



HAL
open science

A Global Homogeneity Test for High-Dimensional Linear Regression

Camille Charbonnier, Nicolas Verzelen, Fanny Villers

► **To cite this version:**

Camille Charbonnier, Nicolas Verzelen, Fanny Villers. A Global Homogeneity Test for High-Dimensional Linear Regression. 2013. hal-00851592v1

HAL Id: hal-00851592

<https://hal.science/hal-00851592v1>

Preprint submitted on 15 Aug 2013 (v1), last revised 16 Jun 2014 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Global Homogeneity Test for High-Dimensional Linear Regression

Camille Charbonnier^{*}, Nicolas Verzelen[†] and Fanny Villers[‡]

Abstract: This paper is motivated by the comparison of genetic networks based on microarray samples. The aim is to test whether the differences observed between two inferred Gaussian graphical models come from real differences or arise from estimation uncertainties. Adopting a neighborhood approach, we consider a two-sample linear regression model with random design and propose a procedure to test whether these two regressions are the same. Relying on multiple testing and variable selection strategies, we develop a testing procedure that applies to high-dimensional settings where the number of covariates p is larger than the number of observations n_1 and n_2 of the two samples. Both type I and type II errors are explicitly controlled from a non-asymptotic perspective and the test is proved to be minimax adaptive to the sparsity. The performances of the test are evaluated on simulated data. Moreover, we illustrate how this procedure can be used to compare genetic networks on Hess *et al* breast cancer microarray dataset.

AMS 2000 subject classifications: Primary 62H15; secondary 62P10.

Keywords and phrases: Gaussian graphical model, two-sample hypothesis testing, high-dimensional statistics, multiple testing, adaptive testing, minimax hypothesis testing, detection boundary.

1. Introduction

The recent flood of high-dimensional data has motivated the development of a vast range of sparse estimators, in particular a large variety of derivatives from the Lasso. If theoretical guarantees have been provided in terms of prediction, estimation and selection performances (among a lot of others [6, 42, 28]), only a rather small proportion of the research effort has focused on quantifying the uncertainty surrounding the estimate *on a given data set with given design proportions*, be it in terms of confidence intervals or parametric hypothesis testing schemes guaranteeing a control on type I errors. Yet, quantifying the uncertainty is essential in applications where further experiments or developments rely on selected models and estimated coefficients.

This paper is mainly motivated by the validation of differences observed between Gaussian graphical models inferred from transcriptomic data from two subpopulations ([26],[15],[11]) when looking for potentially new drug or knock-out targets [20]. Following the development of differential analysis techniques, there is now a surging need for statistical validations of differential regulations between pairs of conditions. Of course, graph theory comes with a vast literature about graph comparisons. Yet, we would like to stress that the objective here is not to compare two graphical structures *taken for granted*, but to test whether the divergences in estimated graphical structures could come from *estimation uncertainties* or unveil actual differences between biological mechanisms. Adopting a neighborhood selection approach [26], hypothesis testing in the Gaussian graphical model can be solved via multiple hypothesis testing in the usual linear regression framework [40].

Therefore in the sequel we keep this motivation in mind but adopt the more general theoretical framework of high-dimensional linear regression. Formally, we consider the

^{*}e-mail: camille.charbonnier@ensae.org

[†]e-mail: nicolas.verzelen@supagro.inra.fr

[‡]e-mail: fanny.villers@upmc.fr

following statistical model

$$Y^{(1)} = X^{(1)}\beta^{(1)} + \epsilon^{(1)} \quad (1)$$

$$Y^{(2)} = X^{(2)}\beta^{(2)} + \epsilon^{(2)}, \quad (2)$$

where the size p row vectors $X^{(1)}$ and $X^{(2)}$ follow Gaussian distributions $\mathcal{N}(0_p, \Sigma^{(1)})$ and $\mathcal{N}(0_p, \Sigma^{(2)})$, whose covariance matrices remain unknown. The noise components $\epsilon^{(1)}$ and $\epsilon^{(2)}$ are independent from the design matrices and follow a centered Gaussian distribution with unknown standard deviations $\sigma^{(1)}$ and $\sigma^{(2)}$. In this formal setting, our objective is to develop a test for the equality of $\beta^{(1)}$ and $\beta^{(2)}$ which remains valid in high-dimension.

1.1. Related results

The literature on high-dimensional two-sample tests is very light. In the context of high-dimensional two-sample comparison of means, [4, 33, 10, 25] have introduced global tests to compare the means of two high-dimensional Gaussian vectors with unknown variance. Recently, [8, 23] developed a two-sample test for covariance matrices of two high-dimensional vectors.

In contrast, the one-sample analog of our problem has recently attracted a lot of attention, offering as many theoretical bases for extension to the two-sample problem. In fact, the high-dimensional linear regression tests for the nullity of coefficients can be interpreted as a limit of the two-sample test in the case where $\beta^{(2)}$ is known to be zero, and the sample size n_2 is considered infinite so that we perfectly know the distribution of the second sample.

There are basically two different objectives in high-dimensional linear testing: the local approach and the global approach. In the local approach, one considers the p tests of the nullity of the coefficients $\mathcal{H}_{0,i} : \beta_i^{(1)} = 0$ ($i = 1, \dots, p$), and the objective is to control error measures such as the false discovery rate of the resulting multiple testing procedures. In a way, one aims to assess the statistical significance of the variables. This can be achieved by providing a confidence region of $\beta^{(1)}$. In his seminal paper, [37] suggests such a confidence region for the Lasso estimator but his region inappropriately gives a null variance for all coefficients which are set to zero. A somewhat answer to this issue would be provided by Bayesian approaches like the Bayesian Lasso [21], which provides posterior credible intervals for each coefficient. Two recent papers have also addressed these issues from a frequentist point of view. [44] provides robust confidence intervals for each individual component of β building upon the Lasso and the scaled Lasso [2, 35, 36]. Independently, [7] develop a similar idea, building upon the Ridge estimator. Another line of work in the local approach, amounts to derive p -values for the nullity of the coefficients. Indeed, [43] suggests to split the sample in half and apply model selection on the first half in order to test for the significance of each coefficient using the usual combination of ordinary least squares and Student t-test on a model of reasonable size on the second half. To reduce the dependency of the results to the splitting, [27] advocate to use half-sampling B times, and aggregate the B p -values obtained for variable j in a way which controls either the family-wise error rate or false discovery rate.

In the global approach, the objective is to test the null hypothesis $\mathcal{H}_0 : \beta^{(1)} = 0$. If the global approach is clearly less informative than approaches providing individual significance tests like [27, 44, 7], global approaches can reach better performances for fewer sample sizes. The idea of [41], based upon the work of [5], is to approximate the alternative $\mathcal{H}_1 : \beta^{(1)} \neq 0$ by a collection of tractable alternatives $\{\mathcal{H}_1^S : \exists j \in S, \beta_j^{(1)} \neq 0, S \in \mathcal{S}\}$ working on subsets $S \subset \{1, \dots, p\}$ of reasonable size. The null hypothesis is rejected if the null hypothesis \mathcal{H}_0^S is rejected for at least one the collection $S \in \mathcal{S}$. If the resulting

procedure is computationally intensive, it is non-asymptotically minimax adaptive to the unknown sparsity of $\beta^{(1)}$, that is it achieves the optimal rate of detection without any assumption on the population covariance $\Sigma^{(1)}$ of the covariates. Another series of work relies on the higher-criticism. Higher-criticism was originally introduced in orthonormal designs [14], but has been proved to reach optimal detection rates in high-dimensional linear regression as well [3, 19]. In the end, higher-criticism is highly competitive in terms of computing time, and achieves the asymptotic rate of detection with the optimal constants. However, these nice properties require strong assumptions on the design.

While writing this paper, we came across the parallel work of Städler and Mukherjee [34], which nicely adapts the *screen and clean* procedure of Wasserman and Roder [43] as well as Meinshausen *et al.* [27] to the two-sample framework. Most importantly, because they estimate the supports of sample-specific estimators and joint estimator separately in the screening step, they resort to an elegant estimation of the p -values for the non-nested likelihood ratio test in the cleaning step. Yet, they do not provide any theoretical controls on type I error control or power for their overall testing strategy.

In this paper, we will build our testing strategy upon the global approach developed by [5] and [41]. Contrary to the *screen and clean* procedure, which can suffer in both steps from power losses following half-sampling of already small samples, this approach can be proved to achieve optimal rates of detection.

1.2. Testing hypotheses and form of the design

Suppose that we observe an n_1 -sample of $(Y^{(1)}, X^{(1)})$ and an n_2 -sample of $(Y^{(2)}, X^{(2)})$ noted $\mathbf{Y}^{(1)}$, $\mathbf{Y}^{(2)}$, $\mathbf{X}^{(1)}$, and $\mathbf{X}^{(2)}$. Defining analogously $\epsilon^{(1)}$ and $\epsilon^{(2)}$, we obtain the decompositions $\mathbf{Y}^{(1)} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon^{(1)}$ and $\mathbf{Y}^{(2)} = \mathbf{X}^{(2)}\beta^{(2)} + \epsilon^{(2)}$. The objective is to test whether models (1) and (2) are the same, that is

$$\begin{cases} \mathcal{H}_0 : \beta^{(1)} = \beta^{(2)}, \quad \sigma^{(1)} = \sigma^{(2)}, \quad \text{and} \quad \Sigma^{(1)} = \Sigma^{(2)} \\ \mathcal{H}_1 : \beta^{(1)} \neq \beta^{(2)}, \quad \sigma^{(1)} \neq \sigma^{(2)}. \end{cases} \quad (3)$$

In the null hypothesis, we have included the assumption that the population covariances of the covariates are equal ($\Sigma^{(1)} = \Sigma^{(2)}$). This choice of assumption is primarily motivated by our final objective to derive homogeneity tests for Gaussian graphical models.

Furthermore, our assumption can be interpreted as an intermediary case between two fixed designs settings: design equality ($\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$) and arbitrary different design ($\mathbf{X}^{(1)} \neq \mathbf{X}^{(2)}$). In the first case, the two-sample problem amounts to a one-sample problem by considering $\tilde{\mathbf{Y}} = \mathbf{Y}^{(1)} - \mathbf{Y}^{(2)}$ and it has therefore been studied in the aforementioned literature. The second case is extremely difficult. Indeed, we show in Section 7 that, in this particular setting, no test can do better than random guess. Below, we shall prove that the testing problem (3) is much simpler than fixed and different design and at the same time more versatile than design equality.

1.3. Our contribution

In this paper, we introduce a novel two-sample testing procedure for testing the homogeneity of two high-dimensional regression models (3). This test, which is built upon the work of [5], is completely data-driven and its type I error is explicitly controlled. Furthermore, it is computationally amenable in a large p and small n setting. Interestingly, the procedure does not require any half-sampling steps which are known to decrease the

robustness when the sample size is small. Finally, we prove that this procedure is minimax adaptive to the sparsity of $\beta^{(1)}$ and $\beta^{(2)}$ from a non-asymptotic point of view. Below, we describe the ideas underlying our approach.

Likelihood ratio statistics used to test such hypotheses like \mathcal{H}_0 in the classical *large n, small p* setting are untractable on high-dimensional datasets for the mere reason that the maximum likelihood estimator is not itself defined under high-dimensional design proportions. Our approach approximates the untractable high-dimensional test by a multiple testing construction, similarly to the strategy developed by [5] in order to derive statistical tests against non-parametric alternatives and adapted to one sample tests for high-dimensional linear regression in [41]. The testing strategy relies on the fundamental assumption that either the true supports of $\beta^{(1)}$ and $\beta^{(2)}$ are sparse or that their difference $\beta^{(1)} - \beta^{(2)}$ is sparse, so that the test can be successfully led in a subset $S^* \subset \{1, \dots, p\}$ of variables with reasonable size, compared to the sample sizes n_1 and n_2 . Of course, this low dimensional subset S^* is unknown. The whole objective of the testing strategy is to achieve similar rates of detection (up to a logarithmic constant) as an oracle test which would know in advance the optimal low-dimensional subset S^* .

If S stands for any subset of $\{1, \dots, p\}$ satisfying $2|S| \leq n_1 \wedge n_2$, we define the following restricted linear regression model :

$$\begin{cases} \mathbf{Y}^{(1)} &= \mathbf{X}_S^{(1)} \beta_S^{(1)} + \epsilon_S^{(1)} \\ \mathbf{Y}^{(2)} &= \mathbf{X}_S^{(2)} \beta_S^{(2)} + \epsilon_S^{(2)}, \end{cases}$$

where $X_S^{(1)}$ and $X_S^{(2)}$ represent the restriction of the random vectors $X^{(1)}$ and $X^{(2)}$ to covariates indexed by S , with covariance structures $\Sigma_S^{(1)}$ and $\Sigma_S^{(2)}$ respectively. Of course, $\epsilon_S^{(1)}$ and $\epsilon_S^{(2)}$ follow centered Gaussian distributions with new unknown conditional standard deviations $\sigma_S^{(1)}$ and $\sigma_S^{(2)}$. Coefficients $\beta_S^{(1)}$ and $\beta_S^{(2)}$ represent the coefficients of the orthogonal linear projection of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ upon the spaces generated by $\mathbf{X}_S^{(1)}$ and $\mathbf{X}_S^{(2)}$ respectively. We now state the test hypotheses in reduced dimension:

$$\begin{cases} \mathcal{H}_{0,S} : & \beta_S^{(1)} = \beta_S^{(2)}, \quad \sigma_S^{(1)} = \sigma_S^{(2)}, \quad \text{and} \quad \Sigma_S^{(1)} = \Sigma_S^{(2)}, \\ \mathcal{H}_{1,S} : & \beta_S^{(1)} \neq \beta_S^{(2)}, \quad \text{or} \quad \sigma_S^{(1)} \neq \sigma_S^{(2)}. \end{cases}$$

Of course, there is no reason in general for $\beta_S^{(1)}$ and $\beta_S^{(2)}$ to coincide with the restrictions of $\beta^{(1)}$ and $\beta^{(2)}$ to S , even less in high-dimension since variables in S can be in all likelihood correlated with covariates in S^c . Yet, as exhibited by Lemma 1.1, there is still a strong link between the collection of low dimension hypotheses $\mathcal{H}_{0,S}$ and the global null hypothesis \mathcal{H}_0 .

Lemma 1.1. *The hypothesis \mathcal{H}_0 implies $\mathcal{H}_{0,S}$ for any subset $S \subset \{1, \dots, p\}$.*

Proof. Under \mathcal{H}_0 , the random vectors of size $p+1$ $(Y^{(1)}, X^{(1)})$ and $(Y^{(2)}, X^{(2)})$ follow the same distribution. Hence, for any subset S , $Y^{(1)}$ follows conditionally on $X_S^{(1)}$ the same distribution as $Y^{(2)}$ conditionally on $X_S^{(2)}$. In other words, $\beta_S^{(1)} = \beta_S^{(2)}$. \square

By contraposition, it suffices to reject at least one of the $\mathcal{H}_{0,S}$ hypotheses to reject the global null hypothesis. This fundamental observation motivates our testing procedure. The idea is to build a multiple testing procedure that considers the testing problems $\mathcal{H}_{0,S}$ against $\mathcal{H}_{1,S}$ for a collection of subsets \mathcal{S} . Obviously, it would be prohibitive in terms of algorithmic complexity to test for each null hypothesis $\mathcal{H}_{0,S}$ for each $S \subset \{1, \dots, p\}$, since there would be 2^p such sets. As a result, we restrain ourselves to a relatively small

collection of hypotheses $\{\mathcal{H}_{0,S}, S \in \widehat{\mathcal{S}}\}$, where the collection of supports $\widehat{\mathcal{S}}$ is potentially data-driven. If the collection \mathcal{S} is judiciously selected, then we can manage not to lose too much power compared to the exhaustive search.

Concretely, we proceed in three steps :

1. We define new parametric statistics for testing $\mathcal{H}_{0,S}$ against $\mathcal{H}_{1,S}$. These statistics are related to the likelihood ratio statistic between the conditional distributions $Y^{(1)}|X_S^{(1)}$ and $Y^{(2)}|X_S^{(2)}$. Even if the distribution of these relevant statistics are free from any unknown parameter, it is computationally intensive to estimate the p -value by Monte-Carlo. This is why we provide an explicit and non-asymptotic upper bound of the p -values corresponding to these statistics.
2. We define algorithms aiming to select a data-driven collection of subsets $\widehat{\mathcal{S}}$ identified as most informative for our testing problem. These collections rely on the regularization path of the Lasso applied to a reparametrized linear model.
3. We define two calibration procedures which both guarantee a control on type-I error:
 - we use a Bonferroni calibration which is both computationally and conceptually simple;
 - we define a calibration procedure based upon permutations to reach a fine tuning of multiple testing calibration in practice, for an increase in empirical power.

After a short clarification of the notations, we devote Section 2 to the description of the adaptive likelihood-ratio procedure, along with theoretical controls of type-I error. Section 3 provides a non-asymptotic control of the power. Then, we derive that our testing procedure is minimax adaptive to the sparsity. Additional results on the power are postponed to Section 5. Section 4 provides simulated experiments comparing the performances of the suggested procedures along with an illustration of how to use this strategy to compare Gaussian graphical models inferred from microarray samples. Section 7 provides additional details about the technique used in Section 2 to control the quantiles of the likelihood-ratio statistic. Finally, all the proofs are postponed to Section 8.

1.4. Notation

In the sequel, ℓ_p norms are denoted $|\cdot|_p$, except for the l_2 norm which is referred as $\|\cdot\|$ to alleviate notations. For any positive definite matrix Σ , $\|\cdot\|_\Sigma$ denotes the Euclidean norm associated with the scalar product induced by Σ : for every vector x , $\|x\|_\Sigma^2 = x^\top \Sigma x$. Besides, for every set S , $|S|$ denote its cardinality. For any integer k , \mathbf{I}_k stands for the identity matrix of size k . For any square matrix A , $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ denote respectively the maximum and minimum eigenvalues of A . When the context makes it obvious, we may omit to mention A to alleviate notations and use φ_{\max} and φ_{\min} instead. Moreover, \mathbf{Y} refers to the size $n_1 + n_2$ concatenation of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ and \mathbf{X} refers to the size $(n_1 + n_2) \times p$ the concatenation of $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

To finish with, L refers to a positive numerical constant that may vary from line to line.

2. Adaptive Homogeneity Tests

We now turn to the description of the three major elements required by our strategy:

1. a parametric statistic to test the hypotheses $\mathcal{H}_{0,S}$;
2. a well-targeted data-driven collection of models $\widehat{\mathcal{S}}$;
3. a calibration procedure guaranteeing the control on type I error.

2.1. Parametric Test Statistic

Likelihood-based Statistic. In the following, $\mathcal{L}^{(1)}$ (resp. $\mathcal{L}^{(2)}$) denotes the log-likelihood of the first (resp. second) sample normalized by n_1 (resp. n_2). Given a subset $S \subset \{1, \dots, p\}$ of size smaller than $n_1 \wedge n_2$, $(\widehat{\beta}_S^{(1)}, \widehat{\sigma}_S^{(1)})$ stands for the maximum likelihood estimator of $(\beta^{(1)}, \sigma^{(1)})$ among vectors β whose supports are included in S . Similarly, we note $(\widehat{\beta}_S^{(2)}, \widehat{\sigma}_S^{(2)})$ for the maximum likelihood corresponding to the second sample.

We introduce a new parametric statistic taking the form of a two-sample likelihood-ratio, measuring the adequacy of sample-specific estimators to the opposite sample. To do so, let us define the likelihood ratio in sample i between an arbitrary pair (β, σ) and the corresponding sample-specific estimator $(\widehat{\beta}_S^{(i)}, \widehat{\sigma}_S^{(i)})$:

$$\mathcal{D}_{n_i}^{(i)}(\beta, \sigma) := \mathcal{L}_{n_i}^{(i)}(\widehat{\beta}_S^{(i)}, \widehat{\sigma}_S^{(i)}) - \mathcal{L}_{n_i}^{(i)}(\beta, \sigma).$$

With this definition, $\mathcal{D}_{n_1}^{(1)}(\widehat{\beta}^{(2)}, \widehat{\sigma}^{(2)})$ measures how far $(\widehat{\beta}^{(2)}, \widehat{\sigma}^{(2)})$ is from $(\widehat{\beta}^{(1)}, \widehat{\sigma}^{(1)})$ in terms of likelihood within sample 1.

We now define the following symmetrized statistic:

$$F_S := 2 \left[\mathcal{D}_{n_1}^{(1)}(\widehat{\beta}^{(2)}, \widehat{\sigma}^{(2)}) + \mathcal{D}_{n_2}^{(2)}(\widehat{\beta}^{(1)}, \widehat{\sigma}^{(1)}) \right]. \quad (4)$$

The statistic F_S amounts to comparing the estimators $(\widehat{\beta}_S^{(1)}, \widehat{\sigma}_S^{(1)})$ and $(\widehat{\beta}_S^{(2)}, \widehat{\sigma}_S^{(2)})$ through their corresponding log-likelihoods. In order to simplify the forthcoming analysis, we decompose the test statistic F_S into the sum of three terms $F_{S,1} + F_{S,2} + F_{S,3}$, where

$$\begin{aligned} F_{S,1} &= -2 + \frac{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2} + \frac{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1} \\ F_{S,2} &= \frac{\|\mathbf{X}_S^{(2)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1} \\ F_{S,3} &= \frac{\|\mathbf{X}_S^{(1)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}. \end{aligned}$$

This decomposition highlights the different distances at stake in F_S . While the first term $F_{S,1}$ evaluates the discrepancies in terms of conditional variances, the last two terms $F_{S,2}$ and $F_{S,3}$ address the comparison of $\beta^{(1)}$ to $\beta^{(2)}$.

We characterize the distribution of F_S via the distribution of each of these terms under $\mathcal{H}_{0,S}$, given in Proposition 2.1. To simplify notations, we denote by g the non-negative function defined on \mathbb{R}^+ mapping x to $-2 + x + 1/x$.

Proposition 2.1 (Conditional distributions of $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$ under $\mathcal{H}_{0,S}$).

1. Let Z denote a Fisher random variable with $(n_1 - |S|, n_2 - |S|)$ degrees of freedom. Then, under the null hypothesis,

$$F_{S,1} | \mathbf{X}_S \underset{\mathcal{H}_{0,S}}{\sim} g \left[Z \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} \right].$$

2. Let Z_1 and Z_2 be two centered and independent Gaussian vectors with covariance $\mathbf{X}_S^{(2)} \left[(\mathbf{X}_S^{(1)*} \mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)*} \mathbf{X}_S^{(2)})^{-1} \right] \mathbf{X}_S^{(2)*}$ and $I_{n_1 - |S|}$. Then, under the null hypothesis,

$$F_{S,2} | \mathbf{X}_S \underset{\mathcal{H}_{0,S}}{\sim} \frac{\|Z_1\|^2/n_2}{\|Z_2\|^2/n_1}.$$

A symmetric result holds for $F_{S,3}$.

Because the statistics F_S and $F_{S,i}$, $i = 1, \dots, 3$ are naturally increasing with the size of model S , the only way to calibrate the multiple testing step over a collection of models of various sizes is to convert the statistics to a unique common scale. The most natural is to convert observed $F_{S,i}$'s into p -values. In the sequel, we note $\bar{Q}_{1,|S|}(u | \mathbf{X}_S)$ (resp. $\bar{Q}_{2,|S|}(u | \mathbf{X}_S)$ and $\bar{Q}_{3,|S|}(u | \mathbf{X}_S)$) for the conditional probability that $F_{S,1}$ (resp. $F_{S,2}$ and $F_{S,3}$) is larger than u . With this notation, $\bar{Q}_{i,|S|}(F_{S,i} | \mathbf{X}_S)$ denotes the p -value associated with $F_{S,i}$, conditional on \mathbf{X}_S .

Although the distributions identified in Proposition 2.1 are not familiar distributions with ready-to-use quantile tables, they all share the advantage that they do not depend on any unknown quantity, such as design variances $\Sigma^{(1)}$ and $\Sigma^{(2)}$, noise variances $\sigma^{(1)}$ and $\sigma^{(2)}$, or even true signals $\beta^{(1)}$ and $\beta^{(2)}$. By Proposition 2.1, the p -value $\bar{Q}_{1,|S|}(x | \mathbf{X}_S)$ is easily computed from distribution function of a Fisher random variable. Since the conditional distribution of $F_{S,2}$ given X_S only depends on $|S|$, n_1 , n_2 , and \mathbf{X}_S , one could compute an estimation of the p -value $\bar{Q}_2(u | X_S)$ associated with an observed value u by Monte-Carlo simulations. However, this approach is computationally prohibitive for large collections of subsets S . This is why we use instead an explicit upper bound of $\bar{Q}_{2,|S|}(u | \mathbf{X}_S)$ based on Laplace method, as given by Proposition 2.2.

Proposition 2.2 (Upper-bound on $F_{S,2}$ and $F_{S,3}$ quantiles). *Let us note $a = (a_1, \dots, a_{|S|})$ the positive eigenvalues of*

$$\frac{n_1}{n_2(n_1 - |S|)} \mathbf{X}_S^{(2)} \left[(\mathbf{X}_S^{(1)*} \mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)*} \mathbf{X}_S^{(2)})^{-1} \right] \mathbf{X}_S^{(2)*}.$$

For any $u > |a|_1$, take

$$\tilde{Q}_{2,|S|}(u | \mathbf{X}_S) := \exp \left[-\frac{1}{2} \sum_{i=1}^{|S|} \log(1 - 2\lambda^* a_i) - \frac{n_1 - |S|}{2} \log \left(1 + \frac{2\lambda^* u}{n_1 - |S|} \right) \right],$$

where λ^* is explicitly defined in Section 7. Then, for any $u > |a|_1$,

$$\bar{Q}_{2,|S|}(u | \mathbf{X}_S) \leq \tilde{Q}_{2,|S|}(u | \mathbf{X}_S).$$

From now on, we use the upper-bounds $\tilde{Q}_{i,|S|}(F_{S,i} | \mathbf{X}_S)$ of the ideal p -values as the test statistics rather than the original statistics $F_{S,i}$. To simplify notations, we also denote by $\tilde{Q}_{1,|S|}$ the true p -value $\bar{Q}_{1,|S|}$ associated with $F_{S,1}$.

2.2. Choices of Test Collections

Many collections \mathcal{S} can be thought of. The ideal collection \mathcal{S} must satisfy the best tradeoff between the inclusion of the maximum number of relevant models S and a reasonable computing time, which is linear in the size $|S|$ of the collection. In the following, we distinguish deterministic and data-driven collections, which we differentiate by adding a hat on data-driven collections $\hat{\mathcal{S}}$.

Deterministic Collections. Among deterministic collections of tests, the most straightforward collections consist of all size- k subsets of $\{1, \dots, p\}$, which we denote \mathcal{S}_k . This kind of family is interesting in at least two ways. First, it neglects none of the variables: we cannot miss any signal. Second, it provides collections of tests which are independent from the data, thereby reducing the risk of overfitting. However, as we allow the model size k or total number of candidate variables p to grow, these deterministic families can rapidly reach unreasonable sizes. Admittedly, \mathcal{S}_1 always remains feasible, but reducing the search to models of size 1 can be costly in terms of power. As a variation on size k models, an interesting collection in terms of theoretical developments is the collection of all models of size smaller than k , denoted $\mathcal{S}_{\leq k} = \bigcup_{j=1}^k \mathcal{S}_j$.

Data-driven Collections. In order to investigate models of varying sizes while keeping the size of the collection moderate, we suggest to derive data-driven collections of tests $\widehat{\mathcal{S}}$. The idea is to start from a deterministic family \mathcal{S} and define an algorithm mapping $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ to some data-driven collection $\widehat{\mathcal{S}} \subset \mathcal{S}$ of restricted size. In practice, we start from $\mathcal{S}_{\leq D_{\max}}$, where $D_{\max} = \lfloor (n_1 \wedge n_2)/2 \rfloor$, and derive the collection $\widehat{\mathcal{S}}$ from the Lasso regularization path of a reparametrized joint regression model, presented in Equation (5).

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \theta_*^{(1)} \\ \theta_*^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix}. \quad (5)$$

In this reparametrized model, $\theta_*^{(1)}$ captures the mean effect $(\beta^{(1)} + \beta^{(2)})/2$, while $\theta_*^{(2)}$ captures the discrepancy between the sample-specific effect $\beta^{(i)}$ and the mean effect $\theta_*^{(1)}$, that is to say $\theta_*^{(2)} = (\beta^{(1)} - \beta^{(2)})/2$. Combining this reparametrization with variable selection by the Lasso, we aim to select, on the one hand, variables presenting strong common effects through $\theta_*^{(1)}$, on the other hand, variables presenting strong diverging effects through $\theta_*^{(2)}$. We build two families of models from this reparametrized model: first, the increasing family $\widehat{\mathcal{S}}_L^{(2)}$ of variables included by the Lasso in the $\theta_*^{(2)}$ part, by order of activation, second the increasing family $\widehat{\mathcal{S}}_L^{(1)}$ of variables included by the Lasso algorithm, independently from its activation in the $\theta_*^{(2)}$ or $\theta_*^{(1)}$ part.

Given $\lambda > 0$, we write \widehat{T}_λ for the support of the Lasso estimator of $\theta_* = (\theta_*^{(1)}, \theta_*^{(2)})$ with tuning parameter λ . Then we build the subsets $\widehat{\mathcal{S}}_\lambda^{(1)}$ and $\widehat{\mathcal{S}}_\lambda^{(2)}$ of $\{1, \dots, p\}$ by

$$i \in \widehat{\mathcal{S}}_\lambda^{(1)} \Leftrightarrow (i \in \widehat{T}_\lambda \text{ or } i + p \in \widehat{T}_\lambda), \quad i \in \widehat{\mathcal{S}}_\lambda^{(2)} \Leftrightarrow (i + p \in \widehat{T}_\lambda).$$

Denote by $\lambda_1, \lambda_2, \dots, \lambda_{k_{\max}}$ the parameter of the Lasso regularization path for regression model (5). Here k_{\max} is the smallest integer q such that $|\widehat{\mathcal{S}}_{\lambda_{q+1}}^{(1)}| > D_{\max}$. This allows us to build the two following families from the reparametrized model:

$$\widehat{\mathcal{S}}_L^{(1)} = \{\widehat{\mathcal{S}}_{\lambda_k}^{(1)}; k = 1, \dots, k_{\max}\}, \quad \widehat{\mathcal{S}}_L^{(2)} = \{\widehat{\mathcal{S}}_{\lambda_k}^{(2)}; k = 1, \dots, k_{\max}\}.$$

The justification of the second subset family is that we want to focus on variables which have disagreeing effects between the two samples. However, the divergence between effects might only appear conditionally on other variables with similar effects, this is why the first subset family is chosen to include both types of variables. In the end, we consider the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$, consisting of the reunion of both subset families and \mathcal{S}_1 ,

$$\widehat{\mathcal{S}}_{\text{Lasso}} := \widehat{\mathcal{S}}_L^{(1)} \cup \widehat{\mathcal{S}}_L^{(2)} \cup \mathcal{S}_1. \quad (6)$$

Of course, this part of the testing strategy is highly flexible: any other relevant model selection strategy can be used.

2.3. Combining the parametric statistics

The objective of this Section is to calibrate a multiple testing procedure based on the p -values $\{(\tilde{Q}_{1,|S|}(F_{S,i}|\mathbf{X}_S), \tilde{Q}_{2,|S|}(F_{S,i}|\mathbf{X}_S), \tilde{Q}_{3,|S|}(F_{S,i}|\mathbf{X}_S)), S \in \hat{\mathcal{S}}\}$, so that the type-I error remains smaller than a chosen level α .

When using a data-driven model collection, we must take good care of preventing the risk overfitting which results from using the same dataset both for model selection and hypothesis testing. In that purpose, we consider a given a deterministic collection \mathcal{S} of subsets and assume that the data-driven collection $\hat{\mathcal{S}}$ results from a fixed algorithm mapping $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ to $\hat{\mathcal{S}} \subset \mathcal{S}$. For the sake of simplicity, we assume in the two following sections that $\emptyset \not\subset \mathcal{S}$, which merely means that we do not include in the collection of tests the raw comparison of $\text{Var}(\mathbf{Y}^{(1)})$ to $\text{Var}(\mathbf{Y}^{(2)})$.

Testing Procedure Given a model collection $\hat{\mathcal{S}}$ and a sequence $\hat{\alpha} = \{\alpha_{i,S}, i = 1, 2, 3, S \in \hat{\mathcal{S}}\}$, we define the following test function:

$$T_{\hat{\mathcal{S}}}^{\hat{\alpha}} = \begin{cases} 1 & \text{if } \exists S \in \hat{\mathcal{S}}, \exists i \in \{1, 2, 3\} \quad \tilde{Q}_{i,|S|}(F_{S,i}|\mathbf{X}_S) \leq \alpha_{i,S}. \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

In other words, the test function rejects the global null if there exists at least one model $S \in \mathcal{S}$ such that at least one of the three p -values is below the corresponding threshold $\alpha_{i,S}$.

The next two paragraphs define two different calibrations for multiple testing over the collection of parametric tests.

Bonferroni Calibration (B). The collection of weights $\hat{\alpha}^B = \{\alpha_{i,S}, S \in \mathcal{S}\}$ is chosen such that

$$\sum_{S \in \mathcal{S}} \sum_{i=1}^3 \alpha_{i,S} \leq \alpha. \quad (8)$$

For the collection $\mathcal{S}_{\leq k}$, or any data-driven collection derived from $\mathcal{S}_{\leq k}$, a natural choice is

$$\alpha_{1,S} = \frac{\alpha}{2k} \binom{|S|}{p}^{-1}, \alpha_{2,S} = \alpha_{3,S} = \frac{\alpha}{4k} \binom{|S|}{p}^{-1}, \quad (9)$$

which puts as much weight to the comparison of the conditional variances ($F_{S,1}$) and the comparison the coefficients ($(F_{S,2}, F_{S,3})$). Similarly, for the collection $\hat{\mathcal{S}}_{\text{Lasso}}$, a natural choice is (9) with k replaced by $\lfloor (n_1 \wedge n_2)/2 \rfloor$. Alternatively, one can give a Bayesian flavor to the choice of the weights $\alpha_{i,S}, S \in \mathcal{S}$.

Denote by $T_{\hat{\mathcal{S}}}^B$ the multiple testing procedure associated with the collection of models $\hat{\mathcal{S}}$ and the weight sequence $\hat{\alpha}^B$. Proposition 2.3 shows that $T_{\hat{\mathcal{S}}}^B$ is a test of size α .

Proposition 2.3 (Size of $T_{\hat{\mathcal{S}}}^B$). *The test function $T_{\hat{\mathcal{S}}}^B$ satisfies $\mathbb{P}_{\mathcal{H}_0}[T_{\hat{\mathcal{S}}}^B < 0] \leq \alpha$.*

Remark 2.1 (Bonferroni correction on \mathcal{S} and not on $\hat{\mathcal{S}}$). *Note that even though we restrict ourselves to the collection $\hat{\mathcal{S}}$, the Bonferroni correction must be applied to the initial deterministic collection \mathcal{S} including $\hat{\mathcal{S}}$. Indeed, if we replace the condition (8) by the condition $\sum_{S \in \hat{\mathcal{S}}} \sum_{i=1}^3 \alpha_{i,S} \leq \alpha$, then the size of the corresponding is not constrained anymore to be smaller than α . This is due to the fact that we use the same data set to select $\hat{\mathcal{S}} \subset \mathcal{S}$ and to perform the multiple testing procedure. As a simple example, consider any deterministic collection \mathcal{S} and the data-driven collection*

$$\hat{\mathcal{S}} = \left\{ \arg \min_{S \in \mathcal{S}} \min_{i=1,2,3} \tilde{Q}_{i,|S|}(F_{S,i}|\mathbf{X}_S) \right\},$$

meaning that we select in $\widehat{\mathcal{S}}$ the subset S that minimizes the p -values of the parametric tests. Thus, computing $T_{\widehat{\mathcal{S}}}^B$ for this particular collection $\widehat{\mathcal{S}}$ is equivalent to performing a multiple testing procedure on \mathcal{S} .

If procedure $T_{\widehat{\mathcal{S}}}^B$ is computationally and conceptually simple, the size of the corresponding test can be much lower than α because of three difficulties:

1. Independently from our problem, Bonferroni corrections are known to be too conservative under dependence of the test statistics.
2. As emphasized by Remark 2.1, while the Bonferroni correction needs to be based on the whole collection \mathcal{S} , only the subsets $S \in \widehat{\mathcal{S}}$ are considered. Provided we could afford the computational cost of testing all subsets within \mathcal{S} , this loss cannot be compensated for if we use the Bonferroni correction.
3. As underlined in Subsection 2.1, for computational reasons, we do not consider in (7) the conditional p -value $Q_{2,|S|}(F_{S,2}|\mathbf{X}_S)$ and $Q_{3,|S|}(F_{S,3}|\mathbf{X}_S)$ but only upper bounds $\widetilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S)$ and $\widetilde{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S)$ of them. We therefore overestimate the type I error due to $F_{S,2}$ and $F_{S,3}$.

We address the three aforementioned issues applying a permutation approach.

Calibration by permutation (P). The collection of weights $\widehat{\alpha}^P = \{\alpha_{i,S}, S \in \mathcal{S}\}$ is chosen such that each $\alpha_{i,S}$ remains inversely proportional to $\binom{p}{|S|}$ in order to put all subsets sizes at equal footage. We also maintain an equal Bonferroni correction at the p -value level. In other words, we choose a collection of weights of the form

$$\alpha_{i,S} = \widehat{C}_i \binom{p}{|S|}^{-1}, \quad (10)$$

where \widehat{C}_i 's are calibrated by permutation to control the type I error of the global test.

Given a permutation π of the set $\{1, \dots, n_1 + n_2\}$, one gets \mathbf{Y}^π and \mathbf{X}^π by permuting the components of \mathbf{Y} and the rows of \mathbf{X} . This allows to us to get a new sample $(\mathbf{Y}^{\pi,(1)}, \mathbf{Y}^{\pi,(2)}, \mathbf{X}^{\pi,(1)}, \mathbf{X}^{\pi,(2)})$. Using this new sample, we compute a new collection $\widehat{\mathcal{S}}^\pi$ and parametric statistics $F_{S,1}^\pi, F_{S,2}^\pi, F_{S,3}^\pi$, respectively. We note \mathcal{P} the uniform distribution over the permutations of size $n_1 + n_2$.

We define \widehat{C}_1 as the $\alpha/2$ -quantiles with respect to \mathcal{P} of

$$\min_{S \in \widehat{\mathcal{S}}^\pi} \left\{ \widetilde{Q}_{1,|S|} (F_{S,1}^\pi | \mathbf{X}_S^\pi) \binom{p}{|S|} \right\}.$$

Similarly, $\widehat{C}_2 = \widehat{C}_3$ are the $\alpha/2$ -quantiles with respect to \mathcal{P} of

$$\min_{S \in \widehat{\mathcal{S}}^\pi} \left\{ \left(\widetilde{Q}_{2,|S|} (F_{S,2}^\pi | \mathbf{X}_S^\pi) \wedge \widetilde{Q}_{3,|S|} (F_{S,3}^\pi | \mathbf{X}_S^\pi) \right) \binom{p}{|S|} \right\}.$$

In practice, we estimate the quantiles \widehat{C}_i by sampling a large number N of permutations. Proposition 2.4 proves that the test $T_{\widehat{\mathcal{S}}}^P$ associated with the weight sequence $\widehat{\alpha}^P$ allows to control the type-I error rate at level α .

Proposition 2.4 (Size of $T_{\widehat{\mathcal{S}}}^P$). *The test function $T_{\widehat{\mathcal{S}}}^P$ satisfies*

$$\alpha/2 \leq \mathbb{P}_{\mathcal{H}_0} \left[T_{\widehat{\mathcal{S}}}^P < 0 \right] \leq \alpha.$$

Remark 2.2. Through the three constants \widehat{C}_1 , \widehat{C}_2 and \widehat{C}_3 , this permutation approach corrects simultaneously for the losses mentioned earlier due to the Bonferroni correction, in particular the restriction to a data-driven class \widehat{S} and the upper bounds of $\overline{Q}_{S,2}$ and $\overline{Q}_{S,3}$.

Yet, the level of $T_{\widehat{S}}^P$ is not exactly α because we treat separately the two p -values and apply a Bonferroni correction. It would be possible to calibrate all the statistics simultaneously in order to constrain the size of the corresponding test to be exactly α . However, this last approach would favor the statistic $F_{S,1}$ too much, because we would put on the same level the exact p -value $\overline{Q}_{S,1}$ and the upper bounds $\widetilde{Q}_{S,2}$ and $\widetilde{Q}_{S,3}$.

3. Power and Adaptation to Sparsity

Let us fix some number $\delta \in (0, 1)$. The objective is to investigate the set of parameters $(\beta^{(1)}, \sigma^{(1)}, \beta^{(2)}, \sigma^{(2)})$ that enforce the power of the test to exceed $1 - \delta$. We focus here on the Bonferroni calibration (B) procedure because the analysis is easier. Section 4 will illustrate that the permutation calibration (P) outperforms the Bonferroni calibration (B) in practice. In the sequel, $A \lesssim B$ (resp. $A \gtrsim B$) means that for some constant $L(\alpha, \delta)$ that only depends on α and δ , $A \leq L(\alpha, \delta)B$ (resp. $A \geq L(\alpha, \delta)B$).

3.1. Symmetrized Kullback-Leibler divergence

Intuitively, the test T_S^B should reject \mathcal{H}_0 with large probability when $(\beta^{(1)}, \sigma^{(1)})$ is far from $(\beta^{(2)}, \sigma^{(2)})$ in some sense. A classical way of measuring the divergence between two distributions is the Kullback-Leibler discrepancy. In the sequel, we note $\mathcal{K} [\mathbb{P}_{Y^{(1)}|X}; \mathbb{P}_{Y^{(2)}|X}]$ the Kullback discrepancy between the conditional distribution of $Y^{(1)}$ given $X^{(1)} = X$ and conditional distribution of $Y^{(2)}$ given $X^{(2)} = X$. Then, we denote \mathcal{K}_1 the expectation of this Kullback divergence when $X \sim \mathcal{N}(0_p, \Sigma^{(1)})$. Exchanging the roles of $\Sigma^{(1)}$ and $\Sigma^{(2)}$, we also define \mathcal{K}_2 :

$$\mathcal{K}_1 := \mathbb{E}_{X^{(1)}} \left\{ \mathcal{K} [\mathbb{P}_{Y^{(1)}|X}; \mathbb{P}_{Y^{(2)}|X}] \right\}, \quad \mathcal{K}_2 := \mathbb{E}_{X^{(2)}} \left\{ \mathcal{K} [\mathbb{P}_{Y^{(2)}|X}; \mathbb{P}_{Y^{(1)}|X}] \right\}.$$

The sum $\mathcal{K}_1 + \mathcal{K}_2$ forms a semidistance with respect to $(\beta^{(1)}, \sigma^{(1)})$ and $(\beta^{(2)}, \sigma^{(2)})$ as proved by the following decomposition

$$2(\mathcal{K}_1 + \mathcal{K}_2) = \left(\frac{\sigma^{(1)}}{\sigma^{(2)}} \right)^2 + \left(\frac{\sigma^{(2)}}{\sigma^{(1)}} \right)^2 - 2 + \frac{\|\beta^{(2)} - \beta^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} + \frac{\|\beta^{(2)} - \beta^{(1)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2}.$$

When $\Sigma^{(1)} \neq \Sigma^{(2)}$, we quantify the discrepancy between these covariance matrices by

$$\varphi_{\Sigma^{(1)}, \Sigma^{(2)}} := \varphi_{\max} \left\{ \sqrt{\Sigma^{(2)}}(\Sigma^{(1)})^{-1}\sqrt{\Sigma^{(2)}} + \sqrt{\Sigma^{(1)}}(\Sigma^{(2)})^{-1}\sqrt{\Sigma^{(1)}} \right\}.$$

3.2. Power for a deterministic collection \mathcal{S}

First, we control the power of T_S^B for a deterministic collection $\mathcal{S} = \mathcal{S}_{\leq k}$ (with some $k \leq (n_1 \wedge n_2)/2$) and the Bonferroni calibration weights $\widehat{\alpha}_{i,S}$ as in (9). Results for arbitrary deterministic collections \mathcal{S} are postponed to Section 5. For any $\beta \in \mathbb{R}^p$, $|\beta|_0$ refers to the size of its support and $|\beta|$ stands for the vector $(|\beta_i|), i = 1, \dots, p$. We consider the two following assumptions

A.1 : $\log(1/(\alpha\delta)) \lesssim n_1 \wedge n_2.$

A.2 : $|\beta^{(1)}|_0 + |\beta^{(2)}|_0 \lesssim k \wedge \left(\frac{n_1 \wedge n_2}{\log(p)} \right), \quad \log(p) \leq n_1 \wedge n_2.$

Remark 3.1. Condition **A.1** requires that the type I and type II errors under consideration are not exponentially smaller than the sample size. Condition **A.2** tells us that the number of non-zero components of $\beta^{(1)}$ and $\beta^{(2)}$ has to be smaller than $(n_1 \wedge n_2)/\log(p)$. This requirement has been shown [39] to be minimal to obtain fast rates of testing of the form (11) in the specific case $\beta^{(2)} = 0$, $\sigma^{(1)} = \sigma^{(2)}$ and $n_2 = \infty$.

Theorem 3.1 (Power of $T_{S_{\leq k}}^B$). Assuming that **A.1** and **A.2** hold, $\mathbb{P}[T_{S_{\leq k}}^B = 1] \geq 1 - \delta$ as long as

$$\mathcal{K}_1 + \mathcal{K}_2 \gtrsim \varphi_{\Sigma^{(1)}, \Sigma^{(2)}} \frac{\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \vee 1\} \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n_1 \wedge n_2}. \quad (11)$$

If we further assume that $\Sigma^{(1)} = \Sigma^{(2)} := \Sigma$, then $\mathbb{P}[T_{S_{\leq k}}^B = 1] \geq 1 - \delta$ as long as

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \gtrsim \frac{|\beta^{(1)} - \beta^{(2)}|_0 \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n_1 \wedge n_2}. \quad (12)$$

Remark 3.2. The condition $\Sigma^{(1)} = \Sigma^{(2)}$ is not necessary to control the power of $T_{S_{\leq k}}^B$ in terms of $|\beta^{(1)} - \beta^{(2)}|_0$ as in (12). However, the expression (12) would become far more involved.

Remark 3.3. Before assessing the optimality of Theorem 3.1, let us briefly compare the two rates of detection (11) and (12). According to (11), $T_{S_{\leq k}}^B$ is powerful as soon as the symmetrized Kullback distance is large compared $\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0\} \log(p) / (n_1 \wedge n_2)$. In contrast, (12) tells us that $T_{S_{\leq k}}^B$ is powerful when $\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2 / (\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}])$ is large compared to the sparsity of the difference: $|\beta^{(1)} - \beta^{(2)}|_0 \log(p) / (n_1 \wedge n_2)$.

When $\beta^{(1)}$ and $\beta^{(2)}$ have many non-zero coefficients in common, $|\beta^{(1)} - \beta^{(2)}|_0$ is much smaller than $|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0$. Furthermore, the left-hand side of (12) is of the same order as $\mathcal{K}_1 + \mathcal{K}_2$ when $\Sigma^{(1)} = \Sigma^{(2)}$, $\sigma^{(1)} = \sigma^{(2)}$ and $\|\beta^{(i)}\|_{\Sigma} / \sigma^{(i)} \lesssim 1$ for $i = 1, 2$, that is when the conditional variances are equal and when the signals $\|\beta^{(i)}\|_{\Sigma}$ are at most at the same order as the noises levels $\sigma^{(i)}$. In such a case, (12) outperforms (11) and only the sparsity of the difference $\beta^{(1)} - \beta^{(2)}$ plays a role in the detection rates. Below, we prove that both (11) and (12) are both optimal from a minimax point of view but on different sets.

Proposition 3.2 (Minimax lower bounds). Assume that $p \geq 5$, $\Sigma^{(1)} = \Sigma^{(2)} = I_p$, fix some $\gamma > 0$, and fix (α, δ) such that $\alpha + \delta < 53\%$. There exist two constants $L(\gamma)$ and $L'(\gamma)$ such that the following holds.

- For all $1 \leq s \leq p^{1/2-\gamma}$ **no** level- α test has a power larger than $1 - \delta$ simultaneously over all s -sparse vectors $(\beta^{(1)}, \beta^{(2)})$ satisfying **A.2** and

$$\mathcal{K}_1 + \mathcal{K}_2 \geq L(\alpha, \delta, \gamma) \frac{s}{n_1 \wedge n_2} \log(p). \quad (13)$$

- For all $1 \leq s \leq p^{1/2-\gamma}$, **no** level- α test has a power larger than $1 - \delta$ simultaneously over all sparse vectors $(\beta^{(1)}, \beta^{(2)})$ satisfying **A.2**, $|\beta^{(1)} - \beta^{(2)}|_0 \leq s$, $\|\beta^{(1)}\|_{I_p} \leq \sigma^{(1)}$, $\|\beta^{(2)}\|_{I_p} \leq \sigma^{(2)}$ and

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{I_p}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \geq L'(\alpha, \delta, \gamma) \frac{s}{n_1 \wedge n_2} \log(p). \quad (14)$$

The proof (in Section 8) is a straightforward application of minimax lower bounds obtained for the one-sample testing problem [41, 3].

Remark 3.4. Equation (11) together with (13) tells us that $T_{\mathcal{S}_{\leq k}}^B$ simultaneously achieves (up to a constant) the optimal rates of detection over s -sparse vectors $\beta^{(1)}$ and $\beta^{(2)}$ for all

$$s \lesssim k \wedge p^{1/2-\gamma} \wedge \frac{n_1 \wedge n_2}{\log(p)},$$

for any $\gamma > 0$. If the minimax lower bound is only proved for $\Sigma^{(1)} = \Sigma^{(2)} = I_p$, the detection rate (11) of $T_{\mathcal{S}_{\leq k}}^B$ is valid for any $(\Sigma^{(1)}, \Sigma^{(2)})$.

Remark 3.5. Equation (12) together with (14) tells us that $T_{\mathcal{S}_{\leq k}}^B$ simultaneously achieves (up to a constant) the optimal rates of detection over s -sparse differences $\beta^{(1)} - \beta^{(2)}$ satisfying $\frac{\|\beta^{(1)}\|_{\Sigma}}{\sigma^{(1)}} \vee \frac{\|\beta^{(2)}\|_{\Sigma}}{\sigma^{(2)}} \leq 1$ for all $s \lesssim k \wedge p^{1/2-\gamma} \wedge \frac{n_1 \wedge n_2}{\log(p)}$.

Remark 3.6 (Informal justification of the introduction of the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$). If we look at the proof of Theorem 3.1, we observe that the power (11) is achieved by the statistics $(F_{S_{\cup,1}}, F_{S_{\cup,2}}, F_{S_{\cup,3}})$ where S_{\cup} is the union of the support of $\beta^{(1)}$ and $\beta^{(2)}$. In contrast, (12) is achieved by the statistics $(F_{S_{\Delta,1}}, F_{S_{\Delta,2}}, F_{S_{\Delta,3}})$ where S_{Δ} is the support of $\beta^{(1)} - \beta^{(2)}$. Intuitively, the idea underlying the collection $\widehat{\mathcal{S}}_L^{(1)}$ in the definition (6) of $\widehat{\mathcal{S}}_{\text{Lasso}}$ is to estimate S_{\cup} , while the idea underlying the collection $\widehat{\mathcal{S}}_L^{(2)}$ is to estimate S_{Δ} .

3.3. Power of $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B$

For the sake of simplicity, we restrict here to the case $n_1 = n_2 := n$, more general results being postponed to Section 5. The test $T_{\mathcal{S}_{\leq n/2}}^B$ is computationally expensive (non polynomial with respect to p). The collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ has been introduced to fix this burden. We consider $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B$ with the prescribed Bonferroni calibration weights $\widehat{\alpha}_{i,S}$ (as in (9) with k replaced by $\lfloor (n_1 \wedge n_2)/2 \rfloor$). In the statements below, $\psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(1)}, \psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(2)}, \dots$ refer to positive quantities that only depend on the largest and the smallest eigenvalues of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. Consider the additional assumptions

$$\mathbf{A.3} : \quad |\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \lesssim \psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(1)} \frac{n}{\log(p)}.$$

$$\mathbf{A.4} : \quad |\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \lesssim \psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(2)} \sqrt{\frac{n}{\log(p)}}.$$

Theorem 3.3. Assuming that **A.1** and **A.3** hold, we have $\mathbb{P}[T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B = 1] \geq 1 - \delta$ as long as

$$\mathcal{K}_1 + \mathcal{K}_2 \gtrsim \psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(3)} \frac{\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \vee 1\} \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n}. \quad (15)$$

If $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$ and if **A.1** and **A.4** hold, then $\mathbb{P}[T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B = 1] \geq 1 - \delta$ as long as

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \gtrsim \psi_{\Sigma, \Sigma}^{(4)} \frac{|\beta^{(1)} - \beta^{(2)}|_0 \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n}. \quad (16)$$

Remark 3.7. The rates of detection (15) and the sparsity condition **A.3** are analogous to (11) and Condition **A.2** in Theorem 3.1 for $T_{\mathcal{S}_{\leq (n_1 \wedge n_2)/2}}^B$. The second result (16) is also similar to (12). As a consequence, $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B$ is minimax adaptive to the sparsity of $(\beta_1, \beta^{(2)})$ and of $\beta^{(1)} - \beta^{(2)}$.

Remark 3.8. Dependencies of **A.3**, **A.4**, (15) and (16) on $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are unavoidable because the collection $\widehat{\mathcal{S}}_{Lasso}$ is based on the Lasso estimator which require design assumptions to work well [9]. Nevertheless, one can improve all these dependencies using restricted eigenvalues instead of largest eigenvalues. This and other extensions are considered in Section 5.

4. Numerical Experiments

This section evaluates the performances of the suggested test statistics along with aforementioned test collections and calibrations on simulated linear regression datasets. An illustration of how to adapt this testing strategy to the global comparison of Gaussian graphical models inferred from real transcriptomic data is given in the second part.

4.1. Synthetic Linear Regression Data

In order to calibrate the difficulty of the testing task, we simulate our data according to the rare and weak parametrization adopted in [3]. We choose a large but still reasonable number of variables $p = 200$, and restrict ourselves to cases where the number of observations $n = n_1 = n_2$ in each sample remains smaller than p . The sparsity of sample-specific coefficients $\beta^{(1)}$ and $\beta^{(2)}$ is parametrized by the number of non zero common coefficients $p^{1-\eta}$ and the number of non zero coefficients $p^{1-\eta_2}$ which are specific to $\beta^{(2)}$. The magnitude μ_r of all non zero coefficients is set to a common value of $\sqrt{2r \log p}$, where we let the magnitude parameter range from $r = 0$ to $r = 0.5$:

$$\begin{aligned} \beta^{(1)} &= \begin{pmatrix} \underbrace{\mu_r \ \mu_r \ \dots \ \mu_r}_{p^{1-\eta} \text{ common coefficients}} & \mathbf{0} \dots \mathbf{0} & \mathbf{0} \dots \mathbf{0} \end{pmatrix} \\ \beta^{(2)} &= \begin{pmatrix} \underbrace{\mu_r \ \mu_r \ \dots \ \mu_r}_{p^{1-\eta} \text{ common coefficients}} & \underbrace{\mu_r \ \dots \ \mu_r}_{p^{1-\eta_2} \text{ sample-2-specific coefficients}} & \mathbf{0} \dots \mathbf{0} \end{pmatrix} \end{aligned}$$

We consider three sample sizes $n = 25, 50, 100$, and generate two sub-samples of equal size $n_1 = n_2 = n$ according to the following sample specific linear regression models:

$$\begin{cases} \mathbf{Y}^{(1)} &= \mathbf{X}^{(1)}\beta^{(1)} + \varepsilon^{(1)}, \\ \mathbf{Y}^{(2)} &= \mathbf{X}^{(2)}\beta^{(2)} + \varepsilon^{(2)}. \end{cases}$$

Design matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are generated by multivariate Gaussian distributions, $\mathbf{X}_i^{(j)} \sim \mathcal{N}(0, \Sigma^{(j)})$ with varying choices of $\Sigma^{(j)}$, as detailed below. Noise components $\varepsilon_i^{(1)}$ and $\varepsilon_i^{(2)}$ are generated independantly from $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ according to a standard centered Gaussian distribution.

The next two paragraphs detail the different design scenarios under study as well as test statistics, collections and calibrations in competition. Each experiment is repeated 1000 times.

Design Scenarios Under Study.

Sparsity Patterns. We study six different sparsity patterns as summarized in Table 1. The first two are meant to validate type I error control. The last four allow us to compare the performances of the various test statistics, collections and calibrations under different sparsity levels and proportions of shared coefficients. In all cases, the choices of sparsity parameters η and η_2 lead to strong to very strong levels of sparsity. The last column of Table 1 illustrates the signal sparsity patterns of $\beta^{(1)}$ and $\beta^{(2)}$ associated with each scenario. In scenarios 1 and 2, sample-specific signals share little, if not none, non zero

coefficient. In scenarios 3 and 4, sample-specific coefficients show some overlap. Scenario 4 is the most difficult one since the number of sample-2-specific coefficients is much smaller than the number of common non zero coefficients: the sparsity of the difference between $\beta^{(1)}$ and $\beta^{(2)}$ is much smaller than the global sparsity of $\beta^{(2)}$. This explains why the illustration in the last column might be misleading: the two patterns are not equal but do actually differ by only one covariate.

Beyond those six varying sparsity patterns, we consider three different correlation structures $\Sigma^{(1)}$ and $\Sigma^{(2)}$ for the generation of the design matrix. In all three cases, we assume that $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$. On top of the basic orthogonal matrix $\Sigma^{(1)} = \Sigma^{(2)} = I_p$, we investigate two randomly generated correlation structures.

Power Decay Correlation Structure. First, we consider a power decay correlation structure such that $\Sigma_{i,j} = \rho^{|i-j|}$. Since the sparsity pattern of $\beta^{(1)}$ and $\beta^{(2)}$ is linked to the order of the covariates, we randomly permute at each run the columns and rows of Σ in order to make sure the correlation structure is independent from the sparsity pattern.

Gaussian Graphical Model Structure. Second, we simulate correlation structures corresponding to a Gaussian graphical model with an affiliation structure between three clusters, as generated by the `GGMselect` R package. A new structure is generated at each run.

Both random correlation structures are calibrated such that, on average, each covariate is correlated with 10 other covariates with correlations above 0.2 in absolute value. This corresponds to fixing ρ at a value of 0.75 in the power decay correlation structure and the intra-cluster connectivity coefficient to 5% in the Gaussian graphical model structure.

Test statistics, collections and calibrations in competition In the following, we present the results of the proposed test statistics combined with two test collections, namely a deterministic and data-driven model collection, respectively \mathcal{S}_1 and $\widehat{\mathcal{S}}_{\text{Lasso}}$, as well as with a Bonferroni (**B**) or Permutation (**P**) calibration (computed with 100 random permutations).

Furthermore, to put those results in perspective, we compare the suggested test statistic to the usual likelihood ratio statistic for the equality of $\beta_S^{(1)}$ and $\beta_S^{(2)}$, which follows a Fisher distribution with $|S|$ and $n_1 + n_2 - 2|S|$ degrees of freedom, for a given support $|S|$ of reduced dimension:

$$Fis_S = \frac{\|\mathbf{Y} - \mathbf{X}_S \widehat{\beta}_S\|^2 - \|\mathbf{Y}^{(1)} - \mathbf{X}_S^{(1)} \widehat{\beta}_S^{(1)}\|^2 - \|\mathbf{Y}^{(2)} - \mathbf{X}_S^{(2)} \widehat{\beta}_S^{(2)}\|^2}{\|\mathbf{Y}^{(1)} - \mathbf{X}_S^{(1)} \widehat{\beta}_S^{(1)}\|^2 + \|\mathbf{Y}^{(2)} - \mathbf{X}_S^{(2)} \widehat{\beta}_S^{(2)}\|^2} \frac{n_1 + n_2 - 2|S|}{|S|}, \quad (17)$$

where $\widehat{\beta}_S$ is the maximum likelihood estimator restricted to covariates in support S on the concatenated sample (\mathbf{X}, \mathbf{Y}) . If this statistic Fis_S is able to detect differences between $\beta^{(1)}$ and $\beta^{(2)}$, it is not really suited for detecting differences between the standard deviatons $\sigma^{(1)}$ and $\sigma^{(2)}$.

The Fisher statistic Fis_S is adapted to the high-dimensional framework similarly as the suggested statistics $(F_{S,1}, F_{S,2}, F_{S,3})$, except that exact p -values are available. The corresponding test with a collection $\widehat{\mathcal{S}}$ and a Bonferroni (resp. permutation) calibration is denoted $T_{\widehat{\mathcal{S}}}^{B,\text{Fisher}}$ ($T_{\widehat{\mathcal{S}}}^{P,\text{Fisher}}$).

Validation of Type I Error Control










Setting	η	# common	η_2	# $\beta^{(2)}$ specific	Signals
\mathcal{H}_{00}	-	0	-	0	$\beta^{(1)}$ _____
					$\beta^{(2)}$ _____
\mathcal{H}_0	5/8	7	-	0	$\beta^{(1)}$ 
					$\beta^{(2)}$ 
1	-	0	5/8	7	$\beta^{(1)}$ _____
					$\beta^{(2)}$ 
2	7/8	1	5/8	7	$\beta^{(1)}$ 
					$\beta^{(2)}$ 
3	5/8	7	5/8	7	$\beta^{(1)}$ 
					$\beta^{(2)}$ 
4	5/8	7	7/8	1	$\beta^{(1)}$ 
					$\beta^{(2)}$ 

TABLE 1

Summary of the six different sparsity patterns under study.

Control Under the Global Null Hypothesis \mathcal{H}_{00} . Table 2 presents level checks under a restricted null hypothesis \mathcal{H}_{00} , such that $\beta^{(1)} = \beta^{(2)} = 0$, along with 95% Gaussian confidence intervals.

As expected, the Bonferroni calibration combined with the majoration of quantiles or data-driven test collections is, by far, much too conservative. Even with the likelihood ratio statistic, for which we know the exact p -value, it is unthinkable to use Bonferroni calibration as soon as we adopt data-driven test collections instead of deterministic ones.

Control Under the Global Equality of Non Null Coefficients \mathcal{H}_0 . Figures 1 and 2 present level checks under \mathcal{H}_0 but with non null $\beta^{(1)} = \beta^{(2)} \neq 0$, under respectively orthogonal and non-orthogonal correlation structures. Conclusions are perfectly similar to the case \mathcal{H}_{00} : all methods behave well, except the Bonferroni calibration which is as conservative as expected for $T_{\hat{\Sigma}}^*$ (using any test collection) and for $T_{\hat{\Sigma}}^{*,\text{Fisher}}$ as soon as we use the data-driven test collection $\hat{\Sigma}_{\text{Lasso}}$ instead of the deterministic collection \mathcal{S}_1 .

Power Analysis. Figure 3 represents power performances for the suggested test $T_{\hat{\Sigma}}^*$ and the usual likelihood ratio test $T_{\hat{\Sigma}}^{*,\text{Fisher}}$ combined with either \mathcal{S}_1 or $\hat{\Sigma}_{\text{Lasso}}$ test collections using a calibration by permutation under an orthogonal covariance matrix Σ . Figure 4 represents equivalent results for power decay and GGM covariance structures.

In the absence of common coefficients (scenarios 1 and 2), the suggested test $T_{\hat{\Sigma}}^*$ reaches 100% power from very low signal magnitudes and small sample sizes. Compared to the test based on usual likelihood ratio statistics, which does not reach more than 40% power when $n = 25$ given the signal magnitudes under consideration, the suggested statistics proves itself extremely efficient. Under these settings as well, any subset of size 1 containing one

Model collection	\mathcal{S}_1		$\widehat{\mathcal{S}}_{Lasso}$	
	(B)	(P)	(B)	(P)
Calibration				
n= 25	1 ± 0.6	6.9 ± 1.6	0 ± 0	6.9 ± 1.6
n= 50	1.8 ± 0.8	5.8 ± 1.4	0 ± 0	6 ± 1.5
n= 100	1.0 ± 0.6	7.4 ± 1.6	0.1 ± 0.2	7.4 ± 1.6

(a) Tests $T_{\mathcal{S}}^*$

Model collection	\mathcal{S}_1		$\widehat{\mathcal{S}}_{Lasso}$	
	(B)	(P)	(B)	(P)
Calibration				
n= 25	5.5 ± 1.4	6.8 ± 1.6	0.5 ± 0.4	6.5 ± 1.5
n= 50	4.5 ± 1.3	5.5 ± 1.4	0.1 ± 0.2	5.3 ± 1.4
n= 100	4.8 ± 1.3	0.1 ± 1.5	6.6 ± 0.2	6.5 ± 1.5

(b) Tests $T_{\mathcal{S}}^{*,Fisher}$

TABLE 2

Estimated test levels in percentage along with 95% Gaussian confidence interval (in percentage) under \mathcal{H}_0 for the seven different strategies, based upon 1000 simulations.

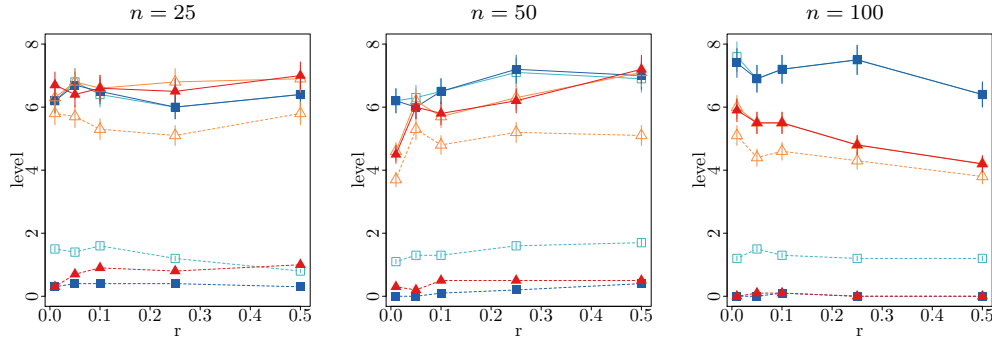


FIGURE 1. Estimated test levels in percentage under \mathcal{H}_0 for the six different strategies for varying magnitudes of common non null coefficients, based upon 1000 simulations. Bonferroni calibration in dotted lines, calibration by permutation in plain lines. Blue squares represent the suggested test $T_{\mathcal{S}}^*$, red triangles stand for the Fisher test $T_{\mathcal{S}}^{*,Fisher}$. The deterministic collection \mathcal{S}_1 is drawn in plain points, while the data-driven collection $\widehat{\mathcal{S}}_{Lasso}$ is in empty points.

of the variables activated in only $\beta^{(2)}$ can suffice to reject the null, which is why collection \mathcal{S}_1 performs actually very well when associated with $(F_{S,1}, F_{S,2}, F_{S,3})$ and not so badly when associated with F_{i_S} .

However, in more complex settings 3 and 4, where larger subsets are required to correct for strong and numerous common effects, subset collection $\widehat{\mathcal{S}}_{Lasso}$ performs much better than the collection \mathcal{S}_1 .

Figure 4 provide similar results under respectively power decay correlated designs and GGM-like correlated designs for a sample size of $n = 50$, leading to similar conclusions as in the uncorrelated case.

4.2. Real Transcriptomic Breast Cancer Data

The procedure developed in Section 2 can be adapted to the case of Gaussian graphical models as in [40]. Adopting a neighborhood selection approach, estimation of the Gaussian

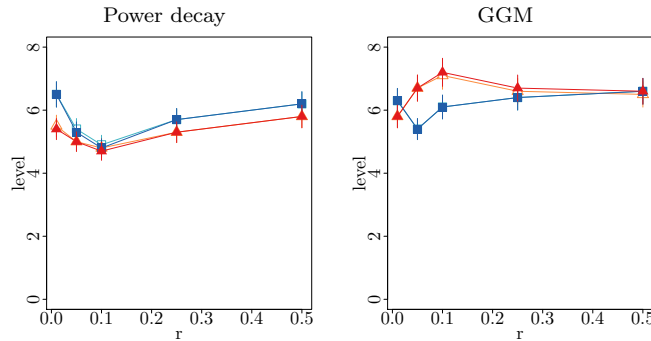


FIGURE 2. Estimated test levels in percentage under \mathcal{H}_0 for the 4 strategies calibrated by permutation for varying magnitudes of common non null coefficients, based upon 1000 simulations, under power decay and GGM correlation structures when $n = 50$. Blue squares represent the suggested test $T_{\hat{S}}^*$, red triangles stand for the Fisher test $T_{\hat{S}}^{*,\text{Fisher}}$. The deterministic collection \mathcal{S}_1 is drawn in plain points, while the data-driven collection $\hat{\mathcal{S}}_{\text{Lasso}}$ is in empty points.

graphical model amounts to the estimation of p independent linear regressions. Therefore, the idea is to run for each gene in the network a neighborhood test conducted at level α/p , thereby correcting for multiple testing.

We apply this strategy to the full (training and validation) breast cancer dataset studied by [18] and [29], whose training subset was originally published in [31]. The full dataset consists of microarray gene expression profiles from 133 patients with stage I-III breast cancer undergoing preoperative chemotherapy. A majority of patients ($n=99$) presented residual disease (RD), while 34 patients demonstrated a pathologic complete response (pCR). The common objective of [18] and [29] was to develop a predictor of complete response to treatment from gene expression profiling. In particular, [18] identified an optimal predictive subset of 30 probes, mapping to 26 distinct genes.

[1] inferred Gaussian graphical models among those 26 genes on each patient class using weighted neighborhood selection. The corresponding graphs of conditional dependencies for medium regularization are presented in Figure 5. Those two graphs happen to differ dramatically from one another. The question we tackle is whether those differences remain when taking into account estimation uncertainties.

We run for each of the 26 genes a neighborhood test $T_{\hat{\mathcal{S}}_{\text{Lasso}}}^P$ at level $0.05/26$. We associate to each neighborhood test a p -value computed as the fraction of the 10000 permutation values of the statistic that are less than the observed test statistic.

Most of the graph estimation methods proposed in the literature, such as the procedure of [1] leading to Figure 5, rely on the assumption that observations are i.i.d. Yet the training and validation datasets have been collected and analysed separately by two different clinical centers. We therefore start by checking whether the pooled sample can be considered as homogeneous. Within each group of patients (RD and pCR), we lead a test for the homogeneity of Gaussian graphical models between the training and validation subsets.

Within pCR patients (3), two neighborhood tests corresponding to CA12 and PDGFRA are rejected at level $0.05/26$. Within RD patients (4), half of the neighborhoods happen to differ significantly between the training and validation datasets. Genes CA12 and JMJD2B are responsible for the rejection of respectively seven and six neighborhoods.

Because of these surprisingly significant divergences between training and validation subsets, we restrict the subsequent analysis to the training set ($n=82$ patients, among which 61 RD and 21 pCR patients).

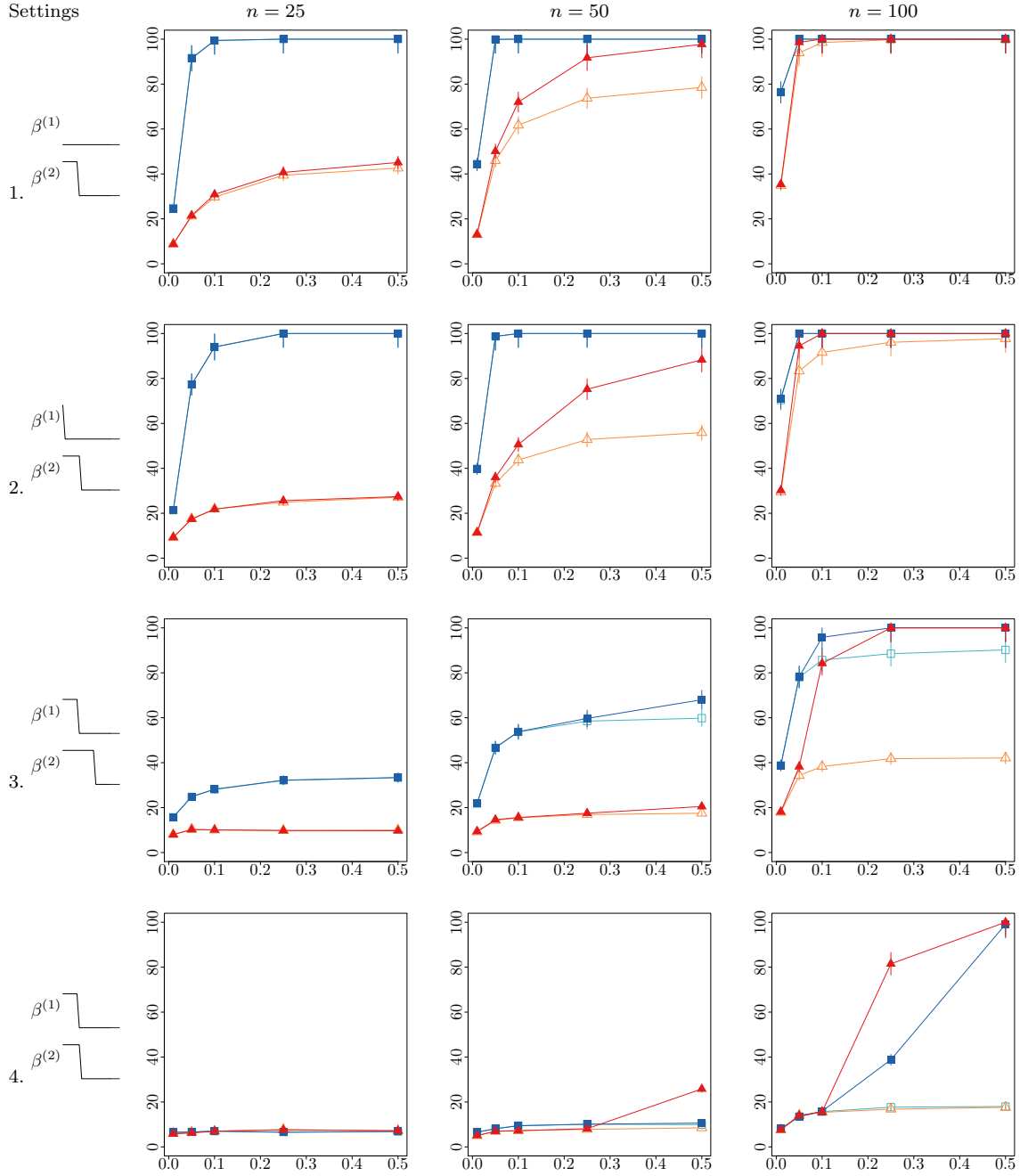


FIGURE 3. Power (in percentage) as a function of signal magnitude parameter r for the suggested test $T_{\hat{S}}^*$ and the test $T_{\hat{S}}^{*,\text{Fisher}}$ based on the likelihood ratio, combined with \mathcal{S}_1 or $\hat{\mathcal{S}}_{Lasso}$ test collections and a calibration by permutation, for various sparsity pattern under the assumption of uncorrelated designs $\Sigma^{(1)} = \Sigma^{(2)} = I_p$. Blue squares represent the suggested test $T_{\hat{S}}^*$, red triangles stand for the Fisher test $T_{\hat{S}}^{*,\text{Fisher}}$. The deterministic collection \mathcal{S}_1 is drawn in plain points, while the data-driven collection $\hat{\mathcal{S}}_{Lasso}$ is in empty points.

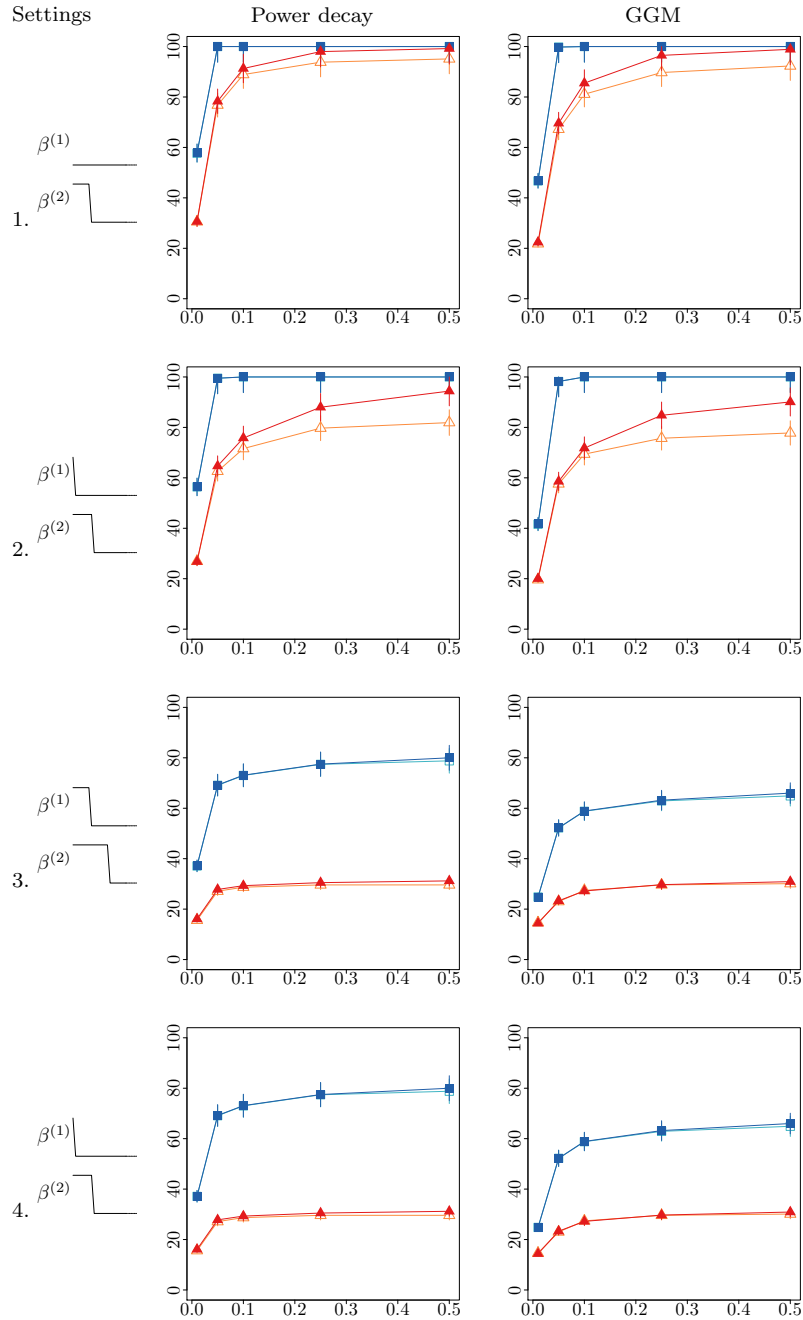


FIGURE 4. Power (in percentage) as a function of signal magnitude parameter r for the suggested test $T_{\hat{S}}^*$ and the test $T_{\hat{S}}^{*,\text{Fisher}}$ based on the likelihood ratio, combined with S_1 or \hat{S}_{Lasso} test collections and a calibration by permutation, for various sparsity patterns under power decay and GGM correlated designs, at $n = 50$ observations. Blue squares represent the suggested test $T_{\hat{S}}^*$, red triangles stand for the Fisher test $T_{\hat{S}}^{*,\text{Fisher}}$. The deterministic collection S_1 is drawn in plain points, while the data-driven collection \hat{S}_{Lasso} is in empty points.

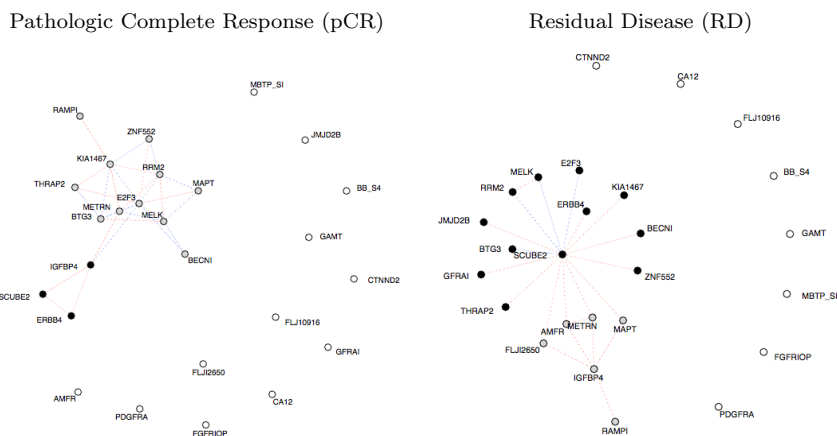


FIGURE 5. Graphs of conditional dependencies among the 26 genes selected by [18] on patients with pathologic complete response or residual disease with medium regularization as presented in Figure 3 of [1].

	AMFR	BB_S4	BECN1	BTG3	CA12	CTNND2	E2F3
decision	0	0	0	0	1	0	0
<i>p</i> -value	0.0492	0.0072	0.1972	1	0.0018	0.0100	0.1080
	ERBB4	FGFRIOP	FLJ10916	FLJ2650	GAMT	GFRAI	IGFBP4
decision 0	0	0	0	0	0	0	0
<i>p</i> -value	0.5610	0.0242	0.2542	0.0312	0.1158	0.5318	0.0458
	JMJD2B	KIA1467	MAPT	MBTP_SI	MELK	METRN	PDGFRA
decision	0	0	0	0	0	0	1
<i>p</i> -value	0.0128	0.0272	0.0178	0.0062	0.5602	1	0.0012
	RAMP1	RRM2	SCUBE2	THRAP2	ZNF552		
decision	0	0	0	0	0		
<i>p</i> -value	0.0444	0.0022	0.2372	0.0228	0.0028		

TABLE 3

Homogeneity test between training and test samples among pCR patients. Summary of test decisions after Bonferroni multiple testing correction and *p*-values for each neighborhood test. The suggested statistic is combined with a data-driven model collection 10000 permutations. The *p*-value is computed as the fraction of the permutation values of the statistic that are less than the observed test statistic.

To roughly check that we got rid of the underlying heterogeneity, we create an artificial dataset under H_0 by permutation of the patients, regardless of their class. No neighborhood test is rejected at a level corrected for multiple testing. We also cut the group of patients with residual disease artificially in half. When testing for the difference between the two halves, no significant heterogeneity remains, whatever the neighborhood.

Within the training set, the comparison of Gaussian graphical structures between pCR and RD patients leads to the rejection of all neighborhood tests after Bonferroni correction for multiple testing of the 26 neighborhoods, as summarized in Table 5. RRM2, MAPT and MELK genes appear as responsible for the rejection of respectively nine, nine and four of these neighborhood tests. Quite interestingly, these three genes have all been described in clinical literature as new promising drug targets. [17] exhibited inhibitors of RRM2 expression, which reduced *in vitro* and *in vivo* cell proliferation. [32] led functional biology experiments validating the relationship between MAPT expression levels and response to therapy, suggesting to inhibit its expression to increase sensitivity to treatment. More recently, [12] developed a therapeutic candidate inhibiting MELK expression that was

	AMFR	BB_S4	BECNI	BTG3	CA12	CTNND2	E2F3
decision	0	1	1	0	1	0	0
p-value	0.0046	<0.0001	<0.0001	0.0202	<0.0001	0.0684	0.0428
	ERBB4	FGFRIOP	FLJ10916	FLJI2650	GAMT	GFRAI	IGFBP4
decision	0	1	1	1	1	0	0
p-value	0.26	<0.0001	<0.0001	0.002	<0.0001	0.3606	0.389
	JMJD2B	KIA1467	MAPT	MBTP_SI	MELK	METRN	PDGFRA
decision	1	1	0	1	0	0	1
p-value	<0.0001	2e-04	0.006	6e-04	0.1556	0.1054	<0.0001
	RAMP1	RRM2	SCUBE2	THRAP2	ZNF552		
decision	0	0	0	1	1		
pvalue23	0.2288	0.2988	0.3552	<0.0001	<0.0001		

TABLE 4

Homogeneity test between training and test samples among RD patients. Summary of test decisions after Bonferroni multiple testing correction and p-values for each neighborhood test. The suggested statistic is combined with a data-driven model collection 10000 permutations. The p-value is computed as the fraction of the permutation values of the statistic that are less than the observed test statistic.

proved to suppress the growth of tumour-initiating cells in mice with various cancer types, including breast cancer.

	AMFR	BB_S4	BECNI	BTG3	CA12	CTNND2	E2F3
decision	1	1	1	1	1	1	1
p-value	<0.0001	<0.0001	4e-04	<0.0001	2e-04	<0.0001	<0.0001
rejected model	RRM2	RRM2	MAPT	MAPT	MAPT	RRM2	MAPT
	ERBB4	FGFRIOP	FLJ10916	FLJI2650	GAMT	GFRAI	IGFBP4
decision	1	1	1	1	1	1	1
p-value	<0.0001	4e-04	4e-04	<0.0001	<0.0001	<0.0001	<0.0001
rejected model	MELK	MAPT	RRM2	MAPT	RRM2	BTG3	MELK
	JMJD2B	KIA1467	MAPT	MBTP_SI	MELK	METRN	PDGFRA
decision	1	1	1	1	1	1	1
p-value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
rejected model	MAPT	MELK	RRM2	E2F3	MAPT	MELK	RRM2
	RAMP1	RRM2	SCUBE2	THRAP2	ZNF552		
decision	1	1	1	1	1		
p-value	<0.0001	<0.0001	<0.0001	<0.0001	2e-04		
rejected model	RRM2	MAPT	BTG3	E2F3	RRM2		

TABLE 5

Summary of neighborhood tests between RD and pCR patients within the training set ($n=82$). Decision is made at level 0.05/26 to correct for multiple testing. The suggested statistic is combined with a data-driven model collection 10000 permutations. The p-value is computed as the fraction of the permutation values of the statistic that are less than the observed test statistic.

For comprehensiveness, we add that similar analysis of the validation set ($n=51$ patients, among which 38 RD and 13 pCR patients) leads to the identification of only 9 significantly altered neighborhoods between pCR and RD patients 6. This difference in the number of significantly altered neighborhoods can be explained by the reduced size of the sample. Yet, genes responsible for the rejection of the tests differ from those identified on the training set. In particular, five of the significant tests are rejected because of SCUBE2, which has been recently recognised as a novel tumor suppressor gene [24].

	AMFR	BB_S4	BECNI	BTG3	CA12	CTNND2	E2F3
decision	0	0	0	1	0	0	1
p-value	0.0024	0.0028	0.0048	0.0018	0.0028	0.0082	<0.0001
rejected model	-	-	-	SCUBE2	-	-	METR
	ERBB4	FGFRIOP	FLJ10916	FLJI2650	GAMT	GFRAI	IGFBP4
decision	1	0	1	0	0	1	1
p-value	0.0014	0.0072	8e-04	0.0142	0.0046	8e-04	2e-04
rejected model	SCUBE2	-	SCUBE2	-	-	E2F3	SCUBE2
	JMJD2B	KIA1467	MAPT	MBTP_SI	MELK	METR	PDGFRA
decision	0	1	0	0	0	1	0
p-value	0.0054	0.0018	0.0032	0.0078	0.0036	4e-04	0.0104
rejected model	-	SCUBE2	-	-	-	E2F3	-
	RAMPI	RRM2	SCUBE2	THRAP2	ZNF552		
decision	0	0	1	0	0		
p-value	0.0056	0.0034	2e-04	0.0024	0.006		
rejected model	-	-	FLJ10916	-	-		

TABLE 6

Summary of neighborhood tests between RD and pCR patients within the validation set ($n=51$). Decision is made at level 0.05/26 to correct for multiple testing. The suggested statistic is combined with a data-driven model collection 10000 permutations. The p-value is computed as the fraction of the permutation values of the statistic that are less than the observed test statistic.

5. Additional results

This section provides two additional results to the power analysis of Section 3. Theorem 5.1 extends Theorem 3.1 from deterministic collections of the form $\mathcal{S}_{\leq k}$ to any deterministic collection \mathcal{S} , unveiling a bias/variance-like trade-off linked to the cardinality S of collection \mathcal{S} . In a second part, Propositions 5.2 and 5.3 explicit the dependency of the constants in Theorem 3.3 on $\Sigma^{(1)}$ and $\Sigma^{(2)}$ through largest and smallest sparse eigenvalues and compatibility constants.

5.1. Power of $T_{\mathcal{S}}^B$

Let us provide a general analysis of the power of $T_{\mathcal{S}}^B$ for arbitrary deterministic collections \mathcal{S} . To do so, we need to consider the Kullback discrepancy between the conditional distribution of $Y^{(1)}$ given $X_S^{(1)} = X_S$ and the conditional distribution of $Y^{(2)}$ given $X_S^{(2)} = X_S$, which we denote $\mathcal{K} [\mathbb{P}_{Y^{(1)}|X_S}; \mathbb{P}_{Y^{(2)}|X_S}]$. For short, we respectively note $\mathcal{K}_1(S)$ and $\mathcal{K}_2(S)$

$$\begin{aligned} \mathcal{K}_1(S) &:= \mathbb{E}_{X_S^{(1)}} \left\{ \mathcal{K} [\mathbb{P}_{Y^{(1)}|X_S}; \mathbb{P}_{Y^{(2)}|X_S}] \right\} , \\ \mathcal{K}_2(S) &:= \mathbb{E}_{X_S^{(2)}} \left\{ \mathcal{K} [\mathbb{P}_{Y^{(2)}|X_S}; \mathbb{P}_{Y^{(1)}|X_S}] \right\} . \end{aligned}$$

Intuitively, $\mathcal{K}_1(S) + \mathcal{K}_2(S)$ corresponds to some distance between the regression of $Y^{(1)}$ given $X_S^{(1)}$ and of $Y^{(2)}$ given $X_S^{(2)}$. Noting $\Sigma_S^{(1)}$ (resp. $\Sigma_S^{(2)}$) the restriction of $\Sigma^{(1)}$ (resp. $\Sigma^{(2)}$) to indices in S , we define

$$\varphi_S := \varphi_{\max} \left\{ \sqrt{\Sigma_S^{(2)}} (\Sigma_S^{(1)})^{-1} \sqrt{\Sigma_S^{(2)}} + \sqrt{\Sigma_S^{(1)}} (\Sigma_S^{(2)})^{-1} \sqrt{\Sigma_S^{(1)}} \right\} . \quad (18)$$

Theorem 5.1 (Power of $T_{\mathcal{S}}^B$ for any deterministic \mathcal{S}). *For any $S \in \mathcal{S}$, we note $\alpha_S = \min_{i=1,2,3} \alpha_{i,S}$. The power of $T_{\mathcal{S}}^B$ is larger than $1 - \delta$ as long as there exists $S \in \mathcal{S}$ such that $|S| \lesssim n_1 \wedge n_2$ and*

$$1 + \log[1/(\delta\alpha_S)] \lesssim n_1 \wedge n_2 , \quad (19)$$

and

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \gtrsim \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{1}{\alpha_S \delta} \right) \right]. \quad (20)$$

Remark 5.1. Let us note $\Delta(S)$ the right hand side of (20). According to Theorem 5.1, The term $\Delta(S)$ plays the role of a variance term and therefore increases with the cardinality S . Furthermore, the term $\mathcal{K}_1 - \mathcal{K}_1(S) + \mathcal{K}_2 - \mathcal{K}_2(S)$ plays the role of a bias. Let us note \mathcal{S}^* the subcollection of \mathcal{S} made of sets S satisfying (19). According to theorem 5.1, $T_{\mathcal{S}}^B$ is powerful as long as $\mathcal{K}_1 + \mathcal{K}_2$ is larger (up to constants)

$$\inf_{S \in \mathcal{S}^*} \{ \mathcal{K}_1 - \mathcal{K}_1(S) + \mathcal{K}_2 - \mathcal{K}_2(S) \} + \Delta(S) \quad (21)$$

Such a result is comparable to oracle inequalities obtained in estimation since the test $T_{\mathcal{S}}^B$ is powerful when the Kullback loss $\mathcal{K}_1 + \mathcal{K}_2$ is larger than the trade-off (21) between a bias-like term and a variance-like term without requiring the knowledge of this trade-off in advance. We refer to [5] for a thorough comparison between oracle inequalities in model selection and second type error terms of this form.

5.2. Sharper analysis of $T_{\mathcal{S}_{Lasso}}^B$

Given a matrix \mathbf{X} , an integer k , and a number M , one respectively defines the largest and smallest eigenvalues of order k , the compatibility constants $\kappa[M, k, \mathbf{X}]$ and $\eta[M, k, \mathbf{X}]$ (see [38]) by

$$\begin{aligned} \Phi_{k,+}(\mathbf{X}) &= \sup_{\theta, 1 \leq |\theta|_0 \leq k} \frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2}, & \Phi_{k,-}(\mathbf{X}) &= \inf_{\theta, 1 \leq |\theta|_0 \leq k} \frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2}, \\ \kappa[M, k, \mathbf{X}] &= \min_{T, \theta: |T| \leq k, \theta \in \mathcal{C}(M, T)} \left\{ \frac{\|\mathbf{X}\theta\|}{\|\theta\|} \right\}, \\ \eta[M, k, \mathbf{X}] &= \min_{T, \theta: |T| \leq k, \theta \in \mathcal{C}(M, T)} \left\{ \sqrt{k} \frac{\|\mathbf{X}\theta\|}{|\theta|_1} \right\}, \end{aligned} \quad (22)$$

where $\mathcal{C}(M, T) = \{ \theta : |\theta_{T^c}|_1 < M |\theta_T|_1 \}$. Given an integer k , define

$$\begin{aligned} \gamma_{\Sigma^{(1)}, \Sigma^{(2)}, k} &:= \frac{\bigwedge_{i=1,2} \kappa^2 \left[\mathfrak{G}, k_*, \sqrt{\Sigma^{(i)}} \right]}{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}, \\ \gamma'_{\Sigma^{(1)}, \Sigma^{(2)}, k} &:= \frac{\bigvee_{i=1,2} \Phi_{k_*,+}^2(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}}) \bigwedge_{i=1,2} \kappa^2 \left[\mathfrak{G}, k, \sqrt{\Sigma^{(i)}} \right]}, \end{aligned}$$

that measure the closeness to orthogonality of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. Theorem 3.3 is straightforward consequence of the two following results.

Proposition 5.2. *There exist four positive constants L^* , L_1^* , L_2^* , and L_3^* such that following holds. Define k_* as the largest integer that satisfies*

$$(k_* + 1) \log(p) \leq L^*(n_1 \wedge n_2), \quad (23)$$

and assume that

$$1 + \log [1/(\alpha\delta)] < L_1^*(n_1 \wedge n_2). \quad (24)$$

The hypothesis \mathcal{H}_0 is rejected by $T_{\mathcal{S}_{Lasso}}^B$ with probability larger than $1 - \delta$ for any $(\beta^{(1)}, \beta^{(2)})$ satisfying

$$|\beta^{(1)}|_0 + |\beta^{(2)}|_0 \leq L_2^* \gamma_{\Sigma^{(1)}, \Sigma^{(2)}, k} k_* \left(\frac{n_1}{n_2} \wedge \frac{n_2}{n_1} \right). \quad (25)$$

and

$$\mathcal{K}_1 + \mathcal{K}_2 \geq L_3^* \gamma'_{\Sigma^{(1)}, \Sigma^{(2)}, k_*} \frac{(|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \vee 1) \log(p) + \log\{1/(\alpha\delta)\}}{n_1 \wedge n_2} \left(\frac{n_1}{n_2} \vee \frac{n_2}{n_1} \right).$$

This proposition tells us that $T_{\mathcal{S}_{Lasso}}^B$ behaves nearly as well as what has been obtained in (11) for $T_{\mathcal{S}_{\leq (n_1 \wedge n_2)/2}}^B$, at least when n_1 and n_2 are of the same order.

In the next proposition, we assume that $\Sigma^{(1)} = \Sigma^{(2)} := \Sigma$. Given an integer k , define

$$\tilde{\gamma}_{\Sigma, k} := \frac{\kappa[6, k, \sqrt{\Sigma}] \Phi_{k, -}^{1/2}(\sqrt{\Sigma})}{\Phi_{1, +}(\sqrt{\Sigma})}, \quad \tilde{\gamma}_{\Sigma, k}^{(2)} := \frac{\kappa^2 [6, k, \sqrt{\Sigma}]}{\Phi_{k, +}(\sqrt{\Sigma})}, \quad \tilde{\gamma}_{\Sigma, k}^{(3)} := \frac{\Phi_{1, +}^2(\sqrt{\Sigma})}{\kappa^2 [6, k, \sqrt{\Sigma}]}.$$

Proposition 5.3. *Let us assume that $\Sigma^{(1)} = \Sigma^{(2)} := \Sigma$. There exist five positive constants L^* , \tilde{L}^* , L_1^* , L_2^* , and L_3^* such that following holds. Define k_* and \tilde{k}_* as the largest positive integers that satisfy*

$$\begin{aligned} (k_* + 1) \log(p) &\leq L^*(n_1 \wedge n_2), \\ \tilde{k}_* &\leq \tilde{L}^* \tilde{\gamma}_{\Sigma, k_*} \left[\frac{n_1 \wedge n_2}{|n_1 - n_2|} \wedge \sqrt{\frac{n_1 \wedge n_2}{\log(p)}} \right], \end{aligned} \quad (26)$$

with the convention $x/0 = \infty$. Assume that

$$1 + \log[1/(\alpha\delta)] < L_1^*(n_1 \wedge n_2).$$

The hypothesis \mathcal{H}_0 is rejected by $T_{\mathcal{S}_{Lasso}}^B$ with probability larger than $1 - \delta$ for any $(\beta^{(1)}, \beta^{(2)})$ satisfying

$$|\beta^{(1)}|_0 + |\beta^{(2)}|_0 \leq L_2^* \tilde{\gamma}_{\Sigma, \tilde{k}_*}^{(2)} \tilde{k}_*. \quad (27)$$

and

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{\text{Var}(Y^{(1)}) \wedge \text{Var}(Y^{(2)})} \geq L_3^* \tilde{\gamma}_{\Sigma, k_*}^{(3)} \left[(|\beta^{(1)} - \beta^{(2)}|_0 \vee 1) \log(p) + \log\{1/(\alpha\delta)\} \right].$$

Remark 5.2. *The definition (26) of \tilde{k}_* together with Condition (27) restrict the number of non-zero components $|\beta^{(1)}|_0 + |\beta^{(2)}|_0$ to be small in front of $(n_1 \wedge n_2)/|n_1 - n_2|$. This technical assumption enforces the design matrix in the reparametrized model (5) to be almost block-diagonal and allows us to control efficiently the Lasso estimator $\hat{\theta}_{\lambda}^{(2)}$ of $\theta_*^{(2)}$ for some $\lambda > 0$ (see the proof in Section 8 for further details). Still, this is not clear to what extent this assumption is necessary.*

6. Discussion

In this paper, we develop an adaptive likelihood-ratio test which reaches minimax high-dimensional rates of testing to compare two linear regressions. To control explicitly the type I error of the global test, a two-step calibration method has been proposed. First, the p -values of each individual test $(F_{S, i}, i \in \{1, 2, 3\}, S \in \hat{\mathcal{S}})$ are explicitly upper bounded (Proposition 2.2), then the thresholds (\hat{C}_i) are calibrated using a permutation method with weights given by (10). Using a naive permutation approach without properly bounding the p -values or without correcting the weights as in (10) would have favored large subsets S in the global procedure.

In the spirit of [5], our type II error analysis is completely non-asymptotic. However, the numerical constants involved in the bounds are clearly not optimal. Another line of

work initiated by [14] considers an asymptotic but high-dimensional framework and aims at providing detection rates with optimal constants. For instance [3, 19] have derived such results in the one-sample high-dimensional linear regression testing problem under strong assumptions on the design matrices. In our opinion, both analyses are complementary. If deriving sharp detection rates (under perhaps stronger assumptions on the covariance) is a stimulating open problem, it is beyond the scope of our paper.

The Kullback discrepancies considered in the power analysis of the test depend on $\beta^{(1)}$ and $\beta^{(2)}$ through the prediction distances $\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma^i}$, $i = 1, 2$ rather than the l_2 distance $\|\beta^{(1)} - \beta^{(2)}\|$. On the one hand, such a dependency on the prediction abilities is natural, as our testing procedure relies on the likelihood ratio. On the other hand, it is possible to characterize the power of our testing procedures as in Theorems 3.1 and 3.3 in terms of the distance $\|\beta^{(1)} - \beta^{(2)}\|$ by inverting $\Sigma^{(1)}$ and $\Sigma^{(2)}$ at $\beta^{(1)} - \beta^{(2)}$. However, the inversion would lead to an additional factor of the form $\Phi_{|\beta^{(1)} - \beta^{(2)}|_0, -}^{-1}(\sqrt{\Sigma^{(i)}})$ in the testing rates.

In terms of interpretation, even though our procedure adopts a global testing approach through prediction distances, our real dataset example illustrates that identifying which subset in the collection is responsible for rejecting the null hypothesis provides clues into which specific coefficients are most likely to differ between samples.

Thinking of gene network inference by Gaussian graphical modeling, the high levels of correlations encountered within transcriptomic datasets and the potential number of missing variables result in highly unstable graphical estimations. Our global testing approach provides a way to validate whether sample-specific graphs eventually share comparable predictive abilities or disclose genuine structural changes. Such a statistical validation is obviously crucial before translating any graphical analysis into further biological experiments. Interestingly, the three main genes pointed out by our testing strategy have been validated as promising therapeutic targets by functional biology experiments.

Finally, this test should also facilitate the validation of the fundamental i.i.d. assumption across multiple samples, paving the way to pooled analyses when possible. In that respect, we draw attention to the significant heterogeneity detected between the training and validation sets of the well-known Hess *et al* dataset, suggesting that these samples should be used as originally intended. Methods which require i.i.d. observations should only be applied with caution to this dataset if considered as a unique homogeneous sample.

7. Technical Details

7.1. Two-sample testing for fixed and different designs

Proposition 7.1. *Consider \mathbf{X}_1 and \mathbf{X}_2 as fixed and assume that $\sigma^{(1)} = \sigma^{(2)} = 1$. If the $(n_1 + n_2 \times p)$ matrix formed by \mathbf{X}_1 and \mathbf{X}_2 has rank $n_1 + n_2$, then any test T based on the data (\mathbf{Y}, \mathbf{X}) satisfies:*

$$\sup_{\beta \in \mathbb{R}^p} \mathbb{P}_{\beta, \beta} [T = 1] + \inf_{\beta^{(1)} \neq \beta^{(2)} \in \mathbb{R}^p} \mathbb{P}_{\beta^{(1)}, \beta^{(2)}} [T = 0] \geq 1 .$$

In other words, any level- α test T has a type II error larger than $1 - \alpha$, and this uniformly over $\beta^{(1)}$ and $\beta^{(2)}$. Consequently, any test in this setting cannot perform better than complete random guess.

Proof. Using the rank condition, we derive that for any vector (a, b) in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, there exists $\beta \in \mathbb{R}^p$ such that $\mathbf{X}^{(1)}\beta = a$ and $\mathbf{X}^{(2)}\beta = b$. Consequently, under the null hypothesis, $(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ follows any distributions $\mathcal{N}(a, I_{n_1}) \otimes \mathcal{N}(b, I_{n_2})$ with (a, b) arbitrary in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. Hence, for any $\beta^{(1)} \neq \beta^{(2)} \in \mathbb{R}^p$, the distribution $\mathbb{P}_{\beta^{(1)}, \beta^{(2)}}$ of $(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$ is not distinguishable from the null hypothesis. The result follows.

□

7.2. Upper bounds of the quantiles

This Section explicits the upper bounds $\tilde{Q}_{2,|S|}(u|\mathbf{X}_S)$ and $\tilde{Q}_{3,|S|}(u|\mathbf{X}_S)$. Because of the symmetry between $F_{S,2}$ and $F_{S,3}$, we only provide developments for $F_{S,2}$. Let us note $a = (a_1, \dots, a_{|S|})$ the positive eigenvalues of

$$\frac{n_1}{n_2(n_1 - |S|)} \mathbf{X}_S^{(2)} \left[(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)\top} \mathbf{X}_S^{(2)})^{-1} \right] \mathbf{X}_S^{(2)\top}.$$

Definition 7.2 (Recall of the definition of the upper-bound $\tilde{Q}_{2,|S|}(u|\mathbf{X}_S)$). *Consider some number $u > |a|_1$. If all the components of a are equal, then we take*

$$\lambda^* = \frac{u - |a|_1}{2u(|a|_\infty + \frac{|a|_1}{n_1 - |S|})}$$

If a is not a constant vector, then we define λ^* by

$$\begin{aligned} b &:= \frac{|a|_1 u}{|a|_\infty (n_1 - |S|)} + u + \frac{\|a\|^2}{|a|_\infty} - |a|_1, \\ \Delta &:= b^2 - \frac{4u(u - |a|_1)}{(n_1 - |S|)|a|_\infty} \left(|a|_1 - \frac{\|a\|^2}{|a|_\infty} \right), \end{aligned} \quad (28)$$

$$\lambda^* := \frac{1}{\frac{4u}{n_1 - |S|} \left(|a|_1 - \frac{\|a\|^2}{|a|_\infty} \right)} (b - \sqrt{\Delta}) \quad (29)$$

We recall that $\tilde{Q}_{2,|S|}(u|\mathbf{X}_S)$ is defined as follows

$$\tilde{Q}_{2,|S|}(u|\mathbf{X}_S) := \exp \left[-\frac{1}{2} \sum_{i=1}^{|S|} \log(1 - 2\lambda^* a_i) - \frac{n_1 - |S|}{2} \log \left(1 + \frac{2\lambda^* u}{n_1 - |S|} \right) \right].$$

Proof of Proposition 2.2. For the sake of simplicity, we note $N = n_1 - |S|$, $(Z_1, \dots, Z_{|S|})$ a standard Gaussian random vector and W_N a χ^2 random variable with N degrees of freedom. We apply Laplace method to upper bound $\mathbb{P}[F_{S,2} \geq u]$:

$$\begin{aligned} \mathbb{P}[F_{S,2} \geq u] &= \mathbb{P} \left[\sum_{i=1}^{|S|} a_i Z_i^2 \geq u W_N / N \right] \leq \inf_{\lambda > 0} \mathbb{E} \exp \left[\lambda \sum_{i=1}^{|S|} a_i Z_i^2 - \lambda u W_N / N \right] \\ &\leq \inf_{0 < \lambda < |a|_\infty / 2} \exp[\psi_u(\lambda)], \end{aligned}$$

where

$$\psi_u(\lambda) = -\frac{1}{2} \sum_{i=1}^{|S|} \log(1 - 2\lambda a_i) - \frac{N}{2} \log \left(1 + \frac{2\lambda u}{N} \right).$$

The sharpest upper-bound is given by the value λ^* which minimizes $\psi_u(\lambda)$. We obtain an approximation of λ^* by cancelling the second-order approximation of its derivative. Deriving ψ_u gives

$$\psi'_u(\lambda) = \sum_{i=1}^{|S|} \frac{a_i}{1 - 2\lambda a_i} - \frac{u}{1 + \frac{2\lambda u}{N}},$$

which admits the following second order approximation :

$$|a|_1 + \frac{2\lambda\|a\|^2}{1-2|a|_\infty\lambda} - \frac{u}{1+\frac{2\lambda u}{N}}. \quad (30)$$

Cancelling this quantity amounts to solving a polynomial equation of the second degree. The smallest solution of this equation leads to the desired λ^* . \square

8. Proofs

Additional Notations. Given a subset S , $\Pi_S^{(1)}$ (resp. $\Pi_S^{(2)}$) stands for the orthogonal projection onto the space spanned by the rows of $\mathbf{X}_S^{(1)}$ (resp. $\mathbf{X}_S^{(2)}$). Moreover, $\Pi_{S^\perp}^{(1)}$ denotes the orthogonal projection along the space spanned by the rows of $\mathbf{X}_S^{(1)}$.

8.1. Distributions of $F_{S,1}$, $F_{S,2}$ and $F_{S,3}$ (Proposition 2.1)

Let us consider the regression of $Y^{(1)}$ (resp. $Y^{(2)}$) with respect to $X_S^{(1)}$ (resp. $X_S^{(2)}$):

$$Y^{(1)} = X_S^{(1)}\beta_S^{(1)} + \epsilon_S^{(1)}, \quad Y^{(2)} = X_S^{(2)}\beta_S^{(2)} + \epsilon_S^{(2)}.$$

Under the null hypothesis $\mathcal{H}_{0,S}$, we have $\beta_S^{(1)} = \beta_S^{(2)}$ and $\sigma_S^{(1)} = \sigma_S^{(2)}$. For the sake of simplicity, we write β_S and σ_S for these two quantities. Define the random variable T_1 and T_2 as

$$T_1 = \frac{\|\Pi_{S^\perp}^{(1)}\epsilon_S^{(1)}\|^2}{(n_1 - |S|)\sigma_S^2}, \quad T_2 = \frac{\|\Pi_{S^\perp}^{(2)}\epsilon_S^{(2)}\|^2}{(n_2 - |S|)\sigma_S^2}. \quad (31)$$

Given \mathbf{X} , T_1/T_2 follows a Fisher distribution with $(n_1 - |S|, n_2 - |S|)$ degrees of freedom. Observing that under the null hypothesis

$$F_{S,1} = -2 + \frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} + \frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)}$$

allows us to prove the first assertion of Proposition 2.1. Let us turn to the second statistic:

$$F_{S,2} = \frac{n_1}{n_2(n_1 - |S|)} \frac{U}{T_1},$$

where

$$U = \frac{\|\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(2)\top}\mathbf{X}_S^{(2)})^{-1}\mathbf{X}_S^{(2)\top}\epsilon_S^{(2)} - \mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top}\mathbf{X}_S^{(1)})^{-1}\mathbf{X}_S^{(1)\top}\epsilon_S^{(1)}\|^2}{\sigma_S^2}.$$

Given \mathbf{X} , U is independent from T_1 since T_1 is a function of $\Pi_{S^\perp}^{(1)}\epsilon_S^{(1)}$ while U is a function of $(\epsilon_S^{(2)}, \Pi_S^{(1)}\epsilon_S^{(1)})$. Furthermore, U is the squared norm of a centered Gaussian vector with covariance

$$\mathbf{X}_S^{(2)} \left[(\mathbf{X}_S^{(1)\top}\mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)\top}\mathbf{X}_S^{(2)})^{-1} \right] \mathbf{X}_S^{(2)\top}.$$

8.2. Calibrations

Proof of Proposition 2.3. By definition of the p -values $\tilde{Q}_{i,|S|}$, we have under \mathcal{H}_0 for each $S \in \mathcal{S}$ and each $i \in \{1, 2, 3\}$

$$\mathbb{P}_{\mathcal{H}_0} \left[\tilde{Q}_{i,|S|} (F_{S,i}|\mathbf{X}_S) \leq \alpha_{i,S}|\mathbf{X}_S \right] \leq \alpha_{i,S}.$$

Applying a union bound and integrating with respect to \mathbf{X} allows us to control the type I error:

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_0}[T_{\hat{\mathcal{S}}}^B = 1] &= \mathbb{E} \left[\sum_{S \in \hat{\mathcal{S}}} \sum_{i=1}^3 \mathbb{P} \left[\tilde{Q}_{i,|S|}(F_{S,i}|\mathbf{X}_S) < \alpha_{i,S} \right] \right] \\ &\leq \sum_{S \in \mathcal{S}} \sum_{i=1}^3 \mathbb{P} \left[\tilde{Q}_{i,|S|}(F_{S,i}|\mathbf{X}_S) < \alpha_{i,S} \right] \\ &\leq \sum_{S \in \mathcal{S}} \sum_{i=1}^3 \mathbb{E}_{\mathbf{X}_S} \left[\mathbb{P} \left[\tilde{Q}_{i,|S|}(F_{S,i}|\mathbf{X}_S) < \alpha_{i,S} \right] \right] \leq \sum_{S \in \mathcal{S}} \alpha_{i,S} \leq \alpha, \end{aligned}$$

where we have upper bounded the sum over the random collection \mathcal{S} by the sum over \mathcal{S} . \square

Proof of Proposition 2.4. Consider $i \in \{1, 2, 3\}$. Under \mathcal{H}_0 , the distributions of

$$\begin{aligned} &\min_{S \in \hat{\mathcal{S}}_\pi} \left\{ \tilde{Q}_{1,|S|}(F_{S,1}(\pi)|\mathbf{X}_S^\pi) \binom{p}{|S|} \right\}, \\ &\min_{S \in \hat{\mathcal{S}}_\pi} \left\{ \left(\tilde{Q}_{2,|S|}(F_{S,2}(\pi)|\mathbf{X}_S^\pi) \wedge \tilde{Q}_{3,|S|}(F_{S,3}(\pi)|\mathbf{X}_S^\pi) \right) \binom{p}{|S|} \right\} \end{aligned}$$

are invariant with respect to the permutation π . Hence, we derive

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_0} \left[\min_{S \in \hat{\mathcal{S}}} \tilde{Q}_{1,|S|}(F_{S,1}|\mathbf{X}_S) \binom{p}{|S|} \leq \hat{C}_1 \middle| \mathbf{X}_S \right] &= \alpha/2, \\ \mathbb{P}_{\mathcal{H}_0} \left[\min_{S \in \hat{\mathcal{S}}_\pi} \left\{ \left(\tilde{Q}_{2,|S|}(F_{S,2}(\pi)|\mathbf{X}_S^\pi) \wedge \tilde{Q}_{3,|S|}(F_{S,3}(\pi)|\mathbf{X}_S^\pi) \right) \binom{p}{|S|} \right\} \leq \hat{C}_2 \middle| \mathbf{X}_S \right] &= \alpha/2. \end{aligned}$$

Applying a union bound and integrating with respect to \mathbf{X} allows us to conclude. \square

8.3. Proof of Theorem 5.1

The objective is to exhibit a subset for which the power of T_S^B is larger than $1 - \delta$. This subset is such that the distance between the two sample-specific distributions is large enough that we can actually reject the null hypothesis with large probability. As exposed in Theorem 5.1, we rely on the semi-distances $\mathcal{K}_1(S) + \mathcal{K}_2(S)$ for $S \in \mathcal{S}$:

$$2(\mathcal{K}_1(S) + \mathcal{K}_2(S)) = \left(\frac{\sigma_S^{(1)}}{\sigma_S^{(2)}} \right)^2 + \left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}} \right)^2 - 2 + \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_S^{(2)})^2} + \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(1)}}^2}{(\sigma_S^{(1)})^2}. \quad (32)$$

The proof is split into five main lemmas. First, we upper bound $\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S)$, $\tilde{Q}_{2,|S|}^{-1}(x|\mathbf{X}_S)$, and $\tilde{Q}_{3,|S|}^{-1}(x|\mathbf{X}_S)$ in Lemmas 8.1, 8.2 and 8.3. Then, we control the deviations of $F_{S,1}$, $F_{S,2}$, and $F_{S,3}$ under $\mathcal{H}_{1,S}$ in Lemmas 8.4 and 8.5. In the sequel, we call \mathbf{S}' the subcollection of \mathbf{S} made of subsets S satisfying $|S| \leq (n_1 \wedge n_2)/2$ and

$$\log(12/\delta) < L_1^\bullet(n_1 \wedge n_2), \quad \log(1/\alpha_S) \leq L_2^\bullet(n_1 \wedge n_2), \quad |S| \leq L_3^\bullet \quad (33)$$

where the numerical constants L_1^\bullet , L_2^\bullet , and L_3^\bullet only depend on L_2^* in (40) and on the constants introduced in Lemmas 8.1–8.5. These conditions allow us to fix the constants in the statement (19) of Theorem 5.1.

Lemma 8.1 (Upper-bound of $\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S)$). *There exists a positive universal constant L such that the following holds. Consider some $0 < x < 1$ such that $16 \log(2/x) \leq n_1 \wedge n_2$. For any subset S of size smaller than $(n_1 \wedge n_2)/2$, we have*

$$\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S) \leq L \left\{ \left(\frac{|S|(n_1 - n_2)}{n_1 n_2} \right)^2 + \log(2/x) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}. \quad (34)$$

We recall that $a = (a_1, \dots, a_{|S|})$ denotes the positive eigenvalues of

$$\frac{n_1}{n_2(n_1 - |S|)} \mathbf{X}_S^{(2)} \left[(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{-1} + (\mathbf{X}_S^{(2)\top} \mathbf{X}_S^{(2)})^{-1} \right] \mathbf{X}_S^{(2)\top}.$$

Lemma 8.2 (Upper-bound of $\tilde{Q}_{2,|S|}^{-1}(x|\mathbf{X}_S)$). *There exist two positive universal constants L_1 and L_2 such that the following holds. If $|a|_1 < u \leq (n_1 - |S|)|a|_\infty$ and if $|S| \leq L_1 n_1$,*

$$\log \left[\tilde{Q}_{2,|S|}^{-1}(u|\mathbf{X}_S) \right] \leq - \frac{(u - |a|_1)^2}{4[|a|_\infty(u - |a|_1) + \|a\|^2]} + \frac{(u - |a|_1)u^3}{2(n_1 - |S|)[|a|_\infty(u - |a|_1) + \|a\|^2]}.$$

For any $0 < x < 1$, satisfying

$$L_2 \log(1/x) \leq n_1 - |S|, \quad (35)$$

we have the following upper bound

$$\tilde{Q}_{2,|S|}^{-1}(x|\mathbf{X}_S) \leq |a|_\infty \left[2|S| + 2\sqrt{2|S| \log(1/x)} + 8 \log(1/x) \right]. \quad (36)$$

Lemma 8.3 (Upper-bound of $|a|_\infty$). *There exist two positive universal constants L_1 and L_2 such that the following holds. Consider δ a positive number satisfying $L_1 \log(4/\delta) < n_1 \wedge n_2$. With probability larger than $1 - \delta/2$, we have*

$$|a|_\infty \leq L_2 \left[\frac{1}{n_2} + \frac{\varphi_{\max} \left\{ \sqrt{\Sigma_S^{(2)}} (\Sigma_S^{(1)})^{-1} \sqrt{\Sigma_S^{(2)}} \right\}}{n_1} \right].$$

Lemma 8.4 (Deviations of $F_{S,1}$). *There exist three positive universal constants L_1 , L_2 and L_3 such that the following holds. Assume that $L_1 \log(1/\delta) \leq n_1 \wedge n_2$. With probability larger than $1 - \delta$, we have*

$$F_{S,1} \geq L_2 \left(\frac{(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2}{\sigma_S^{(1)} \sigma_S^{(2)}} \right)^2 - L_3 \left[|S|^2 \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} \right) + \log \left(\frac{1}{\delta} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]. \quad (37)$$

Lemma 8.5 (Deviations of $F_{S,2}$). *There exist two positive universal constants L_1 and L_2 such that the following holds. Assume that*

$$L_1 \log(12/\delta) < n_1 \wedge n_2. \quad (38)$$

With probability larger than $1 - \delta/2$, we have

$$F_{S,2} \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{8(\sigma_S^{(1)})^2} - \log(6/\delta) L_2 \left[\frac{1}{n_2} \frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2} + \frac{\varphi_S}{n_1} \right], \quad (39)$$

where φ_S is defined in (18).

Consider some $S \in \mathbf{S}'$. Combining Lemmas 8.1 and 8.4, we derive that $\tilde{Q}_{1,|S|}(F_{S,1}|\mathbf{X}_S) \leq \alpha_S$ holds with probability larger than $1 - \delta$ if

$$\frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} \geq L \left[|S|^2 \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} \right) + \log[1/(\alpha_S \delta)] \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right].$$

Similarly, combining Lemmas 8.2, 8.3, and 8.5, we derive that $\tilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \leq \alpha_S$ with probability larger than $1 - \delta$ if

$$\frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_S^{(1)})^2} \geq L'_1 (\varphi_S + 1) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{6}{\delta \alpha_S} \right) \right] + \frac{L'_2}{n_2} \left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}} \right)^2 \log \left(\frac{6}{\delta} \right).$$

A symmetric result holds for $\tilde{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S)$.

Consequently, $\tilde{Q}_{1,|S|}(F_{S,1}|\mathbf{X}_S) \wedge \tilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S) \wedge \tilde{Q}_{3,|S|}(F_{S,3}|\mathbf{X}_S) \leq \alpha_S$ with probability larger than $1 - \delta$ if

$$\begin{aligned} \mathcal{K}_1(S) + \mathcal{K}_2(S) &\geq L_1^* \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{6}{\alpha_S \delta} \right) \right] \\ &\quad + L_2^* \log(6/\delta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[\left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}} \right)^2 + \left(\frac{\sigma_S^{(1)}}{\sigma_S^{(2)}} \right)^2 \right]. \end{aligned} \quad (40)$$

Since we assume that $4L_2^* \log(6/\delta) \leq n_1 \wedge n_2$ in (33), the last condition is fulfilled if

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \geq L^* \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) [|S| + \log\{6/(\alpha_S \delta)\}].$$

We now proceed to the proof of the five previous lemmas.

Proof of Lemma 8.1. Let $u \in (0, 1)$ and $\bar{F}_{D,N}^{-1}(u)$ be the $1 - u$ quantile of a Fisher random variable with D and N degrees of freedom. According to [5], we have

$$\bar{F}_{D,N}^{-1}(u) \leq 1 + 2\sqrt{\left(\frac{1}{D} + \frac{1}{N} \right) \log \left(\frac{1}{u} \right)} + \left(\frac{N}{2D} + 1 \right) \left[\exp \left(\frac{4}{N} \log \left(\frac{1}{u} \right) \right) - 1 \right].$$

Let us assume that $8/N \log(1/u) \leq 1$. By convexity of the exponential function it holds that

$$\bar{F}_{D,N}^{-1}(u) \leq 1 + 2\sqrt{\left(\frac{1}{D} + \frac{1}{N} \right) \log \left(\frac{1}{u} \right)} + \left(\frac{4}{D} + \frac{8}{N} \right) \log \left(\frac{1}{u} \right).$$

Recall T_1 and T_2 defined in (31). Under hypothesis \mathcal{H}_0 ,

$$\frac{T_1}{T_2} \sim \text{Fisher}(n_1 - |S|, n_2 - |S|).$$

Consider some $x > 0$ such that $[8/(n_1 - |S|) \vee 8/(n_2 - |S|)] \log(2/x) \leq 1$. Then, with probability larger than $1 - x/2$ we have,

$$\begin{aligned} \frac{T_1 n_2 (n_1 - |S|)}{T_2 n_1 (n_2 - |S|)} &\leq \left(1 + \frac{|S|(n_1 - n_2)}{n_1(n_2 - |S|)} \right) \left(1 + 8\sqrt{\frac{\log(2/x)}{n_1 - |S|}} + 8\sqrt{\frac{\log(2/x)}{n_2 - |S|}} \right) \\ &\leq \left(1 + \frac{|S|(n_1 - n_2)}{n_1(n_2 - |S|)} \right) \left(1 + 12\sqrt{\frac{\log(2/x)}{n_1}} + 12\sqrt{\frac{\log(2/x)}{n_2}} \right) \leq L, \end{aligned}$$

since $|S| \leq (n_1 \wedge n_2)/2$. Similarly, with probability at least $1 - x/2$, we have

$$\frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} \leq \left[\left(1 + \frac{|S|(n_2 - n_1)}{n_2(n_1 - |S|)} \right) \left(1 + 12\sqrt{\frac{\log(2/x)}{n_1}} + 12\sqrt{\frac{\log(2/x)}{n_2}} \right) \right] \wedge L. \quad (41)$$

Depending on the sign of $\frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} - 1$, we apply one of the two following identities:

$$\begin{aligned} \frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} + \frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} - 2 &= \left(\frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} - 1 \right)^2 \frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)}, \\ \frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} + \frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} - 2 &= \left(\frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} - 1 \right)^2 \frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)}. \end{aligned}$$

Combining the different bounds, we conclude that with probability larger than $1 - x$,

$$\begin{aligned} F_{S,1} &:= \frac{T_1}{T_2} \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} + \frac{T_2}{T_1} \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} - 2 \\ &\leq L \left[\left(\frac{|S|(n_1 - n_2)}{n_1 n_2} \right)^2 + \log(2/x) \frac{n_1 + n_2}{n_1 n_2} \right]. \end{aligned}$$

□

Proof of Lemma 8.2. As in the proof of Proposition 2.2, we note $N = n_1 - |S|$. Recall that $\tilde{Q}_{2,|S|}(u|\mathbf{X}_S)$ is defined as $\exp \psi_u(\lambda^*)$ (see Definition 7.2). We start by upper-bounding $\psi_u(\lambda^*)$, which proves the first upper-bound of the logarithm of the tail probability $\log \tilde{Q}_{2,|S|}(u|\mathbf{X}_S)$. We then exhibit a value u_x such that $\psi_{u_x}(\lambda^*) \leq \log x$.

Upper-bound of the tail probability. Since Equation (30) is increasing with respect to λ and with respect to N , λ^* decreases with N . Consequently,

$$\lambda^* \leq \lambda_+ := \frac{u - |a|_1}{2 \left[|a|_\infty (u - |a|_1) + \|a\|^2 \right]}.$$

By convexity, $1 - \sqrt{1 - x} \geq x/2$ for any $0 \leq x \leq 1$. Applying this inequality, we upper bound $\sqrt{\Delta}$ and derive that

$$\lambda^* \geq \lambda_- := \frac{u - |a|_1}{2 \left[|a|_\infty (u - |a|_1) + \|a\|^2 + \frac{|a|_1 u}{N} \right]}.$$

Since $u \leq N|a|_\infty$, $2\lambda^*u \leq N$. Observing that $-\log(1 - 2x)/2 \leq x + x^2/(1 - 2x)$ for any $0 < x < 1/2$ and that $\log(1 + x) \geq x - x^2$ for any $x > 0$, we derive

$$\begin{aligned} \psi_u(\lambda^*) &\leq |a|_1 \lambda_+ + \frac{\lambda_+^2 \|a\|^2}{1 - 2|a|_\infty \lambda_+} - \lambda^* u + 2 \frac{(\lambda^*)^2 u^2}{N} \\ &\leq -\frac{(u - |a|_1)^2}{4 \left[|a|_\infty (u - |a|_1) + \|a\|^2 \right]} + \frac{2\lambda_+^2 u^2}{N} + (\lambda_+ - \lambda_-)u \\ &\leq -\frac{(u - |a|_1)^2}{4 \left[|a|_\infty (u - |a|_1) + \|a\|^2 \right]} + \frac{(u - |a|_1)u^3}{2N \left[|a|_\infty (u - |a|_1) + \|a\|^2 \right]^2}. \quad (42) \end{aligned}$$

Upper-bound of the quantile. Let us turn to the upper bound of $\tilde{Q}_{2,|S|}^{-1}(x|\mathbf{X}_S)$. Consider u_x the solution larger than $|a|_1$ of the equation

$$\frac{(u - |a|_1)^2}{4[|a|_\infty(u - |a|_1) + \|a\|^2]} = 2 \log(1/x) ,$$

and observe that

$$2\|a\|\sqrt{\log(1/x)} \leq u_x - |a|_1 \leq 2\sqrt{2}\|a\|\sqrt{\log(1/x)} + 8|a|_\infty \log(1/x) .$$

Choosing L_1 and L_2 large enough in the condition $|S| \leq L_1 n_1$ and in condition (35) leads us to $u_x \leq N|a|_\infty$. We now prove that $\psi_{u_x \vee 2|a|_1}(\lambda^*) \leq \log x$. If $u_x \geq 2|a|_1$, then $u_x^3 \leq 8(u_x - |a|_1)^3$ and it follows from (42) that

$$\psi_{u_x}(\lambda^*) \leq \log(1/x) \left[-2 + \frac{2^8 \log(1/x)}{N} \right] \leq -\log(1/x)$$

if we take L_2 large enough in Condition (35). If $u_x \leq 2|a|_1$, then $|a|_1^2/(|a|_\infty|a|_1 + \|a\|^2) \geq 8 \log(1/x)$ and

$$\psi_{u_x \vee 2|a|_1}(\lambda^*) \leq -\frac{|a|_1^2}{4[|a|_\infty|a|_1 + \|a\|^2]} \left[1 - \frac{2^4|a|_1^2}{N[|a|_\infty|a|_1 + \|a\|^2]} \right] \leq -\log(1/x) ,$$

if we take L_1 and L_2 large enough in the two aforementioned condition. since $|S| \leq 2^{-6}n_1$. Thus, we conclude that

$$\tilde{Q}_{2,|S|}^{-1}(x|\mathbf{X}_S) \leq u_x \vee 2|a|_1 \leq |a|_1 + \left[2\sqrt{2}\|a\|\sqrt{\log(1/x)} + 8|a|_\infty \log(1/x) \right] \vee |a|_1 .$$

□

Proof of Lemma 8.3. Upon defining $\mathbf{Z}_S^{(1)} = \mathbf{X}_S^{(1)} (\Sigma_S^{(1)})^{-1/2}$ and $\mathbf{Z}_S^{(2)} = \mathbf{X}_S^{(2)} (\Sigma_S^{(2)})^{-1/2}$, it follows that $\mathbf{Z}_S^{(1)}$ and $\mathbf{Z}_S^{(2)}$ follow standard Gaussian distributions.

$$\begin{aligned} |a|_\infty &\leq \frac{n_1}{n_2(n_1 - |S|)} \left[1 + \varphi_{\max} \left\{ \mathbf{Z}_S^{(2)} \sqrt{\Sigma_S^{(2)} (\Sigma_S^{(1)})^{-1}} (\mathbf{Z}_S^{(1)\top} \mathbf{Z}_S^{(1)})^{-1} \sqrt{(\Sigma_S^{(1)})^{-1} \Sigma_S^{(2)} \mathbf{Z}_S^{(2)\top}} \right\} \right] \\ &\leq \frac{2}{n_2} + 2 \frac{\varphi_{\max}[\mathbf{Z}_S^{(2)\top} \mathbf{Z}_S^{(2)}]}{n_2 \varphi_{\max}[\mathbf{Z}_S^{(1)\top} \mathbf{Z}_S^{(1)}]} \varphi_{\max} \left[\sqrt{\Sigma_S^{(2)} (\Sigma_S^{(1)})^{-1}} \sqrt{\Sigma_S^{(2)}} \right] . \end{aligned}$$

In order to conclude, we control the largest and the smallest eigenvalues of Standard Wishart matrices applying Lemma 8.12. □

Proof of Lemma 8.4. By symmetry, we can assume that $\sigma_S^{(1)}/\sigma_S^{(2)} \geq 1$. Recall the definition of T_1 and T_2 in the proof of Proposition 2.1

CASE 1. Suppose that $T_1/T_2 \geq 1$.

$$\begin{aligned} -2 + \frac{(\sigma_S^{(1)})^2 T_1}{(\sigma_S^{(2)})^2 T_2} + \frac{(\sigma_S^{(2)})^2 T_2}{(\sigma_S^{(1)})^2 T_1} &\geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2 (\sigma_S^{(2)})^2} + \frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2} \left(\frac{T_1}{T_2} - 1 \right) + \frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2} \left(\frac{T_2}{T_1} - 1 \right) \\ &\geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2 (\sigma_S^{(2)})^2} . \end{aligned} \tag{43}$$

CASE 2. Suppose that $T_1/T_2 \leq 1$.

$$\begin{aligned} -2 + \frac{(\sigma_S^{(1)})^2 T_1}{(\sigma_S^{(2)})^2 T_2} + \frac{(\sigma_S^{(2)})^2 T_2}{(\sigma_S^{(1)})^2 T_1} &= \left(\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2} - \frac{T_2}{T_1} \right)^2 \frac{(\sigma_S^{(2)})^2 T_1}{(\sigma_S^{(1)})^2 T_2} \\ &\geq \frac{T_1}{T_2} \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{4(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} \mathbf{1}_{\frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2} - 1 \geq 2\left(\frac{T_2}{T_1} - 1\right)}. \end{aligned}$$

We need to control the deviations of T_2/T_1 . Using bound (41), we get

$$\frac{T_2}{T_1} \leq \left(1 + \frac{|S|(n_2 - n_1)}{n_2(n_1 - |S|)} \right) \left(1 + 12\sqrt{\frac{\log(1/\delta)}{n_1}} + 12\sqrt{\frac{\log(1/\delta)}{n_2}} \right),$$

with probability larger than $1 - \delta$. Since $|S| \leq (n_1 \wedge n_2)/2$, we derive that

$$\frac{T_2}{T_1} - 1 \leq \frac{2|S|}{n_1} + 24\sqrt{\frac{\log(1/\delta)}{n_1}} + 24\sqrt{\frac{\log(1/\delta)}{n_2}} \leq 3,$$

for L_1 large enough in the statement of the lemma. In conclusion, we have

$$-2 + \frac{(\sigma_S^{(1)})^2 T_1}{(\sigma_S^{(2)})^2 T_2} + \frac{(\sigma_S^{(2)})^2 T_2}{(\sigma_S^{(1)})^2 T_1} \geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{16(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2}, \quad (44)$$

with probability larger than $1 - \delta$, as long as

$$\frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} \geq L \left[\frac{|S|^2}{n_1^2} + \frac{|S|^2}{n_2^2} + \log(1/\delta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]. \quad (45)$$

Combining (43), (44), and (45), we derive

$$-2 + \frac{(\sigma_S^{(1)})^2 T_1}{(\sigma_S^{(2)})^2 T_2} + \frac{(\sigma_S^{(2)})^2 T_2}{(\sigma_S^{(1)})^2 T_1} \geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{16(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} - L \left[\frac{|S|^2}{n_1^2} + \frac{|S|^2}{n_2^2} + \log(1/\delta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right],$$

with probability larger than $1 - \delta$. \square

Proof of Lemma 8.5. We want to lower bound the random variable $F_{S,2} = \frac{Rn_1}{(\sigma_S^{(1)})^2 T_1 (n_1 - |S|)}$ where R is defined by

$$R := \|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)}) + \Pi_S^{(2)}\epsilon_S^{(2)} - \mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(1)\top}\epsilon_S^{(1)}\|^2/n_2.$$

Let us first work conditionally to $\mathbf{X}_S^{(1)}$ and $\mathbf{X}_S^{(2)}$. Upon defining the Gaussian vector W by

$$W \sim \mathcal{N}\left[0, (\sigma_S^{(2)})^2 \Pi_S^{(2)} + (\sigma_S^{(1)})^2 \mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top}\mathbf{X}_S^{(1)})^{(-1)}\mathbf{X}_S^{(2)\top}\right],$$

we get $R = \|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)}) + W\|^2/n_2$. We have the following lower bound:

$$\begin{aligned} R &\geq \left(\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\| + \left\langle W, \frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|} \right\rangle \right)^2 / n_2 \\ &\geq \frac{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|^2}{2n_2} - \frac{1}{n_2} \left\langle W, \frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|} \right\rangle^2 \end{aligned}$$

The random variable $\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|^2 / \|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2$ follows a χ^2 distribution with n_2 degrees of freedom. Given $(\mathbf{X}_S^{(1)}, \mathbf{X}_S^{(2)})$, $\left\langle W, \frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|} \right\rangle^2$ is proportional to a χ^2 distributed random variable with one degree of freedom and its variance is smaller than $(\sigma_S^{(2)})^2 + \varphi_{\max}[\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{(-1)} \mathbf{X}_S^{(2)\top}] (\sigma_S^{(1)})^2$. Applying Lemma 8.11, we derive that with probability larger than $1 - x/6$,

$$R \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{2} \left[1 - 2\sqrt{\frac{\log(12/x)}{n_2}} \right] - 4\frac{\log(12/x)}{n_2} \left[(\sigma_S^{(2)})^2 + (\sigma_S^{(1)})^2 \varphi_{\max}\{\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{(-1)} \mathbf{X}_S^{(2)\top}\} \right].$$

Using the upper bound $|S| \leq (n_1 \wedge n_2)/2$ and Lemma 8.12, we control the last term

$$\varphi_{\max}\left[\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{(-1)} \mathbf{X}_S^{(2)\top}\right] \leq L\varphi_S \frac{n_2}{n_1},$$

with probability larger than $1 - 2\exp[-(n_1 \wedge n_2)L']$. If we take the constant L_1 large enough in condition (38), then we get

$$R \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{4} - \log(12/\delta) L \left[\frac{(\sigma_S^{(2)})^2}{n_2} + \frac{(\sigma_S^{(1)})^2}{n_1} \varphi_S \right], \quad (46)$$

with probability larger than $1 - \delta/3$.

Let us now upper bound the random variable $T_1(n_1 - |S|)/n_1$. Since $(n_1 - S)T_1$ follows a χ^2 distribution with $n_1 - |S|$ degrees of freedom, we derive from Lemma 8.11 that

$$T_1(n_1 - |S|)/n_1 \leq 1 + 2\sqrt{\frac{\log(6/\delta)}{n_1}} + \frac{2}{n_1} \log(6/\delta) \leq 2, \quad (47)$$

with probability larger than $1 - \delta/6$. Gathering (46) and (47), we conclude that

$$F_{S,2} \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{8(\sigma_S^{(1)})^2} - \log(6/\delta) L \left[\left(\frac{\sigma_S^{(2)}}{\frac{1}{n_2}\sigma_S^{(1)}} \right)^2 + \frac{\varphi_S}{n_1} \right],$$

with probability larger than $1 - \delta/2$. □

8.4. Proof of Theorem 3.1: Power of $T_{S \leq k}^B$

This proposition is a straightforward corollary of Theorem 5.1. Consider the subsets S_{\cup} and S_{Δ} of $\{1, \dots, p\}$ such that S_{\cup} is the union of the support of $\beta^{(1)}$ and $\beta^{(2)}$ and S_{Δ} is the supports of $\beta^{(2)} - \beta^{(1)}$. Assume first that S_{\cup} and S_{Δ} are non empty. By Definition (9) of the weights, we have

$$\log\left(\frac{1}{\alpha_{i,S_{\cup}}}\right) \leq \log(4k) + \log(1/\alpha) + |S_{\cup}| \log(p) \leq 2|S_{\cup}| \log(p) + \log(1/\alpha).$$

A similar upper bound holds for $\log(1/\alpha_{i,S_{\Delta}})$. If we choose the numerical constants large enough in Conditions A.1 and A.2, then the sets S_{\cup} and S_{Δ} follow the conditions of Theorem 5.1.

Applying Theorem 5.1, we derive that $T_{S_{\leq k}}^B$ rejects \mathcal{H}_0 with probability larger than $1 - \delta$ when

$$\mathcal{K}_1(S_{\cup}) + \mathcal{K}_2(S_{\cup}) \geq \varphi_{S_{\cup}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S_{\cup}| + \log \left(\frac{1}{\alpha_{S_{\cup}}} \right) \right].$$

Observing that $\varphi_{S_{\cup}} \leq \varphi_{\Sigma^{(1)}, \Sigma^{(2)}}$, $\mathcal{K}_1(S_{\cup}) = \mathcal{K}_1$, $\mathcal{K}_2(S_{\cup}) = \mathcal{K}_2$ and that $|S_{\cup}| \leq |\beta^{(1)}|_0 + |\beta^{(2)}|_0$ allows to prove the first result. Let us turn to the second result. According to Theorem 5.1, $T_{S_{\leq k}}^B$ rejects \mathcal{H}_0 with probability larger than $1 - \delta$ when

$$\mathcal{K}_1(S_{\Delta}) + \mathcal{K}_2(S_{\Delta}) \geq \varphi_{S_{\Delta}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S_{\Delta}| + \log \left(\frac{1}{\alpha_{S_{\Delta}}} \right) \right].$$

Since $\mathcal{K}_1(S_{\Delta}) + \mathcal{K}_2(S_{\Delta}) \geq \frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{2[\text{Var}(Y^{(1)}) \wedge \text{Var}(Y^{(2)})]}$ and since $|S_{\Delta}| = |\beta^{(1)} - \beta^{(2)}|_0$, the second result follows.

If $S_{\cup} = \emptyset$, then we can consider any subset of size 1 to prove the first result. If $S_{\Delta} = \emptyset$, then $\beta^{(1)} = \beta^{(2)}$ and the second result does not tell us anything.

8.5. Proof of Proposition 5.2

For simplicity, we assume in the sequel that $\beta^{(1)} \neq 0$ or $\beta^{(2)} \neq 0$, the case $\beta^{(1)} = \beta^{(2)} = 0$ being handled by any set $S \in \mathcal{S}_1 \subset \widehat{\mathcal{S}}_{\text{Lasso}}$.

This proof is divided into two main steps. First, we prove that with large probability the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ contains some set \widehat{S}_{λ} close to the union S_{\cup} of the supports of $\beta^{(1)}$ and $\beta^{(2)}$. Then, we show that the statistics $(F_{\widehat{S}_{\lambda,1}}, F_{\widehat{S}_{\lambda,2}}, F_{\widehat{S}_{\lambda,3}})$ allow to reject \mathcal{H}_0 with large probability.

Recall that the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ is based on the Lasso regularization path of the following heteroscedastic Gaussian linear model,

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \theta_*^{(1)} \\ \theta_*^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix} \quad (48)$$

which we denote for short $\mathbf{Y} = \mathbf{W}\theta_* + \boldsymbol{\epsilon}$. Given a tuning parameter λ , $\widehat{\theta}_{\lambda}$ refers to the Lasso estimator of θ :

$$\widehat{\theta}_{\lambda} = \arg \inf_{\theta \in \mathbb{R}^{2p}} \|\mathbf{Y} - \mathbf{W}\theta\|^2 + \lambda|\theta|_1.$$

In order to analyze the Lasso solution $\widehat{\theta}_{\lambda}$, we need to control how \mathbf{W} acts on sparse vectors.

Lemma 8.6 (Control of the design \mathbf{W}). *If we take the constants L^* , L_1^* , and L_2^* in Proposition 5.2 small enough then the following holds. The event*

$$\mathcal{A} := \left\{ \forall \theta \text{ s.t. } |\theta|_0 \leq k_*, \frac{1}{2} \leq \frac{\|\mathbf{X}^{(1)}\theta\|^2}{n_1\|\theta\|_{\Sigma^{(1)}}^2} \leq 2 \text{ and } \frac{1}{2} \leq \frac{\|\mathbf{X}^{(2)}\theta\|^2}{n_2\|\theta\|_{\Sigma^{(2)}}^2} \leq 2 \right\} \\ \cap \left\{ \frac{\kappa \left[6, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \mathbf{X}^{(1)}/\sqrt{n_1} \right]}{\kappa \left[6, |\beta^{(1)}|_0 + |\beta^{(2)}|_0, \sqrt{\Sigma^{(1)}} \right]} \wedge \frac{\kappa \left[6, |\theta_*|_0, \mathbf{X}^{(2)}/\sqrt{n_1} \right]}{\kappa \left[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}} \right]} \geq 2^{-3} \right\}$$

has probability larger than $1 - \delta/4$. Furthermore, on the event \mathcal{A} ,

$$\Phi_{k,+}(\mathbf{W}) \leq 4(n_1 + n_2) \left[\Phi_{k,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k,+}(\sqrt{\Sigma^{(2)}}) \right], \\ \Phi_{k,-}(\mathbf{W}) \geq (n_1 \wedge n_2) \left[\Phi_{k,-}(\sqrt{\Sigma^{(1)}}) \wedge \Phi_{k,-}(\sqrt{\Sigma^{(2)}}) \right],$$

for any $k \leq k_*$.

The following property is a slight variation of Lemma 11.2 in [38] and Lemma 3.2 in [16].

Lemma 8.7 (Behavior of the Lasso estimator $\widehat{\theta}_\lambda$). *If we take L_2^* in Proposition 5.2 small enough then the following holds. The event*

$$\mathcal{B} = \left\{ |\mathbf{W}^T \boldsymbol{\epsilon}|_\infty \leq 2(\sigma^{(1)} \vee \sigma^{(2)}) \sqrt{2\Phi_{1,+}(\mathbf{W}) \log(p)} \right\}$$

occurs with probability larger than $1 - 1/p$. Assume that

$$\lambda \geq 8(\sigma^{(1)} \vee \sigma^{(2)}) \sqrt{2\Phi_{1,+}(\mathbf{W}) \log(p)} .$$

Then, on the event $\mathcal{A} \cap \mathcal{B}$ we have

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq L_1 \frac{\lambda^2 / (n_1 \wedge n_2)}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0 , \quad (49)$$

$$|\widehat{\theta}_\lambda|_0 \leq L_2 \frac{n_1 \vee n_2}{n_1 \wedge n_2} \frac{\Phi_{k_*,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k_*,+}(\sqrt{\Sigma^{(2)}})}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0 \leq k_*/2 . \quad (50)$$

In the sequel, we fix

$$\lambda = 16(\sigma^{(1)} \vee \sigma^{(2)}) \sqrt{2(n_1 + n_2) \left[\Phi_{1,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{1,+}(\sqrt{\Sigma^{(2)}}) \right] \log(p)} .$$

and we consider the set \widehat{S}_λ defined by the union of the support of $\widehat{\theta}_\lambda^{(1)}$ and $\widehat{\theta}_\lambda^{(2)}$. On the event $\mathcal{A} \cap \mathcal{B}$, Lemma 8.7 tells us that $|\widehat{S}_\lambda| \leq k_*$. Thus, \widehat{S}_λ belongs to the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$. We shall prove that

$$\min_{i \in \{1,2,3\}} \widetilde{Q}_{i,|\widehat{S}_\lambda|} \left(F_{\widehat{S}_\lambda,i} \mid \mathbf{X}_{\widehat{S}_\lambda} \right) < \alpha_{i,\widehat{S}_\lambda}$$

with probability larger than $1 - \delta/2$. In the following lemma, we compare $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$ to $\mathcal{K}_1 + \mathcal{K}_2$. Note $R_{\Sigma^{(1)}, \Sigma^{(2)}} = \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})} \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]}$.

Lemma 8.8. *On the event $\mathcal{A} \cap \mathcal{B}$, we have*

$$L \left[\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \right] \geq 1 \wedge \left[\mathcal{K}_1 + \mathcal{K}_2 - L' R_{\Sigma^{(1)}, \Sigma^{(2)}} \frac{|S_U|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p) \right] .$$

Then, we closely follow the arguments of Theorem 5.1 to state that $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B$ rejects \mathbf{H}_0 with large probability as long as $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$ is large enough.

Lemma 8.9. *If on the event $\mathcal{A} \cap \mathcal{B}$, we have*

$$\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \geq L\varphi_{\widehat{S}_\lambda} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|\widehat{S}_\lambda| \log(p) + \log \left(\frac{1}{\alpha\delta} \right) + \log(p) \right] ,$$

then, $\min_{i \in \{1,2,3\}} \widetilde{Q}_{i,|\widehat{S}_\lambda|} (F_{\widehat{S}_\lambda,i} \mid \mathbf{X}_{\widehat{S}_\lambda}) < \alpha_{i,\widehat{S}_\lambda}$ with probability larger than $1 - \delta/2$.

We derive from (50) that on the event $\mathcal{A} \cap \mathcal{B}$,

$$|\widehat{S}_\lambda| \leq L' \frac{n_1 \vee n_2}{n_1 \wedge n_2} \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} |S_\cup|.$$

Since $|S_\cup| \leq |\beta^{(1)}|_0 + |\beta^{(2)}|_0$, it follows from Condition (25) that $|\widehat{S}_\lambda| \leq k_*$. Gathering Lemmas 8.8 and 8.9 allows us to conclude if we take L_3^* in Proposition 5.2 large enough.

Proof of Lemma 8.6. In order to bound $\mathbb{P}(\mathcal{A})$, we apply Lemma 8.12 to simultaneously control $\varphi_{\max}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})$, $\varphi_{\max}(\mathbf{X}_S^{(2)\top} \mathbf{X}_S^{(2)})$, $\varphi_{\min}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})$, and $\varphi_{\min}(\mathbf{X}_S^{(2)\top} \mathbf{X}_S^{(2)})$ for all sets S of size k_* . Combining a union bound with Conditions (23) and (24) allows us to prove that

$$\mathbb{P} \left[\left\{ \forall \theta \text{ s.t. } |\theta|_0 \leq k_*, \ 1/2 \leq \frac{\|\mathbf{X}^{(1)}\theta\|^2}{n_1 \|\theta\|_{\Sigma^{(1)}}^2} \leq 2 \text{ and } 1/2 \leq \frac{\|\mathbf{X}^{(2)}\theta\|^2}{n_2 \|\theta\|_{\Sigma^{(2)}}^2} \leq 2 \right\} \right] \geq 1 - \delta/8$$

Applying Corollary 1 in [30], we derive that there exist three positive constant c_1, c_2 and c_3 such that the following holds. With probability larger than $1 - c_1 \exp[-c_2(n_1 \wedge n_2)]$, we have

$$\bigwedge_{i=1,2} \frac{\kappa[6, |\theta_*|_0, \mathbf{X}^{(i)}/\sqrt{n_i}]}{\kappa[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} \geq 2^{-3},$$

if $|\theta_*|_0 \log(p) < c_3 \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} (n_1 \wedge n_2)$. Hence, we conclude that $\mathbb{P}[\mathcal{A}] \geq 1 - \delta/4$.

Consider an integer $k \leq k_*$ and a k -sparse vector $\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix}$ in \mathbb{R}^{2p} . Under event \mathcal{A} , we have

$$\begin{aligned} \|\mathbf{W}\theta\|^2 &= \|\mathbf{X}^{(1)}(\theta^{(1)} + \theta^{(2)})\|^2 + \|\mathbf{X}^{(2)}(\theta^{(1)} - \theta^{(2)})\|^2 \\ &\leq 2n_1 \|\theta^{(1)} + \theta^{(2)}\|_{\Sigma^{(1)}}^2 + 2n_2 \|\theta^{(1)} - \theta^{(2)}\|_{\Sigma^{(2)}}^2 \\ &\leq 4(n_1 + n_2) \left[\Phi_{k,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k,+}(\sqrt{\Sigma^{(2)}}) \right] \|\theta\|^2 \\ \|\mathbf{W}\theta\|^2 &\geq \frac{1}{2} \left[n_1 \|\theta^{(1)} + \theta^{(2)}\|_{\Sigma^{(1)}}^2 + n_2 \|\theta^{(1)} - \theta^{(2)}\|_{\Sigma^{(2)}}^2 \right] \\ &\geq (n_1 \wedge n_2) \left[\Phi_{k,-}(\sqrt{\Sigma^{(1)}}) \wedge \Phi_{k,-}(\sqrt{\Sigma^{(2)}}) \right] \|\theta\|^2. \end{aligned}$$

□

Proof of Lemma 8.7. Observe that the variance of $[\mathbf{W}^\top \epsilon]_i$ given \mathbf{W} is smaller than $\Phi_{1,+}(\mathbf{W})(\sigma^{(1)} \vee \sigma^{(2)})^2$. Using a union bound and the deviations of the Gaussian distribution, it follows that $\mathbb{P}(\mathcal{B}) \geq 1 - 1/p$.

Recall the definition of $\eta[\cdot, \cdot]$ in (22). A slight variation of Lemma 11.2 in [38] ensures that

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq L \frac{\lambda^2}{\eta^2[3, |\theta_*|_0, \mathbf{W}]} |\theta_*|_0 \quad (51)$$

on event \mathcal{B} . Fix $k = |\theta_*|_0$ and consider some $\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix} \in \mathcal{C}(3, T)$ with $|T| = k$. Define $T' \subset \{1, \dots, p\}$ by $i \in T'$ if $i \in T$ or $i + p \in T$. We have

$$\begin{aligned} |(\theta^{(1)} + \theta^{(2)})_{T^c}|_1 \vee |(\theta^{(1)} - \theta^{(2)})_{T^c}|_1 &\leq |\theta_{T^c}^{(1)}|_1 + |\theta_{T^c}^{(2)}|_1 \leq |\theta_{T^c}|_1 \leq 3|\theta_T|_1 \\ &\leq 3 \left[|\theta_{T'}^{(1)}|_1 + |\theta_{T'}^{(2)}|_1 \right] \\ &\leq 6 \left[|(\theta^{(1)} + \theta^{(2)})_{T'}|_1 \vee |(\theta^{(1)} - \theta^{(2)})_{T'}|_1 \right] \end{aligned}$$

It follows that $\theta^{(1)} + \theta^{(2)} \in \mathcal{C}(6, T')$ or $\theta^{(1)} - \theta^{(2)} \in \mathcal{C}(6, T')$. By symmetry, we assume that $|(\theta^{(1)} + \theta^{(2)})_{T'}|_1 \geq |(\theta^{(1)} - \theta^{(2)})_{T'}|_1$. Let us lower bound the l_1 norm of $\theta^{(1)} + \theta^{(2)}$ in terms of θ .

$$2|\theta^{(1)} + \theta^{(2)}|_1 \geq \left[|(\theta^{(1)} + \theta^{(2)})_{T'}|_1 + |(\theta^{(1)} - \theta^{(2)})_{T'}|_1 \right] \geq |\theta_T|_1 \geq \frac{|\theta|_1}{4},$$

since θ belongs to $\mathcal{C}(3, T)$. Thus, we derive the lower bound

$$\begin{aligned} \frac{k\|\mathbf{W}\theta\|^2}{|\theta|_1^2} &\geq \frac{k\|\mathbf{X}^{(1)}(\theta^{(2)} + \theta^{(1)})\|^2}{|\theta|_1^2} + \frac{k\|\mathbf{X}^{(2)}(\theta^{(2)} - \theta^{(1)})\|^2}{|\theta|_1^2} \\ &\geq \frac{(n_1 \wedge n_2)|\theta^{(2)} + \theta^{(1)}|_1^2}{|\theta|_1^2} \left[\bigwedge_{i=1,2} \eta^2(6, k, \mathbf{X}^{(i)}/\sqrt{n_i}) \right] \\ &\geq L(n_1 \wedge n_2) \left[\bigwedge_{i=1,2} \kappa^2(6, k, \mathbf{X}^{(i)}/\sqrt{n_i}) \right] \\ &\geq L(n_1 \wedge n_2) \left[\kappa^2(6, k, \sqrt{\Sigma^{(1)}}) \wedge \kappa^2(6, k, \sqrt{\Sigma^{(2)}}) \right], \end{aligned}$$

where the last inequality proceeds from Lemma 8.6. We conclude that

$$L'\kappa^2[3, |\theta_*|_0, \mathbf{W}] \geq (n_1 \wedge n_2) \left[\kappa^2(6, k, \sqrt{\Sigma^{(1)}}) \wedge \kappa^2(6, k, \sqrt{\Sigma^{(2)}}) \right].$$

Gathering this bound with (51), it follows that

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq \frac{L'\lambda^2/(n_1 \wedge n_2)}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0,$$

which allows us to prove (49). Lemma 3.1 in [16] tells us that on event \mathcal{B} ,

$$\lambda^2|\widehat{\theta}_\lambda|_0 \leq 16\Phi_{|\widehat{\theta}_\lambda|_0, +}(\mathbf{W})\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2.$$

Gathering the last two bounds and Lemma 8.6, we obtain

$$|\widehat{\theta}_\lambda|_0 \leq L \frac{\Phi_{|\widehat{\theta}_\lambda|_0, +}(\mathbf{W})}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0. \quad (52)$$

Recall that $|\theta_*|_0 \leq |\beta^{(1)}|_0 + |\beta^{(2)}|_0$. The upper-bound $\Phi_{|\widehat{\theta}_\lambda|_0, +}(\mathbf{W}) \leq (1 + |\widehat{\theta}_\lambda|_0/k_*)\Phi_{k_*, +}(\mathbf{W})$ and Lemma 8.6 enforce

$$\begin{aligned} |\widehat{\theta}_\lambda|_0 &\leq L \frac{n_1 \vee n_2}{n_1 \wedge n_2} \frac{\Phi_{k_*, +}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k_*, +}(\sqrt{\Sigma^{(2)}})}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0 \left[1 + \frac{|\widehat{\theta}_\lambda|_0}{k_*} \right] \\ &\leq (k_* + |\widehat{\theta}_\lambda|_0)/2, \end{aligned}$$

where the last inequality holds if we take L_*^* in (25) small enough. Hence, $|\widehat{\theta}_\lambda|_0 \leq k_*$. Coming back to (52), we prove (50). \square

Proof of Lemma 8.8. Given the Lasso estimator $\widehat{\theta}_\lambda$ of θ_* in model (48), we define $\widehat{\beta}_\lambda^{(1)}$ and $\widehat{\beta}_\lambda^{(2)}$ by

$$\widehat{\beta}_\lambda^{(1)} = \widehat{\theta}_\lambda^{(1)} + \widehat{\theta}_\lambda^{(2)}, \quad \widehat{\beta}_\lambda^{(2)} = \widehat{\theta}_\lambda^{(1)} - \widehat{\theta}_\lambda^{(2)}.$$

On event $\mathcal{A} \cap \mathcal{B}$, we upper bound the difference between $(\beta^{(1)}, \beta^{(2)})$ and $(\widehat{\beta}_\lambda^{(1)}, \widehat{\beta}_\lambda^{(2)})$.

$$\begin{aligned} & \|\beta^{(1)} - \widehat{\beta}_\lambda^{(1)}\|_{\Sigma^{(1)}}^2 + \|\beta^{(2)} - \widehat{\beta}_\lambda^{(2)}\|_{\Sigma^{(2)}}^2 \\ & \leq 2 \left[\left\| \frac{\mathbf{X}^{(1)}}{\sqrt{n_1}} (\beta^{(1)} - \widehat{\beta}_\lambda^{(1)}) \right\|^2 + \left\| \frac{\mathbf{X}^{(2)}}{\sqrt{n_2}} (\beta^{(2)} - \widehat{\beta}_\lambda^{(2)}) \right\|^2 \right] \\ & \leq \frac{2}{n_1 \wedge n_2} \|\mathbf{W}(\theta_* - \widehat{\theta}_\lambda)\|^2 \\ & \leq L \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} \frac{|S_{\cup}|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p) (\sigma^{(1)} \vee \sigma^{(2)})^2, \end{aligned}$$

where the last inequality follows from Lemma 8.7. Let us now lower bound the Kullback discrepancy $2[\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)]$ which equals

$$\left(\frac{\sigma_{\widehat{S}_\lambda}^{(1)}}{\sigma_{\widehat{S}_\lambda}^{(2)}} \right)^2 + \left(\frac{\sigma_{\widehat{S}_\lambda}^{(1)}}{\sigma_{\widehat{S}_\lambda}^{(2)}} \right)^2 - 2 + \frac{\|\beta_{\widehat{S}_\lambda}^{(2)} - \beta_{\widehat{S}_\lambda}^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_{\widehat{S}_\lambda}^{(1)})^2} + \frac{\|\beta_{\widehat{S}_\lambda}^{(2)} - \beta_{\widehat{S}_\lambda}^{(1)}\|_{\Sigma^{(1)}}^2}{(\sigma_{\widehat{S}_\lambda}^{(2)})^2}.$$

CASE 1: $\frac{\sigma^{(1)} \vee \sigma^{(2)}}{\sigma^{(1)} \wedge \sigma^{(2)}} \geq \sqrt{2}$. By symmetry, we can assume that $\sigma^{(1)} > \sigma^{(2)}$.

$$\begin{aligned} (\sigma_{\widehat{S}_\lambda}^{(1)})^2 &= (\sigma^{(1)})^2 + \|\beta^{(1)} - \beta_{\widehat{S}_\lambda}^{(1)}\|_{\Sigma^{(1)}}^2 \geq (\sigma^{(1)})^2 \\ (\sigma_{\widehat{S}_\lambda}^{(2)})^2 &= (\sigma^{(2)})^2 + \|\beta^{(2)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(2)}}^2 \leq (\sigma^{(2)})^2 + \|\beta^{(2)} - \widehat{\beta}_\lambda^{(2)}\|_{\Sigma^{(2)}}^2 \\ &\leq (\sigma^{(2)})^2 + L \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} \frac{|S_{\cup}|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p) (\sigma^{(1)})^2 \quad (53) \\ &\leq (\sigma^{(2)})^2 + \frac{(\sigma^{(1)})^2}{4}, \end{aligned}$$

where we used conditions (23) and (25) in the last inequality assuming that we have taken L^* and L_2^* small enough in these two conditions. This enforces

$$2 \left[\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \right] \geq \frac{1}{12}.$$

CASE 2: $\frac{\sigma^{(1)} \vee \sigma^{(2)}}{\sigma^{(1)} \wedge \sigma^{(2)}} \leq \sqrt{2}$. Let us note

$$A = 2L \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} \frac{|S_{\cup}|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p),$$

with L as in (53). Arguing as in Case 1, we derive that

$$\begin{aligned} (\sigma_{\widehat{S}_\lambda}^{(1)})^2 &\leq (\sigma^{(1)})^2 [1 + A] \leq 2(\sigma^{(1)})^2, \\ (\sigma_{\widehat{S}_\lambda}^{(1)})^2 &\leq (\sigma^{(2)})^2 [1 + A] \leq 2(\sigma^{(2)})^2. \end{aligned}$$

Let us lower bound $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$ in terms of $\mathcal{K}_1 + \mathcal{K}_2$. First, we consider the ratio of

the variances

$$\begin{aligned}
\frac{(\sigma_{\widehat{S}_\lambda}^{(1)})^2}{(\sigma_{\widehat{S}_\lambda}^{(2)})^2} + \frac{(\sigma_{\widehat{S}_\lambda}^{(2)})^2}{(\sigma_{\widehat{S}_\lambda}^{(1)})^2} - 2 &\geq \left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} \right] / (1+A) - 2 \\
&\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} - 2 - \frac{A}{1+A} \left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} \right] \\
&\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} - 2 - 3A. \tag{54}
\end{aligned}$$

Let us now lower bound the remaining part of $\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda)$. For $i = 1, 2$, $|\beta^{(i)} - \widehat{\beta}_\lambda^{(i)}|_0 \leq |\theta_*|_0 + |\widehat{\theta}_\lambda|_0 \leq k_*$ by Lemma 8.7 and Condition (25).

$$\begin{aligned}
&\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} + \frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} \\
&\leq \frac{3}{(\sigma^{(1)})^2 \wedge (\sigma^{(2)})^2} \sum_{i=1}^2 \left[\|\beta^{(1)} - \beta_{\widehat{S}_\lambda}^{(1)}\|_{\Sigma^{(i)}}^2 + \|\beta^{(2)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(i)}}^2 + \|\beta_{\widehat{S}_\lambda}^{(1)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(i)}}^2 \right] \\
&\leq L_1 \left[\frac{\|\beta_{\widehat{S}_\lambda}^{(1)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} + \frac{\|\beta_{\widehat{S}_\lambda}^{(1)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} \right] \\
&\quad + \frac{L_2}{(\sigma^{(1)} \wedge \sigma^{(2)})^2} \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} \left[\sum_{i=1}^2 \|\beta^{(i)} - \widehat{\beta}_\lambda^{(i)}\|_{\Sigma^{(i)}}^2 \right] \\
&\leq L_1 \left[\frac{\|\beta_{\widehat{S}_\lambda}^{(1)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} + \frac{\|\beta_{\widehat{S}_\lambda}^{(1)} - \beta_{\widehat{S}_\lambda}^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} \right] + L_2 \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} A
\end{aligned}$$

Gathering the last inequality with (54) yields

$$\mathcal{K}_1(\widehat{S}_\lambda) + \mathcal{K}_2(\widehat{S}_\lambda) \geq L_1 [\mathcal{K}_1 + \mathcal{K}_2] - L_2 \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} A.$$

□

Proof of Lemma 8.9. For any non empty set S of size smaller or equal to k_* , define $\delta_S = \delta \left(2 \binom{|S|}{p} k_* \right)^{-1}$. If we take L^* and L_1^* in (23-24) small enough, then $1 + \log[1/(\alpha_S \delta_S)] / (n_1 \wedge n_2)$ is smaller than some constant L small enough so that we can apply Theorem 5.1. Arguing as in the proof of this Theorem, we derive that

$$\mathbb{P} \left[\min_{i \in \{1,2,3\}} \widetilde{Q}_{i,S}(F_{S,i} | \mathbf{X}_S) < \alpha_S \right] \geq 1 - \delta_S$$

if

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \geq L\varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| \log(p) + \log \left(\frac{1}{\alpha_S} \right) + \log(p) \right].$$

Applying a union bound over all sets S of size smaller or equal to k_* allows us to prove

$$\mathbb{P} \left[\min_{i \in \{1,2,3\}} \widetilde{Q}_{i,\widehat{S}_\lambda}(F_{\widehat{S}_\lambda,i} | \mathbf{X}_{\widehat{S}_\lambda}) < \alpha_{\widehat{S}_\lambda} \right] \geq 1 - \delta.$$

□

8.6. Proof of Proposition 5.3

This proof follows the same steps as above. Taking \tilde{L}^* small enough, we can assume that $n_1 \vee n_2 \leq 2(n_1 \wedge n_2)$. Rewrite the linear regression model $\mathbf{Y} = \mathbf{W}\theta_* + \epsilon$ as follows:

$$\mathbf{Y} = \mathbf{W}^{(1)}\theta_*^{(1)} + \mathbf{W}^{(2)}\theta_*^{(2)} + \epsilon .$$

From the definition of the Lasso estimator $\hat{\theta}_\lambda = \begin{pmatrix} \hat{\theta}_\lambda^{(1)} \\ \hat{\theta}_\lambda^{(2)} \end{pmatrix}$, we derive that $\hat{\theta}_\lambda^{(2)}$ is the solution of the following minimization problem:

$$\arg \min_{\theta \in \mathbb{R}^p} \|\epsilon + \mathbf{W}^{(2)}\theta_*^{(2)} + \mathbf{W}^{(1)}(\theta_*^{(1)} - \hat{\theta}_\lambda^{(1)}) - \mathbf{W}^{(2)}\theta\| + \lambda|\theta'|_1 . \quad (55)$$

We fix

$$\lambda = 16(\sigma^{(1)} \vee \sigma^{(2)})\sqrt{2(n_1 + n_2)\Phi_{1,+}(\sqrt{\Sigma})\log(p)} .$$

and we suppose that event $\mathcal{A} \cap \mathcal{B}$ (defined in the last proof) holds. Recall that $\mathbb{P}[\mathcal{A} \cap \mathcal{B}] \geq 1 - \delta/4 - 1/p$. We consider the set $\hat{\mathcal{S}}_\lambda^{(2)}$ defined as the support of $\hat{\theta}_\lambda^{(2)}$. Note that $\hat{\mathcal{S}}_\lambda^{(2)} \in \hat{\mathcal{S}}_L^{(2)} \subset \hat{\mathcal{S}}_{\text{Lasso}}$.

Lemma 8.10. *If we take constants \tilde{L}^* and L_2^* in Proposition 5.3 small enough, then the following holds. There exists an \mathcal{C} of probability larger than $1 - 1/p$ such that, under $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$, we have*

$$|\mathbf{W}^{(2)\top}\mathbf{W}^{(1)}(\theta_*^{(1)} - \hat{\theta}_\lambda^{(1)})|_\infty \leq \lambda/8 \quad (56)$$

It follows that on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$:

$$\left| \mathbf{W}^{(2)\top} \left[\epsilon + \mathbf{W}^{(1)}(\theta_*^{(1)} - \hat{\theta}_\lambda^{(1)}) \right] \right|_\infty \leq \lambda/4$$

Arguing as in the proof of Lemma 8.7 and taking L_2^* small enough, we derive that under $\mathcal{A} \cap \mathcal{B}$,

$$\|\mathbf{W}^{(2)}(\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)})\|^2 \leq L_1 \frac{\lambda^2/(n_1 \wedge n_2)}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} |\theta_*^{(2)}|_0 , \quad (57)$$

$$|\hat{\theta}_\lambda^{(2)}|_0 \leq L_2 \frac{\Phi_{k_*,+}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} |\theta_*^{(2)}|_0 \leq \tilde{k}_*/2 \leq k_*/2 . \quad (58)$$

This allows us to upper bound $\|\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)}\|_\Sigma^2$ under event \mathcal{A} .

$$\begin{aligned} \|\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)}\|_\Sigma^2 &\leq \frac{L}{n_1 \wedge n_2} \left[\|\mathbf{X}^{(1)}(\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)})\|^2 + \|\mathbf{X}^{(2)}(\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)})\|^2 \right] \\ &\leq \frac{L}{n_1 \wedge n_2} \|\mathbf{W}^{(2)}(\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)})\|^2 . \end{aligned}$$

Pythagorean inequality then gives

$$\begin{aligned} \|\beta^{(1)} - \beta^{(2)}\|_\Sigma^2 &= \|\beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(1)} - \beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(2)}\|_\Sigma^2 + \|\beta^{(1)} - \beta^{(2)} - \beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(1)} + \beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(2)}\|_\Sigma^2 \\ &\leq \|\beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(1)} - \beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(2)}\|_\Sigma^2 + \|\theta_*^{(2)} - \hat{\theta}_\lambda^{(2)}\|_\Sigma^2 \\ &\leq \|\beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(1)} - \beta_{\hat{\mathcal{S}}_\lambda^{(2)}}^{(2)}\|_\Sigma^2 + L \frac{|\theta_*^{(2)}|_0 \log(p)}{n_1 \wedge n_2} \frac{\Phi_{1,+}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} (\sigma^{(1)} \vee \sigma^{(2)})^2 , \end{aligned}$$

where we use the two previous upper bounds in the last line. Consequently, we obtain

$$\mathcal{K}_1(\widehat{S}_\lambda^{(2)}) + \mathcal{K}_2(\widehat{S}_\lambda^{(2)}) \geq L \frac{\|\beta^{(1)} - \beta^{(2)}\|_\Sigma^2}{\text{Var}(Y^{(1)}) \vee \text{Var}(Y^{(2)})} - L' \frac{|\theta_*^{(2)}|_0 \log(p)}{n_1 \wedge n_2} \frac{\Phi_{1,+}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]}.$$

Applying Lemma 8.9 to $\widehat{S}_\lambda^{(2)}$, using (58) and taking L_3^* large enough then allows us to conclude.

Proof of Lemma 8.10. Given any matrix A , we define the norm $\|A\|_\infty = \max_{i,j} |A_{i,j}|$. Suppose that we are under events $\mathcal{A} \cap \mathcal{B}$ defined previously. Arguing as in the proof of Lemma 8.7, we derive that $|\theta_*|_0 + |\widehat{\theta}_\lambda|_0 \leq \tilde{k}_*$ and

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq L_1 \frac{\lambda^2}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma}](n_1 \wedge n_2)} \tilde{k}_*. \quad (59)$$

Thus, $|\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)}|_0 \leq \tilde{k}_*$ and we derive

$$\begin{aligned} \left| \mathbf{W}^{(2)\top} \mathbf{W}^{(1)} \left(\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)} \right) \right|_\infty &= \left| \left(\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)} \right) \left(\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)} \right) \right|_\infty \\ &\leq \|\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)}\| \sqrt{\tilde{k}_*} \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \\ &\leq \frac{\|\mathbf{W}(\theta_* - \widehat{\theta})\|}{\sqrt{\Phi_{k_*, -}(\mathbf{W})}} \sqrt{\tilde{k}_*} \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \\ &\leq L \frac{\lambda \tilde{k}_* \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty}{\sqrt{n_1 \wedge n_2} \kappa[6, |\theta_*|_0, \sqrt{\Sigma}] \sqrt{\Phi_{k_*, -}(\mathbf{W})}}, \end{aligned} \quad (60)$$

where we used (59) in the last line.

Combining deviations inequality for χ^2 distributions (Lemma 8.11) and for Gaussian distributions and a union bound, we derive that

$$\|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \leq \Phi_{1,+}(\sqrt{\Sigma}) \left[|n_1 - n_2| + L \sqrt{(n_1 \vee n_2) \log(p)} \right], \quad (61)$$

with probability larger than $1 - 1/p$. Consider some θ with $|\theta|_0 \leq k_*$. When event \mathcal{A} defined in Lemma 8.6 holds, we have

$$\begin{aligned} \frac{\|\mathbf{W}\theta\|^2}{\|\theta\|^2} &= \frac{\|\mathbf{X}^{(1)}(\theta^{(1)} + \theta^{(2)})\|^2}{\|\theta\|^2} + \frac{\|\mathbf{X}^{(2)}(\theta^{(1)} - \theta^{(2)})\|^2}{\|\theta\|^2} \\ &\geq \frac{\Phi_{k_*, -}(\sqrt{\Sigma})}{2} \frac{n_1 \|\theta^{(1)} + \theta^{(2)}\|^2 + n_2 \|\theta^{(1)} - \theta^{(2)}\|^2}{\|\theta\|^2} \\ &\geq \Phi_{k_*, -}(\sqrt{\Sigma})(n_1 \wedge n_2). \end{aligned}$$

Let us note $T_\Sigma = \frac{\Phi_{1,+}(\sqrt{\Sigma})}{\kappa[6, k_*, \sqrt{\Sigma}] \Phi_{k_*, -}^{1/2}(\sqrt{\Sigma})}$. Gathering the last upper bound with (60) and (61), we get

$$\left| \mathbf{W}^{(2)\top} \mathbf{W}^{(1)} \left(\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)} \right) \right|_\infty \leq L \lambda \tilde{k}_* \left[\frac{|n_1 - n_2|}{n_1 \wedge n_2} + \sqrt{\frac{\log(p)}{n_1 \wedge n_2}} \right] T_\Sigma,$$

since $n_1 \vee n_2 \leq 2(n_1 \wedge n_2)$. Taking \tilde{L}^* small enough in definition (26) of \tilde{k}_* allows us to conclude. \square

8.7. Proof of Proposition 3.2

By symmetry, we can assume that $n_1 \leq n_2$. Let us fix $\beta^{(2)} = 0$ and $\sigma^{(2)} = 1$. Fix some positive integer $s \leq p^{1/2-\gamma}$ and fix $r \in (0, 1/\sqrt{2})$.

We consider the test of hypotheses $\mathcal{H}_0 : \beta^{(1)} = 0, \sigma^{(1)} = 1$ against $\mathcal{H}_1 : |\beta^{(1)}|_0 = s, \|\beta^{(1)}\| = r^2$, and $\sigma^{(1)} = \sqrt{1-r^2}$. Note that for this problem, the data $(\mathbf{Y}^{(2)}, \mathbf{X}^{(2)})$ do not bring any information on the hypotheses. This one-sample testing problem is a specific case of the two-sample testing problem considered in the proposition. Thus, a minimax lower bound for the one-sample problem provides us a minimax lower bound for the two-sample problem.

According to Theorem 4.3 in [41], no level α test has power larger than $1 - \delta$ if

$$\frac{r^2}{1-r^2} \leq \frac{s}{2n_1} \log \left(1 + \frac{p}{s^2} + \sqrt{\frac{2p}{s^2}} \right)$$

Since $s \leq p^{1/2-\gamma}$, no level α test has power larger than $1 - \delta$ if

$$\frac{r^2}{1-r^2} \leq \gamma \frac{|s|}{n_1} \log(p) . \quad (62)$$

By Assumption (A.2), one may assume that that the right-hand side term is smaller than $1/2$. Observe that

$$2(\mathcal{K}_1 + \mathcal{K}_2) = \frac{2r^2}{1-r^2} \quad \text{and} \quad \frac{\|\beta^{(1)} - \beta^{(2)}\|_{I_p}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} = r^2 \geq \frac{1}{2} \frac{r^2}{1-r^2} ,$$

for $r \leq \sqrt{2}$. The result follows.

8.8. Technical lemmas

In this section, some useful deviation inequalities for χ^2 random variables [22] and for Wishart matrices [13] are reminded.

Lemma 8.11. *For any integer $d > 0$ and any positive number x ,*

$$\begin{aligned} \mathbb{P} \left(\chi^2(d) \leq d - 2\sqrt{dx} \right) &\leq \exp(-x) , \\ \mathbb{P} \left(\chi^2(d) \geq d + 2\sqrt{dx} + 2x \right) &\leq \exp(-x) . \end{aligned}$$

Lemma 8.12. *Let $Z^\top Z$ be a standard Wishart matrix of parameters (n, d) with $n > d$. For any positive number x ,*

$$\mathbb{P} \left\{ \varphi_{\min}(Z^\top Z) \geq n \left(\left\{ 1 - \sqrt{\frac{d}{n}} - x \right\} \vee 0 \right) \right\} \leq \exp(-nx^2/2) ,$$

and

$$\mathbb{P} \left[\varphi_{\max}(Z^\top Z) \leq n \left(1 + \sqrt{\frac{d}{n}} + x \right)^2 \right] \leq \exp(-nx^2/2) .$$

Acknowledgements

We are grateful to Christophe Giraud for fruitful discussions. The research of N. Verzelen is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration).

References

- [1] AMBROISE, C., CHIQUET, J., AND MATIAS, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electronic Journal of Statistics* 3, 205–238.
- [2] ANTONIADIS, A. (2010). Comments on: l1-penalization for mixture regression models. *Test* 19, 257–258.
- [3] ARIAS-CASTRO, E., CANDÈS, E., AND PLAN, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *Annals of Statistics* 39, 2533–2556.
- [4] BAI, Z. AND SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6, 311–329.
- [5] BARAUD, Y., HUET, S., AND LAURENT, B. (2003). Adaptive tests of linear hypotheses by model selection. *Annals of Statistics* 31, 225–251.
- [6] BICKEL, P., RITOV, Y., AND TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* 37, 1705–1732.
- [7] BÜHLMANN, P. (2012). Statistical significance in high-dimensional linear models. arXiv:1202.137.
- [8] CAI, T., LIU, W., AND XIA, Y. (2011). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings.
- [9] CANDÈS, E. AND PLAN, Y. (2007). Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics* 37, 2145–2177.
- [10] CHEN, S. AND QIN, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* 38, 808–835.
- [11] CHIQUET, J., GRANDVALET, Y., AND AMBROISE, C. (2011). Inferring multiple graphical structures. *Statistics and Computing* 21, 4, 537–553.
- [12] CHUNG, S., SUZUKI, H., MIYAMOTO, T., TAKAMATSU, N., TATSUGUCHI, A., UEDA, K., KIJIMA, K., NAKAMURA, Y., AND MATSUO, Y. (2012). Development of an orally-administrative melk-targeting inhibitor that suppresses the growth of various types of human cancer. *Oncotarget* 3, 1629–40.
- [13] DAVIDSON, K. R. AND SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces, Vol. I*. North-Holland, Amsterdam, 317–366. [MR1863696 \(2004f:47002a\)](#)
- [14] DONOHO, D. AND JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixture. *Annals of Statistics* 32, 962–994.
- [15] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* 9, 3, 432–441.
- [16] GIRAUD, C., HUET, S., AND VERZELEN, N. (2012). Supplement to ‘High-dimensional regression with unknown variance’.
- [17] HEIDEL, J., LIU, J., YEN, Y., ZHOU, B., HEALE, B., ROSSI, J., BARTLETT, D., AND DAVIS, M. (2007). Potent sirna inhibitors of ribonucleotide reductase subunit rrm2 reduce cell proliferation in vitro and in vivo. *Clinical Cancer Research* 13.
- [18] HESS, K., ANDERSON, K., SYMMANS, W., VALERO, V., IBRAHIM, N., MEJIA, J., BOOSER, D., THERIAULT, R., BUZDAR, U., DEMPSEY, P., ROUZIER, R., SNEIGE, N., ROSS, J., VIDAURRE, T., GÓMEZ, H., HORTOBAGYI, G., AND PUSTZAI, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology* 24, 26, 4236–4244.
- [19] INGSTER, Y., TSYBAKOV, A., AND VERZELEN, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics* 4, 1476–1526.
- [20] JEANMOUGIN, M., GUEDJ, M., AND AMBROISE, C. (2011). Defining a robust biological prior from pathway analysis to drive network inference. *Journal de la Société*

- Française de Statistique* 152, 97–110.
- [21] KYUNG, M., GILL, J., GHOSH, M., AND CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis* 5, 369–412.
- [22] LAURENT, B. AND MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 28, 5, 1302–1338. [MR1805785 \(2002c:62052\)](#)
- [23] LI, J. AND CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* 40, 2, 908–940. <http://dx.doi.org/10.1214/12-AOS993>. [MR2985938](#)
- [24] LIN, Y., CHEN, C., CHENG, C., AND YANG, R. (2011). Domain and functional analysis of a novel breast tumor suppressor protein, scube2. *Journal of Biological Chemistry* 29, 27039–47.
- [25] LOPES, M., JACOB, L., AND WAINWRIGHT, M. (2011). A more powerful two-sample test in high dimensions using random projection. In *NIPS*.
- [26] MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* 34, 3, 1436–1462. [MR2278363](#)
- [27] MEINSHAUSEN, N., MEIER, L., AND BÜHLMANN, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* 104, 1671–1681.
- [28] MEINSHAUSEN, N. AND YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of statistics* 37, 246–270.
- [29] NATOWICZ, R., INCITTI, R., HORTA, E., CHARLES, B., GUINOT, P., YAN, K., COUTANT, C., ANDRE, F., PUSZTAI, L., AND ROUZIER, R. (2008). Prediction of the outcome of preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete responses. *BMC Bioinformatics* 9.
- [30] RASKUTTI, G., WAINWRIGHT, M., AND YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research* 11, 2241–2259. [MR2719855 \(2011h:62272\)](#)
- [31] ROUZIER, R., PEROU, C., SYMMANS, F., IBRAHIM, N., CRISTOFANILLI, M., ANDERSON, K., HESS, K., STEC, J., AYERS, M., WAGNER, P., MORANDI, P., FAN, C., RABIUL, I., ROSS, J. S., HORTOBAGYI, G., AND PUSZTAI, L. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research* 11.
- [32] ROUZIER, R., RAJAN, R., WAGNER, P., HESS, K., GOLD, D., STEC, J., AYERS, M., ROSS, J., ZHANG, P., BUCHHOLZ, T., KUERER, H., GREEN, M., ARUN, B., HORTOBAGYI, G., SYMMANS, W., AND PUSZTAI, L. (2005). Microtubule-associated protein tau: a marker of paclitaxel sensitivity in breast cancer. *Proceedings of the National Academy of Sciences* 102, 8315–8320.
- [33] SRIVASTAVA, M. AND DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* 99, 386–402.
- [34] STÄDLER, N. AND MUKHERJEE, S. (2012). Two-sample testing in high-dimensional models.
- [35] SUN, T. AND ZHANG, C. (2010). Comments on: l1-penalization for mixture regression models. *Test* 19, 270–275.
- [36] SUN, T. AND ZHANG, C. (2011). Scaled sparse linear regression. [arXiv:1104.4595](#).
- [37] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.
- [38] VAN DE GEER, S. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* 3, 1360–1392.
- [39] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high-dimensional phenomena. *Electron. J. Stat.* 6, 38–90.
- [40] VERZELEN, N. AND VILLERS, F. (2009). Tests for gaussian graphical models. *Com-*

- put. Statist. Data Anal.* 53, 1894–1905.
- [41] VERZELEN, N. AND VILLERS, F. (2010). Goodness-of-fit tests for high-dimensional gaussian linear models. *Annals of Statistics* 38, 704–752.
 - [42] WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* 55.
 - [43] WASSERMAN, L. AND ROEDER, K. (2009). High dimensional variable selection. *Annals of Statistics* 37, 2178–2201.
 - [44] ZHANG, C. AND ZHANG, S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. arXiv:1110.2563.