussian estimators coupled with consistent estimators for the covariance matrix (see [35] for a similar discussion and [29] for such an asymptotic result but for knockoffs).

7 Acknowledgments

LJ and SS were partially supported by DMS-2045981.

References

- [1] 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.
- [2] E. Arias-Castro, E. J. Candès, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.
- [3] S. Bakin. Adaptive regression and model selection in data mining problems, 1999.
- [4] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.
- [5] 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.
- [6] 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.
- [7] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics: Tenth World Congress*, 3:245–295, 2013.
- [8] J. G. Bryan and P. D. Hoff. Smaller p-values in genomics studies using distilled auxiliary information. *Biostatistics*, 24(1):193–208, 2021.
- [9] E. J. 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: Statistical Methodology*, 80(3):551–577, 2018.
- [10] C. Dai, B. Lin, X. Xing, and J. S. Liu. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 118(544):2503–2520, 2022.
- [11] Z. Ding and D. Dunson. Using prior information to boost power in correlation structure support recovery, 2021.
- [12] B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [13] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, 1925.
- [14] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection, 2014.
- [15] D. A. Freedman. Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228, 1981.
- [16] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso, 2010.

- [17] 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.
- [18] 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, 1993.
- [19] J. Hajek. Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33(3):1124–1147, 1962.
- [20] S. Han, N. Wang, Y. Guo, F. Tang, L. Xu, Y. Ju, and L. Shi. Application of sparse representation in bioinformatics. *Frontiers in Genetics*, 12(12), 2021.
- [21] P. D. Hoff. Smaller p-values via indirect information. *Journal of the American Statistical Association*, 117(539):1254–1269, 2022.
- [22] P. D. Hoff and C. Yu. Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, 13(1):94–119, 2019.
- [23] D. Huang and L. Janson. Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics*, 48(5):3021–3042, 2020.
- [24] L. A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, 43(5):1449–1458, 1972.
- [25] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(82):2869–2909, 2014.
- [26] 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.
- [27] J. Lee and Z. Ren. Boosting e-bh via conditional calibration, 2024.
- [28] L. Lei and P. J. Bickel. An assumption-free exact test for fixed-design linear models with exchangeable errors. *Biometrika*, 108(2):397–412, 2020.
- [29] X. Li and W. Fithian. Whiteout: when do fixed-x knockoffs fail?, 2021.
- [30] Y. Luo, W. Fithian, and L. Lei. Improving knockoffs with conditional calibration, 2022.
- [31] 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.
- [32] 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, 1938.
- [33] 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*, 86(1):122–154, 2024.

- [34] S. 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.
- [35] S. Sengupta and L. Janson. The ℓ -test: leveraging sparsity in the gaussian linear model for improved inference, 2025.
- [36] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
- [37] A. Spector and L. Janson. Powerful knockoffs via minimizing reconstructability. *The Annals of Statistics*, 50(1):252–276, 2022.
- [38] M. A. Stephens. Goodness-of-fit and sufficiency: exact and approximate tests. *Methodology and Computing in Applied Probability*, 14(3):785–791, 2012.
- [39] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
- [40] J. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29(2):614–623, 1958.
- [41] 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.
- [42] 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.
- [43] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.
- [44] 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.
- [45] H. Zhang and Y. Lin. Component selection and smoothing in smoothing spline analysis of variance models. *Institute of Statistics Mimeo Series*, 2003.
- [46] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.
- [47] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Extending the ℓ -test

A.1 Challenges of analytic p-value characterization

Theorem C.8 in Appendix C establishes a functional relationship between $\hat{\beta}_{1:k}^\lambda$ and $\mathbf{u}_{1:k}$ by extending the inverse characterization result of Theorem 2.1 in [35]. After a change of variables, this result yields the density of $\hat{\beta}_{1:k}^\lambda \mid \mathbf{S}^{(1:k)}$ under $H_{1:k}$ and therefore allows us to express the p-value of the conditional test that rejects for large values of (3) as follows. Letting $l > 0$ denote the observed test statistic,

$$\begin{aligned} P_{H_{1:k}} \left(\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda \right\| \geq l \mid \mathbf{S}^{(1:k)} \right) &= \int_{\{\mathbf{b}: \mathbf{b}^T \mathbf{M} \mathbf{b} \geq l^2\}} f_{\hat{\beta}_{1:k}^\lambda}^{H_{1:k}}(\mathbf{b} \mid \mathbf{S}^{(1:k)}) d\mathbf{b} \\ &= \int_{\{\mathbf{b}: \mathbf{b}^T \mathbf{M} \mathbf{b} \geq l^2\}} f_{\mathbf{u}_{1:k}}(f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b})) \cdot \left| \frac{\partial f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b})}{\partial \mathbf{b}} \right| d\mathbf{b}, \text{ where} \\ \mathbf{M} &= \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k}. \end{aligned}$$

In the second equality, note that although $f_{\mathbf{S}^{(1:k)}}^{-1}$ is nondifferentiable on the measure-zero set where the LASSO estimate inside (see Theorem C.8) changes its active set, this does not affect the validity of the change-of-variables formula for the integral, so we apply the Jacobian expression in the usual way. Evaluating this integral, however, is intractable. Even for moderately large k and a well-behaved integrand, evaluating a k -dimensional integral over an ellipsoidal region is computationally demanding. In our setting, the problem is compounded by the fact that evaluating the integrand is an extremely computationally intensive task that requires solving a LASSO to obtain $\Lambda_{1:k}(\mathbf{b}, \frac{\mathbf{b}}{\|\mathbf{b}\|})$ as well as computing the determinant of the matrix derivative $\frac{\partial f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b})}{\partial \mathbf{b}}$.

When $k = 1$, [35] circumvents this issue by showing that $\hat{\beta}_1^\lambda$ is a monotone increasing function of u_1 . This monotonicity allows the p-value to be mapped from $\hat{\beta}_1^\lambda$ -space to u_1 -space as follows

$$\begin{aligned} P_{H_{1:k}} \left(|\hat{\beta}_1^\lambda| \geq l \mid \mathbf{S}^{(1)} \right) &= P_{H_{1:k}}(\hat{\beta}_1^\lambda \geq l \mid \mathbf{S}^{(1)}) + P_{H_{1:k}}(\hat{\beta}_1^\lambda \leq -l \mid \mathbf{S}^{(1)}) \\ &= P(u_1 \geq f_{\mathbf{S}^{(1)}}^{-1}(l)) + P(u_1 \leq f_{\mathbf{S}^{(1)}}^{-1}(-l)), \end{aligned}$$

which reduces to evaluating the CDF of u_1 via its one-to-one mapping to a t -distribution, $\frac{\sqrt{n-du_1}}{\sqrt{1-u_1^2}} \sim t_{n-d}$. Crucially, this approach avoids the need to integrate a complex integrand that runs a LASSO.

When $k > 1$, the relationship between $\hat{\beta}_{1:k}^\lambda$ and $\mathbf{u}_{1:k}$ becomes substantially harder to characterize. The matrix derivative $\nabla \hat{\beta}_{1:k}^\lambda(\mathbf{u}_{1:k})$, evaluated explicitly in Proposition C.7, is quite complex and does not have any special structure (e.g. symmetry, diagonality, positive definiteness). Moreover, unlike in the $k = 1$ case where the monotonicity property of $\frac{\partial \hat{\beta}_1^\lambda}{\partial u_1}$ enabled a straightforward transformation from $\hat{\beta}_1^\lambda$ -space to u_1 -space, here it is unclear what kind of analytic information about $\hat{\beta}_{1:k}^\lambda$ as a function of $\mathbf{u}_{1:k}$ would even facilitate such a transformation, given the additional complexity of the p-value expression in $\hat{\beta}_{1:k}^\lambda$ -space when $k > 1$.

A.2 Recentered $\mathbf{u}_{1:k}$ -test

While the exact characterization of (3) conditional on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$ is restricted to $k = 1$, [35] also mentions a recentered u_1 -test that approximates the ℓ -test well and requires only a small, fixed number of LASSOs to perform. A multivariate analogue can be obtained by using Theorem C.8 and recentering $\mathbf{u}_{1:k}$ by the center of the ellipsoidal region $f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{0})$. This multivariate $\mathbf{u}_{1:k}$ -test, however, has two shortcomings. First, while its performance in many settings is similar to the test that uses the p-value in (5), it is not hard to find settings where it significantly underperforms. Second, even when it does perform well, if one wants to do multiple testing or for any other reason be able to obtain precise p-values, then one needs a large number of MC samples which, even without needing a LASSO for each, can become expensive. To make sparsity-informed coefficient group testing practical, we address both of these problems with new methods in this paper.

B Power of the L -test

In Section 3.2, we claimed that the premultiplier in the L -test statistic takes on different structure depending on how variance is distributed among the first k predictors $\mathbf{X}_{1:k}$. In this section, we formally establish those properties. Recall that the L -test premultiplier is given by

$$\phi = A(\tilde{\beta}_{1:k}^\lambda, \lambda) = \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \left[\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)}) \mathbf{X}_{1:k} + \frac{n\lambda}{\|\tilde{\beta}_{1:k}^\lambda\|} \mathbf{I} \right]^{-1} \mathbf{X}_{1:k}^T \mathbf{V}_{1:k}.$$

We study it in two cases—when the block $\mathbf{X}_{1:k}$ is orthonormal and when it has rank 1. These two setups correspond to opposite extreme situations where the predictors provide maximally distinct vs. completely redundant information.

Throughout the analysis below, we assume $\beta_{-1:k}$ is sparse so that we can expect the active set $\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)$ defined in Lemma 2.1 to be small. Because only a small fraction of the predictors are active in this case, the projection $\mathbf{P}_{\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)}$ acts in a very low-dimensional space, capturing only about $|\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)|/n$ of the variance, which is negligible when $|\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)|$ is small. In this sense,

$$\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}(\tilde{\beta}_{1:k}^\lambda)}) \mathbf{X}_{1:k} \approx \mathbf{X}_{1:k}^T \mathbf{X}_{1:k}, \quad (12)$$

an approximation we will make in both cases below.

B.1 Orthonormal case

Since the columns of $\mathbf{X}_{1:k}$ are orthonormal, $\mathbf{X}_{1:k}^T \mathbf{X}_{1:k} = \mathbf{I}$. Moreover, for $i = 1, \dots, k$,

$$\frac{(\mathbf{I} - \mathbf{P}_{-1:i}) \mathbf{X}_i}{\|(\mathbf{I} - \mathbf{P}_{-1:i}) \mathbf{X}_i\|} = \frac{\mathbf{X}_i}{\|\mathbf{X}_i\|} \implies \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} = \mathbf{I}.$$

Making approximation (12) and substituting these expressions for $\mathbf{X}_{1:k}^T \mathbf{X}_{1:k}$ and $\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}$, we get that

$$\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda) \approx \left(1 + \frac{n\lambda}{\|\tilde{\boldsymbol{\beta}}_{1:k}^\lambda\|} \right)^{-1} \mathbf{I}.$$

Thus, in the case that the variance is evenly distributed across predictors, the premultiplier has essentially no impact and the recentering vector plays the primary role.

B.2 Rank-1 case

Making approximation (12), taking $\mathbf{X}_{1:k} = \mathbf{v}\mathbf{b}^T$, and applying the Sherman-Morrison formula, we have for some constant $C \in \mathbb{R}$

$$\mathbf{A}(\tilde{\boldsymbol{\beta}}_{1:k}^\lambda, \lambda) \approx C \cdot \mathbf{V}_{1:k}^T \mathbf{v} (\mathbf{V}_{1:k}^T \mathbf{v})^T.$$

The matrix $\mathbf{V}_{1:k}^T \mathbf{v} (\mathbf{V}_{1:k}^T \mathbf{v})^T$ is symmetric and has rank 1, so it has one non-zero eigenvalue corresponding to eigenvector $\mathbf{V}_{1:k}^T \mathbf{v}$ and $k - 1$ zero eigenvalues. Writing out $\mathbf{u}_{1:k}$, we have that

$$\begin{aligned} \mathbf{u}_{1:k} &= \frac{1}{\hat{\sigma}_{1:k}} \mathbf{V}_{1:k}^T (\mathbf{X}_{1:k} \boldsymbol{\beta}_{1:k} + \boldsymbol{\epsilon}), \text{ by Lemma C.1} \\ &= \frac{1}{\hat{\sigma}_{1:k}} \mathbf{V}_{1:k}^T (\mathbf{v}\mathbf{b}^T \boldsymbol{\beta}_{1:k} + \boldsymbol{\epsilon}) \\ &= \frac{1}{\hat{\sigma}_{1:k}} ((\mathbf{b}^T \boldsymbol{\beta}_{1:k}) \mathbf{V}_{1:k}^T \mathbf{v} + \mathbf{V}_{1:k}^T \boldsymbol{\epsilon}). \end{aligned}$$

If the true $\boldsymbol{\beta}_{1:k}$ and the PC direction $\mathbf{b}$ that explains all the variance in $\mathbf{X}_{1:k}$ are well alligned (ie. the inner product $\mathbf{b}^T \boldsymbol{\beta}_{1:k}$ is large in magnitude), we can expect the eigenvector $\mathbf{V}_{1:k}^T \mathbf{v}$ to be approximately parallel to $\mathbf{u}_{1:k}$.

C Proofs of results

C.1 Sampling from $\mathbf{y} \mid \mathcal{S}^{(1:k)}$ under $H_{1:k}$

Lemma C.1. *For the linear model in (1), define $\hat{\mathbf{y}}_{1:k} = \mathbf{P}_{-1:k} \mathbf{y}$ and $\hat{\sigma}_{1:k}^2 = \|(\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y}\|^2$, and recall the matrix $\mathbf{V} \in \mathbb{R}^{n \times (n-d+k)}$ described following Equation (2). There exists a unique vector $\mathbf{u} \in \mathbb{R}^{n-d+k}$, such that $\|\mathbf{u}\| = 1$ and the following decomposition holds:*

$$\mathbf{y} = \hat{\mathbf{y}}_{1:k} + \hat{\sigma}_{1:k} \mathbf{V} \mathbf{u}.$$

Moreover,

$$\mathbf{u} \mid \mathcal{S}^{(1:k)} \stackrel{H_{1:k}}{\sim} \text{Unif}(\mathbb{S}^{n-d+k-1}).$$

Proof. Since $\mathbf{V}$ is full-column rank, there exists a unique vector $\mathbf{u}$ such that

$$\mathbf{V}\mathbf{u} = \frac{\mathbf{y} - \hat{\mathbf{y}}_{1:k}}{\hat{\sigma}_{1:k}}.$$

In particular, $\mathbf{u} = \mathbf{V}^T \left(\frac{\mathbf{y} - \hat{\mathbf{y}}_{1:k}}{\hat{\sigma}_{1:k}} \right)$ with $\|\mathbf{u}\| = \|\mathbf{V}\mathbf{u}\| = \left\| \frac{\mathbf{y} - \hat{\mathbf{y}}_{1:k}}{\hat{\sigma}_{1:k}} \right\| = \frac{1}{\hat{\sigma}_{1:k}} \|\mathbf{y} - \hat{\mathbf{y}}_{1:k}\| = 1$ since $\mathbf{V}$ is an orthogonal transformation and hence preserves norms. This proves the decomposition. The conditional distribution of $\mathbf{u}$ follows from the fact that

$$\mathbf{y} \mid \mathbf{S}^{(1:k)} \stackrel{H_{1:k}}{\sim} \hat{\mathbf{y}}_{1:k} + \hat{\sigma}_{1:k} \mathbf{V}\mathbf{u}, \text{ where } \mathbf{u} \sim \text{Unif}(\mathbb{S}^{n-d+k-1}).$$

The proof of this result is identical to that of Proposition E.1 in [30], except we use $\mathbf{P}_{-1:k}$, $(\mathbf{I} - \mathbf{P}_{-1:k})$, and $\mathbf{V}$ accordingly. $\square$

C.2 Conditional F -test statistic

Lemma C.2. *Let $\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}}$ denote the first k OLS estimate coefficients. The F -test for $H_{1:k}$ in model (1) is equivalent to the test that rejects for large values of*

$$\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \right\|$$

conditional on $\mathbf{S}^{(1:k)}$. This latter test is also equivalent to the conditional test rejecting for large values of $\|\mathbf{u}_{1:k}\|$.

Proof. The proof of Lemma C.2 relies on a series of sub-lemmas. The first gives a representation for a subvector of the OLS that extends the convenient representation for one OLS coefficient $\hat{\boldsymbol{\beta}}_{j,\text{OLS}} = \frac{\mathbf{X}_j^T (\mathbf{I} - \mathbf{P}_{-j}) \mathbf{y}}{\|(\mathbf{I} - \mathbf{P}_{-j}) \mathbf{X}_j\|^2}$ used in [14].

Lemma C.3. *The first k OLS coefficients can be written as*

$$\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} = \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y},$$

where $\mathbf{S} = \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k}$.

Proof of Lemma C.3. We start by decomposing the OLS estimate into block matrices as follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{OLS}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \left(\begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} \begin{pmatrix} \mathbf{X}_{1:k} & \mathbf{X}_{-1:k} \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} \mathbf{y} \\ &= \begin{pmatrix} \mathbf{X}_{1:k}^T \mathbf{X}_{1:k} & \mathbf{X}_{1:k}^T \mathbf{X}_{-1:k} \\ \mathbf{X}_{-1:k}^T \mathbf{X}_{1:k} & \mathbf{X}_{-1:k}^T \mathbf{X}_{-1:k} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} \mathbf{y} \\ &= \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} \mathbf{y} \text{ [where } \mathbf{A}, \mathbf{B}, \mathbf{C} \text{ and } \mathbf{D} \text{ represent the block-matrices they are replacing]} \\ &= \begin{pmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{G} & \mathbf{H} \end{pmatrix} \begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} \mathbf{y} \text{ (say)} \\ &= \begin{pmatrix} \mathbf{E} \mathbf{X}_{1:k}^T + \mathbf{F} \mathbf{X}_{-1:k}^T \\ \mathbf{G} \mathbf{X}_{1:k}^T + \mathbf{H} \mathbf{X}_{-1:k}^T \end{pmatrix} \mathbf{y}. \end{aligned}$$

In particular, letting $\mathbf{S} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C} = \mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}$ denote the Schur complement of $\mathbf{D}$, then $\mathbf{E} = \mathbf{S}^{-1}$ and $\mathbf{F} = -\mathbf{S}^{-1}\mathbf{B}\mathbf{D}^{-1}$. Note that the block inversion is well-defined because $\mathbf{D}$ and $\mathbf{S}$ are both invertible. The invertibility of $\mathbf{D}$ follows from the fact that $\mathbf{X}_{-1:k}$ is also full column-rank, while that of $\mathbf{S}$ follows from the fact that so is $(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}$. To see the latter fact, note that for any vector $\mathbf{v}$, $(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}\mathbf{v} = \mathbf{0}$ implies that $\mathbf{X}_{1:k}\mathbf{v}$ lies in the column-space of $\mathbf{X}_{-1:k}$. Because the columns of $\mathbf{X}_{1:k}$ are linearly independent from that of $\mathbf{X}_{-1:k}$, this can only happen if $\mathbf{v} = \mathbf{0}$; thereby showing that $(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}$ is full column-rank.

Next, from the above decomposition of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$, it follows that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} &= (\mathbf{E}\mathbf{X}_{1:k}^T + \mathbf{F}\mathbf{X}_{-1:k}^T)\mathbf{y} \\ &= (\mathbf{S}^{-1}\mathbf{X}_{1:k}^T - \mathbf{S}^{-1}\mathbf{X}_{1:k}^T\mathbf{X}_{-1:k}(\mathbf{X}_{-1:k}^T\mathbf{X}_{-1:k})^{-1}\mathbf{X}_{-1:k}^T)\mathbf{y} \\ &= \mathbf{S}^{-1}\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}.\end{aligned}$$

□

Lemma C.4. *The F -test statistic for $H_{1:k}$ in model (1) can be written as*

$$F := \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \right\|^2 / k}{\left(\hat{\sigma}_{1:k}^2 - \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \right\|^2 \right) / (n - d)}. \quad (13)$$

Proof of Lemma C.4. We start by manipulating the classic F -test statistic as follows

$$\begin{aligned}F &= \frac{(\|\mathbf{y} - \mathbf{P}_{-1:k}\mathbf{y}\|^2 - \|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2)/k}{\|\mathbf{y} - \mathbf{P}\mathbf{y}\|^2/(n - d)} \\ &= \frac{\|\mathbf{P}\mathbf{y} - \mathbf{P}_{-1:k}\mathbf{y}\|^2/k}{\|(\mathbf{I} - \mathbf{P})\mathbf{y}\|^2/(n - d)} \\ &= \frac{\|\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2/k}{\|(\mathbf{I} - \mathbf{P})\mathbf{y}\|^2/(n - d)} \\ &= \frac{\|\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2/k}{\|(\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2/(n - d)} \\ &= \frac{\|\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2/k}{(\hat{\sigma}_{1:k}^2 - \|\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}\|^2)/(n - d)}.\end{aligned}$$

By expanding the projection matrix $\mathbf{P}$, decomposing $\mathbf{X}^T$ into block matrices, and using

Lemma C.3, we can rewrite $\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}$ in terms of $\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}}$ as follows

$$\begin{aligned}
\mathbf{P}(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y} &= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} \mathbf{X}_{1:k}^T \\ \mathbf{X}_{-1:k}^T \end{pmatrix} (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y} \\
&= \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \begin{pmatrix} \mathbf{S}\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \\ \mathbf{0} \end{pmatrix} \\
&= (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1})_{1:k} \mathbf{S}\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \\
&= (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1})_{1:k} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \\
&= (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1})_{1:k} \mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \\
&= \mathbf{P} \mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}},
\end{aligned}$$

where we use the fact that $\mathbf{V} \mathbf{V}^T = (\mathbf{I} - \mathbf{P}_{-1:k})$ since the columns of $\mathbf{V}$ form an orthonormal basis for the orthogonal complement of the column space of $\mathbf{X}_{-1:k}$. The definition of $\mathbf{V}$ implies that the columns of $\mathbf{V}_{-1:k}$ are orthogonal to the column space of $\mathbf{X}$, so $\mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k} = \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k}$. We also have that $\mathbf{P} \mathbf{V}_{1:k} = \mathbf{V}_{1:k}$ since the columns of $\mathbf{V}_{1:k}$ lie in the column space of $\mathbf{X}$. Thus,

$$\mathbf{P} \mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} = \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}}.$$

Substituting this expression into the F -test statistic and noting that $\mathbf{V}_{1:k}$ is orthogonal and hence norm-preserving, we find that

$$F = \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \right\|^2 / k}{\left(\hat{\sigma}_{1:k}^2 - \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} \right\|^2 \right) / (n - d)}.$$

□

Lemma C.5. *The first k OLS coefficients can be written in terms of the first k coefficients of $\mathbf{u}$ from (6) as follows*

$$\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} = \hat{\sigma}_{1:k} (\mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1} \mathbf{u}_{1:k}.$$

Proof of Lemma C.5. By Lemma C.3, we have that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} &= \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y} \\
&= \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) (\hat{\mathbf{y}}_{1:k} + \hat{\sigma}_{1:k} \mathbf{V} \mathbf{u}), \text{ by Lemma C.1} \\
&= \hat{\sigma}_{1:k} \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{V} \mathbf{u} \\
&= \hat{\sigma}_{1:k} \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{V} \mathbf{u},
\end{aligned}$$

because the columns of $\mathbf{V}$ lie in the space orthogonal to the columns of $\mathbf{X}_{-1:k}$. Since the columns of $\mathbf{V}_{-1:k}$ are orthogonal to the column space of $\mathbf{X}$, $\mathbf{X}_{1:k}^T \mathbf{V} \mathbf{u} = \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{u}_{1:k}$. Moreover, $\mathbf{S}^{-1} = (\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k})^{-1} = (\mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k})^{-1} = (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1}$. Thus,

$$\hat{\boldsymbol{\beta}}_{1:k,\text{OLS}} = \hat{\sigma}_{1:k} (\mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{u}_{1:k} = \hat{\sigma}_{1:k} (\mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1} \mathbf{u}_{1:k}.$$

□

Lemma C.6. *The F-test statistic is independent of $\mathbf{S}^{(1:k)}$ under $H_{1:k}$.*

Proof of Lemma C.6. By Lemma C.4, the F-test statistic can be written as

$$\begin{aligned} F &= \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2 / k}{\left(\hat{\sigma}_{1:k}^2 - \left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2 \right) / (n - d)} \\ &= \frac{\frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2}{\hat{\sigma}_{1:k}^2} / k}{\left(1 - \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2}{\hat{\sigma}_{1:k}^2} \right) / (n - d)}. \end{aligned}$$

It suffices to show that $\frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2}{\hat{\sigma}_{1:k}^2}$ is independent of $\mathbf{S}^{(1:k)}$ under $H_{1:k}$, which we will demonstrate by showing that this expression's conditional distribution is the same as its marginal distribution under $H_{1:k}$. By Lemma C.5 and C.1, under $H_{1:k}$,

$$\left. \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2}{\hat{\sigma}_{1:k}^2} \right| \mathbf{S}^{(1:k)} \sim \left\| \mathbf{u}_{1:k} \right\|^2 \mid \mathbf{S}^{(1:k)} \sim \left\| \tilde{\mathbf{u}}_{1:k} \right\|^2,$$

where $\tilde{\mathbf{u}} \sim \text{Unif}(\mathbb{S}^{n-d+k-1})$.

Next, under $H_{1:k}$, we can write

$$\mathbf{y} = \mathbf{X}_{-1:k} \boldsymbol{\beta}_{-1:k} + \boldsymbol{\epsilon}, \text{ for some } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Recall $(\mathbf{I} - \mathbf{P}_{-1:k}) = \mathbf{V} \mathbf{V}^T$ and $\mathbf{S}^{-1} = (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1}$. Since the columns of $\mathbf{V}_{-1:k}$ are orthogonal to the column space of $\mathbf{X}$, we also have $\mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T = \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T$. Thus,

$$\begin{aligned} \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \hat{\boldsymbol{\beta}}_{1:k, \text{OLS}} \right\|^2}{\hat{\sigma}_{1:k}^2} &= \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y} \right\|^2}{\hat{\sigma}_{1:k}^2}, \text{ by Lemma C.3} \\ &= \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T \mathbf{y} \right\|^2}{\left\| \mathbf{V} \mathbf{V}^T \mathbf{y} \right\|^2} \\ &= \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{X}_{1:k} \mathbf{S}^{-1} \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{y} \right\|^2}{\left\| \mathbf{V}^T \mathbf{y} \right\|^2} \\ &= \frac{\left\| \mathbf{V}_{1:k}^T \mathbf{y} \right\|^2}{\left\| \mathbf{V}^T \mathbf{y} \right\|^2} \\ &= \left\| \mathbf{u}_{1:k}^* \right\|^2, \end{aligned}$$

where $\mathbf{u}^* = \frac{\mathbf{V}_{1:k}^T \mathbf{y}}{\left\| \mathbf{V}^T \mathbf{y} \right\|} \sim \text{Unif}(\mathbb{S}^{n-d+k-1})$ since $\mathbf{V}^T \mathbf{y} = \mathbf{V}^T \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n-d+k})$. Thus,

$$F \perp \mathbf{S}^{(1:k)} \text{ under } H_{1:k}.$$

□

By Lemmas C.4 and C.6, the F -test is equivalent to the test that rejects for large values of (13) conditional on $\mathbf{S}^{(1:k)}$. Conditional on $\mathbf{S}^{(1:k)}$, however, (13) is a monotonically increasing function of (4). Thus, the F -test is equivalent to the test that rejects for large values of (4) conditional on $\mathbf{S}^{(1:k)}$, giving the result. The last part of the lemma can easily be seen by substituting the expression for $\hat{\beta}_{1:k,\text{OLS}}$ from Lemma C.5 into (4). $\square$

C.3 $\hat{\beta}_{1:k}^\lambda$ as a function of $\mathbf{u}_{1:k}$

In an attempt to demonstrate why it is difficult to analytically study $\hat{\beta}_{1:k}^\lambda$ as a function of $\mathbf{u}_{1:k}$, we state and prove a result below that specifies the matrix derivative of $\hat{\beta}_{1:k}^\lambda$ with respect to $\mathbf{u}_{1:k}$. Different elements of the matrix derivative shown in Proposition C.7 exhibit nice properties. For example, $\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{\mathcal{A}(\mathbf{u}_{1:k})})\mathbf{X}_{1:k}$ is positive definite, the inverted matrix is symmetric, and $\mathbf{X}_{1:k}^T\mathbf{V}_{1:k}$ is a diagonal matrix with positive diagonal entries. Nevertheless, a study of the composite matrix reveals that it does not exhibit any clear useful properties (ie. symmetric, positive definite, sparsity, low-rank structure) that would lend insight into $\hat{\beta}_{1:k}^\lambda$'s behavior as a function of $\mathbf{u}_{1:k}$.

Proposition C.7. *Suppose that $\mathbf{u}_{1:k} \in \mathbb{R}^k$ is such that $\hat{\beta}_{1:k}^\lambda(\mathbf{u}_{1:k}) \neq \mathbf{0}$. Then,*

$$\nabla \hat{\beta}_{1:k}^\lambda(\mathbf{u}_{1:k}) = \hat{\sigma}_{1:k} \left(\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{\mathcal{A}(\mathbf{u}_{1:k})})\mathbf{X}_{1:k} + \frac{n\lambda}{\|\hat{\beta}_{1:k}^\lambda\|} \left(\mathbf{I} - \frac{\hat{\beta}_{1:k}^\lambda(\hat{\beta}_{1:k}^\lambda)^T}{\|\hat{\beta}_{1:k}^\lambda\|^2} \right) \right)^{-1} \mathbf{X}_{1:k}^T \mathbf{V}_{1:k}, \text{ where}$$

$$\mathcal{A}(\mathbf{u}_{1:k}) = \left\{ i \in \{k+1, \dots, d\} : \hat{\beta}_i^\lambda(\mathbf{u}_{1:k}) \neq 0 \right\}$$

Proof. To show how $\hat{\beta}_{1:k}^\lambda$ can be written as a function of $\mathbf{u}_{1:k}$, we start by first rewriting the objective function associated with the group LASSO in (2) using a similar approach to that in the proof of Theorem 2.1 in [35]. Recalling that $\mathbf{y} = \hat{\mathbf{y}}_{1:k} + \hat{\sigma}_{1:k}\mathbf{V}\mathbf{u}$ from (6) and letting $\hat{\mathbf{e}}_{1:k} = \hat{\sigma}_{1:k}\mathbf{V}\mathbf{u}$, we can decompose the error term $\hat{\mathbf{e}}_{1:k}$ as $\hat{\mathbf{e}}_{1:k} = \hat{\mathbf{e}}_{\parallel} + \hat{\mathbf{e}}_{\perp}$, where

$$\begin{aligned} \hat{\mathbf{e}}_{\parallel} &= (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}(\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k})^{-1}\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k})\hat{\mathbf{e}}_{1:k} \\ &= (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}\hat{\beta}_{1:k,\text{OLS}}, \text{ by Lemma C.3} \\ \hat{\mathbf{e}}_{\perp} &= (\mathbf{I} - (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}(\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k})^{-1}\mathbf{X}_{1:k}^T(\mathbf{I} - \mathbf{P}_{-1:k}))\hat{\mathbf{e}}_{1:k} \end{aligned}$$

Note that $\hat{\mathbf{e}}_{\parallel}$ is the projection of $\hat{\mathbf{e}}_{1:k}$ onto the column span of $(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}$ and $\hat{\mathbf{e}}_{\perp}$ is perpendicular to the column space of $\mathbf{X}$. The objective function in (2) can thus be written as

$$\begin{aligned} \|\mathbf{y} - \mathbf{x}\beta\|^2 + 2n\lambda\|\beta_{1:k}\|_2 + 2n\lambda\|\beta_{-1:k}\|_1 &= \|\hat{\mathbf{y}}_{1:k} + \hat{\mathbf{e}}_{\parallel} + \hat{\mathbf{e}}_{\perp} - \mathbf{X}\beta\|^2 + 2n\lambda\|\beta_{1:k}\|_2 + 2n\lambda\|\beta_{-1:k}\|_1 \\ &= \|\hat{\mathbf{y}}_{1:k} - \mathbf{P}_{-1:k}\mathbf{X}_{1:k}\beta_{1:k} - \mathbf{X}_{-1:k}\beta_{-1:k}\|^2 + \|\hat{\mathbf{e}}_{\perp}\|^2 \\ &\quad + \left\| (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}(\hat{\beta}_{1:k,\text{OLS}} - \beta_{1:k}) \right\|^2 + 2n\lambda\|\beta_{1:k}\|_2 \\ &\quad + 2n\lambda\|\beta_{-1:k}\|_1 \\ &= \tilde{f}(\beta; \mathbf{y}, \mathbf{X}) + \left\| (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}(\hat{\beta}_{1:k,\text{OLS}} - \beta_{1:k}) \right\|^2, \end{aligned}$$

where $\tilde{f}$ denotes the expression it is replacing. Defining the function

$$\begin{aligned}\hat{\beta}^\lambda(\mathbf{a}) &= \arg \min_{\beta \in \mathbb{R}^d} \left(\tilde{f}(\beta; \mathbf{y}, \mathbf{X}) + \|(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}(\mathbf{a} - \beta_{1:k})\|^2 \right) \\ &= \arg \min_{\beta \in \mathbb{R}^d} U(\mathbf{a}, \beta),\end{aligned}$$

we see that $\hat{\beta}_{1:k}^\lambda = \hat{\beta}_{1:k}^\lambda(\hat{\beta}_{1:k, \text{OLS}}) = \hat{\beta}_{1:k}^\lambda(\hat{\sigma}_{1:k}(\mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1} \mathbf{u}_{1:k})$ by Lemma C.5. Then, by the chain rule, we have that

$$\nabla \hat{\beta}_{1:k}^\lambda(\mathbf{u}_{1:k}) = \nabla \hat{\beta}_{1:k}^\lambda(\hat{\beta}_{1:k, \text{OLS}}) \cdot \nabla \hat{\beta}_{1:k, \text{OLS}}(\mathbf{u}_{1:k}).$$

$\nabla \hat{\beta}_{1:k, \text{OLS}}(\mathbf{u}_{1:k}) = \hat{\sigma}_{1:k}(\mathbf{V}_{1:k}^T \mathbf{X}_{1:k})^{-1}$, so all we need is $\nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a})$. We start by establishing that $\hat{\beta}^\lambda(\mathbf{a})$ is continuous in $\mathbf{a}$ by invoking Lemma 3 from [35], the proof of which relies on an application of Berge's maximum theorem. The first assumption of this lemma is satisfied since $U(\mathbf{a}, \beta)$ is a continuous function of its arguments. For the second assumption, let $\mathbf{a}^* \in \mathbb{R}^k$, $d > 0$, and $\epsilon = \sup_{\mathbf{a} \in B(\mathbf{a}^*, d)} U(\mathbf{a}, \mathbf{0})$, and note that $U(\mathbf{a}, \beta) \geq \tilde{f} \geq 2n\lambda(\|\beta_{1:k}\|_2 + \|\beta_{-1:k}\|_1) \geq 2n\lambda\|\beta\|$. Take $M = \frac{\epsilon}{2n\lambda}$ and note that if $\|\beta\| > M$, then $\inf_{\mathbf{a} \in B(\mathbf{a}^*, d)} U(\mathbf{a}, \beta) \geq 2n\lambda\|\beta\| > \epsilon$, as desired. It follows that $\hat{\beta}^\lambda(\mathbf{a})$ is continuous in $\mathbf{a}$.

Assume that $\mathbf{u}_{1:k}$ is such that $\hat{\beta}_{1:k}^\lambda \neq \mathbf{0}$, and let $\mathcal{A}(\mathbf{a}) = \{i \in \{k+1, \dots, d\} : \hat{\beta}_i^\lambda(\mathbf{a}) \neq 0\}$. Using $[k]$ as shorthand for $\{1, \dots, k\}$, since $\hat{\beta}^\lambda(\mathbf{a})$ minimizes U , it follows that $\hat{\beta}_{[k] \cup \mathcal{A}(\mathbf{a})}^\lambda$ is a minimizer of $U(\mathbf{a}, \gamma)$, where γ is a vector of length $k + |\mathcal{A}(\mathbf{a})|$. Here we abuse notation slightly: $U(\mathbf{a}, \beta)$ is originally defined with $\beta \in \mathbb{R}^p$ as its second argument, while γ denotes the subvector of parameters indexed by $[k] \cup \mathcal{A}(\mathbf{a})$. Thus, $U(\mathbf{a}, \gamma)$ is to be interpreted as evaluating U at the vector in $\mathbb{R}^p$ that agrees with γ on $[k] \cup \mathcal{A}(\mathbf{a})$ and is zero elsewhere.

Because all of the entries of $\hat{\beta}_{[k] \cup \mathcal{A}(\mathbf{a})}^\lambda$ are non-zero, $U(\mathbf{a}, \gamma)$ is differentiable in γ at $\hat{\beta}_{[k] \cup \mathcal{A}(\mathbf{a})}^\lambda$. The first-order conditions indicate that for $i \in \mathcal{A}(\mathbf{a})$,

$$\begin{aligned}\frac{\partial U}{\partial \gamma_i} \Big|_{\gamma = \hat{\beta}_{[k] \cup \mathcal{A}(\mathbf{a})}^\lambda} &= 0 \\ \implies -\mathbf{X}_i^T \left(\hat{\mathbf{y}}_{1:k} - \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda(\mathbf{a}) - \mathbf{X}_{\mathcal{A}(\mathbf{a})} \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \right) + n\lambda \text{sign}(\hat{\beta}_i^\lambda(\mathbf{a})) &= 0.\end{aligned}$$

Thus,

$$-\mathbf{X}_{\mathcal{A}(\mathbf{a})}^T \left(\hat{\mathbf{y}}_{1:k} - \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda(\mathbf{a}) - \mathbf{X}_{\mathcal{A}(\mathbf{a})} \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \right) + n\lambda \text{sign}(\hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a})) = \mathbf{0}.$$

Note that $\hat{\beta}^\lambda(\mathbf{a})$ is continuous in the neighborhood of $\mathbf{a}$ in which the active set $\mathcal{A}(\mathbf{a})$ does not change and hence the signs of the active variables also do not change. So, differentiating both sides of the equation above with respect to $\mathbf{a}$ yields

$$\begin{aligned}\mathbf{X}_{\mathcal{A}(\mathbf{a})}^T \left(\mathbf{P}_{-1:k} \mathbf{X}_{1:k} \nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a}) + \mathbf{X}_{\mathcal{A}(\mathbf{a})} \nabla \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \right) &= \mathbf{0} \\ \implies \nabla \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) &= -(\mathbf{X}_{\mathcal{A}(\mathbf{a})}^T \mathbf{X}_{\mathcal{A}(\mathbf{a})})^{-1} \mathbf{X}_{\mathcal{A}(\mathbf{a})}^T \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a})\end{aligned}\tag{14}$$

Similarly, for $i \in [k]$,

$$\begin{aligned}
& \left. \frac{\partial U}{\partial \gamma_i} \right|_{\gamma = \hat{\beta}_{[k] \cup \mathcal{A}(\mathbf{a})}^\lambda} = 0 \\
& \implies -\mathbf{X}_i^T \left(\hat{\mathbf{y}}_{1:k} - \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda(\mathbf{a}) - \mathbf{X}_{\mathcal{A}(\mathbf{a})} \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \right) \\
& \quad + n\lambda \frac{\hat{\beta}_i^\lambda(\mathbf{a})}{\|\hat{\beta}_{1:k}^\lambda(\mathbf{a})\|} - \mathbf{X}_i^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} \left(\mathbf{a} - \hat{\beta}_{1:k}^\lambda(\mathbf{a}) \right) = 0 \\
& \implies -\mathbf{X}_{1:k}^T \left(\hat{\mathbf{y}}_{1:k} - \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \hat{\beta}_{1:k}^\lambda(\mathbf{a}) - \mathbf{X}_{\mathcal{A}(\mathbf{a})} \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \right) \\
& \quad + n\lambda \frac{\hat{\beta}_{1:k}^\lambda(\mathbf{a})}{\|\hat{\beta}_{1:k}^\lambda(\mathbf{a})\|} - \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} \left(\mathbf{a} - \hat{\beta}_{1:k}^\lambda(\mathbf{a}) \right) = 0
\end{aligned}$$

Differentiating again with respect to $\mathbf{a}$ yields

$$\begin{aligned}
& \mathbf{X}_{1:k}^T \mathbf{P}_{-1:k} \mathbf{X}_{1:k} \nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a}) + \mathbf{X}_{1:k}^T \mathbf{X}_{\mathcal{A}(\mathbf{a})} \nabla \hat{\beta}_{\mathcal{A}(\mathbf{a})}^\lambda(\mathbf{a}) \\
& \quad + n\lambda \frac{1}{\|\hat{\beta}_{1:k}^\lambda(\mathbf{a})\|} \left(\mathbf{I} - \frac{\hat{\beta}_{1:k}^\lambda(\mathbf{a}) \hat{\beta}_{1:k}^\lambda(\mathbf{a})^T}{\|\hat{\beta}_{1:k}^\lambda(\mathbf{a})\|^2} \right) \nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a}) \\
& \quad - \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} + \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} \nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a}) = 0
\end{aligned} \tag{15}$$

Substituting (14) into (15) and rearranging, we get

$$\nabla \hat{\beta}_{1:k}^\lambda(\mathbf{a}) = \left(\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}(\mathbf{a})}) \mathbf{X}_{1:k} + \frac{n\lambda}{\|\hat{\beta}_{1:k}^\lambda\|} \left(\mathbf{I} - \frac{\hat{\beta}_{1:k}^\lambda (\hat{\beta}_{1:k}^\lambda)^T}{\|\hat{\beta}_{1:k}^\lambda\|^2} \right) \right)^{-1} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k}$$

Note that $\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{X}_{1:k} = \mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T \mathbf{X}_{1:k} = \mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T \mathbf{X}_{1:k}$ since $\mathbf{V} \mathbf{V}^T = (\mathbf{I} - \mathbf{P}_{-1:k})$ and the columns of $\mathbf{V}_{-1:k}$ are orthogonal to the column space of $\mathbf{X}$ by definition of $\mathbf{V}$. So, combining the above with $\nabla \hat{\beta}_{1:k, \text{OLS}}(\mathbf{u}_{1:k})$ delivers the result. $\square$

C.4 Inverse of $\hat{\beta}_{1:k}^\lambda$ as a function of $\mathbf{u}_{1:k}$

Theorem C.8. For any $\mathbf{b} \in \mathbb{R}^k$ and $\epsilon \in B_k(\mathbf{0}, 1)$, the unit ball centered at the origin in $\mathbb{R}^k$, define the functions

$$\hat{\beta}_{-1:k}^\lambda(\mathbf{b}) = \arg \min_{\beta_{-1:k} \in \mathbb{R}^{d-k}} \left(\frac{1}{2n} \|\mathbf{y} - \mathbf{X}_{1:k} \mathbf{b} - \mathbf{X}_{-1:k} \beta_{-1:k}\|_2^2 + \lambda \|\beta_{-1:k}\|_1 \right) \tag{16}$$

$$\Lambda_{1:k}(\mathbf{b}, \epsilon) = \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \left(-\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{1:k} \mathbf{b} - \mathbf{X}_{-1:k} \hat{\beta}_{-1:k}^\lambda(\mathbf{b})) + n\lambda \epsilon \right)$$

Then, the function $f_{\mathbf{S}(1:k)} : \mathbb{R}^k \rightarrow \mathbb{R}^k$ defined via its inverse as

$$f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}) = \begin{cases} \Lambda_{1:k}(\mathbf{b}, \frac{\mathbf{b}}{\|\mathbf{b}\|}), & \text{if } \mathbf{b} \neq \mathbf{0} \\ \{\Lambda_{1:k}(\mathbf{0}, \epsilon) : \|\epsilon\| \leq 1\}, & \text{if } \mathbf{b} = \mathbf{0} \end{cases}$$

satisfies $\hat{\beta}_{1:k}^\lambda = f_{\mathbf{S}^{(1:k)}}(\mathbf{u}_{1:k})$.

Proof. In order to prove Theorem C.8, we first need the intermediate result below, following the same approach taken in the proof of Theorem 2.1 in [35].

Lemma C.9. *For data $(\mathbf{y}, \mathbf{X})$ with $\mathbf{X}$ full column-rank and $\lambda > 0$, define $\hat{\beta}_*^\lambda(\mathbf{b})$ to be $\mathbf{b}$ for the first k coordinates and $\hat{\beta}_{-1:k}^\lambda(\mathbf{b})$ on the rest, where $\hat{\beta}_{-1:k}^\lambda(\mathbf{b})$ is defined as in the statement of Theorem C.8. Also let*

$$f_\lambda(\mathbf{y}; \beta) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda(\|\beta_{1:k}\|_2 + \|\beta_{-1:k}\|_1)$$

be the group LASSO objective function. Then, for any $\mathbf{b} \in \mathbb{R}^k$, the following are equivalent statements:

1. $\mathbf{0} \in \partial_{\beta_{1:k}} f_\lambda(\mathbf{y}; \beta)|_{\beta=\hat{\beta}_*^\lambda(\mathbf{b})}$
2. $\hat{\beta}^\lambda = \hat{\beta}_*^\lambda(\mathbf{b})$
3. $\hat{\beta}_{1:k}^\lambda = \mathbf{b}$

Proof of Lemma C.9. First, we show equivalence between (1) and (2), so assume $\mathbf{0} \in \partial_{\beta_{1:k}} f_\lambda(\mathbf{y}; \beta)|_{\beta=\hat{\beta}_*^\lambda(\mathbf{b})}$. The function $f_\lambda(\mathbf{y}; (\beta_{-1:k} = \hat{\beta}_{-1:k}^\lambda(\mathbf{b}), \beta_{1:k}))$ is convex in $\beta_{1:k}$, so since $\mathbf{0} \in \partial_{\beta_{1:k}} f_\lambda(\mathbf{y}; \beta)|_{\beta=\hat{\beta}_*^\lambda(\mathbf{b})} \iff \mathbf{0} \in \partial_{\beta_{1:k}} f_\lambda(\mathbf{y}; (\beta_{-1:k} = \hat{\beta}_{-1:k}^\lambda(\mathbf{b}), \beta_{1:k}))|_{\beta_{1:k}=\mathbf{b}}$,

$$\mathbf{b} = \arg \min_{\beta_{1:k}} \left(f_\lambda(\mathbf{y}; (\beta_{-1:k} = \hat{\beta}_{-1:k}^\lambda(\mathbf{b}), \beta_{1:k})) \right).$$

By definition of $\hat{\beta}_{-1:k}^\lambda(\mathbf{b})$, we also have that

$$\hat{\beta}_{-1:k}^\lambda(\mathbf{b}) = \arg \min_{\beta_{-1:k}} (f_\lambda(\mathbf{y}; (\beta_{-1:k}, \beta_{1:k} = \mathbf{b}))).$$

The above implies that running a blockwise coordinate descent with the blocks $[1 : k]$ and $\{[j]\}_{(k+1) \leq j \leq d}$ starting at $\hat{\beta}_*^\lambda(\mathbf{b})$ will result in iterates being constant at $\hat{\beta}_*^\lambda(\mathbf{b})$. The algorithm for the optimization of the sparse group LASSO taking $\lambda_2 = 0$ in [16] implies that blockwise coordinate descent over groups will result in the minimizer. Thus, $\hat{\beta}^\lambda = \hat{\beta}_*^\lambda(\mathbf{b})$. Assuming (2), since $\hat{\beta}^\lambda$ is the minimizer of the objective $f_\lambda(\mathbf{y}; \beta)$, the sub-gradient of the objective with respect to each coordinate, in particular the first k , evaluated at the minimizer must contain 0. Since $\hat{\beta}^\lambda = \hat{\beta}_*^\lambda(\mathbf{b})$, condition (1) follows.

We now show equivalence between (2) and (3). That (2) implies (3) follows from the definition of $\hat{\beta}_*^\lambda(\mathbf{b})$. So, assuming (3), the blockwise coordinate descent argument from before implies

$$\hat{\beta}_{-1:k}^\lambda = \arg \min_{\beta_{-1:k}} \left(f_\lambda(\mathbf{y}; (\beta_{-1:k}, \beta_{1:k} = \hat{\beta}_{1:k}^\lambda)) \right)$$

Then since $\hat{\beta}_{1:k}^\lambda = \mathbf{b}$ by assumption, we have that $\hat{\beta}_{-1:k}^\lambda = \hat{\beta}_{-1:k}^\lambda(\mathbf{b})$. So, $\hat{\beta}^\lambda = \hat{\beta}_*^\lambda(\mathbf{b})$. $\square$

We now proceed with the proof of Theorem C.8 by using the decomposition in (6) from Lemma C.1 to decompose the group LASSO objective function as follows

$$\begin{aligned} f_\lambda(\mathbf{y}; \boldsymbol{\beta}) &= \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}_{1:k}\|_2 + \lambda \|\boldsymbol{\beta}_{-1:k}\|_1 \\ &= \frac{1}{2n} \|\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{-1:k}\boldsymbol{\beta}_{-1:k} - \mathbf{P}_{-1:k}\mathbf{X}_{1:k}\boldsymbol{\beta}_{1:k}\|_2^2 + \frac{1}{2n} \|\hat{\sigma}_{1:k}\mathbf{V}\mathbf{u} - (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}\boldsymbol{\beta}_{1:k}\|_2^2 \\ &\quad + \lambda \|\boldsymbol{\beta}_{1:k}\|_2 + \lambda \|\boldsymbol{\beta}_{-1:k}\|_1. \end{aligned}$$

The subdifferential of $f_\lambda(\mathbf{y}; \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}_{1:k}$ is given by

$$\begin{aligned} \partial_{\boldsymbol{\beta}_{1:k}} f_\lambda(\mathbf{y}; \boldsymbol{\beta}) &= -\frac{1}{n}(\mathbf{P}_{-1:k}\mathbf{X}_{1:k})^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{-1:k}\boldsymbol{\beta}_{-1:k} - \mathbf{P}_{-1:k}\mathbf{X}_{1:k}\boldsymbol{\beta}_{1:k}) \\ &\quad - \frac{1}{n}((\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k})^T (\hat{\sigma}_{1:k}\mathbf{V}\mathbf{u} - (\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{X}_{1:k}\boldsymbol{\beta}_{1:k}) \\ &\quad + \lambda s(\boldsymbol{\beta}_{1:k}), \text{ where} \\ s(\boldsymbol{\beta}_{1:k}) &= \begin{cases} \frac{\boldsymbol{\beta}_{1:k}}{\|\boldsymbol{\beta}_{1:k}\|}, & \text{if } \boldsymbol{\beta}_{1:k} \neq \mathbf{0} \\ \{\mathbf{v} \in \mathbb{R}^k : \|\mathbf{v}\| \leq 1\}, & \text{otherwise} \end{cases} \end{aligned}$$

The above simplifies as follows

$$\begin{aligned} \partial_{\boldsymbol{\beta}_{1:k}} f_\lambda(\mathbf{y}; \boldsymbol{\beta}) &= -\frac{1}{n}\mathbf{X}_{1:k}^T \hat{\mathbf{y}}_{1:k} - \frac{\hat{\sigma}_{1:k}}{n}\mathbf{X}_{1:k}^T \mathbf{V}\mathbf{u} + \frac{\mathbf{X}_{1:k}^T \mathbf{X}_{-1:k}\boldsymbol{\beta}_{-1:k} + \mathbf{X}_{1:k}^T \mathbf{X}_{1:k}\boldsymbol{\beta}_{1:k}}{n} + \lambda s(\boldsymbol{\beta}_{1:k}) \\ &= -\frac{1}{n}\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}\boldsymbol{\beta}) - \frac{\hat{\sigma}_{1:k}}{n}\mathbf{X}_{1:k}^T \mathbf{V}\mathbf{u} + \lambda s(\boldsymbol{\beta}_{1:k}). \end{aligned}$$

We can write $\mathbf{X}_{1:k}^T \mathbf{V}\mathbf{u}$ as $\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}\mathbf{u}_{1:k}$ since the columns of $\mathbf{V}_{-1:k}$ are orthogonal to the column space of $\mathbf{X}$. We then have that

$$\partial_{\boldsymbol{\beta}_{1:k}} f_\lambda(\mathbf{y}; \boldsymbol{\beta}) = -\frac{1}{n}\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}\boldsymbol{\beta}) - \frac{\hat{\sigma}_{1:k}}{n}\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}\mathbf{u}_{1:k} + \lambda s(\boldsymbol{\beta}_{1:k}).$$

Setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^\lambda * (\mathbf{b})$ for $\mathbf{b} \in \mathbb{R}^k$ and applying Lemma C.9, we have the following equivalences

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{1:k}^\lambda = \mathbf{b} &\iff \mathbf{0} \in \partial_{\boldsymbol{\beta}_{1:k}} f_\lambda(\mathbf{y}; \boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^\lambda(\mathbf{b})} \\ &\iff \mathbf{u}_{1:k} \in \frac{1}{\hat{\sigma}_{1:k}}(\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \left(-\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{1:k}\mathbf{b} - \mathbf{X}_{-1:k}\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})) + n\lambda s(\mathbf{b}) \right). \end{aligned}$$

Equivalently, using $\Lambda_{1:k}(\mathbf{b}, \boldsymbol{\epsilon})$ as defined in the statement of Theorem C.8, we find that $\hat{\boldsymbol{\beta}}_{1:k}^\lambda = f_{\mathbf{S}^{(1:k)}}(\mathbf{u}_{1:k})$, where

$$f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}) = \begin{cases} \Lambda_{1:k}(\mathbf{b}, \frac{\mathbf{b}}{\|\mathbf{b}\|}), & \text{if } \mathbf{b} \neq \mathbf{0} \\ \{\Lambda_{1:k}(\mathbf{0}, \boldsymbol{\epsilon}) : \|\boldsymbol{\epsilon}\| \leq 1\}, & \text{if } \mathbf{b} = \mathbf{0} \end{cases}.$$

□

C.5 Proof of Lemma 2.1

Proof. From Theorem C.8, the function $f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}; r)$ specified in the lemma's statement is given by

$$f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}; r) = \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \left(-\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{1:k} \mathbf{b} - \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})) + n\lambda \frac{\mathbf{b}}{r} \right).$$

We start by establishing that $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})$ is continuous in $\mathbf{b}$. Let $U(\mathbf{b}, \boldsymbol{\beta}_{-1:k})$ denote the objective function in (16), and note that it is a continuous function of its arguments. Also, $U(\mathbf{b}, \boldsymbol{\beta}_{-1:k}) \geq \lambda \|\boldsymbol{\beta}_{-1:k}\|$. So letting $\mathbf{b}^* \in \mathbb{R}^k$, $d > 0$, and $\epsilon = \sup_{\mathbf{b} \in B(\mathbf{b}^*, d)} U(\mathbf{b}, \mathbf{0})$, if we take $M = \frac{\epsilon}{\lambda}$, then $\|\boldsymbol{\beta}_{-1:k}\| > M$ implies that $\inf_{\mathbf{b} \in B(\mathbf{b}^*, d)} U(\mathbf{b}, \boldsymbol{\beta}_{-1:k}) \geq \lambda \|\boldsymbol{\beta}_{-1:k}\| > \epsilon$. By Lemma 3 from [35], $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})$ is continuous in $\mathbf{b}$.

We will have that $\nabla \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}) = \mathbf{0}$ if $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}) = \mathbf{0}$ so consider a value of $\mathbf{b} \in \mathbb{R}^k$ such that $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}) \neq \mathbf{0}$, and define the active set $\mathcal{A}(\mathbf{b}) = \{i \in \{k+1, \dots, d\} : \hat{\beta}_i^\lambda(\mathbf{b}) \neq 0\}$. Since $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})$ minimizes U , it follows that $\hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b})$ is a minimizer of $U(\mathbf{b}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is a vector of length $|\mathcal{A}(\mathbf{b})|$. Because all of the entries of $\hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b})$ are non-zero, $U(\mathbf{b}, \boldsymbol{\gamma})$ is differentiable in $\boldsymbol{\gamma}$ at $\hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b})$. Thus, the first-order conditions indicate that for $i \in \mathcal{A}(\mathbf{b})$,

$$\begin{aligned} \frac{\partial}{\partial \gamma_i} U(\mathbf{b}, \boldsymbol{\gamma}) \Big|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b})} &= 0 \\ \implies -\mathbf{X}_i^T (\mathbf{y} - \mathbf{X}_{1:k} \mathbf{b} - \mathbf{X}_{\mathcal{A}(\mathbf{b})} \hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b})) + 2n\lambda \text{sign}(\hat{\beta}_i^\lambda(\mathbf{b})) &= 0 \end{aligned}$$

Note that $\hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})$ is continuous in the neighborhood of $\mathbf{b}$ in which the active set $\mathcal{A}(\mathbf{b})$ does not change and hence the signs of the active variables also do not change. So, differentiating both sides of the equation above with respect to $\mathbf{b}$ yields,

$$\begin{aligned} \mathbf{X}_i^T \mathbf{X}_{1:k} + \mathbf{X}_i^T \mathbf{X}_{\mathcal{A}(\mathbf{b})} \nabla \hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b}) &= 0, \text{ for all } i \in \mathcal{A}(\mathbf{b}) \\ \implies \mathbf{X}_{\mathcal{A}(\mathbf{b})}^T \mathbf{X}_{1:k} + \mathbf{X}_{\mathcal{A}(\mathbf{b})}^T \mathbf{X}_{\mathcal{A}(\mathbf{b})} \nabla \hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b}) &= 0 \\ \implies \nabla \hat{\boldsymbol{\beta}}_{\mathcal{A}(\mathbf{b})}^\lambda(\mathbf{b}) &= -(\mathbf{X}_{\mathcal{A}(\mathbf{b})}^T \mathbf{X}_{\mathcal{A}(\mathbf{b})})^{-1} \mathbf{X}_{\mathcal{A}(\mathbf{b})}^T \mathbf{X}_{1:k}. \end{aligned}$$

With $\nabla \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b})$ determined, the result follows from simple manipulations. $\square$

C.6 Proof of Corollary 2.2

Proof. By Lemma 2.1, for $\mathbf{b}$ in a sufficiently small neighborhood of $\mathbf{b}^*$ such that $\|\mathbf{b}\| = \|\mathbf{b}^*\|$,

$$f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}; \|\mathbf{b}^*\|) = \nabla f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|) \mathbf{b} + \boldsymbol{\nu}(\mathbf{b}^*), \text{ where } \boldsymbol{\nu}(\mathbf{b}^*) = f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|) - \nabla f_{\mathbf{S}(1:k)}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|) \mathbf{b}^*.$$

All we need to show is that $\boldsymbol{\nu}(\mathbf{b}^*)$ simplifies to the form specified in the corollary's statement. Observe that

$$\begin{aligned}
f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|) - \nabla f_{\mathbf{S}^{(1:k)}}^{-1}(\mathbf{b}^*; \|\mathbf{b}^*\|) \mathbf{b}^* &= \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \\
&\quad \cdot \left(-\mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{1:k} \mathbf{b}^* - \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}^*)) + n\lambda \frac{\mathbf{b}^*}{\|\mathbf{b}^*\|} \right) \\
&\quad - \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \left[\mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}(\mathbf{b}^*)}) \mathbf{X}_{1:k} + \frac{n\lambda}{\|\mathbf{b}^*\|} \mathbf{I} \right] \mathbf{b}^* \\
&= -\frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T \hat{\mathbf{y}}_{1:k} + \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}^*) \\
&\quad + \frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T \mathbf{P}_{\mathcal{A}(\mathbf{b}^*)} \mathbf{X}_{1:k} \mathbf{b}^* \\
&= -\frac{1}{\hat{\sigma}_{1:k}} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k})^{-1} \mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}^*) - \mathbf{P}_{\mathcal{A}(\mathbf{b}^*)} \mathbf{X}_{1:k} \mathbf{b}^*) \\
&= -\frac{1}{\hat{\sigma}_{1:k}} (\mathbf{V}_{1:k}^T \mathbf{X}_{1:k}) \mathbf{S}^{-1} \mathbf{X}_{1:k}^T (\hat{\mathbf{y}}_{1:k} - \mathbf{X}_{-1:k} \hat{\boldsymbol{\beta}}_{-1:k}^\lambda(\mathbf{b}^*) - \mathbf{P}_{\mathcal{A}(\mathbf{b}^*)} \mathbf{X}_{1:k} \mathbf{b}^*).
\end{aligned}$$

Using the notation in the corollary statement, the result follows from the line above. $\square$

C.7 Proof of Proposition 2.4

Proof. Since $\mathbf{u}_{1:k}$ is rotationally invariant,

$$\|\mathbf{u}_{1:k} - \mathbf{c}\| \sim \|\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1\|,$$

where $\mathbf{e}_1 = (1, 0, \dots, 0) \in \mathbb{R}^k$. Let $\mathbf{u}_{1:k} = \frac{\mathbf{Z}_{1:k}}{\|\mathbf{Z}\|}$, where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n-d+k})$. Then,

$$\begin{aligned}
\|\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1\|^2 &= (\mathbf{u}_{1:k}^T - \|\mathbf{c}\| \mathbf{e}_1^T)(\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1) \\
&= \|\mathbf{u}_{1:k}\|^2 - 2\|\mathbf{c}\| u_1 + \|\mathbf{c}\|^2 \\
&= \frac{Z_1^2 + \dots + Z_k^2}{Z_1^2 + \dots + Z_{n-d+k}^2} - 2\|\mathbf{c}\| \frac{Z_1}{\sqrt{Z_1^2 + \dots + Z_{n-d+k}^2}} + \|\mathbf{c}\|^2.
\end{aligned}$$

Let $Z_1 = SX$, where S is a random sign and $X \sim \chi_1$, so that $\{S, X, Z_2, \dots, Z_{n-d+k}\}$ are jointly independent. Also, let $G_1 = Z_1^2$, $G_2 = Z_2^2 + \dots + Z_k^2$, and $G_3 = Z_{k+1}^2 + \dots + Z_{n-d+k}^2$.

Then, we have that

$$\begin{aligned}
\frac{Z_1^2 + \dots + Z_k^2}{Z_1^2 + \dots + Z_{n-d+k}^2} &= \frac{G_1 + G_2}{G_1 + G_2 + G_3} \\
&= B_2 \sim \text{Beta}\left(\frac{k}{2}, \frac{n-d}{2}\right) \\
\frac{Z_1}{\sqrt{Z_1^2 + \dots + Z_{n-d+k}^2}} &= \frac{SX}{\sqrt{Z_1^2 + \dots + Z_{n-d+k}^2}} \\
&= S \sqrt{\frac{Z_1^2}{Z_1^2 + \dots + Z_{n-d+k}^2}} \\
&= S \sqrt{\frac{G_1}{G_1 + G_2 + G_3}} \\
&= S \sqrt{B_1 B_2}, \text{ where } B_1 = \frac{G_1}{G_1 + G_2} \sim \text{Beta}\left(\frac{1}{2}, \frac{k-1}{2}\right) \\
\implies \|\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1\|^2 &= B_2 - 2\|\mathbf{c}\| S \sqrt{B_1 B_2} + \|\mathbf{c}\|^2.
\end{aligned}$$

Note that $B_1 \perp B_2$ is a standard result, and $S \perp \{B_1, B_2\}$ since B_1 and B_2 are functions of $\{X, Z_2, \dots, Z_{n-d+k}\}$ and S was chosen to be independent of $\{X, Z_2, \dots, Z_{n-d+k}\}$. Thus, $\{S, B_1, B_2\}$ are jointly independent. Letting $W = \|\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1\|^2$, the density $f_W(w)$ is

$$\begin{aligned}
f_W(w) &= f_{W|S=1}(w) \frac{1}{2} + f_{W|S=-1}(w) \frac{1}{2} \\
&= \frac{1}{2}(f_{W_1}(w) + f_{W_2}(w)), \text{ where} \\
W_1 &= B_2 + 2\|\mathbf{c}\| \sqrt{B_1 B_2} + \|\mathbf{c}\|^2 \\
W_2 &= B_2 - 2\|\mathbf{c}\| \sqrt{B_1 B_2} + \|\mathbf{c}\|^2.
\end{aligned}$$

We will now determine the densities of W_1 and W_2 . Define the invertible function

$$\begin{aligned}
g_1(B_1, B_2) &= (W_1, B_2) \\
\implies g_1^{-1}(W_1, B_2) &= \left(\frac{1}{B_2} \left(\frac{W_1 - B_2 - \|\mathbf{c}\|^2}{2\|\mathbf{c}\|} \right)^2, B_2 \right).
\end{aligned}$$

By change of variables,

$$\begin{aligned}
f_{W_1, B_2}(w, t) &= f_{B_1, B_2}(g_1^{-1}(w, t)) \left| \frac{w - t - \|\mathbf{c}\|^2}{2t\|\mathbf{c}\|^2} \right| \\
&= h(\sqrt{w}, t),
\end{aligned}$$

where $h(w, t)$ is defined in the statement of Proposition 2.4. It follows that

$$f_{W_1}(w) = \begin{cases} \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^{w - \|\mathbf{c}\|^2} h(\sqrt{w}, t) dt & \text{if } \|\mathbf{c}\|^2 \leq w \leq \|\mathbf{c}\|^2 + 1 \\ \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^1 h(\sqrt{w}, t) dt & \text{if } \|\mathbf{c}\|^2 + 1 < w \leq (\|\mathbf{c}\| + 1)^2 \\ 0 & \text{otherwise} \end{cases}$$

Proceeding in a similar manner, define the invertible function

$$g_2(B_1, B_2) = (W_2, B_2) \\ \implies g_2^{-1}(W_2, B_2) = \left(\frac{1}{B_2} \left(\frac{\|\mathbf{c}\|^2 + B_2 - W_2}{2\|\mathbf{c}\|} \right)^2, B_2 \right).$$

By change of variables,

$$f_{W_2, B_2}(w, t) = f_{B_1, B_2}(g_2^{-1}(w, t)) \left| \frac{\|\mathbf{c}\|^2 + t - w}{2t\|\mathbf{c}\|^2} \right| \\ = h(\sqrt{w}, t).$$

It follows that

$$f_{W_2}^{\|\mathbf{c}\| \leq 1}(w) = \begin{cases} \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^{\|\mathbf{c}\|^2 + 2\sqrt{w}\|\mathbf{c}\| + w} h(\sqrt{w}, t) dt & \text{if } 0 \leq w \leq (\|\mathbf{c}\| - 1)^2 \\ \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^1 h(\sqrt{w}, t) dt & \text{if } (\|\mathbf{c}\| - 1)^2 < w \leq \|\mathbf{c}\|^2 \\ \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^1 h(\sqrt{w}, t) dt & \text{if } \|\mathbf{c}\|^2 < w \leq \|\mathbf{c}\|^2 + 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{W_2}^{\|\mathbf{c}\| > 1}(w) = \begin{cases} \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^1 h(\sqrt{w}, t) dt & \text{if } (\|\mathbf{c}\| - 1)^2 < w \leq \|\mathbf{c}\|^2 \\ \int_{\|\mathbf{c}\|^2 - 2\sqrt{w}\|\mathbf{c}\| + w}^1 h(\sqrt{w}, t) dt & \text{if } \|\mathbf{c}\|^2 < w \leq \|\mathbf{c}\|^2 + 1 \\ 0 & \text{otherwise} \end{cases}.$$

Putting these pieces together, we can obtain the density of W , and after applying change of variables, we find that the density of the random variable $Z = \sqrt{W} = \|\mathbf{u}_{1:k} - \|\mathbf{c}\| \mathbf{e}_1\|$ is

$$f_Z(z) = \mathbb{I}[\|\mathbf{c}\| \leq 1] \{zg(z)\mathbb{I}[z \in [0, \|\mathbf{c}\| + 1]]\} \\ + \mathbb{I}[\|\mathbf{c}\| > 1] \{zg(z)\mathbb{I}[z \in [\|\mathbf{c}\| - 1, \|\mathbf{c}\| + 1]]\} \\ g(z) = \begin{cases} \int_{\|\mathbf{c}\|^2 - 2z\|\mathbf{c}\| + z^2}^{\|\mathbf{c}\|^2 + 2z\|\mathbf{c}\| + z^2} h(z, t) dt & \text{if } 0 \leq z \leq \|\mathbf{c}\| - 1 \\ \int_{\|\mathbf{c}\|^2 - 2z\|\mathbf{c}\| + z^2}^1 h(z, t) dt & \text{if } \|\mathbf{c}\| - 1 < z \leq \|\mathbf{c}\| + 1 \end{cases}$$

□

C.8 Density of $\|\mathbf{u}_{1:k} - \boldsymbol{\nu}\| \mid \mathbf{S}^{(1:k)}$ under $H_{1:k}$ in the special case $k = 1$

Proof. The approach taken in the proof of Proposition 2.4 does not apply to the special case $k = 1$, so this section addresses this case. For any $z \in \mathbb{R}$, with some rewriting, we have that

$$P(|u_1 - c| \geq z) = 1 - P(u_1 \leq z + c) + P(u_1 \leq -z + c). \quad (17)$$

By similar reasoning from the proof of Proposition 2.4, $u_1 \sim S\sqrt{B}$ where S is a random sign and $B \sim \text{Beta}(\frac{1}{2}, \frac{n-d}{2})$ with $S \perp B$. It follows that for $a \in \mathbb{R}$,

$$\begin{aligned} P(u_1 \leq a) &= \frac{1}{2} \left(P(\sqrt{B} \leq a) + P(-\sqrt{B} \leq a) \right) \\ &= \frac{1}{2} \left(1 + \text{sign}(a) F_B(a^2) \right), \end{aligned}$$

where $F_B(\cdot)$ denotes the CDF of B . Taking $a = c \pm z$ and substituting into (17), after some simplification, we find that

$$P(|u_1 - c| \geq z) = 1 \mp \frac{1}{2} \text{sign}(c \pm z) F_B((c \pm z)^2).$$

□

C.9 Oracle test equivalence

Denote the cumulative distribution function (CDF) for the distribution of O_m from (10) conditional on $\mathbf{S}^{(1:k)}$ under $H_{1:k}$ evaluated at some $x \in \mathbb{R}$ by

$$F_{O_m}(x \mid \mathbf{S}^{(1:k)}) := P_{H_{1:k}}(O_m \leq x \mid \mathbf{S}^{(1:k)}).$$

We can then use $\bar{F}_{O_m}(O_m \mid \mathbf{S}^{(1:k)})$ as a valid p -value for $H_{1:k}$ that rejects for large O_m conditional on $\mathbf{S}^{(1:k)}$, where

$$\bar{F}_{O_m}(x \mid \mathbf{S}^{(1:k)}) := 1 - F_{O_m}(x \mid \mathbf{S}^{(1:k)}).$$

The following result formalizes how we can think of the oracle test as being equivalent in some sense to the test that uses the p -value $\bar{F}_{O_m}(O_m \mid \mathbf{S}^{(1:k)})$ in the limit as $m \rightarrow \infty$.

Proposition C.10. *Let T denote the one-sided t -test statistic for H_1 in model (9), and let $F_T(\cdot)$ denote the CDF of the t -distribution with $n - d + k - 1$ degrees of freedom so that $\bar{F}_T(T) := 1 - F_T(T)$ represents the p -value of the test. Also let $\bar{F}_{O_m}(O_m \mid \mathbf{S}^{(1:k)})$ denote the p -value associated with the conditional test for $H_{1:k}$ in model (1) that rejects for large O_m . Then,*

$$\bar{F}_{O_m}(O_m \mid \mathbf{S}^{(1:k)}) \xrightarrow{a.s.} \bar{F}_T(T) \text{ as } m \rightarrow \infty.$$

Proof. We start by recharacterizing the one-sided t -test for $H_1 : \gamma_1 = 0$ in the model $(\mathbf{y}, \tilde{\mathbf{X}})$ specified in (9). Note that the statistic $\mathbf{S}^{(1:k)}$ is also the minimal sufficient statistic for this model under the null $H_1 : \gamma_1 = 0$. In the proof of Lemma 2.2 of [35], it is shown that the one-sided t -test statistic T is a monotone increasing function of u_1 conditional on $\mathbf{S}^{(1:k)}$.

So, the conditional one-sided t -test is equivalent to the conditional test using u_1 . Since $u_1 = \frac{1}{\hat{\sigma}_{1:k}} \frac{\tilde{\mathbf{X}}_1^T(\mathbf{I} - \mathbf{P}_{-1:k})}{\|(\mathbf{I} - \mathbf{P}_{-1:k})\tilde{\mathbf{X}}_1\|} \mathbf{y}$, however, the conditional one-sided test using u_1 is the same as the conditional one-sided test using $\tilde{\mathbf{X}}_1^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}$. Moreover, since T is independent of $\mathbf{S}^{(1:k)}$, which follows from theory on ancillary statistics and is also proved in Appendix D.4 of [35], the standard one-sided t -test is equivalent to the conditional one-sided t -test and hence the conditional test using $\tilde{\mathbf{X}}_1^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}$.

The square of the test statistic O_m defined in (10) can be written as

$$\|\mathbf{u}_{1:k}\|^2 + 2m \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \right) (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \mathbf{u}_{1:k} + m^2 \left\| (\mathbf{V}_{1:k}^T \mathbf{X}_{1:k}) \left(\frac{\beta_{1:k}}{\|\beta_{1:k}\|} \right) \right\|^2$$

The third term does not depend on the data $\mathbf{y}$, so the test that rejects for large values of O_m conditional on $\mathbf{S}^{(1:k)}$ is equivalent to the conditional test that rejects for large values of the test statistic

$$\left(\mathbf{u}_{1:k}^T + 2m \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \right) (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \right) \mathbf{u}_{1:k}.$$

Recall $\mathbf{u}_{1:k} = \frac{1}{\hat{\sigma}_{1:k}} \mathbf{V}_{1:k}^T \mathbf{y}$ and $\mathbf{X}_{1:k}^T \mathbf{V}_{1:k} \mathbf{V}_{1:k}^T = \mathbf{X}_{1:k}^T \mathbf{V} \mathbf{V}^T = \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k})$. The realized p-value for the test that rejects for large values of O_m conditional on $\mathbf{S}^{(1:k)}$ is

$$\begin{aligned} P_{H_{1:k}} & \left(\left(\tilde{\mathbf{u}}_{1:k}^T + 2m \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \right) (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \right) \tilde{\mathbf{u}}_{1:k} \geq \left(\mathbf{u}_{1:k}^T + 2m \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \right) (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \right) \mathbf{u}_{1:k} \middle| \mathbf{S}^{(1:k)} \right) \\ & = P_{H_{1:k}} \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \tilde{\mathbf{u}}_{1:k} \geq \frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} (\mathbf{X}_{1:k}^T \mathbf{V}_{1:k}) \mathbf{u}_{1:k} + \frac{D}{m} \middle| \mathbf{S}^{(1:k)} \right) \\ & = P_{H_{1:k}} \left(\frac{1}{\hat{\sigma}_{1:k}} \frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \tilde{\mathbf{y}} \geq \frac{1}{\hat{\sigma}_{1:k}} \frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y} + \frac{D}{m} \middle| \mathbf{S}^{(1:k)} \right), \text{ where} \\ D & = \frac{1}{2} (\|\mathbf{u}_{1:k}\|^2 - \|\tilde{\mathbf{u}}_{1:k}\|^2). \end{aligned}$$

Letting A_m denote the event corresponding to the last probability expression above, we see that the sequence of events $A_1, A_2, \dots$ are nested. So, by the continuity of probability, as $m \rightarrow \infty$ the probability above converges to

$$P_{H_{1:k}} \left(\frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \tilde{\mathbf{y}} \geq \frac{\beta_{1:k}^T}{\|\beta_{1:k}\|} \mathbf{X}_{1:k}^T (\mathbf{I} - \mathbf{P}_{-1:k}) \mathbf{y} \middle| \mathbf{S}^{(1:k)} \right).$$

Since $\tilde{\mathbf{X}}_1 = \frac{\mathbf{X}_{1:k} \beta_{1:k}}{\|\beta_{1:k}\|}$, this is the p-value associated with the conditional test using $\tilde{\mathbf{X}}_1^T(\mathbf{I} - \mathbf{P}_{-1:k})\mathbf{y}$ that we previously justified is equivalent to the one-sided t -test for $H_1 : \gamma_1 = 0$ in the model $(\mathbf{y}, \tilde{\mathbf{X}})$ from (9). Using $\bar{F}_{O_m}(\cdot \mid \mathbf{S}^{(1:k)})$ and $\bar{F}_T(\cdot)$ as defined in the statement of Proposition C.10, it follows that

$$\bar{F}_{O_m}(O_m \mid \mathbf{S}^{(1:k)}) \xrightarrow{a.s.} \bar{F}_T(T) \text{ as } m \rightarrow \infty.$$

□

D Choice of λ and $\mathbf{b}^*$

In Section 2.2, we recommended choosing the tuning parameter λ and $\mathbf{b}^*$ for the L -test by cross-validating on $(\tilde{\mathbf{y}}, \mathbf{X})$, where $\tilde{\mathbf{y}}$ is a draw from the distribution $\mathbf{y} \mid \mathcal{S}^{(1:k)}$ under $H_{1:k}$. Taking a similar approach to [35], to motivate this choice, in the first section below, we consider two conventions used to pick the optimal λ in CV (with $\mathbf{b}^*$ then taken to be the group LASSO estimate at that chosen λ) and conclude that the min rule delivers the best results. In the following section, we quantify the randomness that selecting λ and $\mathbf{b}^*$ in this manner introduces into the L -test p-value. We also compare the L -test that performs CV on $(\tilde{\mathbf{y}}, \mathbf{X})$ to the L -test that performs CV on $(\hat{\mathbf{y}}_{1:k}, \mathbf{X})$, which would introduce no variability into the test. While cross-validating on the full dataset $(\mathbf{y}, \mathbf{X})$ would invalidate the p-value, we still consider how this invalid test would do since it serves as a good benchmark for the performance we might hope to achieve with our valid L -test.

D.1 Min rule vs. 1se rule

In this section, we compare two conventions used to select the optimal hyperparameter in CV: the min rule, which selects the λ corresponding to the minimum cross-validated error, and the 1se rule, which selects the largest λ resulting in a cross-validated error within one standard deviation of the minimum cross-validated error. Figure 7 compares the performance of the L -test that performs CV on $(\tilde{\mathbf{y}}, \mathbf{X})$ using these two rules and shows that the min rule performs noticeably better in all settings considered.

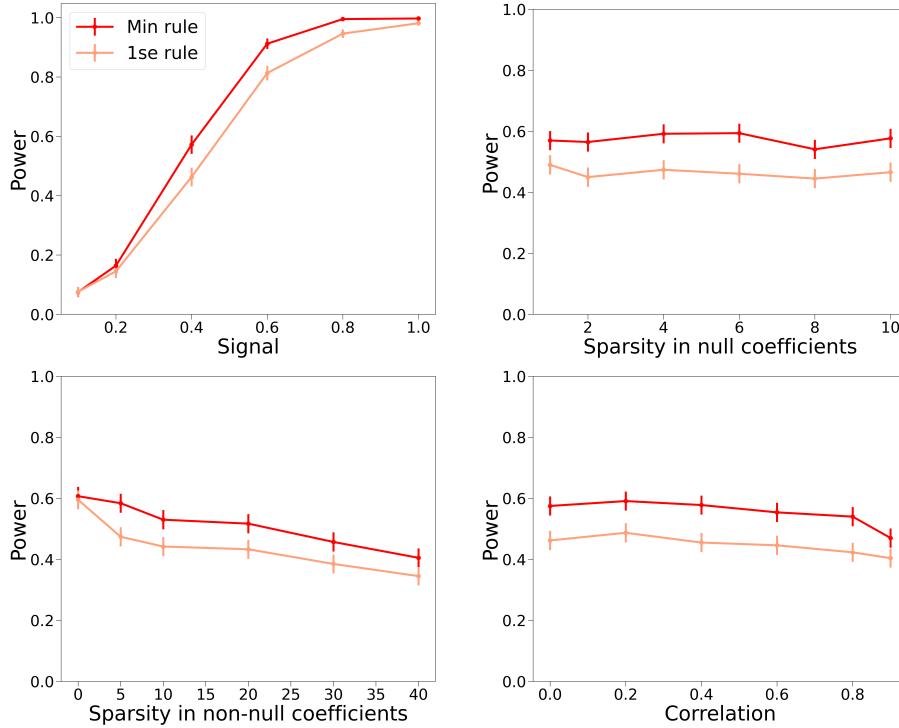


Figure 7: Min vs. 1se rule L -tests. Panel settings are the same as those in Figure 2.

D.2 Randomness in L -test

To assess the randomness due to cross-validating on $(\tilde{\mathbf{y}}, \mathbf{X})$, we run $m = 100$ replications for each of the first five signals in the top-left panel of Figure 2, denoting the i -th replication's dataset by $(\mathbf{y}^{(i)}, \mathbf{X}^{(i)})$. For each replication, we draw 100 samples of $\tilde{\mathbf{y}}$, yielding p-values p_{ij} corresponding to the j -th sample for dataset i .

The overall standard deviation of the p-values given by $\sqrt{\text{Var}(p_{i,j})}$ and the standard deviation conditioned on a particular dataset given by $\sqrt{\text{E}(\text{Var}(p_{ij} | (\mathbf{y}^{(i)}, \mathbf{X}^{(i)})))}$ are estimated using

$$\sqrt{\frac{1}{m^2 - 1} \sum_{i,j} (p_{ij} - \bar{p})^2} \quad \text{and} \quad \sqrt{\frac{1}{m(m-1)} \sum_{i,j=1}^m (p_{ij} - \bar{p}_i)^2},$$

respectively, where $\bar{p} = \frac{1}{m^2} \sum_{i,j=1}^m p_{ij}$ and $\bar{p}_i = \frac{1}{m} \sum_{j=1}^m p_{ij}$. This entire procedure is repeated 200 times and the means and associated standard errors of the overall and within standard deviation estimates are plotted in Figure 8. We conduct this experiment for both the L -test and MC-free test from Section 2.3, as the small p-values in the highest signal setting would typically be computed using the MC-free test per our recommendation.

As shown in Figure 8, sampling $\tilde{\mathbf{y}}$ accounts for at most 33% and 16% of the overall standard deviation (11% and 2.7% of the overall variance) for the L -test and MC-free test, respectively. This variance contribution comes down even further to less than 3% and 1% for the L -test and MC-free test, respectively, when we take a mean of 5 repeated λ 's and $\mathbf{b}^*$'s. Repeating the experiment with $k = 1$ reproduced the results of [35] ($\approx 1\%$ of overall variance), so increasing k raised the variability slightly, though it did not worsen further when we tried $k = 20$.

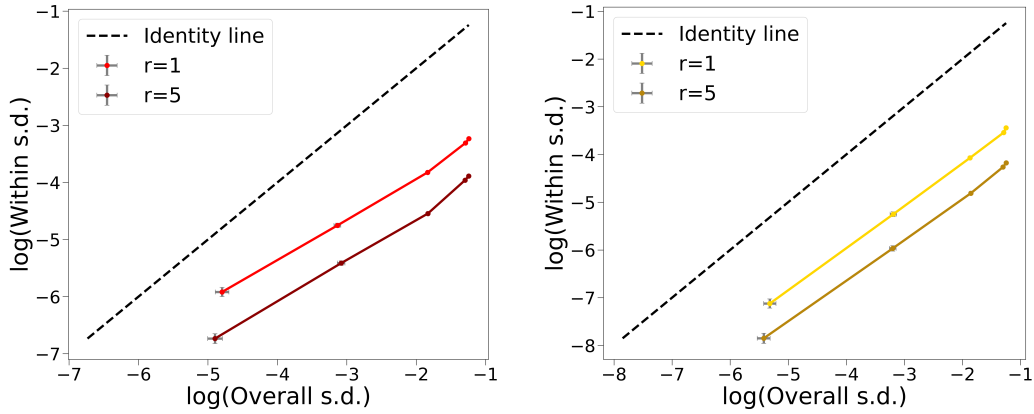


Figure 8: Plot showing the randomness introduced by conditional sampling of $\tilde{\mathbf{y}}$ in the L -test (left) and MC-free test (right) for each of the first five signal settings in Figure 2. For each test, r repetitions of λ and $\mathbf{b}^*$ are averaged to compute the penalty and point estimate.

Even though this randomness is negligible, in pursuit of an approach that introduces no randomness into the performance of the L -test, we considered the L -test that does CV on $(\hat{\mathbf{y}}_{1:k}, \mathbf{X})$. In Figure 9, we compare the performance of the L -tests that do CV on $(\tilde{\mathbf{y}}, \mathbf{X})$

and $(\hat{\mathbf{y}}_{1:k}, \mathbf{X})$ to the invalid test that does CV on $(\mathbf{y}, \mathbf{X})$. We see that the L -test that does CV on $(\tilde{\mathbf{y}}, \mathbf{X})$ does better than the one that cross-validates on $(\hat{\mathbf{y}}_{1:k}, \mathbf{X})$ and in fact does pretty much just as well as the invalid test, demonstrating that our recommended method does very close to the best we could hope to do.

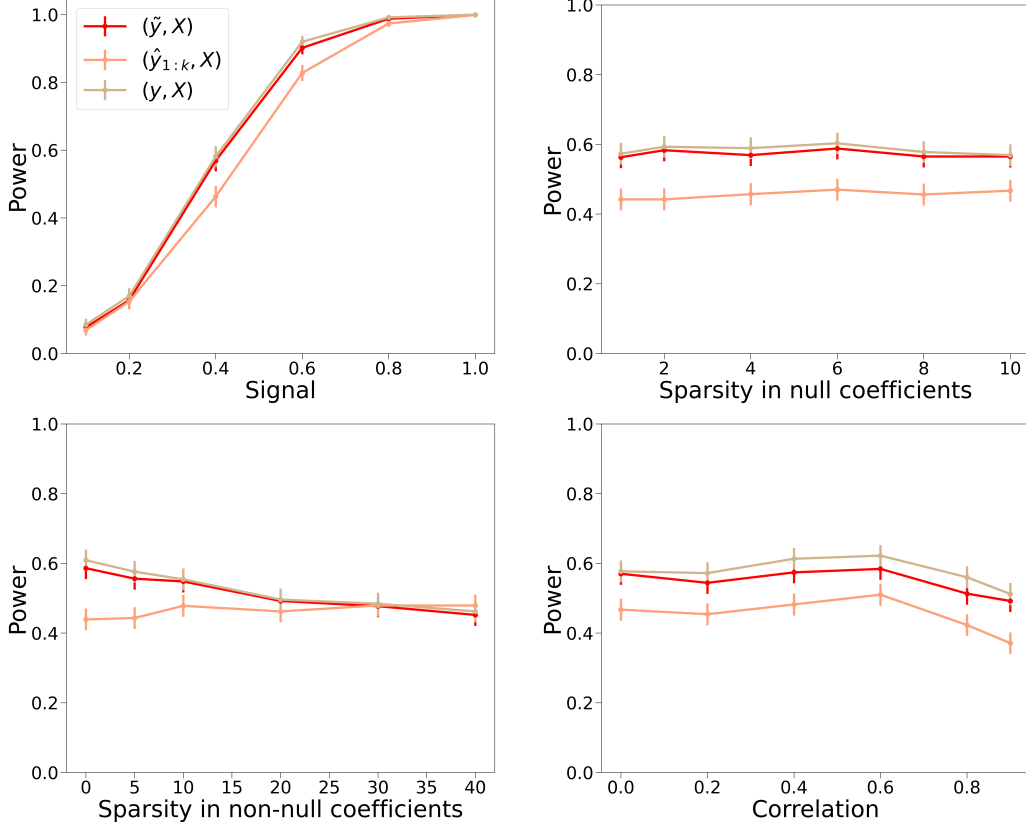


Figure 9: L -tests with λ chosen by cross-validating on $(\mathbf{y}, \mathbf{X})$, $(\tilde{\mathbf{y}}, \mathbf{X})$, and $(\hat{\mathbf{y}}_{1:k}, \mathbf{X})$. Panel settings are the same as those in Figure 2.

E Additional simulation studies

E.1 Explicit construction for $\mathbf{V}$

Recall the matrix $\mathbf{V} \in \mathbb{R}^{n \times (n-d+k)}$ described following Equation (2). We can construct such an orthonormal matrix by taking $\mathbf{V}_i = \frac{(\mathbf{I} - \mathbf{P}_{-1:i})\mathbf{X}_i}{\|(\mathbf{I} - \mathbf{P}_{-1:i})\mathbf{X}_i\|}$ for $i \in \{1, \dots, k\}$. Then, letting $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\tilde{\mathbf{V}}^T$ denote the singular value decomposition of $\mathbf{X}$, define $\mathbf{V}_{(k+1):n-d+k} = \mathbf{U}_{(d+1):n}$. It is easy to check that the resulting columns of $\mathbf{V}$ are an orthonormal basis that spans a space orthogonal to the column space of $\mathbf{X}_{-1:k}$. But, since $\mathbf{V}$ has $n - d + k$ columns, its column space is the orthogonal complement of the column space of $\mathbf{X}_{-1:k}$. This particular construction of $\mathbf{V}$ is used in all simulation experiments.

E.2 Visualization of $f_{\mathcal{S}(1:k)}^{-1}$ from Theorem C.8

In Section 2.2, we explained that our approximation for the piecewise affine function $f_{\mathcal{S}(1:k)}^{-1}(\mathbf{b}; r)$ was motivated by visualizations in which we found it to appear globally affine. Figure 10 shows examples of such visualizations in both low and high correlation regimes, where we can see that the contours appear ellipsoidal.

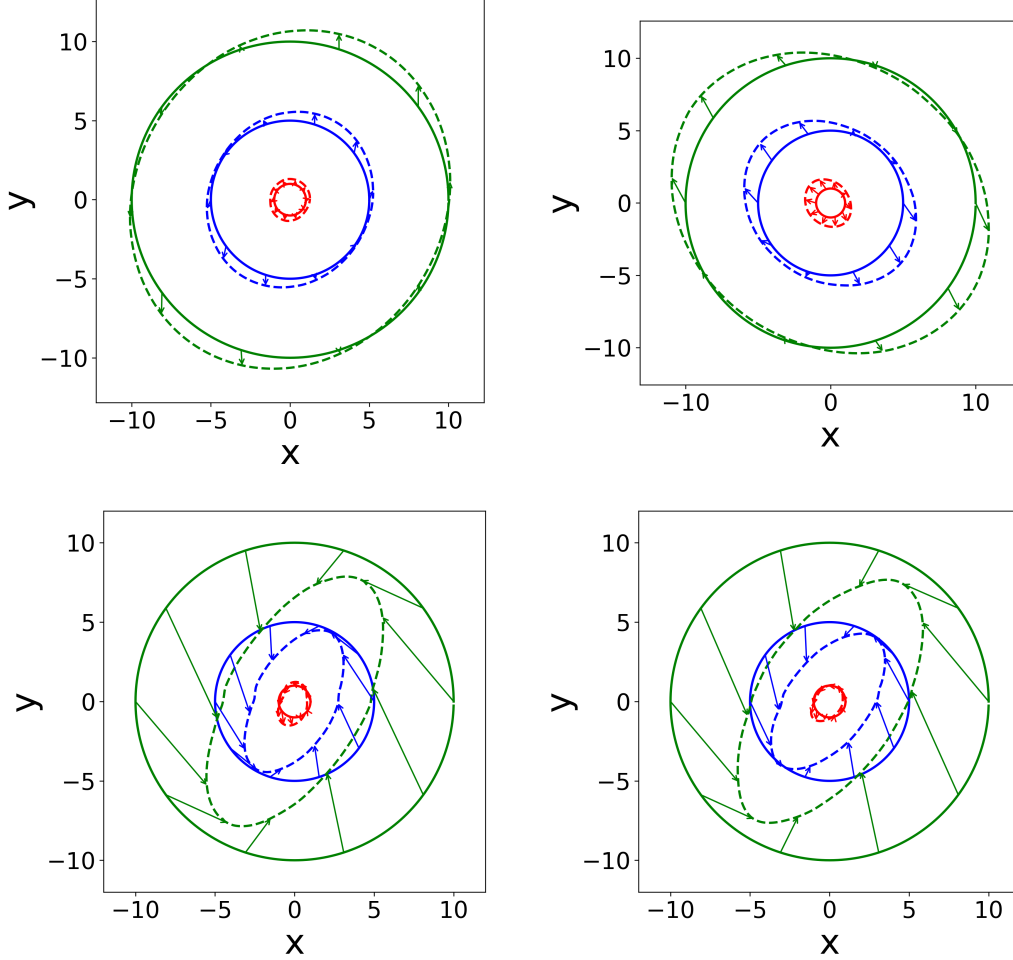


Figure 10: Plots show $f_{\mathcal{S}(1:k)}^{-1}$ from Theorem C.8 applied to three concentric circles. The arrows point to the corresponding outputs. The parameters are $n = 100$, $d = 50$, $k = 2$, $A = 0.3$, $k_1 = k$, $k_2 = 5$, and $\rho = 0$ in the top row while $\rho = 0.9$ in the bottom row. Each row shows two instantiations of $f_{\mathcal{S}(1:k)}^{-1}$, corresponding to two different data generations.

E.3 P-value visualizations

Building on the theoretical intuition in Section 3 and p-value geometry shown in Figure 1, we empirically visualize the p-value mass of the L -, oracle, and F -tests in Figure 11. At low correlation, the L -test's nearly circular ellipsoid gains power mainly from recentering;

at high correlation, variance concentrates along few PCs, producing an eccentric ellipsoid nearly orthogonal to $\mathbf{u}_{1:k}$ and yielding near-oracle power.

E.4 Power plots

In Figure 12, we extend the experiments from Section 4.1 by further comparing the tests' powers in a large model setting with $n = 1000$ and $d = 500$ and models that are close to unidentifiability (ie. $n \approx d$).

Figure 14 further demonstrates the similarity between the L -test and PC-test in the block-orthogonal design setting studied in Section 4.2. To show that the way both the PC-test and premultiplier $\mathbf{A}(\tilde{\beta}_{1:k}^\lambda, \lambda)$ exploit the geometry of $\mathbf{X}_{1:k}$ for power gains is not specific to block orthogonality, Figure 15 compares the PC-test with a test that applies $\mathbf{A}(\tilde{\beta}_{1:k}^\lambda, \lambda)$ directly to $\mathbf{u}_{1:k}$ (without recentering)—we call this the ϕ -test—in a standard design setting. Interestingly, the bottom panel shows a setting where the L -test underperforms the group LASSO MC test slightly, suggesting greater robustness of the latter to anti-alignment between $\beta_{1:k}$ and the top PCs of $\mathbf{X}_{1:k}$. Nevertheless, the gap is small, and the L -test closely matches the group LASSO MC test's performance in nearly all other simulations and our real data analysis.

When we first introduced the L -test, we mentioned that we could have considered constructing a test based on LASSO-type estimators other than the group LASSO. Figure 16 compares the powers of the L -test and tests that use the p-value in (5) but with the group LASSO swapped out for the LASSO [39], elastic net [47], or sparse group LASSO [36]. For these tests, just as for the L -test, hyperparameter tuning is done on $(\tilde{\mathbf{y}}, \mathbf{X})$, where $\tilde{\mathbf{y}} \sim \mathbf{y} \mid \mathcal{S}^{(1:k)}$ under $H_{1:k}$, using 10-fold cross-validation. For the elastic net and sparse group LASSO, hyperparameter tuning is performed for both the penalty and the ℓ^1 -ratio.

The L -test and sparse group LASSO test outperform the F -test and almost always have higher power than the LASSO and elastic net tests, which perform almost identically. The only exception is the setting where the null is 1-sparse, exactly when we would expect these latter estimates to perform well. While the L -test could in theory be modified to approximate the MC test that instead uses the statistic obtained by substituting the LASSO or elastic net for $\hat{\beta}_{1:k, \text{OLS}}$ in (4) to achieve even higher power in this setting, it represents a relatively narrow scenario, unlikely to occur in practice when testing groups of coefficients, where one typically expects more than a single nonzero element.

Almost all of the tests show worse performance as sparsity among the non-null parameters increases. The only exception to this trend is the initial poor performance of the sparse group LASSO test when $k_2 = 0$. In this setting, the sparse group LASSO often returns the zero vector, resulting in large p-values. The other LASSO-based estimators do not exhibit this behavior because they penalize coefficients differently: the LASSO, with individual-level penalties, is able to identify the strong signals in k_1 out of the first k coefficients, while the group LASSO, with group-level penalties, is able to select the first group. The sparse group LASSO combines both penalties, and their interaction drives all coefficients to zero in this setting.

E.5 Robustness

This section extends the results from Section 4.3 by considering additional settings for each of the four types of model violation. These settings are described below. All settings are tested in both low ($\rho = 0$) and high ($\rho = 0.5$) correlation regimes.

1. **Heavy-tailed errors:** For each (n, d) pair, we take the errors to be independently and identically distributed from the t_ν -distribution with ν degrees of freedom for $\nu \in \{2, 5, 10, 15, 20, 30\}$. Errors are standardized by the theoretical standard deviation of the t-distribution except when $\nu = 2$, since there is no finite second moment in this case. Smaller values of ν correspond to larger tails, and taking $\nu = \infty$ results in standard normal tails.
2. **Skewed errors:** Same as the heavy-tailed error experiments, except errors are taken to be independently and identically distributed from the Gamma distribution with scale parameter 1 and shape parameter α for $\alpha \in \{1, 2, 4, 6, 8, 10\}$. Errors are standardized. Larger values of α correspond to greater symmetry.
3. **Heteroskedastic errors:** Similar to the prior two settings, except errors are generated independently in the following manner. Letting r_i denote the mean of the i -th row of $\mathbf{X}$,

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

We test $\eta^2 \in \{0.01, 0.25, 0.5, 1, 4, 8\}$, with larger departures from 1 corresponding to a greater degree of heteroskedasticity.

4. **Non-linearity:** Unlike the prior settings, errors are taken to be independently and identically distributed from the normal distribution with mean 0 and variance σ^2 , but the design matrix is taken to be $\mathbf{X}^\delta$, where the ij -th entry is defined as $(X^\delta)_{ij} = \text{sign}(X_{ij})|X_{ij}|^\delta$. We test $\delta \in \{0.3, 0.5, 1, 2, 3, 4\}$, where larger departures from 1 correspond to greater non-linearity in the model.

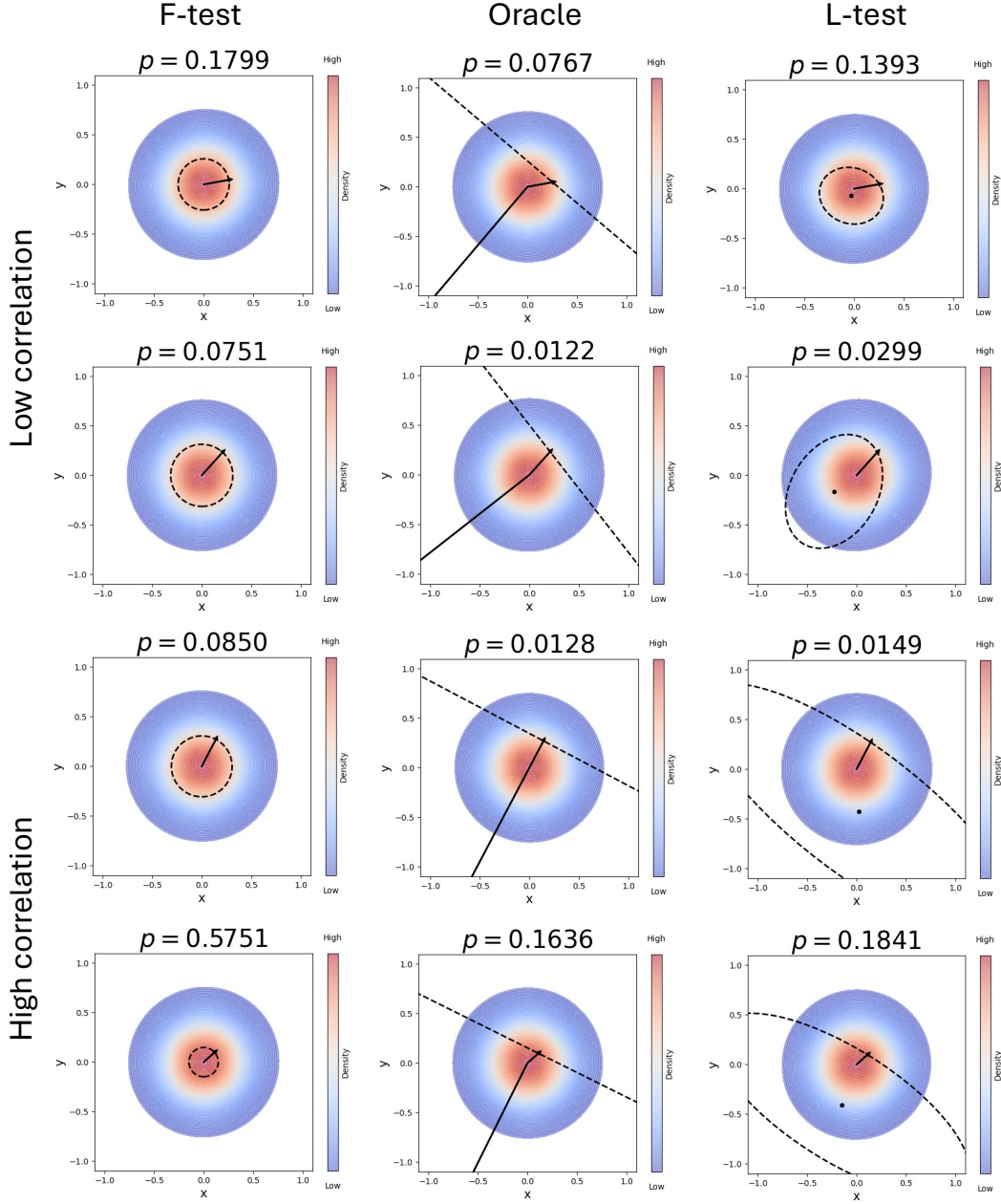


Figure 11: Heatmaps show the density of $\tilde{\mathbf{u}}_{1:k}$ for $\tilde{\mathbf{u}} \sim \text{Unif}(\mathbb{S}^{n-d+k})$. The black arrows correspond to the observed $\mathbf{u}_{1:k}$. In the left column, the F -test p-value mass is the region outside the dashed circle. In the middle column, the oracle test p-value mass is the region outside the dashed hyperplane drawn perpendicular to the oracle's recentering vector in black and through $\mathbf{u}_{1:k}$. In the right column, the L -test's p-value mass is the region outside the dashed ellipsoid. The parameters are $n = 100$, $d = 50$, $k = 2$, $A = 0.3$, $k_2 = 5$, and $\beta_{1:k}$ dense and nonnegative. The first two rows take $\rho = 0$ while the last two take $\rho = 0.9$ and show the p-value geometry on two data generations.

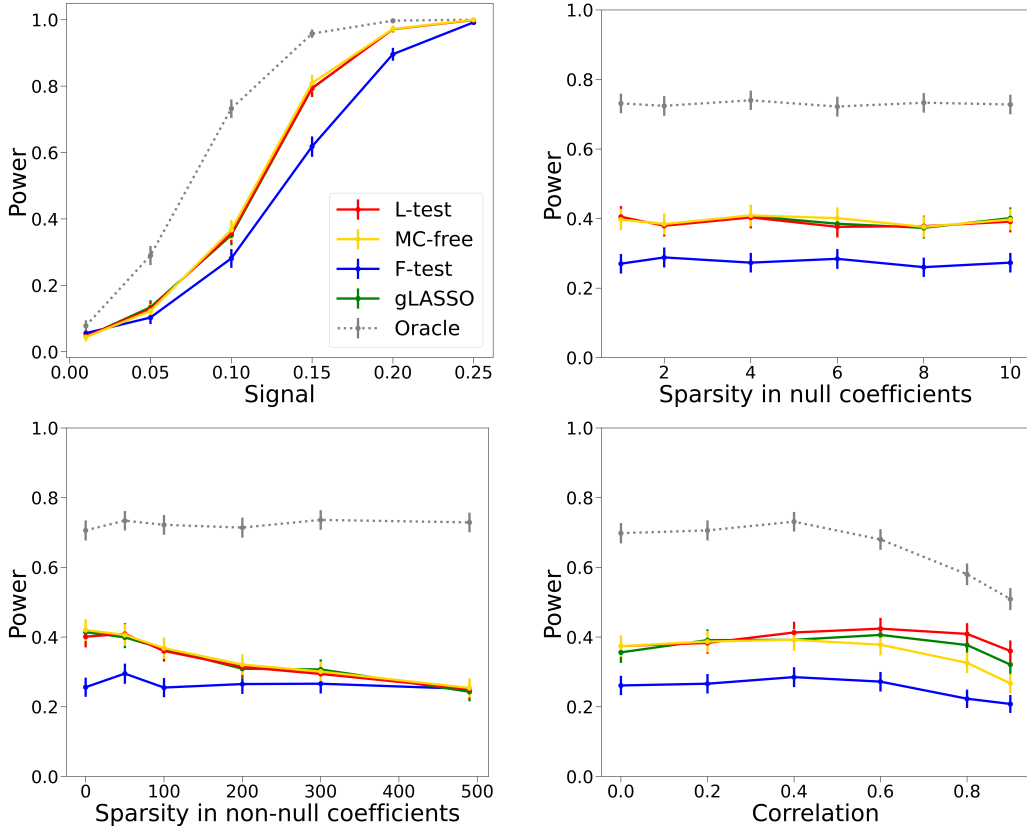


Figure 12: Comparison of test powers in large model. In all panels, $n = 1000$, $d = 500$, and $k = 10$. The top left panel takes $k_1 = k$, $k_2 = 49$, $\rho = 0$ and varies A ; top right takes $A = 0.1$, $k_2 = 49$, and $\rho = 0$ and varies k_1 ; bottom left takes $A = 0.1$, $k_1 = 4$, $\rho = 0$ and varies k_2 ; bottom right takes $A = 0.1$, $k_1 = 4$, $k_2 = 49$ and varies ρ .

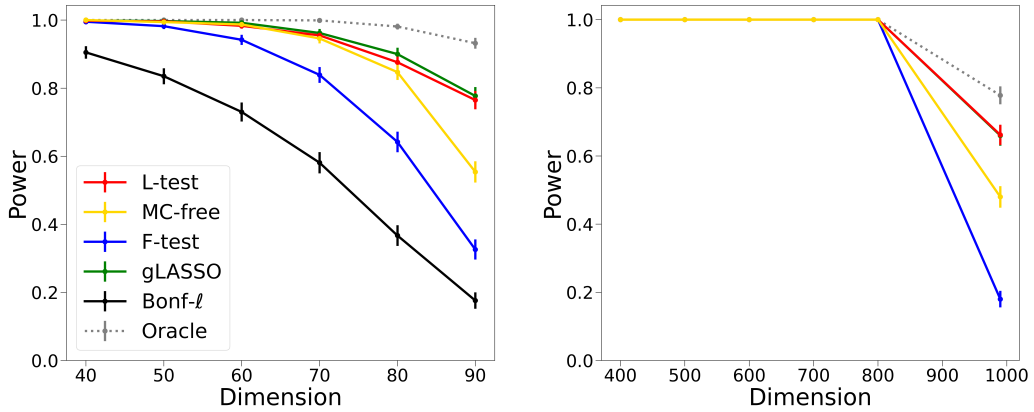


Figure 13: Comparison of test powers in high-dimensional models. The left panel takes $n = 100$, $k = 10$, $A = 0.8$, $k_1 = k$, and $\rho = 0$ and varies k_2 and d to maintain 10% non-null sparsity; the right does the same but takes $n = 1000$, $k = 10$, $A = 0.6$, $k_1 = k$, and $\rho = 0$.

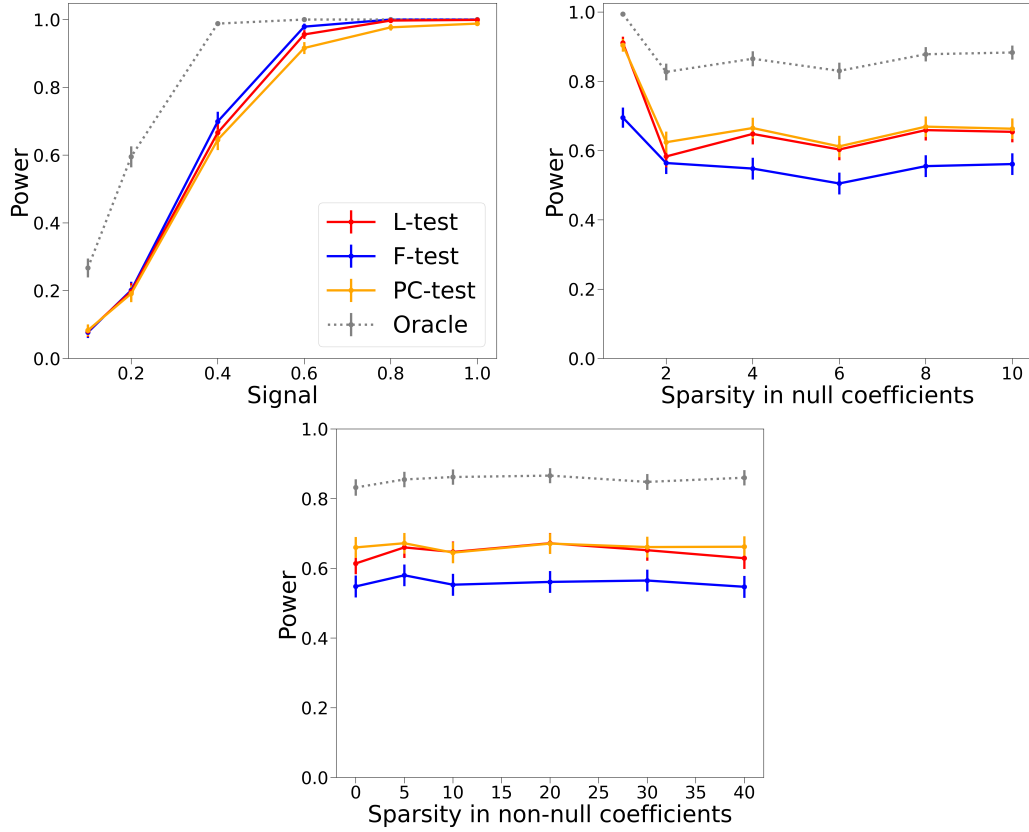


Figure 14: In all panels, we orthogonalize $\mathbf{X}_{1:k}$ to $\mathbf{X}_{-1:k}$ by using the design $(\mathbf{X}_{1:k} - (\mathbf{I} - \mathbf{P}_{1:k})\mathbf{X}_{-1:k})$, followed by standardization. Panel settings are the same as those in Figure 2, except top right and bottom take $\rho = 0.9$.

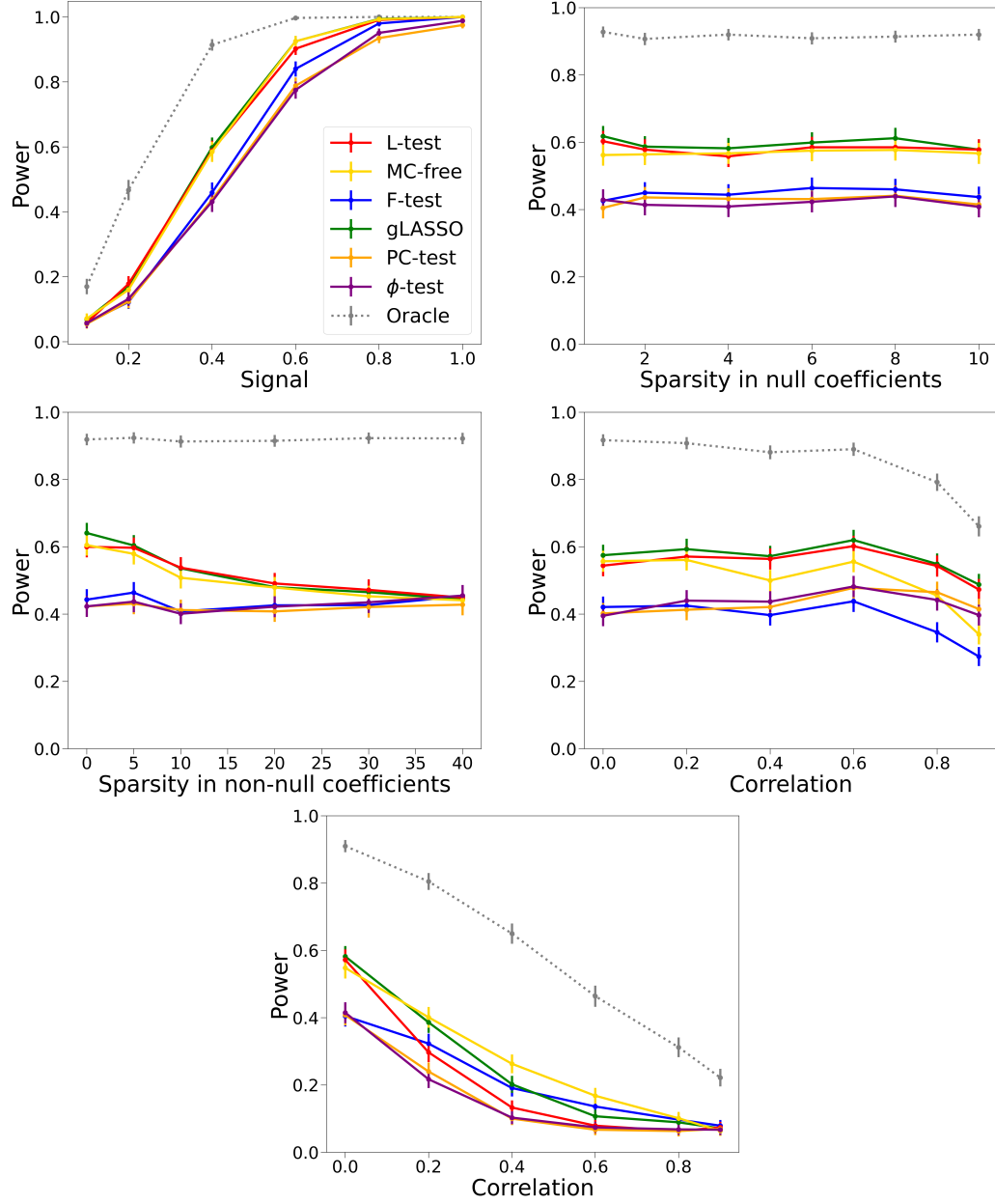


Figure 15: Top four panel settings are the same as those in Figure 2. The bottom panel is the same as the middle panel of Figure 3, except using a non-orthogonal design.

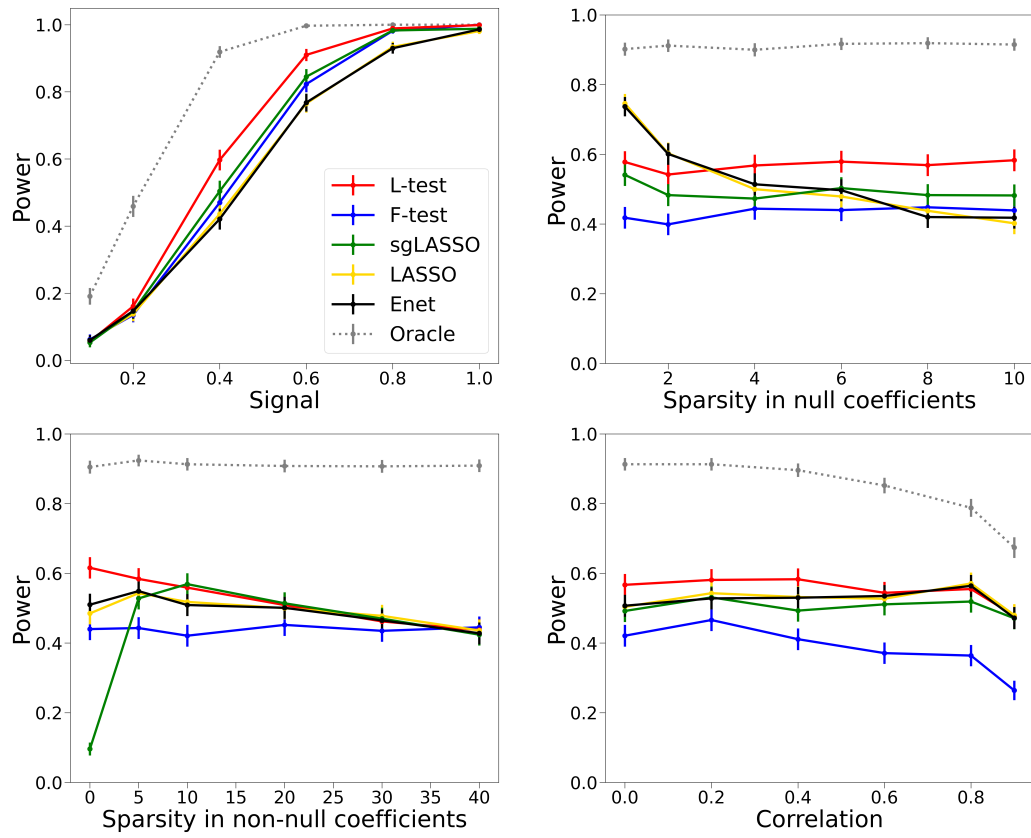


Figure 16: Comparison of L -test with other LASSO-based tests. Same settings as in Figure 2.

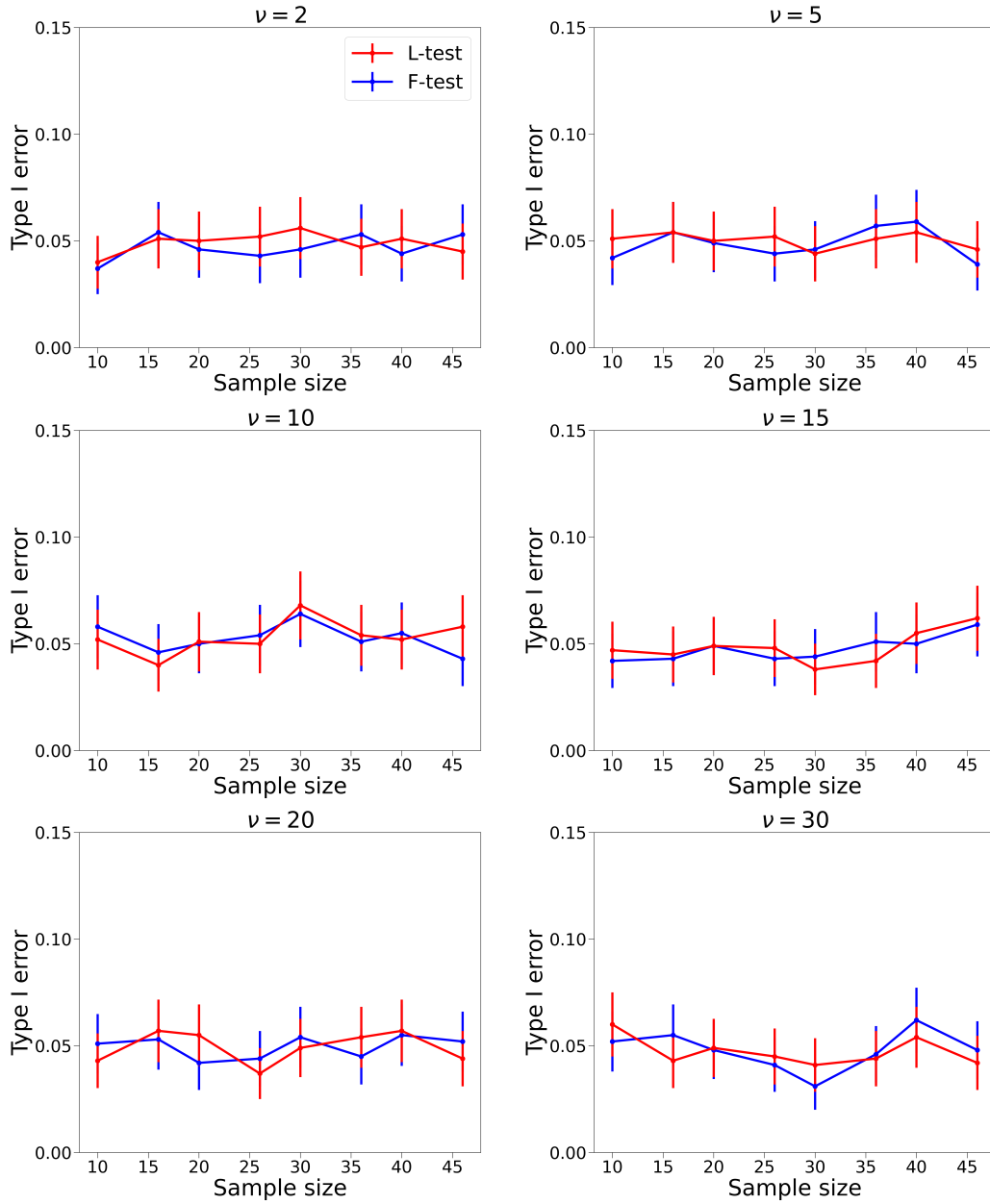


Figure 17: $\rho = 0$. F -test and L -test size for t -distributed errors with varying degrees of freedom.

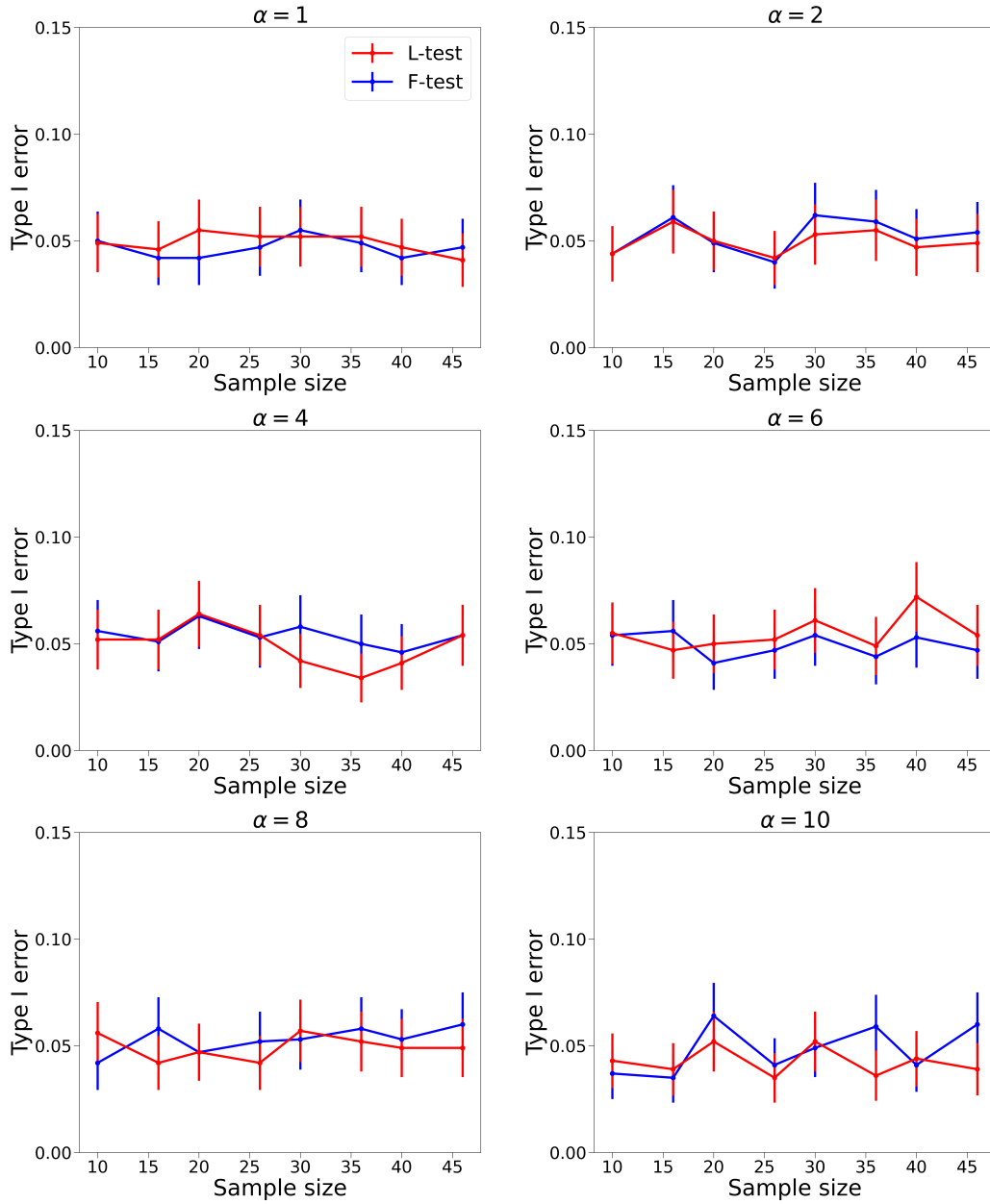


Figure 18: $\rho = 0$. F -test and L -test size for Gamma distributed errors with varying shape parameter (scale parameter = 1).

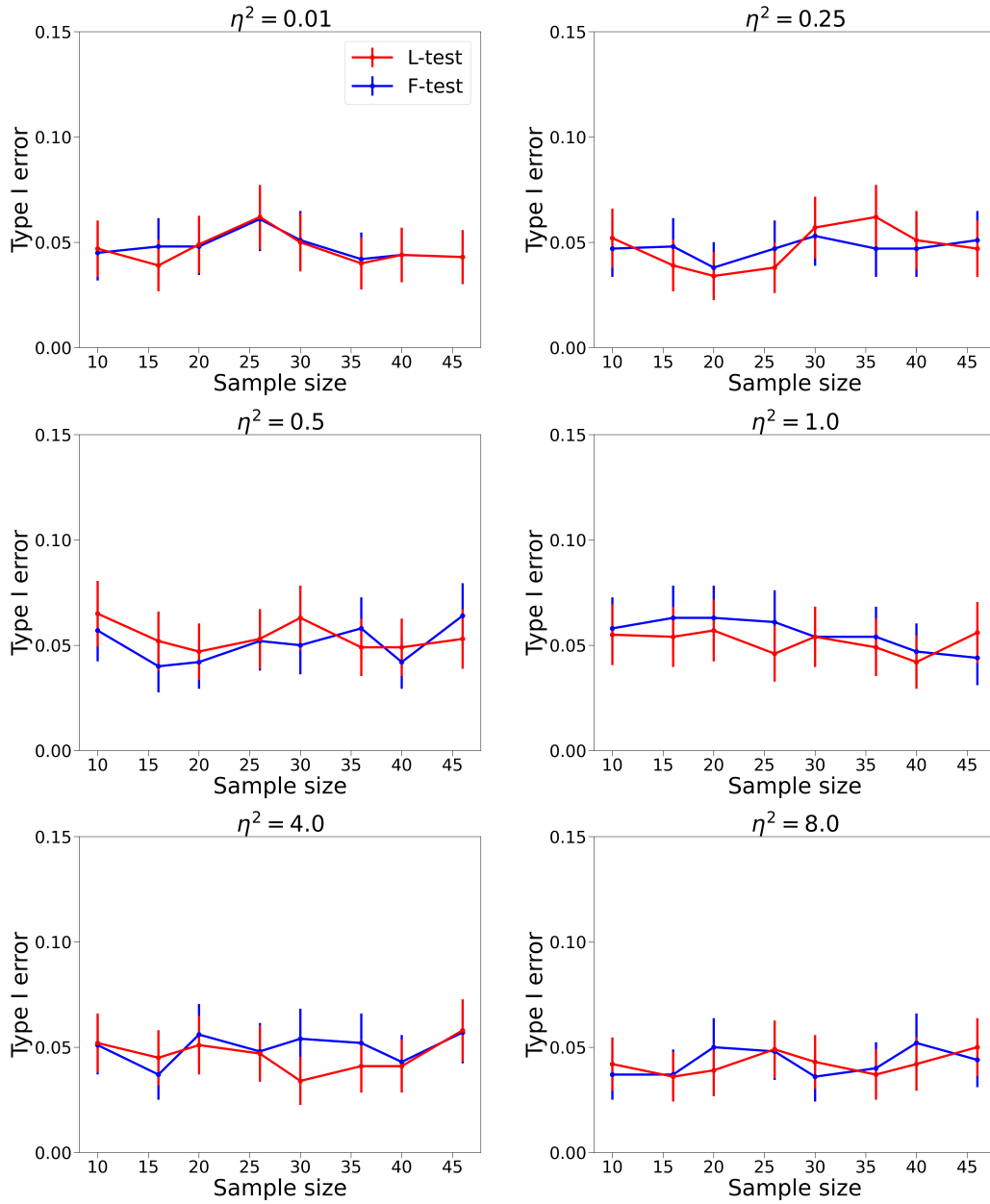


Figure 19: $\rho = 0$. F -test and L -test size for heteroskedastic errors with departures of η^2 from 1 indicating a greater degree of heteroskedasticity.

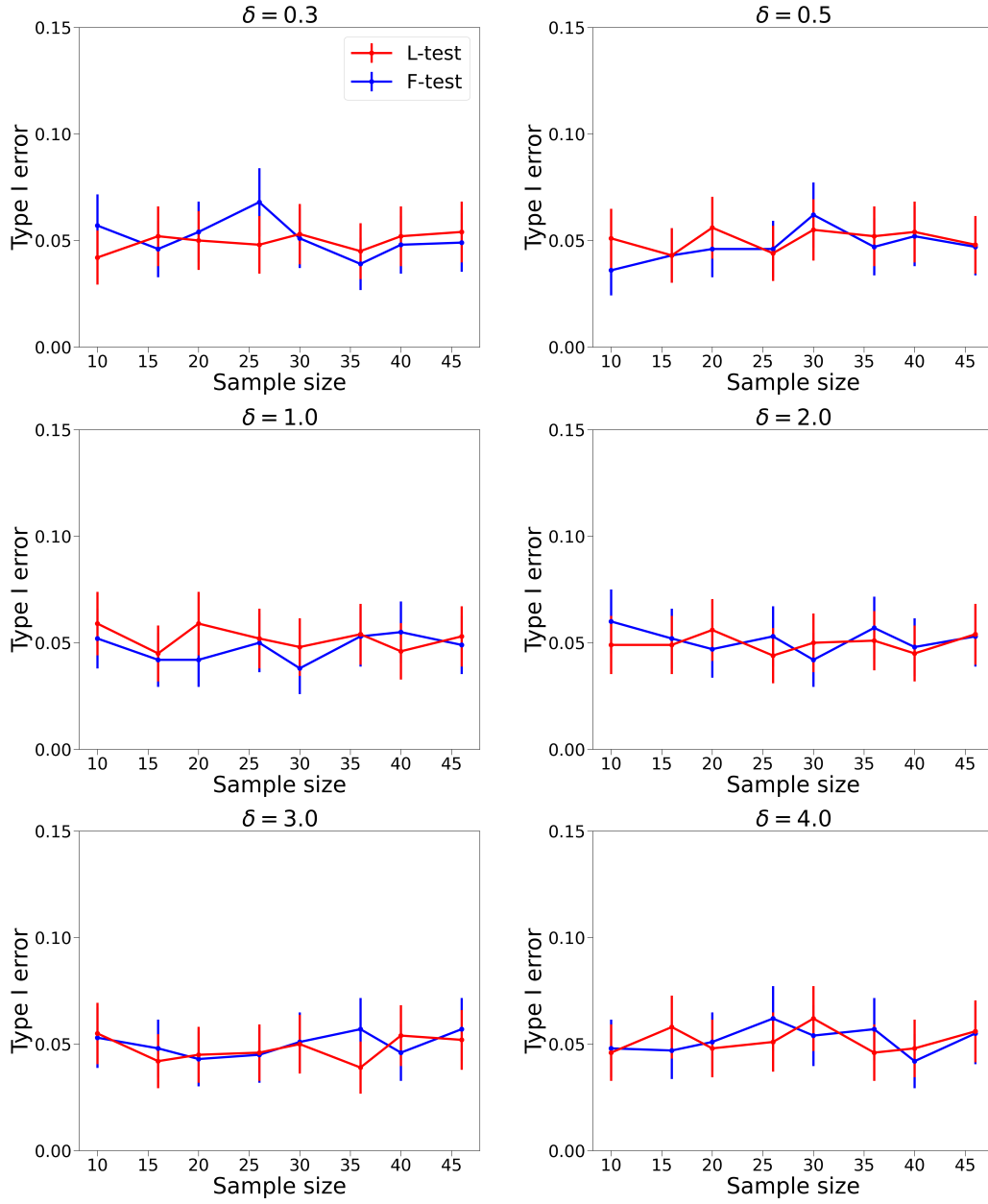


Figure 20: $\rho = 0$. F -test and L -test size under non-linearity, where departures of δ from 1 indicate a greater degree of non-linearity.

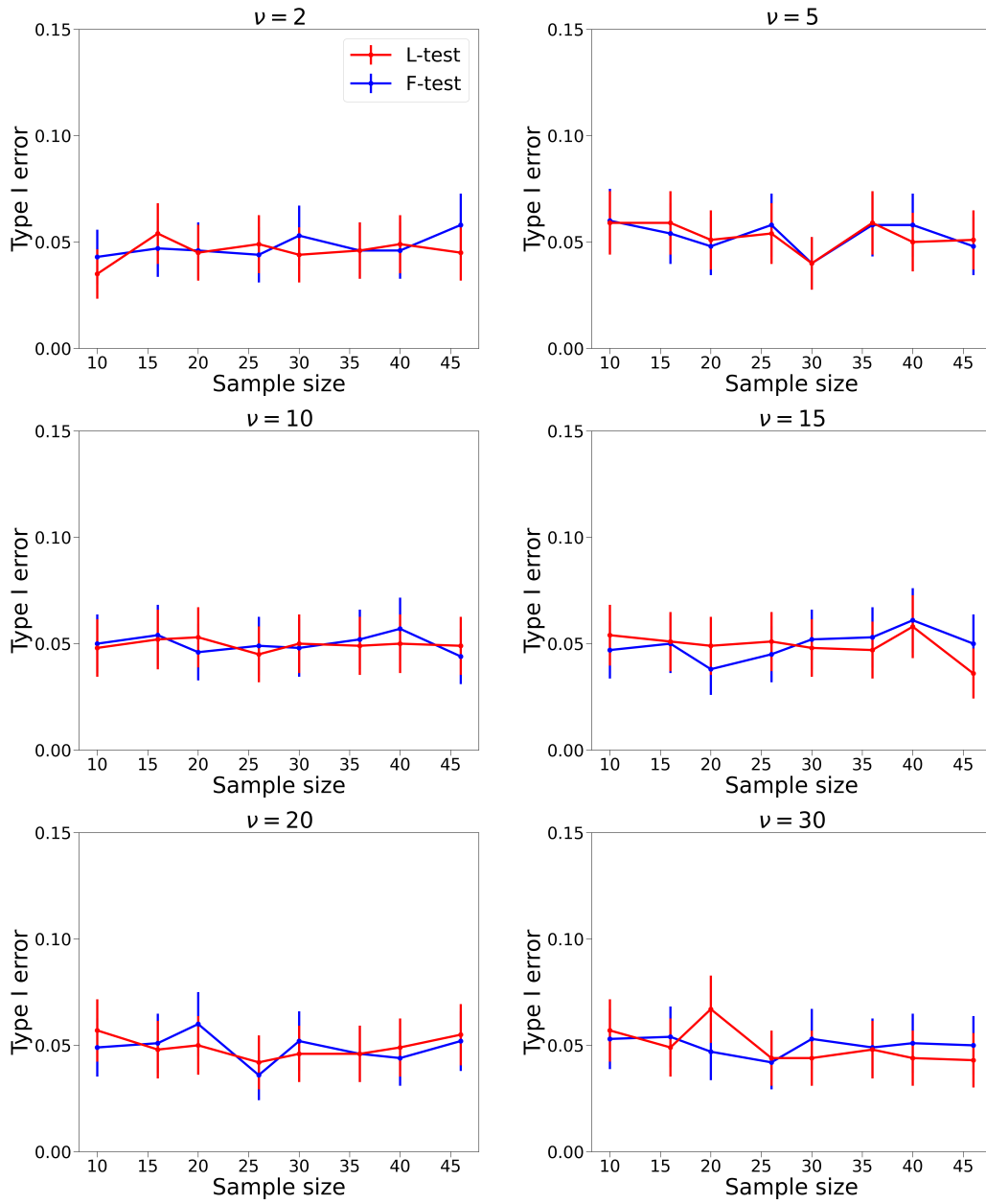


Figure 21: Same as Figure 17 but with $\rho = 0.5$.

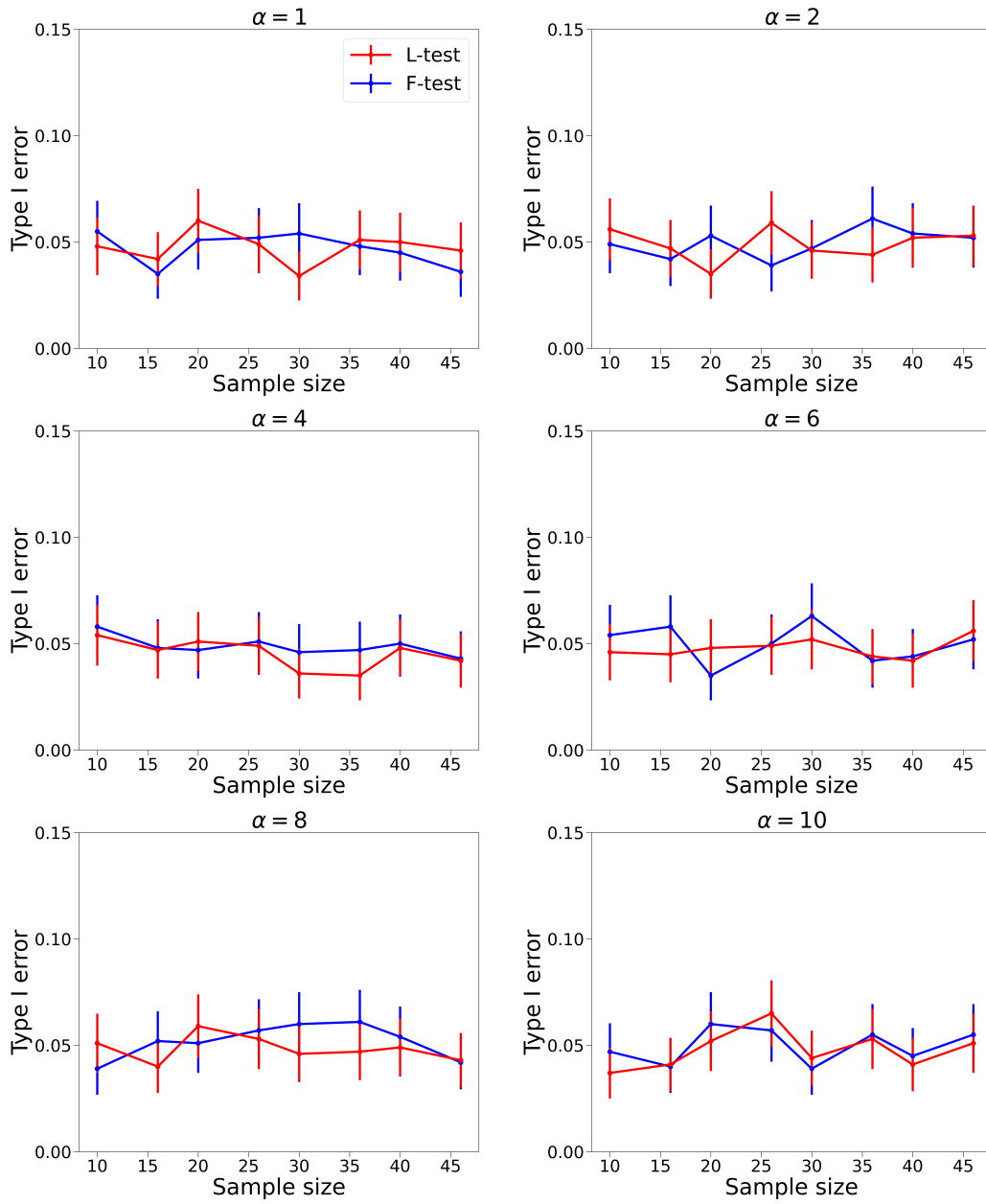


Figure 22: Same as Figure 18 but with $\rho = 0.5$.

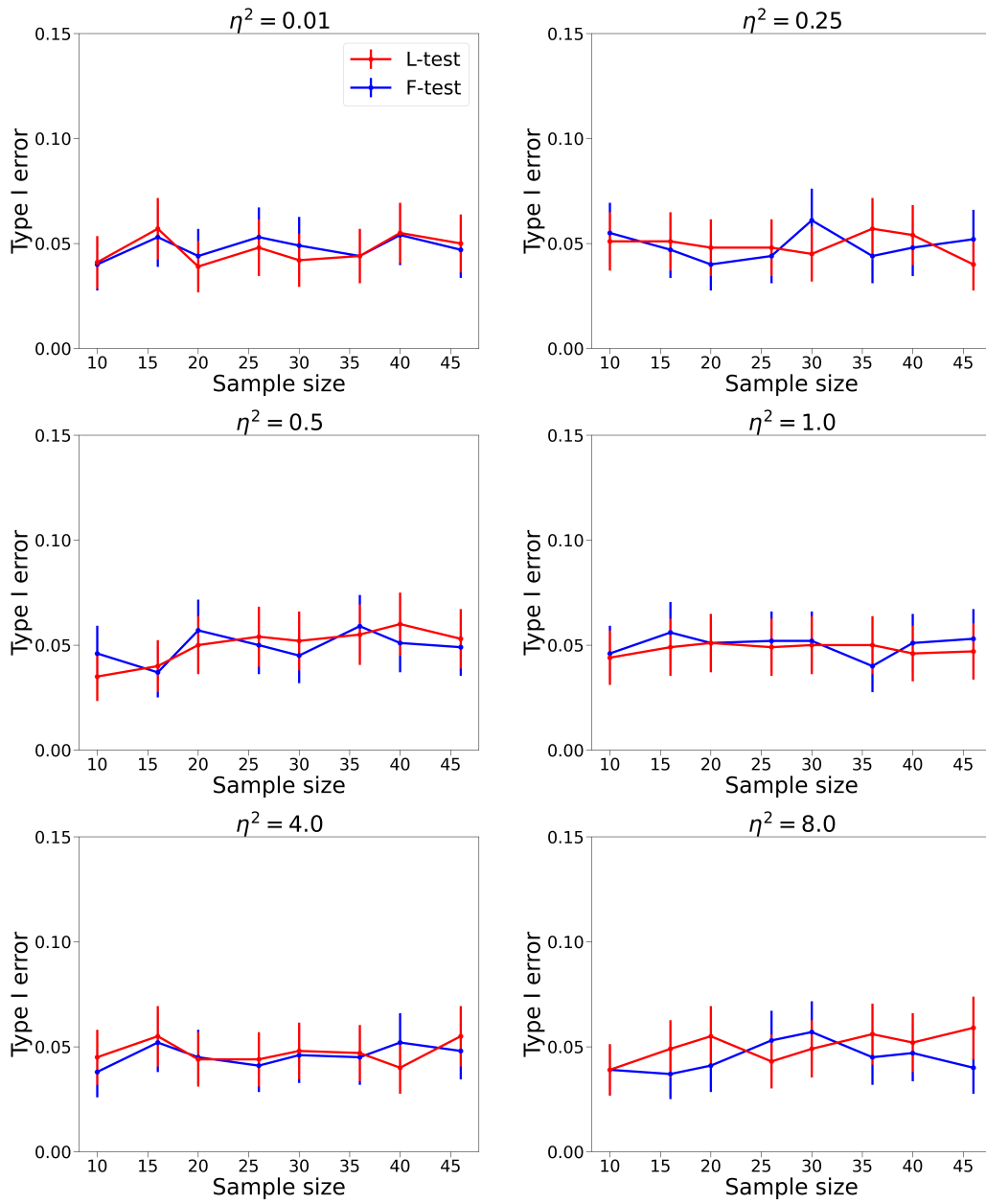


Figure 23: Same as Figure 19 but with $\rho = 0.5$.

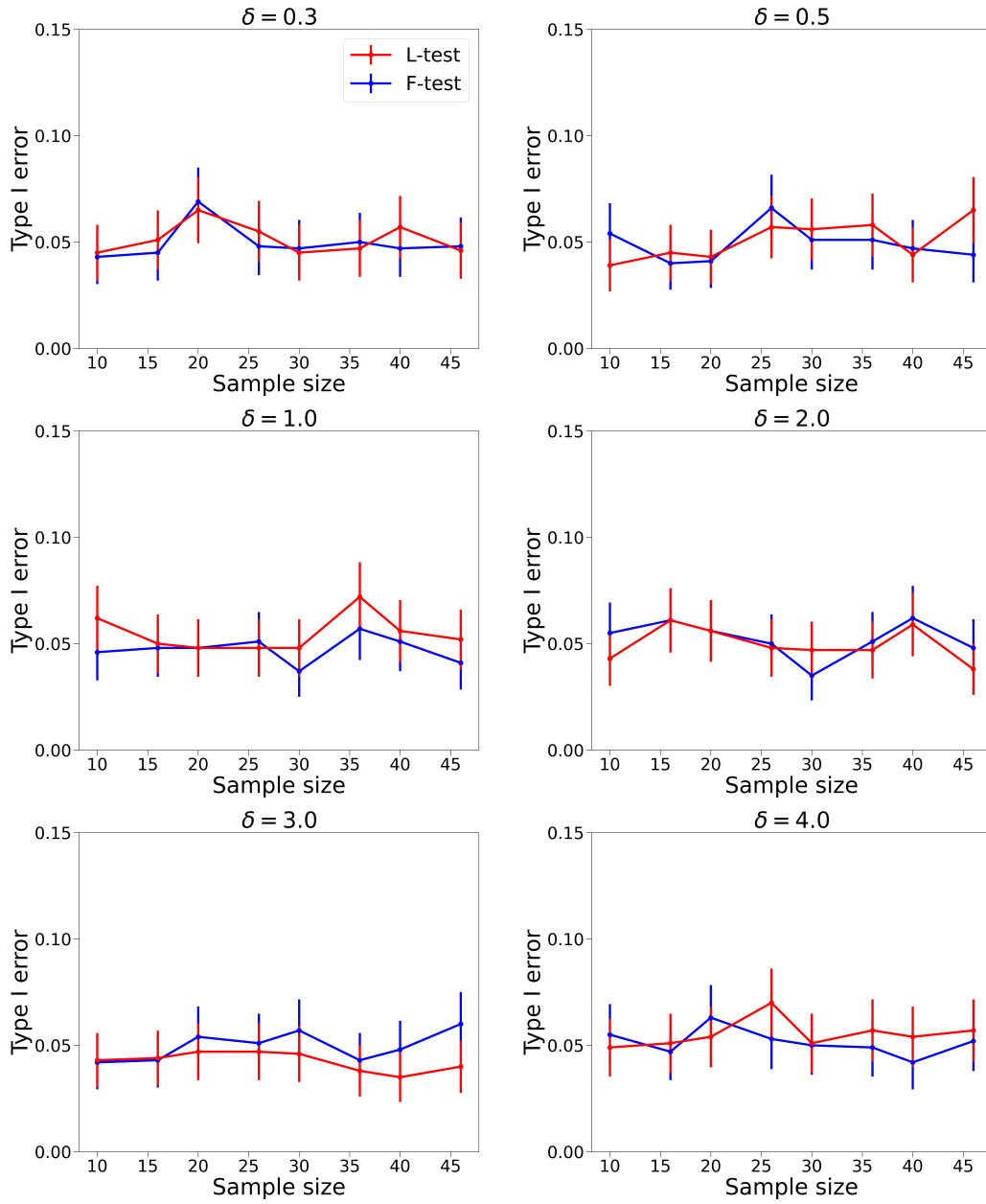


Figure 24: Same as Figure 20 but with $\rho = 0.5$.