Blockwise and coordinatewise thresholding to combine tests of different natures in modern ANOVA

Sylvain Sardy

Abstract

We derive new tests for fixed and random ANOVA based on a thresholded point estimate. The pivotal quantity is the threshold that sets all the coefficients of the null hypothesis to zero. Thresholding can be employed coordinatewise or blockwise, or both, which leads to tests with good power properties under alternative hypotheses that are either sparse or dense.

Keywords: ANOVA; multiple comparisons test; mixed effects; sparsity; thresholding.

1 Introduction

Analysis of variance (ANOVA) has something in common with thresholding regression in that ANOVA tests a null hypothesis that some parameters are equal to zero, and thresholding performs model selection by setting some coefficients to zero. This paper exploits this link to derive ANOVA tests based on thresholding.

While ANOVA and tests belong to the general knowledge of a statistician, thresholding is a more recent concept that we now review. A simple way to introduce thresholding and thresholding test is to consider the canonical regression model

$$Y_n = \theta_n + \epsilon_n \quad \text{for} \quad n = 1, \ldots, N, \tag{1}$$

where the noise is independent standard Gaussian. A thresholded point estimate of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ is obtained by applying a function η_λ to the data $\mathbf{Y} = (Y_1, \ldots, Y_N)$, that is,

$$\hat{\boldsymbol{\theta}} = \eta_\lambda(\mathbf{Y}) \tag{2}$$

with the property that some or all entries of the estimate are null for a large enough threshold λ. Defining $(x)_+ = \max(x, 0)$, thresholding can be performed:

- coordinatewise, for instance with soft-thresholding [Donoho and Johnstone, 1994]: considering each n in turn, estimate θ_n with

$$\eta_\lambda^{\text{soft}}(Y_n) = \left(1 - \frac{\lambda}{|Y_n|}\right)_+ Y_n =: \hat{\theta}_n. \tag{3}$$

- blockwise, for instance with truncated James-Stein thresholding [James and Stein, 1961]: considering all entries of $\mathbf{Y}$ together, estimate $\boldsymbol{\theta}$ with
$$\eta_{\lambda}^{\mathrm{JS}+}(\mathbf{Y}) = \left(1 - \frac{\lambda}{\|\mathbf{Y}\|_2^2}\right)_+ \mathbf{Y} =: \hat{\boldsymbol{\theta}}. \tag{4}$$

The choice of the threshold λ plays an important role in the quality of the estimation in regression (see for instance [Donoho and Johnstone, 1994], [Donoho and Johnstone, 1995], [Tibshirani, 1996], [Yuan and Lin, 2006], [Efron, 2004] and [Zou et al., 2007]) or to control the false discovery rate [Benjamini and Hochberg, 1995].

In this article we are interested in linear analysis of variance and testing null hypotheses regarding factors and continuous covariates. We derive new powerful tests based on a thresholded point estimate of the coefficients, and we choose the threshold λ_α for the test to have the desired level α. Since we commented that thresholding can be performed coordinatewise or blockwise, we derive tests based on either coordinate or block thresholding. Or hybrids of both. Two tests are used extensively in ANOVA: Tukey's multiple comparisons test and Fisher F-test. The first one is related to coordinate thresholding and the latter to block thresholding. One goal is to combine Tukey- and Fisher-like tests.

We illustrate the link between thresholding and testing on the canonical model (1) and derive two tests for $H_0: \ \theta_1 = \ldots = \theta_N = 0$ using a thresholded point estimate. For simplicity we assume for now unit standard deviation for the Gaussian noise. The two tests are:

- based on coordinatewise thresholding (3): clearly, for a sample $\mathbf{y}$, $\lambda_{\mathbf{y}} = \max_{n=1,\ldots,N} |y_n| = \|\mathbf{y}\|_\infty$ is the smallest and finite threshold that sets all the estimated parameters to zero. Letting F_{Λ_0} be the distribution of that statistics under the null hypothesis, and choosing $\lambda_\alpha = F_{\Lambda_0}^{-1}(1-\alpha)$, the test that rejects H_0 if $\lambda_{\mathbf{y}} > \lambda_\alpha$ or, equivalently if at least one entry of the coordinatewise thresholded point estimate $\hat{\boldsymbol{\theta}}(\lambda_\alpha)$ is different from zero has the desired level α. Here $\lambda_\alpha = -\Phi^{-1}([1-\exp\{\log(1-\alpha)/N\}]/2)$ has a closed form expression, otherwise one can estimate it by Monte Carlo. Arias-Castro et al. [2011] call this test a max-test for an obvious reason.
- based on blockwise thresholding (4): likewise, for a sample $\mathbf{y}$, $\lambda_{\mathbf{y}} = \|\mathbf{y}\|_2^2$ is the smallest and finite threshold that sets all the estimated parameters to zero. Under the null, that statistics $\Lambda_0 \sim \chi^2_N$. Therefore choosing the threshold $\lambda_\alpha = F_{\chi^2_N}^{-1}(1-\alpha)$, the test that rejects H_0 if $\hat{\boldsymbol{\theta}}(\lambda_\alpha)$ is different from the null vector leads to another test of level α.

Arias-Castro et al. [2011] studied the asymptotic relative power of both tests under either a dense or a sparse alternative hypothesis H_1. According to their definition, H_1 is dense or sparse whether the parameters $\boldsymbol{\theta}$ satisfy either $\{\|\boldsymbol{\theta}\|_2^2 \geq B\}$ or $\{\|\boldsymbol{\theta}\|_\infty \geq A\}$ for some positive lower bounds B and A, respectively. They proved that the max-test has more power under the sparse alternative.

Another goal of this paper is to derive thresholding-based tests for ANOVA considering either a coordinate or a block thresholding strategy, and see how they relate to and improve on existing tests. Our tests are based on thresholding estimators developed for linear models, and in particular lasso [Tibshirani, 1996], grouped lasso [Yuan and Lin, 2006] and smooth blockwise iterative thresholding

[Sardy, 2012]. We show how the threshold parameter λ_α of these estimators can be determined for the thresholding test to have the desired level α. Lockhart et al. [2013] consider a sequential approach of testing the significance of the successive lasso coefficients.

Section 2 starts will the simple one-way ANOVA, derives a blockwise and coordinatewise tests, and investigate their relative power under a dense and a sparse alternative. To take the best of both (dense and sparse) alternative worlds, Section 3 derives a single α-level test, called O&+ test, that is nearly as powerful as either the blockwise or the coordinatewise tests under both alternatives. Section 4 is concerned with Tukey multiple comparisons test, and proves that it amounts to a coordinate thresholding test. The latter has the advantage of having the exact desired level not only in the balanced situation (like Tukey's), but also in the unbalanced one (where Tukey's is conservative), albeit a Monte Carlo estimate of the threshold. Section 5 presents the general framework for iterative thresholding-based tests for ANOVA, for which some coefficients may be thresholded coordinatewise, blockwise or both, depending on the nature of the parameters (fixed effects, interactions, random effects) and on the nature of the alternative hypothesis considered (dense or sparse). Section 6 applies the new test to a real data set modeled by mixed effects. Section 7 proposes another selection of the threshold based on an extension of the universal threshold to satisfy both good estimation and model selection properties. Section 8 draws some conclusions.

2 One-way ANOVA: two tests

To fix notation, consider one-way ANOVA with T treatments and R replications

$$Y_{tr} = \mu_t + \epsilon_{tr}, \quad t = 1, \dots, T,\ r = 1, \dots, R, \tag{5}$$

where $\epsilon_{tr} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$ and the total number of observations is $N = TR$. In matrix notation, (5) is $\mathbf{Y} = X\boldsymbol{\mu} + \boldsymbol{\epsilon}$, where X is an $N \times T$ matrix with

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1R} \\ Y_{21} \\ \vdots \\ Y_{2R} \\ \vdots \\ Y_{T1} \\ \vdots \\ Y_{TR} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & 0 \\ 1 & 0 & 0 & \dots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

The matrix has orthogonal columns since $X^{\mathrm{T}} X = R\ I_T$. For testing

$$H_0 : \mu_1 = \ldots = \mu_T (= \mu) \quad \text{against} \quad H_1 : \text{for at least one } t,\ \mu_t \neq \mu \tag{6}$$

the most common approach is Fisher's test based on the pivot

$$F_{\mathbf{Y}_0} = \frac{(\mathrm{RSS}_{H_0} - \mathrm{RSS})/(T-1)}{\mathrm{RSS}/(N-T)} \sim \mathcal{F}_{T-1, N-T},$$

where $\mathcal{F}_{d_1,d_2}$ is the Fisher distribution with degrees of freedom d_1 and d_2, $\text{RSS} = \|\mathbf{Y}_0 - X\hat{\boldsymbol{\mu}}^{\text{LS}}\|_2^2$ with $\hat{\boldsymbol{\mu}}^{\text{LS}} = (X^{\text{T}}X)^{-1}X^{\text{T}}\mathbf{Y}_0$, $\text{RSS}_{H_0} = \|\mathbf{Y}_0 - \bar{Y}_0\mathbf{1}\|_2^2$ and $\mathbf{Y}_0 \sim \text{N}(\mu\mathbf{1}, \sigma^2 I_N)$. Next section shows that the same test can be derived based on block thresholding.

2.1 Blockwise thresholding test

It is well known that model (5) and test (6) are equivalent to testing

$$H_0 : \theta_1 = \ldots = \theta_T = 0 \quad \text{against} \quad H_1 : \text{for at least one } t,\ \theta_t \neq 0 \tag{7}$$

for the model

$$\mathbf{y} = \mu\mathbf{1} + X\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{8}$$

As opposed to (6), the formulation (7) of the null hypothesis is sparse in the sense that the parameters of the null hypothesis are all zero. This motivates the following test based on thresholding. For a given level α, the idea is to derive a thresholded point estimate $\hat{\boldsymbol{\theta}}(\lambda_\alpha)$ and to control the threshold λ_α such that the point estimate is the null vector with the desired probability $1-\alpha$ under the null hypothesis. The parameter μ is not to be tested, so we first calculate

$$\mathbf{y}_A = \mathbf{y} - P_A\mathbf{y} \quad \text{with} \quad P_A = A(A^{\text{T}}A)^{-1}A^{\text{T}} \tag{9}$$

to remove the contribution of the $N \times 1$ matrix $A = \mathbf{1}$ corresponding to μ. Here P_A is the projection matrix in the range of A. Then, given a positive threshold λ and a smoothness parameter $s \geq 1$, the block threshold estimate [Sardy, 2012]

$$\hat{\boldsymbol{\theta}}(\lambda) \quad = \quad \left(1 - \tfrac{\lambda}{\|X^{\text{T}}\mathbf{y}_A\|_2}\right)_+^s \tfrac{1}{R}X^{\text{T}}\mathbf{y}_A \tag{10}$$

generalizes truncated James-Stein's thresholding (4) and has the property that $\hat{\boldsymbol{\theta}}(\lambda) = \mathbf{0}$ iff $\lambda \geq \|X^{\text{T}}\mathbf{y}_A\|_2$. Note that $1/R$ in (10) stems from the inverse of $X^{\text{T}}X$. Observing that $\|X^{\text{T}}\mathbf{y}_A\|_2/\sigma$ is a pivot leads to the following theorem.

Theorem 1 (Block thresholding test): Consider model (8) for which we test (7) at a prescribed level α. Let $\mathbf{Y}_0 \sim \text{N}(\mu\mathbf{1}, \sigma^2 I_N)$ be the distribution of $\mathbf{Y}$ under H_0 and let $\hat{\sigma}$ be a positive estimate of σ such that $\hat{\sigma}/\sigma$ is a pivot. Then $\Lambda_{0,2} = \|X^{\text{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_2/\hat{\sigma}(\mathbf{Y}_0)$ is a pivot with distribution $F^O_{\Lambda_{0,2}}$. Defining the thresholded point estimate $\hat{\boldsymbol{\theta}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,2})$ in (10) for the observations $\mathbf{y}$ and setting the threshold $\lambda_{\alpha,2}$ such that $F^O_{\Lambda_{0,2}}(\lambda_{\alpha,2}) = 1-\alpha$, then the test

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{\theta}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,2}) \neq \mathbf{0} \\ 0 & \text{otherwise} \end{cases}$$

has level α. Finally, letting $\hat{\sigma}^2(\mathbf{y})$ be the standard unbiased estimate of variance, the block thresholding test is equivalent to Fisher test with the relation $\lambda^2_{\alpha,2} = R(T-1)F^{-1}_{\mathcal{F};T-1,N-T}(1-\alpha)$, where $F_{\mathcal{F}}$ is the cdf of the Fisher distribution.

Proof: $\Lambda_{0,2} = (\|U_0\|_2/\sigma)/(\hat{\sigma}(\mathbf{Y}_0)/\sigma)$, where $U_0 = X^{\text{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0) \sim \text{N}(\mathbf{0}, \sigma^2(RI_T - R^2/NJ_T))$ with J_T the $T \times T$ matrix of ones. So the distribution of ratio $\Lambda_{0,2}$ does not depend on (μ, σ). Moreover

$$\text{E}_{H_0}\phi(\mathbf{y}) = \text{P}\left(\hat{\boldsymbol{\theta}}(\frac{\mathbf{Y}_0}{\hat{\sigma}(\mathbf{Y}_0)}; \lambda_{\alpha,2}) \neq \mathbf{0}\right) = \text{P}\left(\frac{\|\mathbf{U}_0\|_2}{\hat{\sigma}(\mathbf{Y}_0)} > \lambda_{\alpha,2}\right) = 1 - F^O_{\Lambda_{0,2}}(\lambda_{\alpha,2}) = \alpha.$$

The equivalence to Fisher's test using the standard unbiased estimate of variance for σ^2 is straightforward. $\square$

By equivalence with Fisher's test, the distribution $F^O_{\Lambda_{0,2}}$ is known when using the standard unbiased estimate of variance, so that the $(1-\alpha)$-quantile $\lambda_{\alpha,2}$ can be calculated from the quantile of Fisher's distribution. In the more complex situations we will consider in the following, $F^O_{\Lambda_{0,2}}$ does not have an explicit expression. This is for instance the case with this test if another estimate of variance, say a robust one, possibly dependent of the numerator, is employed. In that case $\lambda_{\alpha,2}$ can be estimated by Monte-Carlo simulation.

2.2 Coordinatewise thresholding test

Instead of blocking the T treatment parameters into one block of size T and thresholding blockwise, thresholding could be performed on T blocks of size one, known as coordinatewise thresholding. For blocks of size one, smooth blockwise iterative thresholding [Sardy, 2012] defines a point estimate as a solution to a set of nonlinear equations

$$\left\{ \begin{array}{rcl} \hat\theta_t(\lambda) & = & \left(1-\frac{\lambda}{|\mathbf{x}_t^{\rm T}\mathbf{r}_t|}\right)_+^s \frac{1}{R}\mathbf{x}_t^{\rm T}\mathbf{r}_t \quad \text{with} \quad \mathbf{r}_t=\mathbf{y}_A-\sum_{i\neq t}\mathbf{x}_i\hat\theta_i(\lambda) \\ & & t=1,\ldots,T \end{array} \right. \tag{11}$$

for a given positive threshold λ and smoothness parameter $s \geq 1$. Note that $1/R$ in (11) stems from the inverse of $\mathbf{x}_t^{\rm T}\mathbf{x}_t = R$ for all t. For $s=1$, this defines the lasso estimate in a way that makes thresholding clearly visible: we recognize soft-thresholding (3) applied to least squares estimates on the partial residuals.

Moreover the estimate $\hat{\boldsymbol\theta}(\lambda) = (\hat\theta_1(\lambda),\ldots,\hat\theta_T(\lambda))$ satisfies the property that $\hat\theta_t(\lambda)=0$ for all $t=1,\ldots,T$ iff $\lambda \geq \|X^{\rm T}\mathbf{y}_A\|_\infty = \max_{t=1,\ldots,T}|\mathbf{x}_t^{\rm T}\mathbf{y}_A|$. This is a particular case of Lemma 1 proved in Section 5. Observing that $\|X^{\rm T}\mathbf{y}_A\|_\infty/\sigma$ is a pivot leads to the following theorem, which proof is similar to that of Theorem 1.

Theorem 2 (Coordinate thresholding test): Consider model (8) for which we test (7) at a prescribed level α. Let $\mathbf{Y}_0 \sim {\rm N}(\mu\mathbf{1}, \sigma^2 I_N)$ be the distribution of $\mathbf{Y}$ under H_0 and let $\hat\sigma$ be a positive estimate of σ such that $\hat\sigma/\sigma$ is a pivot. Then $\Lambda_{0,\infty} = \|X^{\rm T}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_\infty/\hat\sigma(\mathbf{Y}_0)$ is a pivot with distribution $F^+_{\Lambda_{0,\infty}}$. Defining the thresholded point estimate $\hat{\boldsymbol\theta}(\mathbf{y}/\hat\sigma(\mathbf{y});\lambda_{\alpha,\infty})$ solution to (11) for the observations $\mathbf{y}$ and setting the threshold $\lambda_{\alpha,\infty}$ such that $F^+_{\Lambda_{0,\infty}}(\lambda_{\alpha,\infty}) = 1-\alpha$, then the test

$$\phi(\mathbf{y}) = \left\{ \begin{array}{ll} 1 & \text{if } \hat\theta_t(\mathbf{y}/\hat\sigma(\mathbf{y});\lambda_{\alpha,\infty}) \neq 0 \text{ for at least one } t \in \{1,\ldots,T\} \\ 0 & \text{otherwise} \end{array} \right.$$

has level α.

2.3 Power analysis of both tests under two alternatives

We now compare the power of the two tests proposed in Sections 2.1 and 2.2. Given a level α, the thresholds

$$\lambda_{\alpha,2} = \sqrt{R}\sqrt{F^{-1}_{\chi^2_T}(1-\alpha)} \quad \text{and} \quad \lambda_{\alpha,\infty} = -\sqrt{R}\Phi^{-1}\left(\frac{1-(1-\alpha)^{1/T}}{2}\right) \tag{12}$$

respectively confer the blockwise (based on the 2-norm) and coordinatewise (based on the ∞-norm) tests the desired level. We assume here that μ and σ are known for simplicity, but the conclusions below would hold if both μ and σ were estimated. Arias-Castro et al. [2011] prove the asymptotic result that the test based on coordinate thresholding behaves differently than Fisher's test in terms of power depending whether the alternative hypothesis is dense or sparse.

We consider two alternatives: a dense alternative of the form

$$H_1^D : \boldsymbol{\theta} = \theta(\pm 1, \ldots, \pm 1),$$

and a sparse alternative of the form

$$H_1^S : \boldsymbol{\theta} = \theta(\pm 1, 0, \ldots, 0),$$

where $\theta \in \mathrm{I\!R}$. The power of both tests as a function of θ under both alternatives is reported in Table 1. Figure 1 plots power as a function of θ for $T = 5$ treatments and $R = 10$ replications. This corroborates the asymptotic results of Arias-Castro et al. [2011]: for a dense alternative, Fisher/block-test is more powerful, while for a sparse alternative coordinate-test is more powerful.

Table 1: Power of blockwise and coordinatewise thresholding tests at a level α under a dense and sparse alternative. Notation: $\Delta_\theta\Phi(\lambda; R) = \Phi((\lambda - R\theta)/\sqrt{R}) - \Phi((-\lambda - R\theta)/\sqrt{R})$.

	dense	sparse
block $\lambda = \lambda_{\alpha,2}$	$1 - F_{\chi^2_{T,RT\theta^2}}(\lambda^2/R)$	$1 - F_{\chi^2_{T,R\theta^2}}(\lambda^2/R)$
coordinate $\lambda = \lambda_{\alpha,\infty}$	$1 - \{\Delta_\theta\Phi(\lambda; R)\}^T$	$1 - \Delta_\theta\Phi(\lambda; R)\{\Delta_0\Phi(\lambda; R)\}^{T-1}$

3 One way ANOVA: the O&+ test

The relative power of both tests calls for a single test that would be of level α while being as powerful as the best between the block- and coordinate-tests. The following test approaches that goal by defining a point estimate based on joint block- and a coordinate-thresholding on the same parameters.

We now explain the notation F^O, F^+ and $F^{O\&+}$. In dimension two, the ℓ_2-ball employed by Fisher block-test is a circle symbolized by 'O', the two canonical directions employed by coordinate thresholding are the horizontal and vertical directions symbolized by '+', and so we call the joint test the O&+ test.

To define the O&+ test, consider the concatenated matrix $[X, X]$ and two sets of coefficients $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Block $\boldsymbol{\theta}_1$ into one block and treat the second coordinatewise. For reason we will explain, rescale the matrix X into $X_1 = XD_1$ and $X_2 = XD_2$, where D_1 and D_2 are diagonal. Finally consider the point estimate $\hat{\boldsymbol{\theta}}(\lambda) = (\hat{\boldsymbol{\theta}}_1(\lambda), \hat{\theta}_{2,1}(\lambda), \ldots, \hat{\theta}_{2,T}(\lambda))$ defined as a solution to

$$\left\{ \begin{array}{rcl} \hat{\boldsymbol{\theta}}_1(\lambda) & = & \left(1 - \frac{\lambda}{\|X_1^{\mathrm{T}}\mathbf{r}_1\|_2}\right)_+^s \frac{1}{R} D_1^{-2} X_1^{\mathrm{T}} \mathbf{r}_1 \quad \text{with} \quad \mathbf{r}_1 = \mathbf{y}_A - X_2 \hat{\boldsymbol{\theta}}_2(\lambda) \\ \hat{\theta}_{2,t}(\lambda) & = & \left(1 - \frac{\lambda}{|\mathbf{x}_{2,t}^{\mathrm{T}}\mathbf{r}_{2,t}|}\right)_+^s \frac{1}{R d_{2,t}^2} \mathbf{x}_{2,t}^{\mathrm{T}} \mathbf{r}_{2,t} \quad \text{with} \\ & & \mathbf{r}_{2,t} = \mathbf{y}_A - X_1 \boldsymbol{\theta}_1 - \sum_{i \neq t} \mathbf{x}_{2,i} \hat{\theta}_{2,i}(\lambda) \quad t = 1, \ldots, T \end{array} \right. \tag{13}$$

for a given positive threshold λ and smoothness parameter $s \geq 1$, where $\mathbf{y}_A = \mathbf{y} - P_A\mathbf{y}$ as before. The solution to the system is unique if $s > 1$ [Sardy, 2012], and has the property that $\hat{\boldsymbol{\theta}}(\lambda) = \mathbf{0}$ iff $\lambda \geq \max\{\|X_1^{\mathrm{T}}\mathbf{y}_A\|_2, \|X_2^{\mathrm{T}}\mathbf{y}_A\|_\infty\}$ for $s \geq 1$. This is a particular case of Lemma 1 proved below, which leads to the following test.

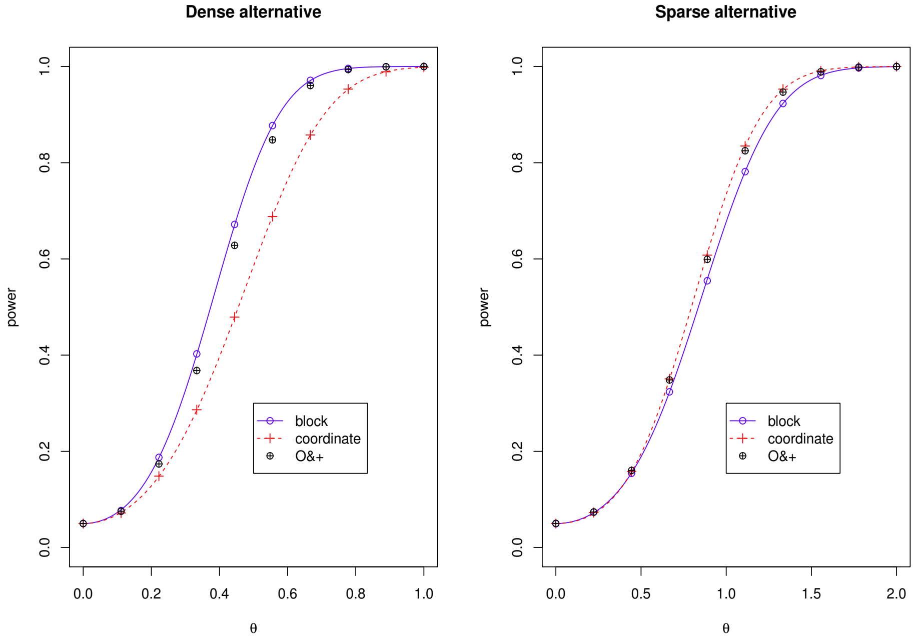

Figure 1: Power analysis of the three thresholding tests at a prescribed level $\alpha = .05$ for $T = 5$ treatments and $R = 10$ replications: block, coordinate and O&+ tests. The lines are the theoretical powers of Table 1 and the dots are empirical probabilities estimated by Monte-Carlo. Note that each test starts at power $\alpha = 0.05$ for $\theta = 0$, as expected.

Theorem 3 (O&+ test): Consider model (8) for which we test (7) at a prescribed level α. Let $\mathbf{Y}_0 \sim \mathrm{N}(\mu\mathbf{1}, \sigma^2 I_N)$ be the distribution of $\mathbf{Y}$ under H_0 and let $\hat{\sigma}$ be a positive estimate of σ such that $\hat{\sigma}/\sigma$ is a pivot. Then

$$\Lambda_{0,(2,\infty)} = \max\{\|X_1^{\mathrm{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_2)/\hat{\sigma}(\mathbf{Y}_0), \|X_2^{\mathrm{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_\infty/\hat{\sigma}(\mathbf{Y}_0)\} \tag{14}$$

is a pivot with distribution $F^{\oplus}_{\Lambda_{0,(2,\infty)}}$. Defining the thresholded point estimate $\hat{\boldsymbol{\theta}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,(2,\infty)})$ solution to (13) for the observations $\mathbf{y}$ and setting the threshold $\lambda_{\alpha,(2,\infty)}$ such that $F^{\oplus}_{\Lambda_{0,(2,\infty)}}(\lambda_{\alpha,(2,\infty)}) = 1 - \alpha$, then the test

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{\theta}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,(2,\infty)}) \neq \mathbf{0} \\ 0 & \text{otherwise} \end{cases}$$

has level α.

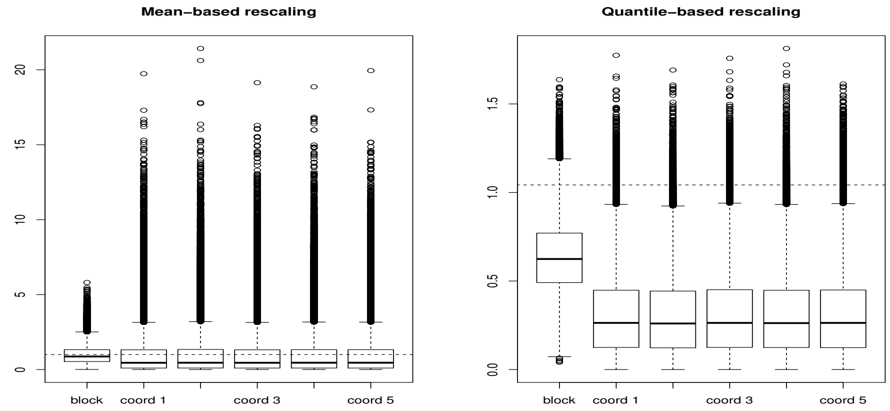

Figure 2: Illustration of mean-based rescaling (left) and quantile-based rescaling (right) for one-way ANOVA with $T = 5$ treatments and $R = 10$ replications. The series of 6 boxplots represents the empirical distribution of the components of the pivot $\Lambda_{0,(2,\infty)}$ defined in (14) under the null: the first one corresponds to realizations of $\|X_1^{\rm T}\mathbf{Y}_0\|_2$ and the other 5 correspond to realizations of $|\mathbf{x}_{2,t}^{\rm T}\mathbf{Y}_0|$ for $t = 1, \ldots, T$. We observe that the first boxplot on the right figure has the advantage of having upper extremes in the magnitude of the other five.

Following on Section 2.3 and Table 1 we consider the power of the joint-test under both dense and sparse alternatives as a function of θ (again assuming μ and σ are known). Figure 1 shows the power function of the block- and coordinate-tests (curve), and estimate them as well for values of θ on a grid with a Monte-Carlo simulation. We also report the empirical power of the O&+ test on the same grid, which has the remarkable property of performing under both alternatives almost as well as the best of the two individual tests.

To achieve with a single test a power nearly as good as the best of the two individual tests, rescaling the design matrix X by D_1 and D_2 is a crucial step. To allow the joint-test to have the same sensitivity whether the alternative is of the dense or sparse type, we perform the following quantile-based rescaling: we let the matrix corresponding to the block coefficients $\boldsymbol{\theta}_1$ be $X_1 = XD_1$ where D_1 is diagonal with entries $1/\lambda_{\alpha,2}$; likewise we let $X_2 = XD_2$ where D_2 is diagonal with entries $1/\lambda_{\alpha,\infty}$ for the coefficients $\boldsymbol{\theta}_2$ thresholded coordinatewise. The theoretical values of $\lambda_{\alpha,2}$ and $\lambda_{\alpha,\infty}$ are known (12) in our simple setting, but in more complex settings, we rely on a Monte-Carlo simulation to estimate them. Figure 2 illustrates the advantage of quantile-based rescaling (right), as opposed to mean-based rescaling (left) proposed by Yuan and Lin [2006] for group lasso. The left plot shows that, under the null, the boxplots are centered around their means (horizontal dotted line); because of that rescaling, the distribution of the block statistics $\|X_1^{\rm T}\mathbf{Y}_0\|_2$ (first boxplot from left) has its largest observations significantly lower than those of the coordinate statistics $|\mathbf{x}_{2,1}^{\rm T}\mathbf{Y}_0|, \ldots, |\mathbf{x}_{2,T}^{\rm T}\mathbf{Y}_0|$ (second to sixth boxplots). Consequently, with an alternative hypothesis of the dense type, the joint-test would have low power with that rescaling. Instead, the right plot of Figure 2 shows how quantile rescaling centers the distributions of the block and coordinate statistics around their $(1-\alpha)$-quantile (horizontal dotted line), hence providing the joint test with a homogeneous sensitivity under both dense and sparse alternatives.

4 Tukey multiple comparisons test

When the null hypothesis (6) is rejected, Tukey [1953] is interested in identifying which null hypotheses

$$H_0^{(t,t')}:\ \mu_t - \mu_{t'} = 0, \quad t = 1, \ldots, T,\ t' = t+1, \ldots, T \tag{15}$$

caused rejection of (6). His test is dual to intervals

$$\bar{y}_t - \bar{y}_{t'} - \frac{\mathcal{T}_{T,N-T}(\alpha)\hat{\sigma}(\mathbf{y})}{\sqrt{(R_t + R_{t'})/2}} \leq \mu_t - \mu_{t'} \leq \bar{y}_t - \bar{y}_{t'} + \frac{\mathcal{T}_{T,N-T}(\alpha)\hat{\sigma}(\mathbf{y})}{\sqrt{(R_t + R_{t'})/2}} \tag{16}$$

based on the Studentized range distribution $\mathcal{T}_{T,N-T}$, where $\hat{\sigma}^2(\mathbf{y})$ is the unbiased estimate of variance. Here $R_t > 0$ is the number of replication in treatment t (not necessarily all equal to R). The test rejects $H_0^{(t,t')}$ if zero is not covered by its corresponding interval (16). The level α of the test satisfies that none of the $T(T-1)/2$ tests are rejected with probability $1-\alpha$ under the null hypothesis (6). While the level is exact when $R_t = R$ for all t, the test is conservative [Hayter, 1984] for unbalanced designs, that is when the number of replication R_t differs between treatments.

Tukey's multiple comparisons test can also be derived based on thresholding and its level can be set to α, even when the number of replication R_t differs between treatments. To see that, first note that the design matrix is orthogonal with $X^{\mathrm{T}}X = \mathrm{diag}(R_1, \ldots, R_T)$. Then let $E = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and Δ be the $T(T-1)/2 \times T$ matrix such that $\boldsymbol{\delta} = \Delta\boldsymbol{\mu}$ are the pairwise differences. Left multiplying (2) be ΔE implies $\tilde{\mathbf{y}} = X\boldsymbol{\delta} + \tilde{\boldsymbol{\varepsilon}}$ with $X = I$, $\tilde{\mathbf{y}} = \Delta E\mathbf{y}$ and $\tilde{\boldsymbol{\varepsilon}} = \Delta E\boldsymbol{\varepsilon}$.

The coordinate thresholding estimate defined in (11) can now be applied to that latter model. First rescaling must be performed: the diagonal $D^2 = \mathrm{diag}(\sigma^2 \Delta E E^{\mathrm{T}} \Delta^{\mathrm{T}})$ of the covariance matrix of $\tilde{\boldsymbol{\varepsilon}}$ is not constant, unless $R_t = R$ for all $t = 1, \ldots, T$. So we standardize $X = I$ such that the marginals of $X^{\mathrm{T}}\tilde{\mathbf{y}}$ are Gaussian with identical variance under H_0, by multiplying X by D^{-1}. Since the matrix $X = D^{-1}$ is diagonal, the coordinate thresholding estimate defined in (11) has the closed form expression

$$\hat{\delta}_{(t,t')}(\lambda) = \left(1 - \frac{\lambda}{|\tilde{y}_{(t,t')}/d_{(t,t')}|}\right)_+ d_{(t,t')}\tilde{y}_{(t,t')}, \tag{17}$$

for all pairs (t,t'). This thresholded point estimate leads to the following test.

Theorem 4 (Coordinate thresholding test and Tukey multiple comparisons test). Consider model (5) for which we test all null hypotheses (15) at a prescribed level α. Let $\mathbf{Y}_0 \sim \mathrm{N}(\mu\mathbf{1}, \sigma^2 I_N)$ be the distribution of $\mathbf{Y}$ in (5) and let $\hat{\sigma}^2$ be the standard unbiased estimate of the variance. Then $\Lambda_{0,\infty} = \|D^{-1}\Delta E\mathbf{Y}_0\|_\infty / \hat{\sigma}(\mathbf{Y}_0)$ is a pivot with distribution $F^+_{\Lambda_{0,\infty}}$. Defining the

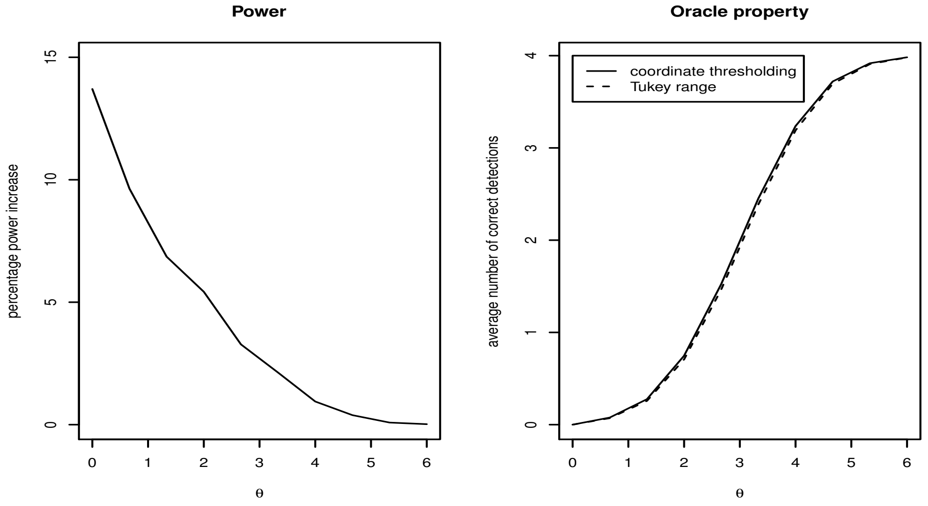

Figure 3: Results of Monte Carlo simulation for Tukey's multiple comparisons test for one-way ANOVA model (8) with $T = 5$ groups and sparse alternatives of the form $H_1 : \ \boldsymbol{\theta} = (\theta, 0, 0, 0, 0)$ with $\theta \in [0, 6]$ The number of replication is $(1, 5, 9, 10, 10)$ in each of the five groups. Left plot: percentage increase in power between the exact thresholding test (coordinate thresholding) and the conservative Tukey test. Right plot: average number of correct detections as a function of θ, the maximum possible being $T - 1 = 4$.

thresholded point estimate $\hat{\boldsymbol{\delta}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,\infty})$ in (17) for the observations $\mathbf{y}$ and setting the threshold $\lambda_{\alpha,\infty}$ such that $F^{+}_{\Lambda_{0,\infty}}(\lambda_{\alpha,\infty}) = 1 - \alpha$, then the test

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \hat{\delta}_{(t,t')}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_{\alpha,\infty}) \neq 0 \text{ for at least one } (t, t') \\ 0 & \text{otherwise} \end{cases}$$

has level α. Moreover if $R_t = R$ for all treatments, this test is equivalent to Tukey's range test with the relation $\mathcal{T}_{T,N-T}(\alpha) = \lambda_{\alpha,\infty}\sqrt{2}$, where $N = RT$.

Proof: the proof that this test has level α is straightforward. For the equivalence to Tukey multiple comparisons test when $R_t = R$ for all treatments t, note that on the one hand Tukey's test rejects when at least one interval (16) does not cover zero, or equivalently when $|\bar{y}_t - \bar{y}_{t'}| - \mathcal{T}_{T,N-T}(\alpha)\hat{\sigma}(\mathbf{y})/\sqrt{R} \leq 0$ for at least one pair (t, t'). On the other hand, the thresholding test rejects when $|\tilde{y}_{(t,t')}/d_{(t,t')}/\hat{\sigma}(\mathbf{y})| \leq \lambda_{\alpha,\infty}$, where $\tilde{y}_{(t,t')} = (\Delta E\mathbf{y})_{(t,t')} = \bar{y}_t - \bar{y}_{t'}$, and all $d^2_{(t,t')} = 2/R$ when $R = R_t$. So the thresholding test rejects when $|\bar{y}_t - \bar{y}_{t'}| \leq \lambda_{\alpha,\infty}\hat{\sigma}(\mathbf{y})\sqrt{2/R}$. So the two tests are equivalent and $\mathcal{T}_{T,N-T}(\alpha) = \lambda_{\alpha,\infty}\sqrt{2}$. □

We illustrate the gain in power and oracle property of the thresholding test in comparison to Tukey's multiple comparisons test on the same Monte Carlo simulation as in Section 3, except that the number of replication in each of the $T = 5$ treatments is $(1, 5, 9, 10, 10)$ instead of $R_t = R = 10$ for all treatments. The alternative hypothesis of the form $H_1 : \ \boldsymbol{\theta} = (\theta, 0, 0, 0, 0)$ is indexed by the parameter θ in the range $[0, 6]$. In that setting, Tukey's test is conservative as we observe on the left plot of Figure 3 where the percentage increase of power is plotted as a function of θ. The right plot informs on the number of correct detections, with a target value of $T-1$ non-zero entries guessed correctly. We see on the graph that the thresholding test has slighlty more correct detections on average than Tukey's conservative test.

If it is not clear to the statistician whether the alternative hypothesis is sparse or dense, then Tukey multiple comparisons test could be turned into an O&+ test as well to have more power under both alternatives.

Finally we considered pairwise contrasts, but the thresholding test could easily be implemented to test other contrasts.

5 Thresholding test for general ANOVA models

We now consider general ANOVA models which can be written in the form

$$\mathbf{y} = A\mathbf{b} + X\boldsymbol{\theta} + \sigma\mathbf{z} \quad \text{with} \quad \left\{ \begin{array}{l} X = [X_1 \ldots X_Q], \\ \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q), \\ \mathbf{z} \sim \mathrm{N}(\mathbf{0}, I_N). \end{array} \right. \tag{18}$$

The matrix A corresponds to the nuisance parameters $\mathbf{b}$. In the one-way ANOVA considered in the previous section, $A = \mathbf{1}$ and b is the intercept coefficients, but A can include a large number of parameters we do not want to test. The $N \times P$ matrix X corresponds to the parameters of interest, and is the horizontal concatenation of matrices $X_1, \ldots, X_Q$ corresponding to effects $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$, each of respective size P_q such that $\sum_{q=1}^{Q} P_q = P$. For instance in a one-way ANOVA plus random effects, $\boldsymbol{\theta}_1$ are the main and $\boldsymbol{\theta}_2$ random effects, as in the application of Section 6. Another example of concatenated matrices is the joint test of Section 3 with coefficients $(\boldsymbol{\theta}_1, \theta_{21}, \ldots, \theta_{2T})$ with corresponding matrices $X = [X_1 \mathbf{x}_{21} \ldots \mathbf{x}_{2T}]$. We assume that matrices A and $X_1, \ldots, X_Q$ are all full rank; X needs not be full rank. Moreover we assume the space of the nuisance parameters are not too large, in the sense that all X_q must have some components outside the range of A.

Our goal is to test the null hypothesis (7), that is test

$$H_0: \ \boldsymbol{\theta}_1 = \mathbf{0}, \ \ldots, \boldsymbol{\theta}_Q = \mathbf{0} \tag{19}$$

at a desired level α. To derive a threshold-based test, we must define a point estimate that thresholds and is uniquely defined. Sardy [2012] guarantees uniqueness of a thresholding estimator with a linear and invertible reparametrization of (18) into

$$\mathbf{y} = A\mathbf{b} + \tilde{X}\boldsymbol{\gamma} + \sigma\mathbf{z} \quad \text{with} \quad \left\{ \begin{array}{l} \tilde{X} = [\tilde{X}_1 \ldots \tilde{X}_Q], \\ \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_Q), \\ \mathbf{z} \sim \mathrm{N}(\mathbf{0}, I_N), \end{array} \right. \tag{20}$$

where each $\tilde{X}_q$ must satisfy the orthogonality condition: $\tilde{X}_q^{\mathrm{T}} \tilde{X}_q = d_q^2 I_{P_q}$ with $d_q > 0$. Group lasso is also defined with this condition [Yuan and Lin, 2006], but not necessarily uniquely. Orthogonalization can be achieved with a QR or SVD decomposition, for instance. Testing (19) is equivalent to testing

$$H_0: \ \boldsymbol{\gamma}_1 = \mathbf{0}, \ \ldots, \boldsymbol{\gamma}_Q = \mathbf{0}. \tag{21}$$

For a given threshold $\lambda > 0$ and a smoothness parameter $s \geq 1$, we introduce the thresholded point estimate $\hat{\boldsymbol{\gamma}}$ defined as a solution (not necessarily unique unless $s > 1$) to the following nonlinear system

$$\left\{\begin{array}{lcl} \hat{\boldsymbol{\gamma}}_q(\lambda) & = & \left(1 - \frac{\lambda}{\|\tilde{X}_q^{\mathrm{T}}\mathbf{r}_q\|_2}\right)_+^s \tilde{X}_q^{\mathrm{T}}\mathbf{r}_q/d_q \\ & & \text{with} \quad \mathbf{r}_q = \mathbf{y}_A - \sum_{q'\neq q}\tilde{X}_{q'}\hat{\boldsymbol{\gamma}}_{q'}(\lambda), \quad q \in \mathcal{Q}_{\text{block}} \\ \hat{\gamma}_{q,t}(\lambda) & = & \left(1 - \frac{\lambda}{|\tilde{\mathbf{x}}_{q,t}^{\mathrm{T}}\mathbf{r}_{q,t}|}\right)_+^s \tilde{\mathbf{x}}_{q,t}^{\mathrm{T}}\mathbf{r}_{q,t}/d_q \\ & & \text{with} \quad \mathbf{r}_{q,t} = \mathbf{y}_A - \tilde{\mathbf{X}}_{-(q,t)}\hat{\gamma}_{-q,t}(\lambda), \quad q \in \mathcal{Q}_{\text{coord}},\ t = 1, \ldots, P_q \end{array}\right. \tag{22}$$

where $\mathbf{y}_A$ is defined in (9) and $\mathcal{Q}_{\text{block}}$ are the indexes of blocked variables (resp., $\mathcal{Q}_{\text{coord}}$ for variables thresholded coordinatewise), and $\mathbf{X}_{-(q,t)}$ is the matrix X without column t of block q [Sardy, 2012]. This point estimate has the following important property.

Lemma 1: Considering system (22) for given $s \geq 1$ and $\lambda > 0$, then

$$\hat{\boldsymbol{\gamma}}_q(\lambda) = \mathbf{0} \text{ for all } q \in \mathcal{Q}_{\text{block}} \cup \mathcal{Q}_{\text{coord}} \tag{23}$$

if and only if

$$\lambda \geq \max\{\max_{q\in\mathcal{Q}_{\text{block}}} \|\tilde{X}_q^{\mathrm{T}}\mathbf{y}_A\|_2, \max_{q\in\mathcal{Q}_{\text{coord}}} \|\tilde{X}_q^{\mathrm{T}}\mathbf{y}_A\|_\infty\}. \tag{24}$$

In this case, $\hat{\boldsymbol{\gamma}}(\lambda)$ is uniquely defined.

Proof: the implication is straightforward. For the converse, the choice of λ in (24) implies $\mathbf{0}$ is a solution. The proof is complete if the zero solution is unique. When $s > 1$, Sardy [2012] proved uniqueness of the solution to (22). When $s = 1$, (22) are the first order optimality conditions to the hybrid lasso-grouped lasso defined as solution to

$$\min_{\boldsymbol{\gamma}} \frac{1}{2}\|\mathbf{y}_A - \tilde{X}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_{\mathcal{Q}},$$

where $\|\boldsymbol{\gamma}\|_{\mathcal{Q}} = \sum_{q\in\mathcal{Q}_{\text{block}}}\|\boldsymbol{\gamma}_q\|_2 + \sum_{q\in\mathcal{Q}_{\text{coord}}}\|\boldsymbol{\gamma}_q\|_1$ is a hybrid-norm. This cost function $C(\boldsymbol{\gamma})$ of the above minimization problem is convex in $\boldsymbol{\gamma}$, and the solution may not be unique in that case. In fact if two solutions $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ exist then their convex combinations $\boldsymbol{\gamma}_\delta = \delta\boldsymbol{\gamma}_1 + (1-\delta)\boldsymbol{\gamma}_2$ are also solutions, that is $C(\boldsymbol{\gamma}_1) = C(\boldsymbol{\gamma}_2) = C(\boldsymbol{\gamma}_\delta) = C^*$ for all $\delta \in [0, 1]$. But suppose $\tilde{X}\boldsymbol{\gamma}_1 \neq \tilde{X}\boldsymbol{\gamma}_2$, then the strict convexity of $\|\cdot\|_2^2$ and convexity of the $\mathcal{Q}$-norm imply that

$$\begin{aligned} C(\boldsymbol{\gamma}_\delta) &= \frac{1}{2}\|\delta(\mathbf{y}_A - \tilde{X}\boldsymbol{\gamma}_1) + (1-\delta)(\mathbf{y}_A - \tilde{X}\boldsymbol{\gamma}_2)\|_2^2 + \lambda\|\delta\boldsymbol{\gamma}_1 + (1-\delta)(\boldsymbol{\gamma})_2\|_{\mathcal{Q}} \\ &< \frac{1}{2}\{\delta\|\mathbf{y}_A - \tilde{X}\boldsymbol{\gamma}_1\|_2^2 + (1-\delta)\|\mathbf{y}_A - \tilde{X}\boldsymbol{\gamma}_2\|_2^2\} + \lambda(\delta\|\boldsymbol{\gamma}_1\|_{\mathcal{Q}} + (1-\delta)\|\boldsymbol{\gamma}_1\|_{\mathcal{Q}}) \\ &= C^*. \end{aligned}$$

This contradiction implies that any two solutions must satisfy $\tilde{X}\boldsymbol{\gamma}_1 = \tilde{X}\boldsymbol{\gamma}_2$, and have the same least squares cost. So their penalty term must be equal: $\lambda\|\boldsymbol{\gamma}_1\|_{\mathcal{Q}} = \lambda\|\boldsymbol{\gamma}_2\|_{\mathcal{Q}}$. Since $\boldsymbol{\gamma}_1 = \mathbf{0}$ is a solution, its $\mathcal{Q}$-norm is zero. Then necessarily $\boldsymbol{\gamma}_2 = \mathbf{0}$. □

The following theorem proposes a test of level α and shows it is a thresholding test since it amounts to testing whether a thresholded point estimate is null or not. Its proof is based on Lemma 1.

Theorem 5 (Thresholding test for ANOVA): Consider model (18) for which we test (19) at a prescribed level α. Assume A is full column rank and let P_A be the projection in the range of A. Let $\mathbf{Y}_0 \sim \mathrm{N}(A\mathbf{b}, \sigma^2 I_N)$ be the distribution of $\mathbf{Y}$ under H_0 and let $\hat{\sigma}^2$ be the standard unbiased estimate of variance. Then

$$\Lambda_0 = \max\{\max_{q\in\mathcal{Q}_{\text{block}}} \|\tilde{X}_q^{\mathrm{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_2, \max_{q\in\mathcal{Q}_{\text{coord}}} \|\tilde{X}_q^{\mathrm{T}}(\mathbf{Y}_0 - P_A\mathbf{Y}_0)\|_\infty\}/\hat{\sigma}(\mathbf{Y}_0) \tag{25}$$

is a pivot with distribution F_{Λ_0}. Letting $\lambda_\alpha = F_{\Lambda_0}^{-1}(1-\alpha)$, then the test

$$\phi(\mathbf{y}) = \begin{cases} 1 & \text{if } \lambda_\alpha < \max\{ \quad \max_{q\in\mathcal{Q}_{\text{block}}} \|\tilde{X}_q^{\mathrm{T}}(\mathbf{y} - P_A\mathbf{y})\|_2/\hat{\sigma}(\mathbf{y}), \\ & \qquad\qquad\qquad \max_{q\in\mathcal{Q}_{\text{coord}}} \|\tilde{X}_q^{\mathrm{T}}(\mathbf{y} - P_A\mathbf{y})\|_\infty/\hat{\sigma}(\mathbf{y})\} \\ 0 & \text{otherwise} \end{cases}$$

has level α. Moreover defining the thresholded point estimate $\hat{\boldsymbol{\gamma}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_\alpha)$ as a solution to (22) for the observation $\mathbf{y}$, then the test

$$\tilde{\phi}(\mathbf{y}) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{\gamma}}(\mathbf{y}/\hat{\sigma}(\mathbf{y}); \lambda_\alpha) \neq \mathbf{0} \\ 0 & \text{otherwise} \end{cases} = \phi(\mathbf{y}).$$

Rescaling the block matrices X_q after orthonormalizing them is a crucial step as we illustrated in Section 3. Quantile rescaling allows the Q blocks to contribute equally to the distribution of the pivot Λ_0 in (25), regardless of their sizes. Quantile rescaling is defined as follows. Given a block q and the orthonormalization W_q of X_q (with QR or SVD), quantile rescaling applies the same factor to W_q and leads to the rescaled matrix:

- $\tilde{X}_q = W_q d_q$, where $1/d_q = \lambda_{\alpha,2}^{(q)}$ is the $(1-\alpha)$-quantile of the distribution of $\|W_q^{\mathrm{T}}\mathbf{y}_A\|_2$ under the null, if the corresponding coefficients $\boldsymbol{\theta}$ are thresholded blockwise;
- $\tilde{X}_q = W_q d_q$, where $1/d_q = \lambda_{\alpha,\infty}^{(q)}$ is the $(1-\alpha)$-quantile of the distribution of $\|W_q^{\mathrm{T}}\mathbf{y}_A\|_\infty$ under the null, if the corresponding coefficients $\boldsymbol{\theta}$ are thresholded coordinatewise.

In the case $P > N$, we propose the following estimate of the standard deviation σ. Let $X = UDV^{\mathrm{T}}$ be the singular value decomposition of X, the design matrix of rank $R \leq N$. Then $\hat{\boldsymbol{\gamma}}^{\mathrm{LS}} = D\hat{\boldsymbol{\beta}}^{\mathrm{LS}} \sim \mathrm{N}(\boldsymbol{\gamma}, \sigma^2 I_R)$, where $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}$ is the least squares estimate with the transformed matrix XV, the matrix of principal component regression. In eigen directions of small singular values d_r, the true coefficients γ_r should essentially be zero. So we propose to estimate the standard deviation with $\hat{\sigma} = \mathrm{MAD}(|\hat{\gamma}_{p_0}^{\mathrm{LS}}|, \ldots, |\hat{\gamma}_R^{\mathrm{LS}}|)$ for p_0 large enough, say $p_0 = \lfloor R/2 \rfloor$. If P is prohibitively large to prevent an SVD, then its columns can be sampled to create sample matrices of a reasonable size, and repeated estimations of σ can then be aggregated into one. Note that this resampling procedure should be reproduced under the null to determine the appropriate threshold.

6 Application

We illustrate the thresholding test on a real data set modeled with mixed-effects [Pinheiro and Bates, 2000]. The effort y_{ij} required (on the Borg scale) to arise from a stool is measured for $J = 9$ different subjects each using $I = 4$ different types of stools. A linear mixed-effects model can be written as

$$y_{ij} = \theta_0 + X\boldsymbol{\theta}_1 + Z\boldsymbol{\theta}_2 + \epsilon_{ij},$$

where X model the fixed effects for Types (4 columns) and Z models the random effect for Subjects (9 columns). The noise is assumed i.i.d. $\mathrm{N}(0, \sigma^2)$ and $\boldsymbol{\theta}_2$ is believed to be independent realizations from $\mathrm{N}(0, \sigma_2)$. The goal is to test

$$H_0: \ \theta_{1,1} = \theta_{1,2} = \theta_{1,3} = \theta_{1,4} = 0 \quad \text{and} \quad \sigma_2 = 0,$$

or equivalently

$$H_0: \ \theta_{1,1} = \theta_{1,2} = \theta_{1,3} = \theta_{1,4} = 0 \quad \text{and} \quad \boldsymbol{\theta}_2 = \mathbf{0}.$$

So we employ the thresholded test coordinatewise for the fixed effects and blockwise for the random effect. Table 2 reports the result at a level $\alpha = 0.05$. The joint coordinate and block thresholding test rejects the null hypothesis because $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are both declared significantly different from zero. Moreover the test provides the information that level 3 of the type of stool is not significant.

After choosing the contrast $\sum_{i=1}^{4} \theta_{1,i} = 0$, the `lme` procedure available in `R` also declares σ_2 significantly different from zero, and $\theta_{1,3}$ not significant.

Table 2: ErgoStool data: thresholded estimate at a level $\alpha = 0.05$. The four fixed effects are thresholded coordinatewise and the random effects blockwise.

Fixed	θ_0		$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$	$\theta_{1,4}$			
	10.2		-0.81	1.38	0	-0.14			
Random	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$	$\theta_{2,4}$	$\theta_{2,5}$	$\theta_{2,6}$	$\theta_{2,7}$	$\theta_{2,8}$	$\theta_{2,9}$
	0.52	0.52	0.11	-0.30	-0.50	-0.03	0.11	-0.57	-0.10

7 Extension

Instead of pure testing, Yuan and Lin [2006] are interested in model selection with good mean squared error performance. To that aim, the universal threshold of Donoho and Johnstone [1994] can be adapted to our thresholding procedure. Recall that for wavelet smoothing, the matrix X is orthonormal and the universal threshold $\lambda = \sqrt{2 \log N}$ has the property to recover the true zero-vector with high probability since $\mathrm{P}(\hat{\boldsymbol{\theta}}_{\lambda_N} = \mathbf{0} \mid H_0) \approx 1 - 1/\sqrt{\pi \log N}$. Donoho et al. [1995] also showed minimax results for this choice of threshold. In the situation where X is not orthonormal but is rather a complex ANOVA matrix, one can define the quantile universal threshold.

Definition: Consider model (18) where $\boldsymbol{\theta}$ is segmented into Q groups. The quantile universal threshold (QUT) for the point estimate (22) is λ_Q such that

$$F_{\Lambda_0}(\lambda_Q) = 1 - \alpha \quad \text{with} \quad \alpha = 1/\sqrt{\pi \log Q}, \tag{26}$$

where F_{Λ_0} is the null distribution of the smallest threshold Λ_0 defined in (25) that sets to zero all estimated coefficients when the true ones are null.

We investigate the model selection and predictive performance of employing the quantile universal threshold (26) to threshold the parameters of the linear model (18). We consider Models III and IV used by Yuan and Lin [2006] to compare various estimators. Letting $\epsilon \sim \mathrm{N}(0, 2^2)$, then

- Model III has 2 factors out of $Q = 16$: let $Z_1, \ldots, Z_{16}, W$ be i.i.d. standard Gaussian and $X_i = (Z_i + W)/\sqrt{2}$, then generate 100 samples from

  $$Y = \{X_3^3 + X_3^2 + X_3\} + \{\frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6\} + \epsilon,$$

  There is a total of 16 groups of size 3.

- Model IV has 3 factors out of $Q = 20$: let $X_1, \ldots, X_{20}$ be generated as in Model III, and let $X_{11}, \ldots, X_{20}$ be trichotomized as 0, 1 and 2 if smaller then $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$ or in between, then generate 100 samples from

  $$Y = \{X_3^3 + X_3^2 + X_3\} + \{\frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6\} + \{2I(X_{11} = 0) + I(X_{11} = 1)\} + \epsilon,$$

  There is a total of 20 groups, half of size 3 and half of size 2.

We compare the performance of three estimators, least squares, grouped lasso and smooth blockwise iterative thresholding (22) for $s = 1$ (SBITE) and quantile rescaling based on three criteria: the number of factors selected, the MSE on the coefficients $\boldsymbol{\theta}$, and, of marginal interest for ANOVA, the predictive MSE on $\boldsymbol{\mu} = X\boldsymbol{\theta}$. We estimate these quantities by taking the average over 200 runs of a Monte-Carlo simulation. The results are summarized in Table 3. The selected threshold λ for group-lasso is either C_p or oracle [Yuan and Lin, 2006]. The SBITE estimator uses the quantile universal threshold (26) instead. The empirical results point to the excellent performance of SBITE with the quantile universal threshold both for the estimation of the number of factors and the MSE of the estimated ANOVA coefficients.

8 Conclusions

Thresholding tests alleviate three problems of standard ANOVA tests: they do not require to specify types of contraint, they are not sequential and they do not require exact knowledge of the distribution of complex pivots but simply require an estimate of the critical value, for instance by Monte Carlo. For the first time, block and coordinate thresholding is employed jointly to combine tests of various natures on the same parameters and therefore increase the power of the test under different alternatives. Hence, observing that Fisher's test comes from block thresholding and Tukey's test comes from coordinate thresholding, we filled a possible continuum between these two tests by developing hybrid tests based on ℓ_2- and ℓ_∞-norms, essentially tests based on combined F- and t-tests. More generally ℓ_p-tests could be derived.

How to put variables into groups is the choice of the statistician based not only on the nature of the parameters (fixed or random effects, main effects, interaction effects) but also on the type of alternative hypothesis he or she wants to test (sparse or dense).

Table 3: Results of Monte-Carlo simulation of [Yuan and Lin, 2006, Table 1, p. 61]. In bold, the best between C_p and QUT.

	Least squares	Group lasso (C_p/oracle)	SBITE (QUT)
Model III			
Estimated number of factors out of 16. True=2.	16	11/7.5	**3.7**
Model error			
on $\boldsymbol{\theta}$	7.2	1.5/0.7	**0.6**
on $X\boldsymbol{\theta}$	7.5	1.5/0.9	**1.4**
Model IV			
Estimated number of factors out of 20. True=3.	20	15/10	**5.2**
Model error			
on $\boldsymbol{\theta}$	15	3.4/2.1	**2.9**
on $X\boldsymbol{\theta}$	5.7	**1.6**/1.1	2.0

By deriving new tests based on thresholding in a linear ANOVA setting, this paper is the extension and practical implementation of group lasso and the max-test. The level of the test can be set to any desired level, which was not addressed by the original group lasso. We also showed that the proposed quantile rescaling is crucial to insure that parameters are democratically represented in the test whether they belong to a block of large or small size.

Combining dependent tests of various natures could be done for the higher criticism and false discovery rate approaches as well, by extending the work of Donoho and Jiashun [2004] and Benjamini and Hochberg [1995] with p-values related to dependent t-tests and F-tests.

R code is available upon request.

9 Acknowledgements

I thank David Donoho for an insightful discussion and for proposing the name 'O&+ test', Olivier Renaud for interesting discussions about ANOVA, Caroline Giacobino for her helpful rigor, and Florian Stern [Stern, 2012] and Laura Turbatu for their contribution to an early development of the idea.

References

E. Arias-Castro, E. J. Candes, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological*, 57:289–300, 1995.

D. Donoho and J. Jiashun. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224, 1995.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995.

B. Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

A. J. Hayter. A proof of the conjecture that the tukey-kramer multiple comparisons procedure is conservative. *The Annals of Statistics*, 12:61–75, 1984.

W. James and C. Stein. Estimation with quadratic loss. In Jerzy Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pages 361–379. University of California Press, 1961.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *arXiv:1301.7161*, 2013.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer-Verlag New York, Inc., 2000.

S. Sardy. Smooth blockwise iterative thresholding: a smooth fixed point estimator based on the likelihood's block gradient. *Journal of the American Statistical Association*, 107:800–813, 2012.

F. Stern. Lasso et estimateurs déerivés: Applications en analyse de la variance. Master's thesis, Université de Strasbourg, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Methodological*, 58:267–288, 1996.

J. W. Tukey. *The problem of multiple comparisons*. Chapman and Hall, New York, 1953. The Collected Works of John W. Tukey VIII. Multiple comparisons: 1948-1983.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 68(1):49–67, 2006.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35:2173–2192, 2007.