Global Testing under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism

Ery Arias-Castro*, Emmanuel J. Candès† and Yaniv Plan‡

Abstract

Testing for the significance of a subset of regression coefficients in a linear model, a staple of statistical analysis, goes back at least to the work of Fisher who introduced the analysis of variance (ANOVA). We study this problem under the assumption that the coefficient vector is sparse, a common situation in modern high-dimensional settings. Suppose we have p covariates and that under the alternative, the response only depends upon on the order of $p^{1-\alpha}$ of those, $0 \le \alpha \le 1$. Under moderate sparsity levels, i.e. $0 \le \alpha \le 1/2$, we show that ANOVA is essentially optimal under some conditions on the design. This is no longer the case under strong sparsity constraints, i.e. $\alpha > 1/2$. In such settings, a multiple comparison procedure is often preferred and we establish its optimality when $\alpha \ge 3/4$. However, these two very popular methods are suboptimal, and sometimes powerless, under moderately strong sparsity where $1/2 < \alpha < 3/4$. We suggest a method based on the Higher Criticism that is powerful in the whole range $\alpha > 1/2$. This optimality property is true for a variety of designs, including the classical (balanced) multi-way designs and more modern '$p > n$' designs arising in genetics and signal processing. In addition to the standard fixed effects model, we establish similar results for a random effects model where the nonzero coefficients of the regression vector are normally distributed.

MSC 2000: Primary 62G10, 94A13; secondary 62G20.

Keywords: Detecting a sparse signal, analysis of variance, higher criticism, minimax detection, incoherence, random matrices, suprema of Gaussian processes, compressive sensing.

1 Introduction

1.1 The analysis of variance

Testing whether a subset of covariates have any linear relationship with a quantitative response has been a staple of statistical analysis since Fisher introduced the analysis of variance (ANOVA) in the 1920's [14]. Fisher developed ANOVA in the context of agricultural trials and the test has since then been one of the central tools in the statistical analysis of experiments [35]. As a consequence, there are countless situations in which it is routinely used; in particular, in the analysis of clinical trials [36] or in that of cDNA microarray experiments [7, 25, 37] to name just two important areas of biostatistics.

To begin with, consider the simplest design known as the one-way layout,

$$y_{ij} = \mu + \tau_j + z_{ij},$$

*Department of Mathematics, University of California, San Diego [eariasca@ucsd.edu] (corresponding author)

†Departments of Mathematics and of Statistics, Stanford University [candes@stanford.edu]

‡Applied and Computational Mathematics, California Institute of Technology [plan@caltech.edu]

where y_{ij} is the ith observation in group j, τ_j is the main effect for the jth treatment, and the z_{ij}'s are measurement errors assumed to be i.i.d. zero-mean normal variables. The goal is of course to determine whether there is any difference between the treatments. Formally, assuming there are p groups, the testing problem is

$$\begin{array}{ll} H_0: & \tau_1 = \tau_2 = \ldots = \tau_p = 0, \\ H_1: & \text{At least one } \tau_i \neq 0. \end{array}$$

The classical one-way analysis of variance is based on the well-known F-test calculated by all statistical software packages. A characteristic of ANOVA is that it tests for a *global* null and does not result in the identification of which τ_i's are nonzero.

Taking within-group averages reduces the model to

$$y_j = \beta_j + z_j, \quad j = 1, \ldots, p, \tag{1.1}$$

where $\beta_j = \mu + \tau_j$ and the z_j's are independent zero-mean Gaussian variables. If we suppose that the grand mean has been removed, so that the overall mean effect vanishes, i.e. $\mu = 0$, then the testing problem becomes

$$\begin{array}{ll} H_0: & \beta_1 = \beta_2 = \ldots = \beta_p = 0, \\ H_1: & \text{At least one } \beta_j \neq 0. \end{array} \tag{1.2}$$

In order to discuss the power of ANOVA in this setting, assume for simplicity that the variances of the error terms in (1.1) are known and identical, so that ANOVA reduces to a chi-square test that rejects for large values of $\sum_j y_j^2$. As explained before, this test does not identify which of the β_j's are nonzero but it has great power in the sense that it maximizes the minimum power against alternatives of the form $\{\boldsymbol{\beta} : \sum_j \beta_j^2 \geq B\}$ where $B > 0$. Such an appealing property may be shown via invariance considerations, see [32] and [27, Chap. 7-8].

1.2 Multiple testing and sparse alternatives

A different approach to the same testing problem is to test each individual hypothesis $\beta_j = 0$ versus $\beta_j \neq 0$, and combine these tests by applying a Bonferroni-type correction. One way to implement this idea is by computing the minimum P-value and compare it with a threshold adjusted to achieve a desired significance level. When the variances of the z_j's are identical, this is equivalent to rejecting the null when

$$\text{Max}(\mathbf{y}) = \max_j |y_j| \tag{1.3}$$

exceeds a given threshold. From now on, we will refer to this procedure as the Max test. Because ANOVA is such a well established method, it might surprise the reader – but not the specialist – to learn that there are situations where the Max test, though apparently naive, outperforms ANOVA by a wide margin. Suppose indeed that $z_j \sim \mathcal{N}(0, 1)$ in (1.1) and consider an alternative of the form $\max_j |\beta_j| \geq A$ where $A > 0$. In this setting, ANOVA requires A to be at least as large as $p^{1/4}$ to provide small error probabilities whereas the Max test only requires A to be on the order of $(2 \log p)^{1/2}$. When p is large, the difference is very substantial. Later in the paper, we shall prove that in an asymptotic sense, the Max test maximizes the minimum power against alternatives of this form. The key difference between these two different classes of alternatives resides in the kind of configurations of parameter values which make the likelihoods under H_0 and H_1 very close. For the alternative $\{\boldsymbol{\beta} : \sum_j \beta_j^2 \geq B\}$, the likelihood functions are hard to distinguish when the entries of $\boldsymbol{\beta}$ are of about the same size (in absolute value). For the other, namely, $\{\boldsymbol{\beta} : \max_j |\beta_j| \geq A\}$, the likelihood functions are hard to distinguish when there is a single nonzero coefficient equal to $\pm A$.

Multiple hypothesis testing with sparse alternatives is now commonplace, in particular in computational biology where the data is high-dimensional and we typically expect that only a few of the many measured variables actually contribute to the response – only a few assayed treatments may have a positive effect. For instance, DNA microarrays allow the monitoring of expression levels in cells for thousands of genes simultaneously. An important question is to decide whether some genes are differentially expressed; that is, whether or not there are genes whose expression levels are associated with a response such as the absence/presence of prostate cancer. A typical setup is that the data for the ith individual consists of a response or covariate y_i (indicating whether this individual has a specific disease or not) and a gene expression profile y_{ji}, $1 \leq j \leq p$. A standard approach consists in computing, for each gene j, a statistic T_j for testing the null hypothesis of equal mean expression levels and combining them with some multiple hypothesis procedure [12, 13]. A possible and simple model in this situation may assume $T_j \sim \mathcal{N}(0, 1)$ under the null while $T_j \sim \mathcal{N}(\beta_j, 1)$ under the alternative. Hence, we are in our sparse detection setup since one typically expects only a few genes to be differentially expressed. Despite the form of the alternative, ANOVA is still a popular method for testing the global null in such problems [25, 37].

1.3 This paper

Our exposition has thus far concerned simple designs, namely, the one-way layout or sparse mean model. This paper, however, is concerned with a much more general problem: we wish to decide whether or not a response depends linearly upon a few covariates. We thus consider the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}, \tag{1.4}$$

with an n-dimensional response $\mathbf{y} = (y_1, \ldots, y_n)$, a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (assumed to have full rank), and a noise vector, assumed to be i.i.d. standard normal. The decision problem (1.2) is whether all the β_i's are zero or not. We briefly pause to remark that statistical practitioners are familiar with the ANOVA derived F-statistic – also known as the model adequacy test – that software packages routinely provide for testing H_0. Our concern, however, is not at all model adequacy but rather, we view the test of the global null as a detection problem. In plain English, we would like to know whether there is signal or whether the data is just noise. A more general problem is to test whether a subset of coordinates of $\boldsymbol{\beta}$ are all zero or not, and as is well known, ANOVA is in this setup the most popular tool for comparing nested models. We emphasize that our results also apply to such general model comparisons, as we shall see later.

There are many applications of high-dimensional setups in which a response may depend upon only a few covariates. We give a few examples in the life sciences and in engineering; there are, of course, many others.

- *Genetics.* A single nucleotide polymorphism (SNP) is a form of DNA variation that occurs when at a single position in the genome, multiple (typically two) different nucleotides are found with positive frequency in the population of reference. One then collects information about allele counts at polymorphic locations. Almost all common SNPs have only two alleles so that one records a variable x_{ij} on individual i taking values in $\{0, 1, 2\}$ depending upon how many copies of, say, the rare allele one individual has at location j. One also records a quantitative trait y_i. Then the problem is to decide whether or not this quantitative trait has a genetic background. In order to scan the entire genome for a signal, one needs to screen between 300,000 and 1,000,000 SNPs. However, if the trait being measured has a genetic background, it will be typically regulated by a small number of genes. In this example, n is typically in the thousands while p is in the hundreds of thousands. The standard approach is to test each hypothesis $H_j : \beta_j \neq 0$ by using a statistic depending on the least-squares estimate $\hat{\beta}_j$ obtained by fitting the simple linear regression model

$$y_i = \hat{\beta}_0 + \hat{\beta}_j x_{ij} + r_{ij}. \tag{1.5}$$

The global null is then tested by adjusting the significance level to account for the multiple comparisons, effectively implementing a Max test; see [33, 39] for example.

- *Communications.* A multi-user detection problem typically assumes a linear model of the form (1.4), where the jth column of $\mathbf{X}$, denoted $\mathbf{x}_j$, is the channel impulse response for user j so that the received signal from the jth user is $\beta_j \mathbf{x}_j$ (we have $\beta_j = 0$ in case user j is not sending any message). Note that the mixing matrix $\mathbf{X}$ is often modeled as random with i.i.d. entries. In a strong noise environment, we might be interested in knowing whether information is being transmitted (some β_j's are not zero) or not. In some applications, it is reasonable to assume that only a few users are transmitting information at any given time. Standard methods include the matched filter detector, which corresponds to the Max test applied to $\mathbf{X}^T \mathbf{y}$, and linear detectors, which correspond to variations of the ANOVA F-test [21].

- *Signal detection.* The most basic problem in signal processing concerns the detection of a signal $S(t)$ from the data $y(t) = S(t) + z(t)$ where $z(t)$ is white noise. When the signal is non-parametric, a popular approach consists in modeling $S(t)$ as a (nearly) sparse superposition of waveforms taken from a dictionary $\mathbf{X}$, which leads to our linear model (1.4) (the columns of $\mathbf{X}$ are elements from this dictionary). For instance, to detect a multi-tone signal, one would employ a dictionary of sinusoids; to detect a superposition of radar pulses, one would employ a time-frequency dictionary [30, 31]; and to detect oscillatory signals, one would employ a dictionary of chirping signals. In most cases, these dictionaries are massively overcomplete so that we have more candidate waveforms than the number of samples, i. e. $p > n$. Sparse signal detection problems abound, e.g. the detection of cracks in materials [41], of hydrocarbon from seismic data [6] and of tumors in medical imaging [23].

- *Compressive sensing.* The sparse detection model may also arise in the area of compressive sensing [4, 5, 9], a novel theory which asserts that it is possible to accurately recover a (nearly) sparse signal – and by extension, a signal that happens to be sparse in some fixed basis or dictionary – from the knowledge of only a few of its random projections. In this context, the $n \times p$ matrix $\mathbf{X}$ with $n \ll p$ may be a random projection such as a partial Fourier matrix or a matrix with i.i.d. entries. Before reconstructing the signal, we might be interested in testing whether there is any signal at all in the first place.

All these examples motivate the study of two classes of sparse alternatives.

1. *Sparse fixed effects model (SFEM).* Under the alternative, the regression vector $\boldsymbol{\beta}$ has at least S nonzero coefficients exceeding A in absolute value.

2. *Sparse random effects model (SREM).* Under the alternative, the regression vector $\boldsymbol{\beta}$ has at least S nonzero coefficients assumed to be i.i.d. normal with zero mean and variance τ^2.

In both models, we set $S = p^{1-\alpha}$, where $\alpha \in (0, 1)$ is the sparsity exponent. Our purpose is to study the performance of various test statistics for detecting such alternatives.[1]

[1] We will sometimes put a prior on the support of $\boldsymbol{\beta}$ and on the signs of its nonzero entries in SFEM.

1.4 Prior work

To introduce our results and those of others, we need to recall a few familiar concepts from statistical decision theory. From now on, Ω denotes a set of alternatives, namely, a subset of $\mathbb{R}^p \setminus \{0\}$ and π is a prior on Ω. The Bayes risk of a test $T = T(\mathbf{X}, \mathbf{y})$ for testing $\boldsymbol{\beta} = \mathbf{0}$ versus $\boldsymbol{\beta} \sim \pi$ when H_0 and H_1 occur with the same probability, is defined as the sum of its probability of type I error (false alarm) and its average probability of type II error (missed detection). Mathematically,

$$\mathrm{Risk}_\pi(T) := \mathbb{P}_{\mathbf{0}}(T = 1) + \pi\left[\mathbb{P}_{\boldsymbol{\beta}}(T = 0)\right], \tag{1.6}$$

where $\mathbb{P}_{\boldsymbol{\beta}}$ is the probability distribution of $\mathbf{y}$ given by the model (1.4) and $\pi[\cdot]$ is the expectation with respect to the prior π. If we consider the linear model in the limit of large dimensions, i. e. $p \to \infty$ and $n = n(p) \to \infty$, and a sequence of priors $\{\pi_p\}$, then we say that a sequence of tests $\{T_{n,p}\}$ is asymptotically *powerful* if $\lim_{p\to\infty} \mathrm{Risk}_{\pi_p}(T_{n,p}) = 0$. We say that it is asymptotically *powerless* if $\liminf_{p\to\infty} \mathrm{Risk}_{\pi_p}(T_{n,p}) \geq 1$. When no prior is specified, the risk is understood as the worst-case risk defined as

$$\mathrm{Risk}(T) := \mathbb{P}_{\mathbf{0}}(T = 1) + \max_{\boldsymbol{\beta}\in\Omega} \mathbb{P}_{\boldsymbol{\beta}}(T = 0).$$

With our modeling assumptions, ANOVA for testing $\boldsymbol{\beta} = \mathbf{0}$ versus $\boldsymbol{\beta} \neq \mathbf{0}$ reduces to the chi-square test that rejects for large values of $\|\mathbf{Py}\|^{\mathbf{2}}$, where $\mathbf{P}$ is the orthogonal projection onto the range of $\mathbf{X}$. Since under the alternative, $\|\mathbf{Py}\|^{\mathbf{2}}$ has the chi-square distribution with $\min(n,p)$ degrees of freedom and noncentrality parameter $\|\mathbf{X}\boldsymbol{\beta}\|^2$, a simple argument shows that ANOVA is asymptotically powerless when

$$\|\mathbf{X}\boldsymbol{\beta}\|^2 / \sqrt{\min(n,p)} \to 0, \tag{1.7}$$

and asymptotically powerful if the same quantity tends to infinity.

Consider the sparse fixed effects alternative now. We prove that ANOVA is still essentially optimal under mild levels of sparsity corresponding to $\alpha \in [0, 1/2]$ but not under strong sparsity where $\alpha \in (1/2, 1]$. In the sparse mean model (1.1) where $\mathbf{X}$ is the identity, ANOVA is suboptimal requiring A to grow as a power of p; this is simply because (1.7) becomes $A^2 S/\sqrt{p} \to 0$ when all the nonzero coefficients are equal to A in absolute value. In contrast, the Max test is asymptotically powerful when A is on the order of $\sqrt{\log p}$ but is only optimal under very strong sparsity, namely, for $\alpha \in [3/4, 1]$. It is possible to improve on the Max test in the range $\alpha \in (1/2, 3/4)$ and we now review the literature which only concerns the sparse mean model, $\mathbf{X} = \mathbf{I}_p$. Set

$$\rho^*(\alpha) = \begin{cases} \alpha - 1/2, & 1/2 < \alpha < 3/4, \\ (1 - \sqrt{1-\alpha})^2, & 3/4 \leq \alpha < 1. \end{cases} \tag{1.8}$$

Then Ingster [22] showed that if $A = \sqrt{2r \log p}$ with $r < \rho^*(\alpha)$ fixed as $p \to \infty$, then all sequences of tests are asymptotically powerless. In the other direction, he showed that there is an asymptotically powerful sequence of tests if $r > \rho^*(\alpha)$. See also the work of Jin [24]. Donoho and Jin [8] analyzed a number of testing procedures in this setting, and in particular, the Higher Criticism of Tukey which rejects for large values of

$$HC^*(\mathbf{y}) = \sup_{t>0} \frac{\#\{i : |y_i| > t\} - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}},$$

where $\bar{\Phi}$ denotes the survival function of a standard normal random variable. They showed that the Higher Criticism is powerful within the detection region established by Ingster. Hall and Jin

[17, 18] have recently explored the case where the noise may be correlated, i.e. $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ and the covariance matrix $\mathbf{V}$ is known and has full rank. Letting $\mathbf{V} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$ be a Cholesky factorization of the covariance matrix, one can whiten the noise in $\mathbf{y} = \boldsymbol{\beta} + \mathbf{z}$ by multiplying both sides by $\mathbf{L}^{-\mathbf{1}}$, which yields $\tilde{\mathbf{y}} = \mathbf{L}^{-\mathbf{1}}\boldsymbol{\beta} + \tilde{\mathbf{z}}$; $\tilde{\mathbf{z}}$ is now white noise, and this is a special case of the linear model (1.4). When the design matrix is triangular with coefficients decaying polynomially fast away from the diagonal, [18] proves that the detection threshold remains unchanged, and that a form of Higher Criticism still achieves asymptotic optimality.

There are few other theoretical results in the literature, among which [15] develops a locally most powerful (score) test in a setting similar to SREM; here, 'locally' means that this property only holds for values of τ sufficiently close to zero. The authors do not provide any minimal value of τ that would guarantee the optimality of their method. However, since their score test resembles the ANOVA F-test, we suggest that it is only optimal for very small values of τ corresponding to mild levels of sparsity; i.e. $\alpha < 1/2$.

In the signal processing literature, a number of applied papers consider the problem of detecting a signal expressed as a linear combination in a dictionary [6, 16, 41]. However, the extraction of the salient signal is often the end goal of real signal processing applications so that research has focused on estimation rather than pure detection. As a consequence, one finds a literature entirely focused on estimation rather than on testing whether the data is just white noise or not. Examples of pure detection papers include [11] where the authors consider that $\boldsymbol{\beta}$ is known under the alternative so that the alternative is simple, and [19] which assumes that $\boldsymbol{\beta}$ is approximately known and examines the performance of the corresponding matched filter. Finally, [34] proposes a Bayesian approach for the detection of sparse signals in a sensor network for which the design matrix is assumed to have some polynomial decay in terms of the distance between sensors.

1.5 Our contributions

We show that if the predictor variables are not too correlated, there is a sharp detection threshold in the sense that no test is essentially better than a coin toss when the signal strength is below this threshold, and that there are statistics which are asymptotically powerful when the signal strength is above this threshold. This threshold is the same as that one gets for the sparse mean problem. Therefore, this work extends the earlier results and methodologies cited above [8, 17, 18, 22, 24], and is applicable to the modern high-dimensional situation where the number of predictors may greatly exceed the number of observations.

A simple condition under which our results hold is a low-coherence assumption.[2] Let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be the column vectors of $\mathbf{X}$, assumed to be normalized; this assumption is merely for convenience since it simplifies the exposition, and is not essential. Then if a large majority of all pairs of predictors have correlation less than γ with $\gamma = O(p^{-1/2+\varepsilon})$ for each $\varepsilon > 0$ (the real condition is weaker), then the results for the sparse mean model (1.1) apply almost unchanged. Interestingly, this is true even when the ratio between the number of observations and the number of variables is negligible, i.e. $n/p \to 0$. In particular, $A = \sqrt{2\rho^*(\alpha)\log p}$ is the sharp detection threshold for SFEM (sparse fixed effects model). Moreover, applying the Higher Criticism, not to the values of $\mathbf{y}$, but to those of $\mathbf{X}^T\mathbf{y}$ is asymptotically powerful as soon as the nonzero entries of $\boldsymbol{\beta}$ are above this threshold; this is true for all $\alpha \in (1/2, 1]$. In contrast, the Max test applied to $\mathbf{X}^T\mathbf{y}$ is only optimal in the region $\alpha \in [3/4, 1]$. We derive the sharp threshold for SREM as well, which is at $\tau = \sqrt{\alpha/(1-\alpha)}$. We show that the Max tests and the Higher Criticism are essentially optimal in

[2]Although we are primarily interested in the modern $p > n$ setup, our results apply *regardless* of the values of p and n.

this setting as well for all $\alpha \in (1/2, 1]$, that is, they are both asymptotically powerful as soon as the signal-to-noise ratio permits.

Before continuing, it may be a good idea to give a few examples of designs obeying the low-coherence assumption (weak correlations between most of the predictor variables) since it plays an important role in our analysis.

- *Orthogonal designs.* This is the situation where the columns of $\mathbf{X}$ are orthogonal so that $\mathbf{X}^T\mathbf{X}$ is the $p \times p$ identity matrix (necessarily, $p \leq n$). Here the coherence is of course the lowest since $\gamma(\mathbf{X}) = 0$.

- *Balanced, one-way designs.* As in a clinical trial comparing p treatments, assume a balanced, one-way design with k replicates per treatment group and with the grand mean already removed. This corresponds to the linear model (1.4) with $n = pk$ and, since we assume the predictors to have norm 1,

$$\mathbf{X} = \frac{1}{\sqrt{k}} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix} \in \mathbb{R}^{n \times p}, \tag{1.9}$$

  where each vector in this block representation is k-dimensional. This is in fact an example of orthogonal design. Note that our results apply even under the standard constraint $\mathbf{1}^T\boldsymbol{\beta} = 0$.

- *Concatenation of orthonormal bases.* Suppose that $p = nk$ and that $\mathbf{X}$ is the concatenation of k orthonormal bases in $\mathbb{R}^n$ jointly used as to provide an efficient signal representation. Then our result applies provided that $k = O(n^\varepsilon)$, $\forall \varepsilon > 0$ and that our bases are mutually incoherent so that γ is sufficiently small (for examples of incoherent bases, see e.g. [10]).

- *Random designs.* As in some compressive sensing and communications applications, assume that $\mathbf{X}$ has i.i.d. normal entries[3] with columns subsequently normalized (the column vectors are sampled independently and uniformly at random on the unit sphere). Such a design is close to orthogonal since $\gamma \leq \sqrt{5(\log p)/n}$ with high probability. This fact follows from a well-known concentration inequality for the uniform distribution on the sphere [26]. The exact same bound applies if the entries of $\mathbf{X}$ are instead i.i.d. Rademacher random variables.

We return to the discussion of our statistics and note that the Higher Criticism and the Max test applied to $\mathbf{X}^T\mathbf{y}$ are exceedingly simple methods with a straightforward implementation running in $O(np)$ flops. This brings us to two important points.

1. In the classical sparse mean model, Bonferroni-type multiple testing (the Max test) is not optimal when the sparsity level is moderately strong, i.e. when $1/2 < \alpha < 3/4$ [8]. This has direct implications in the fields of genetics and genomics where this is the prevalent method. The same is true in our more general model and it implies, for example, that the matched filter detector in wireless multi-user detection is suboptimal in the same sparsity regime.

   We elaborate on this point because this carries an important message. When the sparsity level is moderately strong, the Higher Criticism method we propose is powerful in situations where the signal amplitude is so weak that the Max test is powerless. *This says that one can detect a linear relationship between a response* $\mathbf{y}$ *and a few covariates even though those*

[3] This is a frequently discussed channel model in communications.

covariates that are most correlated with $\mathbf{y}$ *are not even in the model.* Put differently, if we assign a P-value to each hypothesis $\beta_j = 0$ (computed from a simple linear regression as discussed earlier), then *the case against the null is not in the tail of these P-values but in the bulk*; that is, the smallest P-values may not carry any information about the presence of a signal. In the situation we describe, the smallest P-values most often correspond to true null hypotheses, sometimes in such a way that the false discovery rate (FDR) cannot be controlled at any level below 1; and yet, the Higher Criticism has full power.

2. Though we developed the idea independently, the Higher Criticism applied to $\mathbf{X}^T\mathbf{y}$ is similar to the innovated Higher Criticism of Hall and Jin [18], which is specifically designed for time series. Not surprisingly, our results and arguments bear some resemblance with those of Hall and Jin [18]. We have already explained how their results apply when the design matrix is triangular (and in particular square) and has sufficiently rapidly decaying coefficients away from the diagonal. Our results go much further in the sense that 1) they include designs that are far from being triangular or even square, and 2) they include designs with coefficients that do not necessarily follow any ordered decay pattern. On the technical side, Hall and Jin astutely reduce matters to the case where the design matrix is banded, which greatly simplifies the analysis. In the general linear model, it is not clear how a similar reduction would operate especially when $n < p$ – at the very least, we do not see a way – and one must deal with more intricate dependencies in the noise term $\mathbf{X}^T\mathbf{z}$.

As we have remarked earlier, we have discussed testing the global null $\boldsymbol{\beta} = \mathbf{0}$ whereas some settings obviously involve nuisance parameters as in the comparison of nested models. Examples of nuisance parameters include the grand mean in a balanced, one-way design or, more generally, the main effects or lower-order interactions in a multi-way layout. In signal processing, the nuisance term may represent clutter as opposed to noise. In general, we have

$$\mathbf{y} = \mathbf{X}^{(0)}\boldsymbol{\beta}^{(0)} + \mathbf{X}^{(1)}\boldsymbol{\beta}^{(1)} + \mathbf{z},$$

where $\boldsymbol{\beta}^{(0)}$ is the vector of nuisance parameters, and $\boldsymbol{\beta}^{(1)}$ the vector we wish to test. Our results concerning the performance of ANOVA, the Higher Criticism or the Max test apply provided that the column spaces of $\mathbf{X}^{(0)}$ and $\mathbf{X}^{(1)}$ be sufficiently far apart. This occurs in lots of applications of interest. In the case of the balanced, multi-way design, these spaces are actually orthogonal. In signal processing, these spaces will also be orthogonal if the column space of $\mathbf{X}^{(0)}$ spans the low-frequencies while we wish to detect the presence of a high-frequency signal. The general mechanism which allows us to automatically apply our results is to simply assume that $\mathbf{P_0}\mathbf{X}^{(\mathbf{1})}$, where $\mathbf{P_0}$ is the orthogonal projector with the range of $\mathbf{X}^{(0)}$ as null space, obeys the conditions we have for $\mathbf{X}$.

1.6 Organization of the paper

The paper is organized as follows. In Section 2, we consider orthogonal designs and state results for the classical setting where no sparsity assumption is made on the regression vector $\boldsymbol{\beta}$, and the setting where $\boldsymbol{\beta}$ is mildly sparse. In Section 3, we study designs in which *most* pairs of predictor variables are only weakly correlated; this part contains our main results. In Section 4, we focus on some examples of designs with full correlation structure, in particular multi-way layouts with embedded constraints. Section 5 complements our study with some numerical experiments, and we close the paper with a short discussion, namely, Section 6. Finally, the proofs are gathered in the Appendix.

1.7 Notation

We provide a brief summary of the notations used throughout the paper. Set $[p] = \{1, \dots, p\}$ and for a subset $\mathcal{J} \subset [p]$, let $|\mathcal{J}|$ be its cardinality. Bold upper (resp. lower) case letters denote matrices (resp. vectors), and the same letter not bold represents its coefficients, e.g. a_j denotes the jth entry of $\mathbf{a}$. For an $n \times p$ matrix $\mathbf{A}$ with column vectors $\mathbf{a}_1, \dots, \mathbf{a}_p$, and a subset $\mathcal{J} \subset [p]$, $\mathbf{A}_\mathcal{J}$ denotes the n-by-$|\mathcal{J}|$ matrix with column vectors $\mathbf{a}_j$, $j \in \mathcal{J}$. Likewise, $\mathbf{a}_\mathcal{J}$ denotes the vector $(a_j, j \in \mathcal{J})$. The Euclidean norm of a vector is $\|\mathbf{a}\|$ and the sup-norm $\|\mathbf{a}\|_\infty$. For a matrix $\mathbf{A} = (a_{ij})$, $\|\mathbf{A}\|_\infty = \sup_{i,j} |a_{ij}|$, and this needs to be distinguished from $\|\mathbf{A}\|_{\infty,\infty}$, which is the operator norm induced by the sup norm, $\|\mathbf{A}\|_{\infty,\infty} = \sup_{\|\mathbf{x}\|_\infty \le 1} \|\mathbf{A}\mathbf{x}\|_\infty$. The Frobenius (Euclidean) norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F$. Φ (resp. ϕ) denotes the cumulative distribution (resp. density) function of a standard normal random variable, and $\bar{\Phi}$ its survival function. For brevity, we say that $\boldsymbol{\beta}$ is S-sparse if $\boldsymbol{\beta}$ has exactly S nonzero coefficients. Finally, we say that a random variable $X \sim F_X$ is stochastically smaller than $Y \sim F_Y$, denoted $X \le^{\text{sto}} Y$, if $F_X(t) \ge F_Y(t)$ for all scalar t.

2 Orthogonal Designs

This section introduces some results for the orthogonal design in which the columns of $\mathbf{X}$ are orthonormal, i.e. $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. While from the analysis viewpoint there is little difference with the case where $\mathbf{X}$ is the identity matrix, this is of course a special case of our general results, and this section may also serve as a little warm-up. Our first result, which is a special case of Proposition 2, determines the range of sparse alternatives for which ANOVA is essentially optimal.

Proposition 1 *Suppose $\mathbf{X}$ is orthogonal and that the sparsity exponent obeys $\alpha \in [0, 1/2]$.*

1. *In SFEM, all sequences of tests are asymptotically powerless if $A^2 S/p^{1/2} \to 0$.*
2. *In SREM, the conclusion is the same if $\tau^2 S/p^{1/2} \to 0$.*

Returning to our earlier discussion, it follows from (1.7) and the lower bound $\|\mathbf{X}\beta\|^2 = \|\beta\|^2 \ge A^2 S$ that ANOVA has full asymptotic power whenever $A^2 S/p^{1/2} \to \infty$. Therefore, comparing this with the content of Proposition 1 reveals that ANOVA is essentially optimal in the moderately sparse range corresponding to $\alpha \in [0, 1/2]$.

The second result of this section is that under an $n \times p$ orthogonal design, the detection threshold is the same as if $\mathbf{X}$ were the identity. We need a little bit of notation to develop our results. As in [8], define

$$\rho_{\text{Max}}(\alpha) = (1 - \sqrt{1 - \alpha})^2,$$

and observe that with $\rho^*(\alpha)$ as in (1.8),

$$\begin{cases} \rho^*(\alpha) < \rho_{\text{Max}}(\alpha), & 1/2 \le \alpha < 3/4, \\ \rho^*(\alpha) = \rho_{\text{Max}}(\alpha), & 3/4 \le \alpha \le 1. \end{cases}$$

We will also set a detection threshold for SREM defined by

$$\rho^*_{\text{rand}}(\alpha) = \sqrt{\alpha/(1 - \alpha)}. \tag{2.1}$$

With these definitions, the following theorem compares the performance of the Higher Criticism and the Max test.

Theorem 1 *Suppose $\mathbf{X}$ is orthogonal and assume the sparsity exponent obeys $\alpha \in (1/2, 1]$.*

1. *In SFEM, all sequences of tests are asymptotically powerless if* $A = \sqrt{2r \log p}$ *with* $r < \rho^*(\alpha)$. *Conversely, the Higher Criticism applied to* $|\mathbf{x}_1^T \mathbf{y}|, \ldots, |\mathbf{x}_p^T \mathbf{y}|$ *is asymptotically powerful if* $r > \rho^*(\alpha)$. *Also, the Max test is asymptotically powerful if* $r > \rho_{\mathrm{Max}}(\alpha)$ *and powerless if* $r < \rho_{\mathrm{Max}}(\alpha)$.

2. *In SREM, all sequences of tests are asymptotically powerless if* $\tau < \rho^*_{\mathrm{rand}}(\alpha)$. *Conversely, both the Higher Criticism and the Max test applied to* $|\mathbf{x}_1^T \mathbf{y}|, \ldots, |\mathbf{x}_p^T \mathbf{y}|$ *are asymptotically powerful if* $\tau > \rho^*_{\mathrm{rand}}(\alpha)$.

In the upper bounds, r *and* τ *are fixed while* $p \to \infty$.

To be absolutely clear, the statements for SFEM may be understood either in the worst-case risk sense or under the uniform prior on the set of S-sparse vectors with nonzero coefficients equal to $\pm A$. For SREM, the prior simply selects the support of $\boldsymbol{\beta}$ uniformly at random. After multiplying the observation by $\mathbf{X}^T$, matters are reduced to the case of the identity design for which the performance of the Higher Criticism and the Max test have been established in SFEM [8]. The result for the sparse random model is new and appears in more generality in Theorem 4.

To conclude, the situation concerning orthogonal designs is very clear. In SFEM for instance, if the sparsity level is such that $\alpha \le 1/2$, then ANOVA is asymptotically optimal whereas the Higher Criticism is optimal if $\alpha > 1/2$ (in contrast, the Max test is only optimal in the range $\alpha \ge 3/4$).

3 Weakly Correlated Designs

We begin by introducing a model of design matrices in which most of the variables are only weakly correlated. Our model depends upon two parameters, and we say that a $p \times p$ correlation matrix $\mathbf{C}$ belongs to the class $\mathcal{S}_p(\gamma, \Delta)$ if and only if it obeys the following two properties.

- *Strong correlation property.* This requires that for all $j \neq k$,

  $$|c_{jk}| \le 1 - (\log p)^{-1}.$$

  That is, *all* the correlations are bounded above by $1 - (\log p)^{-1}$. In the limit of large p, this is not an assumption and we will later explain how one can relax this even further.

- *Weak correlation property.* This is the main assumption and this requires that for all j,

  $$|\{k : |c_{jk}| > \gamma\}| \le \Delta.$$

  Note that for $\gamma \le 1$, $\Delta \ge 1$ since $c_{jj} = 1$. Fix a variable $\mathbf{x}_j$. Then at most $\Delta - 1$ other variables have a correlation exceeding γ with $\mathbf{x}_j$.

Our only real condition caps the number of variables that can have a correlation with any other above a threshold γ. An orthogonal design belongs to $\mathcal{S}_p(0, 1)$ since all the correlations vanish. With high probability, the Gaussian and Rademacher designs described earlier belong to $\mathcal{S}_p(\gamma, 1)$ with $\gamma = \sqrt{5(\log p)/n}$.

3.1 Lower bound on the detectability threshold

The main result of this paper is that if the predictor variables are not highly correlated, meaning that the quantities γ and Δ above are sufficiently small, then there are computable detection thresholds for our sparse alternatives that are very similar or identical to those available for orthogonal designs.

We begin by studying lower bounds and for SFEM, these may be understood either in a worst-case sense or under the prior where $\boldsymbol{\beta}$ is uniformly distributed among all S-sparse vectors with nonzero coefficients equal to $\pm A$. For SREM, these hold under a prior generating the support uniformly at random. We first consider mildly sparse alternatives.

Proposition 2 *Assume the sparsity exponent obeys* $\alpha \in [0, 1/2]$, *and suppose that* $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma, 1)$.

1. *In SFEM, all sequences of tests are asymptotically powerless if* $A^2 S(p^{-1/2} + \gamma \log p) \to 0$.
2. *In SREM, the conclusion is the same if* $\tau^2 S(p^{-1/2} + \gamma) \to 0$.

In Proposition 2 we have required that $\Delta = 1$ in order to derive sharp results. Moving now to sparser alternatives, we allow for Δ to increase with p, although very slowly, while the condition on γ remains essentially the same.

Theorem 2 *Assume the sparsity exponent obeys* $\alpha \in (1/2, 1]$, *and suppose that* $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma, \Delta)$ *with the following parameter asymptotics: 1)* $\Delta = O(p^\varepsilon)$, *for all* $\varepsilon > 0$, *and 2)* $\gamma p^{1-\alpha} (\log p)^4 \to 0$.

- *In SFEM, all sequences of tests are asymptotically powerless if* $A = \sqrt{2r \log p}$ *with* $r < \rho^*(\alpha)$.
- *In SREM, all sequences of tests are asymptotically powerless if* $\tau < \rho^*_{\text{rand}}(\alpha)$.

The result is essentially the same in the case of a balanced, multi-way design with the usual linear constraints. We comment on this point at the end of the proof of Theorem 2.

The reader may be surprised to see that the number n of observations does not explicitly appear in the above lower bounds. The sample size appears implicitly, however, since it must be large enough for the class $\mathcal{S}_p(\gamma, \Delta)$ to be nonempty. Assume $\Delta = 1$ for instance and that $p \geq n$. Then by the lower bound [38, Eq. (12)], we have

$$\gamma \geq \sqrt{(p-n)/(np)}. \tag{3.1}$$

For instance, $\gamma \geq 1/\sqrt{2n}$ if $p \geq 2n$.

As a technical aside, we remark that the lower bounds hold under the strong correlation assumption

$$|c_{jk}| \leq 1 - \delta,$$

for any $\delta < 1$, provided that $\gamma \delta^{-2} p^{1-\alpha} (\log p)^{3/2} \to 0$. We shall prove this more general statement, and the theorem is thus a special case corresponding to $\delta = (\log p)^{-1}$.

3.2 Upper bound on the detectability threshold

We now turn to upper bounds and unless stated otherwise, these assume the following models:

- For SFEM, we assume that $\boldsymbol{\beta}$ has a support generated uniformly at random and that its monzero coefficients have random signs.
- For SREM, we assume that $\boldsymbol{\beta}$ has a support generated uniformly at random.

We require that the support of $\boldsymbol{\beta}$ be generated uniformly at random and, in SFEM, that the signs of its coefficient be also random to rule out situations where cancellations occur making the signal strength potentially too small (and possibly vanish) to allow for reliable detection.

We begin by studying the performance of ANOVA when the alternative is not that sparse. We state our result for $\Delta = 1$ in accordance with the lower bound (Proposition 2) although the result holds when Δ obeys $\Delta = O(p^\varepsilon)$ for all $\varepsilon > 0$.

Proposition 3 *Assume the sparsity exponent obeys $\alpha \in [0, 1/2]$ and that $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma, 1)$.*

- *In SFEM, ANOVA is asymptotically powerful when $A^2 S/\sqrt{\min(n,p)} \to \infty$ and $\gamma \log p \to 0$.*
- *In SREM, ANOVA is asymptotically powerful when $\tau^2 S/\sqrt{\min(n,p)} \to \infty$ and $\gamma \to 0$.*

Compared with Proposition 2, the condition on γ is substantially weaker. More importantly, there appears to be a major discrepancy when n is negligible compared to p because $\sqrt{\min(n,p)}$ replaces $\sqrt{p}$. This is illusory, however, as the lower bound on γ displayed in (3.1) implies that the condition on A in Proposition 2 matches that of Proposition 3 up a $\log p$ factor.

Turning to sparser alternatives, we apply the Higher Criticism to $\mathbf{X}^T\mathbf{y}$ and for $t > 0$, put

$$H(t) = \frac{|\{j : |\mathbf{x}_j^T\mathbf{y}| > t\}| - 2p\bar{\Phi}(t)}{\sqrt{2p\bar{\Phi}(t)(1 - 2\bar{\Phi}(t))}}.$$

The Innovated Higher Criticism of Hall and Jin [18] resembles $\sup_{t>0} H(t) := HC^*(\mathbf{X}^T\mathbf{y})$, the main difference being that they apply a threshold to the entries of $\mathbf{X}$ before multiplying by $\mathbf{X}^T$. Here, to facilitate the analysis, we search for the maximum on a discrete grid and define

$$H^*(s) = \max\{H(t) : t \in [s, \sqrt{5\log p}] \cap \mathbb{N}\}.$$

Theorem 3 *Assume the sparsity exponent obeys $\alpha \in (1/2, 1]$ and that $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma, \Delta)$ with the following parameter asymptotics: 1) $\Delta = O(p^\varepsilon)$, for all $\varepsilon > 0$; 2) $\gamma^2 p^{1-\alpha}(\log p)^4 \to 0$ and 3) $\gamma^3 = O(p^{\varepsilon+5\alpha-4})$, for all $\varepsilon > 0$.*

- *In SFEM, the test based on $H^*(s_p(\alpha))$ with $s_p(\alpha) := \sqrt{2\min(1, 4\rho^*(\alpha))\log p}$ is asymptotically powerful against any alternative defined by $S = p^{1-\alpha'}$ with $\alpha' \geq \alpha$ and $A = \sqrt{2r\log p}$ with $r > \rho^*(\alpha')$.*
- *In SREM, the test based on $H^*(\sqrt{2\log p})$ is asymptotically powerful when $\tau > \rho^*_{\rm rand}(\alpha)$ regardless of $\alpha \in (1/2, 1]$, and without condition 3).*

In SREM, the conclusion is an immediate consequence of the behavior of the Max test stated in Theorem 4 and we, therefore, omit the proof. Having said this, the remarks below apply to SFEM.

1. The condition on γ is weaker than the condition required in Theorem 2, although the two conditions get ever closer as α approaches $1/2$.
2. The test based on $H^*(\sqrt{2\log p})$ is asymptotically powerful for all $\alpha \in [3/4, 1]$ (this test is closely related to the Max test).
3. Under the stronger assumption $\gamma = O(p^{-1/2+\varepsilon})$, for all $\varepsilon > 0$, the test rejecting for large values of $H^*(\sqrt{\log p/\log\log p})$, is asymptotically powerful for all values of $\alpha \in (1/2, 1]$.
4. Other discretizations in the definition of H^* would yield the same result. In fact, we believe the result holds without any discretization but we were not able to establish this in general. However, suppose that $p = kn$ and that $\mathbf{X}$ is the concatenation of k orthonormal bases. If $k = O(n^\varepsilon)$, for all $\varepsilon > 0$, the result holds without any discretization, meaning that rejecting for large values of $\sup_{t>0} H(t)$ is asymptotically powerful under the same conditions. This comes from leveraging the behavior (under the null) of the Higher Criticism – detailed in [8] – for each basis.

Turning our attention to the Max test now, the results available for orthogonal designs remain valid under similar conditions on the matrix $\mathbf{X}$.

Theorem 4 *Assume the sparsity exponent obeys* $\alpha \in (1/2, 1]$ *and that* $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma, \Delta)$ *with the following parameter asymptotics: 1)* $\Delta = O(p^\varepsilon)$*, for all* $\varepsilon > 0$*; 2)* $\gamma^2 p^{1-\alpha}(\log p)^2 \to 0$*.*

- *In SFEM, the Max test is asymptotically powerful if* $A \geq \sqrt{2r\log p}$ *with* $r > \rho_{\mathrm{Max}}(\alpha)$*, and powerless if* $r < \rho_{\mathrm{Max}}(\alpha)$*.*
- *In SREM, the Max test is asymptotically powerful for a fixed signal level obeying* $\tau > \rho^*_{\mathrm{rand}}(\alpha)$*. This holds regardless of* $\alpha \in (1/2, 1]$*.*

This theorem of course justifies the assertion made in the Introduction, which stated that one could detect a linear relationship between the response and a few covariates even though those covariates that were mostly correlated with the response were not in the model. In the proof, we use fine asymptotic results for the maximum of correlated normal random variables due to Berman [2].

3.3 Normal designs

A common assumption in multivariate statistics is that the rows of the design matrix are independent draws from the multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$. Our results apply provided that $\mathbf{\Sigma}$ obeys the assumptions about $\mathbf{X}^T\mathbf{X}$.

Corollary 1 *Suppose the rows of* $\mathbf{X}$ *are independent samples from* $\mathcal{N}(0, \mathbf{\Sigma})$*, and* $\mathbf{\Sigma} \in \mathcal{S}_p(\gamma, \Delta)$ *(the columns are normalized). Then the conclusions of Theorems 2, 3 and 4 are all valid provided that* $\sqrt{n^{-1}\log p}$ *obeys the conditions imposed on* γ*.*

We remark that if the columns are not normalized so that the rows of $\mathbf{X}$ are independent samples from $\mathcal{N}(0, \mathbf{\Sigma})$, the same result holds with a threshold A replaced by $A/\sqrt{n}$. This holds because the norm of each column is sharply concentrated around $\sqrt{n}$.

4 Some Special Designs

We now consider correlation matrices which have a substantial portion of large entries. In general, the detection threshold may depend upon some fine details of $\mathbf{X}$ but we give here some representative results applying to situations of interest.

We first examine the simple, yet important and useful example of constant correlation, where $\mathbf{x}_j^T\mathbf{x}_k = 1$ if $j = k$, and $= \gamma$ if $j \neq k$.[4] We impose $0 < \gamma < 1$ to make sure that $\mathbf{X}^T\mathbf{X}$ is at least positive definite as $p \to \infty$ (this implies that $\mathbf{X}^T\mathbf{X}$ has full rank which in turn imposes $p \leq n$). The balanced one-way design has this structure since it can be modeled by the matrix

$$\mathbf{X} = \frac{1}{\sqrt{2k}} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \\ -\mathbf{1} & -\mathbf{1} & \cdots & -\mathbf{1} \end{bmatrix},$$

where each vector in this block representation is k-dimensional. Without further assumptions on $\boldsymbol{\beta}$, this design is equivalent to (1.9) with the constraint $\mathbf{1}^T\boldsymbol{\beta} = 0$, except for the normalization. With this definition, $\mathbf{X}^T\mathbf{X}$ has diagonal entries equal to 1 and off-diagonal entries equal to 1/2 so we are in the setting – with $\gamma = 1/2$ – of our next result below.

[4] Whether such a family of vectors exists for special values of γ is a nontrivial matter, and we refer the reader to the literature on equiangular lines, see [28] for example.

Theorem 5 *Suppose that* $\mathbf{x}_j^T \mathbf{x}_k$ *is equal to* 1 *if* $j = k$ *and* γ *otherwise, and that the sparsity exponent obeys* $\alpha \in (1/2, 1]$*. Then without further assumption, the conclusions of Theorems 2, 3 and 4 remain valid with the bounds on* A *and* τ *divided by* $\sqrt{1-\gamma}$*.*

The balanced, one-way design may be seen either as an orthogonal design with a linear constraint, or a constant-correlation design without any constraint. More generally, a multi-way design is easily defined as an orthogonal design with a set of linear constraints. Specifically, suppose the coordinates of $\boldsymbol{\beta}$ are indexed by an m-dimensional index vector, so that

$$\boldsymbol{\beta} = (\beta_{\mathbf{j}} : \mathbf{j} = (j_1, \dots, j_m),\ j_s \in [p_s]), \quad p = \prod_{s=1}^{m} p_s.$$

We assume the design is balanced with k replicates per cell so that $n = pk$. With any fixed order on the index set, say the lexicographic order, the design matrix is the same as in the balanced, one-way design (1.9). Here, $\boldsymbol{\beta}$ obeys the linear constraints

$$\sum_{s \neq t} \sum_{j_s=1}^{p_s} \beta_{j_1 \cdots j_m} = 0, \tag{4.1}$$

for all $j_t \in [p_t]$ and $t \in [m]$ (there are $\sum_{t=1}^{m} p_t$ constraints). As in the balanced, one-way design, Theorem 1 applies to the balanced, multi-way design. The argument for the lower bound is at the end of the proof of Theorem 2. The proof of the upper bounds is exactly as in the case of any other orthogonal design. Finally, embedding the linear constraints into the design matrix leads to a family of designs with a 'full' correlation structure with off-diagonal elements which, in general, are not of the same magnitude unless the design is one-way.

5 Numerical Experiments

This section complements our study with some numerical simulations, which illustrate the empirical performance for finite sample sizes. Here, $\mathbf{X}$ is an $n \times p$ Gaussian design with i.i.d. standard normal entries, and normalized columns. We study fixed effects and investigate the performance of ANOVA, the Higher Criticism[5] and the Max test. We also compare the detection limits with those available in the case of the $p \times p$ identity design, since the theory developed in Corollary 1 predicts that the detection boundaries are asymptotically identical (provided n grows sufficiently rapidly).

We performed simulations with matrices of sizes $500 \times 10{,}000$, $2{,}000 \times 10{,}000$, $1{,}000 \times 100{,}000$ and $5{,}000 \times 100{,}000$, various sparsity levels, and strategically selected values of r. Each data point corresponds to an average over 1,000 trials in the case where $p = 10{,}000$, and over 500 trials when $p = 100{,}000$. A new design matrix is sampled for each trial. The performance of each of the three methods is computed in terms of its best (empirical) risk defined as the sum of probabilities of type I and II errors achievable across all thresholds. The results are reported in Figures 1 and 2. As expected, the detection thresholds for the Gaussian design are quite close to those available for the identity design. The performance of ANOVA improves very quickly as the sparsity decreases, dominating the Max test with $S = \sqrt{p}$; its performance also improves as n becomes smaller, in accordance with (1.7). The performance of the Max test follows the opposite pattern, degrading as S increases. Interestingly, the Higher Criticism remains competitive across the different sparsity levels.

[5] We do not use the discretization here.

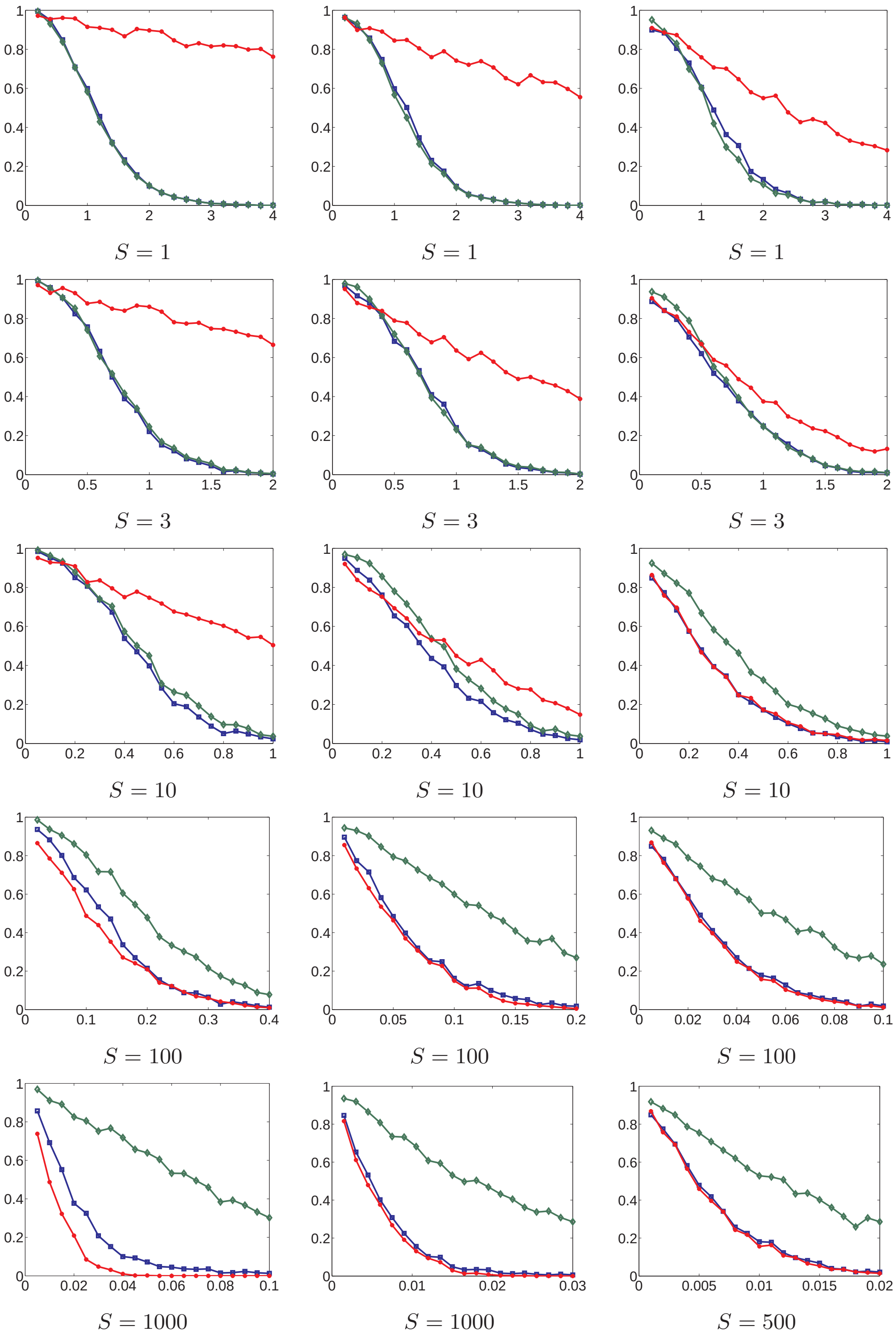

Figure 1: Left column: identity design with $p = 10{,}000$. Middle column: Gaussian design with $p = 10{,}000$ and $n = 2{,}000$. Right column: Gaussian design with $p = 10{,}000$ and $n = 500$. Sparsity level S is indicated below each plot. In each plot, the empirical risk (based on 1,000 trials) of each method [ANOVA (red bullets); Higher Criticism (blue squares); Max test (green diamonds)] is plotted against r (note the different scales).

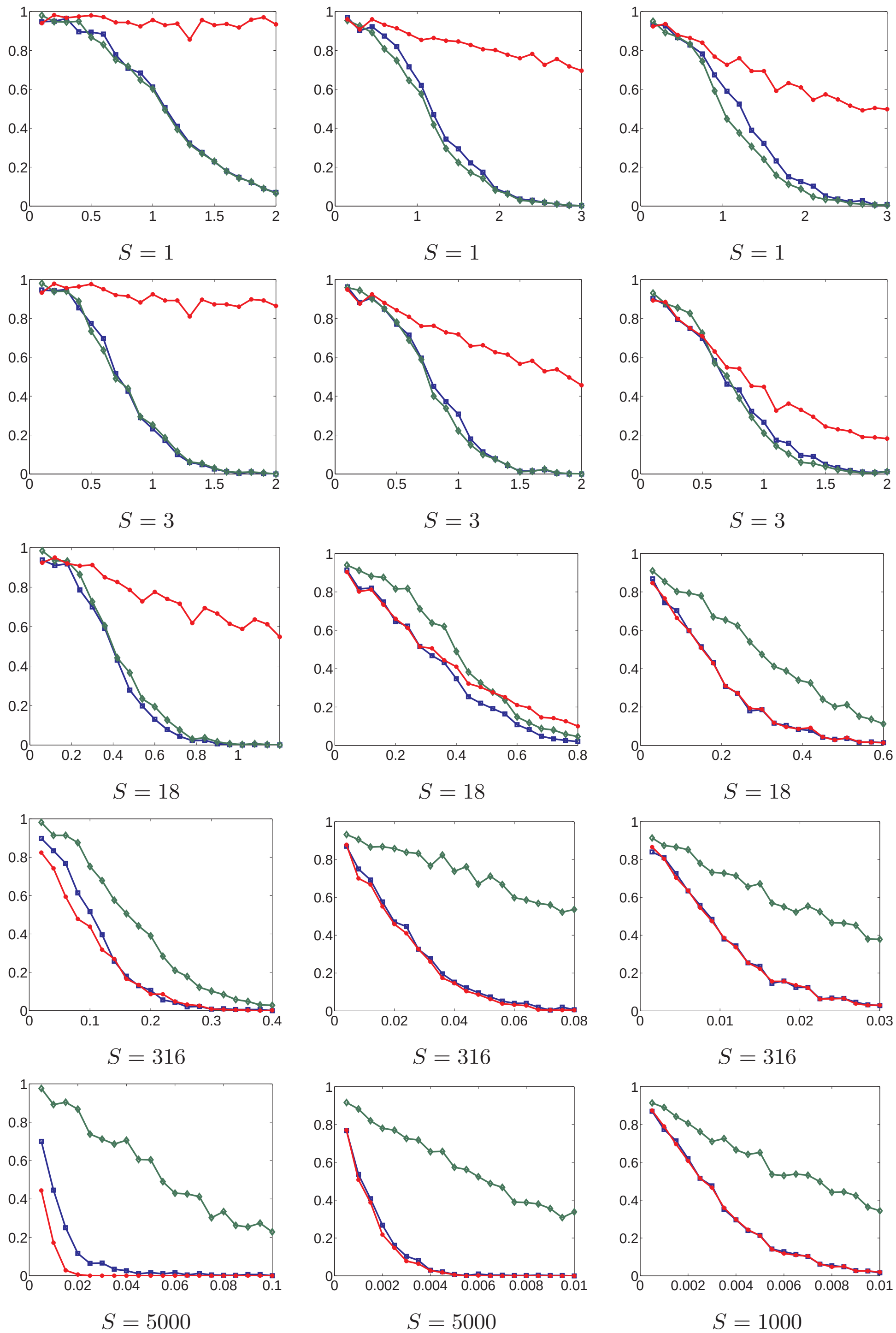

Figure 2: Left column: identity design with $p = 100{,}000$. Middle column: Gaussian design with $p = 100{,}000$ and $n = 5{,}000$. Right column: Gaussian design with $p = 100{,}000$ and $n = 1{,}000$. Sparsity level S is indicated below each plot. In each plot, the empirical risk (based on 500 trials) of each method [ANOVA (red bullets); Higher Criticism (blue squares); Max test (green diamonds)] is plotted against r (note the different scales).

6 Discussion

It is possible to extend our results to setups with correlated errors, with known covariance. As discussed in Section 1, suppose $\mathbf{z}$ in (1.4) is $\mathcal{N}(\mathbf{0}, \mathbf{V})$. We may then whiten the noise by multiplying both sides of (1.4) by $\mathbf{L}^{-1}$, where $\mathbf{L}\mathbf{L}^T$ is a Cholesky decomposition of $\mathbf{V}$. This leads to a model of the form

$$\mathbf{y} = \mathbf{L}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{z},$$

which is our problem with $\mathbf{L}^{-1}\mathbf{X}$ instead of $\mathbf{X}$. In some situations, the noise covariance matrix may not be known and we refer to [18] for a brief discussion of this issue.

Although several generalizations are possible, an interesting open problem is to determine the detection boundary for a given sequence of designs $\{\mathbf{X}_{n\times p}\}$ with n and p growing to infinity. We have seen that if most of the predictor variables are only weakly correlated, then the detection boundary is as if the predictors were orthogonal. Similar conclusions for certain types of square designs in which $n = p$ are also presented in the work of Hall and Jin [18]. Although we introduced some sharp results in Section 4 corresponding to some important design matrices, the class of matrices for which we have definitive answers is still quite limited. We hope other researchers will engage this area of research and develop results towards a general theory.

A Proofs of the Main Results

Below, $\mathcal{J}(\boldsymbol{\beta})$ denotes the support of $\boldsymbol{\beta}$, i.e. the subset indexing the coordinates of $\boldsymbol{\beta}$ that are nonzero. Whenever possible, we will routinely omit the dependence on $\boldsymbol{\beta}$ to lighten the notation. Throughout, C denotes a numerical constant whose value may change with each appearance.

A.1 Preliminaries

We will use the following results multiple times. The first is classical [40].

Lemma 1 *For $t \geq 1$,*

$$\left(1 - \frac{1}{t^2}\right)\frac{\phi(t)}{t} \leq \bar{\Phi}(t) \leq \frac{\phi(t)}{t}.$$

In particular, when $t \geq 1$ and $a \to 0$ such that $at \to 0$, $\bar{\Phi}(t + a) = (1 + O(at))\bar{\Phi}(t)$.

Another useful result is this:

Lemma 2 *For $U, V \sim \mathcal{N}(0, 1)$ with $\mathrm{Cov}(U, V) = \rho$, and $a, b \in \mathbb{R}$,*

$$\mathbb{E}\left(e^{aU}\mathbf{1}_{\{V>b\}}\right) \leq e^{a^2/2}\bar{\Phi}(b - a\rho).$$

Proof. This comes from a direct integration, see [18, Lem. A.10]. □

Next, $\mathrm{Hyp}(N, m, n)$ denotes the hypergeometric distribution counting the number of red balls in n draws from an urn containing m red balls out of N. The following result compares the hypergeometric distribution with a related binomial distribution, which is in general simpler to work with.

Lemma 3 $\mathrm{Hyp}(N, m, n)$ *is stochastically smaller than* $\mathrm{Bin}(n, m/(N - m))$.

Proof. Suppose the balls are picked one by one without replacement. At each stage, the probability of selecting a red ball is smaller than $m/(N-m)$. The result follows. □

The likelihood ratio corresponding to a prior π on the set of alternatives is given by

$$W = \pi\left[\exp(\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \|\mathbf{X}\boldsymbol{\beta}\|^2/2)\right], \tag{A.1}$$

where $\pi[\cdot]$ denotes the expectation with respect to π. By the Neyman-Pearson fundamental Lemma [27, Th. 3.2.1], the likelihood ratio test $T := \{W > 1\}$ minimizes the average risk (1.6). Moreover, we have the fundamental lower bound

$$\mathrm{Risk}_\pi(T) \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}\left(W^2\right) - 1}.$$

Therefore, all sequences of tests are powerless if $\limsup \mathbb{E}\left(W^2\right) = 1$ and $\mathbb{E}$ denotes expectation with respect to $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

A.2 Proof of Proposition 2

Let π be a prior on the set of alternatives. As argued in Section A.1, it is enough to show that $\limsup \mathbb{E}\left(W^2\right) = 1$. By Fubini's theorem we may integrate with respect to $\mathbf{y}$ first to obtain

$$\mathbb{E}\left(W^2\right) = \pi^{\otimes 2}\left[\exp(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}')\right] = \pi^{\otimes 2}\left[\mathrm{I}\cdot\mathrm{II}\right],$$

where $\boldsymbol{\beta}, \boldsymbol{\beta}' \sim^{\text{iid}} \pi$, and

$$\mathrm{I} = \exp(\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} - \mathbf{I})\boldsymbol{\beta}'), \quad \mathrm{II} = \exp(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}')$$

Then by the Cauchy-Schwartz inequality, it is enough to show that

$$\pi^{\otimes 2}\left[\mathrm{I}^2\right] \leq 1 + o(1), \quad \pi^{\otimes 2}\left[\mathrm{II}^2\right] \leq 1 + o(1).$$

Let $\mathcal{J}, \mathcal{J}'$ denote the respective supports of $\boldsymbol{\beta}, \boldsymbol{\beta}'$.

A.2.1 Fixed effects model

Lemma 4 *Fix $\mathbf{R} \in \mathbb{R}^{q\times q}$ and let $\boldsymbol{\xi}, \boldsymbol{\xi}'$ be independent samples from the uniform distribution on $\{-1,1\}^q$. Then*

$$\mathbb{E}\left(\exp(\boldsymbol{\xi}^T\mathbf{R}\boldsymbol{\xi}')\right) \leq \exp(2\|\mathbf{R}\|_F\log(3q)) + \|\mathbf{R}\|_F.$$

For the fixed effects model, let π be the uniform prior on the set of S-sparse vectors with all nonzero coefficients equal to $\pm A$. For $\pi^{\otimes 2}\left[\mathrm{I}^2\right]$, we apply Lemma 4 with $\boldsymbol{\xi} = \boldsymbol{\beta}_\mathcal{J}/A$, $\boldsymbol{\xi} = \boldsymbol{\beta}'_{\mathcal{J}'}/A$ and $\mathbf{R} = A^2(\mathbf{X}_\mathcal{J}^T\mathbf{X}_{\mathcal{J}'} - \mathbf{I}_S)$. Since $\mathbf{R}$ has Frobenius norm bounded by $A^2\gamma S$, this gives

$$\pi^{\otimes 2}\left[\mathrm{I}^2\right] \leq \exp(4A^2\gamma S\log p) + 2A^2\gamma S.$$

It follows from our assumptions that $A^2\gamma S\log p = o(1)$ and, therefore, $\pi^{\otimes 2}\left[\mathrm{I}^2\right] \leq 1 + o(1)$. Next,

$$\pi^{\otimes 2}\left[\mathrm{II}^2\right] = \pi^{\otimes 2}\left[\exp(4A^2U - 2A^2|\mathcal{J}\cap\mathcal{J}'|)\right],$$

where

$$U = |\{j \in \mathcal{J}\cap\mathcal{J}' : \beta_j = \beta'_j\}|.$$

Given $|\mathcal{J} \cap \mathcal{J}'|$, U is binomial with $|\mathcal{J} \cap \mathcal{J}'|$ trials and a probability of success equal to 1/2. The expression for the binomial moment generating function gives

$$\pi^{\otimes 2}\left[\mathrm{II}^2\right] = \pi^{\otimes 2}\left[\left(1 - \frac{1}{2} + \frac{1}{2}e^{4A^2}\right)^{|\mathcal{J}\cap\mathcal{J}'|} e^{-2A^2|\mathcal{J}\cap\mathcal{J}'|}\right] = \pi^{\otimes 2}\left[\cosh(2A^2)^{|\mathcal{J}\cap\mathcal{J}'|}\right].$$

Under $\pi^{\otimes 2}$, $|\mathcal{J} \cap \mathcal{J}'| \sim \mathrm{Hyp}(p, S, S)$, which is stochastically bounded by $\mathrm{Bin}(S, S/(p - S))$ by Lemma 3. Hence, using $\cosh(s) = 1 + s^2/2 + o(s^2)$ for small s, for A small enough and p large enough we get

$$\pi^{\otimes 2}\left[\cosh(2A^2)^{|\mathcal{J}\cap\mathcal{J}'|}\right] \le \left(1 - \frac{S}{p-S} + \frac{S}{p-S}\cosh(2A^2)\right)^S \le \left(1 + \frac{4S}{p}A^4\right)^S,$$

and the last expression tends to one if $A^4S^2/p \to 0$. Hence, $\pi^{\otimes 2}\left[\mathrm{II}^2\right] \le 1 + o(1)$.

A.2.2 Random effects model

For the random effects model, π is the prior on S-sparse vectors with uniform support and nonzero coefficients independently sampled from $\mathcal{N}(0, \tau^2)$. The following result plays the role of Lemma 4.

Lemma 5 *Let $\mathbf{R}$ be a $q \times q$ matrix with zeroes on the diagonal. Then for $\boldsymbol{\xi}, \boldsymbol{\xi}' \sim^{\mathrm{iid}} \mathcal{N}(\mathbf{0}, \mathbf{I})$,*

$$\mathbb{E}\left(\exp(\boldsymbol{\xi}^T\mathbf{R}\boldsymbol{\xi}')\right) \le 1 + 2\|R\|_F^2,$$

provided that $\|\mathbf{R}\|_F^2 \le 1/2$.

Using Lemma 5 as we used Lemma 4, this time combined with the fact that $\tau^2\gamma S = o(1)$ gives $\pi^{\otimes 2}\left[\mathrm{I}^2\right] \le 1 + o(1)$. For the other term, observe that for each $t \in (0, 1)$ and $z, z' \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1)$,

$$\mathbb{E}\left(\exp(tzz')\right) = (1 - t^2)^{-1/2},$$

which gives

$$\pi^{\otimes 2}\left[\mathrm{II}^2\right] = \pi^{\otimes 2}\left[(1 - 4\tau^4)^{-|\mathcal{J}\cap\mathcal{J}'|/2}\right].$$

Since $|\mathcal{J}\cap\mathcal{J}'|$ is stochastically bounded by $\mathrm{Bin}(S, S/(p-S))$ as before and $(1-s)^{-1/2} = 1+s/2+o(s)$ for small s, for τ small enough and p large enough, we have

$$\pi^{\otimes 2}\left[(1 - 4\tau^4)^{-|\mathcal{J}\cap\mathcal{J}'|/2}\right] \le \left(1 - \frac{S}{p-S} + \frac{S}{p-S}(1 - 4\tau^4)^{-1/2}\right)^S \le \left(1 + \frac{4S}{p}\tau^4\right)^S,$$

The last term tends to 1 if $\tau^4S^2/p \to 0$. Hence, $\pi^{\otimes 2}\left[\mathrm{II}^2\right] \le 1 + o(1)$.

A.3 Proof of Proposition 3

Let π be the appropriate prior on the set of alternatives. It suffices to prove that (1.7) holds with high probability under π. We have

$$\|\mathbf{X}\boldsymbol{\beta}\|^2 = U + \boldsymbol{\beta}^T\boldsymbol{\beta}, \quad U := \boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X} - \mathbf{I})\boldsymbol{\beta}.$$

Since $\boldsymbol{\beta}^T\boldsymbol{\beta} \ge A^2S$ in SFEM (resp. $\boldsymbol{\beta}^T\boldsymbol{\beta} = (1 + o_P(1))\tau^2S$ in SREM), and the distribution of U is symmetric about 0, it suffices to prove that $U \le A^2S/2$ (resp. $\le \tau^2S/2$), with high probability under π.

A.3.1 Fixed effects model

For the fixed effects model, let π be the uniform prior on the set of S-sparse vectors with all nonzero coefficients equal to $\pm A$. By Markov's inequality, we have

$$\pi\left[U > A^2S/2\right] \leq \pi\left[\exp(U - A^2S/2)\right].$$

We then apply Lemma 4 and obtain

$$\pi\left[\exp(U - A^2S/2)\right] \leq (\exp(2A^2\gamma S\log p) + A^2\gamma S)\exp(-A^2S/2) \to 0,$$

since $\gamma\log p \to 0$ and $A^2S \to \infty$.

A.3.2 Random effects model

For the random effects model, π is the prior on S-sparse vectors with uniform support and with nonzero coefficients independently sampled from $\mathcal{N}(0,\tau^2)$. In general, we cannot apply Markov's inequality together with Lemma 5 since we may have $\tau^2\gamma S \to \infty$. We use Chebychev's inequality instead:

$$\pi\left[U > \tau^2S/2\right] \leq \frac{\pi\left[U^2\right]}{(\tau^2S/2 - \pi\left[U\right])_+^2}.$$

For the first moment, the 'trace trick' gives

$$\pi\left[U\right] = \tau^2\mathrm{trace}(\mathbf{X}^T\mathbf{X} - \mathbf{I}) = 0.$$

For the second moment, a direct calculation shows that

$$\pi\left[U^2\right] = 2\tau^4\|\mathbf{X}^T\mathbf{X} - \mathbf{I}\|_F^2.$$

Since $\|\mathbf{X}^T\mathbf{X} - \mathbf{I}\|_F \leq \gamma S$, applying Chebychev's inequality gives

$$\pi\left[U > \tau^2S/2\right] \leq 8\gamma^2 \to 0.$$

A.4 Proof of Theorem 2

The structure of the proof parallels that of [18, Th. 3.1], and operates via the so-called moment method. The main difference is that we cannot reduce matters to banded matrices as in [18, Sec. 10.2].

A.4.1 Fixed effects model

Without loss of generality, we assume that $\gamma \geq 1/p$. For $j \in [p]$, set $\mathcal{I}_j = \{k : |\mathbf{x}_j^T\mathbf{x}_k| \geq \gamma\}$ and observe that $|\mathcal{I}_j| \leq \Delta$. Define the symmetric relation $j \leftrightarrow k$ if and only if $j \in \mathcal{I}_k$ (or equivalently, iff $k \in \mathcal{I}_j$), and for a subset $\mathcal{J} \subset [p]$, $k \leftrightarrow \mathcal{J}$ if $k \in \mathcal{I}_j$ for some $j \in \mathcal{J}$. Recall the bound

$$\Delta = O(p^\varepsilon), \quad \forall\varepsilon > 0. \tag{A.2}$$

Let π_0 be the uniform prior on the set of S-sparse vectors with all nonzero coefficients equal to $\pm A$, and let π be the restriction of π_0 to those vectors with support $\mathcal{J}$ obeying

$$|\mathcal{I}_j \cap \mathcal{J}| \leq 1, \text{for all } j \in [p]. \tag{A.3}$$

The advantage of using π instead of π_0 is that the correlations indexed by $\mathcal{J}$ satisfying (A.3) are small, yielding the bounds

$$A\mathbf{1}_{\{j\in\mathcal{J}\}} - A\gamma S \le |\mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}| \le A\mathbf{1}_{\{j\leftrightarrow\mathcal{J}\}} + A\gamma S. \tag{A.4}$$

(Although this bound is sufficient for our purposes, it can and will be refined later in (A.9).) In fact, π and π_0 are essentially equivalent. The following result is the equivalent of [18, Lem. A.8], and justifies our use of π.

Lemma 6 *When $\mathcal{J}$ is uniformly distributed among the subsets of $[p]$ of size S,*

$$\mathbb{P}\left(\exists j : |\mathcal{I}_j \cap \mathcal{J}| \ge 2\right) \le 2S\Delta/p = O(p^{1-2\alpha+\varepsilon}), \quad \forall\varepsilon > 0.$$

Proof. The union bound together with $|\mathcal{I}_j \cap \mathcal{J}| \sim \mathrm{Hyp}(p, |\mathcal{I}_j|, S)$ and Lemma 3, give get

$$\mathbb{P}\left(\exists j \in [p] : |\mathcal{I}_j \cap \mathcal{J}| \ge 2\right) \le p\mathbb{P}\left(\mathrm{Bin}(S, \Delta/(p-\Delta)) \ge 2\right) \le (S\Delta)^2/p$$

for p large enough. Above, we are using the relation $\mathbb{P}\left(\mathrm{Bin}(m,q) \ge 2\right) = (1+o(1))(mq)^2/2$ valid when $mq \to 0$. The conclusion follows from (A.2). □

Let $\mathbf{u} = \mathbf{X}^T\mathbf{z}$ and $t_p = \sqrt{2\log p}$, and define $D = \{|u_j| \le t_p, \forall j = 1,\dots,p\}$. Recall that $\mathbb{E}$ denotes the expectation with respect to $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $\pi[\cdot]$ that with respect to $\boldsymbol{\beta} \sim \pi$. Now a standard refinement – the details are omitted – shows that the risk goes to 1 if $\mathbb{E}\left(W\mathbf{1}_{D^c}\right) = o(1)$ and $\mathbb{E}\left(W^2\mathbf{1}_D\right) \le 1 + o(1)$. We first show that $\mathbb{E}(W\mathbf{1}_{D^c}) \to 0$. We have

$$\begin{aligned}
\mathbb{E}(W\mathbf{1}_{D^c}) &\le \sum_{j\in[p]} \mathbb{E}\left(W\mathbf{1}_{\{|u_j|>t_p\}}\right) \\
&\le \sum_{j\in[p]} \pi\left[\bar{\Phi}(t_p - |\mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}|)\right] \\
&\le (1+o(1)) \sum_{j\in[p]} \pi\left[\bar{\Phi}(t_p - A\mathbf{1}_{\{\mathcal{I}_j\cap\mathcal{J}\neq\emptyset\}})\right] \\
&\le (1+o(1))\left(p\bar{\Phi}(t_p) + \Delta S\bar{\Phi}(t_p - A)\right) \\
&\le C\left(\frac{1}{\sqrt{\log p}} + \Delta p^{1-\alpha-(1-\sqrt{r})^2}\right).
\end{aligned}$$

The second inequality uses Lemma 2 combined with $\mathbf{u}^T\boldsymbol{\beta} \sim \mathcal{N}(0, \|\mathbf{X}\boldsymbol{\beta}\|^2)$, $u_j \sim \mathcal{N}(0,1)$ and $\mathrm{Cov}(\mathbf{u}^T\boldsymbol{\beta}, u_j) = \mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}$. The third uses (A.4) and Lemma 1 together with $t_pA\gamma S \to 0$. The fourth uses the equivalence $\mathcal{I}_j\cap\mathcal{J} \neq \emptyset \Leftrightarrow j \in \cup_{k\in\mathcal{J}}\mathcal{I}_k$ and the definition of Δ, while the last uses Lemma 1. We conclude that the last expression is $o(1)$ since $1-\alpha-(1-\sqrt{r})^2 < 0$ for $r < \rho^*(\alpha)$, and (A.2) holds.

We now turn to proving that $\mathbb{E}\left(W^2\mathbf{1}_D\right) \le 1 + o(1)$. By Fubini's theorem, we have

$$\mathbb{E}(W^2\mathbf{1}_D) = \pi^{\otimes 2}\mathbb{E}\left(\exp\left(\mathbf{u}^T(\boldsymbol{\beta}+\boldsymbol{\beta}') - \|\mathbf{X}\boldsymbol{\beta}\|^2/2 - \|\mathbf{X}\boldsymbol{\beta}'\|^2/2\right)\prod_{j\in[p]}\mathbf{1}_{\{|u_j|\le t_p\}}\right). \tag{A.5}$$

If $\mathcal{J}$ and $\mathcal{J}'$ are the respective supports of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, which are i.i.d. samples from π, define $\mathcal{K} = \mathcal{J} \cup \mathcal{J}'$ and

$$\mathcal{L} = \{(j,k) \in \mathcal{J}\times\mathcal{J}' \text{ or } \mathcal{J}'\times\mathcal{J} : j \neq k,\, j \leftrightarrow k\},$$
$$\mathcal{K}_0 = \{j \in \mathcal{K} : \nexists k \in \mathcal{K} \text{ such that } (j,k) \in \mathcal{L}\}, \quad \mathcal{K}_1 = \mathcal{K}_0 \setminus (\mathcal{J}\cap\mathcal{J}').$$

The pairs in $\mathcal{L}$ are unordered. The covariance of $\mathbf{u}_\mathcal{K}$ is $\mathbf{X}_\mathcal{K}^T\mathbf{X}_\mathcal{K}$. The factorization that follows is an extension of the Cholesky decomposition.

Lemma 7 *Under the condition $\gamma S/\delta^2 = o(1)$, there is an invertible (upper triangular) matrix* $\mathbf{M}$ *obeying*

$$\|\mathbf{M}-\mathbf{I}\|_\infty \le 2\gamma/\delta, \text{ and } \|\mathbf{M}-\mathbf{I}\|_{\infty,\infty},\ \|\mathbf{M}^{-1}-\mathbf{I}\|_{\infty,\infty} \le 5\gamma S/\delta,$$

such that $\mathbf{X}_{\mathcal{K}}^T\mathbf{X}_{\mathcal{K}} = \mathbf{M}^T\mathbf{N}\mathbf{M}$*, where* $\mathbf{N}$ *is the matrix with ones on the diagonal,* $\mathbf{x}_j^T\mathbf{x}_k$ *at* $(j,k)\in\mathcal{L}$ *and zeros elsewhere.*

Define $\mathbf{v} = \mathbf{M}^{-T}\mathbf{u}_{\mathcal{K}} \sim \mathcal{N}(\mathbf{0},\mathbf{N})$, $\mathbf{a} = \mathbf{M}\boldsymbol{\beta}_{\mathcal{K}}$ (resp. $\mathbf{a}' = \mathbf{M}\boldsymbol{\beta}'_{\mathcal{K}}$), and set $\omega = 5\gamma S/\delta$. Lemma 7 gives

$$\mathbb{E}(W^2\mathbf{1}_D) \quad = \quad \pi^{\otimes 2}\mathbb{E}\left(\exp\left(\mathbf{v}^T(\mathbf{a}+\mathbf{a}') - \mathbf{a}^T\mathbf{N}\mathbf{a}/2 - (\mathbf{a}')^T\mathbf{N}\mathbf{a}'/2\right)\prod_{j\in[p]}\mathbf{1}_{\{|v_j|\le(1-\omega)^{-1}t_p\}}\right).$$

By restricting the product to $j\notin\mathcal{K}_1$, we obtain the bound

$$\mathbb{E}(W^2\mathbf{1}_D) \le \pi^{\otimes 2}\left[\mathrm{I}\cdot\mathrm{II}\cdot\mathrm{III}\right],$$

where

$$\mathrm{I} = \prod_{j\in\mathcal{K}_1}\mathbb{E}\left(\exp\left((a_j+a_j')v_j - \frac{a_j^2}{2} - \frac{(a_j')^2}{2}\right)\right),$$

$$\mathrm{II} = \prod_{j\in\mathcal{J}\cap\mathcal{J}'}\mathbb{E}\left(\exp\left((a_j+a_j')v_j - \frac{a_j^2}{2} - \frac{(a_j')^2}{2}\right)\mathbf{1}_{\{|v_j|\le(1-\omega)^{-1}t_p\}}\right),$$

$$\begin{aligned}\mathrm{III} = \prod_{(j,k)\in\mathcal{L}}\mathbb{E}\Bigg(\exp\Bigg((a_j+a_j')v_j + (a_k+a_k')v_k - \frac{a_j^2+a_k^2+2(\mathbf{x}_j^T\mathbf{x}_k)a_ja_k}{2}\\ - \frac{(a_j')^2+(a_k')^2+2(\mathbf{x}_j^T\mathbf{x}_k)a_j'a_k'}{2}\Bigg)\mathbf{1}_{\{|v_j|,|v_k|\le(1-\omega)^{-1}t_p\}}\Bigg).\end{aligned}$$

A direct calculation gives

$$\mathrm{I} = \prod_{j\in\mathcal{K}_1}\exp(a_ja_j') = \exp\left(\sum_{j\in\mathcal{K}_1}a_ja_j'\right).$$

For II, we use Lemma 2 and Lemma 7 which together imply that $|a_j-\beta_j|, |a_j'-\beta_j'| \le \omega A = o(1)$, and Lemma 1, to obtain

$$\begin{aligned}\mathrm{II} &\le \prod_{j\in\mathcal{J}\cap\mathcal{J}'}\exp(a_ja_j')\Phi((1-\omega)^{-1}t_p - |a_j+a_j'|)\\ &= (1+O(\omega\log p))^{|\mathcal{J}\cap\mathcal{J}'|}\prod_{j\in\mathcal{J}\cap\mathcal{J}'}\exp(\beta_j\beta_j')\Phi(t_p - |\beta_j+\beta_j'|)\\ &\le (1+O(\omega\log p))^{|\mathcal{J}\cap\mathcal{J}'|}\left(e^{A^2}\Phi(t_p-2A)\right)^{|\{j\in\mathcal{J}\cap\mathcal{J}':\beta_j=\beta_j'\}|}.\end{aligned}$$

Observe that either $(j,k) \in \mathcal{J} \times \mathcal{J}'$ or $(j,k) \in \mathcal{J}' \times \mathcal{J}$, but not both. This together with the Cauchy-Schwartz inequality give

$$\begin{aligned}\text{III} &\leq \prod_{(j,k)\in\mathcal{L}} \exp\left(2(a_j a_j' + a_k a_k') + \frac{1}{2}(a_j^2 + (a_j')^2 + a_k^2 + (a_k')^2 - (\mathbf{x}_j^T \mathbf{x}_k)(a_j a_k + a_j' a_k'))\right) \\ &\qquad \times \Phi((1-\omega)^{-1} t_p - 2|a_j + a_j'|)^{1/2} \Phi((1-\omega)^{-1} t_p - 2|a_k + a_k'|)^{1/2} \\ &= (1 + O(\omega \log p))^{|\mathcal{L}|} \left(e^{A^2} \Phi(t_p - 2A)\right)^{|\mathcal{L}|}.\end{aligned}$$

Lemma 6 followed by an application of Hölder's Inequality yields

$$\mathbb{E}(W^2 \mathbf{1}_D) = (1+o(1))\pi_0^{\otimes 2}\left[\text{I} \cdot \text{II} \cdot \text{III}\right] \leq (1+o(1))\pi_0^{\otimes 2}[\text{I}^f]^{1/f}\ \pi_0^{\otimes 2}\left[\text{II}^g \cdot \text{III}^g\right]^{1/g},$$

where $f, g > 0$ are arbitrary conjugate reals obeying $1/f + 1/g = 1$.

We deal with $\pi_0^{\otimes 2}\left[\text{I}^f\right]$ first. Since $\beta_j \beta_j' = 0$ for all $j \in \mathcal{K}_1$, we have

$$\sum_{j\in\mathcal{K}_1} a_j a_j' = \boldsymbol{\beta}_{\mathcal{J}_1}^T \mathbf{X}_{\mathcal{J}_1}^T \mathbf{X}_{\mathcal{J}_1'} \boldsymbol{\beta}'_{\mathcal{J}_1'},$$

where $\mathcal{J}_1 = \mathcal{J} \cap \mathcal{K}_1$ and $\mathcal{J}_1' = \mathcal{J}' \cap \mathcal{K}_1$. We apply Lemma 4 with $\boldsymbol{\xi} = \boldsymbol{\beta}/A$, $\boldsymbol{\xi}' = \boldsymbol{\beta}'/A$ and $\mathbf{R} = fA^2 \mathbf{X}_{\mathcal{J}_1}^T \mathbf{X}_{\mathcal{J}_1'}$. Noting that $\|\mathbf{R}\|_F \leq fA^2\gamma S$, since $\|\mathbf{X}_{\mathcal{J}_1}^T \mathbf{X}_{\mathcal{J}_1'}\|_\infty \leq \gamma$ and $|\mathcal{J}_1| = |\mathcal{J}_1'| \leq S$, we obtain

$$\pi_0^{\otimes 2}[\text{I}^f] \leq \exp(2fA^2\gamma S \log p) + fA^2\gamma S.$$

Since $A^2\gamma S \log p = o(1)$, for any fixed $f > 0$, $\pi_0^{\otimes 2}\left[\text{I}^f\right] \leq 1 + o(1)$.

We now turn our attention to $\pi_0^{\otimes 2}\left[\text{II}^g \cdot \text{III}^g\right]$ and show that for g close enough to 1, this quantity is bounded above by $1 + o(1)$. Note that $e^{A^2}\Phi(t_p - 2A) \leq p^d$, where $d = 2r$ if $r \geq 1/4$, and $d = 1 - 2(1-\sqrt{r})^2$ otherwise; in particular, $d \geq 0$. Therefore,

$$\pi_0^{\otimes 2}\left[\text{II}^g \cdot \text{III}^g\right] \leq \pi_0^{\otimes 2}\left[((1+o(1))p^{gd})^{|\mathcal{J}\cap\mathcal{J}'| + |L|}\right].$$

Since $\left|\cup_{j\in\mathcal{J}} \mathcal{I}_j\right| \leq S\Delta$, $|\mathcal{J} \cap \mathcal{J}'| + |L|$ is stochastically bounded by $\text{Hyp}(p, S\Delta, S)$ under π_0, which in turn is stochastically bounded by $\text{Bin}(S, S\Delta/(p - S\Delta))$. Hence, plugging the expression for the moment generating function of the binomial distribution gives

$$\pi_0^{\otimes 2}\left[\text{II}^g \cdot \text{III}^g\right] \leq \left(1 + (1+o(1))\frac{S\Delta}{p - S\Delta} p^{gd}\right)^S.$$

It then suffices to check that for g close enough to 1, $S(S\Delta/p)p^{gd} = o(1)$. This holds since Δ is smaller than any power of p by (A.2) and because $S(S/p)p^d = p^{1-2\alpha+d}$, with $1 - 2\alpha + d < 0$. The latter holds because $r < \rho^*(\alpha)$ so that $1 - 2\alpha + gd < 0$ provided g is sufficiently close to 1.

Balanced, multi-way designs

The approach is essentially the same except that we use the prior π_0 restricted to the set of S-sparse vectors with coordinates $\pm A$ satisfying the linear constraints (4.1). This may put a constraint on S (e.g. that S be even in the one-way design), which can be assumed without loss of generality; the reason is that it is only the order of S that matters, $\log S \sim (1-\alpha)\log p$. The key point here is that the signs of the coordinates are now irrelevant as the orthogonality eliminates the term η.

A.4.2 Random effects model

The proof parallels the previous one. Introduce π_0 and π with the same respective distributions on the support as before but with nonzero sampled entries now i.i.d. $\mathcal{N}(0,\tau^2)$. Note that $\mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}$ is normal with mean zero and variance $\tau^2\sum_{k\in\mathcal{J}}(\mathbf{x}_j^T\mathbf{x}_k)^2$, so that

$$\tau\mathbf{1}_{\{j\in\mathcal{J}\}}|z| \overset{\text{sto}}{\leq} |\mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}| \overset{\text{sto}}{\leq} \tau(\mathbf{1}_{\{j\leftrightarrow\mathcal{J}\}}+\gamma S)|z|, \quad z\sim\mathcal{N}(0,1). \tag{A.6}$$

holds instead of (A.4).

Lemma 8 *For $t>0$ and $z,z'\overset{\text{iid}}{\sim}\mathcal{N}(0,1)$,*

$$\mathbb{E}\bar{\Phi}(t-\tau|z|)\leq 2\exp\left(-\frac{t^2}{2(\tau^2+1)}\right),$$

and

$$\mathbb{E}\left(e^{zz'}\bar{\Phi}(t-\tau|z+z'|)\right)\leq 2\exp\left(-\frac{t^2(\tau^2-1)}{2(\tau^2+1)}\right).$$

Proof. Just as for Lemma 2, the proof is by direct calculation. □

Armed with this auxiliary result, we write

$$\begin{aligned}\mathbb{E}(W\mathbf{1}_{D^c}) &\leq (1+o(1))\left(p\bar{\Phi}(t_p)+\Delta S\mathbb{E}\bar{\Phi}(t_p-\tau|z|)\right)\\ &\leq C\left(\frac{1}{\sqrt{\log p}}+\Delta p^{1-\alpha-1/(1+\tau^2)}\right).\end{aligned}$$

where the first inequality uses (A.6), and the second uses the first part of the lemma. Because $(\tau^2+1)(1-\alpha)<1$, we conclude that $\mathbb{E}(W\mathbf{1}_{D^c})=o(1)$.

The arguments for bounding $\mathbb{E}(W^2\mathbf{1}_D)$ are essentially the same as in the fixed effects model except in two places. To begin with, we use Lemma 5 instead of Lemma 4. For the first term I, we apply Lemma 5 with $\mathbf{R}=\mathbf{X}_{\mathcal{J}}^T\mathbf{X}_{\mathcal{J}'}$ and use $\|\mathbf{R}\|_F^2\leq\gamma^2S^2=o(1)$ to get $\pi_0^{\otimes 2}\left[\mathrm{I}^f\right]^{1/f}\leq 1+o(1)$ for any fixed $f>0$. For II and III, we use the second part of Lemma 8 in place of Lemma 2 obtaining the following bound, valid for any fixed $g>0$,

$$\pi_0^{\otimes 2}\left[\mathrm{II}^g\cdot\mathrm{III}^g\right]\leq\left(1+(1+o(1))\frac{S\Delta}{p-S\Delta}p^{g(\tau^2-1)/(\tau^2+1)}\right)^S.$$

Assumption (A.2) gives for all $\varepsilon>0$,

$$S\frac{S\Delta}{p-S\Delta}p^{g(\tau^2-1)/(\tau^2+1)}=O(p^{1-2\alpha+g(\tau^2-1)/(\tau^2+1)+\varepsilon}).$$

Choose $\varepsilon=1-\alpha-1/(\tau^2+1)$. When $g=1$, the exponent is then equal to $3(1-\alpha-1/(\tau^2+1))<0$ and, therefore, the exponent is negative when g is sufficiently close to 1. From this, we conclude that for g sufficiently close to 1, $\pi_0^{\otimes 2}\left[\mathrm{II}^g\cdot\mathrm{III}^g\right]\leq 1+o(1)$.

A.5 Proof of Theorem 3

We use the notation introduced in Section A.4.1. By Lemma 6, we may assume that the support $\mathcal{J}$ of $\boldsymbol{\beta}$ obeys (A.3). We simply examine the expected value and variance of $H(t)$ under the null and the alternative for t in the range defined for $H^*(s_p(\alpha))$, i.e. $t = \sqrt{2r\log p}$ with $\min(1, 4\rho^*(\alpha)) \le r \le 5/2$. We will use the following lemmas. The first result bounds the tail of a bivariate normal distribution and its first part is due to Li and Shao [29].

Lemma 9 *For z and z' standard normal with $\zeta = \operatorname{Cov}(z, z')$, and for $a, a' \in \mathbb{R}$ and $t > 0$, we have*

$$\operatorname{Cov}(\mathbf{1}_{\{|z+a|>t\}}, \mathbf{1}_{\{|z'+a'|>t\}}) \le |\zeta| \exp\left(-\frac{(t-|a|)^2 + (t-|a'|)^2}{2(1+|\zeta|)}\right).$$

Moreover, there is a numerical constant $C > 0$ such that if $t \ge 1$ and $|\zeta|, |a|, |a'| \le 1/t$,

$$\operatorname{Cov}(\mathbf{1}_{\{|z+a|>t\}}, \mathbf{1}_{\{|z'+a'|>t\}}) \le C|\zeta|(a^2 + (a')^2 + |\zeta|)t^2 e^{-t^2}.$$

The second result is a bound on the inner product between a fixed vector and a random vector with coordinates ± 1. It is a direct consequence of Hoeffding's inequality in the form of [20, Eq. (2.6)].

Lemma 10 *Fix $\mathbf{r} \in \mathbb{R}^q$ and let $\boldsymbol{\xi}$ be a random vector with symmetric i.i.d. ± 1 entries. Then*

$$\mathbb{P}\left(|\mathbf{r}^T\boldsymbol{\xi}| > t\|\mathbf{r}\|\right) \le 2e^{-t^2/2}.$$

Recall that the variables $(\mathbf{x}_1^T\mathbf{y}, \ldots, \mathbf{x}_p^T\mathbf{y})$ are jointly normal with $a_j := \mathbb{E}(\mathbf{x}_j^T\mathbf{y}) = \mathbf{x}_j^T\mathbf{X}\boldsymbol{\beta}$ and $c_{jk} := \operatorname{Cov}(\mathbf{x}_j^T\mathbf{y}, \mathbf{x}_k^T\mathbf{y}) = \mathbf{x}_j^T\mathbf{x}_k$. As usual, the subscript 0 (resp. 1) denotes the expectation under the null hypothesis (resp. alternative).

A.5.1 Under the null

Clearly, $\mathbb{E}_0 H(t) = 0$. Next, assuming that p is large enough so that $\bar{\Phi}(t) \le 1/4$,

$$\begin{aligned}
\operatorname{Var}_0 H(t) &\le 1 + \frac{1}{p\bar{\Phi}(t)} \sum_{j \ne k} \operatorname{Cov}(\mathbf{1}_{\{|\mathbf{x}_j^T\mathbf{z}|>t\}}, \mathbf{1}_{\{|\mathbf{x}_k^T\mathbf{z}|>t\}}) \\
&\le 1 + \frac{1}{p\bar{\Phi}(t)} \left(p(\Delta - 1)e^{-t^2/2} + Cp^2\gamma^2 t^2 e^{-t^2}\right) \\
&\le \nu := C(\log p)^2 \left(\Delta + \gamma^2 p^{2(3-4\alpha)_+}\right).
\end{aligned}$$

The first line uses the definition of $H(t)$. The second line comes from decomposing the sum into $j \leftrightarrow k$, for which we use the first part of Lemma 9, and $j \not\leftrightarrow k$, for which we use the second part of Lemma 9. The third line comes from Lemma 1 and the range of t under consideration.

Since $\mathbb{E}_0 H(t) = 0$ for all t, Chebyshev's inequality and the union bound give

$$\mathbb{P}_0(H^*(s_p(\alpha)) > (\log p)\nu^{1/2}) \le \frac{1}{\log p} \to 0.$$

We then use $(\log p)\nu^{1/2}$ as the critical value for the test based on $H^*(s_p(\alpha))$.

A.5.2 Under the alternative

Recall that $|\mathcal{I}_j \cap \mathcal{J}| \leq 1$ for all j, with probability at least $1 - 2p^{-1}$. Then applying Lemma 10 and using the union bound gives

$$|a_j - \beta_j| \leq (4 \log p)^{1/2} \gamma S^{1/2} \|\boldsymbol{\beta}\|_\infty, \quad \forall j \in \mathcal{J}, \tag{A.7}$$

and

$$|a_j| \leq (4 \log p)^{1/2} \gamma S^{1/2} \|\boldsymbol{\beta}\|_\infty, \quad \forall j \not\leftrightarrow \mathcal{J}. \tag{A.8}$$

Therefore, we may assume that (A.7) and (A.8) hold.

Let j be such that $|\beta_j| = \max_k |\beta_k|$ and assume that $|\beta_j| \geq \sqrt{6 \log p}$. Then as in (A.4), but now using (A.7), we have

$$|\mathbf{x}_j^T \mathbf{y}| \geq |\mathbf{x}_j^T \mathbf{X} \boldsymbol{\beta}| - |\mathbf{x}_j^T \mathbf{z}| \geq (1 - (4 \log p)^{1/2} \gamma S^{1/2}) |\beta_j| - |\mathbf{x}_j^T \mathbf{z}|.$$

Since $(\log p)^{1/2} \gamma S^{1/2} = o(1)$ and $|\mathbf{x}_j^T \mathbf{z}| = O_P(1)$, we see that $|\mathbf{x}_j^T \mathbf{y}| > \sqrt{5 \log p}$ with high probability, in which case

$$H([\sqrt{5 \log p}]) \geq \frac{1 - 2p\bar{\Phi}([\sqrt{5 \log p}])}{\sqrt{2p\bar{\Phi}([\sqrt{5 \log p}])}} \geq p^{3/4},$$

where we have used Lemma 1. Further, since $\gamma S^{1/2} = o(1)$ and (A.2) holds, we have $(\log p)\nu^{1/2} = o(p^{3/4})$. Therefore, because $H^*(s_p(\alpha)) \geq H([\sqrt{5 \log p}])$, the test based on $H^*(s_p(\alpha))$ rejects the null hypothesis with high probability.

We now focus on an alternative obeying $A \leq |\beta_j| \leq \sqrt{6 \log p}$ for all $j \in \mathcal{J}$, where $|\mathcal{J}| = S = p^{1-\alpha}$ and $A = \sqrt{2r \log p}$, $r > \rho^*(\alpha)$. Then the following refinement of (A.4) holds

$$|\beta_j| \mathbf{1}_{\{j \in \mathcal{J}\}} - \omega \leq |a_j| \leq |\beta_k| \mathbf{1}_{\{j \leftrightarrow k \in \mathcal{J}\}} + \omega, \tag{A.9}$$

with $\omega := 5(\log p)\gamma S^{1/2}$. Let $t = [\sqrt{2r_1 \log p}]$ with $r_1 = \min(1, 4r)$. For the expectation,

$$\mathbb{E}_1 H(t) \geq \frac{1}{\sqrt{2p\bar{\Phi}(t)}} \sum_{j \in [p]} (\bar{\Phi}(t + a_j) + \bar{\Phi}(t - a_j) - 2\bar{\Phi}(t)).$$

For any $t > 0$, the function $a \to \bar{\Phi}(t + a) + \bar{\Phi}(t - a) - 2\bar{\Phi}(t)$ is non-negative so that we may restrict the sum to $j \in \mathcal{J}$ to obtain a lower bound. Moreover, the same function exceeds $\bar{\Phi}(t + |a|) - 2\bar{\Phi}(t)$. Combined, these observations yield

$$\mathbb{E}_1 H(t) \geq \frac{1}{\sqrt{2p\bar{\Phi}(t)}} \sum_{j \in \mathcal{J}} \bar{\Phi}(t - |a_j|) - S\sqrt{2\bar{\Phi}(t)/p}.$$

Since $S = o(p^{1/2})$, Lemma 1 with $t\omega = o(1)$ gives

$$\mathbb{E}_1 H(t) \geq (1 + o(1)) \frac{\Lambda}{\sqrt{2p\bar{\Phi}(t)}} + o(1), \quad \Lambda := \sum_{j \in \mathcal{J}} \bar{\Phi}(t - |\beta_j|).$$

We will use the following lower bound:

$$\Lambda \geq C \left(\#\{j \in \mathcal{J} : |\beta_j| > t\} + \frac{\lambda}{\sqrt{\log p}} \right), \quad \lambda := \sum_{j \in \mathcal{J}} e^{-(t - |\beta_j|)^2/2}, \tag{A.10}$$

obtained using $\bar\Phi(t) \geq 1/2$ for $t > 0$ and Lemma 1. Note that

$$\Lambda \geq S\bar\Phi(t - A) \geq C(\log p)^{-1/2} p^{1-\alpha-(\sqrt{r_1}-\sqrt{r})^2}, \quad \text{with } 1 - \alpha - (\sqrt{r_1} - \sqrt{r})^2 > 0, \tag{A.11}$$

since $r > \rho^*(\alpha)$ and the definition of r_1, so that $\Lambda \to \infty$ as a power of p.

We immediately observe that $(\log p)\nu^{1/2}$ is negligible compared to $\mathbb{E}_1 H(t)$ by comparing (A.10) and (A.11), and by the condition $\gamma = o(p^{2\alpha-3/2})$. Therefore, by Chebyshev's inequality, it suffices to show that $\mathbb{E}_1 H(t)/\sqrt{\mathrm{Var}_1 H(t)} \to \infty$. We bound the variance under the alternative as we did under the null. Assuming p is large enough so that $\bar\Phi(t) \leq 1/4$, we have

$$\mathrm{Var}_1 H(t) \leq 1 + \frac{1}{p\bar\Phi(t)} \sum_{j\neq\ell} \mathrm{Cov}(\mathbf{1}_{\{|\mathbf{x}_j^T\mathbf{z}+a_j|>t\}}, \mathbf{1}_{\{|\mathbf{x}_\ell^T\mathbf{z}+a_\ell|>t\}}).$$

We show that the sum above is negligible compared with Λ^2. We divide that (double) sum into two parts.

The first part is over $j \leftrightarrow \mathcal{J}$. Let $k \in \mathcal{J}$ be such that $j \leftrightarrow k$. We partition the sum over ℓ into four groups: $\ell = k$, $\ell \leftrightarrow j$ with $\ell \neq k$, $\ell \not\leftrightarrow j$ with $\ell \in \mathcal{J}$ and $\ell \not\leftrightarrow j$ with $\ell \notin \mathcal{J}$. Using the first part of Lemma 9, we have

$$\mathrm{Cov}(\mathbf{1}_{\{|\mathbf{x}_j^T\mathbf{z}+a_j|>t\}}, \mathbf{1}_{\{|\mathbf{x}_\ell^T\mathbf{z}+a_\ell|>t\}}) \leq |c_{j\ell}| \exp\left(-\frac{(t-|a_j|)^2 + (t-|a_\ell|)^2}{2(1+|c_{j\ell}|)}\right). \tag{A.12}$$

Note that $t\omega = o(1)$. By (A.9), $|a_j| \leq |\beta_k| + \omega$. When $\ell = k$, we use $|c_{jk}| \leq 1$ and apply (A.9) to get $|a_k| \leq |\beta_k| + \omega$, to bound (A.12) by $2e^{-(t-|\beta_k|)^2/2}$. (The factor 2 bounds a term of order $(1 + O(t\omega))$.) When $\ell \leftrightarrow j$ with $\ell \neq k$, then $\ell \not\leftrightarrow \mathcal{J}$ by (A.3) so that $|a_\ell| \leq \omega$ by (A.9). Together with $|c_{j\ell}| \leq 1$, these bound (A.12) by $2e^{-((t-|\beta_k|)^2+t^2)/4}$. When $\ell \not\leftrightarrow j$ with $\ell \in \mathcal{J}$, we use $|c_{j\ell}| \leq \gamma$ and $|a_\ell| \leq |\beta_\ell| + \omega$ by (A.9) to bound (A.12) by $2\gamma e^{-((t-|\beta_k|)^2+(t-|\beta_\ell|)^2)/2}$. When $\ell \not\leftrightarrow j$ with $\ell \notin \mathcal{J}$, we use $|c_{j\ell}| \leq \gamma$ and $|a_\ell| \leq \omega$ by (A.9) to bound (A.12) by $2e^{-((t-|\beta_k|)^2+t^2)/2}$. Summing over ℓ, it follows from $|\mathcal{I}_j| \leq \Delta$ for all j, and from the assumption about Δ that the bound

$$e^{-(t-|\beta_k|)^2/2} + \Delta e^{-((t-|\beta_k|)^2+t^2)/4} + \gamma \sum_{\ell\in\mathcal{J}} e^{-((t-|\beta_k|)^2+(t-|\beta_\ell|)^2)/2} + \gamma p e^{-((t-|\beta_k|)^2+t^2)/2},$$

holds up to a multiplicative numerical factor. Then using $|\mathcal{I}_k| \leq \Delta$ and the definition of λ, the sum over $j \leftrightarrow \mathcal{J}$ is bounded by

$$\Delta\lambda + \Delta^2 e^{-t^2/4} \sum_{k\in\mathcal{J}} e^{-(t-|\beta_k|)^2/4} + \Delta\gamma\lambda^2 + \Delta\gamma p e^{-t^2/2}\lambda,$$

up to a multiplicative numerical factor. We show that each term is negligible compared to Λ^2. For the first term, we use (A.10) with (A.11), and (A.2). For the second, we use (A.10), the bound

$$e^{-t^2/4} \sum_{k\in\mathcal{J}} e^{-(t-|\beta_k|)^2/4} \leq e^{-t^2/4} \#\{j \in \mathcal{J} : |\beta_j| > t\} + \lambda,$$

and the fact that Λ goes to infinity as a power of p. For the third, we use (A.10) with (A.11), (A.2) and the fact that γ goes to zero as a power of p. For the fourth, we use $\gamma = o(p^{2\alpha-3/2})$ in addition to (A.10) with (A.11), and (A.2).

The second part is over $j \not\leftrightarrow \mathcal{J}$. We partition the sum over ℓ into three groups: $\ell \leftrightarrow j$, $\ell \not\leftrightarrow j$ with $\ell \in \mathcal{J}$ and $\ell \not\leftrightarrow j$ with $\ell \notin \mathcal{J}$. When $\ell \leftrightarrow j$, we use $|c_{j\ell}| \leq 1$ and $|a_j|, |a_\ell| \leq \omega$ by (A.9) to bound (A.12) by $2e^{-t^2/2}$. (Again, the factor 2 bounds a term of order $(1 + O(t\omega))$.) When $\ell \not\leftrightarrow j$ with $\ell \leftrightarrow \mathcal{J}$, we use $|c_{j\ell}| \le \gamma$ and $|a_j| \le \omega$ and $|a_\ell| \le |\beta_\ell| + \omega$ by (A.9) to bound (A.12) by $2\gamma e^{-(t^2+(t-|\beta_\ell|)^2)/4}$. When $\ell \not\leftrightarrow j$ with $\ell \not\leftrightarrow \mathcal{J}$, we use the second part of Lemma 9 together with $|c_{j\ell}| \le \gamma$ and $|a_j|, |a_\ell| \le \omega$ by (A.9) to obtain

$$\mathrm{Cov}(\mathbf{1}_{\{|\mathbf{x}_j^T\mathbf{z}+a_j|>t\}}, \mathbf{1}_{\{|\mathbf{x}_\ell^T\mathbf{z}+a_\ell|>t\}}) \le C\gamma(\omega^2 + \gamma)t^2 e^{-t^2}.$$

Summing over ℓ, we obtain the bound

$$\Delta e^{-t^2/2} + \gamma \sum_{\ell \in \mathcal{J}} e^{-(t^2+(t-|\beta_\ell|)^2)/2} + p\gamma(\omega^2 + \gamma)t^2 e^{-t^2},$$

which holds up to a multiplicative numerical factor. The sum over such j's is, therefore, bounded by

$$\Delta p e^{-t^2/2} + \gamma p e^{-t^2/2}\lambda + p^2\gamma(\omega^2 + \gamma)t^2 e^{-t^2},$$

up to a multiplicative numerical factor. Using similar arguments, the first two terms are seen to be negligible compared to Λ^2. The last term is also negligible compared to Λ^2 by the definition of ω and the assumption $\gamma^3 = O(p^{\varepsilon+5\alpha-4})$ for all $\varepsilon > 0$.

A.6 Proof of Theorem 4

We use the notation introduced in Section A.5 and also define $M = \|\mathbf{X}^T\mathbf{y}\|_\infty = \max_j |\mathbf{x}_j^T\mathbf{y}|$ so that the Max test rejects for large values of M. The lower bound below is an immediate consequence of [2, Th. 3.1].

Lemma 11 *Let $z_{1q}, \dots, z_{qq}$ be an array of standard normal random variables with covariances obeying $\max_{j\ne k} |\mathrm{Cov}(z_{jq}, z_{kq})| = o(\log q)^{-1}$, and set*

$$\kappa_q = \sqrt{2\log q} - \frac{\log\log q + \log(4\pi)}{2\sqrt{2\log q}}.$$

Then for any fixed $s \in \mathbb{R}$,

$$\mathbb{P}\left(\max\{z_{1q}, \dots, z_{qq}\} \le \kappa_q + \frac{s}{\sqrt{2\log q}}\right) \to \exp(-e^{-s}), \quad q \to \infty.$$

A.6.1 Sparse fixed effects model

We first show that the Max test with rejection region $\{M > \sqrt{2\log p}\}$ is asymptotically powerful when $r > \rho_{\mathrm{Max}}(\alpha)$. A simple application of the union bound and Lemma 1 give $\mathbb{P}\left(M > \sqrt{2\log p}\right) \to 0$. Under the alternative, let $\mathcal{J}^+ = \{j : a_j > 0\}$ and assume without loss of generality that $|\mathcal{J}^+| \ge S/2$. Reasoning as in Section A.5, we establish (A.9) and use the lower bound to obtain $a_j \ge A - \omega$ for all $j \in \mathcal{J}^+$. Recall that $\omega = o(1)$. Then

$$M \ge \max_{j\in\mathcal{J}^+} \mathbf{x}_j^T\mathbf{y} \ge A + \max_{j\in\mathcal{J}^+} \mathbf{x}_j^T\mathbf{z} + o_P(1).$$

We now apply Lemma 11 to $\{\mathbf{x}_j^T\mathbf{z} : j \in \mathcal{J}^+\}$ whose pairwise correlations are bounded in absolute value by $\gamma = o(\log|\mathcal{J}^+|)$. This gives $\max_{j\in\mathcal{J}^+} \mathbf{x}_j^T\mathbf{z} \ge \sqrt{2\log|\mathcal{J}^+|} + o_P(1)$. Since $|\mathcal{J}^+| \ge \frac{1}{2}p^{1-\alpha}$, we have

$$M \ge \max_{j\in\mathcal{J}^+} \mathbf{x}_j^T\mathbf{y} \ge (\sqrt{r} + \sqrt{1-\alpha})\sqrt{2\log p} + o_P(1).$$

Finally, the condition $r > \rho_{\text{Max}}(\alpha)$ implies $\sqrt{r} + \sqrt{1-\alpha} > 1$ so that the probability of Type II error goes to zero as $p \to \infty$.

We now prove that the Max test is *not* asymptotically powerful when $r < \rho_{\text{Max}}(\alpha)$. This is a little more challenging and requires using the lower bound in Lemma 1. This is done in two steps. Let $\{t_p\}$ be any sequence of thresholds obeying $\mathbb{P}(M > t_p) \to 0$ under H_0. Lemma 11 states that this holds if and only if $t_p = \kappa_p + s_p/\sqrt{\log p}$ where $s_p \to -\infty$.

- We first show that under the alternative, we still have

$$\mathbb{P}\left(\max_{j \not\leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{y}| > t_p\right) \to 0.$$

Indeed, the upper bound in (A.9) gives that with high probability, it holds that $|a_j| \le \omega$ uniformly over all $j \not\leftrightarrow \mathcal{J}$ and, therefore,

$$\mathbb{P}\left(\max_{j \not\leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{y}| > t_p\right) \le \mathbb{P}\left(\max_{j \not\leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{z}| > t_p - \omega\right) + o(1).$$

Since $\omega = O(\log p)^{-1/2}$, $t_p - \omega$ is still of the form $\kappa_p + s'_p/\sqrt{\log p}$ with $s'_p \to -\infty$, and the same Lemma 11 implies that the probability term in the right-hand side tends to zero as well.

- Second, if $r < \rho_{\text{Max}}(\alpha)$, then

$$\mathbb{P}\left(\max_{j \leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{y}| > t_p\right) \to 0.$$

The reasoning is the same. The upper bound in (A.9) gives that $|a_j| \le A + \omega$ holds – with high probability – for all $j \not\leftrightarrow \mathcal{J}$ and, therefore,

$$\mathbb{P}\left(\max_{j \leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{y}| > t_p\right) \le \mathbb{P}\left(\max_{j \leftrightarrow \mathcal{J}} |\mathbf{x}_j^T \mathbf{z}| > t_p - A - \omega\right) + o(1).$$

Define $q = \Delta S$, which is an upper bound on $|\{j : j \leftrightarrow \mathcal{J}\}|$, and consider $\varepsilon > 0$ small enough so that $1 - \sqrt{r} > \sqrt{1 - \alpha + \varepsilon}$. This exists since $r < \rho_{\text{Max}}(\alpha)$. On the one hand, $t_p - A - \omega = (1 - \sqrt{r})\sqrt{2\log p} + o(1)$, while on the other hand, $\kappa_p \le \sqrt{1-\alpha+\varepsilon} \cdot \sqrt{2\log p}$ by (A.2). Hence, $t_p - A - \omega$ is of the form $\kappa_p + s'_p/\sqrt{\log p}$ with $s'_p \to -\infty$. The same Lemma 11 implies that the probability term in the right-hand side tends to zero as well.

A.6.2 Random effects model

We use the same threshold $\sqrt{2\log p}$. Under the alternative, the distribution of $\mathbf{x}_j^T \mathbf{y}$ for $j \in \mathcal{J}$ is given by

$$\mathbf{x}_j^T \mathbf{y} = \mathbf{x}_j^T \mathbf{X}\boldsymbol{\beta} + \mathbf{x}_j^T \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \tau^2 \|\mathbf{x}_j^T \mathbf{X}\|^2 + 1),$$

where $\|\mathbf{x}_j^T \mathbf{X}\|^2 \ge 1$. Further, for $j, k \in \mathcal{J}$ distinct,

$$\left|\text{Cov}(\mathbf{x}_j^T \mathbf{y}, \mathbf{x}_k^T \mathbf{y})\right| = \left|\pi_0 \left[\mathbf{x}_j^T (\tau^2 \mathbf{X}_{\mathcal{J}} \mathbf{X}_{\mathcal{J}}^T + \mathbf{I})\mathbf{x}_k\right]\right| \le \tau^2((S-2)\gamma^2 + 2\gamma) + \gamma = o(\log S)^{-1},$$

where we used Lemma 6 to obtain the upper bound. Hence, it follows from Lemma 11 that

$$M \ge \max_{j \in \mathcal{J}} |\mathbf{x}_j^T \mathbf{y}| \ge (1 + o(1))\sqrt{2(\tau^2+1)\log S} = (1+o(1))\sqrt{2(\tau^2+1)(1-\alpha)\log p}$$

with high probability. This establishes Theorem 4.

A.7 Proof of Corollary 1

Since the norm of a standard normal vector in $\mathbb{R}^n$ is sharply concentrated around $\sqrt{n}$ (this can be seen from well-known deviations bounds for chi-square variables), we may just as well work with $\mathbf{X}$ generated by sampling n independent row vectors from $\mathcal{N}(\mathbf{0}, n^{-1}\mathbf{C})$. The following inequality bounds the difference between the empirical covariance matrix and its expected value [3, Lem. A.3]: there is a constant $B > 0$ such that

$$\mathbb{P}\left(\max_{j,k\in[p]} \left|\mathbf{x}_j^T\mathbf{x}_k - c_{jk}\right| > B\sqrt{\frac{\log p}{n}}\right) \to 0.$$

Thus, if $\mathbf{C} \in \mathcal{S}_p(\gamma, \Delta)$, $\mathbf{X}^T\mathbf{X} \in \mathcal{S}_p(\gamma', \Delta')$ with $\delta' \geq \delta - B\sqrt{\frac{\log p}{n}}$ and $\gamma' \leq \gamma + B\sqrt{\frac{\log p}{n}}$. Under the conditions of Corollary 1, δ' and γ' obey the conditions of Theorem 2. The corollary follows.

A.8 Proof of Theorem 5

The critical ingredient is this:

Lemma 12 *[1] Let $z_1, \dots, z_n$ be a Gaussian vector with $\mathrm{var}(z_j) = 1$ and $\mathrm{Cov}(z_j, z_k) = \gamma$, $j \neq k$. Then there exist $v_1, \dots, v_n$ and v i.i.d. $\mathcal{N}(0, 1)$ variables such that $z_j = (1-\gamma)^{1/2}v_j + \gamma^{1/2}v$.*

The proof of the lower bound follows that of Theorem 2. We focus on the sparse fixed effects model; the case of the sparse random effects model may be deduced from this as done in the proof of Theorem 2. For simplicity, redefine $\mathbf{X}$ to be $(1-\gamma)^{-1/2}\mathbf{X}$, and A to be $(1-\gamma)^{1/2}A$. Assume for convenience that S is even and let π be the uniform prior on the set of S-sparse vectors with all nonzero coefficients equal to $\pm A$ and $S/2$ positive entries. Let $\mathbf{u} = \mathbf{X}^T\mathbf{z}$ as before. Applying Lemma 12 to $\mathbf{u}$, one finds $v_1, \dots, v_p$ and v i.i.d. $\mathcal{N}(0, 1)$ such that $u_j = v_j + \gamma^{1/2}(1-\gamma)^{-1/2}v$. Put $\mathbf{v} = (v_1, \dots, v_p)$. Since $\mathbf{1}^T\boldsymbol{\beta} = 0$, $\mathbf{u}^T\boldsymbol{\beta} = \mathbf{v}^T\boldsymbol{\beta}$ and $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$, the likelihood ratio is given by

$$W = \pi\left[\exp(\mathbf{v}^T\boldsymbol{\beta} - \|\boldsymbol{\beta}\|^2/2)\right].$$

This reduces matters to the case of the identity design, except for the fact that π is restricted to vectors satisfying $\mathbf{1}^T\boldsymbol{\beta} = 0$. This subtlety is unimportant since the sign of the coefficients does not play any role when the design is orthogonal.

For the upper bound, we have

$$\mathbf{X}^T\mathbf{y} = \boldsymbol{\beta} + \gamma^{1/2}(1-\gamma)^{-1/2}v\mathbf{1} + \mathbf{v},$$

and it is straightforward to see that the middle term being of order $O_P(1)$ does not change the performance of the Higher Criticism test and/or that of the Max test.

B Proofs of the auxiliary results

B.1 Proof of Lemma 7

Put $c_{jk} = \mathbf{x}_j^T\mathbf{x}_k$ for short. The matrix $\mathbf{M} = (m_{jk})$ solves the following system of equations:

$$c_{jk} = \sum_{a\in\mathcal{K}} m_{aj}m_{ak} + \sum_{(a,b)\in\mathcal{L}} (m_{aj}m_{bk} + m_{bj}m_{ak})c_{ab}, \quad \forall j, k \in \mathcal{K}. \tag{B.1}$$

Without loss of generality, assume that $\mathcal{K} = \{1, \dots, |\mathcal{K}|\}$ and $\mathcal{K}_0 = \{1, \dots, |\mathcal{K}_0|\}$, by permuting indices if necessary, and $\mathcal{L} = \{(2s, 2s+1) : s > |\mathcal{K}_0|/2\}$, assuming $|\mathcal{K}_0|$ is odd. We use an induction. Let $\omega \to 0$ such that $\gamma S/(\delta^2 \omega) \to 0$. For $k = 1$, (B.1) reads $m_{11}^2 = c_{11} = 1$, so that we may choose $m_{11} = 1$. Now assume that for some $j \in \mathcal{K}_0$ with $j \le k$, $|m_{aj}|, |m_{ak}| \le 2\gamma$ for all $a < j$; and $|m_{aa} - 1| \le \omega\gamma$ for all $a \le j$. The term m_{jk} appears in only one term in the right-hand side of (B.1), and all the other terms are bounded by $(2\gamma)^2$ by the induction hypothesis. (Recall that $|c_{ab}| \le 1$ for all (a, b).) Therefore, since $|\mathcal{K}| \le 2S$, we have

$$c_{jk} = m_{jj}\, m_{jk} + R,$$

where $|R| \le 8\gamma^2 S \le \omega\gamma$. If $j < k$, we have $|m_{jj} - 1| \le \omega\gamma$ by the induction hypothesis and $|c_{jk}| \le \gamma$ because $(j, k) \notin \mathcal{L}$. Thus,

$$|m_{jk}| \le (1 - \omega\gamma)^{-1}(\gamma + \omega\gamma) \le 2\gamma.$$

Therefore, the first induction proceeds and we conclude that $|m_{jk}| \le 2\gamma$ for each $j \in \mathcal{K}_0$ with $j < k$. Further, taking $j = k$, we see that $m_{kk}^2 = c_{kk} - R = 1 - R$, again with$|R| \le \omega\gamma$, so that (assuming $m_{kk} > 0$) $|m_{kk} - 1| \le \omega\gamma$. Therefore, the second induction proceeds and we conclude that

$$\forall k \in \mathcal{K}, j \in \mathcal{K}_0 : |m_{jk}| \le 2\gamma \text{ and } |m_{jj} - 1| \le \omega\gamma. \tag{B.2}$$

We now consider the case where $j \notin \mathcal{K}_0$. Assume that for some j obeying $|\mathcal{K}_0|/2 \le j \le k$, $|m_{aj}|, |m_{ak}| \le 2\gamma/\delta$ for $a < j$ with $(a, j) \notin \mathcal{L}$; $|m_{ab}| \le \omega\gamma/\delta$ if $(a, b) \in \mathcal{L}$ and $\max(a, b) < j$; and $|m_{aa} - 1|, |m_{aa} - 1| \le \omega\gamma/\delta$ for $a \le j$. If $j = 2s$ and $k > j + 1$, (B.2) and the induction hypothesis give

$$\begin{aligned} c_{jk} &= m_{jj}\, m_{jk} + m_{jj}\, m_{j+1,k}\, c_{j,j+1} + R_1, \\ c_{j+1,k} &= m_{j,j+1}\, m_{jk} + m_{j+1,j+1}\, m_{j+1,k} + (m_{j,j+1}\, m_{j+1,k} + m_{j+1,j+1}\, m_{j,k})c_{j,j+1} + R_2, \end{aligned}$$

where

$$|R_1|, |R_2| \le |\mathcal{K}_0| 4\gamma^2 + |\mathcal{L}| 4\gamma^2/\delta^2 \le 8\gamma^2 S/\delta^2 = o(\omega\gamma).$$

We solve for m_{jk} and $m_{j+1,k}$, and obtain

$$\begin{aligned} m_{jk} &= c_{jk}/m_{jj} - c_{j,j+1}\, m_{j+1,k} - R_1/m_{jj}, \\ ((1 - c_{j,j+1}^2)m_{j+1,j+1} + c_{j,j+1}\, m_{j,j+1})\, m_{j+1,k} &= c_{j+1,k} - R_2 - (c_{jk} - R_1)c_{j,j+1}/m_{jj}. \end{aligned}$$

Since $|c_{jk}|, |c_{j+1,k}| \le \gamma$ and $|c_{j,j+1}| \le 1 - \delta$ and since $|m_{jj} - 1|, |m_{j+1,j+1} - 1|, |m_{j,j+1}|$ are all bounded by $\omega\gamma/\delta$ (induction hypothesis) with $\omega\gamma/\delta = o(\delta)$, we find that $|m_{jk}|, |m_{j+1,k}| \le 2\gamma/\delta$. Therefore, the first induction proceeds and we conclude that $|m_{jk}| \le 2\gamma/\delta$ for any $j < k$. If $j + 1 = k$, we have

$$\begin{aligned} c_{jk} &= m_{jj}\, m_{jk} + m_{jj}\, m_{kk}\, c_{jk} + R_1, \\ 1 &= m_{jk}^2 + m_{kk}^2 + (m_{jk}\, m_{kk} + m_{kk}\, m_{jk})c_{jk} + R_2. \end{aligned}$$

We solve for m_{jk} and m_{kk}, and obtain

$$\begin{aligned} m_{jk} &= c_{jk}(1/m_{jj} - m_{kk}) - R_1/m_{jj}, \\ (1 - c_{jk}^2)m_{kk}^2 &= 1 - ((c_{jk} - R_1)/m_{jj})^2 - R_2. \end{aligned}$$

Since $|c_{jk}| \le 1 - \delta$ and $|m_{jj} - 1| \le \omega\gamma/\delta$ (induction hypothesis), we get that $|m_{kk} - 1|, |m_{jk}| \le \omega\gamma/\delta$. If $j = k$, we have

$$1 = c_{kk} = m_{kk}^2 + R_1,$$

so that $|m_{kk} - 1| \leq \omega\gamma$. Therefore, the second induction proceeds and we conclude that

$$\forall j, k \in \mathcal{K} : |m_{jk}| \leq 2\gamma/\delta \text{ and } |m_{jj} - 1| \leq \omega\gamma/\delta. \tag{B.3}$$

To conclude the proof of Lemma 7, we compute

$$\|\mathbf{M} - \mathbf{I}\|_{\infty,\infty} = \max_{j\in\mathcal{K}} \ |m_{jj} - 1| + \sum_{k\in\mathcal{K}, k\neq j} |m_{jk}| \leq 4\gamma S/\delta,$$

and deduce from this that $\|\mathbf{M}^{-1} - \mathbf{I}\|_{\infty,\infty} \leq 5\gamma S/\delta$; this last step uses the lemma below.

Lemma 13 *Any positive semidefinite matrix* $\mathbf{M}$ *obeying* $\|\mathbf{M} - \mathbf{I}\|_{\infty,\infty} < 1$ *is invertible and*

$$\|\mathbf{M}^{-1} - \mathbf{I}\|_{\infty,\infty} \leq \frac{\|\mathbf{M} - \mathbf{I}\|_{\infty,\infty}}{1 - \|\mathbf{M} - \mathbf{I}\|_{\infty,\infty}}.$$

Proof. The proof expresses $\mathbf{M}^{1/2}$ and $\mathbf{M}^{-1}$ as a power series of $(\mathbf{M} - \mathbf{I})$ and just uses the fact that $\|\cdot\|_{\infty,\infty}$ is an (induced) operator norm. We skip the details. □

B.2 Proof of Lemma 4

We first show that

$$\mathbb{P}\left(|\boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}'| > t\|\mathbf{R}\|_F\right) \leq 3qe^{-t/2}. \tag{B.4}$$

Let r_{jk} be the (j,k) coefficient of $\mathbf{R}$, and set $B = \|\mathbf{R}\|_F$, $B_j = \sqrt{\sum_k r_{jk}^2}$. We have

$$\eta := \boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}' = \sum_{j,k} r_{jk}\xi_j\xi'_k = \sum_j \xi_j\lambda_j, \quad \lambda_j := \sum_k r_{jk}\xi'_k$$

and

$$\mathbb{P}\left(|\eta| > tB\right) \leq \mathbb{P}\Big(|\sum_j \xi_j\lambda_j| > tB, \text{ given } |\lambda_j| \leq t^{1/2}B_j^{1/2}, \forall j\Big) + \sum_j \mathbb{P}\left(|\lambda_j| > t^{1/2}B_j^{1/2}\right).$$

We first apply Hoeffding's inequality [20, Eq. (2.6)] to bound the second term:

$$\mathbb{P}\left(|\lambda_j| > t^{1/2}B_j^{1/2}\right) \leq 2e^{-t/2}.$$

We also apply the same inequality to bound the first term, using the independence between the sequence ξ and ξ'. Conditionally,

$$\mathbb{P}\Big(|\sum_j \xi_j\lambda_j| > t^{1/2}\sqrt{\sum_j \lambda_j^2}\Big) \leq 2e^{-t/2}.$$

Now (B.4) follows from $t^{1/2}\sqrt{\sum_j \lambda_j^2} \leq tB$ which holds when $|\lambda_j| \leq t^{1/2}B_j^{1/2}$ for all j, together with the union bound.

Armed with (B.4), we proceed as follows:

$$\begin{aligned}
\mathbb{E}\left(\exp(\boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}')\right) &\leq \frac{1}{2} + \frac{1}{2}\int_0^\infty \mathbb{P}\left(|\boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}'| \geq \log t\right) dt \\
&\leq \frac{1}{2} + \frac{1}{2}\int_0^\infty \min\left[1,\ 3qt^{-1/(2\|\mathbf{R}\|_F)}\right] dt \\
&= \frac{1}{2} + \frac{1}{2}\left(\exp(2\|\mathbf{R}\|_F \log(3q)) + 2\|\mathbf{R}\|_F\right) \\
&\leq \exp(2\|\mathbf{R}\|_F \log(3q)) + \|\mathbf{R}\|_F.
\end{aligned}$$

The first inequality uses the symmetry about 0 of the distribution of $\boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}'$. The second inequality uses (B.4). The rest is a direct calculation.

B.3 Proof of Lemma 5

Integrating with respect to $\boldsymbol{\xi}'$ first, evaluates the moment generating function of $\boldsymbol{\xi}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at $\mathbf{R}^T \boldsymbol{\xi}$. This gives

$$\mathbb{E}\left(e^{\boldsymbol{\xi}^T \mathbf{R} \boldsymbol{\xi}'}\right) = \mathbb{E}\left(e^{\boldsymbol{\xi}^T \mathbf{R}^T \mathbf{R} \boldsymbol{\xi}}\right) = \det(\mathbf{I} - \mathbf{R}^T \mathbf{R})^{-1/2}.$$

Recall that $\|\mathbf{M}\|$ is the usual operator norm $\mathbf{M}$. We show that, for any positive semidefinite matrix $\mathbf{M}$ such that $\|\mathbf{M}\| < 1/2$,

$$\det(\mathbf{I} - \mathbf{M}) \geq \exp(-2\,\mathrm{trace}(\mathbf{M})). \tag{B.5}$$

Define $f(a) = \det(\mathbf{I} - a\mathbf{M})$, which is well-defined for $a \in [0, 1]$ and positive. The function f is differentiable, and

$$\frac{d}{da} \log f(a) = -\mathrm{trace}((\mathbf{I} - a\mathbf{M})^{-1}\mathbf{M}).$$

Therefore, by the mean value theorem, there is $a \in [0, 1]$ such that

$$\log \det(\mathbf{I} - \mathbf{M}) = -\mathrm{trace}((\mathbf{I} - a\mathbf{M})^{-1}\mathbf{M}).$$

Let $(\lambda_1, \ldots, \lambda_q)$ be the eigenvalues of $\mathbf{M}$. It follows from

$$(\mathbf{I} - a\mathbf{M})^{-1} = \sum_{s=0}^{\infty} a^s \mathbf{M}^s,$$

and $\mathrm{trace}(\mathbf{M}^s) = \sum_{j=1}^{q} \lambda_j^s$ that

$$\mathrm{trace}((\mathbf{I} - a\mathbf{M})^{-1}\mathbf{M}) \leq \sum_{j=1}^{q} \sum_{s=0}^{\infty} a^s \lambda_j^{s+1} = \sum_{j=1}^{q} \frac{\lambda_j}{1 - a\lambda_j} \leq 2 \sum_{j=1}^{q} \lambda_j = 2\,\mathrm{trace}(\mathbf{M}).$$

The inequality (B.5) follows from this.

Applying (B.5) with $\mathbf{M} = \mathbf{R}^T \mathbf{R}$ ($\|\mathbf{M}\| \leq \|\mathbf{R}\|_F^2 < 1/2$) gives

$$\det(\mathbf{I} - \mathbf{R}^T \mathbf{R})^{-1/2} \leq \exp(\mathrm{trace}(\mathbf{R}^T \mathbf{R})) \leq 1 + 2\|\mathbf{R}\|_F^2.$$

B.4 Proof of Lemma 9

The first part is a special case of [29, Cor. 2.1]. For the second part, we have

$$\mathbb{P}\left(z > s, z' > s'\right) = \frac{1}{2\pi\sqrt{1 - \zeta^2}} \int_s^{\infty} \int_{s'}^{\infty} e^{-\frac{x^2 + y^2 + 2\zeta xy}{2(1 - \zeta^2)}} dx dy.$$

Expand the integrand as a function of ζ (this is valid by dominated convergence). After some calculations, this gives

$$\mathbb{P}\left(z > s, z' > s'\right) = \mathbb{P}\left(z > s\right) \mathbb{P}\left(z' > s'\right) + \zeta \int_s^{\infty} x e^{-\frac{x^2}{2}} dx \int_{s'}^{\infty} y e^{-\frac{y^2}{2}} dy$$
$$+ O(\zeta^2) \int_s^{\infty} x^2 e^{-\frac{x^2}{2}} dx \int_{s'}^{\infty} y^2 e^{-\frac{y^2}{2}} dy,$$

where $O(\zeta^2)$ indicates a term of order ζ^2 independent of (s, s'). Hence,

$$\mathbb{P}\left(z > s, z' > s'\right) = \mathbb{P}\left(z > s\right)\mathbb{P}\left(z' > s'\right) + \zeta\phi(s)\phi(s') + O(\zeta^2)ss'\phi(s)\phi(s').$$

We apply this identity to each term in

$$\begin{aligned}\mathbb{P}\left(|z+a| > t, |z'+a'| > t\right) = \mathbb{P}\left(z > t-a, z' > t-a'\right) + \mathbb{P}\left(z > t-a, -z' > t+a'\right) \\ + \mathbb{P}\left(-z > t+a, z' > t-a'\right) + \mathbb{P}\left(-z > t+a, -z' > t+a'\right).\end{aligned}$$

Observing that $\mathrm{Cov}(z, -z') = \mathrm{Cov}(-z, z') = -\zeta$, we have

$$\begin{aligned}&\mathbb{P}\left(|z+a| > t, |z'+a'| > t\right) - \mathbb{P}\left(|z+a| > t\right)\mathbb{P}\left(|z'+a'| > t\right) \\ &\qquad = \zeta(\phi(t+a) - \phi(t-a))(\phi(t+a') - \phi(t-a')) + O(\zeta^2)t^2\phi(t)^2 \\ &\qquad = \zeta(4aa't^2 + O((a^3 + (a')^3)t^3))\phi(t)^2 e^{-(a^2+(a')^2)/2} + O(\zeta^2)t^2\phi(t)^2.\end{aligned}$$

The first equality uses $\phi(t \pm a), \phi(t \pm a') = O(\phi(t))$. The second uses a Taylor development of the exponential function. Both equalities use the fact that $|a|, |a'| \leq 1/t$. This proves the claim.

Acknowledgements

This work was partially supported by an ONR grant N00014-09-1-0258. We would like to thank Chiara Sabatti for stimulating discussions and for suggesting improvements on an earlier version of the manuscript, and Ewout van den Berg for help with the simulations.

References

[1] S. M. Berman. Equally correlated random variables. *Sankhyā Ser. A*, 24:155–156, 1962.

[2] S. M. Berman. Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.*, 35:502–516, 1964.

[3] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227, 2008.

[4] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, Dec. 2006.

[5] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete Fourier information. *IEEE Trans. Info. Theory*, 52(2):489–509, Feb. 2006.

[6] J. P. Castagna, S. Sun, and R. W. Siegfried. Instantaneous spectral analysis: Detection of low-frequency shadows associated with hydrocarbons. *The Leading Edge*, 22(2):120–127, 2003.

[7] G. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32:490–495, 2002.

[8] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[9] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[10] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[11] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk. Sparse signal detection from incoherent projections. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 3:III–III, May 2006.

[12] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, 18(1):71–103, 2003.

[13] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160, 2001.

[14] R. A. Fisher. *Statistical methods for research workers*. Hafner Publishing Co., New York, 1973. Fourteenth edition—revised and enlarged.

[15] J. J. Goeman, S. A. van de Geer, and H. C. van Houwelingen. Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(3):477–493, 2006.

[16] R. Gribonval and E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Process.*, 51(1):101–111, 2003.

[17] P. Hall and J. Jin. Properties of higher criticism under strong dependence. *Ann. Statist.*, 36(1):381–402, 2008.

[18] P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, 38(3):1686–1732, 2010.

[19] J. Haupt and R. Nowak. Compressive sampling for signal detection. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[20] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[21] M. Honig. *Advances in multiuser detection*. Wiley-IEEE Press, 2009.

[22] Y. I. Ingster. Minimax detection of a signal for ℓ_n^l balls. *Math. Methods Statist.*, 7:401–428, 1999.

[23] D. James, B. D. Clymer, and P. Schmalbrock. Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging*, 13(6):876–881, 2001.

[24] J. Jin. *Detecting and Estimating Sparse Mixtures*. PhD thesis, Stanford University, 2003.

[25] M. Kerr, M. Martin, and G. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000.

[26] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[27] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.

[28] P. W. H. Lemmens and J. J. Seidel. Equiangular lines. *J. Algebra*, 24:494–512, 1973.

[29] W. Li and Q. Shao. A normal comparison inequality and its applications. *Probability Theory and Related Fields*, 122(4):494–508, 2002.

[30] S. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2008.

[31] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[32] T. Mathew and B. K. Sinha. Optimum tests for fixed effects and variance components in balanced models. *J. Amer. Statist. Assoc.*, 83(401):133–135, 1988.

[33] M. McCarthy, G. Abecasis, L. Cardon, D. Goldstein, J. Little, J. Ioannidis, and J. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[34] J. Meng, H. Li, and Z. Han. Sparse Event Detection in Wireless Sensor Networks using Compressive Sensing. In *43rd Annual Conference on Information Sciences and Systems (CISS)*, 2009.

[35] D. C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons Inc., Hoboken, NJ, seventh edition, 2009.

[36] S. Piantadosi. *Clinical trials*. Wiley-Interscience, Hoboken, NJ, second edition, 2005.

[37] D. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32:502–508, 2002.

[38] T. Strohmer and R. W. Heath, Jr. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.

[39] C. Willer, S. Sanna, A. Jackson, A. Scuteri, L. Bonnycastle, R. Clarke, S. Heath, N. Timpson, S. Najjar, H. Stringham, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, 40(2):161–169, 2008.

[40] D. Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.

[41] G. Zhang, S. Zhang, and Y. Wang. Application of adaptive time-frequency decomposition in ultrasonic NDE of highly-scattering materials. *Ultrasonics*, 38(10):961–964, 2000.