Electronic Journal of Statistics
Vol. 14 (2020) 174–206
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1656

Estimation of linear projections of non-sparse coefficients in high-dimensional regression

David Azriel

Technion – Israel Institute of Technology, Haifa, Israel
e-mail: davidazr@technion.ac.il

Armin Schwartzman

University of California, San Diego, La Jolla, CA,
e-mail: armins@ucsd.edu

Abstract: In this work we study estimation of signals when the number of parameters is much larger than the number of observations. A large body of literature assumes for these kind of problems a sparse structure where most of the parameters are zero or close to zero. When this assumption does not hold, one can focus on low-dimensional functions of the parameter vector. In this work we study one-dimensional linear projections. Specifically, in the context of high-dimensional linear regression, the parameter of interest is $\boldsymbol{\beta}$ and we study estimation of $\mathbf{a}^T\boldsymbol{\beta}$. We show that $\mathbf{a}^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimator, using pseudo-inverse when $p > n$, is minimax and admissible. Thus, for linear projections no regularization or shrinkage is needed. This estimator is easy to analyze and confidence intervals can be constructed. We study a high-dimensional dataset from brain imaging where it is shown that the signal is weak, non-sparse and significantly different from zero.

MSC 2010 subject classifications: Primary 62J05, 60K35; secondary 62P10.
Keywords and phrases: High-dimensional regression, linear projections.

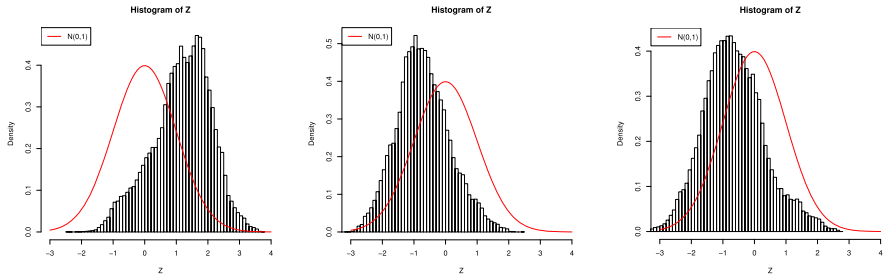
Received February 2019.

Contents

1. Introduction

This research emerged from analysis of a high-dimensional dataset obtained from brain imaging. The data belongs to a study of cortical thickness of adults who had a diagnosis of attention deficit/hyperactivity disorder (ADHD) as children (Proal et al., 2011). The dataset consists of cortical thickness measurements for about 80000 cortical voxels, obtained from magnetic resonance imaging (MRI) scans, as well as demographic and behavioral measurements, for each of 139 individuals. In this study, it had been noticed by Reiss et al. (2012) that z-scores corresponding to the voxelwise relationship between cortical thickness and ADHD diagnosis did not follow the theoretical standard normal distribution. Instead, the distribution of z-scores exhibited a substantial shift away from zero, indicating a possible widespread cortical thinning over the brain for individuals

with ADHD. It is unclear, however, whether those results could have been caused by correlation between voxels rather than by a real relationship with clinical diagnosis (Azriel and Schwartzman, 2015).

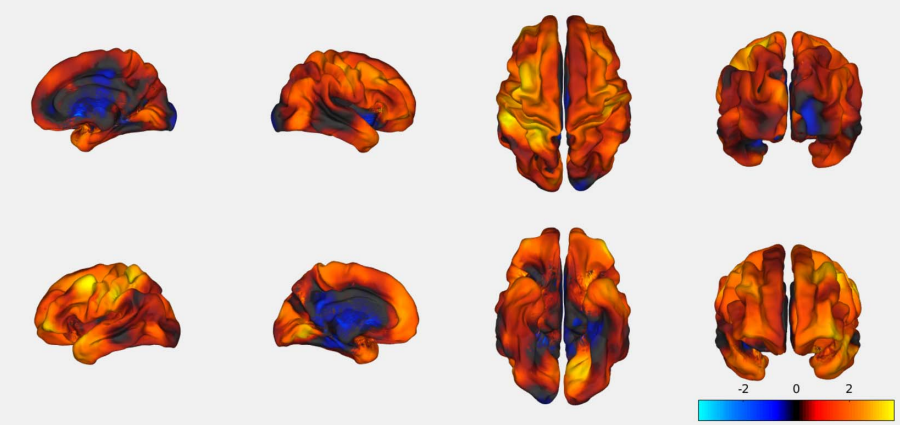


(a) Data (b) Simulation - no signal (c) Simulation - signal

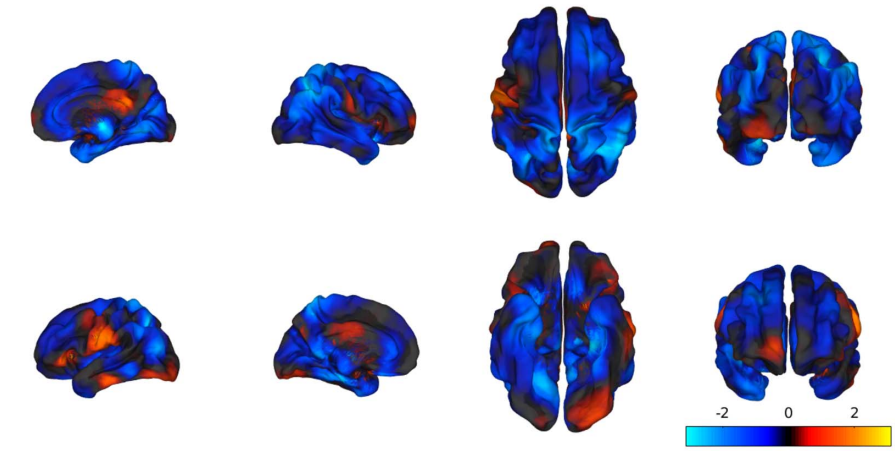
FIG 1. *Histogram of the z-scores of the real data (a) and of the simulated data (b,c), where in (b) there is no signal and in (c) there is. The red line illustrates the standard normal density.*

Figure 1 shows histograms of (a) the actual z-scores (standardized to be distributed $N(0,1)$ under the global null hypothesis) and (b,c) the z-scores in simulated data sets; see Section 2.2 below for exact definitions. In (b) the response variable is simulated independently (no signal) whereas in (c) the regression coefficients are not zero (simulated from a normal distribution). In all cases, there is a clear departure from the global null, but in (b) the departure is caused only by the correlation structure and in (c) it can be attributed to a true signal. In case (a) it is not clear if there is a signal or not. The fact that the correlation structure can cause shifted z-scores can also be seen in the brain maps in Figure 2; (a) corresponds to Figure 1(a), showing the observed widespread positive z-values, while (b) corresponds to Figure 1(b), showing the simulated widespread negative z-values. Thus, as originally pointed out by Efron (2007) (and later made more precise by Schwartzman (2008) and Azriel and Schwartzman (2015)), even when the theoretical null model is correct, the empirical distribution of the test statistics can be different from the theoretical null distribution simply because of the correlation structure. In this work, we aim at estimating such departures and detecting whether the departure is due to correlation, or to true signal.

In our motivating dataset, there are two groups: 59 adults in whom ADHD had been established in childhood, along with 80 controls. Previous works (Proal et al., 2011; Reiss et al., 2012) considered the resulting multiple hypotheses problem (voxel-wise) and used the FDR criterion to detect areas in the cortical region where the null is rejected. As discussed below, this approach has low power to detect signals that are weak and non-sparse. As an alternative perspective, we set up here the problem as a regression of the response on the cortical thickness measurements as predictors, and attempt to find linear projections of the regression coefficients that will indicate a spatial distribution of the signal. In Azriel and Schwartzman (2015) we studied the distribution of the z-scores under the global null hypothesis accounting for the correlation structure. This allows



(a) Data



(b) Simulation - 1

FIG 2. *Brain maps of the z-scores of the real data (a) and of the simulated data (b).*

us to test the global null, as we show here (in Section 5.2), but not to estimate the signal, which is the interest of the current work.

Consider the linear regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is i.i.d. with mean zero and variance σ^2 and $\mathbf{X}$ is a fixed $n \times p$ matrix. While in many cases, including the brain imaging dataset mentioned above, $\mathbf{X}$ is in fact random, in this work we adopt the conditionality principle (Birnhaum, 1962) and treat it as fixed since it is ancillary to the parameter of interest. Here we focus on the high-dimensional case $p > n$. In this context, the Lasso estimator suggested by Tibshirani (1996) has gained much popularity and many extensions were suggested. That line of research is related to sparsity assumptions where most of the parameters are assumed to be zero or close to zero. Those assumption

are violated in many datasets including the one that motivated our study, as demonstrated by the histogram of Figure 1. Dicker (2014) and Janson et al. (2017) studied inference of signal-to-noise ratio, and of σ^2 in the non-sparse case. For multiple comparisons, the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) is guaranteed to choose the non-zero parameters with controlled error rate, but in our dataset, no significant voxels (at the level of $\alpha = 0.05$) are found by BH. Indeed, the BH estimator is highly variable under strong dependence (Owen, 2005; Schwartzman and Lin, 2011) and has low power when there are many non-zero parameters but all have still small values.

To answer the question of whether a non-sparse signal is present, we aim at estimating $\theta = \mathbf{a}^T\boldsymbol{\beta}$ (when it is identifiable) for a predetermined vector $\mathbf{a}$. When $\mathbf{a}$ is sparse, θ corresponds to a small subset of $\boldsymbol{\beta}$ that is of interest, while for non-sparse $\mathbf{a}$, θ is a global measure of the signal. We show that the estimator $\hat{\theta} = \mathbf{a}^T \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{Y}$, where $\left(\mathbf{X}^T\mathbf{X}\right)^-$ denotes the Moore-Penrose pseudo-inverse, is unbiased, admissible and minimax. Moreover, its distribution is easily derived and therefore one can construct confidence intervals and perform hypothesis testing. We also study the asymptotic behavior of this estimator and show that it is consistent in certain cases. Even if it is not consistent, a confidence interval can be constructed, it just does not shrink to zero.

Our estimator is a linear function of the z-scores, which are defined in Section 2. We first study estimation of linear combinations of the expectation vector of the z-scores in Section 3 and then return to the problem of estimation of θ in Section 4. We analyze the motivating dataset in Section 5. Section 6 concludes with final remarks. The proofs of the theoretical results and additional results are given in the appendices.

2. The use of z-scores

In high-dimensional regression problems, it is common to reduce the data to univariate z-scores, computed from univariate regressions of the outcome on each predictor. In our situation, univariate z-scores are a useful tool not only for data analysis but as a theoretical tool for deriving the desired estimator.

2.1. Model

The brain imaging dataset we study consists of $n = 139$ subjects with information on $p = 81924$ vertices per subject. Let Y_i denote the behavioral assessment of the i-th subject and let $X_i^{(j)}$ to be the cortical thickness in the j-th vertex of the i-th subject. We consider the regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where the i,j-th entry of $\mathbf{X}$ is $X_i^{(j)}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown coefficients and the ε's are i.i.d with mean zero and variance σ^2. Here both the response and the covariates are assumed centered so that there

is no need for an intercept in the model, and so the matrix $\mathbf{X}$ characterizes the substantive predictors. For simplicity, we ignore the weak dependence induced by the centering of the response.

2.2. z-scores

Consider the simple linear regression estimates $\hat{\beta}^{(j)} = \sum_{i=1}^{n} X_i^{(j)} Y_i / s_{jj}$, where $s_{jk} := \sum_{i=1}^{n} X_i^{(j)} X_i^{(k)}$, and define the z-scores

$$\tilde{Z}_j = \frac{\hat{\beta}^{(j)}}{se(\hat{\beta}^{(j)})} = \frac{\sum_{i=1}^{n} X_i^{(j)} Y_i}{\sqrt{s_{jj}}\sigma}.$$

Then we have $E(\hat{\beta}^{(j)}) = \beta^{(j)} + \sum_{k \neq j} s_{jk}\beta^{(k)}/s_{jj}$ and $Var(\hat{\beta}^{(j)}) = \sigma^2/s_{jj}$, so that

$$E(\tilde{Z}_j) = \sqrt{s_{jj}}\frac{\beta^{(j)}}{\sigma} + \frac{1}{\sigma\sqrt{s_{jj}}}\sum_{j \neq k} s_{jk}\beta^{(k)}, \qquad Var(\tilde{Z}_j) = 1,$$

$$Cov(\tilde{Z}_j, \tilde{Z}_k) = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}.$$

Letting $\mathbf{D}$ be a diagonal matrix with $D_{jj} = \sqrt{s_{jj}}$, we can rewrite in matrix form as

$$\tilde{\mathbf{Z}} = \mathbf{D}^{-1}\mathbf{X}^T\mathbf{Y}/\sigma \tag{2}$$

with respective expectation and covariance matrix

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}\mathbf{D}\boldsymbol{\beta}/\sigma, \qquad \tilde{\boldsymbol{\Sigma}} = \mathbf{D}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{D}^{-1}. \tag{3}$$

Let $\mathbf{A}^-$ denote the pseudo-inverse of matrix $\mathbf{A}$. For any vector $\mathbf{a}$ such that $\theta = \mathbf{a}^T\boldsymbol{\beta}$ is identifiable, we show in Proposition 3 below that there exists a vector $\tilde{\mathbf{a}} = \sigma\mathbf{D}\left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{a}$ such that $\mathbf{a}^T\boldsymbol{\beta} = \tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}}$, mapping the estimation of linear functions of $\boldsymbol{\beta}$ to linear functions of $\tilde{\boldsymbol{\mu}}$. Conversely, for any $\tilde{\mathbf{a}}$, we can define $\mathbf{a} = \mathbf{D}\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{a}}/\sigma$ such that $\tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}} = \mathbf{a}^T\boldsymbol{\beta}$, mapping linear functions of $\tilde{\boldsymbol{\mu}}$ to linear functions of $\boldsymbol{\beta}$.

Notice that the pairwise correlations between the z-scores are the pairwise correlations between the cortical thickness measurements at each vertex. Because the regression is conditional on the vertexwise measurements, we take the pairwise correlations as fixed and known. Note too that the definition of $\tilde{\mathbf{Z}}$ involves σ, which is unknown. Estimation of σ^2 is discussed in Section 5; for now it is considered as a known constant.

2.3. The natural estimator

Writing $\theta = \tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}}$, the natural estimator is $\hat{\theta} = \tilde{\mathbf{a}}^T\tilde{\mathbf{Z}}$. As simple as this estimator is, it turns out it is not a bad one. It is unbiased, and we show in Section 3

that, under Model (1) when assuming that the ε's are normal, it is minimax and admissible.

If we write this estimator in terms of the representation $\theta = \mathbf{a}^T\boldsymbol{\beta}$ we obtain the elegant form $\hat{\theta} = \mathbf{a}^T \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{Y}$ (Section 4.2). In order to investigate the properties of this estimator we first study the general problem of estimating $\theta = \tilde{\mathbf{a}}^T\boldsymbol{\mu}$ using $\mathbf{Z}$, which is a vector with mean $\boldsymbol{\mu}$ and finite variance. This is done in the next section. We return to the regression problem in Section 4.

3. Properties of the natural estimator

In this section we study the general model where $\mathbf{Z}$ is a vector of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Our interest is in estimation of $\theta = \tilde{\mathbf{a}}^T\boldsymbol{\mu}$ and we study the properties of the natural estimator $\tilde{\mathbf{a}}^T\mathbf{Z}$. The basic setting and notation is described in Section 3.1. Section 3.2 shows that the natural estimator is minimax and admissible.

3.1. Dimension reduction

Suppose that $\mathbf{Z} := (Z_1, \ldots, Z_p)$ is a vector of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Our interest is in estimating $\theta = \tilde{\mathbf{a}}^T\boldsymbol{\mu}$ and the natural estimator is $\hat{\theta} = \tilde{\mathbf{a}}^T\mathbf{Z}$. We will suppress in the notation the dependence on p when there is no confusion.

Let $\boldsymbol{\Gamma}$ be the matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$, denoted by $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$, and let $\boldsymbol{\Lambda}$ be the diagonal matrix with the corresponding eigenvalues, denoted by $\lambda_1 \geq \cdots \geq \lambda_p$, in the diagonal. When the rank of $\boldsymbol{\Sigma}$, denoted by r, is smaller than p (as in our motivating dataset), then $\boldsymbol{\Sigma}$ has r positive eigenvalues and the rest $p-r$ eigenvalues are 0. In this case, we define $\boldsymbol{\Lambda}$ to be $r \times r$ diagonal matrix with the r positive eigenvalues in the diagonal and we define $\boldsymbol{\Gamma}$ to be $p \times r$ matrix with the r corresponding eigenvectors in the columns. We have that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$.

It is convenient to work with the following dimension reducing transformation $\mathbf{W} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}^T\mathbf{Z}$. The length of $\mathbf{W}$ is $r \leq p$ and

$$E(\mathbf{W}) = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}^T\boldsymbol{\mu} := \boldsymbol{\eta}, \qquad Cov(\mathbf{W}) = \mathbf{I}_{r\times r}.$$

Hence, $\mathbf{W}$ is an uncorrelated, low dimensional representation of $\mathbf{Z}$. Similar to principal components analysis (PCA), $\mathbf{W}$ and $\boldsymbol{\eta}$ can be thought of as the expressions of $\mathbf{Z}$ and $\boldsymbol{\mu}$ in the canonical coordinate system defined by the covariance matrix $\boldsymbol{\Sigma}$.

When $\boldsymbol{\Sigma}$ is of full rank, we can write in this coordinate system, $\theta = \mathbf{b}^T\boldsymbol{\eta}$, where $\mathbf{b} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^T\tilde{\mathbf{a}}$. This can be generalized to the degenerate case $r < p$ as stated in the following proposition, which also shows that there is a one-to-one correspondence between $\mathbf{Z}$ (of length p) and $\mathbf{W}$ (of length r). That is, even though $\mathbf{Z}$ belongs to $\mathbb{R}^p$, it lies in a r-dimensional sub-space. This implies that θ can be estimated using either $\mathbf{Z}$ or $\mathbf{W}$ and both ways are equivalent.

Proposition 1. *Suppose that $r < p$.*

(I) *Let* $\mathbf{\Gamma}_\perp$ *be the* $p \times (p-r)$ *matrix whose columns are the eigenvectors orthogonal to* $\mathbf{\Gamma}$. *Then* $\mathbf{\Gamma}_\perp^T \mathbf{Z} = \mathbf{\Gamma}_\perp^T \boldsymbol{\mu}$ *is a deterministic known vector, and therefore we can assume without loss of generality that it equals* $\mathbf{0}$.

(II) *If* $\mathbf{\Gamma}_\perp^T \boldsymbol{\mu} = 0$, *then the same relation as in the full-rank case holds, namely,* $\theta = \mathbf{b}^T \boldsymbol{\eta}$ *with* $\mathbf{b} = \mathbf{\Lambda}^{1/2} \mathbf{\Gamma}^T \tilde{\mathbf{a}}$.

(III) *If* $\mathbf{\Gamma}_\perp^T \boldsymbol{\mu} = 0$, *then the dimension reduction transformation is one-to-one and we can write* $\mathbf{Z} = \mathbf{\Gamma} \mathbf{\Lambda}^{1/2} \mathbf{W}$.

3.2. Minimaxity and admissibility of the natural estimator

The natural estimator is $\hat{\theta} = \tilde{\mathbf{a}}^T \mathbf{Z} = \mathbf{b}^T \mathbf{W}$. We have that $E(\hat{\theta}) = \theta$ and that $Var(\hat{\theta}) = \tilde{\mathbf{a}}^T \mathbf{\Sigma} \tilde{\mathbf{a}} = \mathbf{b}^T \mathbf{b}$. Hence, the squared-error loss $E(\hat{\theta} - \theta)^2 = \tilde{\mathbf{a}}^T \mathbf{\Sigma} \tilde{\mathbf{a}} = \mathbf{b}^T \mathbf{b}$ does not depend on θ. Also, when $\mathbf{Z}$ is normal, as shown in the proof of the result below, $\hat{\theta}$ is a limit of Bayes rules. These properties imply that $\hat{\theta}$ is minimax and admissible. The result it stated below, and the formal proof is given in Appendix D.

Theorem 1. *Let* $\mathbf{Z} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$ *and consider the natural (maximum likelihood) estimator* $\hat{\theta} = \mathbf{b}^T \mathbf{W} = \tilde{\mathbf{a}}^T \mathbf{Z}$.

(I) (Minimaxity) *The natural estimator is minimax for every p; i.e., for all estimators* $\tilde{\theta}$, $\sup_{\boldsymbol{\eta} \in \mathbb{R}^r} E(\tilde{\theta} - \theta)^2 \geq \mathbf{b}^T \mathbf{b} = \sup_{\boldsymbol{\eta} \in \mathbb{R}^r} E(\hat{\theta} - \theta)^2$.

(II) (Admissibility) *The natural estimator is admissible for every p; i.e., there exists no estimator* $\tilde{\theta}$ *such that* $E(\tilde{\theta} - \theta)^2 \leq \mathbf{b}^T \mathbf{b}$ *for all* $\boldsymbol{\eta} \in \mathbb{R}^r$ *with strict inequality for some* $\boldsymbol{\eta}$.

In Appendix A an asymptotic analysis of the variance of the natural estimator is presented. Specifically, it is shown that under some regularity conditions, when the natural estimator is inconsistent, then so is every other estimator. Yet, even if the variance of $\hat{\theta}$ does not go to zero, a confidence interval can still be constructed; it just does not shrink.

4. High-dimensional regression

In this section we study estimation of linear projections of $\boldsymbol{\beta}$, which is meaningful only when those linear projections are identifiable. Section 4.1 provides a sufficient and necessary condition for identifiability of θ. The results of Section 3.2 are applied to the regression setting in Section 4.2. The natural estimator is compared to the ridge regression estimate in Section 4.3 and a simulation study is reported in Section 4.4. Section 4.5 introduces an asymptotic analysis of the estimator's variance.

4.1. Identifiability

Recall the regression setting described in Section 2. The full vector $\boldsymbol{\beta}$ is not identifiable: if $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}'$ then $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are indistinguishable. Here we are

interested in estimation of $\theta = \mathbf{a}^T\boldsymbol{\beta}$. Proposition 2 below provides a necessary and sufficient condition for identifiability of θ. To state the result, we use the eigendecomposition notation of Section 3.1 for $\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X}/n = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$.

Proposition 2. *Let $\theta = \mathbf{a}^T\boldsymbol{\beta}$ and define the $r \times 1$ vector $\boldsymbol{\alpha} = \boldsymbol{\Gamma}^T\mathbf{a}$. Consider the orthogonal decomposition $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\alpha} + \boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp$; then θ is identifiable iff $\boldsymbol{\alpha}_\perp = \mathbf{0}$.*

The proposition indicates that θ is identifiable iff $\boldsymbol{\Gamma}_\perp^T\mathbf{a} = 0$, i.e., iff $\mathbf{a}$ belongs to the subspace spanned by the columns of $\boldsymbol{\Gamma}$. Another way to understand Proposition 2 is that it specifies the part of θ that is estimable. If we write $\theta = \mathbf{a}^T\boldsymbol{\beta} = \boldsymbol{\alpha}^T(\boldsymbol{\Gamma}^T\boldsymbol{\beta}) + \boldsymbol{\alpha}_\perp^T(\boldsymbol{\Gamma}_\perp^T\boldsymbol{\beta})$ by the orthogonal decomposition of $\mathbf{a}$, we see that θ contains the portion $\boldsymbol{\Gamma}^T\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ that projects $\boldsymbol{\beta}$ onto the subspace spanned by the columns of $\mathbf{X}$ and the portion orthogonal to it. The former is accessible through the observations (linear model) but the latter is not and therefore not estimable. We will see this phenomenon in the data analysis (Section 5.4).

Proposition 2 implies that $\theta = \mathbf{a}^T\boldsymbol{\beta}$ is identifiable when $\mathbf{a} \in \mathbb{R}^p$ belongs to a subspace of dimension r. Since $r \leq n$, when $p > n$, then for "most" $\mathbf{a}$'s, $\theta = \mathbf{a}^T\boldsymbol{\beta}$ is not identifiable. In practice, there are two ways to proceed. First, for a given $\mathbf{a}$ one can consider instead of $\mathbf{a}$ the closest identifiable vector, i.e.,

$$\arg \min_{\tilde{\mathbf{a}} \in \{\boldsymbol{\Gamma}\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^r\}} \|\tilde{\mathbf{a}} - \mathbf{a}\|^2,$$

which by standard linear algebra is $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\mathbf{a}$. Second, one can consider all possible directions, $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r$ (the columns of $\boldsymbol{\Gamma}$) and adjust for multiplicity. These two ways are demonstrated in our motivating dataset, as described in Section 5.4.

4.2. Estimation of θ

We are interested in estimating $\theta = \mathbf{a}^T\boldsymbol{\beta}$. Assuming that θ is identifiable, Proposition 2 implies that $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\alpha}$. The natural estimator of θ is given by the following result. Recall the definitions of the z-score vector $\tilde{\mathbf{Z}}$ and its expectation $\tilde{\boldsymbol{\mu}}$ given by (2) and (3).

Proposition 3. *Assume that θ is identifiable. Let $\tilde{\mathbf{a}} = \frac{\sigma}{n}\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^T\mathbf{a} = \sigma\mathbf{D}\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{a}$, where $\mathbf{A}^-$ denotes the pseudo-inverse of matrix $\mathbf{A}$.*

(I) The parameter of interest $\theta = \mathbf{a}^T\boldsymbol{\beta}$ is equal to $\theta = \tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}}$.

(II) The natural estimator of θ is $\tilde{\mathbf{a}}^T\tilde{\mathbf{Z}} = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{X}^T\mathbf{Y}$. Its variance is $\sigma^2\mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{a}$.

Interestingly, the form of the natural estimator is very similar to that of the usual linear regression estimate in low dimensions, except that here the pseudo-inverse is used because the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible. The form of the variance is a generalization of the variance of $\mathbf{a}^T\hat{\boldsymbol{\beta}}$, when $\hat{\boldsymbol{\beta}}$ is the least squares estimate. Regarding the quality of the natural estimator, Theorem 1 applies and we have the following corollary.

Corollary 1. *Assume Model* (1) *where* $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$ *and consider estimation of* $\theta = \mathbf{a}^T\boldsymbol{\beta}$ *when it is identifiable. The natural estimator,* $\hat{\theta} = \mathbf{a}^T \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{Y}$, *is minimax and admissible among all estimates that depend on* $\mathbf{Y}$ *through* $\mathbf{X}^T\mathbf{Y}$.

4.3. Comparison to Ridge

To illustrate the admissibility and minimaxity of the natural estimator, consider the ridge regression estimate

$$\hat{\boldsymbol{\beta}}_\lambda^R = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y},$$

for a tuning parameter λ. The corresponding estimate of θ is $\mathbf{a}^T\hat{\boldsymbol{\beta}}_\lambda^R$; it coincides with the natural estimator when $\lambda = 0$. The mean square error of $\boldsymbol{\Gamma}^T\hat{\boldsymbol{\beta}}_\lambda^R$ is (Farebrother, 1976)

$$(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}(\sigma^2\boldsymbol{\Lambda} + \lambda^2\boldsymbol{\Gamma}^T\boldsymbol{\beta}^T\boldsymbol{\beta}\boldsymbol{\Gamma})(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1},$$

and therefore the mean square error of $\mathbf{a}^T\hat{\boldsymbol{\beta}}_\lambda^R = \boldsymbol{\alpha}^T\tilde{\boldsymbol{\Gamma}}^T\hat{\boldsymbol{\beta}}^R$ is

$$\begin{aligned}&\boldsymbol{\alpha}^T(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}(\sigma^2\boldsymbol{\Lambda} + \lambda^2\boldsymbol{\Gamma}^T\boldsymbol{\beta}^T\boldsymbol{\beta}\boldsymbol{\Gamma})(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}\boldsymbol{\alpha}\\ &= \boldsymbol{\alpha}^T(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}\sigma^2\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}\lambda^2\boldsymbol{\Gamma}^T\boldsymbol{\beta}^T\boldsymbol{\beta}\boldsymbol{\Gamma}(\boldsymbol{\Lambda} + \lambda\mathbf{I})^{-1}\boldsymbol{\alpha};\end{aligned}$$

the first summand is the variance and the second is the squared bias. The latter is not bounded in $\boldsymbol{\beta}$ and therefore $\mathbf{a}^T\hat{\boldsymbol{\beta}}_0^R$, which is the natural estimator, is minimax. When λ is sufficiently small then the mean square error of $\mathbf{a}^T\hat{\boldsymbol{\beta}}_\lambda^R$ is smaller than that of $\mathbf{a}^T\hat{\boldsymbol{\beta}}_0^R$ (Farebrother, 1976). However, the optimal λ depends on unknown quantities and hence the natural estimator, which is $\mathbf{a}^T\hat{\boldsymbol{\beta}}_0^R$, is still admissible.

4.4. Comparison to Ridge and Lasso via simulations

To illustrate the theoretical results, we compared the natural estimator to the Lasso and Ridge estimates in a simulation study. For Lasso and Ridge estimates, the tuning parameter λ was chosen using cross-validation (computed by the 'parcor' package in R). We chose $n = 100$ and $p = 500$. Each row $\mathbf{X}_i = (X_i^{(1)}, \ldots, X_i^{(p)})$ was sampled from a multivariate normal distribution with mean 0 and exchangeable correlation structure with $\rho = 0.7$ (see Appendix B for exact definitions); $\mathbf{X}$ is fixed across all simulations. For each simulated data sets we sampled $\varepsilon_1, \ldots, \varepsilon_n \sim^{i.i.d} N(0, 1)$. We repeated the above procedure 1000 times and considered model (1) under three scenarios:

- Full $\boldsymbol{\beta}$: $\beta_j = 1/\sqrt{500}$ for $j = 1, \ldots, 500$,
- Half-full $\boldsymbol{\beta}$: $\beta_j = \begin{cases} 2/\sqrt{500} & j = 1, \ldots, 250 \\ 0 & j = 251, \ldots, 500 \end{cases}$,

- Sparse $\boldsymbol{\beta}$: $\beta_j = \begin{cases} 1/\sqrt{5} & j = 1, \ldots, 5 \\ 0 & j = 6, \ldots, 500 \end{cases}$;

under all three scenarios $||\boldsymbol{\beta}|| = 1$. For all scenarios we considered estimation of $\theta = \mathbf{a}^T\boldsymbol{\beta}$ for $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\mathbf{1}$, which is the closest vector (in L_2 sense) to $\mathbf{1}$ spanned by the eigenvectors corresponding to positive eigenvalues, yielding an identifiable θ.

The simulation average and standard deviation of the mean square error (MSE) are reported in Table 1. It is shown that the risk of the natural estimator is constant across all scenarios. For the full and half-full $\boldsymbol{\beta}$ scenarios, the risk of the natural estimator is smaller than both Lasso and Ridge. Under the sparse scenario the risk of the natural estimator and ridge is about the same. Lasso under-performs in all three scenarios but it has the advantage of identifying the non-zero entries of $\boldsymbol{\beta}$ in the sparse scenario. Overall, the results agree with our theoretical findings: the natural estimator estimates θ well in a minimax sense.

TABLE 1
Simulation average (standard error) of the MSE.

Scenario	Natural estimator	Lasso	Ridge
Full $\boldsymbol{\beta}$	0.020 (0.0009)	0.421 (0.0054)	0.040 (0.0015)
Half-full $\boldsymbol{\beta}$	0.020 (0.0009)	0.396 (0.0054)	0.028 (0.0011)
Sparse $\boldsymbol{\beta}$	0.021 (0.0009)	0.024 (0.0011)	0.020 (0.0009)

4.5. Consistency

We now study the asymptotic behavior of $\hat{\theta}$. For the asymptotic analysis, we assume the following regularity conditions (recall that here $\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X}/n$):

$$\text{The diagonal elements of } \boldsymbol{\Sigma} \text{ are bounded from above} \tag{4}$$

$$\text{The positive eigenvalues of } \boldsymbol{\Sigma} \text{ are bounded from below} \tag{5}$$

$$\frac{\boldsymbol{\gamma}_j^T\boldsymbol{\beta}}{\boldsymbol{\gamma}_{j'}^T\boldsymbol{\beta}} \text{ is bounded for every } j, j' \in \{1, \ldots, r\}. \tag{6}$$

$$Var(Y) \text{ and } \sigma^2 \text{ are } O(1). \tag{7}$$

Condition (4) is natural and is also assumed in the asymptotic analysis in Appendix A. Condition (5) guarantees that $\boldsymbol{\Sigma}$ does not become degenerate on the subspace spanned by the columns of $\boldsymbol{\Gamma}$ (which is the relevant subspace; see Proposition 2). This means that the columns of $\mathbf{X}$ are bounded away from being linearly dependent. By (6), $\boldsymbol{\beta}$ is "equally spread out" over all the eigenvectors that correspond to the positive eigenvalues, i.e., $\boldsymbol{\beta}$ is not concentrated on only part of the subspace. This means that $\boldsymbol{\beta}$ is not sparse in the coordinate system defined by $\boldsymbol{\Gamma}$ (although it could be sparse in the original coordinate system). When (7) holds true, then the signal-to-noise ratio is of order of a constant.

Conditions (4) and (5) can be easily verified since $\boldsymbol{\Sigma}$ is fixed. In our motivating dataset, the maximal diagonal element of $\boldsymbol{\Sigma}$ is 0.71 and the minimal positive eigenvalue is 2.80. In contrast, conditions (6) and (7) depend on unknown parameters and cannot be checked directly. While condition (7) is natural, condition

(6) is less so. However, Proposition 4 below gives an asymptotic approximation to $Var(\hat{\theta})$, which depends on these conditions. Thus, in practice, one can compute $Var(\hat{\theta})$ exactly and compare it to the approximation. We do so in the data analysis in Section 5.4 and find the fit between the theoretical and actual variances to be good.

The following proposition states the order of magnitude of the variance of $\hat{\theta}$ for two cases of $\mathbf{a}$.

Proposition 4. *Assume Model (1) and consider estimation of* $\theta = \mathbf{a}^T\boldsymbol{\beta}$ *when it is identifiable. Suppose that* $\theta = O(1)$*, and that (4)–(7) hold true. Consider two cases of* $\boldsymbol{\alpha}$ *(recall that* $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\alpha}$*):*

- **Global average***:* $\alpha_j = c_\alpha O(1)$ *where* c_α *depends on* p, n *and the* $O(1)$ *term is bounded uniformly over all* $j = 1, \ldots, r$ *(i.e.,* $\frac{\alpha_j}{c_\alpha}$ *is bounded uniformly). Then* $Var(\hat{\theta})$ *equals* $O(1)\frac{p}{nr^2}\sum_{j=1}^{r}\frac{1}{\lambda_j}$*.*
- **Single entry***:* $\boldsymbol{\alpha} = c_\alpha \mathbf{e}_j$ *for a scalar* c_α *and where*

$$\mathbf{e}_j := (0, \ldots, 0, \underbrace{1}_{j^{th}\ place}, 0, \ldots, 0).$$

Then $Var(\hat{\theta})$ *equals* $O(1)\frac{p}{n\lambda_j}$*.*

The variance of $\hat{\theta}$ in the global average case depends on the asymptotic behavior of p and $\sum_{j=1}^{r}\frac{1}{\lambda_j}$. For high-dimensional $\mathbf{X}$ $(p > n)$, we have that $r = n$. In this setting suppose that there are K eigenvalues of order p (spike model) and all the remaining eigenvalues are of the same order. Because $\sum_{j>K}\lambda_j = O(p)$, that order must be p/n, and then $\sum_{j=1}^{n}\frac{1}{\lambda_j} = O(n^2/p)$. As a consequence, the variance is $O(1/n)$ and $\hat{\theta}$ is consistent. Also when $\mathbf{X}$ is low-dimensional $(p = o(n))$, then $r = p$. Because all λ_j are bounded from below (Condition (5)), then $\sum_{j=1}^{p}\frac{1}{\lambda_j} \leq O(p)$ and therefore $\hat{\theta}$ is consistent.

For the single entry case, consistency depends on the eigenvalue λ_j and the rate of grow of p. For high-dimensional $\mathbf{X}$ $(p > n)$, when λ_j is "large", i.e., of order of p, then $\hat{\theta}$ is consistent. When $\mathbf{X}$ is low-dimensional $(p = o(n))$, $\hat{\theta}$ is consistent (using condition (5) that λ_j is bounded from below).

As before, regardless of the consistency of $\hat{\theta}$, a confidence interval for θ can still be constructed; it just may not shrink.

5. Analysis of the data

We now return to the brain imaging dataset, which was described in Section 2.

5.1. Estimation of σ^2

In order to calculate the z-scores, σ^2 needs to be estimated. Estimation of σ^2 when $\boldsymbol{\beta}$ is non-sparse is a topic of two recent papers (Dicker, 2014; Janson

et al., 2017). They work under the framework of random $\mathbf{X}$ and require rather restrictive assumptions on the distribution of $\mathbf{X}$. Here we use the simple estimate $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, which is consistent for $Var(Y)$. The latter is an upper bound since $Var(Y) \geq Var(Y|\mathbf{X}) = \sigma^2$. Therefore, the resulting confidence intervals are conservative. The upper bound is tight in the null case when $\boldsymbol{\beta} = 0$.

5.2. Testing for the global null in the brain imaging dataset

In Azriel and Schwartzman (2015) we studied the empirical cumulative distribution of a large number of correlated normals. We also analyzed the above dataset assuming the global null holds true, i.e., that $\boldsymbol{\beta} = 0$.

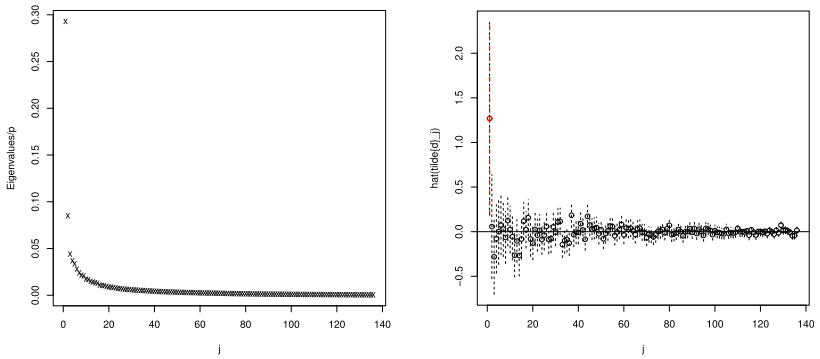
Consider the z-scores (denoted by $\tilde{\mathbf{Z}}$) and its correlation matrix (denoted by $\tilde{\boldsymbol{\Sigma}}$) defined in (2) and (3). Roughly speaking, we showed that if K is the number of "large" eigenvalues of $\tilde{\boldsymbol{\Sigma}}$, i.e., of order p, then there exists a vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^T$, where $\xi_1, \ldots, \xi_K \sim^{i.i.d} N(0,1)$, such that $\tilde{\mathbf{Z}}|\boldsymbol{\xi}$ is weakly correlated (i.e., the Frobenius norm of the correlation matrix divided by p converges to zero). Let $\tilde{m}_i := E(\tilde{Z}_i|\boldsymbol{\xi})$, $i = 1, \ldots, p$, then we have that under the global null $\sum_{i=1}^{p}\tilde{m}_i^2 \sim \sum_{j=1}^{K}\lambda_j\xi_j^2$ where $\xi_1, \ldots, \xi_K \sim^{i.i.d} N(0,1)$.

In the above dataset we found that $K = 2$ works quite well. When $K = 2$ we have that $\sum_{i=1}^{p}\tilde{m}_i^2 = 132018.9$. The two large eigenvalues are $\lambda_1 = 23995.7$ and $\lambda_2 = 6959.6$. Thus, for $T = \lambda_1\xi_1^2 + \lambda_2\xi_2^2$ we have that $P(T > 132018.9) = 0.023$ (computed by simulations), i.e., the p-value for the above test is 0.023, indicating that the global null is rejected at the $\alpha = 0.05$ level. We repeated the above computation with $K = 3, \ldots, 10$ and the resulting p-values were all smaller than 0.03. This result seems to imply that $\boldsymbol{\beta} \neq \mathbf{0}$. Since $\boldsymbol{\beta}$ is a large vector, it cannot be estimated. Instead, below we study its low-dimensional projections.

5.3. Estimation of $\tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}}$

Consider now the mean of the z-scores $\tilde{\boldsymbol{\mu}} = E(\tilde{\mathbf{Z}})$ defined in (3). We first estimate $\theta = \frac{1}{p}\sum_{j=1}^{p}\tilde{\mu}_j$ ($\tilde{\mathbf{a}} = \mathbf{1}/p$). Here $\hat{\theta} = \tilde{\mathbf{a}}^T\tilde{\mathbf{Z}} = 1.173$ and $\tilde{\mathbf{a}}^T\tilde{\boldsymbol{\Sigma}}\tilde{\mathbf{a}} = 0.264$. The latter can be efficiently computed by $\sum_{j=1}^{r}\alpha_j^2\tilde{\lambda}_j$, where $\tilde{\lambda}_j$'s are the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$, computed as the squared singular values of $\boldsymbol{X}\mathbf{D}^{-1}$; in this dataset the rank is $r = 136$. Therefore a confidence interval for θ based on two standard deviations is (0.144, 2.201); notice that it does not include 0. To interpret the result in more natural units, by (19) (see Appendix C), an estimate of the average correlation is $\frac{1}{p}\sum_{j=1}^{p}\frac{\tilde{Z}_j}{\sqrt{n}} = \theta/\sqrt{n} = 0.099$; a confidence interval based on two standard deviations is (0.012, 0.187). This indicates that the average correlation between the outcome and the cortical thickness over the brain is somewhat small, but is still significantly different from zero.

While we get a significant result, the confidence interval for the average θ does not shrink with increasing p. In fact, the confidence interval stabilizes already for relatively small p. For example, taking a random sample of voxels of size $p = 2000$ gives a point estimate of 1.185 and a confidence interval of (0.154,



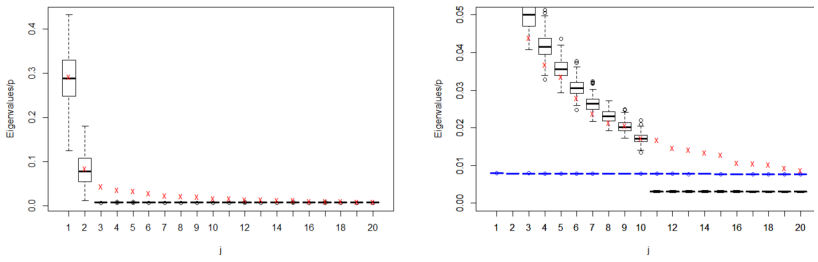
(a) Eigenvalues (b) $\hat{\tilde{d}}_j$

FIG 3. *(a) Plot of* $(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r)/p$. *(b) Plot of* $\hat{\tilde{d}}_j$ *for* $j = 1, \ldots, r$*; a confidence interval is given in dashed lines and the case* $j = 1$ *is colored in red.*

2.216), not far from the results for the full dataset with $p = 81924$. The reason for this, as suggested by the theory, is correlation between the voxels. To assess this, the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ are plotted in Figure 3(a). The first two eigenvalues are much larger than the rest, and it can be checked that in the variance decomposition $\sum_{j=1}^{r} \alpha_j^2 \tilde{\lambda}_j$, the first term captures 99.2% of the variance. This heterogeneity among the eigenvalues is caused by strong correlation; if the voxels were independent, then the eigenvalues would be much more homogeneous.

As a point of reference, if the entries of the matrix $\mathbf{X}$ were i.i.d. $N(0, 1)$, then the range of eigenvalues would be that of the Marchenko-Pastur distribution (Paul and Aue, 2014). For $n = 139$ and $p = 81924$, the Marchenko-Pastur range is $(1 \pm \sqrt{p/n})^2 = (541.8, 638.9)$. In contrast, in our data the range of the eigenvalues is $(\tilde{\lambda}_1, \tilde{\lambda}_p) = (33.4, 23995.7)$; see also Figure 4 (b). In addition, we compared the eigenvalues of $\tilde{\boldsymbol{\Sigma}}$ to a spiked covariance model (Johnstone, 2001), where each row in $\mathbf{X}$ is sampled from independent mean zero normals with variances $(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_K, c, \ldots, c)$ for $c = \frac{p - \sum_{j=1}^{K} \tilde{\lambda}_j}{p - K}$ (so the sum of variances is p). That is, the first K eigenvalues are the same as in $\boldsymbol{\Sigma}$ and the rest are equal. For a literature review on spiked models see Paul and Aue (2014). The parameter K is the number of large eigenvalues of the model. Figure 4 compares the eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{20}$ to the empirical eigenvalues of 200 simulations of the spiked model with $K = 2$ and $K = 10$ and also to 200 simulations of i.i.d. $N(0, 1)$. Although the first K eigenvalues fit, more or less, the empirical ones, the gap between $\tilde{\lambda}_K$ and $\tilde{\lambda}_{K+1}$ is much smaller than in the spiked model. Therefore, while the spiked model compares better to data than the Marchenko-Pastur distribution, there is still non-negligible difference between model's prediction and the observed eigenvalues of $\mathbf{X}$. It looks like that the eigenvalues of the real covariance matrix decay slower than the variances in the spiked model.

Linear combinations of $\tilde{\boldsymbol{\mu}}$ other than the average are also of interest. Since $\tilde{\boldsymbol{\mu}}$ is by definition spanned by the columns $\mathbf{X}$, consider the representation of



(a) $K = 2$ (b) $K = 10$

FIG 4. *Boxplots of 200 simulations of the first 20 eigenvalues of* $\mathbf{X}$ *under the spiked model with* $K = 2$ *(a) and* $K = 10$ *(b). The eigenvalues of the observed* $\mathbf{X}$*, i.e.,* $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{20}$*, are denoted by red 'x'. In Figure (b) boxplots of 200 simulations of i.i.d.* $N(0,1)$ *are plotted in blue. Note that the distribution of the eigenvalues, where it appears constant, it actually decays but very slowly.*

$\tilde{\boldsymbol{\mu}}$ with respect to the eigenvector space, i.e., we write $\tilde{\boldsymbol{\mu}} = \sum_{j=1}^{r} \tilde{\boldsymbol{\gamma}}_j \tilde{d}_j$, where $\tilde{d}_j = \tilde{\boldsymbol{\gamma}}_j^T \tilde{\boldsymbol{\mu}}/p$ are the coefficients of $\tilde{\boldsymbol{\mu}}$ with respect to the eigenvectors $\tilde{\boldsymbol{\gamma}}_1, \ldots, \tilde{\boldsymbol{\gamma}}_{136}$ of $\tilde{\boldsymbol{\Sigma}}$. Figure 3(b) plots the estimates of $\tilde{d}_j$ and the associated confidence intervals. We find that $\hat{\tilde{d}}_1 = 1.268$ and a confidence interval based on two standard deviations is (0.186, 2.351). Note that these values are similar to the estimate and confidence interval for $\theta = \frac{1}{p}\sum_{j=1}^{p} \tilde{\mu}_j$ calculated above. This is not a coincidence; the first eigenvector is very close to the averaging constant vector $\tilde{\mathbf{a}} = \mathbf{1}/p$ with a correlation of 0.938 between the two. For illustration, Figure 5 shows a brain map of the first eigenvector, which is indeed almost constant over the brain.

Back in Figure 3(b), the rest of the $\hat{\tilde{d}}_j$'s for $j > 1$ are much closer to 0. This suggests that most of the signal is contained in the subspace spanned by the first eigenvector $\tilde{\boldsymbol{\gamma}}_1$. However, the variance of $\hat{d}_1$ is also higher than the other $\hat{\tilde{d}}_j$'s. The variance of $\hat{\tilde{d}}_j$ is $\tilde{\boldsymbol{\gamma}}_j^T \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\gamma}}_j = \tilde{\lambda}_j/p$; indeed, the first eigenvalue is much larger than the other eigenvalues as illustrated in Figure 3(a). If we consider $\tilde{d}_j$ for large j, then the signal seems to be weaker, but also the variance is smaller. As indicated in Proposition 5 in Appendix A, when $\tilde{\mathbf{a}}$ is orthogonal to the "large" eigenvectors, the variance is small.

Not only the average signal is significantly different from zero, but one can test the global null hypothesis $\tilde{\boldsymbol{\mu}} = \mathbf{0}$ using the linear projections $\hat{d}_j$'s. Let $T_j = \hat{\tilde{d}}_j / sd(\hat{\tilde{d}}_j) = \tilde{\boldsymbol{\gamma}}_j^T \tilde{\mathbf{Z}}$, then $T_1, \ldots, T_{136}$ are i.i.d $N(0,1)$, under the null (assuming normality). Therefore, under the null, $\max_{j=1,\ldots,136} |T_j|$ is distributed as the maximum absolute value of 136 i.i.d normals. In this case, $\max_{j=1,\ldots,136} |T_j| = 3.3$, which yields a p-value of 0.064 (computed by simulation).

To sum up, our findings indicate that most of the signal $\tilde{\boldsymbol{\mu}}$ is in the direction of the first eigenvector. However, since in this direction the variance is also higher, it is difficult to determine the level of the signal in this direction, although the

confidence interval does not cover zero. This is consistent with the theory given in Appendix A, where it is shown that if $\tilde{\mathbf{a}}$ is not orthogonal to the eigenvectors corresponding to large eigenvalues, then the variance of $\hat{\theta}$ does not shrink to zero.

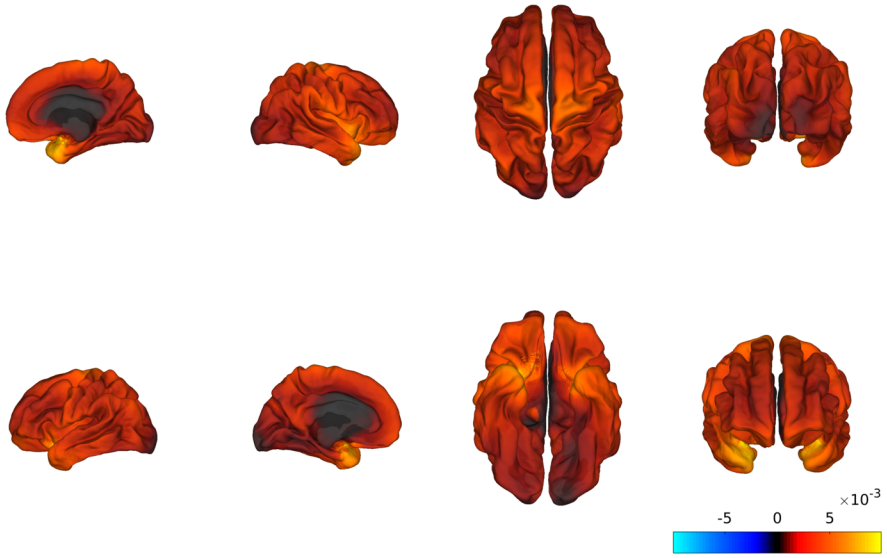


FIG 5. *Brain maps of the values of the first eigenvector.*

5.4. Estimation of $\mathbf{a}^T\boldsymbol{\beta}$

We now consider estimation of $\theta = \mathbf{a}^T\boldsymbol{\beta}$. As in Proposition 4 we study two cases: global average and single entry. Starting with the former, ideally, we wish to estimate the global average $\frac{1}{p}\sum_{j=1}^{p}\beta_j = \mathbf{1}^T\boldsymbol{\beta}/p$. Unfortunately, however, the vector $\mathbf{1}$ is not spanned by the columns of $\boldsymbol{\Gamma}$ and as a consequence the average is not identifiable. We can see this in terms of Proposition 2. Taking the expansion $\mathbf{1} = \boldsymbol{\Gamma}\boldsymbol{\alpha} + \boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp$, we have that

$$\frac{\|\boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp\|^2}{\|\mathbf{1}\|^2} = \frac{\|\mathbf{1} - \boldsymbol{\Gamma}\boldsymbol{\alpha}\|^2}{\|\mathbf{1}\|^2} = 0.0178.$$

The orthogonal component $\boldsymbol{\Gamma}\boldsymbol{\alpha}_\perp$ is not zero, as required by Proposition 2 for identifiability. However, its norm is small relative to the vector $\mathbf{1}$.

As a result, we consider instead the identifiable portion, determined by the vector $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\mathbf{1}/\sqrt{n}$, and estimate $\theta = \mathbf{a}^T\boldsymbol{\beta}$. The vector $\mathbf{a}$ is the closest vector (in L_2 sense) to $\mathbf{1}/\sqrt{n}$ spanned by the columns of $\boldsymbol{\Gamma}$, with a correlation with it of 0.991, so again, the loss in estimation is small. The normalizing constant gives $||\mathbf{a}||^2 = p/n$ and makes it consistent with the normalization of Proposition 4.

For such $\mathbf{a}$ we obtain $\hat{\theta} = \mathbf{a}^T \left(\mathbf{X}^T\mathbf{X}\right)^- \mathbf{X}^T\mathbf{Y} = 1.685$. The variance of the estimator is $\sum_{j=1}^r \tilde{\alpha}_j \lambda_j = 1.627$. Therefore, a confidence interval based on two standard deviations is $(-0.866, 4.236)$, which contains 0. In contrast, the confidence interval of the estimate of $\frac{1}{p}\sum_{j=1}^p \mu_j$ did not contain 0. In this dataset, estimating the average β is harder than estimating the average μ. According to Proposition 4, $Var(\hat{\theta})$ is of the same order as $\frac{p}{n^3}\sum_{j=1}^r \frac{1}{\lambda_j}$. In this dataset the latter expression is 0.47, which is not close to zero.

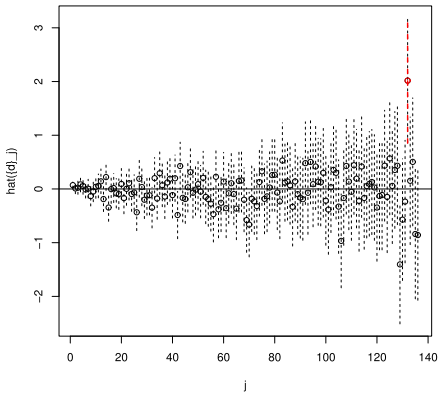


FIG 6. *Plot of $\hat{d}_j$ for $j = 1, \ldots, r$; a confidence interval is given in dashed lines; $j = 132$ is colored in red.*

As in Section 5.3, we are also interested in estimating $d_j = \boldsymbol{\gamma}_j^T\boldsymbol{\beta}$ (which is identifiable). This is the single entry case of Proposition 4. The estimates and confidence intervals are presented in Figure 6. Here, unlike $\tilde{d}_j$ above, the variance of d_j is large for large j, since the eigenvalues appear in the denominator of the expression of the variance $p/(n\lambda_j)$. Also unlike above, the signal does not seem to be very strong on the first eigenvectors. For some j's, however, the estimator of d_j is significantly different from zero; for example, when $j = 132$, $\hat{d}_j/sd(\hat{d}_j) = 3.47$. Here the test for the maximum of $\hat{d}_j/sd(\hat{d}_j)$ yields p-value of 0.035. For illustration, Figure 7 shows a brain map of the eigenvector $\boldsymbol{\gamma}_{132}$. Note that, in contrast with the first eigenvector $\boldsymbol{\gamma}_1$ shown in Figure 5, the eigenvector $\boldsymbol{\gamma}_{132}$ is much more concentrated spatially. While $j = 132$ gives the strongest effect, the vector $\boldsymbol{\beta}$ seems to have smaller components in other eigenvectors as well.

5.5. *Summary of the data analysis and significance*

We have seen that the vector $\tilde{\boldsymbol{\mu}}$ is mostly related to the first eigenvector $\boldsymbol{\gamma}_1$, while the vector $\boldsymbol{\beta}$ is not. The relationship between these two vectors is expressed in (3). Writing that relation as $\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\Sigma}}\mathbf{D}\boldsymbol{\beta}/\sigma = n\mathbf{D}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\boldsymbol{\beta}/\sigma$, we see that the vector $\boldsymbol{\beta}$ is mapped to $\tilde{\boldsymbol{\mu}}$ through the eigenvalues of $\mathbf{X}$. In particular, in this dataset, the first eigenvalue λ_1 is large, causing $\tilde{\boldsymbol{\mu}}$ to be highly aligned with the first eigenvector $\boldsymbol{\gamma}_1$. The fact that the average $\tilde{\boldsymbol{\mu}}$ (or its projection onto the

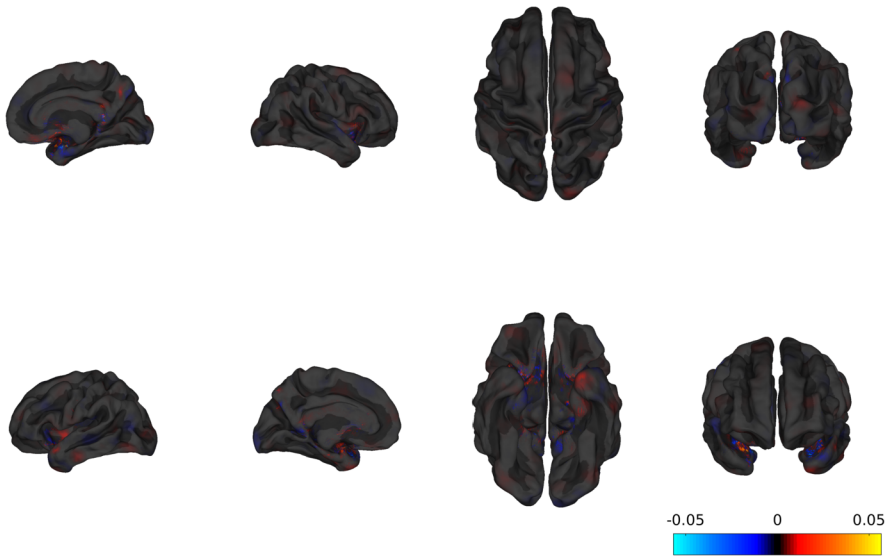


FIG 7. *Brain maps of the values of the* γ_{132}.

first eigenvector) appears to be non-zero, is indicative of the vector $\boldsymbol{\beta}$ being non-zero, with its effect amplified by the first eigenvalue. The vector $\boldsymbol{\beta}$ itself is hidden from the z-scores, yet some linear combinations of it can be estimated by the methods proposed here.

For the vector $\boldsymbol{\beta}$ we found that $\boldsymbol{\gamma}_j^T \hat{\boldsymbol{\beta}}$ is largest when $j = 132$, which is a relatively spatially concentrated eigenvector as presented in Figure 7. This finding might be used to identify areas in the cortical region that are correlated with the behavior assessed in the study.

6. Final remarks

The motivation for this study came from a dataset from brain imaging relating cortical thickness at each voxel to a global behavioral measurement. We aim at estimating one-dimensional linear projections of the coefficient vector $\boldsymbol{\beta}$ without assuming sparsity. In fact, the resulting signal $\tilde{\boldsymbol{\mu}}$ contained in the voxelwise z-scores does not seem to be sparse and traditional high-dimensional methods that aim at identifying the small number of non-zero parameters are not useful. The challenge is to distinguish between real signal and a seemingly high signal due to a correlation effect. Our theoretical results imply that for certain projections and correlation structures these two situations cannot be distinguished consistently.

For a given high-dimensional regression model, the general proposed approach is to estimate $\boldsymbol{\gamma}_j^T \boldsymbol{\beta}$ for all identifiable j's. The results can be used for two types of inference: a global test to check whether $\boldsymbol{\beta} = \mathbf{0}$ and identification of significant one-dimensional projections of $\boldsymbol{\beta}$.

Interestingly, regularization does not play a role in the estimation of linear projections. In the classical setting $\mathbf{Z} \sim N(\boldsymbol{\mu}, \mathbf{I})$, the famous work of Stein (1956) showed that $\mathbf{Z}$ is inadmissible for $\boldsymbol{\mu}$. A better estimate can be obtained by using shrinkage, or equivalently a certain kind of regularization. However, we have shown that $\tilde{\mathbf{a}}^T\mathbf{Z}$ is admissible and minimax for $\tilde{\mathbf{a}}^T\boldsymbol{\mu}$. Thus, when the interest is in the entire vector of $\boldsymbol{\mu}$, regularization is warranted, but for estimation of linear projections of $\boldsymbol{\mu}$ no regularization is required.

In this paper we investigated linear functions of $\boldsymbol{\beta}$ and necessary conditions for when a consistent estimator to $\theta = \mathbf{a}^T\boldsymbol{\beta}$ exists. Other non linear functions are also of interest. For example, in many applications it is desired to estimate the number of non-zero entries of $\boldsymbol{\beta}$ (Chen, 2018) or the squared norm of $\boldsymbol{\beta}$ (Dicker, 2014). We hope that a future study will suggest procedures to estimate such quantities and indicate when consistent estimates exists for non-linear functions of $\boldsymbol{\beta}$.

Appendix A: Consistency results for the natural estimator of Section 3

To state our asymptotic results, we need some more notation. Recall that $\lambda_1 \geq \ldots \geq \lambda_r$ are the positive eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_r$ are the corresponding eigenvectors. Define $\underline{\lambda}_i = \liminf_p \frac{\lambda_i}{p}$, and $\bar{\lambda}_i = \limsup_p \frac{\lambda_i}{p}$; since the variances are assumed bounded, then $\bar{\lambda}_i$ is finite for every i. Further, let

$$\underline{K} = \sum_{i=1}^{\infty} \underline{K}_i \qquad \text{with} \qquad \underline{K}_i = \begin{cases} 1 & \text{if } \underline{\lambda}_i > 0 \\ 0 & \text{if } \underline{\lambda}_i = 0 \end{cases}, \qquad i = 1, 2, \ldots,$$

and

$$\bar{K} = \sum_{i=1}^{\infty} \bar{K}_i \qquad \text{with} \qquad \bar{K}_i = \begin{cases} 1 & \text{if } \bar{\lambda}_i > 0 \\ 0 & \text{if } \bar{\lambda}_i = 0 \end{cases}, \qquad i = 1, 2, \ldots.$$

By definition we have that $\underline{K} \leq \bar{K}$. Notice that $\underline{K}$ could be infinity and that if $\underline{K} < \infty$ then $\underline{K}_i = 1$ for $i \leq \underline{K}$ and $\underline{K}_i = 0$ otherwise; the same for $\bar{K}$. Define $\alpha_i := \tilde{\mathbf{a}}^T\boldsymbol{\gamma}_i$, $i = 1, \ldots, p$; we have that $||\boldsymbol{\alpha}|| = ||\tilde{\mathbf{a}}||$. Let $\underline{\alpha}_i = \liminf_p \sqrt{p}|\alpha_i|$ and $\bar{\alpha}_i = \limsup_p \sqrt{p}|\alpha_i|$. The assumption that $||\tilde{\mathbf{a}}||^2 = O(1/p)$ implies that $\bar{\alpha}_i$ is finite since $p|\alpha_i|^2 \leq p||\boldsymbol{\alpha}||^2 = p||\tilde{\mathbf{a}}||^2 = O(1)$.

The variance of $\hat{\theta}$ is $\mathbf{b}^T\mathbf{b} = \tilde{\mathbf{a}}^T\boldsymbol{\Sigma}\tilde{\mathbf{a}}$. The following proposition provides results on the limiting behavior of $\tilde{\mathbf{a}}^T\boldsymbol{\Sigma}\tilde{\mathbf{a}}$ and of $\mathbf{b}$.

Proposition 5. *Assume that $\mathbf{Z}$ is a vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and consider the natural estimator $\hat{\theta} = \tilde{\mathbf{a}}^T\mathbf{Z}$. Suppose that $||\tilde{\mathbf{a}}||^2 = O(1/p)$ and the variances of $\mathbf{Z}$ are bounded.*

(I) If $\bar{K} > 0$ and $\bar{K} < \infty$, then

(a) We have that $\limsup_p \tilde{\mathbf{a}}^T\boldsymbol{\Sigma}\tilde{\mathbf{a}} = \sum_{i=1}^{\bar{K}} \bar{\lambda}_i\bar{\alpha}_i^2$ and $\liminf_p \tilde{\mathbf{a}}^T\boldsymbol{\Sigma}\tilde{\mathbf{a}} = \sum_{i=1}^{\underline{K}} \underline{\lambda}_i\underline{\alpha}_i^2$.

(b) For $i \le \underline{K}$, $\liminf_p b_i = \sqrt{\underline{\lambda}_i}\underline{\alpha}_i$ *and for* $i \le \bar{K}$, $\limsup_p b_i = \sqrt{\bar{\lambda}_i}\bar{\alpha}_i$.
(c) We have that $\liminf_p \sum_{i=\underline{K}+1}^{r} b_i^2 = 0$ *and* $\limsup_p \sum_{i=\bar{K}+1}^{r} b_i^2 = 0$.

(II) If $\underline{K} = \infty$ *and there exists* i_0 *such that* $\underline{\alpha}_{i_0} > 0$ *then*

$$\limsup_{p\to\infty} \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} \le \sum_{i=1}^{i_0} \bar{\lambda}_i \bar{\alpha}_i^2 + \bar{\lambda}_{i_0+1} p ||\tilde{\mathbf{a}}||^2 \text{ and } \liminf_{p\to\infty} \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} \ge \sum_{i=1}^{i_0} \underline{\lambda}_i \underline{\alpha}_i^2. \tag{8}$$

Part (I) shows that $\limsup_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}}$ can be written as a sum of $\bar{K}$ summands: $\sum_{i=1}^{\bar{K}} \bar{\lambda}_i \bar{\alpha}_i^2$. By definition, $\bar{\lambda}_i > 0$ for $i \in \{1, \ldots, \bar{K}\}$; therefore $\limsup_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}}$ goes to zero iff $\bar{\alpha}_i = 0$ for $i \in \{1, \ldots, \bar{K}\}$. In other words, Proposition 5 implies that $\hat{\theta}$ is consistent iff $\tilde{\mathbf{a}}$ is (asymptotically) orthogonal to the eigenvectors that correspond to the largest $\bar{K}$ eigenvalues. When $\underline{K} = \infty$, then $\tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}}$ is an infinite sum, but it could be bounded by a finite sum. The proof of Proposition 5 implies that if there are no large eigenvalues, i.e., if $\bar{K} = 0$, then consistency of $\hat{\theta}$ follows as stated in the following corollary.

Corollary 2. *If* $\bar{K} = 0$, $||\tilde{\mathbf{a}}||^2 = O(1/p)$ *and the variances of* $\mathbf{Z}$ *are bounded, then* $\limsup_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} = 0$ *and* $\hat{\theta}$ *is consistent.*

Proposition 5 implies that when $||\tilde{\mathbf{a}}||^2$ is of order larger than $O(1/p)$, then $\lim_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} = \infty$ and if $||\tilde{\mathbf{a}}||^2 = o(1/p)$ then $\lim_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} = 0$. Therefore, in the former case, $\hat{\theta}$ is inconsistent for all correlation structures and in the latter case it is always consistent as stated in the corollary below.

Corollary 3. *Assume that the variances of* $\mathbf{Z}$ *are bounded and* $\underline{K} > 0$.

(I) If $||\tilde{\mathbf{a}}||^2$ *is of order larger than* $O(1/p)$, *and there exists* $i \in \{1, \ldots, \underline{K}\}$ *(*$\underline{K}$ *can be infinity) such that* $\underline{\alpha}_i^2 > 0$ *then* $Var(\hat{\theta}) \to \infty$.
(II) If $||\tilde{\mathbf{a}}||^2 = o(1/p)$ *then* $Var(\hat{\theta}) \to 0$.

Parts (I)(b) and (I)(c) in Proposition 5 show that the vector $\mathbf{b}$ is sparse in the sense that only the first $\underline{K}$ or $\bar{K}$ entries are bounded away from zero, while the rest of the entries are close to zero. Proposition 5 implies that when $\bar{K} < \infty$

$$\limsup_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} = \limsup_p \sum_{i=1}^{\bar{K}} b_i^2 \text{ and } \liminf_p \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} = \liminf_p \sum_{i=1}^{\underline{K}} b_i^2.$$

When $\underline{K} = \infty$, $\tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}}$ can be bounded by finite sums as in (8). These finite sums approximations are used below in the proof of Theorem 2, as it reduces the p summands of $\tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}}$ to a bounded number of summands.

Theorem 2 (Inconsistency)**.** *Assume that* $\mathbf{Z}$ *is a vector with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$. *Suppose that* $||\tilde{\mathbf{a}}||^2 = O(1/p)$, *the variances of* $\mathbf{Z}$ *are bounded and there exists* $i_0 \in \{1, \ldots, \underline{K}\}$ *(*$\underline{K}$ *can be infinity) such that* $\underline{\alpha}_{i_0}^2 > 0$. *Then there exists no estimator* $\tilde{\theta}$ *that satisfies* $E(\tilde{\theta} - \theta)^2 \to 0$ *for all* $\boldsymbol{\eta}$. *(More precisely, there exists no sequence of estimators* $\{\tilde{\theta}_p(\mathbf{W}_p)\}_{p=1}^{\infty}$ *that satisfies for each*

sequence of parameters $\{\boldsymbol{\eta}_p\}_{p=1}^{\infty}$*, that* $E\{\tilde{\theta}_p(\mathbf{W}_p) - \theta_p(\boldsymbol{\eta}_p)\}^2 \to_{p\to\infty} 0$*, where* $\mathbf{W}_p \sim N_r(\boldsymbol{\eta}_p, I)$ *and* $\theta_p(\boldsymbol{\eta}_p) = \mathbf{b}_p^T \boldsymbol{\eta}_p$*.)*

Theorem 2 implies that when $\hat{\theta}$ is inconsistent as in Proposition 5, then so is every other estimator. The proof uses a result of Bickel (1981), who shows in the context of estimation of the mean of a single normal observation, Z, when the parameter space is bounded, that Z is "almost" minimax. In our context, this implies that $\hat{\theta}$ is approximately minimax in the bounded case (it is exactly minimax when the parameter space is unbounded). If a consistent estimator existed, then it would be uniformly close to zero on bounded sets, contradicting the approximated minimaxity of $\hat{\theta}$.

Appendix B: Exchangeable correlation

In this section we demonstrate the theory of Section 3 through a study of the specific example of equal correlation. We also compare the natural estimate to empirical Bayes estimator. Demonstrating the admissibility of the natural estimator, it is shown that for some θ's the empirical Bayes estimator is better and for other θ's the natural estimator is better, but there is no consistent way to distinguish between the cases.

B.1. Exchangeable correlation model

Suppose that $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has an exchangeable correlation structure, i.e., for $i \neq j$, $\Sigma_{ij} = Cov(Z_i, Z_j) = \rho$ for a constant $\rho > 0$ and $\Sigma_{ii} = Var(Z_i) = 1$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{pmatrix}. \tag{9}$$

In this case, $\mathbf{Z}$ can be constructed as

$$\mathbf{Z} = \boldsymbol{\mu} + \sqrt{\rho} V \mathbf{1} + \sqrt{1-\rho}\boldsymbol{\varepsilon}, \tag{10}$$

where $\boldsymbol{\varepsilon} := (\varepsilon_1, \dots, \varepsilon_p)^T$, $\mathbf{1} = (1, 1, \dots, 1)^T$ and $V, \varepsilon_1, \dots, \varepsilon_p$ are i.i.d $N(0,1)$. Suppose further that we want to estimate $\frac{1}{p}\sum_{i=1}^{p} \mu_i$ (i.e., $\tilde{\mathbf{a}} = \mathbf{1}/p$). The natural estimate is $\hat{\theta} = \tilde{\mathbf{a}}^T \boldsymbol{Z} = \frac{1}{p}\sum_{j=1}^{p} Z_j$. Multiplying (10) by $\tilde{\mathbf{a}}$ yields $\hat{\theta}$ on the right hand-side, and the left hand-side gives $\theta + \sqrt{\rho}V + \sqrt{1-\rho}\bar{\varepsilon}$ which has the irreducible variance ρ. On the other hand, if we choose $\tilde{\mathbf{a}}$ that is orthogonal to the $\mathbf{1}$ vector, then θ can be estimated consistently since the term $\sqrt{\rho}V$ disappears.

B.2. Consistent and inconsistent estimates under exchangeable correlation

Suppose that $\{Z_i\}_{i=1}^{\infty}$ is a sequence of exchangeable random variables with correlation $\rho \geq 0$, i.e., for each fixed p the covariance matrix $\boldsymbol{\Sigma}$ is (9). The eigenvalues

of $\boldsymbol{\Sigma}$ are $\lambda_1 = \rho p + 1 - \rho$, with corresponding eigenvector equal to $\boldsymbol{\gamma}_1 = \mathbf{1}/\sqrt{p}$, and $\lambda_2 = \cdots = \lambda_p = 1 - \rho$, hence, there is only one large eigenvalue, i.e., $\underline{K} = \bar{K} = 1$, and $\bar{\lambda}_1 = \lim_p \frac{\lambda_1}{p} = \lim_p \frac{\rho p + 1 - \rho}{p} = \rho = \underline{\lambda}_1$. Consider the following two cases for the vector $\tilde{\mathbf{a}}$.

Parallel case: If one is interested in estimating $\theta = \bar{\mu} = \frac{1}{p}\sum_{i=1}^p \mu_i$ (i.e., $\tilde{\mathbf{a}} = \mathbf{1}/p$), then $\alpha_1 = \tilde{\mathbf{a}}^T \boldsymbol{\gamma}_1 = 1/\sqrt{p}$ and $\bar{\alpha}_1 = \underline{\alpha}_1 = 1$. The natural estimator is $\hat{\theta} = \frac{1}{p}\sum_i Z_i$ and

$$\liminf_p Var\left(\frac{1}{p}\sum_{i=1}^p Z_i\right) = \underline{\lambda}_1 \underline{\alpha}_1 = \rho = \bar{\lambda}_1 \bar{\alpha}_1 = \limsup_p Var\left(\frac{1}{p}\sum_{i=1}^p Z_i\right),$$

as in the analysis of the previous subsection. That is, the variance of $\hat{\theta}$ converges to $\rho > 0$ and therefore, by Theorem 2, no consistent estimator exists for $\bar{\mu}$ when the correlation structure is (9).

Considering the representation of θ as $\mathbf{b}^T \boldsymbol{\eta}$ then,

$$b_1 = \lambda_1^{1/2} \boldsymbol{\gamma}_1^T \tilde{\mathbf{a}} = \sqrt{\rho p + 1 - \rho}\frac{1}{\sqrt{p}}\mathbf{1}^T\mathbf{1}/p = \frac{\sqrt{\rho p + 1 - \rho}}{\sqrt{p}} \approx \sqrt{\rho}, \tag{11}$$

and, $b_i = 0$ for $i \geq 2$, since $\boldsymbol{\gamma}_i^T \tilde{\mathbf{a}} = 0$ for $i \geq 2$. Also,

$$Z_1 = \lambda_1^{-1/2} \boldsymbol{\gamma}_1^T \mathbf{Z} = \frac{\frac{1}{\sqrt{p}}\sum_{i=1}^p Z_i}{\sqrt{\rho p + 1 - \rho}} = \frac{1}{\sqrt{\rho}}\bar{Y}\left\{1 + O\left(\frac{1}{\sqrt{p}}\right)\right\} \text{ and similarly}$$
$$\eta_1 = \frac{1}{\sqrt{\rho}}\bar{\mu}\left\{1 + O\left(\frac{1}{\sqrt{p}}\right)\right\};$$

thus, $\hat{\theta} = b_1 Z_1 = \bar{Y}$ and

$$\hat{\theta} \sim N(\theta, b_1^2) \tag{12}$$

with b_1 given by (11).

Orthogonal case: On the other hand, if one is interested in estimating $\sum_{i=1}^p a_i \mu_i$ with $\sum_{i=1}^p a_i = 0$ (and $||\tilde{\mathbf{a}}||^2 = 1/p$) then $\tilde{\mathbf{a}}$ is orthogonal to the leading eigenvector $\boldsymbol{\gamma}_1$, i.e., $\alpha_1 = \tilde{\mathbf{a}}^T \boldsymbol{\gamma}_1 = 0$. Therefore,

$$Var(\hat{\theta}) = Var\left(\sum_{i=1}^p a_i Z_i\right) = \sum_{i=1}^p \lambda_i \alpha_i^2 = \sum_{i=2}^p \lambda_i \alpha_i^2 = (1-\rho)\sum_{i=2}^p \alpha_i^2$$
$$= (1-\rho)||\tilde{\mathbf{a}}||^2 = \frac{1-\rho}{p},$$

i.e., for this $\tilde{\mathbf{a}}$, $\hat{\theta}$ is $\sqrt{p}$-consistent.

B.3. Exchangeable correlation – two blocks

Suppose now that $\boldsymbol{\Sigma}$ consists of two blocks of the form (9), the first one of size $p_1 \times p_1$ and the second of size $p_2 \times p_2$ with $p_1 + p_2 = p$. We assume that there is a

constant correlation between the two blocks ρ_B (which can be zero or negative). Thus,

$$\mathbf{\Sigma} = \begin{pmatrix} \underbrace{\begin{matrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & \ddots & \rho_1 \\ \rho_1 & \rho_1 & 1 \end{matrix}}_{p_1} & \rho_B \\ \rho_B & \underbrace{\begin{matrix} 1 & \rho_2 & \rho_2 \\ \rho_2 & \ddots & \rho_2 \\ \rho_2 & \rho_2 & 1 \end{matrix}}_{p_2} \end{pmatrix}.$$

The matrix $\mathbf{\Sigma}$ has two eigenvalues

$$\lambda_{1,2} = \frac{\xi_1 + \xi_2}{2} \pm \left\{ \left(\frac{\xi_1 - \xi_2}{2} \right)^2 + p_1 p_2 \rho_B^2 \right\}^{1/2}$$

where $\xi_j = 1 + \rho_j\{p_j - 1\}$, $j = 1, 2$, and the other $p - 2$ eigenvalues are either $1 - \rho_1$ or $1 - \rho_2$. For large p, positive semi-definiteness requires that $\rho_B^2 \leq \rho_1\rho_2$. On the boundary, when $\rho_B^2 = \rho_1\rho_2$, λ_2 converges to a positive constant and $\underline{K} = \bar{K} = 1$.

We now concentrate on the case $\rho_B^2 < \rho_1\rho_2$. Then,

$$\begin{aligned} \underline{\lambda}_{1,2} &= \bar{\lambda}_{1,2} \\ &= \lim_p \frac{\lambda_{1,2}}{p} = \frac{\pi\rho_1 + (1-\pi)\rho_2}{2} \pm \left\{ \left(\frac{\pi\rho_1 - (1-\pi)\rho_2}{2} \right)^2 + \pi(1-\pi)\rho_B^2 \right\}^{1/2}, \end{aligned} \tag{13}$$

i.e., $\lambda_{1,2}$ are of order p and $\underline{K} = \bar{K} = 2$. The corresponding eigenvectors are

$$(\underbrace{x_j, \ldots, x_j}_{m_1 \text{ times}}, \underbrace{y_j, \ldots, y_j}_{m_2 \text{ times}})^T, \quad j = 1, 2,$$

where when $\rho_B \neq 0$, x_j and y_j are given by

$$\begin{aligned} x_1 &= \frac{-\sqrt{p_2}\rho_B}{\sqrt{(\xi_1 - \lambda_1)^2 + p_1 p_2 \rho_B^2}}, \qquad y_1 = \frac{(\xi_1 - \lambda_1)/\sqrt{p_2}}{\sqrt{(\xi_1 - \lambda_1)^2 + p_1 p_2 \rho_B^2}}, \\ x_2 &= \frac{(\xi_2 - \lambda_2)/\sqrt{p_1}}{\sqrt{(\xi_2 - \lambda_2)^2 + p_1 p_2 \rho_B^2}}, \qquad y_2 = \frac{-\sqrt{p_1}\rho_B}{\sqrt{(\xi_2 - \lambda_2)^2 + p_1 p_2 \rho_B^2}}. \end{aligned}$$

We have thus,

$$\begin{aligned} \lim_p \sqrt{p} x_1 &= \frac{-\sqrt{1-\pi}\rho_B}{\sqrt{(\pi\rho_1 - \bar{\lambda}_1)^2 + \pi(1-\pi)\rho_B^2}}, \\ \lim_p \sqrt{p} y_1 &= \frac{(\pi\rho_1 - \bar{\lambda}_1)/\sqrt{1-\pi}}{\sqrt{(\pi\rho_1 - \bar{\lambda}_1)^2 + \pi(1-\pi)\rho_B^2}}, \end{aligned}$$

$$\lim_p \sqrt{p}x_2 = \frac{\{(1-\pi)\rho_2 - \bar{\lambda}_2\}/\sqrt{\pi}}{\sqrt{\{(1-\pi)\rho_2 - \bar{\lambda}_2\}^2 + \pi(1-\pi)\rho_B^2}},$$
$$\lim_p \sqrt{p}y_2 = \frac{-\sqrt{\pi}\rho_B}{\sqrt{\{(1-\pi)\rho_2 - \bar{\lambda}_2\}^2 + \pi(1-\pi)\rho_B^2}}. \tag{14}$$

If one is interested in estimating $\theta = \frac{1}{p}\sum_{i=1}^{p}\mu_i$ (i.e., $\tilde{\mathbf{a}} = (1/p, \ldots, 1/p)$), then $\alpha_1 = \frac{p_1x_1+p_2y_1}{p}$, $\alpha_2 = \frac{p_1x_2+p_2y_2}{p}$ and the limit is

$$\underline{\alpha}_1 = \bar{\alpha}_1 = \pi \lim_p \sqrt{p}x_1 + (1-\pi)\lim_p \sqrt{p}y_1 \text{ and}$$
$$\underline{\alpha}_2 = \bar{\alpha}_2 = \pi \lim_p \sqrt{p}x_2 + (1-\pi)\lim_p \sqrt{p}y_2.$$

Therefore, $\bar{\alpha}_1 = 0$ and $\bar{\alpha}_2 = 0$, implies that

$$\bar{\lambda}_1 - \pi\rho_1 + \pi\rho_B = 0 \text{ and, } \bar{\lambda}_2 - (1-\pi)\rho_2 + (1-\pi)\rho_B = 0. \tag{15}$$

However, Proposition 6 below implies that there exists no set of parameters $(\rho_1, \rho_2, \rho_B, \pi)$ for which (15) is satisfied, and therefore $\bar{\mu}$ cannot be consistently estimated for this correlation structure.

Suppose now that the interest is in estimating the difference between the means of each block, i.e., $\theta = \frac{1}{p_1}\sum_{i=1}^{p_1}\mu_i - \frac{1}{p_2}\sum_{i=1}^{p_2}\mu_i$. In this case, $\alpha_1 = x_1$ and $\alpha_2 = y_2$ and the limit is given by (14). Therefore, $\underline{\alpha}_1 = \underline{\alpha}_2 = 0$ iff $\rho_B = 0$ and in this case θ can be consistently estimated iff $\rho_B = 0$, i.e., the blocks are independent.

Proposition 6. *There exists no set of parameters* $(\rho_1, \rho_2, \rho_B, \pi)$ *for* $\rho_B \neq 0$ *that satisfies* (15).

B.4. Empirical Bayes estimator under exchangeable correlation

Empirical Bayes (EB) estimators are found useful in many high-dimensional situations (Efron, 2010) and therefore are potential candidates to improve the natural estimator $\hat{\theta}$. However, Theorem 1 implies that EB estimates cannot uniformly improve $\hat{\theta}$. Furthermore, it implies that there is no consistent way to identify the cases where EB estimates are better. In this section we compare the EB estimator to $\hat{\theta}$ under the exchangeable correlation structure (9).

To define the EB estimator, we start with the Bayesian estimator of $\theta = \mathbf{b}^T\boldsymbol{\eta}$, where $\mathbf{W} \sim N(\boldsymbol{\eta}, \mathbf{I})$. If $\eta_1, \ldots, \eta_p$ were i.i.d $N(\alpha, \tau^2)$ then the Bayes estimator of θ would be the posterior expectation of θ, which is equal to $\mathbf{b}^T\hat{\boldsymbol{\eta}}^B$, where $\hat{\boldsymbol{\eta}}^B = (\hat{\eta}_1^B, \ldots, \hat{\eta}_p^B)$ for $\hat{\eta}_i^B = \frac{\alpha}{\tau^2+1} + \frac{\tau^2 W_i}{\tau^2+1}$. The EB estimator is $\hat{\theta}^{EB} = \mathbf{b}^T\hat{\boldsymbol{\eta}}^{EB}$, where $\hat{\eta}_i^{EB} = \frac{\hat{\alpha}}{\hat{\tau}^2+1} + \frac{\hat{\tau}^2 W_i}{\hat{\tau}^2+1}$, with $\hat{\alpha} = \bar{W}$ and $\hat{\tau}^2 = \max\big(0, \frac{1}{p}\sum_{i=1}^{p}(W_i - \bar{W})^2 - 1\big)$.

The mean square error

$$E(\hat{\theta}^{EB} - \theta)^2 = E\left[\mathbf{b}^T(\hat{\boldsymbol{\eta}}^{EB} - \boldsymbol{\eta})\right]^2 \tag{16}$$

may be evaluated as follows. Without any assumptions on the structure of $\boldsymbol{\eta}$, we may define $\alpha = \bar{\eta}$ and $\tau^2 = \frac{1}{p}\sum_{i=1}^{p}(\eta_i - \bar{\eta})^2$, and we have that $\hat{\alpha} - \alpha = O_p(1/\sqrt{p})$ and $\hat{\tau}^2 - \tau^2 = O_p(1/\sqrt{p})$. Hence, to evaluate (16) we may use the approximation

$$\begin{aligned}\hat{\eta}_i^{EB} - \eta_i = \frac{1}{\hat{\tau}^2+1}\bar{Z} + \frac{\hat{\tau}^2}{\hat{\tau}^2+1}W_i - \eta_i &= \frac{1}{\tau^2+1}\bar{\eta} + \frac{\tau^2}{\tau^2+1}W_i - \eta_i + O_p(1/\sqrt{p}) \\ &= \frac{1}{\tau^2+1}\left\{\tau^2(W_i - \eta_i) + \bar{\eta} - \eta_i\right\} + O_p(1/\sqrt{p}). \quad (17)\end{aligned}$$

Suppose now the correlation structure (9) and the parallel case $\theta = \bar{\mu}$. Then, by (11), $b_1 = \sqrt{\rho}\left\{1 + \frac{(1-\rho)/\rho}{p}\right\}^{1/2}$ and $b_2 = \cdots = b_p = 0$. Therefore, in this case $\hat{\theta}^{EB} = b_1\hat{\eta}_1^{EB}$. When $\hat{\alpha} - \alpha$ and $\hat{\tau} - \tau$ are uniformly integrable, (17) implies that the mean squared error (16) is

$$E(\hat{\theta}^{EB} - \theta)^2 = \frac{b_1^2\{\tau^4 + (\eta_1 - \bar{\eta})^2\}}{(\tau^2+1)^2} + O(1/\sqrt{p}).$$

From (12), the mean squared error of $\hat{\theta}$ is b_1^2. Therefore, $\hat{\theta}^{EB}$ is better than $\hat{\theta}$ when

$$\tau^4 + (\eta_1 - \bar{\eta})^2 < (\tau^2+1)^2 \Longleftrightarrow (\eta_1 - \bar{\eta})^2 < 2\tau^2 + 1. \quad (18)$$

In practice, one cannot verify condition (18), since there is no consistent estimate for η_1 and $\bar{\eta}$. In other words, there is no consistent way to know when $\hat{\theta}^{EB}$ is better. On one extreme, if $\eta_1 = \cdots = \eta_p$, then the left hand-side of (18) converges to 0, while the right hand-side converges to 1, so the risk of $\hat{\theta}^{EB}$ converges to 0. On the other hand, if $(\eta_1 - \bar{\eta})^2$ is large, i.e., when η_1 is distant from the other η's, then $\hat{\theta}$ is better.

To illustrate this point we simulated 1000 times two scenarios with $p = 1000$, $\theta = \bar{\mu} = 5$ and the correlation structure (9) with $\rho = 0.6$. In the first scenario we chose $\boldsymbol{\mu}$ such that $\boldsymbol{\eta}$ is constant (using the relation $\boldsymbol{\eta} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}^T\boldsymbol{\mu}$) and $\bar{\mu} = 5$, while in the second scenario $\boldsymbol{\mu} = (4, 6, 4, 6, \ldots, 4, 6)$. Under the first scenario, $\tau^2 = (\eta_1 - \bar{\eta})^2 = 0$ and therefore (18) is satisfied. For the second scenario, $(\eta_1 - \bar{\eta})^2 = 41.2$ and $\tau^2 = 6.5$; hence, (18) is violated. Table 2 shows the simulation results. Indeed, $\hat{\theta}^{EB}$ has smaller risk than $\hat{\theta}$ for the first scenario but not for the second scenario.

TABLE 2
Simulation results of $E(\hat{\theta} - \theta)^2$ and $E(\hat{\theta}^{EB} - \theta)^2$. Confidence intervals based on two standard deviations are given in parentheses.

Scenario	Estimator	Mean square error (MSE)	Asymptotic MSE
Constant $\boldsymbol{\eta}$	$\hat{\theta}$	0.60 (0.55, 0.66)	0.6
Constant $\boldsymbol{\eta}$	$\hat{\theta}^{EB}$	$1.12 \times 10^{-3}(0.96 \times 10^{-3}, 1.28 \times 10^{-3})$	0
$\boldsymbol{\mu} = (4, 6, 4, 6, \ldots, 4, 6)$	$\hat{\theta}$	0.57 (0.51, 0.62)	0.6
$\boldsymbol{\mu} = (4, 6, 4, 6, \ldots, 4, 6)$	$\hat{\theta}^{EB}$	2.26 (2.16, 2.36)	2.28

B.5. Empirical covariance matrix under exchangeable correlation

Recall the regression setting of Section 2 and suppose that each row of the matrix $\mathbf{X}$ is sampled independently from a distribution with covariance matrix (9). The results of Fan and Wang (2017) indicate that the leading eigenvalue of the sample covariance is of order p, as of the original distribution, but it is biased upward. In terms of our notation, the sample covariance satisfies $\bar{K} = 1$, but λ_1 is greater than the leading eigenvalue $\rho p + 1 - \rho$ of the true underlying distribution and still of order p. This is the case considered in the Simulations Section 4.4.

Appendix C: Vertex-wise correlations

Closely related to the z-scores $\tilde{\mathbf{Z}}$ is the vector of voxelwise correlations $\hat{\boldsymbol{\rho}} = \sigma\tilde{\mathbf{Z}}/[\sqrt{n}\widehat{Sd}(Y)]$ with entries

$$\frac{\sigma}{\sqrt{n}\ \widehat{Sd}(Y)}\tilde{Z}_j = \frac{\sum_{i=1}^n X_i^{(j)} Y_i/n}{\sqrt{s_{jj}/n}\ \widehat{Sd}(Y)}, \tag{19}$$

equal to the observed correlation between the row vector $(X_1^{(j)}, \ldots, X_n^{(j)})$ and $\mathbf{Y}$. Ignoring the error in the estimation of the standard deviation of Y for large n, we have that $\hat{\boldsymbol{\rho}}$ is approximately normal with mean $\sigma\tilde{\boldsymbol{\mu}}/[\sqrt{n}\widehat{Sd}(Y)]$. We may estimate linear projections of this vector, such as the average voxelwise correlation.

Appendix D: Proofs

D.1. Proof of Proposition 1

(I). Write $\mathbf{Z}$ in the coordinate system of the eigenvectors

$$\mathbf{Z} = \boldsymbol{\Gamma}\mathbf{Z}_{\boldsymbol{\Gamma}} + \boldsymbol{\Gamma}_\perp \mathbf{Z}_{\boldsymbol{\Gamma}_\perp}.$$

The variance of the second part is zero since $\mathbf{Z}_{\boldsymbol{\Gamma}_\perp} = \boldsymbol{\Gamma}_\perp^T\mathbf{Z}$ and

$$Var(\boldsymbol{\Gamma}_\perp^T\mathbf{Z}) = \boldsymbol{\Gamma}_\perp^T Cov(\mathbf{Z})\boldsymbol{\Gamma}_\perp = \boldsymbol{\Gamma}_\perp^T\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_\perp = 0.$$

Therefore, $E(\mathbf{Z}_{\boldsymbol{\Gamma}_\perp}) = \boldsymbol{\Gamma}_\perp^T\boldsymbol{\mu}$ is a known vector.

(II). Part I implies that $\boldsymbol{\mu}$ lies within the r-dimensional space spanned by the columns of $\boldsymbol{\Gamma}$. Thus, if we write $\tilde{\mathbf{a}}$ in the coordinate system of the eigenvectors $\tilde{\mathbf{a}} = \boldsymbol{\Gamma}\tilde{\mathbf{a}}_{\boldsymbol{\Gamma}} + \boldsymbol{\Gamma}_\perp\tilde{\mathbf{a}}_{\boldsymbol{\Gamma}_\perp}$, then

$$\theta = \tilde{\mathbf{a}}^T\boldsymbol{\mu} = \{\boldsymbol{\Gamma}\tilde{\mathbf{a}}_{\boldsymbol{\Gamma}} + \boldsymbol{\Gamma}_\perp\tilde{\mathbf{a}}_{\boldsymbol{\Gamma}_\perp}\}^T\boldsymbol{\mu} = \tilde{\mathbf{a}}_{\boldsymbol{\Gamma}}^T\boldsymbol{\Gamma}^T\boldsymbol{\mu}.$$

Therefore, we can write in this coordinate system, $\theta = \mathbf{b}^T\boldsymbol{\eta}$, where $\mathbf{b} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^T\tilde{\mathbf{a}}$ since

$$\mathbf{b}^T\boldsymbol{\eta} = \tilde{\mathbf{a}}^T\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}^T\boldsymbol{\mu} = \tilde{\mathbf{a}}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\mu} = \tilde{\mathbf{a}}_{\boldsymbol{\Gamma}}^T\boldsymbol{\Gamma}^T\boldsymbol{\mu} = \theta.$$

(III). Since $\mathbf{Z} = \boldsymbol{\Gamma}\mathbf{Z}_{\boldsymbol{\Gamma}}$, then $\mathbf{W} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Gamma}^T\mathbf{Z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{Z}_{\boldsymbol{\Gamma}}$ implies that $\mathbf{Z}_{\boldsymbol{\Gamma}} = \boldsymbol{\Lambda}^{1/2}\mathbf{W}$. We conclude that $\mathbf{Z} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\mathbf{W}$ since $\mathbf{Z} = \boldsymbol{\Gamma}\mathbf{Z}_{\boldsymbol{\Gamma}}$. □

D.2. Proof of Theorem 1

The proof follow similar ideas as in Proposition 10.4.2–10.4.4 in Bickel and Doksum (1977) that show that $\bar{Z}$ is minimax and admissible for estimation of a standard normal mean.

Part (I) Let $\pi_k(\boldsymbol{\eta})$ be the density of the prior $\eta_1, \ldots, \eta_r \sim^{i.i.d} N(0, k)$. The posterior distribution of η_i given $\mathbf{W} = \mathbf{w}$ is $N\left(\frac{w_i k}{k+1}, \frac{k}{k+1}\right)$. The Bayes estimator is $\mathbf{b}^T\hat{\boldsymbol{\eta}}^{Bayes}$ (the dependence on k is suppressed in the notation), where $\hat{\boldsymbol{\eta}}^{Bayes} := (\hat{\eta}_1^{Bayes}, \ldots, \hat{\eta}_p^{Bayes})$ for $\hat{\eta}_i^{Bayes} := \frac{W_i k}{k+1}$. The Bayes risk is

$$r_k := \tilde{E}(\mathbf{b}^T\hat{\boldsymbol{\eta}}^{Bayes} - \theta)^2 = \int \tilde{E}\left\{\mathbf{b}^T(\hat{\boldsymbol{\eta}}^{Bayes} - \boldsymbol{\eta})|\mathbf{W} = \mathbf{w}\right\}^2 f(\mathbf{w})d\mathbf{w}$$
$$= \frac{k}{k+1}\mathbf{b}^T\mathbf{b},$$

where $\tilde{E}$ denotes expectation with respect to the joint density of θ and $\mathbf{Z}$ according to induced probability measure of the Bayesian framework and $f(\mathbf{W})$ is the marginal density of $\mathbf{Z}$ under this probability measure. The Bayes risk of the natural estimator is

$$r_k(\hat{\theta}) = \int_{\boldsymbol{\eta}} E(\hat{\theta} - \theta)^2\pi_k(\boldsymbol{\eta})d\boldsymbol{\eta} = \int_{\boldsymbol{\eta}} \mathbf{b}^T\mathbf{b}\pi_k(\boldsymbol{\eta})d\boldsymbol{\eta} = \mathbf{b}^T\mathbf{b}.$$

Hence, $r_k = r_k(\hat{\theta}) - \frac{\mathbf{b}^T\mathbf{b}}{k+1}$. Since $\hat{\theta}$ is constant risk, this implies that for any estimator $\tilde{\theta}$,

$$\sup_{\boldsymbol{\eta}\in\mathbb{R}^r} E(\tilde{\theta} - \theta)^2 \geq \int_{\boldsymbol{\eta}} E(\tilde{\theta} - \theta)^2\pi_k(\boldsymbol{\eta})d\boldsymbol{\eta} \geq r_k$$
$$= r_k(\hat{\theta}) - \frac{\mathbf{b}^T\mathbf{b}}{k+1} = \sup_{\boldsymbol{\eta}\in\mathbb{R}^r} E(\hat{\theta} - \theta)^2 - \frac{\mathbf{b}^T\mathbf{b}}{k+1}.$$

Taking limit as $k \to \infty$ implies the result.

Part (II) Suppose $\hat{\theta}$ is inadmissible, then, there exists a better estimator $\tilde{\theta}$, that satisfies $E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$ for every $\boldsymbol{\eta} \in \mathbb{R}^r$ and by continuity (see e.g., Bickel and Doksum, 1977, P. 429), there exists a box $\mathbf{B} = [a_1, b_1] \times \cdots \times [a_r, b_r]$ such that $E(\hat{\theta} - \theta)^2 - E(\tilde{\theta} - \theta)^2 \geq \varepsilon$ for certain $\varepsilon > 0$ and for every $\boldsymbol{\eta} \in \mathbf{B}$. Therefore,

$$\int_{\mathbb{R}^r} \left\{E(\hat{\theta} - \theta)^2 - E(\tilde{\theta} - \theta)^2\right\}\pi_k(\boldsymbol{\eta})d\boldsymbol{\eta} \geq \int_{\mathbf{B}} \left\{E(\hat{\theta} - \theta)^2 - E(\tilde{\theta} - \theta)^2\right\}\pi_k(\boldsymbol{\eta})d\boldsymbol{\eta}$$
$$\geq \varepsilon \int_{\mathbf{B}} \pi_k(\boldsymbol{\eta})d\boldsymbol{\eta},$$

which implies that

$$\varepsilon \leq \frac{r_k(\hat{\theta}) - r_k(\tilde{\theta})}{\int_{\mathbf{B}} \pi_k(\boldsymbol{\eta}) d\boldsymbol{\eta}} = \frac{\mathbf{b}^T\mathbf{b} - r_k(\tilde{\theta})}{\int_{\mathbf{B}} \pi_k(\boldsymbol{\eta}) d\boldsymbol{\eta}} \leq \frac{\mathbf{b}^T\mathbf{b} - r_k}{\int_{\mathbf{B}} \pi_k(\boldsymbol{\eta}) d\boldsymbol{\eta}}. \tag{20}$$

However, $\mathbf{b}^T\mathbf{b} - r_k = \frac{\mathbf{b}^T\mathbf{b}}{k+1}$ and $\int_{\mathbf{B}} \pi_k(\boldsymbol{\eta}) d\boldsymbol{\eta}$ is of order of $1/\sqrt{k}$; hence, the right hand-side of (20) converges to zero as $k \to \infty$ in contradiction. □

D.3. Proof of Theorem 2

Suppose that a consistent estimator exists. Denote $f_p(\theta) := E(\tilde{\theta} - \theta)^2$; $f_p(\theta)$ converges pointwise to 0. Therefore, by Egorov's theorem, for each finite interval $[-B, B]$ and $\delta > 0$, $f_p(\theta) \to 0$ uniformly on $\theta \in [-B, B] \setminus A_\delta$, where A_δ has Lebesgue measure δ. We set $B = B_0$ for certain "large" B_0, which is defined now. To this end, we use the following result from Bickel (1981): for $X \sim N(\xi, 1)$ when the parameter space is $\xi \in [-\tilde{B}, \tilde{B}]$, the minimax risk rate is $1 - \frac{\pi^2}{\tilde{B}^2} + o(1/\tilde{B}^2)$ attained by a Bayes rule with respect to a certain continuous prior, which we denote by $\pi_{\tilde{B}}(\xi)$. Let $\tilde{B}$ be such that the minimax risk is $> 1/2$ and let B_0 be such that $|\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2|^{1/2}\sqrt{i_0}\tilde{B} \leq B_0$ for every p (recall that $\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2$ is bounded). Summing up,

$$\sup_{\theta \in [-B_0, B_0] \setminus A_\delta} E(\tilde{\theta} - \theta)^2 \to_{p\to\infty} 0. \tag{21}$$

We will now show a contradiction to (21). For every p define the set $\bar{\mathbf{B}}_p := [-\tilde{B}, \tilde{B}]^{i_0} \times \{0\}^{r-i_0}$. Then for every $\boldsymbol{\eta}_p \in \bar{\mathbf{B}}_p$ satisfies $\theta(\boldsymbol{\eta}_p) = \mathbf{b}_p^T\boldsymbol{\eta}_p \in [-B_0, B_0]$, since

$$\begin{aligned} |\mathbf{b}_p^T\boldsymbol{\eta}_p| &= \left|\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i\{\boldsymbol{\eta}_p\}_i\right| \leq \left|\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2\right|^{1/2}\left|\sum_{i=1}^{i_0}\{\boldsymbol{\eta}_p\}_i^2\right|^{1/2} \\ &\leq \left(\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2\right)^{1/2}\sqrt{i_0}\tilde{B} \leq B_0. \end{aligned}$$

Define the set $\mathbf{I}_\delta \subseteq \bar{\mathbf{B}}_p$ such that $\boldsymbol{\eta}_p \in \mathbf{I}_\delta \Leftrightarrow \boldsymbol{\eta}_p^T\mathbf{b}_p \in A_\delta$. Then the set $\mathbf{I}_\delta$ has an arbitrarily small Lebesgue measure as A_δ (since $\mathbf{I}_\delta \subseteq \bar{\mathbf{B}}_p$ and $\bar{\mathbf{B}}_p$ is a i_0-dimensional box).

Define $\boldsymbol{\pi}_p$ to be the density of the prior $\{\boldsymbol{\eta}_p\}_i \sim^{i.i.d} \pi_{\tilde{B}}$ (recall the above definition of $\pi_{\tilde{B}}$ as the least favorable prior) for $i = 1, \ldots, \bar{K}$ and $\{\boldsymbol{\eta}_p\}_i \equiv 0$ for $i > \bar{K}$. The prior $\boldsymbol{\pi}_p$ is defined on $\bar{\mathbf{B}}_p$ and its Bayes risk is $r_{\boldsymbol{\pi}} := \sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2\{1 - \frac{\pi^2}{\tilde{B}^2} + o(1/\tilde{B}^2)\} > \frac{1}{2}\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2$. Therefore,

$$\begin{aligned} \sup_{\theta \in [-B_0, B_0] \setminus A_\delta} E(\tilde{\theta} - \theta)^2 &\geq \sup_{\boldsymbol{\eta} \in \bar{\mathbf{B}}_p \setminus \mathbf{I}_\delta} E(\tilde{\theta} - \theta)^2 \geq \int_{\bar{\mathbf{B}}_p \setminus \mathbf{I}_\delta} E(\tilde{\theta} - \theta)^2 \boldsymbol{\pi}_p(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_p \\ &= \int_{\bar{\mathbf{B}}_p} E(\tilde{\theta} - \theta)^2 \boldsymbol{\pi}_p(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_p - \int_{\mathbf{I}_\delta} E(\tilde{\theta} - \theta)^2 \boldsymbol{\pi}_p(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_p \end{aligned}$$

$$\geq r_{\boldsymbol{\pi}} - \int_{\mathbf{I}_\delta} E(\tilde{\theta}-\theta)^2 \boldsymbol{\pi}_p(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_p$$
$$> \frac{1}{2}\sum_{i=1}^{i_0} \{\mathbf{b}_p\}_i^2 - \int_{\mathbf{I}_\delta} E(\tilde{\theta}-\theta)^2 \boldsymbol{\pi}_p(\boldsymbol{\eta}_p) d\boldsymbol{\eta}_p.$$

Now, $E(\tilde{\theta}-\theta)^2$ is bounded for $\boldsymbol{\eta}_p \in \mathbf{I}_\delta$ (since $|\theta|$ is bounded by B_0) and $\mathbf{I}_\delta$ has an arbitrarily small Lebesgue measure, then

$$\sup_{\theta \in [-B_0, B_0] \setminus A_\delta} E(\tilde{\theta}-\theta)^2 \geq \frac{1}{4}\sum_{i=1}^{i_0} \{\mathbf{b}_p\}_i^2 \geq \frac{1}{4}\sum_{i=1}^{i_0} \underline{\alpha}_i^2 \underline{\lambda}_i - \varepsilon, \tag{22}$$

where the last inequality holds true since $\sum_{i=1}^{i_0}\{\mathbf{b}_p\}_i^2 = \sum_{i=1}^{i_0} \frac{\lambda_i}{p}\{p\alpha_i\}^2$ and from the definitions $\underline{\alpha}_i$ and $\underline{\lambda}_i$. Inequality (22) holds for large enough p and for arbitrarily small $\varepsilon > 0$ and $\sum_{i=1}^{i_0} \underline{\alpha}_i^2 \underline{\lambda}_i$ is positive by assumption, in contradiction to (21). □

D.4. Proof of Proposition 2

One needs to verify when $\mathbf{a}^T\boldsymbol{\beta}_1 = \theta_1 \neq \theta_2 = \mathbf{a}^T\boldsymbol{\beta}_2$ implies that $X\boldsymbol{\beta}_1 \neq X\boldsymbol{\beta}_2$. We have

$$0 \neq \mathbf{a}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = (\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T + \boldsymbol{\alpha}_\perp^T\boldsymbol{\Gamma}_\perp^T)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2). \tag{23}$$

Necessity: Suppose that $\boldsymbol{\alpha}_\perp = 0$, then (23) implies that $\boldsymbol{\Gamma}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \neq 0$. Using the SVD decomposition $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^T$, this implies that

$$\mathbf{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{\Gamma}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \neq 0.$$

Sufficiency: Suppose that $\boldsymbol{\alpha}_\perp \neq 0$. Consider $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ such that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp$. Then,

$$\mathbf{a}^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = (\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^T + \boldsymbol{\alpha}_\perp^T\boldsymbol{\Gamma}_\perp^T)\boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp = ||\boldsymbol{\alpha}_\perp||^2 > 0$$

but

$$\mathbf{X}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = \mathbf{X}\boldsymbol{\Gamma}_\perp\boldsymbol{\alpha}_\perp = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\underbrace{\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}_\perp}_{=0}\boldsymbol{\alpha}_\perp = 0.$$ □

D.5. Proof of Proposition 3

To show part (I), recall that $\tilde{\mathbf{a}} = \frac{\sigma}{n}\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^T\mathbf{a}$ and $\tilde{\boldsymbol{\mu}} = \frac{1}{\sigma}\tilde{\boldsymbol{\Sigma}}\mathbf{D}\boldsymbol{\beta} = \frac{1}{\sigma}\mathbf{D}^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \frac{n}{\sigma}\mathbf{D}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\boldsymbol{\beta}$. Therefore, $\tilde{\mathbf{a}}^T\tilde{\boldsymbol{\mu}} = \mathbf{a}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\beta}$. Since $\mathbf{a} = \boldsymbol{\Gamma}\boldsymbol{\alpha}$, then $\mathbf{a}^T\boldsymbol{\Gamma} = \boldsymbol{\alpha}^T$ and $\mathbf{a}^T\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\beta} = \boldsymbol{\alpha}^T\boldsymbol{\Gamma}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}$.

For part (II), since $\tilde{\mathbf{Z}} = \mathbf{D}^{-1}\mathbf{X}^T\mathbf{Y}/\sigma$ and $\tilde{\mathbf{a}} = \frac{\sigma}{n}\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^T\mathbf{a}$, the natural estimator can be written as

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{Z}} = \mathbf{a}^T\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^T\mathbf{D}\frac{\sigma}{n}\mathbf{D}^{-1}\mathbf{X}^T\mathbf{Y}/\sigma = \mathbf{a}^T\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}^T\frac{1}{n}\mathbf{X}^T\mathbf{Y} = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{X}^T\mathbf{Y}.$$

We have that $\tilde{\mathbf{\Sigma}} = \mathbf{D}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{D}^{-1} = n\mathbf{D}^{-1}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T\mathbf{D}^{-1}$; hence the variance of the natural estimator is

$$\begin{aligned}\tilde{\mathbf{a}}^T\tilde{\mathbf{\Sigma}}\tilde{\mathbf{a}} &= \frac{\sigma^2}{n^2}\mathbf{a}^T\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^T\mathbf{D}n\mathbf{D}^{-1}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T\mathbf{D}^{-1}\mathbf{D}\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^T\mathbf{a} \\ &= \frac{\sigma^2}{n}\mathbf{a}^T\mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^T\mathbf{a} = \sigma^2\mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^-\mathbf{a}\end{aligned}$$ □

D.6. Proof of Proposition 4

We first show that under (6), (7), we have that $\tilde{\beta}_j = O(1/\sqrt{p})$ uniformly over j, where $\tilde{\boldsymbol{\beta}} = \mathbf{\Gamma}\boldsymbol{\beta}$. By the linear model (1),

$$\frac{1}{n}\sum_{i=1}^n(Y_i - \bar{Y})^2 = \boldsymbol{\beta}^T\frac{1}{n}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}\sum_{i=1}^n\varepsilon_i^2 = \sum_{j=1}^r\lambda_j\tilde{\beta}_j^2 + \frac{1}{n}\sum_{i=1}^n\varepsilon_i^2,$$

where $\tilde{\beta}_j = \boldsymbol{\gamma}_j^T\boldsymbol{\beta}$. Since $\frac{1}{n}\sum_{i=1}^n(Y_i-\bar{Y})^2$ and $\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2$ converge to $Var(Y)$ and σ^2, respectively, which are $O(1)$ by condition (7), then $\sum_{j=1}^r\lambda_j\tilde{\beta}_j^2$ is of order of a constant. Condition (6) now implies that $\sum_{j=1}^r\lambda_j\tilde{\beta}_j^2 = O(1)\tilde{\beta}_1^2\sum_{j=1}^r\lambda_j = O(1)\tilde{\beta}_1^2 p$, and therefore $\tilde{\beta}_j = O(1/\sqrt{p})$ uniformly over j.

We now consider two cases:

Global average: Suppose that $\alpha_j = c_\alpha O(1)$ uniformly for all $j = 1,\ldots,r$. In this case, $\theta = \boldsymbol{\alpha}^T\tilde{\boldsymbol{\beta}} = c_\alpha O(1)r/\sqrt{p}$, because $\tilde{\beta}_j = O(1/\sqrt{p})$. Therefore, the assumption that θ is $O(1)$ implies that $c_\alpha = \sqrt{p}/r$. The variance of $\hat{\theta}$ is

$$\sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{a} = \frac{\sigma^2}{n}\sum_{j=1}^r\frac{\alpha_j^2}{\lambda_j} = O(1)\frac{p}{nr^2}\sum_{j=1}^r\frac{1}{\lambda_j}.$$

Single entry: Suppose that $\boldsymbol{\alpha} = c_\alpha\mathbf{e}_j$. Assuming that $\theta = \mathbf{a}^T\boldsymbol{\beta} = \boldsymbol{\alpha}^T\tilde{\boldsymbol{\beta}} = c_\alpha\tilde{\beta}_j$ is $O(1)$, then $\tilde{\beta}_j = O(1/\sqrt{p})$ implies that $c_\alpha = \sqrt{p}$. The variance of $\hat{\theta}$ is

$$\sigma^2\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{a} = \frac{\sigma^2}{n}\sum_{j=1}^r\frac{\alpha_j^2}{\lambda_j} = \frac{\sigma^2}{n}\frac{c_\alpha}{\lambda_j} = O(1)\frac{p}{n\lambda_j}$$ □

D.7. Proof of Proposition 5

Part (I)(a) We show the result for limsup as liminf is similar. The variance of the estimator is

$$\tilde{\mathbf{a}}^T\mathbf{\Sigma}\tilde{\mathbf{a}} = \tilde{\mathbf{a}}^T\left[\sum_{i=1}^p\lambda_i\boldsymbol{\gamma}_i\{\boldsymbol{\gamma}_i\}^T\right]\tilde{\mathbf{a}} = \sum_{i=1}^{\bar{K}}\lambda_i\alpha_i^2 + \sum_{i=\bar{K}+1}^p\lambda_i\alpha_i^2. \tag{24}$$

The limit of the first term in (24) equals to

$$\limsup_p \sum_{i=1}^{\bar{K}} \lambda_i \alpha_i^2 = \limsup_p \sum_{i=1}^{\bar{K}} \lambda_i / p \left\{\sqrt{p}\alpha_i\right\}^2 = \sum_{i=\bar{K}+1}^{p} \bar{\lambda}_i \bar{\alpha}_i^2.$$

The second term in (24) coverges to zero since

$$\begin{aligned} 0 \leq \sum_{i=\bar{K}+1}^{p} \lambda_i \alpha_i^2 &\leq \lambda_{\bar{K}+1} \sum_{i=\bar{K}+1}^{p} \left\{\alpha_i\right\}^2 \leq \frac{\lambda_{\bar{K}+1}}{p} p ||\boldsymbol{\alpha}||^2 \\ &= \frac{\lambda^{(p)}_{\bar{K}+1}}{p} p ||\tilde{\mathbf{a}}||^2 = \frac{\lambda^{(p)}_{\bar{K}+1}}{p} O(1), \end{aligned} \tag{25}$$

which goes to 0 by the definition of $\bar{K}$.

Part (I)(b) We have that $b_i = \sqrt{\lambda_i}\alpha_i = \sqrt{\lambda_i/p}\sqrt{p}\alpha_i$; the result now follows from the definition of $\underline{\lambda}_i$, $\underline{\alpha}_i$, $\bar{\lambda}_i$ and $\bar{\alpha}_i$.

Part (I)(c) The result follows from the computation in (25).

Part II For limsup we have that,

$$\begin{aligned} \tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} &= \sum_{i=1}^{i_0} \frac{\lambda_i}{p} \left(\sqrt{p}\alpha_i\right)^2 + \sum_{i=i_0+1}^{p} \frac{\lambda_i}{p} \left(\sqrt{p}\alpha_i\right)^2 \\ &\leq \sum_{i=1}^{i_0} \frac{\lambda_i}{p} \left(\sqrt{p}\alpha_i\right)^2 + \frac{\lambda_{i_0+1}}{p} \sum_{i=1}^{\infty} (\sqrt{p}\alpha_i)^2. \end{aligned}$$

Taking limsup from both sides implies the result. For liminf, the inequality

$$\tilde{\mathbf{a}}^T \boldsymbol{\Sigma} \tilde{\mathbf{a}} \geq \sum_{i=1}^{i_0} \frac{\lambda_i}{p} \left(\sqrt{p}\alpha_i\right)^2$$

yields the claim. □

D.8. Proof of Proposition 6

By (13) we have that

$$\bar{\lambda}_1 - \pi\rho_1 = z + \sqrt{z^2 + \pi(1-\pi)\rho_B^2} \text{ and, } \bar{\lambda}_2 - (1-\pi)\rho_2 = z - \sqrt{z^2 + \pi(1-\pi)\rho_B^2},$$

where $z = \{\pi\rho_1 - (1-\pi)\rho_2\}/2$. Therefore, if (15) is satisfied, then

$$z + \sqrt{z^2 + \pi(1-\pi)\rho_B^2} + \pi\rho_B = 0 \text{ and, } z - \sqrt{z^2 + \pi(1-\pi)\rho_B^2} + (1-\pi)\rho_B = 0. \tag{26}$$

Summing up the equations in (26), we obtain $2z = -\rho_B$. Hence, if $z > 0$,

$$\begin{aligned} z + \sqrt{z^2 + \pi(1-\pi)\rho_B^2} + \pi\rho_B &= z + \sqrt{z^2 + 4\pi(1-\pi)z^2} - \pi 2z \\ &> z + z - \pi 2z = 2z(1-\pi) > 0, \end{aligned}$$

in contradiction to the first equation in (26). Similarly, if $z < 0$ ($z = 0$ implies that $\rho_B = 0$ which is not possible),

$$\begin{aligned} z - \sqrt{z^2 + \pi(1-\pi)\rho_B^2} + (1-\pi)\rho_B &= z - \sqrt{z^2 + 4\pi(1-\pi)z^2} - (1-\pi)2z \\ &< z + z - (1-\pi)2z = 2z\pi < 0, \end{aligned}$$

in contradiction to the second equation in (26). □

Acknowledgments

We thank Phil Reiss from Haifa U. for providing the data, Chun Fan from the U. of California, San Diego, for helping with the brain figures, and Larry Brown from U. of Pennsylvania and Asaf Weinstein from Stanford U. for useful discussions. This research was partially supported by NIH grant R01CA157528.

References

Azriel, D., Schwartzman, A. (2015). The Empirical Distribution of a Large Number of Correlated Normal Variables. *The Journal of the American Statistical Association*, **110**, 1217–1228. MR3420696

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300. MR1325392

Bickel, P.J. (1981). Minimax Estimation of the Mean of a Normal Distribution when the Parameter Space is Restricted. *Annals of Statistics*, **9**, 1301–1309. MR0630112

Bickel, P. J, Doksum, K. A. (1977). *Mathematical Statistics: basic ideas and selected topics*, San Francisco: Holden-Day. MR0443141

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *Journal of the American Statistical Association*, **57**, 269–306. MR0138176

Chen, X. (2018). Consistent FDR Estimation for Adaptive Multiple Testing Normal Means under Principal Correlation Structure. arXiv:1410.4275.

Dicker, L. H. (2014). Variance Estimation in High-dimensional Linear Models. *Biometrika*, **101**, 269–284. MR3215347

Efron, B. (2007). Correlation and Large-scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*, **102**, 93–103. MR2293302

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press, 2012. MR2724758

Fan, J., Wang, W. (2017). Asymptotics of Empirical Eigenstructure for High Dimensional Spiked Covariance. *The Annals of Statistics*, **45**, 1342–1374. MR3662457

Farebrother, R. W. (1976). Further Results on the Mean Square Error of Ridge Regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**, 248–250. MR0653156

Janson, L., Foygel Barber, R., Candès, E. J. (2017). EigenPrism: Inference for High-dimensional Signal-to-noise Ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 1037–1065. MR3689308

Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *The Annals of statistics*, **29**, 295–327. MR1863961

Owen, A. B. (2005). Variance of the Number of False Discoveries. *Journal of the Royal Statistical Society: Series B*, **67**, 411–426. MR2155346

Paul, D., Aue, A. (2014). Random Matrix Theory in Statistics: A Review. *Journal of Statistical Planning and Inference*, **150**, 1–29. MR3206718

Proal, E., Reiss, P. T., Klein, R. G., Mannuzza, S., Gotimer, K., Ramos-Olazagasti, M. A., Lerch, J. P., He, Y., Zijdenbos, A., Kelly, C., Milham, M. P., Castellanos, F. X. (2011). Brain Gray Matter Deficits at 33-year Follow-up in Adults With Attention-deficit/hyperactivity Disorder Established in Childhood. *Archives of general psychiatry*, **68**, 1122–1134.

Reiss, P. T., Schwartzman, A., Lu, F., Huang, L., Proal, E. (2012). Paradoxical Results of Adaptive False Discovery Rate Procedures in Neuroimaging Studies. *Neuroimage*, **63**, 1833–1840.

Schwartzman, A. (2008). Empirical Null and False Discovery Rate Inference for Exponential Families. *Annals of Applied Statistics*, **2**, 1332–1359. MR2655662

Schwartzman, A., Lin, X. (2011). The Effect of Correlation in False Discovery Rate Estimation. *Biometrika*, **98**, 199–214. MR2804220

Stein, C. (1956). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 197–206. MR0084922

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B*, **58**, 267–288. MR1379242

Vershynin, R. (2012). Introduction to the Non-asymptotic Analysis of Random Matrices. *Compressed sensing*, 210–268, Cambridge Univ. Press, Cambridge. MR2963170