

A global homogeneity test for high-dimensional linear regression

Camille Charbonnier

Inserm U1079, CNR-MAJ Rouen University Hospital

22 boulevard Gambetta

76183 Rouen, France

e-mail: camille.charbonnier1@univ-rouen.fr

Nicolas Verzelen

INRA, UMR 729 MISTEA

2 Place Viala, BÂT. 29

F-34060 Montpellier, France

e-mail: nicolas.verzelen@inra.fr

and

Fanny Villers

LPMA, University Pierre et Marie Curie

4 place Jussieu

75005 Paris, France

e-mail: fanny.villers@upmc.fr

Abstract: This paper is motivated by the comparison of genetic networks inferred from high-dimensional datasets originating from high-throughput Omics technologies. 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.

Received August 2013.

1. Introduction

The recent flood of high-dimensional data has motivated the development of a vast range of sparse estimators for linear regressions, in particular a large variety of derivatives from the Lasso. Although theoretical guarantees have been provided in terms of prediction, estimation and selection performances (among a lot of others [6, 36, 50]), the research effort has only recently turned to the construction of high-dimensional confidence intervals or parametric hypothesis testing schemes [7, 25, 31, 33, 52]. Yet, quantifying the confidence surrounding coefficient estimates and selected covariates is essential in areas of application where these will nourish further targeted investigations.

In this paper we consider the two-sample linear regression model with Gaussian random design.

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

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

In this statistical model, 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.

Suppose that we observe an n_1 -sample of $(Y^{(1)}, X^{(1)})$ and an n_2 -sample of $(Y^{(2)}, X^{(2)})$ noted $\mathbf{Y}^{(1)}, \mathbf{X}^{(1)}$, and $\mathbf{Y}^{(2)}, \mathbf{X}^{(2)}$, with n_1 and n_2 remaining smaller than p . Defining $\epsilon^{(1)}$ and $\epsilon^{(2)}$ analogously, we obtain the decompositions $\mathbf{Y}^{(1)} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon^{(1)}$ and $\mathbf{Y}^{(2)} = \mathbf{X}^{(2)}\beta^{(2)} + \epsilon^{(2)}$. Given these observations, we want to test whether models (1) and (2) are the same, that is

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

In the null hypothesis, we include the assumption that the population covariances of the covariates are equal ($\Sigma^{(1)} = \Sigma^{(2)}$), whereas under the alternative hypothesis the population covariances are not required to be the same. This choice of assumptions is primarily motivated by our final objective to derive homogeneity tests for Gaussian graphical models (see below). A discussion of the design hypotheses is deferred to Section 7.

1.1. Connection with two-sample Gaussian graphical model testing

This testing framework is mainly motivated by the validation of differences observed between Gaussian graphical models (modelling regulation networks) inferred from high-throughput Omics data from two samples [12, 18, 34] when looking for new potential drug or knock-out targets [26]. Following the development of univariate differential analysis techniques, there is now a surging

demand for the detection of differential regulations between pairs of conditions (treated vs. placebo, diseased vs. healthy, exposed vs. controls ...).

The special case when the structure of the network itself is known and only edge weights between genes have to be estimated has already been addressed by the literature (e.g. [41]). Only a few papers have tackled the issue of comparing two estimated networks, facing the difficulty of disentangling estimation errors from true differences in the underlying networks. A first attempt at solving this problem can be found in [19]. Assuming the issue of high-dimensional network estimation solved by use of correlations, regularized partial correlations or partial least square estimation, the authors defined three measures of the differences between networks either in terms of modularity changes or of average edge weights between genes. Another work [13] derived a testing strategy from empirical Bayes posterior distributions of edge weights. However, none of these two papers provide theoretical guarantees on type-I error control or on detection properties of the suggested methods, in particular in a high-dimensional context.

We suggest to build upon our two-sample high-dimensional linear regression testing scheme to derive a global test for the equality of Gaussian graphical models inferred under pairs of conditions. We keep in mind our two main objectives: meeting the practical challenges associated with high-dimension and characterizing the performances of our procedure from a theoretical point of view.

Formally speaking, the global two-sample GGM testing problem is defined as follows. Consider two Gaussian random vectors $Z^{(1)} \sim \mathcal{N}(0, [\Omega^{(1)}]^{-1})$ and $Z^{(2)} \sim \mathcal{N}(0, [\Omega^{(2)}]^{-1})$. The conditional independence graphs are characterized by the non-zero entries of the precision matrices $\Omega^{(1)}$ and $\Omega^{(2)}$ [28]. Given an n_1 -sample of $Z^{(1)}$ and an n_2 -sample of $Z^{(2)}$, the objective is to test

$$\mathcal{H}_0^G : \Omega^{(1)} = \Omega^{(2)} \quad \text{versus} \quad \mathcal{H}_1^G : \Omega^{(1)} \neq \Omega^{(2)}, \quad (4)$$

where $\Omega^{(1)}$ and $\Omega^{(2)}$ are assumed to be sparse (most of their entries are zero). This testing problem is global as the objective is to assess a statistically significant difference between the two distributions. If the test is rejected, a more ambitious objective is to infer the entries where the precision matrices differ (i.e. $\Omega_{i,j}^{(1)} \neq \Omega_{i,j}^{(2)}$).

Adopting a neighborhood selection approach [34] as recalled in Section 6, high-dimensional GGM estimation can be solved by multiple high-dimensional linear regressions. As such, two-sample GGM testing (4) can be solved via multiple two-sample hypothesis testing as (3) in the usual linear regression framework. This extension of two-sample linear regression tests to GGMs is described in Section 6.

1.2. Related work

The literature on high-dimensional two-sample tests is very light. In the context of high-dimensional two-sample comparison of means, [3, 11, 32, 42] have introduced global tests to compare the means of two high-dimensional Gaussian vectors with unknown variance. Recently, [8, 29] developed two-sample tests 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: a local and a global approach. In the local approach, one considers the p tests for the nullity of each coefficient $\mathcal{H}_{0,i} : \beta_i^{(1)} = 0$ ($i = 1, \dots, p$) with the purpose of controlling error measures such as the false discovery rate of the resulting multiple testing procedures. In a way, one aims to assess the individual statistical significance of each of the variables. This can be achieved by providing a confidence region for $\beta^{(1)}$ [7, 25, 31, 33, 52]. Another line of work derives p -values for the nullity of each of the coefficients. Namely, [51] suggests a screen and clean procedure based upon half-sampling. Model selection is first applied upon a random half of the sample in order to test for the significance of each coefficient using the usual combination of ordinary least squares and Student t -tests on a model of reasonable size on the remaining second half. To reduce the dependency of the results to the splitting, [35] advocate to use half-sampling B times and then 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 $\mathcal{H}_0 : \beta^{(1)} = 0$. Although global approaches are clearly less informative than approaches providing individual significance tests like [7, 35, 52], they can reach better performances from smaller sample sizes. Such a property is of tremendous importance when dealing with high-dimensional datasets. The idea of [49], 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 sizes. The null hypothesis is rejected if the null hypothesis \mathcal{H}_0^S is rejected for at least one of the subsets $S \in \mathcal{S}$. Admittedly, the resulting procedure is computationally intensive. Nonetheless 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 higher-criticism. This last testing framework was originally introduced in orthonormal designs [16], but has been proved to reach optimal detection rates in high-dimensional linear regression as well [2, 24]. 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.

The testing strategy we describe in this paper is built upon the global approach developed by [5] and [49].

While writing this paper, we came across the parallel work of Städler and Mukherjee, which deals similarly with both frameworks of two-sample linear regression [43] and two-sample graphical model comparison [44]. Their approach

shares the idea of sample-splitting and dimensionality reduction with the *screen and clean* procedure in its simple-split [51] and multi-split [35] versions, in a global testing perspective however. A detailed comparison of [43, 44] with our contribution is deferred to simulations (Section 5) and discussion (Section 7).

1.3. Our contribution

Our suggested approach stems from the fundamental assumption that either the true supports of $\beta^{(1)}$ and $\beta^{(2)}$ are sparse or 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 that would know in advance the optimal low-dimensional subset S^* .

Concretely, we proceed in three steps:

1. We define algorithms to select a data-driven collection of subsets $\widehat{\mathcal{S}}$ identified as most informative for our testing problem, in an attempt to circumvent the optimal subset S^* .
2. New parametric statistics related to the likelihood ratio statistic between the conditional distributions $Y^{(1)}|X_S^{(1)}$ and $Y^{(2)}|X_S^{(2)}$ are defined for $S \in \widehat{\mathcal{S}}$.
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, allowing us to prove that this procedure is min-max adaptive to the sparsity of $\beta^{(1)}$ and $\beta^{(2)}$ from a non-asymptotic point of view;
 - 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.

The resulting testing procedure 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.

The procedure is described in Section 2 while Section 3 is devoted to technical details, among which theoretical controls on Type I error, as well as some useful empirical tools for interpretation. Section 4 provides a non-asymptotic control of the power. Section 5 provides simulated experiments comparing the performances of the suggested procedures with the approach of [43]. In Section 6, we detail the extension of the procedure to handle the comparison of Gaussian graphical models. The method is illustrated on Transcriptomic Breast Cancer Data. Finally, all the proofs are postponed to Section 8.

The R codes of our algorithms are available at <http://www.proba.jussieu.fr/~villers/> as well as on the EJS website <http://projecteuclid.org/euclid.ejs> as supplementary material of this paper [10].

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 $(n_1 + n_2) \times p$ 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. Description of the testing strategy

Likelihood ratio statistics used to test hypotheses like \mathcal{H}_0 in the classical *large n, small p* setting are intractable 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 intractable 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 [49].

For any subset S of $\{1, \dots, p\}$ satisfying $2|S| \leq n_1 \wedge n_2$, denote $X_S^{(1)}$ and $X_S^{(2)}$ the restrictions of random vectors $X^{(1)}$ and $X^{(2)}$ to covariates indexed by S . Their covariance structure is noted $\Sigma_S^{(1)}$ (resp. $\Sigma_S^{(2)}$). Consider the linear regression of $Y^{(1)}$ onto $X_S^{(1)}$ defined by

$$\begin{cases} Y^{(1)} = X_S^{(1)} \beta_S^{(1)} + \epsilon_S^{(1)} \\ Y^{(2)} = X_S^{(2)} \beta_S^{(2)} + \epsilon_S^{(2)}, \end{cases}$$

where the noise variables $\epsilon_S^{(1)}$ and $\epsilon_S^{(2)}$ are independent from $X_S^{(1)}$ and $X_S^{(2)}$ and follow centered Gaussian distributions with new unknown conditional standard deviations $\sigma_S^{(1)}$ and $\sigma_S^{(2)}$. 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 2.1, there is still a strong link between the collection of low dimensional hypotheses $\mathcal{H}_{0,S}$ and the global null hypothesis \mathcal{H}_0 .

Lemma 2.1. *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. As summarized in Algorithm 1, the idea is to build a well-calibrated 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 $\mathcal{H}_{0,S}$ for every $S \subset \{1, \dots, p\}$, 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 $\widehat{\mathcal{S}}$ is judiciously selected, then we can manage not to lose too much power compared to the exhaustive search.

We now turn to the description of the three major elements required by our overall strategy (see Algorithm 1):

1. a well-targeted data-driven collection of models $\widehat{\mathcal{S}}$ as produced by Algorithm 2;
2. a parametric statistic to test the hypotheses $\mathcal{H}_{0,S}$ for $S \in \widehat{\mathcal{S}}$, we resort actually to a combination of three parametric statistics $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$;
3. a calibration procedure guaranteeing the control on type I error as in Algorithms 3 or 4.

2.1. Choices of test collections (step 1)

The first step of our procedure (Algorithm 1) amounts to selecting a collection $\widehat{\mathcal{S}}$ of subsets of $\{1, \dots, p\}$, also called models. A good collection $\widehat{\mathcal{S}}$ of subsets must satisfy a tradeoff between the inclusion of the maximum number of relevant subsets S and a reasonable computing time for the whole testing procedure, which is linear in the size $|\widehat{\mathcal{S}}|$ of the collection. The construction of $\widehat{\mathcal{S}}$ proceeds in two steps: (i) One chooses a *deterministic* collection \mathcal{S} of models. (ii) One defines an algorithm (called *ModelChoice* in Algorithm 1) mapping the raw data (\mathbf{X}, \mathbf{Y}) to some collection $\widehat{\mathcal{S}}$ satisfying $\widehat{\mathcal{S}} \subset \mathcal{S}$. Even though the introduction of \mathcal{S} as an argument of the mapping could appear artificial at this point, this quantity will be used in the calibration step of the procedure. Our methodology can be applied to any fixed or data-driven collection. Still, we focus here on two particular collections. The first one is useful for undertaking the first steps of the mathematical analysis. For practical applications, we advise to use the second collection.

Deterministic collections like $\mathcal{S}_{\leq k}$ By deterministic, we mean that the model choice step is trivial, in the sense that $ModelChoice(\mathbf{X}, \mathbf{Y}, \mathcal{S}) = \mathcal{S}$. Among deterministic collections, the most straightforward collections consist of all size- k

Algorithm 1 Overall Adaptive Testing Strategy

Require: Data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$, collection \mathcal{S} and desired level α

Step 1 – Choose a collection $\widehat{\mathcal{S}}$ of low-dimensional models (as e.g. $\widehat{\mathcal{S}}_{\text{Lasso}}$ in Algorithm 2)

procedure MODELCHOICE($\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathcal{S}$)

 Define the model collection $\widehat{\mathcal{S}} \subset \mathcal{S}$

end procedure

Step 2 – Compute p-values for each test in low dimension

procedure TEST($\mathbf{X}_S^{(1)}, \mathbf{X}_S^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \widehat{\mathcal{S}}$)

for each subset S in $\widehat{\mathcal{S}}$ **do**

 Compute the p-values $\tilde{q}_{V,S}, \tilde{q}_{1,S}, \tilde{q}_{2,S}$ associated to the statistics $F_{S,V}, F_{S,1}, F_{S,2}$

end for

end procedure

Step 3 – Calibrate decision thresholds as in Algorithms 3 (Bonferroni) or 4 (Permutations)

procedure CALIBRATION($\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \widehat{\mathcal{S}}, \alpha$)

for each subset S in $\widehat{\mathcal{S}}$ and each $i = V, 1, 2$ **do**

 Define a threshold $\alpha_{i,S}$.

end for

end procedure

Step 4 – Final Decision

if there is a least one model S in $\widehat{\mathcal{S}}$ such that there is at least one p-value for which $\tilde{q}_{i,S} < \alpha_{i,S}$ **then**

 Reject the global null hypothesis \mathcal{H}_0

end if

subsets of $\{1, \dots, p\}$, which we denote \mathcal{S}_k . This kind of family provides collections which are independent from the data, thereby reducing the risk of overfitting. However, as we allow the model size k or the 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, we introduce the collection of all models of size smaller than k , denoted $\mathcal{S}_{\leq k} = \bigcup_{j=1}^k \mathcal{S}_j$, which will prove useful in theoretical developments.

Lasso-type collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ Among all data-driven collections, we suggest the Lasso-type collection $\widehat{\mathcal{S}}_{\text{Lasso}}$. Before proceeding to its definition, let us informally discuss the subsets that a “good” collection $\widehat{\mathcal{S}}$ should contain. Let $\text{supp}(\beta)$ denote the support of a vector β . Intuitively, under the alternative hypothesis, good candidates for the subsets are either subsets of $S_{\nabla}^* := \text{supp}(\beta^{(1)}) \cup \text{supp}(\beta^{(2)})$ or subsets of $S_{\Delta}^* := \text{Supp}(\beta^{(1)} - \beta^{(2)})$. The first model S_{∇}^* nicely satisfies $\beta_{S_{\nabla}^*}^{(1)} = \beta^{(1)}$ and $\beta_{S_{\nabla}^*}^{(2)} = \beta^{(2)}$. The second subset has a smaller size than S_{∇}^* and focuses on covariates corresponding to different parameters in the full regression. However, the divergence between effects might only appear conditionally on other variables with similar effects, this is why the first subset S_{∇}^* is also of interest. Obviously, both subsets S_{∇}^* and S_{Δ}^* are unknown. This is why we consider a Lasso methodology that amounts to estimating both S_{∇}^* and S_{Δ}^* in the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$. Concrete details on the construction of $\widehat{\mathcal{S}}_{\text{Lasso}}$ are postponed to Section 3.1.

2.2. Parametric test statistic (step 2)

Given a subset S , we consider the three following statistics to test $\mathcal{H}_{0,S}$ against $\mathcal{H}_{1,S}$:

$$F_{S,V} := -2 + \frac{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2} + \frac{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1}, \quad (5)$$

$$F_{S,1} := \frac{\|\mathbf{X}_S^{(2)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_2}{\|\mathbf{Y}^{(1)} - \mathbf{X}^{(1)}\widehat{\beta}_S^{(1)}\|^2/n_1}, \quad F_{S,2} := \frac{\|\mathbf{X}_S^{(1)}(\widehat{\beta}_S^{(1)} - \widehat{\beta}_S^{(2)})\|^2/n_1}{\|\mathbf{Y}^{(2)} - \mathbf{X}^{(2)}\widehat{\beta}_S^{(2)}\|^2/n_2}. \quad (6)$$

As explained in Section 3, these three statistics derive from the Kullback-Leibler divergence between the conditional distributions $Y^{(1)}|X_S^{(1)}$ and $Y^{(2)}|X_S^{(2)}$. While the first term $F_{S,V}$ evaluates the discrepancies in terms of conditional variances, the last two terms $F_{S,1}$ and $F_{S,2}$ address the comparison of $\beta^{(1)}$ to $\beta^{(2)}$.

Because the distributions of statistics $F_{S,i}$, for $i = V, 1, 2$, under the null depend on the size of 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. Under $H_{0,S}$, the conditional distributions of $F_{S,i}$ for $i = V, 1, 2$ to \mathbf{X}_S are parameter-free and explicit (see Proposition 3.1 in the next section). Consequently, one can define the exact p -values associated to $F_{S,i}$, conditional on \mathbf{X}_S . However, the computation of the p -values require a function inversion, which can be computationally prohibitive. This is why we introduce explicit upper bounds $\tilde{q}_{i,S}$ (Equations (14), (18)) of the exact p -values.

2.3. Combining the parametric statistics (step 3)

The objective of this subsection is to calibrate a multiple testing procedure based on the sequence of p -values $\{(\tilde{q}_{V,S}, \tilde{q}_{1,S}, \tilde{q}_{2,S}), S \in \widehat{\mathcal{S}}\}$, so that the type-I error remains smaller than a chosen level α . In particular, when using a data-driven model collection, we must take good care of preventing the risk of overfitting which results from using the same dataset both for model selection and hypothesis testing.

For the sake of simplicity, we assume in the two following paragraphs that $\emptyset \not\subseteq \mathcal{S}$, 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 $\widehat{\mathcal{S}}$ and a sequence $\widehat{\alpha} = (\alpha_{i,S})_{i=V,1,2, S \in \widehat{\mathcal{S}}}$, we define the test function:

$$T_{\widehat{\mathcal{S}}}^{\widehat{\alpha}} = \begin{cases} 1 & \text{if } \exists S \in \widehat{\mathcal{S}}, \exists i \in \{V, 1, 2\} \quad \tilde{q}_{i,S} \leq \alpha_{i,S}. \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

In other words, the test function rejects the global null if there exists at least one model $S \in \widehat{\mathcal{S}}$ such that at least one of the three p -values is below the corresponding threshold $\alpha_{i,S}$. In Section 3.3, we describe two methods to ade-

quately calibrate thresholds $(\alpha_{i,S})_{S \in \widehat{\mathcal{S}}}$. We first define a Bonferroni procedure, whose conceptual simplicity allows us to derive non-asymptotic type II error bounds of the corresponding tests (Section 4). However, this Bonferroni correction reveals itself to be too conservative in practice, in part paying the price for resorting to data-driven collections and upper bounds on the true p -values. This is why we introduce as a second option the permutation calibration procedure. This second procedure controls the type I error at the nominal level and therefore outperforms the Bonferroni calibration in practice. Nevertheless, the mathematical analysis of the corresponding test becomes more intricate and we are not able to provide sharp type II error bounds.

Remark. In practice, we advocate the use of the Lasso Collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ (Algorithm 2) combined with the permutation calibration method (Algorithm 4). Henceforth, the corresponding procedure is denoted $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^P$.

3. Discussion of the procedure and type I error

In this section, we provide remaining details on the three steps of the testing procedure. First, we describe the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ and provide an informal justification of its definition. Second, we explain the ideas underlying the parametric statistics $F_{S,i}$, $i = V, 1, 2$ and we define the corresponding p -values $\tilde{q}_{i,S}$. Finally, the Bonferroni and permutation calibration methods are defined.

3.1. Collection $\widehat{\mathcal{S}}_{\text{Lasso}}$

We start from $\mathcal{S}_{\leq D_{\max}}$, where in practice $D_{\max} = \lfloor (n_1 \wedge n_2)/2 \rfloor$, and we consider the following reparametrized joint regression model.

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \theta_*^{(1)} \\ \theta_*^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix}. \quad (8)$$

In this new 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 $\theta_*^{(2)} = (\beta^{(1)} - \beta^{(2)})/2$. Consequently, $S_{\Delta}^* := \text{Supp}(\beta^{(1)} - \beta^{(2)}) = \text{supp}(\theta_*^{(2)})$ and $S_{\vee}^* := \text{supp}(\beta^{(1)}) \cup \text{supp}(\beta^{(2)}) = \text{supp}(\theta_*^{(1)}) \cup \text{supp}(\theta_*^{(2)})$. To simplify notations, denote by \mathbf{Y} the concatenation of $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$, as well as by \mathbf{W} the reparametrized design matrix of (8). For a given $\lambda > 0$, the Lasso estimator of θ_* is defined by

$$\widehat{\theta}_\lambda := \begin{pmatrix} \widehat{\theta}_\lambda^{(1)} \\ \widehat{\theta}_\lambda^{(2)} \end{pmatrix} := \arg \min_{\theta \in \mathbb{R}^{2p}} \|\mathbf{Y} - \mathbf{W}\theta\| + \lambda \|\theta\|_1, \quad (9)$$

$$\widehat{V}_\lambda := \text{supp}(\widehat{\theta}_\lambda), \quad \widehat{V}_\lambda^{(i)} := \text{supp}(\widehat{\theta}_\lambda^{(i)}), \quad i = 1, 2. \quad (10)$$

For a suitable choice of the tuning parameter λ and under assumptions of the designs, it is proved [6, 36] that $\widehat{\theta}_\lambda$ estimates well θ_* and \widehat{V}_λ is a good estimator

of $\text{supp}(\theta_*)$. The Lasso parameter λ tunes the amount of sparsity of $\hat{\theta}_\lambda$: the larger the parameter λ , the smaller the support \hat{V}_λ . As the optimal choice of λ is unknown, the collection $\hat{\mathcal{S}}_{\text{Lasso}}$ is built using the collection of all estimators $(\hat{V}_\lambda)_{\lambda>0}$, also called the Lasso regularization path of θ_* . Below we provide an algorithm for computing $\hat{\mathcal{S}}_{\text{Lasso}}$ along with some additional justifications.

Algorithm 2 Construction of the Lasso-type Collection $\hat{\mathcal{S}}_{\text{Lasso}}$

Require: Data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$, Collection $\mathcal{S}_{\leq D_{\max}}$

$$\mathbf{Y} \leftarrow \begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix}$$

$$\mathbf{W} \leftarrow \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix}$$

Compute the function $f : \lambda \mapsto \hat{V}_\lambda$ (defined in (9), (10)) using Lars-Lasso Algorithm [17]
 Compute the decreasing sequences $(\lambda_k)_{1 \leq k \leq q}$ of jumps in f

$k \leftarrow 1$, $\hat{\mathcal{S}}_L^{(1)} \leftarrow \emptyset$, $\hat{\mathcal{S}}_L^{(2)} \leftarrow \emptyset$
while $|\hat{V}_{\lambda_k}^{(1)} \cup \hat{V}_{\lambda_k}^{(2)}| < D_{\max}$ **do**
 $\hat{\mathcal{S}}_L^{(1)} \leftarrow \hat{\mathcal{S}}_L^{(1)} \cup \{\hat{V}_{\lambda_k}^{(1)} \cup \hat{V}_{\lambda_k}^{(2)}\}$
 $\hat{\mathcal{S}}_L^{(2)} \leftarrow \hat{\mathcal{S}}_L^{(2)} \cup \{\hat{V}_{\lambda_k}^{(2)}\}$
 $k \leftarrow k + 1$
end while
 $\hat{\mathcal{S}}_{\text{Lasso}} \leftarrow \hat{\mathcal{S}}_L^{(1)} \cup \hat{\mathcal{S}}_L^{(2)} \cup \mathcal{S}_1$

It is known [17] that the function $f : \lambda \mapsto \hat{V}_\lambda$ is piecewise constant. Consequently, there exist thresholds $\lambda_1 > \lambda_2 > \dots$ such that \hat{V}_λ changes on λ_k 's only. The function f and the collection (λ_k) are computed efficiently using the Lars-Lasso Algorithm [17]. We build two collections of models using $(\hat{V}_{\lambda_k}^{(1)})_{k \geq 1}$ and $(\hat{V}_{\lambda_k}^{(2)})_{k \geq 1}$. Following the intuition described above, for a fixed λ_k , $\hat{V}_{\lambda_k}^{(2)}$ is an estimator of $\text{supp}(\beta^{(1)} - \beta^{(2)})$ while $\hat{V}_{\lambda_k}^{(1)} \cup \hat{V}_{\lambda_k}^{(2)}$ is an estimator of $\text{supp}(\beta^{(1)} \cup \text{supp}(\beta^{(2)}))$. This is why we define

$$\hat{\mathcal{S}}_L^{(1)} := \bigcup_{k=1}^{k_{\max}} \{\hat{V}_{\lambda_k}^{(1)} \cup \hat{V}_{\lambda_k}^{(2)}\}, \quad \hat{\mathcal{S}}_L^{(2)} := \bigcup_{k=1}^{k_{\max}} \{\hat{V}_{\lambda_k}^{(2)}\}, \quad (11)$$

where k_{\max} is the smallest integer q such that $|\hat{V}_{\lambda_{q+1}}^{(1)} \cup \hat{V}_{\lambda_{q+1}}^{(2)}| > D_{\max}$. In the end, we consider the following $\hat{\mathcal{S}}_{\text{Lasso}}$ data-driven family,

$$\hat{\mathcal{S}}_{\text{Lasso}} := \hat{\mathcal{S}}_L^{(1)} \cup \hat{\mathcal{S}}_L^{(2)} \cup \mathcal{S}_1. \quad (12)$$

Recall that \mathcal{S}_1 is the collection of the p models of size 1. Recently, data-driven procedures have been proposed to tune the Lasso and find a parameter $\hat{\lambda}$ is such a way that $\hat{\theta}_{\hat{\lambda}}$ is a good estimator of θ_* (see e.g. [4, 45]). Yet we use the whole regularization path instead of the sole estimator $\hat{\theta}_{\hat{\lambda}}$ because our objective is to find subsets S such that the statistics $F_{S,i}$ are powerful. Consider an example where $\beta^{(2)} = 0$ and $\beta^{(1)}$ contains one large coefficient and many small coefficients. If the sample size is large enough, a well-tuned Lasso estimator will

select several variables. In contrast, the best subset S (in terms of power of $F_{S,i}$) contains only one variable. Using the whole regularization path, we hope to find the best trade-off between sparsity (small size of S) and differences between $\beta_S^{(1)}$ and $\beta_S^{(2)}$. This last remark is formalized in Section 4.4. Finally, the size of the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ is generally linear with $(n_1 \wedge n_2) \vee p$, which makes the computation of $(\tilde{q}_{i,S})_{S \in \widehat{\mathcal{S}}_{\text{Lasso}}, i=V,1,2}$ reasonable.

3.2. Parametric statistics and p -values

3.2.1. Symmetric conditional likelihood

In this subsection, we explain the intuition behind the choice of the parametric statistics $(F_{S,V}, F_{S,1}, F_{S,2})$ defined in Equations (5), (6). Let us denote by $\mathcal{L}^{(1)}$ (resp. $\mathcal{L}^{(2)}$) 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.

Statistics $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$ appear as the decomposition of a two-sample likelihood-ratio, measuring the symmetrical 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. The following symmetrized likelihood statistic can be decomposed into the sum of $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$:

$$2 \left[\mathcal{D}_{n_1}^{(1)}(\widehat{\beta}^{(2)}, \widehat{\sigma}^{(2)}) + \mathcal{D}_{n_2}^{(2)}(\widehat{\beta}^{(1)}, \widehat{\sigma}^{(1)}) \right] = F_{S,V} + F_{S,1} + F_{S,2}. \quad (13)$$

Instead of the three statistics $(F_{S,i})_{i=V,1,2}$, one could use the symmetric likelihood (13) to build a testing procedure. However, we do not manage to obtain an explicit and sharp upper bound of the p -values associated to statistic (13), which makes the resulting procedure either computationally intensive if one estimated the p -values by a Monte-Carlo approach or less powerful if one uses a non-sharp upper bound of the p -values. In contrast, we explain below how, by considering separately $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$, one upper bounds sharply the exact p -values.

3.2.2. Definition of the p -values

Denote by $g(x) = -2 + x + 1/x$ the non-negative function defined on \mathbb{R}^+ . Since the restriction of g to $[1; +\infty)$ is a bijection, we note g^{-1} the corresponding reciprocal function.

Proposition 3.1 (Conditional distributions of $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$ 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,

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

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

$$F_{S,1} | \mathbf{X}_S \underset{\mathcal{H}_{0,S}}{\sim} \frac{\|Z_1\|^2/n_2}{\|Z_2\|^2/n_1}.$$

A symmetric result holds for $F_{S,2}$.

Although the distributions identified in Proposition 3.1 are not all 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 $\beta^{(2)}$. For any $i = V, 1, 2$, $\bar{Q}_{i,|S|}(u | \mathbf{X}_S)$ stands for the conditional probability that $F_{S,i}$ is larger than u under $\mathcal{H}_{0,S}$.

By Proposition 3.1, the exact p -value $\tilde{q}_{V,S} = \bar{Q}_{V,|S|}(F_{S,V} | \mathbf{X}_S)$ associated to $F_{S,V}$ is easily computed from the distribution function of a Fisher random variable:

$$\begin{aligned} \tilde{q}_{V,S} &= \mathcal{F}_{n_1 - |S|, n_2 - |S|} \left[g^{-1} (F_{S,V}) \frac{n_1(n_2 - |S|)}{n_2(n_1 - |S|)} \right] \\ &\quad + \mathcal{F}_{n_2 - |S|, n_1 - |S|} \left[g^{-1} (F_{S,V}) \frac{n_2(n_1 - |S|)}{n_1(n_2 - |S|)} \right], \end{aligned} \quad (14)$$

where $\mathcal{F}_{m,n}(u)$ denotes the probability that a Fisher random variable with (m, n) degrees of freedom is larger than u .

Since the conditional distribution of $F_{S,1}$ given \mathbf{X}_S only depends on $|S|$, n_1 , n_2 , and \mathbf{X}_S , one could compute an estimation of the p -value $\bar{Q}_{1,|S|}(u | \mathbf{X}_S)$ 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 $\bar{Q}_{1,|S|}(u | \mathbf{X}_S)$ based on Laplace method, as given in the definition below and justified in the proof of Proposition 3.3.

Definition 3.2 (Definition of $\tilde{Q}_{1,|S|}$ and $\tilde{Q}_{2,|S|}$). Let us note $a = (a_1, \dots, a_{|S|})$ the positive eigenvalues of

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

For any $u \leq |a|_1$, define $\tilde{Q}_{1,|S|}(u | \mathbf{X}_S) := 1$. For any $u > |a|_1$, take

$$\tilde{Q}_{1,|S|}(u | \mathbf{X}_S) := \exp \left[-\frac{1}{2} \sum_{i=1}^{|S|} \log(1 - 2\lambda^* a_i) - \frac{n_1 - |S|}{2} \log \left(1 + \frac{2\lambda^* u}{n_1 - |S|} \right) \right], \quad (15)$$

where λ_* is defined as follows. If all the components of a are equal, then $\lambda^* := \frac{u - |a|_1}{2u(|a|_\infty + \frac{|a|_1}{n_1 - |S|})}$. If a is not a constant vector, then we define

$$\begin{aligned} b &:= \frac{|a|_1 u}{|a|_\infty (n_1 - |S|)} + u + \frac{\|a\|^2}{|a|_\infty} - |a|_1, \\ \Delta &:= b^2 - \frac{4u(u - |a|_1)}{(n_1 - |S|)|a|_\infty} \left(|a|_1 - \frac{\|a\|^2}{|a|_\infty} \right), \end{aligned} \quad (16)$$

$$\lambda^* := \frac{1}{\frac{4u}{n_1 - |S|} \left(|a|_1 - \frac{\|a\|^2}{|a|_\infty} \right)} \left(b - \sqrt{\Delta} \right). \quad (17)$$

$\tilde{Q}_{2,|S|}$ is defined analogously by exchanging the role of $\mathbf{X}_S^{(1)}$ and $\mathbf{X}_S^{(2)}$.

Proposition 3.3. *For any $u \geq 0$, and for $i = 1, 2$, $\bar{Q}_{i,|S|}(u|\mathbf{X}_S) \leq \tilde{Q}_{i,|S|}(u|\mathbf{X}_S)$.*

Finally we define the approximate p -values $\tilde{q}_{1,S}$ and $\tilde{q}_{2,S}$ by

$$\tilde{q}_{1,S} := \tilde{Q}_{1,|S|}(F_{S,1}|\mathbf{X}_S), \quad \tilde{q}_{2,S} := \tilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S). \quad (18)$$

Although we use similar notations for $\tilde{q}_{i,S}$ with $i = V, 1, 2$, this must not mask the essential difference that $\tilde{q}_{V,S}$ is the *exact* p -value of $F_{S,V}$ whereas $\tilde{q}_{1,S}$ and $\tilde{q}_{2,S}$ only are upper-bounds on $F_{S,1}$ and $F_{S,2}$ p -values. The consequences of this asymmetry in terms of calibration of the test is discussed in the next subsection.

3.3. Comparison of the calibration procedures and type I error

3.3.1. Bonferroni calibration (B)

Recall that a data-driven model collection $\hat{\mathcal{S}}$ is defined as the result of a fixed algorithm mapping a deterministic collection \mathcal{S} and (\mathbf{X}, \mathbf{Y}) to a subcollection $\hat{\mathcal{S}}$. The collection of thresholds $\hat{\alpha}^B = \{\alpha_{i,S}, S \in \hat{\mathcal{S}}\}$ is chosen such that

$$\sum_{S \in \mathcal{S}} \sum_{i=V,1,2} \alpha_{i,S} \leq \alpha. \quad (19)$$

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

$$\alpha_{V,S} := \frac{\alpha}{2k} \binom{p}{|S|}^{-1}, \quad \alpha_{1,S} = \alpha_{2,S} := \frac{\alpha}{4k} \binom{p}{|S|}^{-1}, \quad (20)$$

which puts as much weight to the comparison of the conditional variances ($F_{S,V}$) and the comparison of the coefficients ($F_{S,1}, F_{S,2}$). Similarly for the collection $\hat{\mathcal{S}}_{\text{Lasso}}$, a natural choice is (20) with k replaced by D_{\max} (which equals $\lfloor (n_1 \wedge n_2)/2 \rfloor$ in practice).

Given any data-driven collection $\hat{\mathcal{S}}$, denote by $T_{\hat{\mathcal{S}}}^B$ the multiple testing procedure calibrated by Bonferroni thresholds $\hat{\alpha}^B$ (19).

Proposition 3.4 (Size of $T_{\hat{\mathcal{S}}}^B$). *The test function $T_{\hat{\mathcal{S}}}^B$ satisfies $\mathbb{P}_{\mathcal{H}_0}[T_{\hat{\mathcal{S}}}^B = 1] \leq \alpha$.*

Algorithm 3 Bonferroni Calibration for a collection $\widehat{\mathcal{S}} \subset \mathcal{S}_{\leq D_{\max}}$

Require: maximum model dimension D_{\max} , model collection $\widehat{\mathcal{S}}$, desired level α

for each subset S in $\widehat{\mathcal{S}}$ **do**
 $\alpha_{V,S} \leftarrow \alpha(2D_{\max})^{-1} \binom{p}{|S|}^{-1}$
 $\alpha_{1,S} \leftarrow \alpha(4D_{\max})^{-1} \binom{p}{|S|}^{-1}, \quad \alpha_{2,S} \leftarrow \alpha_{1,S}$
end for

Remark 3.1 (Bonferroni correction on \mathcal{S} and not on $\widehat{\mathcal{S}}$). Note that even though we only compute the statistics $F_{S,i}$ for models $S \in \widehat{\mathcal{S}}$, the Bonferroni correction (19) must be applied to the initial deterministic collection \mathcal{S} including $\widehat{\mathcal{S}}$. Indeed, if we replace the condition (19) by the condition $\sum_{S \in \widehat{\mathcal{S}}} \sum_{i=1}^3 \alpha_{i,S} \leq \alpha$, 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 $\widehat{\mathcal{S}} \subset \mathcal{S}$ and to perform the multiple testing procedure. As a simple example, consider any deterministic collection \mathcal{S} and the data-driven collection

$$\widehat{\mathcal{S}} = \left\{ \arg \min_{S \in \mathcal{S}} \min_{i=V,1,2} \tilde{q}_{i,S} \right\},$$

meaning that $\widehat{\mathcal{S}}$ only contains 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} .

Although 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 3.1, whereas 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 the above subsection, for computational reasons we do not consider the exact p -values of $F_{S,1}$ and $F_{S,2}$ but only upper bounds $\tilde{q}_{1,S}$ and $\tilde{q}_{2,S}$ of them.

In fact, the three aforementioned issues are addressed by the permutation approach.

3.3.2. Calibration by permutation (P)

The collection of thresholds $\widehat{\alpha}^P = \{\alpha_{i,S}, S \in \widehat{\mathcal{S}}\}$ is chosen such that each $\alpha_{i,S}$ remains inversely proportional to $\binom{p}{|S|}$ in order to put all subset sizes at equal footage. In other words, we choose a collection of thresholds of the form

$$\alpha_{i,S} = \widehat{C}_i \binom{p}{|S|}^{-1}, \quad (21)$$

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 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 $\widehat{\mathcal{S}}^\pi$, parametric statistics $(F_{S,i}^\pi)_{i=V,1,2}$ and p -values $(\tilde{q}_{i,S}^\pi)_{i=V,1,2}$. Denote \mathcal{P} the uniform distribution over the permutations of size $n_1 + n_2$.

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

$$\min_{S \in \widehat{\mathcal{S}}^\pi} \left[\tilde{q}_{V,S}^\pi \binom{p}{|S|} \right]. \quad (22)$$

Similarly, $\widehat{C}_1 = \widehat{C}_2$ are the $\alpha/2$ -quantiles with respect to \mathcal{P} of

$$\min_{S \in \widehat{\mathcal{S}}^\pi} \left[(\tilde{q}_{1,S}^\pi \wedge \tilde{q}_{2,S}^\pi) \binom{p}{|S|} \right]. \quad (23)$$

In practice, the quantiles \widehat{C}_i are estimated by sampling a large number B of permutations. The permutation calibration procedure for the Lasso collection $\widehat{\mathcal{S}}_{\text{Lasso}}$ is summarized in Algorithm 4.

Algorithm 4 Calibration by Permutation for $\widehat{\mathcal{S}}_{\text{Lasso}}$

Require: Data $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$, maximum model dimension D_{\max} , number B of permutations, desired level α

for $b = 1, \dots, B$ **do**

 Draw π a random permutation of $\{1, \dots, n_1 + n_2\}$

$\mathbf{X}^{(b)}, \mathbf{Y}^{(b)} \leftarrow \pi$ -permutation of (\mathbf{X}, \mathbf{Y})

procedure LASSOMODELCHOICE($\mathbf{X}^{(1,b)}, \mathbf{X}^{(2,b)}, \mathbf{Y}^{(1,b)}, \mathbf{Y}^{(2,b)}, S_{\leq D_{\max}}$)

 Define $\widehat{\mathcal{S}}_{\text{Lasso}}^{(b)}$ (as in Algorithm 2)

end procedure

procedure TEST($\mathbf{X}^{(1,b)}, \mathbf{X}^{(2,b)}, \mathbf{Y}^{(1,b)}, \mathbf{Y}^{(2,b)}, \widehat{\mathcal{S}}_{\text{Lasso}}^{(b)}$)

for each subset S in $\widehat{\mathcal{S}}_{\text{Lasso}}^{(b)}$ **do**

 Compute the p -values $\tilde{q}_{i,S}^{(b)}$ for $i = V, 1, 2$.

end for

$M_V^{(b)} \leftarrow \min_{S \in \widehat{\mathcal{S}}_{\text{Lasso}}^{(b)}} \tilde{q}_{V,S}^{(b)} \binom{p}{|S|}$

$M_1^{(b)} \leftarrow \min_{S \in \widehat{\mathcal{S}}_{\text{Lasso}}^{(b)}} (\tilde{q}_{1,S}^{(b)} \wedge \tilde{q}_{2,S}^{(b)}) \binom{p}{|S|}$

end procedure

end for

Define \widehat{C}_V as the $\alpha/2$ -quantile of the $(M_V^{(1)}, \dots, M_V^{(B)})$ distribution

Define $\widehat{C}_1 = \widehat{C}_2$ as the $\alpha/2$ -quantile of the $(M_1^{(1)}, \dots, M_1^{(B)})$ distribution

for each subset S in $\widehat{\mathcal{S}}_{\text{Lasso}}$, each $i = V, 1, 2$, **do**

$\alpha_{i,S} \leftarrow \widehat{C}_i \binom{p}{|S|}^{-1}$

end for

Given any data-driven collection $\widehat{\mathcal{S}}$, denote by $T_{\widehat{\mathcal{S}}}^P$ the multiple testing procedure calibrated by the permutation method (21).

Proposition 3.5 (Size of $T_{\hat{S}}^P$). *The test function $T_{\hat{S}}^P$ satisfies*

$$\alpha/2 \leq \mathbb{P}_{\mathcal{H}_0} \left[T_{\hat{S}}^P = 1 \right] \leq \alpha.$$

Remark 3.2. Through the three constants \hat{C}_V , \hat{C}_1 and \hat{C}_2 (Eqs. (22), (23)), the permutation approach corrects simultaneously for the losses mentioned earlier due to the Bonferroni correction, in particular the restriction to a data-driven class \hat{S} and the approximate p -values $\tilde{q}_{1,S}$ and $\tilde{q}_{2,S}$. Yet, the level of $T_{\hat{S}}^P$ is not exactly α because we treat separately the statistics $F_{S,V}$ and $(F_{S,1}, F_{S,2})$ 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 $F_{S,V}$ too much, because we would put on the same level the exact p -value $\tilde{q}_{V,S}$ and the upper bounds $\tilde{q}_{1,S}$ and $\tilde{q}_{2,S}$.

3.4. Interpretation tools

Empirical p -value When using a calibration by permutations, one can derive an empirical p -value $p^{empirical}$ to assess the global significance of the test. In contrast with model and statistic specific p -values $\tilde{q}_{i,S}$, this p -value provides a nominally accurate estimation of the type-I error rate associated with the global multiple testing procedure, every model in the collection and test statistic being considered. It can be directly compared to the desired level α to decide about the rejection or not of the global null hypothesis.

This empirical p -value is obtained as the fraction of the permuted values of the statistic that are less than the observed test statistic. Keeping the notation of Algorithm 4, the empirical p -value for the variance and coefficient parts are given respectively by:

$$\begin{aligned} p_V^{empirical} &= \frac{1}{B} \sum_{b=1}^B \mathbb{1} \left[M_V^{(b)} < \min_{S \in \hat{S}_{Lasso}} \tilde{q}_{V,S} \binom{p}{|S|} \right], \\ p_{1-2}^{empirical} &= \frac{1}{B} \sum_{b=1}^B \mathbb{1} \left[M_1^{(b)} < \min_{S \in \hat{S}_{Lasso}} (\tilde{q}_{1,S} \wedge \tilde{q}_{2,S}) \binom{p}{|S|} \right]. \end{aligned}$$

The empirical p -value for the global test is then given by the following equation.

$$p^{empirical} = 2 \min(p_V^{empirical}, p_{1-2}^{empirical}). \quad (24)$$

Rejected model Moreover, one can keep track of the model responsible for the rejection, unveiling sensible information on which particular coefficients most likely differ between samples. The rejected models for the variance and coefficient parts are given respectively by:

$$S_V^R = \arg \min_{S \in \hat{S}_{Lasso}} \tilde{q}_{V,S} \binom{p}{|S|}$$

$$S_{1-2}^R = \arg \min_{S \in \widehat{S}_{\text{Lasso}}} (\tilde{q}_{1,S} \wedge \tilde{q}_{2,S}) \binom{p}{|S|}$$

We define the rejected model S^R as model S_V^R or S_{1-2}^R according to the smallest empirical p-value $p_V^{\text{empirical}}$ or $p_{1-2}^{\text{empirical}}$.

4. 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. Numerical experiments in Section 5 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 positive constant $L(\alpha, \delta)$ that only depends on α and δ , $A \leq L(\alpha, \delta)B$ (resp. $A \geq L(\alpha, \delta)B$).

We first define the symmetrized Kullback-Leibler divergence as a way to measure the discrepancies between $(\beta^{(1)}, \sigma^{(1)})$ and $(\beta^{(2)}, \sigma^{(2)})$. Then, we consider tests with deterministic collections in Sections 4.2–4.3. We prove that the corresponding tests are minimax adaptive to the sparsity of the parameters or to the sparsity of the difference $\beta^{(1)} - \beta^{(2)}$. Sections 4.4–4.5 are devoted to the analysis of $T_{\widehat{S}_{\text{Lasso}}}^B$. Under stronger assumptions on the population covariances than for deterministic collections, we prove that the performances of $T_{\widehat{S}_{\text{Lasso}}}^B$ are nearly optimal.

4.1. 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

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

Observe that the quantity $\varphi_{\Sigma^{(1)}, \Sigma^{(2)}}$ can be considered as a constant if we assume that the smallest and largest eigenvalues of $\Sigma^{(i)}$ are bounded away from zero and infinity.

4.2. Power of $T_{\mathcal{S}_{\leq k}}^B$

First, we control the power of $T_{\mathcal{S}}^B$ for a deterministic collection $\mathcal{S} = \mathcal{S}_{\leq k}$ (with some $k \leq (n_1 \wedge n_2)/2$) and the Bonferroni calibration thresholds $\hat{\alpha}_{i, \mathcal{S}}$ as in (20). For any $\beta \in \mathbb{R}^p$, $|\beta|_0$ refers to the size of its support. We consider the two following assumptions

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

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

Remark 4.1. 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 [47] to be minimal to obtain fast rates of testing of the form (25) in the specific case $\beta^{(2)} = 0$, $\sigma^{(1)} = \sigma^{(2)}$ and $n_2 = \infty$.

Theorem 4.1 (Power of $T_{\mathcal{S}_{\leq k}}^B$). *Assuming that **A.1** and **A.2** hold, $\mathbb{P}[T_{\mathcal{S}_{\leq k}}^B = 1] \geq 1 - \delta$ as long as*

$$\mathcal{K}_1 + \mathcal{K}_2 \gtrsim \varphi_{\Sigma^{(1)}, \Sigma^{(2)}} \frac{\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \vee 1\} \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n_1 \wedge n_2}. \quad (25)$$

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

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \gtrsim \frac{|\beta^{(1)} - \beta^{(2)}|_0 \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n_1 \wedge n_2}. \quad (26)$$

Remark 4.2. The condition $\Sigma^{(1)} = \Sigma^{(2)}$ is not necessary to control the power of $T_{\mathcal{S}_{\leq k}}^B$ in terms of $|\beta^{(1)} - \beta^{(2)}|_0$ as in (26). However, the expression (26) would become far more involved.

Remark 4.3. Before assessing the optimality of Theorem 4.1, let us briefly compare the two rates of detection (25) and (26). According to (25), $T_{\mathcal{S}_{\leq k}}^B$ is powerful as soon as the symmetrized Kullback distance is large compared to $\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0\} \log(p)/(n_1 \wedge n_2)$. In contrast, (26) tells us that $T_{\mathcal{S}_{\leq k}}^B$ 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 (26) 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 of the same order as the noise levels $\sigma^{(i)}$. In such a case, (26) outperforms (25) and only the sparsity of the difference $\beta^{(1)} - \beta^{(2)}$ plays a role in the detection rates. Below, we prove that (25) and (26) are both optimal from a minimax point of view but on different sets.

Proposition 4.2 (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 $L(\alpha, \delta, \gamma)$ and $L'(\alpha, \delta, \gamma)$ 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

$$\mathcal{K}_1 + \mathcal{K}_2 \geq L(\alpha, \delta, \gamma) \frac{s}{n_1 \wedge n_2} \log(p). \quad (27)$$

- 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**, $|\beta^{(1)} - \beta^{(2)}|_0 \leq s$ and

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{I_p}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \geq L'(\alpha, \delta, \gamma) \frac{s}{n_1 \wedge n_2} \log(p). \quad (28)$$

The proof (in Section 8) is a straightforward application of minimax lower bounds obtained for the one-sample testing problem [2, 49].

Remark 4.4. Equation (25) together with (27) tell 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 $\gamma > 0$. Nevertheless, we only managed to prove the minimax lower bound for $\Sigma^{(1)} = \Sigma^{(2)} = I_p$, implying that, even though the detection rate (25) is not improvable uniformly over all $(\Sigma^{(1)}, \Sigma^{(2)})$, some improvement is perhaps possible for specific covariance matrices. Up to our knowledge, there exist no such results of adaptation to the population covariance of the design even in the one sample problem.

Remark 4.5. Equation (26) together with (28) 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 4.6 (Informal justification of the introduction of the collection $\widehat{\mathcal{S}}_{\text{Lasso}}$). If we look at the proof of Theorem 4.1, we observe that the power (25) is achieved by the statistics $(F_{S_V, V}, F_{S_V, 1}, F_{S_V, 2})$ where S_V is the union of the support of $\beta^{(1)}$ and $\beta^{(2)}$. In contrast, (26) is achieved by the statistics $(F_{S_\Delta, V}, F_{S_\Delta, 1}, F_{S_\Delta, 2})$ 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 (12) of $\widehat{\mathcal{S}}_{\text{Lasso}}$ is to estimate S_V , while the idea underlying the collection $\widehat{\mathcal{S}}_L^{(2)}$ is to estimate S_Δ .

4.3. Power of T_S^B for any deterministic \mathcal{S}

Theorem 4.3 below extends Theorem 4.1 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 of subsets S of collection \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 $\mathcal{K}_2(S)$

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

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

$$\varphi_S := \varphi_{\max} \left\{ \sqrt{\Sigma_S^{(2)} (\Sigma_S^{(1)})^{-1}} \sqrt{\Sigma_S^{(2)}} + \sqrt{\Sigma_S^{(1)} (\Sigma_S^{(2)})^{-1}} \sqrt{\Sigma_S^{(1)}} \right\}. \quad (29)$$

Theorem 4.3 (Power of T_S^B for any deterministic \mathcal{S}). *For any $S \in \mathcal{S}$, we note $\alpha_S = \min_{i=V,1,2} \alpha_{i,S}$. The power of T_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*

$$1 + \log[1/(\delta\alpha_S)] \lesssim n_1 \wedge n_2, \quad (30)$$

and

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \gtrsim \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{1}{\alpha_S \delta} \right) \right]. \quad (31)$$

Remark 4.7. Let us note $\Delta(S)$ the right hand side of (31). According to Theorem 4.3, The $\Delta(S)$ term plays the role of a variance term and therefore increases with the cardinality of S . Furthermore, the $\mathcal{K}_1 - \mathcal{K}_1(S) + \mathcal{K}_2 - \mathcal{K}_2(S)$ term plays the role of a bias. Let us note \mathcal{S}^* the subcollection of \mathcal{S} made of sets S satisfying (30). According to Theorem 4.3, T_S^B is powerful as long as $\mathcal{K}_1 + \mathcal{K}_2$ is larger (up to constants) to

$$\inf_{S \in \mathcal{S}^*} \{ \mathcal{K}_1 - \mathcal{K}_1(S) + \mathcal{K}_2 - \mathcal{K}_2(S) \} + \Delta(S) \quad (32)$$

Such a result is comparable to oracle inequalities obtained in estimation since T_S^B is powerful when the $\mathcal{K}_1 + \mathcal{K}_2$ Kullback loss is larger than the trade-off (32) 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.

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

For the sake of simplicity, we restrict our developments to the case $n_1 = n_2 := n$ in this subsection, more general results being postponed to the next subsection. The $T_{\widehat{\mathcal{S}}_{\leq n/2}}^B$ test is computationally expensive (non polynomial with respect to p). The $\widehat{\mathcal{S}}_{Lasso}$ collection has been introduced to fix this burden. We consider $T_{\widehat{\mathcal{S}}_{Lasso}}^B$ with the prescribed Bonferroni calibration thresholds $\widehat{\alpha}_{i,S}$, as in (20) with k replaced by $\lfloor (n_1 \wedge n_2)/2 \rfloor$. In the following statements, $\psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(1)}$, $\psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(2)}$, \dots refer to positive quantities that only depend on the largest and 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 4.4. *Assuming that **A.1** and **A.3** hold, we have $\mathbb{P}[T_{\widehat{\mathcal{S}}_{Lasso}}^B = 1] \geq 1 - (\delta \vee 4/p)$ as long as*

$$\mathcal{K}_1 + \mathcal{K}_2 \gtrsim \psi_{\Sigma^{(1)}, \Sigma^{(2)}}^{(3)} \frac{\{|\beta^{(1)}|_0 \vee |\beta^{(2)}|_0 \vee 1\} \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n}. \quad (33)$$

*If $\Sigma^{(1)} = \Sigma^{(2)} = \Sigma$ and if **A.1** and **A.4** hold, then $\mathbb{P}[T_{\widehat{\mathcal{S}}_{Lasso}}^B = 1] \geq 1 - (\delta \vee 6/p)$ as long as*

$$\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma}^2}{\text{Var}[Y^{(1)}] \wedge \text{Var}[Y^{(2)}]} \gtrsim \psi_{\Sigma, \Sigma}^{(4)} \frac{|\beta^{(1)} - \beta^{(2)}|_0 \log(p) + \log\left(\frac{1}{\alpha\delta}\right)}{n}. \quad (34)$$

Remark 4.8. The rates of detection (33) and the sparsity condition **A.3** are analogous to (25) and Condition **A.2** in Theorem 4.1 for $T_{\widehat{\mathcal{S}}_{\leq (n_1 \wedge n_2)/2}}^B$. The second result (34) is also similar to (26). As a consequence, $T_{\widehat{\mathcal{S}}_{Lasso}}^B$ is minimax adaptive to the sparsity of $(\beta^{(1)}, \beta^{(2)})$ and of $\beta^{(1)} - \beta^{(2)}$.

Remark 4.9. Dependencies of **A.3**, **A.4**, (33) and (34) on $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are unavoidable because the $\widehat{\mathcal{S}}_{Lasso}$ collection 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 the next subsection.

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

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 [46]) by

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, 1 \leq |\theta|_0 \leq k} \frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2}, \quad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, 1 \leq |\theta|_0 \leq k} \frac{\|\mathbf{X}\theta\|^2}{\|\theta\|^2},$$

$$\begin{aligned}\kappa[M, k, \mathbf{X}] &= \min_{T, \theta: |T| \leq k, \theta \in \mathcal{C}(M, T)} \left\{ \frac{\|\mathbf{X}\theta\|}{\|\theta\|} \right\}, \\ \eta[M, k, \mathbf{X}] &= \min_{T, \theta: |T| \leq k, \theta \in \mathcal{C}(M, T)} \left\{ \sqrt{k} \frac{\|\mathbf{X}\theta\|}{|\theta|_1} \right\},\end{aligned}\quad (35)$$

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

$$\begin{aligned}\gamma_{\Sigma^{(1)}, \Sigma^{(2)}, k} &:= \frac{\bigwedge_{i=1,2} \kappa^2 [6, k_*, \sqrt{\Sigma^{(i)}}]}{\bigvee_{i=1,2} \Phi_{k_*, +}(\sqrt{\Sigma^{(i)}})}, \\ \gamma'_{\Sigma^{(1)}, \Sigma^{(2)}, k} &:= \frac{\bigvee_{i=1,2} \Phi_{k_*, +}^2(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}}) \bigwedge_{i=1,2} \kappa^2 [6, k, \sqrt{\Sigma^{(i)}}]},\end{aligned}$$

which measure the closeness to orthogonality of $\Sigma^{(1)}$ and $\Sigma^{(2)}$. Theorem 4.4 is a straightforward consequence of the following two results.

Proposition 4.5. *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*

$$(k_* + 1) \log(p) \leq L^*(n_1 \wedge n_2), \quad (36)$$

and assume that

$$1 + \log[1/(\alpha\delta)] < L_1^*(n_1 \wedge n_2) \quad \text{and} \quad \delta \geq 4/p. \quad (37)$$

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

$$|\beta^{(1)}|_0 + |\beta^{(2)}|_0 \leq L_2^* \gamma_{\Sigma^{(1)}, \Sigma^{(2)}, k_*} k_* \left(\frac{n_1}{n_2} \wedge \frac{n_2}{n_1} \right). \quad (38)$$

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 $T_{\hat{\mathcal{S}}_{\text{Lasso}}}^B$ behaves nearly as well as what has been obtained in (25) for $T_{\hat{\mathcal{S}}_{\leq (n_1 \wedge n_2)/2}}^B$, 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 4.6. *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*

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

$$\tilde{k}_* \leq \tilde{L}^* \tilde{\gamma}_{\Sigma, k_*} \left[\frac{n_1 \wedge n_2}{|n_1 - n_2|} \wedge \sqrt{\frac{n_1 \wedge n_2}{\log(p)}} \right], \quad (39)$$

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

$$1 + \log [1/(\alpha\delta)] < L_1^*(n_1 \wedge n_2) \quad \text{and} \quad \delta \geq 6/p.$$

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

$$|\beta^{(1)}|_0 + |\beta^{(2)}|_0 \leq L_2^* \tilde{\gamma}_{\Sigma, \tilde{k}_*}^{(2)} \tilde{k}_*. \quad (40)$$

and

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

Remark 4.10. Definition (39) of \tilde{k}_* together with Condition (40) 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 (8) 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.

5. Numerical experiments

This section evaluates the performances of the suggested test statistics along with aforementioned test collections and calibrations on simulated linear regression datasets.

5.1. 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 [2]. 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 $r = 0.5$:

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

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 independently 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 that the correlation structure is independent from the sparsity pattern.

Gaussian graphical model structure Second, we simulate correlation structures with the R package `GGMselect`. The function `simulateGraph` generates covariance matrices corresponding to Gaussian graphical model structure made of

TABLE 1
Summary of the six different sparsity patterns under study

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	$\beta^{(1)}$ _____
					$\beta^{(2)}$ _____

clusters with some intra-cluster and extra-cluster connectivity coefficients. See Section 4 of [20] for more details. 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. With the default option of the function `simulateGraph` the extra-cluster connectivity coefficient is taken five times smaller.

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 $\hat{\mathcal{S}}_{\text{Lasso}}$, 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 our suggested test statistic to the usual Fisher statistic and we compare our approach with the parallel work of [43].

Fisher statistic For a given support $|S|$ of reduced dimension the usual likelihood ratio statistic for the equality of $\beta_S^{(1)}$ and $\beta_S^{(2)}$ follows a Fisher distribution

with $|S|$ and $n_1 + n_2 - 2|S|$ degrees of freedom:

$$\text{Fi}_S = \frac{\|\mathbf{Y} - \mathbf{X}_S \widehat{\beta}_S\|^2 - \|\mathbf{Y}^{(1)} - \mathbf{X}_S^{(1)} \widehat{\beta}_S^{(1)}\|^2 - \|\mathbf{Y}^{(2)} - \mathbf{X}_S^{(2)} \widehat{\beta}_S^{(2)}\|^2}{\|\mathbf{Y}^{(1)} - \mathbf{X}_S^{(1)} \widehat{\beta}_S^{(1)}\|^2 + \|\mathbf{Y}^{(2)} - \mathbf{X}_S^{(2)} \widehat{\beta}_S^{(2)}\|^2} \frac{n_1 + n_2 - 2|S|}{|S|}, \quad (41)$$

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

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

Procedure of Städler and Mukherjee [43] The DiffRegr procedure of Städler and Mukherjee performs two-sample testing between high-dimensional regression models. The procedure is based on sample-splitting: the data is split in two halves, the first one allowing to reduce dimensionality (*screening* step) and the second being used to compute p -values based on a restricted log-likelihood-ratio statistic (*cleaning* step). Contrary to our procedure, which compares the predictive ability of the same model S in both conditions, the DiffReg procedure is based on the selection of three distinct models at the *screening* step, one for each condition separately and one on the joint sample. This approach results in an intricate non-nested model comparison at the *cleaning* step. To increase the stability of the results the splitting step is repeated multiple times and the resulting p -values must be aggregated. We choose to use the p -value calculations based on permutations as it remains computationally reasonable for the regression case, see [43]. The single-splitting and multi-splitting procedures are denoted respectively by SS(perm) and MS(perm).

Validation of type I error control

Control under the global null hypothesis \mathcal{H}_{00} Table 2 presents estimated type I error rates, that is the percentage of simulations for which the null hypothesis is rejected, based upon 1000 simulations under the restricted null hypothesis \mathcal{H}_{00} , where $\beta^{(1)} = \beta^{(2)} = 0$ and under orthogonal correlation structure. The desired level is $\alpha = 5\%$, and the estimated levels are given with a 95% Gaussian confidence interval.

As expected under independence, the combination of the \mathcal{S}_1 collection with Bonferroni correction gives accurate alpha-level when applied to the usual Fisher statistic. On the contrary when applied to the suggested statistics, the use of upper bounds on p -values leads to a strong decrease in observed type-I error. This decrease is exacerbated when using the $\widehat{\mathcal{S}}_{\text{Lasso}}$ collection, since we are accounting for many more models than the number actually tested in order to prevent overfitting. This effect can be seen both on the Fisher statistic and our suggested statistic. Even with the usual Fisher statistic, for which we know the

TABLE 2
 Estimated test levels in percentage along with 95% Gaussian confidence interval
 (in percentage) under \mathcal{H}_{00} based upon 1000 simulations

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

Model collection	\mathcal{S}_1		$\hat{\mathcal{S}}_{Lasso}$	
Calibration	(B)	(P)	(B)	(P)
$n = 25$	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 ± 0.6	7.4 ± 1.6	0.1 ± 0.2	7.3 ± 1.6

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

Model collection	\mathcal{S}_1		$\hat{\mathcal{S}}_{Lasso}$	
Calibration	(B)	(P)	(B)	(P)
$n = 25$	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	6.6 ± 1.5	0.1 ± 0.2	6.5 ± 1.5

(c) DiffRegr procedure

Model collection	SS (perm)	MS (perm)
$n = 25$	4.3 ± 1.3	0.1 ± 0.2
$n = 50$	4.1 ± 1.2	0.2 ± 0.3
$n = 100$	3.5 ± 1.1	0.1 ± 0.2

exact p -value, it is unthinkable to use Bonferroni calibration as soon as we adopt data-driven collections instead of deterministic ones.

On the contrary, a calibration by permutations restores a control of type-I error at the desired nominal level, whatever the test statistic or model collection.

As noted by [43], the multi-splitting procedure yields conservative results in terms of type I error control at level 5%.

Control under the global equality of non null coefficients \mathcal{H}_0 Figures 1 and 3 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 multi-split DiffRegr procedure and the Bonferroni calibration-based procedures $T_{\hat{\mathcal{S}}}^B$ (for any collection $\hat{\mathcal{S}}$) and $T_{\hat{\mathcal{S}}_{Lasso}}^{B,Fisher}$. In particular, the Fisher statistic combined with \mathcal{S}_1 and Bonferroni calibration is more conservative than the desired nominal level under correlated designs.

Power analysis We do not investigate the power of the Bonferroni-based procedures $T_{\hat{\mathcal{S}}}^B$ and $T_{\hat{\mathcal{S}}}^{B,Fisher}$ as they have been shown to be too conservative in the above Type I error analysis. Figure 4 represents power performances for the test $T_{\hat{\mathcal{S}}}^P$ and the usual likelihood ratio test $T_{\hat{\mathcal{S}}}^{P,Fisher}$ combined with either \mathcal{S}_1 or $\hat{\mathcal{S}}_{Lasso}$ test collections using a calibration by permutation under an orthogonal covariance matrix Σ , as well as power performance for the DiffRegr procedure. Figure 5 represents equivalent results for power decay and GGM covariance structures when $n = 50$.

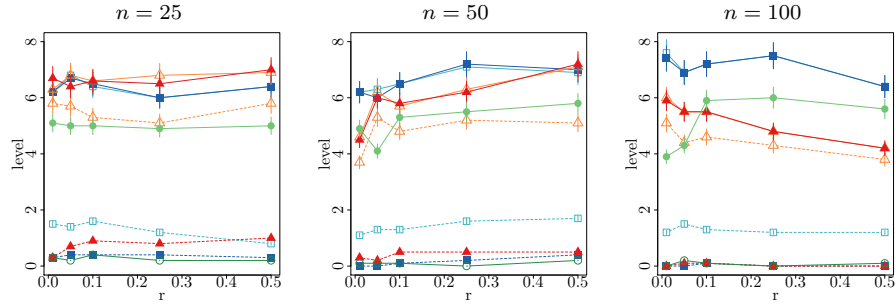


FIG 1. Estimated test levels in percentage under \mathcal{H}_0 for varying magnitudes of common non null coefficients, based upon 1000 simulations. Bonferroni calibration in dotted lines, calibration by permutation in plain lines. Blue squares represent the suggested test $T_{\hat{S}}^*$, red triangles stand for the Fisher test $T_{\hat{S}}^{*,Fisher}$. The deterministic collection \mathcal{S}_1 is drawn in empty points, while the data-driven collection $\hat{\mathcal{S}}_{Lasso}$ is in plain points. Green circles represent the DiffRegr procedure, respectively plain and empty for single-splitting and multi-splitting.

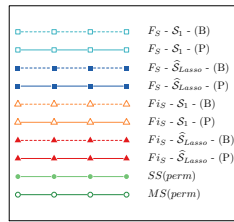


FIG 2. Legend of the procedures under study

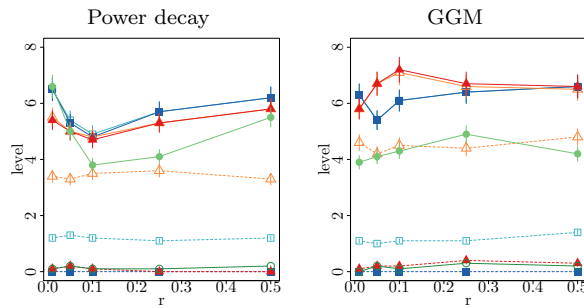


FIG 3. Estimated test levels in percentage under \mathcal{H}_0 for varying magnitudes of common non null coefficients, based upon 1000 simulations, under power decay and GGM correlation structures when $n = 50$. Bonferroni calibration in dotted lines, calibration by permutation in plain lines. Blue squares represent the suggested test $T_{\hat{S}}^*$, red triangles stand for the Fisher test $T_{\hat{S}}^{*,Fisher}$. The deterministic collection \mathcal{S}_1 is drawn in empty points, while the data-driven collection $\hat{\mathcal{S}}_{Lasso}$ is in plain points. Green circles represent the DiffRegr procedure, respectively plain and empty for single-splitting and multi-splitting.

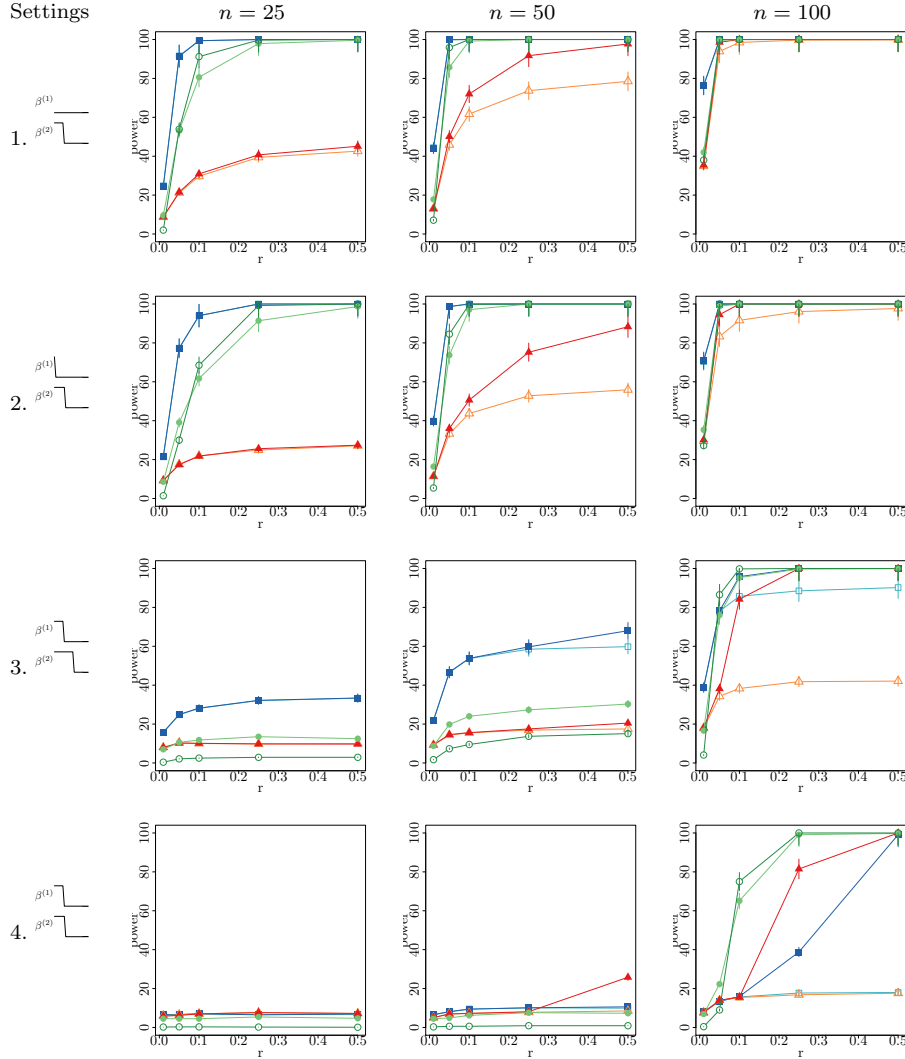


FIG 4. Power (in percentage) as a function of signal magnitude parameter r for various sparsity pattern under the assumption of uncorrelated designs $\Sigma^{(1)} = \Sigma^{(2)} = I_p$. Results for the suggested test $T_{\hat{S}}^P$ and the test $T_{\hat{S}}^{P,Fisher}$, combined with \mathcal{S}_1 or $\hat{\mathcal{S}}_{Lasso}$ test collections. Blue squares represent the suggested test $T_{\hat{S}}^P$, red triangles stand for the Fisher test $T_{\hat{S}}^{P,Fisher}$. The deterministic collection \mathcal{S}_1 is drawn in empty points, while the data-driven collection $\hat{\mathcal{S}}_{Lasso}$ is in plain points. Results for the DiffRegr procedure are represented by green circles, respectively plain and empty for single-splitting and multi-splitting approaches.

In the absence of common coefficients (scenarios 1 and 2), the test $T_{\hat{S}}^P$ 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

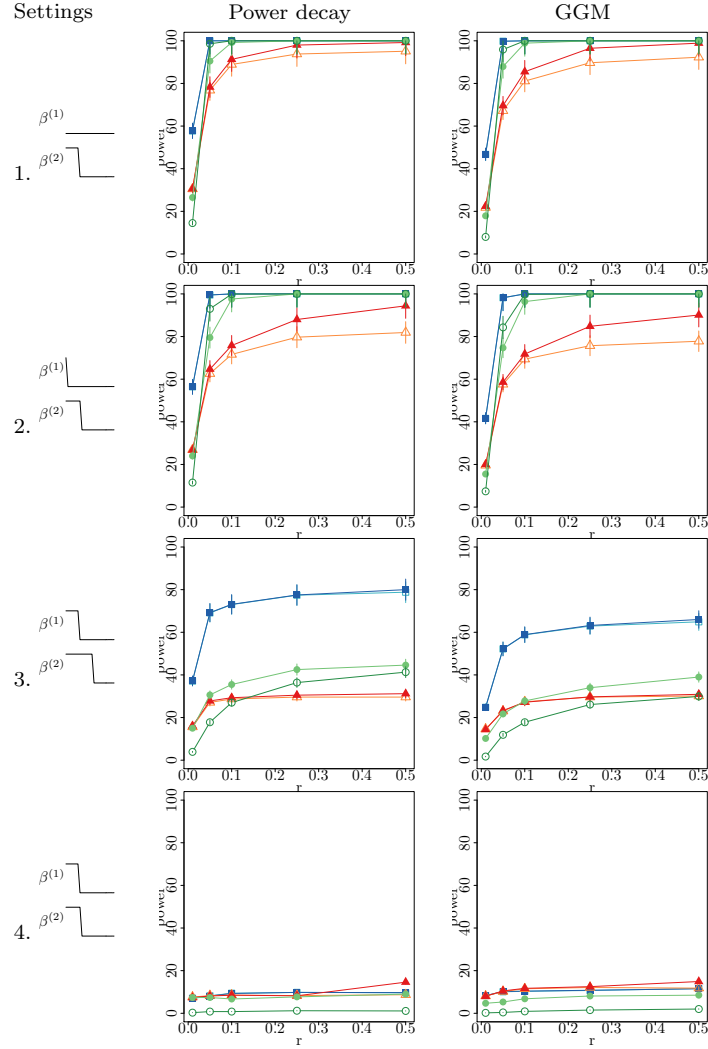


FIG 5. Power (in percentage) as a function of signal magnitude parameter r for various sparsity patterns under power decay and GGM correlated designs, at $n = 50$ observations. Results for the suggested test $T_{\hat{S}}^P$ and the test $T_{\hat{S}}^{P, \text{Fisher}}$, combined with S_1 or \hat{S}_{Lasso} test collections and a calibration by permutation. Blue squares represent the suggested test $T_{\hat{S}}^P$, red triangles stand for the Fisher test $T_{\hat{S}}^{P, \text{Fisher}}$. The deterministic collection S_1 is drawn in empty points, while the data-driven collection \hat{S}_{Lasso} is in plain points. Results for the DiffRegr procedure are represented by green circles, respectively plain and empty for single-splitting and multi-splitting approaches.

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 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 $(F_{S,V}, F_{S,1}, F_{S,2})$ and not so badly when associated with Fi_S .

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

For small n , the test $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^P$ outperforms the DiffRegr procedure. However, when $n = 100$ the procedure DiffRegr performs better than our procedure in the highly challenging setting 4.

Figure 5 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.

6. Application to GGM

The following section explicits the extension of the two-sample linear regression testing framework to the two-sample Gaussian graphical model testing framework. We describe the tools and guidelines for a correct interpretation of the results and illustrate the approach on a typical two-sample transcriptomic dataset.

6.1. How to apply this strategy to GGM testing

Neighborhood selection approach The procedure developed in Section 2 can be adapted to the case of Gaussian graphical models as in [48]. We quickly recall why estimation of the Gaussian graphical model amounts to the estimation of p independent linear regressions when adopting a neighborhood selection approach [34].

Consider two Gaussian random vectors $Z^{(1)} \sim \mathcal{N}(0, [\Omega^{(1)}]^{-1})$ and $Z^{(2)} \sim \mathcal{N}(0, [\Omega^{(2)}]^{-1})$. Their respective conditional independence structures are represented by the graphs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, which consist of a common set of nodes $\Gamma = \{1, \dots, p\}$ and their respective sets of edges $\mathcal{E}^{(1)}$ and $\mathcal{E}^{(2)}$. When speaking of gene regulation networks, each node represents a gene, and edges between genes are indicative of potential regulations. In contrast with gene co-expression networks, edges in Gaussian graphical models do not reflect correlations but partial correlations between gene expression profiles.

Formally, an edge (i, j) belongs to the edge set $\mathcal{E}^{(1)}$ (resp. $\mathcal{E}^{(2)}$) if $Z_i^{(1)}$ (resp. $Z_i^{(2)}$) is independent from $Z_j^{(1)}$ (resp. $Z_j^{(2)}$) conditional on all other variables $Z_{\setminus i, j}^{(1)}$ (resp. $Z_{\setminus i, j}^{(2)}$). When the precision matrix $\Omega^{(k)}$ is nonsingular, the edges are characterized by its non zero entries.

The idea of neighborhood selection is to circumvent the intricate issue of estimating the precision matrix by recovering the sets of edges neighborhood by neighborhood, through the conditional distribution of $Z_i^{(k)}$ given all remaining variables $Z_{\setminus i}^{(k)}$. Indeed, this distribution is again a Gaussian distribution, whose mean is a linear combination of $Z_{\setminus i}^{(k)}$ while its variance is independent from $Z_{\setminus i}^{(k)}$.

Hence, $Z_i^{(k)}$ can be decomposed into the following linear regression:

$$Z_i^{(k)} = \sum_{j \neq i} Z_j^{(k)} \beta_{ij}^{(k)} + \varepsilon_i^{(k)} = Z_{\setminus i}^{(k)} \beta_i^{(k)} + \varepsilon_i^{(k)}, \quad (42)$$

where $\beta_{ij}^{(k)} = -\Omega_{ij}^{(k)} / \Omega_{ii}^{(k)}$ and $\text{Var}[\varepsilon_i^{(k)}] = (\Omega_{ii}^{(k)})^{-1}$.

Given an n_1 -sample of $Z^{(1)}$ and an n_2 -sample of $Z^{(2)}$, we recall that our objective is to test as formalized in (4)

$$\mathcal{H}_0^G : \Omega^{(1)} = \Omega^{(2)} \quad \text{versus} \quad \mathcal{H}_1^G : \Omega^{(1)} \neq \Omega^{(2)}.$$

As a result of Equation (42), testing for the equality of the matrix rows $\Omega_i^{(1)} = \Omega_i^{(2)}$ is equivalent to testing for $\beta_i^{(1)} = \beta_i^{(2)}$ and $\text{Var}[\varepsilon_i^{(1)}] = \text{Var}[\varepsilon_i^{(2)}]$. Denote by $\Sigma_{\setminus i}^{(k)}$ the covariance of $Z_{\setminus i}^{(k)}$. Under the null \mathcal{H}_0^G , we have that for any i , $\Sigma_{\setminus i}^{(1)} = \Sigma_{\setminus i}^{(2)}$. Consequently, we can translate the GGM hypotheses given in Equation (4) into a conjunction of two-sample linear regression tests:

$$\begin{aligned} \mathcal{H}_0^G : \quad & \bigcap_i \left[\beta_i^{(1)} = \beta_i^{(2)}, \Omega_{ii}^{(1)} = \Omega_{ii}^{(2)}, \Sigma_{\setminus i}^{(1)} = \Sigma_{\setminus i}^{(2)} \right] \\ \mathcal{H}_1^G : \quad & \bigcup_i \left[\beta_i^{(1)} \neq \beta_i^{(2)} \right] \cup \left[\Omega_{ii}^{(1)} \neq \Omega_{ii}^{(2)} \right]. \end{aligned} \quad (43)$$

Concretely, we apply the previous two-sample linear regression model with $\mathbf{X}^{(1)} = \mathbf{Z}_{\setminus i}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}_{\setminus i}^{(2)}, \mathbf{Y}^{(1)} = \mathbf{Z}_i^{(1)}$, and $\mathbf{Y}^{(2)} = \mathbf{Z}_i^{(2)}$ for every gene i and combine multiple neighborhood tests using a Bonferroni calibration as presented in Algorithm 5. The equality of $\sigma^{(k)}$'s in \mathcal{H}_0 models the equality of $\Omega_{ii}^{(k)}$'s in \mathcal{H}_0^G while the equality of $\Sigma^{(k)}$'s accounts for the equality of $\Sigma_{\setminus i}^{(k)}$'s.

Algorithm 5 Gaussian Graphical Model Testing Strategy

Require: Data $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$, maximum model dimension D_{max} and desired level α

for each gene $i = 1, \dots, p$ **do**

procedure NEIGHBORHOOD TEST

 Define $\mathbf{X}^{(1)} = \mathbf{Z}_{\setminus i}^{(1)}, \mathbf{X}^{(2)} = \mathbf{Z}_{\setminus i}^{(2)}$

 Define $\mathbf{Y}^{(1)} = \mathbf{Z}_i^{(1)}, \mathbf{Y}^{(2)} = \mathbf{Z}_i^{(2)}$

 Apply the Adaptive Testing Strategy of Algorithm 1 to $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}$

end procedure

end for

Reject the global null hypothesis if at least one Neighborhood Test is rejected at level α/p

Interpretation Because we need $\Omega_{ii}^{(1)} = \Omega_{ii}^{(2)}$ and $\Sigma_{\setminus i}^{(1)} = \Sigma_{\setminus i}^{(2)}$ for every neighborhood in the two-sample GGM null hypothesis \mathcal{H}_0^N (43), the assumptions that $\sigma^{(1)} = \sigma^{(2)}$ and $\Sigma^{(1)} = \Sigma^{(2)}$ in the two-sample linear regression null hypothesis \mathcal{H}_0 (3) are crucial for each neighborhood test to be interpreted correctly. As a result, only the global test can be strictly speaking interpreted in a statistically correct sense.

However in practice, when the global null hypothesis is rejected, our construction of neighborhood tests provides helpful clues on the location of disruptive regulations. In particular, for each rejected neighborhood test i , one can keep track of the rejected model S_R^i , retaining sensible information on which particular regulations are most likely altered between samples.

6.2. Illustration on real transcriptomic breast cancer data

We apply this strategy to the full (training and validation) breast cancer dataset studied by [23] and [37], whose training subset was originally published in [39]. 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 [23] and [37] was to develop a predictor of complete response to treatment from gene expression profiling. In particular, [23] identified an optimal predictive subset of 30 probes, mapping to 26 distinct genes.

Ambroise and co-authors [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 6. 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 $p = 26$ genes a neighborhood test $T_{\hat{S}_{Lasso}}^P$ at level $0.05/26$. We associate to each neighborhood test the empirical p-value defined in Equation (24), which has to be compared to α/p .

Most of the graph estimation methods proposed in the literature, such as the procedure of [1] leading to Figure 6, rely on the assumption that obser-

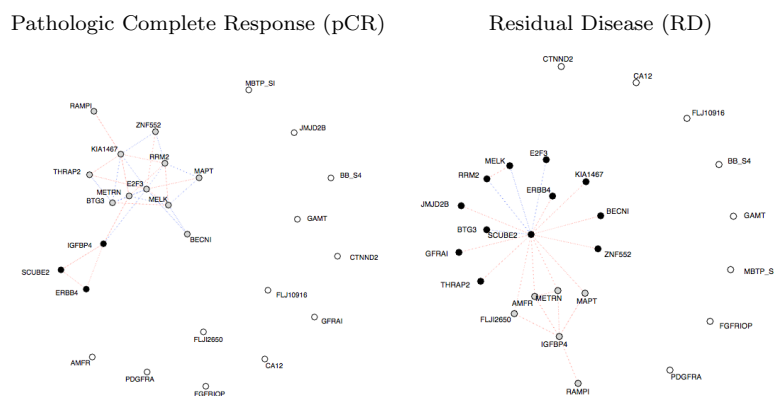


FIG 6. Graphs of conditional dependencies among the 26 genes selected by [23] on patients with pathologic complete response or residual disease with medium regularization as presented in Figure 3 of [1].

TABLE 3
Homogeneity test between training and test samples among pCR patients. Summary of test decisions after Bonferroni multiple testing correction and empirical p-values for each neighborhood test as defined in Section 3.4

	AMFR	BB_S4	BECNI	BTG3	CA12	CTNND2
decision	0	0	0	0	1	0
$p^{empirical}$	0.0492	0.0072	0.1972	1	0.0018	0.0100
	E2F3	ERBB4	FGFRIOP	FLJ10916	FLJI2650	GAMT
decision	0	0	0	0	0	0
$p^{empirical}$	0.1080	0.5610	0.0242	0.2542	0.0312	0.1158
	GFRAI	IGFBP4	JMJD2B	KIA1467	MAPT	MBTP_SI
decision	0	0	0	0	0	0
$p^{empirical}$	0.5318	0.0458	0.0128	0.0272	0.0178	0.0062
	MELK	METRN	PDGFRA	RAMPI	RRM2	SCUBE2
decision	0	0	1	0	0	0
$p^{empirical}$	0.5602	1	0.0012	0.0444	0.0022	0.2372
	THRAP2	ZNF552				
decision	0	0				
$p^{empirical}$	0.0228	0.0028				

vations are i.i.d. Yet the training and validation datasets have been collected and analyzed 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 (Table 3), two neighborhood tests corresponding to CA12 and PDGFRA are rejected at level $0.05/26$. Within RD patients (Table 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).

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. Reference [22] exhibited inhibitors of RRM2 ex-

TABLE 4

Homogeneity test between training and test samples among RD patients. Summary of test decisions after Bonferroni multiple testing correction and empirical p -values for each neighborhood test as defined in Section 3.4

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

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 empirical p -value and the rejected model are defined in Section 3.4

	AMFR	BB_S4	BECNI	BTG3	CA12	CTNND2
decision	1	1	1	1	1	1
$p^{empirical}$	<0.0001	<0.0001	4e-04	<0.0001	2e-04	<0.0001
rejected model	RRM2	RRM2	MAPT	MAPT	MAPT	RRM2
	E2F3	ERBB4	FGFRIOP	FLJ10916	FLJI2650	GAMT
decision	1	1	1	1	1	1
$p^{empirical}$	<0.0001	<0.0001	4e-04	4e-04	<0.0001	<0.0001
rejected model	MAPT	MELK	MAPT	RRM2	MAPT	RRM2
	GFRAI	IGFBP4	JMJD2B	KIA1467	MAPT	MBTP_SI
decision	1	1	1	1	1	1
$p^{empirical}$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
rejected model	BTG3	MELK	MAPT	MELK	RRM2	E2F3
	MELK	METRAN	PDGFRA	RAMPI	RRM2	SCUBE2
decision	1	1	1	1	1	1
$p^{empirical}$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
rejected model	MAPT	MELK	RRM2	RRM2	MAPT	BTG3
	THRAP2	ZNF552				
decision	1	1				
$p^{empirical}$	<0.0001	2e-04				
rejected model	E2F3	RRM2				

pression, which reduced *in vitro* and *in vivo* cell proliferation. Reference [40] led functional biology experiments validating the relationship between MAPT expression levels and response to therapy, suggesting to inhibit its expression to in-

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 empirical p -value and the rejected model are defined in Section 3.4

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

crease sensitivity to treatment. More recently, [14] developed a therapeutic candidate inhibiting MELK expression that was proved to suppress the growth of tumor-initiating cells in mice with various cancer types, including breast cancer.

For comprehensiveness, we add that a 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 (see Table 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 recognized as a novel tumor suppressor gene [30].

7. Discussion

Design hypotheses In this work, we have made two main assumptions on the design matrices:

- (i) The design matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are random.
- (ii) Under the null hypothesis (3), we further suppose that the population covariances $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are equal.

Although this setting is particularly suited to consider the two-sample GGM testing (Section 6), one may wonder whether one can circumvent these two

restrictions. We doubt that this is possible without making the testing problem much more difficult.

First, the formulation (3) allows the null hypothesis to be interpreted as a relevant intermediary case between two extreme fixed design 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 already been thoroughly studied. The second case is on the contrary extremely difficult as illustrated by the proposition below.

Proposition 7.1. *Consider the design matrices $\mathbf{X}^{(1)}$ 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 of $\beta^{(1)} = \beta^{(2)}$ vs $\beta^{(1)} \neq \beta^{(2)}$ based on the data (\mathbf{Y}, \mathbf{X}) satisfies:*

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

where $\mathbb{P}_{\beta^{(1)}, \beta^{(2)}}(\cdot)$ denotes the distribution of $(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)})$. 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.

Under the assumptions of the above proposition, for any $\beta^{(1)} \neq \beta^{(2)}$ there exists some vector β such that $\mathbf{X}^{(1)}\beta^{(1)} = \mathbf{X}^{(1)}\beta$ and $\mathbf{X}^{(2)}\beta^{(2)} = \mathbf{X}^{(2)}\beta$. Consequently, it is impossible to distinguish the null hypothesis from the alternative hypothesis.

Furthermore, if $\Sigma^{(1)} \neq \Sigma^{(2)}$ is allowed in the null (3), then the two-sample testing problem seems to become much more difficult in the sense that it is impossible to reformulate the null hypothesis into a conjunction of low-dimensional hypotheses as done in Lemma 2.1. Indeed, consider the following toy example: $\sigma^{(1)} = \sigma^{(2)} = 1$, $\beta^{(1)} = \beta^{(2)} = (a, 0, 0, \dots)^T$ for some $a > 0$, $\Sigma^{(1)} = I_p$ and $\Sigma^{(2)} = (\rho + \mathbf{1}_{i=j})_{1 \leq i, j \leq p}$ for some $\rho > 0$. Then, for any subset S that does not contain the first component, the parameters $\beta_S^{(1)}$ and $\beta_S^{(2)}$ are different. Consequently, $\beta_S^{(1)} \neq \beta_S^{(2)}$ does not imply that $\beta^{(1)} \neq \beta^{(2)}$ and one should not rule out the parameter equality hypothesis relying on some low-dimensional regressions.

Comparison with related work [43, 44] Städler and Mukherjee propose a very general approach to high-dimensional two-sample testing, being applicable to a wide range of models. In particular this approach allows for the direct comparison of two-sample Gaussian graphical models without adopting a neighborhood selection approach. This avoids the burden of multiple neighborhood linear regression and the multiple testing correction which follows.

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 rate or power for their overall testing strategy.

Finally, it appears in the numerical experiments that our procedure outperforms the DiffRegr Procedure when n is small but that the multi-split procedure shows stable results and performs well even in difficult scenarios.

Non asymptotic bounds and constants 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 [16] considers an asymptotic but high-dimensional framework and aims to provide detection rates with optimal constants. For instance [2, 24] 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. While deriving sharp detection rates (under perhaps stronger assumptions on the covariance) is a stimulating open problem, it is beyond the scope of our paper.

Loss functions and interpretation 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 4.1 and 4.4 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.

Gene network inference 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. In that particular respect, exploiting our multiple testing scheme to pinpoint differences between the networks, or identifying differentially regulated gene-sets within the networks as procedure [44] does, will be extremely useful to exhibit new targets and pathways of interest.

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 subsets of the well-known Hess *et al.*

dataset, suggesting that these samples should be used separately as originally intended. Methods which require i.i.d. observations should only be applied with caution to this dataset if considered as a single large and homogeneous sample.

8. Proofs

8.1. Two-sample testing for fixed and different designs

Proofs of Proposition 7.1. 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. \square

8.2. Upper bounds of the quantiles

Proof of Proposition 3.3. In the sequel, we note $N = n_1 - |S|$. Furthermore $(Z_1, \dots, Z_{|S|})$ denotes a standard Gaussian random vector and W_N is a χ^2 random variable with N degrees of freedom. We apply Laplace method to upper bound $\mathbb{P}[F_{S,1} \geq u]$:

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

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 canceling 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:

$$|a|_1 + \frac{2\lambda \|a\|^2}{1 - 2|a|_\infty \lambda} - \frac{u}{1 + \frac{2\lambda u}{N}}. \quad (44)$$

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

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 $\mathbf{X}_S^{(1)}$. Finally, to use similar notations for each $i \in \{V, 1, 2\}$, we define $\tilde{Q}_{V,|S|}(F_{S,V}|\mathbf{X}_S) := \tilde{q}_{V,S} = \overline{Q}_{V,|S|}(F_{S,V}|\mathbf{X}_S)$ the exact p -value associated to $F_{S,V}$.

8.3. Distributions of $F_{S,V}$, $F_{S,1}$ and $F_{S,2}$ (Proposition 3.1)

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

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

Define the random variable T_1 and T_2 as

$$T_1 = \frac{\|\Pi_{S^\perp}^{(1)}\epsilon_S^{(1)}\|^2}{(n_1 - |S|)(\sigma_S^{(1)})^2}, \quad T_2 = \frac{\|\Pi_{S^\perp}^{(2)}\epsilon_S^{(2)}\|^2}{(n_2 - |S|)(\sigma_S^{(2)})^2}. \quad (45)$$

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

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

allows us to prove the first assertion of Proposition 3.1. Let us turn to the second statistic:

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

where under the null hypothesis $\mathcal{H}_{0,S}$

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

Given \mathbf{X}_S , 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

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

8.4. Calibrations

Proof of Proposition 3.4. By definition of the p -values $\tilde{q}_{i,S}$, we have under \mathcal{H}_0 for each $S \in \mathcal{S}$ and each $i \in \{V, 1, 2\}$

$$\mathbb{P}[\tilde{q}_{i,S} \leq \alpha_{i,S}|\mathbf{X}_S] \leq \alpha_{i,S}.$$

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

$$\begin{aligned} \mathbb{P}[T_{\widehat{\mathcal{S}}}^B = 1] &= \mathbb{P}\left[\exists S \in \widehat{\mathcal{S}}, \exists i \in \{V, 1, 2\}, \tilde{q}_{i,S} \leq \alpha_{i,S}\right] \\ &\leq \sum_{S \in \mathcal{S}} \sum_{i=V,1,2} \mathbb{P}(\tilde{q}_{i,S} < \alpha_{i,S}) \\ &\leq \sum_{S \in \mathcal{S}} \sum_{i=V,1,2} \mathbb{E}[\mathbb{P}(\tilde{q}_{i,S} < \alpha_{i,S} | \mathbf{X}_S)] \leq \sum_{S \in \mathcal{S}} \alpha_{i,S} \leq \alpha, \end{aligned}$$

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

Proof of Proposition 3.5. Consider $i \in \{V, 1, 2\}$. Under \mathcal{H}_0 , the distributions of

$$\begin{aligned} &\min_{S \in \widehat{\mathcal{S}}^\pi} \left[\tilde{q}_{V,S}^\pi \binom{p}{|S|} \right], \\ &\min_{S \in \widehat{\mathcal{S}}^\pi} \left[(\tilde{q}_{1,S}^\pi \wedge \tilde{q}_{2,S}^\pi) \binom{p}{|S|} \right] \end{aligned}$$

are invariant with respect to the permutation π . Hence, we derive that under \mathcal{H}_0

$$\begin{aligned} &\mathbb{P}\left[\min_{S \in \widehat{\mathcal{S}}^\pi} \tilde{q}_{V,S}^\pi \binom{p}{|S|} \leq \widehat{C}_V\right] = \alpha/2, \\ &\mathbb{P}\left[\min_{S \in \widehat{\mathcal{S}}^\pi} (\tilde{q}_{1,S}^\pi \wedge \tilde{q}_{2,S}^\pi) \binom{p}{|S|} \leq \widehat{C}_1\right] = \alpha/2. \end{aligned}$$

Applying an union bound, we conclude that the type I error is smaller than the sum of these two last probabilities. \square

8.5. Proof of Theorem 4.3

Intuitively, the test T_S^B should be powerful when some distance between the two sample-specific distributions is large enough. The objective here is to characterize the minimal distance that enforce the power of the procedure to be larger than $1 - \delta$. As exposed in Theorem 4.3, we rely on the semi-distances $\mathcal{K}_1(S) + \mathcal{K}_2(S)$ for $S \in \mathcal{S}$:

$$\begin{aligned} 2(\mathcal{K}_1(S) + \mathcal{K}_2(S)) &= \left(\frac{\sigma_S^{(1)}}{\sigma_S^{(2)}}\right)^2 + \left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}}\right)^2 - 2 + \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_S^{(1)})^2} \\ &\quad + \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(1)}}^2}{(\sigma_S^{(2)})^2}. \end{aligned} \tag{46}$$

We start from the definition (7) of the test:

$$\mathbb{P}[T_S^B = 1] = \mathbb{P}[\exists S \in \mathcal{S}, \exists i \in \{V, 1, 2\}, \tilde{q}_{i,S} \leq \alpha_{i,S}]$$

$$\geq \sup_{S \in \mathcal{S}} \mathbb{P}[\tilde{q}_{V,S} \wedge \tilde{q}_{1,S} \wedge \tilde{q}_{2,S} \leq \alpha_S],$$

where we recall that $\alpha_S = \min_{i=V,1,2} \alpha_{i,S}$. In the remainder of the proof, we show, that if for some $S \in \mathcal{S}$, both conditions (30) and (31) are fulfilled then $\mathbb{P}[\tilde{q}_{V,S} \wedge \tilde{q}_{1,S} \wedge \tilde{q}_{2,S} \leq \alpha_S]$ is larger than $1 - \delta$.

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

$$\log(12/\delta) < L_1^\bullet(n_1 \wedge n_2), \quad \log(1/\alpha_S) \leq L_2^\bullet(n_1 \wedge n_2), \quad |S| \leq L_3^\bullet \quad (47)$$

where the numerical constants L_1^\bullet , L_2^\bullet , and L_3^\bullet only depend on L_2^* in (54) and on the constants introduced in Lemmas 8.1–8.5. These conditions allow us to fix the constants in the statement (30) of Theorem 4.3.

Lemma 8.1 (Upper-bound of $\tilde{Q}_{V,|S|}^{-1}(x|\mathbf{X}_S)$). *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*

$$\tilde{Q}_{V,|S|}^{-1}(x|\mathbf{X}_S) \leq L \left\{ \left(\frac{|S|(n_1 - n_2)}{n_1 n_2} \right)^2 + \log(2/x) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right\}. \quad (48)$$

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 $\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S)$). *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 $|S| \leq L_1 n_1$,*

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

For any $0 < x < 1$ satisfying

$$L_2 \log(1/x) \leq n_1 - |S|, \quad (49)$$

we have the following upper bound

$$\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S) \leq |a|_\infty \left[2|S| + 2\sqrt{2|S| \log(1/x)} + 8 \log(1/x) \right]. \quad (50)$$

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 $F_{S,V}$). *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*

$$F_{S,V} \geq L_2 \left(\frac{(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2}{\sigma_S^{(1)} \sigma_S^{(2)}} \right)^2 - L_3 \left[|S|^2 \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} \right) + \log \left(\frac{1}{\delta} \right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]. \quad (51)$$

Lemma 8.5 (Deviations of $F_{S,1}$). *There exist two positive universal constants L_1 and L_2 such that the following holds. Assume that*

$$L_1 \log(12/\delta) < n_1 \wedge n_2. \quad (52)$$

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

$$F_{S,1} \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{8(\sigma_S^{(1)})^2} - \log(12/\delta) L_2 \left[\frac{1}{n_2} \frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2} + \frac{\varphi_S}{n_1} \right], \quad (53)$$

where φ_S is defined in (29).

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

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

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

$$\begin{aligned} \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_S^{(1)})^2} &\geq L'_1 (\varphi_S + 1) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{12}{\delta \alpha_S} \right) \right] \\ &\quad + \frac{L'_2}{n_2} \left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}} \right)^2 \log \left(\frac{12}{\delta} \right). \end{aligned}$$

A symmetric result holds for $\tilde{Q}_{2,|S|}(F_{S,2}|\mathbf{X}_S)$. Consequently, we have

$$\mathbb{P}[T_S^B = 1] \geq \mathbb{P}[\tilde{q}_{V,S} \wedge \tilde{q}_{1,S} \wedge \tilde{q}_{2,S} \leq \alpha_S] \geq 1 - \delta$$

as soon as

$$\begin{aligned} \mathcal{K}_1(S) + \mathcal{K}_2(S) &\geq L_1^* \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|S| + \log \left(\frac{12}{\alpha_S \delta} \right) \right] \\ &\quad + L_2^* \log(12/\delta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[\left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}} \right)^2 + \left(\frac{\sigma_S^{(1)}}{\sigma_S^{(2)}} \right)^2 \right]. \end{aligned} \quad (54)$$

Since it is assumed that $4L_2^* \log(12/\delta) \leq n_1 \wedge n_2$ (see (47)), this last condition is fulfilled as soon as

$$\mathcal{K}_1(S) + \mathcal{K}_2(S) \geq L^* \varphi_S \left(\frac{1}{n_1} + \frac{1}{n_2} \right) [|S| + \log\{12/(\alpha_S \delta)\}].$$

This concludes the proof. 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 (45). Given \mathbf{X}_S ,

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

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,

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

where we used in the last inequality that $|S| \leq (n_1 \wedge n_2)/2$. Similarly, with probability at least $1 - x/2$, we have

$$\begin{aligned} & \frac{T_2 n_1(n_2 - |S|)}{T_1 n_2(n_1 - |S|)} \\ & \leq \left[\left(1 + \frac{|S|(n_2 - n_1)}{n_2(n_1 - |S|)} \right) \left(1 + 12\sqrt{\frac{\log(2/x)}{n_1}} + 12\sqrt{\frac{\log(2/x)}{n_2}} \right) \right] \wedge L. \quad (55) \end{aligned}$$

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

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

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

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

□

Proof of Lemma 8.2. As in the proof of Proposition 3.3, we note $N = n_1 - |S|$. Recall that $\tilde{Q}_{1,|S|}(u|\mathbf{X}_S)$ is defined as $\exp \psi_u(\lambda^*)$ (see Definition 3.2). We start by upper-bounding $\psi_u(\lambda^*)$. This will entail the first upper-bound of the tail probability $\log \tilde{Q}_{1,|S|}(u|\mathbf{X}_S)$. We then exhibit a value u_x such that $\psi_{u_x}(\lambda^*) \leq \log x$.

Upper-bound of the tail probability First we provide two bounds on λ_* . Since Equation (44) 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 $u \leq N|a|_\infty$, we have $2\lambda^*u \leq N$. 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

$$\begin{aligned}
\psi_u(\lambda^*) &\leq |a|_1 \lambda_+ + \frac{\lambda_+^2 \|a\|^2}{1 - 2|a|_\infty \lambda_+} - \lambda^* u + 2 \frac{(\lambda^*)^2 u^2}{N} \\
&\leq -\frac{(u - |a|_1)^2}{4[|a|_\infty(u - |a|_1) + \|a\|^2]} + \frac{2\lambda_+^2 u^2}{N} + (\lambda_+ - \lambda_-)u \\
&\leq -\frac{(u - |a|_1)^2}{4[|a|_\infty(u - |a|_1) + \|a\|^2]} + \frac{(u - |a|_1)u^3}{2N[|a|_\infty(u - |a|_1) + \|a\|^2]}. \quad (56)
\end{aligned}$$

Upper-bound of the quantile Let us turn to the upper bound of $\tilde{Q}_{1,|S|}^{-1}(x|\mathbf{X}_S)$. Consider u_x the unique 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 (49) 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 (56) that

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

which, in turn, is smaller than $-\log(1/x)$ if we take L_2 large enough in Condition (49). If $u_x \leq 2|a|_1$, then $|a|_1^2 / (|a|_\infty |a|_1 + \|a\|^2) \geq 8 \log(1/x)$ and

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

which is smaller than $-\log(1/x)$ if we take L_1 and L_2 large enough in the two aforementioned conditions. since $|S| \leq 2^{-6} n_1$. Thus, we conclude that

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

□

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.

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

$$\leq \frac{2}{n_2} + 2 \frac{\varphi_{\max}[\mathbf{Z}_S^{(2)\top} \mathbf{Z}_S^{(2)}]}{n_2 \varphi_{\min}[\mathbf{Z}_S^{(1)\top} \mathbf{Z}_S^{(1)}]} \varphi_{\max} \left[\sqrt{\Sigma_S^{(2)}} (\Sigma_S^{(1)})^{-1} \sqrt{\Sigma_S^{(2)}} \right].$$

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

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 3.1. Given \mathbf{X}_S , T_1/T_2 follows a Fisher distribution with $(n_1 - |S|, n_2 - |S|)$ degrees of freedom. Moreover

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

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

$$\begin{aligned} F_{S,V} &\geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2 (\sigma_S^{(2)})^2} + \frac{(\sigma_S^{(1)})^2}{(\sigma_S^{(2)})^2} \left(\frac{T_1}{T_2} - 1 \right) + \frac{(\sigma_S^{(2)})^2}{(\sigma_S^{(1)})^2} \left(\frac{T_2}{T_1} - 1 \right) \\ &\geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2 (\sigma_S^{(2)})^2}. \end{aligned} \tag{57}$$

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

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

We need to control the deviations of T_2/T_1 . Using the bound (55), 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

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

which is also smaller than 3 if the constant L_1 is large enough in the statement of the lemma. In conclusion, we have

$$\mathbb{P} \left[F_{S,V} \geq \frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{16(\sigma_S^{(1)})^2 (\sigma_S^{(2)})^2} \right] \geq 1 - \delta, \tag{58}$$

as long as

$$\frac{[(\sigma_S^{(1)})^2 - (\sigma_S^{(2)})^2]^2}{(\sigma_S^{(1)})^2(\sigma_S^{(2)})^2} \geq L \left[\frac{|S|^2}{n_1^2} + \frac{|S|^2}{n_2^2} + \log(1/\delta) \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]. \quad (59)$$

Combining (57), (58), and (59), we derive

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

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

Proof of Lemma 8.5. We want to lower bound the random variable $F_{S,1} = \frac{n_1 R}{(\sigma_S^{(1)})^2 T_1(n_1 - |S|)}$ 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 $(\mathbf{X}_S^{(1)}, \mathbf{X}_S^{(2)})$, $\langle W, \frac{\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})}{\|\mathbf{X}_S^{(2)}(\beta_S^{(2)} - \beta_S^{(1)})\|} \rangle^2$ 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 $1 - x/6$,

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

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

$$\mathbb{P} \left[\varphi_{\max} \left[\mathbf{X}_S^{(2)}(\mathbf{X}_S^{(1)\top} \mathbf{X}_S^{(1)})^{(-1)} \mathbf{X}_S^{(2)\top} \right] \leq L\varphi_S \frac{n_2}{n_1} \right] \geq 1 - 2 \exp[-(n_1 \wedge n_2)L'].$$

If we take the constant L_1 large enough in condition (52), then we derive that

$$R \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{4} - \log(12/\delta) L \left[\frac{(\sigma_S^{(2)})^2}{n_2} + \frac{(\sigma_S^{(1)})^2}{n_1} \varphi_S \right], \quad (60)$$

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

$$T_1(n_1 - |S|)/n_1 \leq 1 + 2\sqrt{\frac{\log(6/\delta)}{n_1}} + \frac{2}{n_1} \log(6/\delta) \leq 2, \quad (61)$$

with probability larger than $1 - \delta/6$. Gathering (60) and (61), we conclude that

$$F_{S,1} \geq \frac{\|\beta_S^{(2)} - \beta_S^{(1)}\|_{\Sigma^{(2)}}^2}{8(\sigma_S^{(1)})^2} - \log\left(\frac{12}{\delta}\right) L \left[\frac{1}{n_2} \left(\frac{\sigma_S^{(2)}}{\sigma_S^{(1)}}\right)^2 + \frac{\varphi_S}{n_1} \right],$$

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

8.6. Proof of Theorem 4.1: Power of $T_{S \leq k}^B$

This proposition is a straightforward corollary of Theorem 4.3. Consider the subsets S_\vee and S_Δ of $\{1, \dots, p\}$ such that S_\vee is the union of the support of $\beta^{(1)}$ and $\beta^{(2)}$ and S_Δ is the support of $\beta^{(2)} - \beta^{(1)}$. Assume first that S_\vee and S_Δ are non empty. By Definition (20) of the weights, we have

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

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_\vee and S_Δ follow the conditions of Theorem 4.3.

Applying Theorem 4.3, we derive that $T_{S \leq k}^B$ rejects \mathcal{H}_0 with probability larger than $1 - \delta$ when

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

Observing that $\varphi_{S_\vee} \leq \varphi_{\Sigma^{(1)}, \Sigma^{(2)}}$, $\mathcal{K}_1(S_\vee) = \mathcal{K}_1$, $\mathcal{K}_2(S_\vee) = \mathcal{K}_2$ and that $|S_\vee| \leq |\beta^{(1)}|_0 + |\beta^{(2)}|_0$ concludes the first part of the proof. Let us turn to the second result. According to Theorem 4.3, $T_{S \leq k}^B$ rejects \mathcal{H}_0 with probability larger than $1 - \delta$ when

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

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 $S_V = \emptyset$, then we can consider any subset of size 1 to prove the first result. If $S_\Delta = \emptyset$, then $\beta^{(1)} = \beta^{(2)}$ and Condition (26) cannot be fulfilled.

8.7. Proof of Proposition 4.5

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 $S \in \mathcal{S}_1 \subset \widehat{\mathcal{S}}_{\text{Lasso}}$. Recall that we note $\alpha_S := \min_{i=V,1,2} \alpha_{i,S}$. We start from the basic inequality

$$\mathbb{P} \left[T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B = 0 \right] \leq \mathbb{P} \left[\forall S \in \widehat{\mathcal{S}}_{\text{Lasso}}, \min_{i=V,1,2} \tilde{q}_{i,S} > \alpha_S \right].$$

Denote \widehat{S} any (possibly random) collection of indices. If $\widehat{S} \in \widehat{\mathcal{S}}_{\text{Lasso}}$, then $T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B = 0$ implies $\min_{i=V,1,2} \tilde{q}_{i,\widehat{S}} > \alpha_{\widehat{S}}$. Hence, we get

$$\mathbb{P} \left[T_{\widehat{\mathcal{S}}_{\text{Lasso}}}^B = 0 \right] \leq \mathbb{P} \left[\widehat{S} \notin \widehat{\mathcal{S}}_{\text{Lasso}} \right] + \mathbb{P} \left[\left\{ \min_{i=V,1,2} \tilde{q}_{i,\widehat{S}} > \alpha_{\widehat{S}} \right\} \cap \left\{ \widehat{S} \in \widehat{\mathcal{S}}_{\text{Lasso}} \right\} \right]. \quad (62)$$

In the sequel, we shall construct of subset \widehat{S} belonging to $\widehat{\mathcal{S}}_{\text{Lasso}}$ with large probability and such that \widehat{S} is close enough to S_V^* so that the p values $\min_{i=V,1,2} \tilde{q}_{i,\widehat{S}}$ is smaller than $\alpha_{\widehat{S}}$ 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,

$$\begin{bmatrix} \mathbf{Y}^{(1)} \\ \mathbf{Y}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{(1)} & \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} & -\mathbf{X}^{(2)} \end{bmatrix} \begin{bmatrix} \theta_*^{(1)} \\ \theta_*^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}^{(1)} \\ \boldsymbol{\epsilon}^{(2)} \end{bmatrix} \quad (63)$$

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 the sequel, we fix

$$\lambda_0 = 16(\sigma^{(1)} \vee \sigma^{(2)}) \sqrt{2(n_1 + n_2) \left[\Phi_{1,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{1,+}(\sqrt{\Sigma^{(2)}}) \right] \log(p)}. \quad (64)$$

and we consider the set $\widehat{S} := \widehat{S}_{\lambda_0}$ defined by the union of the support of $\widehat{\theta}_{\lambda_0}^{(1)}$ and $\widehat{\theta}_{\lambda_0}^{(2)}$. By definition of $\widehat{\mathcal{S}}_{\text{Lasso}}^{(1)}$, the subset \widehat{S}_{λ_0} belongs to $\widehat{\mathcal{S}}_{\text{Lasso}}$ as soon as the size of the Lasso estimators \widehat{S}_λ are smaller than k_* for all $\lambda \geq \lambda_0$. In order to bound the probability of this event, we study the behavior of the Lasso estimator $\widehat{\theta}_\lambda$.

Lemma 8.6 (Control of the design \mathbf{W}). *If we take the constants L^* , L_1^* , and L_2^* in Proposition 4.5 small enough then the following holds. The event*

$$\mathcal{A} := \left\{ \forall \theta \text{ s.t. } |\theta|_0 \leq k_*, \ 1/2 \leq \frac{\|\mathbf{X}^{(1)}\theta\|^2}{n_1 \|\theta\|_{\Sigma^{(1)}}^2} \leq 2 \text{ and } 1/2 \leq \frac{\|\mathbf{X}^{(2)}\theta\|^2}{n_2 \|\theta\|_{\Sigma^{(2)}}^2} \leq 2 \right\}$$

$$\bigcap \left\{ \frac{\kappa [6, |\theta_*|_0, \mathbf{X}^{(1)}/\sqrt{n_1}]}{\kappa [6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}]} \wedge \frac{\kappa [6, |\theta_*|_0, \mathbf{X}^{(2)}/\sqrt{n_1}]}{\kappa [6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} \geq 2^{-3} \right\}$$

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

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

for any $k \leq k_*$.

The following property is a slight variation of Lemma 11.2 in [46] and Lemma 3.2 in [21].

Lemma 8.7 (Behavior of the Lasso estimator $\widehat{\theta}_\lambda$). *If we take L_2^* in Proposition 4.5 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

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq L_1 \frac{\lambda^2/(n_1 \wedge n_2)}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0, \quad (65)$$

$$|\widehat{\theta}_\lambda|_0 \leq L_2 \frac{n_1 \vee n_2}{n_1 \wedge n_2} \frac{\Phi_{k_*,+}(\sqrt{\Sigma^{(1)}}) \vee \Phi_{k_*,+}(\sqrt{\Sigma^{(2)}})}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0 \leq k_*/2. \quad (66)$$

We deduce from Lemma 8.7 that, on the event $\mathcal{A} \cap \mathcal{B}$, the cardinal of \widehat{S}_λ is smaller than k_* for all $\lambda \geq \lambda_0$ and \widehat{S}_{λ_0} belongs therefore to $\widehat{S}_L^{(1)} \subset \widehat{S}_{\text{Lasso}}$ (see Eq. (12)). Thus,

$$\mathbb{P} \left[\widehat{S}_{\lambda_0} \notin \widehat{S}_{\text{Lasso}} \right] \leq \mathbb{P}[\mathcal{A}^c] + \mathbb{P}[\mathcal{B}^c] \leq \delta/4 + p^{-1} \leq \delta/2,$$

since δ is supposed to be larger than $p/4$ (Condition (37)). From (62), we deduce that

$$\begin{aligned} \mathbb{P} \left[T_{\widehat{S}_{\text{Lasso}}}^B = 0 \right] &\leq \mathbb{P} \left[\widehat{S}_{\lambda_0} \notin \widehat{S}_{\text{Lasso}} \right] + \mathbb{P} \left[\left\{ \min_{i=V,1,2} \tilde{q}_{i,\widehat{S}} > \alpha_{\widehat{S}} \right\} \cap \left\{ \widehat{S} \in \widehat{S}_{\text{Lasso}} \right\} \right] \\ &\leq \mathbb{P}[(\mathcal{A} \cap \mathcal{B})^c] + \mathbb{P} \left[\left\{ \min_{i=V,1,2} \tilde{q}_{i,\widehat{S}_{\lambda_0}} > \alpha_{\widehat{S}_{\lambda_0}} \right\} \cap \mathcal{A} \cap \mathcal{B} \right]. \quad (67) \end{aligned}$$

Thus, it suffices to prove that the right hand side probability is smaller than $\delta/2$.

In the following lemma, we compare $\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0})$ to $\mathcal{K}_1 + \mathcal{K}_2$. Define

$$R_{\Sigma^{(1)}, \Sigma^{(2)}} = \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})} \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]}.$$

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

$$L \left[\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0}) \right] \geq 1 \wedge \left[\mathcal{K}_1 + \mathcal{K}_2 - L' R_{\Sigma^{(1)}, \Sigma^{(2)}} \frac{|S_V|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p) \right].$$

We show in Lemma 8.9 below that the p -values $\tilde{q}_{i, \widehat{S}_{\lambda_0}}$ are smaller than $\alpha_{\widehat{S}_{\lambda_0}}$ as soon as $\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0})$ is large enough.

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

$$\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0}) \geq L \varphi_{\widehat{S}_{\lambda_0}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \left[|\widehat{S}_{\lambda_0}| \log(p) + \log \left(\frac{1}{\alpha \delta} \right) + \log(p) \right], \quad (68)$$

then, $\mathbb{P}[\{\min_{i \in \{V, 1, 2\}} \tilde{q}_{i, \widehat{S}_{\lambda_0}} > \alpha_{\widehat{S}_{\lambda_0}}\} \cap \mathcal{A} \cap \mathcal{B}] \leq \delta/2$.

Under the event $\mathcal{A} \cap \mathcal{B}$, we derive from (66) that,

$$|\widehat{S}_{\lambda_0}| \leq L' \frac{n_1 \vee n_2}{n_1 \wedge n_2} \frac{\bigvee_{i=1,2} \Phi_{k_*, +}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} |S_V|.$$

If we take the numerical constant L_3^* large enough in Proposition 4.5, it then follows from Lemma 8.8 and from the assumption on $\mathcal{K}_1 + \mathcal{K}_2$ that Condition (68) in Lemma 8.9 is satisfied. Then, gathering this lemma with the bound (67), we conclude that the type II error is smaller than δ .

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 (36) and (37) allows us to prove that

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

Applying Corollary 1 in [38], 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 $|\theta_*|_0 \log(p) < c_3 \frac{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]}{\bigvee_{i=1,2} \Phi_{1, +}(\sqrt{\Sigma^{(i)}})} (n_1 \wedge n_2)$. Since $\log(1/\delta)$ is small in front of $n_1 \wedge n_2$ (Condition (37)), 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

$$\|\mathbf{W}\theta\|^2 = \|\mathbf{X}^{(1)}(\theta^{(1)} + \theta^{(2)})\|^2 + \|\mathbf{X}^{(2)}(\theta^{(1)} - \theta^{(2)})\|^2$$

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

Proof of Lemma 8.7. Observe that the variance of $[\mathbf{W}^\top \boldsymbol{\epsilon}]_i$ 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 η in (35). A slight variation of Lemma 11.2 in [46] ensures that on event \mathcal{B} :

$$\|\mathbf{W}(\widehat{\theta}_\lambda - \theta_*)\|^2 \leq L \frac{\lambda^2}{\eta^2[3, |\theta_*|_0, \mathbf{W}]} |\theta_*|_0. \quad (69)$$

In order to show (65), we need to lower bound $\eta^2[3, |\theta_*|_0, \mathbf{W}]$. Fix $k = |\theta_*|_0$ and consider any $\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \end{pmatrix} \in \mathcal{C}(3, T)$ with $|T| = k$. We shall prove that either $\theta^{(1)} + \theta^{(2)}$ or $\theta^{(1)} - \theta^{(2)}$ belongs to some cone. Then, we will deduce from it a control of the ratio $k\|\mathbf{W}\theta\|^2/|\theta|_1^2$ which will enforce a bound of $\eta^2[3, |\theta_*|_0, \mathbf{W}]$. Define $T' = \{i \in \{1, \dots, p\}, \text{ s.t. } i \in T \text{ or } i + p \in T\}$. We have

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

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

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

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

$$\begin{aligned}
\frac{k\|\mathbf{W}\theta\|^2}{|\theta|_1^2} &\geq \frac{k\|\mathbf{X}^{(1)}(\theta^{(2)} + \theta^{(1)})\|^2}{|\theta|_1^2} + \frac{k\|\mathbf{X}^{(2)}(\theta^{(2)} - \theta^{(1)})\|^2}{|\theta|_1^2} \\
&\geq \frac{(n_1 \wedge n_2)|\theta^{(2)} + \theta^{(1)}|_1^2}{|\theta|_1^2} \left[\bigwedge_{i=1,2} \eta^2 \left(6, k, \mathbf{X}^{(i)} / \sqrt{n_i} \right) \right] \\
&\geq L(n_1 \wedge n_2) \left[\bigwedge_{i=1,2} \kappa^2 \left(6, k, \mathbf{X}^{(i)} / \sqrt{n_i} \right) \right] \\
&\geq L(n_1 \wedge n_2) \left[\kappa^2 \left(6, k, \sqrt{\Sigma^{(1)}} \right) \wedge \kappa^2 \left(6, k, \sqrt{\Sigma^{(2)}} \right) \right], \quad (70)
\end{aligned}$$

where we used in the second line that $\theta^{(1)} + \theta^{(2)} \in \mathcal{C}(6, T')$ and in the third line that $|\theta^{(1)} + \theta^{(2)}|_1 \geq |\theta|_1/8$. The last inequality derives from the definition of the event \mathcal{A} . Exchanging the role of $\theta^{(1)} + \theta^{(2)}$ and $\theta^{(1)} - \theta^{(2)}$, we also prove (70) when $|(\theta^{(1)} + \theta^{(2)})_{T'}|_1 \leq |(\theta^{(1)} - \theta^{(2)})_{T'}|_1$. Thus, we have proved that

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

Gathering this bound with (69), 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 (65).

Let us turn to the proof of (66). Lemma 3.1 in [21] 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

$$|\widehat{\theta}_\lambda|_0 \leq L \frac{\Phi_{|\widehat{\theta}_\lambda|_0, +}(\mathbf{W})}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(1)}}] \wedge \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(2)}}]} |\theta_*|_0. \quad (71)$$

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_2^* in (38) small enough. Hence, $|\widehat{\theta}_\lambda|_0 \leq k_*$. Coming back to (71), we can now replace $\Phi_{|\widehat{\theta}_\lambda|_0, +}(\mathbf{W})$ by $\Phi_{k_*, +}(\mathbf{W})$. We obtain (66). \square

Proof of Lemma 8.8. Given the Lasso estimator $\widehat{\theta}_{\lambda_0}$ of θ_* in model (63), we define $\widehat{\beta}_{\lambda_0}^{(1)}$ and $\widehat{\beta}_{\lambda_0}^{(2)}$ by

$$\widehat{\beta}_{\lambda_0}^{(1)} = \widehat{\theta}_{\lambda_0}^{(1)} + \widehat{\theta}_{\lambda_0}^{(2)}, \quad \widehat{\beta}_{\lambda_0}^{(2)} = \widehat{\theta}_{\lambda_0}^{(1)} - \widehat{\theta}_{\lambda_0}^{(2)}.$$

On event $\mathcal{A} \cap \mathcal{B}$, we upper bound the difference between $(\beta^{(1)}, \beta^{(2)})$ and $(\widehat{\beta}_{\lambda_0}^{(1)}, \widehat{\beta}_{\lambda_0}^{(2)})$.

$$\begin{aligned} &\|\beta^{(1)} - \widehat{\beta}_{\lambda_0}^{(1)}\|_{\Sigma^{(1)}}^2 + \|\beta^{(2)} - \widehat{\beta}_{\lambda_0}^{(2)}\|_{\Sigma^{(2)}}^2 \\ &\leq 2 \left[\left\| \frac{\mathbf{X}^{(1)}}{\sqrt{n_1}} (\beta^{(1)} - \widehat{\beta}_{\lambda_0}^{(1)}) \right\|^2 + \left\| \frac{\mathbf{X}^{(2)}}{\sqrt{n_2}} (\beta^{(2)} - \widehat{\beta}_{\lambda_0}^{(2)}) \right\|^2 \right] \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{n_1 \wedge n_2} \|\mathbf{W}(\theta_* - \hat{\theta}_{\lambda_0})\|^2 \\
&\leq L \frac{\bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} \frac{|S_{\vee}|(n_1 \vee n_2)}{(n_1 \wedge n_2)^2} \log(p) (\sigma^{(1)} \vee \sigma^{(2)})^2, \quad (72)
\end{aligned}$$

where the last inequality follows from Lemmas 8.6 and 8.7. Let us now lower bound the Kullback discrepancy $2[\mathcal{K}_1(\hat{S}_{\lambda_0}) + \mathcal{K}_2(\hat{S}_{\lambda_0})]$ which is equal to

$$\left(\frac{\sigma_{\hat{S}_{\lambda_0}}^{(1)}}{\sigma_{\hat{S}_{\lambda_0}}^{(2)}} \right)^2 + \left(\frac{\sigma_{\hat{S}_{\lambda_0}}^{(1)}}{\sigma_{\hat{S}_{\lambda_0}}^{(2)}} \right)^2 - 2 + \frac{\|\beta_{\hat{S}_{\lambda_0}}^{(2)} - \beta_{\hat{S}_{\lambda_0}}^{(1)}\|_{\Sigma^{(2)}}^2}{(\sigma_{\hat{S}_{\lambda_0}}^{(1)})^2} + \frac{\|\beta_{\hat{S}_{\lambda_0}}^{(2)} - \beta_{\hat{S}_{\lambda_0}}^{(1)}\|_{\Sigma^{(1)}}^2}{(\sigma_{\hat{S}_{\lambda_0}}^{(2)})^2}.$$

The analysis is divided into two cases depending on the discrepancy between the conditional variances $\sigma^{(1)}$ and $\sigma^{(2)}$.

Case 1: $\frac{\sigma^{(1)} \vee \sigma^{(2)}}{\sigma^{(1)} \wedge \sigma^{(2)}} \geq \sqrt{2}$. By symmetry, we can assume that $\sigma^{(1)} > \sigma^{(2)}$. Since $(\sigma_S^{(i)})^2 = (\sigma^{(i)})^2 + \|\beta_S^{(i)} - \beta_S^{(i)}\|_{\Sigma^{(i)}}^2$, we have

$$\begin{aligned}
(\sigma_{\hat{S}_{\lambda_0}}^{(1)})^2 &= (\sigma^{(1)})^2 + \|\beta_{\hat{S}_{\lambda_0}}^{(1)} - \beta_{\hat{S}_{\lambda_0}}^{(1)}\|_{\Sigma^{(1)}}^2 \geq (\sigma^{(1)})^2, \\
(\sigma_{\hat{S}_{\lambda_0}}^{(2)})^2 &= (\sigma^{(2)})^2 + \|\beta_{\hat{S}_{\lambda_0}}^{(2)} - \beta_{\hat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(2)}}^2 \leq (\sigma^{(2)})^2 + \|\beta^{(2)} - \hat{\beta}_{\lambda_0}^{(2)}\|_{\Sigma^{(2)}}^2 \\
&\leq (\sigma^{(2)})^2 + L \frac{|S_{\vee}|(n_1 \vee n_2) \log(p) \bigvee_{i=1,2} \Phi_{1,+}(\sqrt{\Sigma^{(i)}})}{(n_1 \wedge n_2)^2 \bigwedge_{i=1,2} \kappa^2[6, |\theta_*|_0, \sqrt{\Sigma^{(i)}}]} (\sigma^{(1)})^2 \quad (73) \\
&\leq (\sigma^{(2)})^2 + \frac{(\sigma^{(1)})^2}{4},
\end{aligned}$$

where we used conditions (36) and (38) in the last inequality assuming that we have taken L^* and L_2^* small enough in these two conditions. The above inequalities enforce

$$2 \left[\mathcal{K}_1(\hat{S}_{\lambda_0}) + \mathcal{K}_2(\hat{S}_{\lambda_0}) \right] \geq \left(\frac{\sigma_{\hat{S}_{\lambda_0}}^{(1)}}{\sigma_{\hat{S}_{\lambda_0}}^{(2)}} \right)^2 + \left(\frac{\sigma_{\hat{S}_{\lambda_0}}^{(1)}}{\sigma_{\hat{S}_{\lambda_0}}^{(2)}} \right)^2 - 2 \geq \frac{1}{12}.$$

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

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

with L as in (73). Using Conditions (36) and (38), we can assume that $A \leq 1$. Arguing as in Case 1, we derive that

$$(\sigma^{(1)})^2 \leq (\sigma_{\hat{S}_{\lambda_0}}^{(1)})^2 \leq (\sigma^{(1)})^2 [1 + A] \leq 2(\sigma^{(1)})^2,$$

$$(\sigma^{(2)})^2 \leq (\sigma_{\widehat{S}_{\lambda_0}}^{(2)})^2 \leq (\sigma^{(2)})^2 [1 + A] \leq 2(\sigma^{(2)})^2.$$

Let us lower bound $\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0})$ in terms of $\mathcal{K}_1 + \mathcal{K}_2$. First, we consider the ratio of variances.

$$\begin{aligned} \frac{(\sigma_{\widehat{S}_{\lambda_0}}^{(1)})^2}{(\sigma_{\widehat{S}_{\lambda_0}}^{(2)})^2} + \frac{(\sigma_{\widehat{S}_{\lambda_0}}^{(2)})^2}{(\sigma_{\widehat{S}_{\lambda_0}}^{(1)})^2} - 2 &\geq \left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} \right] / (1 + A) - 2 \\ &\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} - 2 - \frac{A}{1 + A} \left[\frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} \right] \\ &\geq \frac{(\sigma^{(1)})^2}{(\sigma^{(2)})^2} + \frac{(\sigma^{(2)})^2}{(\sigma^{(1)})^2} - 2 - 3A. \end{aligned} \quad (74)$$

Let us now lower bound the remaining part of $\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0})$. For $i = 1, 2$, $|\beta^{(i)} - \widehat{\beta}_{\lambda_0}^{(i)}|_0 \leq |\theta_*|_0 + |\widehat{\theta}_{\lambda_0}|_0 \leq k_*$ by Lemma 8.7 and Condition (38). Thus, we obtain

$$\begin{aligned} &\frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} + \frac{\|\beta^{(1)} - \beta^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} \\ &\leq \frac{3}{(\sigma^{(1)})^2 \wedge (\sigma^{(2)})^2} \\ &\quad \times \sum_{i=1}^2 \left[\|\beta^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(1)}\|_{\Sigma^{(i)}}^2 + \|\beta^{(2)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(i)}}^2 + \|\beta_{\widehat{S}_{\lambda_0}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(i)}}^2 \right] \\ &\leq L_1 \left[\frac{\|\beta_{\widehat{S}_{\lambda_0}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} + \frac{\|\beta_{\widehat{S}_{\lambda_0}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} \right] \\ &\quad + \frac{L_2}{(\sigma^{(1)} \wedge \sigma^{(2)})^2} \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} \left[\sum_{i=1}^2 \|\beta^{(i)} - \widehat{\beta}_{\lambda_0}^{(i)}\|_{\Sigma^{(i)}}^2 \right] \\ &\leq L_1 \left[\frac{\|\beta_{\widehat{S}_{\lambda_0}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(1)}}^2}{(\sigma^{(2)})^2} + \frac{\|\beta_{\widehat{S}_{\lambda_0}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}}^{(2)}\|_{\Sigma^{(2)}}^2}{(\sigma^{(1)})^2} \right] + L_2 \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} A, \end{aligned}$$

where we used $\sigma^{(1)} \asymp \sigma^{(2)}$ in the second line and the bound (72) in the last line. Gathering the last inequality with (74) yields

$$\mathcal{K}_1(\widehat{S}_{\lambda_0}) + \mathcal{K}_2(\widehat{S}_{\lambda_0}) \geq L_1 [\mathcal{K}_1 + \mathcal{K}_2] - L_2 \frac{\bigvee_{i=1,2} \Phi_{k_*,+}(\sqrt{\Sigma^{(i)}})}{\bigwedge_{i=1,2} \Phi_{k_*, -}(\sqrt{\Sigma^{(i)}})} A. \quad \square$$

Proof of Lemma 8.9. For any non empty set S of size smaller or equal to k_* , define $\delta_S = \delta(2 \binom{|S|}{p} k_*)^{-1}$. If we take L^* and L_1^* in (36–37) 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 4.3. Arguing as in the proof of this Theorem, we derive that

$$\mathbb{P} \left[\min_{i \in \{V, 1, 2\}} \tilde{q}_{i,S} < \alpha_S \right] \geq 1 - \delta_S$$

if

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

Let us call $\mathcal{S}_{\mathcal{K}}$ the collection of subsets S of size smaller or equal to k_* that satisfy the above inequality. Applying an union bound over all sets S in $\mathcal{S}_{\mathcal{K}}$, we obtain

$$\mathbb{P} \left[\max_{S \in \mathcal{S}_{\mathcal{K}}} \left(\min_{i \in \{V, 1, 2\}} \tilde{q}_{i,S} - \alpha_S \right) > 0 \right] \leq \delta/2.$$

As we assume that $\{\widehat{S}_{\lambda_0} \in \mathcal{S}_{\mathcal{K}}\} \subset \mathcal{A} \cap \mathcal{B}$, we conclude that

$$\mathbb{P} \left[\left\{ \min_{i \in \{V, 1, 2\}} \tilde{q}_{i, \widehat{S}_{\lambda_0}} > \alpha_{\widehat{S}_{\lambda_0}} \right\} \cap \mathcal{A} \cap \mathcal{B} \right] \leq \delta/2. \quad \square$$

8.8. Proof of Proposition 4.6

We follow the same approach as for the previous proof. 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 $\widehat{\theta}_\lambda = \begin{pmatrix} \widehat{\theta}_\lambda^{(1)} \\ \widehat{\theta}_\lambda^{(2)} \end{pmatrix}$, we derive that $\widehat{\theta}_\lambda^{(2)}$ is the solution of the following minimization problem:

$$\arg \min_{\theta \in \mathbb{R}^p} \|\epsilon + \mathbf{W}^{(2)}\theta_*^{(2)} + \mathbf{W}^{(1)}(\theta_*^{(1)} - \widehat{\theta}_\lambda^{(1)}) - \mathbf{W}^{(2)}\theta\| + \lambda|\theta'|_1. \quad (75)$$

We fix

$$\lambda_0 = 16(\sigma^{(1)} \vee \sigma^{(2)}) \sqrt{2(n_1 + n_2)\Phi_{1,+}(\sqrt{\Sigma}) \log(p)}.$$

and we suppose that event $\mathcal{A} \cap \mathcal{B}$ (defined in the proof of Proposition 4.5) holds. Recall that $\mathbb{P}[\mathcal{A} \cap \mathcal{B}] \geq 1 - \delta/4 - 1/p$. Consider the set $\widehat{S}_{\lambda_0}^{(2)}$ defined as the support of $\widehat{\theta}_{\lambda_0}^{(2)}$. Arguing as in the proof of Proposition 4.5, we have $\{\widehat{S}_{\lambda_0}^{(2)} \subset \widehat{S}_{\text{Lasso}}\} \subset (\mathcal{A} \cap \mathcal{B})$ and it suffices to prove that

$$\mathbb{P} \left[\left\{ \min_{i=V, 1, 2} \tilde{q}_{i, \widehat{S}_{\lambda_0}^{(2)}} > \alpha_{\widehat{S}_{\lambda_0}^{(2)}} \right\} \cap \mathcal{A} \cap \mathcal{B} \right] \leq \delta/2. \quad (76)$$

Lemma 8.10. *If we take constants \tilde{L}^* and L_2^* in Proposition 4.6 small enough, then the following holds. There exists an event \mathcal{C} of probability larger than $1 - 1/p$*

such that, under $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$, we have

$$|\mathbf{W}^{(2)\top} \mathbf{W}^{(1)} (\theta_*^{(1)} - \widehat{\theta}_{\lambda_0}^{(1)})|_\infty \leq \lambda_0/8. \quad (77)$$

It follows from the above lemma that on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$

$$\left| \mathbf{W}^{(2)\top} \left[\boldsymbol{\epsilon} + \mathbf{W}^{(1)} (\theta_*^{(1)} - \widehat{\theta}_{\lambda_0}^{(1)}) \right] \right|_\infty \leq \lambda_0/4.$$

Since $\widehat{\theta}_{\lambda_0}^{(2)}$ is the solution of a Lasso minimization problem with “noise” $\boldsymbol{\epsilon} + \mathbf{W}^{(1)}(\theta_*^{(1)} - \widehat{\theta}_{\lambda_0}^{(1)})$ (see (75)), we can argue as in the proof of Lemma 8.7. Taking L_2^* small enough, we derive that on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$,

$$\|\mathbf{W}^{(2)}(\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)})\|^2 \leq L_1 \frac{\lambda_0^2/(n_1 \wedge n_2)}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} |\theta_*^{(2)}|_0, \quad (78)$$

$$|\widehat{\theta}_{\lambda_0}^{(2)}|_0 \leq L_2 \frac{\Phi_{k_*, +}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} |\theta_*^{(2)}|_0 \leq \tilde{k}_*/2 \leq k_*/2. \quad (79)$$

This allows us to upper bound $\|\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)}\|_\Sigma^2$ on event $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$. By definition of \mathcal{A} ,

$$\begin{aligned} \|\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)}\|_\Sigma^2 &\leq \frac{L}{n_1 \wedge n_2} \left[\|\mathbf{X}^{(1)}(\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)})\|^2 + \|\mathbf{X}^{(2)}(\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)})\|^2 \right] \\ &\leq \frac{L}{n_1 \wedge n_2} \|\mathbf{W}^{(2)}(\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)})\|^2. \end{aligned}$$

Pythagorean inequality then gives

$$\begin{aligned} \|\beta^{(1)} - \beta^{(2)}\|_\Sigma^2 &= \|\beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(2)}\|_\Sigma^2 + \|\beta^{(1)} - \beta^{(2)} - \beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(1)} + \beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(2)}\|_\Sigma^2 \\ &\leq \|\beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(2)}\|_\Sigma^2 + \|\theta_*^{(2)} - \widehat{\theta}_{\lambda_0}^{(2)}\|_\Sigma^2 \\ &\leq \|\beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(1)} - \beta_{\widehat{S}_{\lambda_0}^{(2)}}^{(2)}\|_\Sigma^2 + L \frac{|\theta_*^{(2)}|_0 \log(p)}{n_1 \wedge n_2} \frac{\Phi_{1,+}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]} (\sigma^{(1)} \vee \sigma^{(2)})^2, \end{aligned}$$

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

$$\mathcal{K}_1(\widehat{S}_{\lambda_0}^{(2)}) + \mathcal{K}_2(\widehat{S}_{\lambda_0}^{(2)}) \geq L \frac{\|\beta^{(1)} - \beta^{(2)}\|_\Sigma^2}{\text{Var}(Y^{(1)}) \vee \text{Var}(Y^{(2)})} - L' \frac{|\theta_*^{(2)}|_0 \log(p)}{n_1 \wedge n_2} \frac{\Phi_{1,+}(\sqrt{\Sigma})}{\kappa^2[6, \tilde{k}_*, \sqrt{\Sigma}]}.$$

If we take the numerical constant L_3^* large enough in Proposition 4.6, it then follows from the above lower bound that Condition (68) in Lemma 8.9 is satisfied by $\widehat{S}_{\lambda_0}^{(2)}$. Thus, we conclude from Lemma 8.9 that $\mathbb{P}[\{\min_{i=V,1,2} \tilde{q}_{i, \widehat{S}_{\lambda_0}^{(2)}} > \alpha_{\widehat{S}_{\lambda_0}^{(2)}}\} \cap \mathcal{A} \cap \mathcal{B} \cap \mathcal{C}] \leq \delta/3$. (In comparison to the original statement of Lemma 8.9, $\delta/2$ is replaced by $\delta/3$, but this change only impacts the universal constants in the statement of Proposition 4.6) Since $\mathbb{P}[\mathcal{C}^c] \leq 1/p \leq \delta/6$, we have shown (76). The proof is finished.

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 $|\theta_*|_0 + |\hat{\theta}_{\lambda_0}|_0 \leq \tilde{k}_*$ and

$$\|\mathbf{W}(\hat{\theta}_{\lambda_0} - \theta_*)\|^2 \leq L_1 \frac{\lambda_0^2}{\kappa^2[6, |\theta_*|_0, \sqrt{\Sigma}](n_1 \wedge n_2)} \tilde{k}_*. \quad (80)$$

Thus, $|\theta_*^{(1)} - \hat{\theta}_{\lambda_0}^{(1)}|_0 \leq \tilde{k}_*$ and we derive

$$\begin{aligned} \left| \mathbf{W}^{(2)\top} \mathbf{W}^{(1)} \left(\theta_*^{(1)} - \hat{\theta}_{\lambda_0}^{(1)} \right) \right|_\infty &= \left| \left(\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)} \right) \left(\theta_*^{(1)} - \hat{\theta}_{\lambda_0}^{(1)} \right) \right|_\infty \\ &\leq \|\theta_*^{(1)} - \hat{\theta}_{\lambda_0}^{(1)}\| \sqrt{\tilde{k}_*} \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \\ &\leq \frac{\|\mathbf{W}(\theta_* - \hat{\theta})\|}{\sqrt{\Phi_{k_*, -}(\mathbf{W})}} \sqrt{\tilde{k}_*} \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \\ &\leq L \frac{\lambda_0 \tilde{k}_* \|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty}{\sqrt{n_1 \wedge n_2} \kappa[6, |\theta_*|_0, \sqrt{\Sigma}] \sqrt{\Phi_{k_*, -}(\mathbf{W})}}, \end{aligned} \quad (81)$$

where we used (80) in the last line.

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

$$\|\mathbf{X}^{(1)\top} \mathbf{X}^{(1)} - \mathbf{X}^{(2)\top} \mathbf{X}^{(2)}\|_\infty \leq \Phi_{1,+}(\sqrt{\Sigma}) \left[|n_1 - n_2| + L \sqrt{(n_1 \vee n_2) \log(p)} \right], \quad (82)$$

defining event \mathcal{C} , holds with probability larger than $1 - 1/p$. Consider some θ with $|\theta|_0 \leq k_*$. When the 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 (81) and (82), we get

$$\left| \mathbf{W}^{(2)\top} \mathbf{W}^{(1)} \left(\theta_*^{(1)} - \hat{\theta}_{\lambda_0}^{(1)} \right) \right|_\infty \leq L \lambda_0 \tilde{k}_* \left[\frac{|n_1 - n_2|}{n_1 \wedge n_2} + \sqrt{\frac{\log(p)}{n_1 \wedge n_2}} \right] T_\Sigma,$$

since $n_1 \vee n_2 \leq 2(n_1 \wedge n_2)$. Taking \tilde{L}^* small enough in definition (39) of \tilde{k}_* allows us to conclude. \square

8.9. Proof of Proposition 4.2

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 [49], 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

$$\frac{r^2}{1 - r^2} \leq \gamma \frac{|s|}{n_1} \log(p). \quad (83)$$

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.

8.10. Technical lemmas

In this section, some useful deviation inequalities for χ^2 random variables [27] and for Wishart matrices [15] 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, two anonymous reviewers, the Associate Editor and the Editor for careful suggestions. We also thank Nicolas Städler for providing us the code of DiffRegr. The research of N. Verzelen is partly supported by the French Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration). The research of C. Charbonnier is supported by the CNR-MAJ.

Supplementary Material

Supplement to “A global homogeneity test for high-dimensional linear regression”

(doi: [10.1214/15-EJS999SUPP](https://doi.org/10.1214/15-EJS999SUPP); .zip).

References

- [1] AMBROISE, C., CHIQUET, J., AND MATIAS, C. (2009). Inferring sparse Gaussian graphical models with latent structure. *Electron. J. Stat.* **3**, 205–238. <http://dx.doi.org/10.1214/08-EJS314>. MR2495837 (2010f:62009)
- [2] ARIAS-CASTRO, E., CANDÈS, E. J., AND PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39**, 5, 2533–2556. <http://dx.doi.org/10.1214/11-AOS910>. MR2906877 (2012k:62110)
- [3] BAI, Z. AND SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6**, 2, 311–329. MR1399305 (97i:62062)
- [4] BARAUD, Y., GIRAUD, C., AND HUET, S. (2014). Estimator selection in the Gaussian setting. *Ann. Inst. Henri Poincaré Probab. Stat.* **50**, 3, 1092–1119. <http://dx.doi.org/10.1214/13-AIHP539>. MR3224300
- [5] BARAUD, Y., HUET, S., AND LAURENT, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31**, 1, 225–251. <http://dx.doi.org/10.1214/aos/1046294463>. MR1962505 (2004a:62091)
- [6] BICKEL, P. J., RITOV, Y., AND TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 4, 1705–1732. <http://dx.doi.org/10.1214/08-AOS620>. MR2533469 (2010j:62118)
- [7] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 4, 1212–1242. <http://dx.doi.org/10.3150/12-BEJSP11>. MR3102549
- [8] CAI, T., LIU, W., AND XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108**, 501, 265–277. <http://dx.doi.org/10.1080/01621459.2012.758041>. MR3174618
- [9] CANDÈS, E. J. AND PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37**, 5A, 2145–2177. <http://dx.doi.org/10.1214/08-AOS653>. MR2543688 (2010j:62017)
- [10] CHARBONNIER, C., VERZELEN, N., AND VILLERS, F. (2015). Supplement to “A global homogeneity test for high-dimensional linear regression.” <http://dx.doi.org/10.1214/15-EJS999SUPP>.
- [11] CHEN, S. X. AND QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 2, 808–835. <http://dx.doi.org/10.1214/09-AOS716>. MR2604697 (2011c:62187)
- [12] CHIQUET, J., GRANDVALET, Y., AND AMBROISE, C. (2011). Inferring multiple graphical structures. *Stat. Comput.* **21**, 4, 537–553. <http://dx.doi.org/10.1007/s11222-010-9191-2>. MR2826691 (2012h:62017)

- [13] CHU, J., LAZARUS, R., CAREY, V., AND RABY, B. (2011). A statistical framework for differential network analysis from microarray data. *BMC Systems Biology* **5**.
- [14] 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**, 1629–1640.
- [15] 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 \(2004f:47002a\)](#)
- [16] DONOHO, D. AND JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 3, 962–994. <http://dx.doi.org/10.1214/009053604000000265>. [MR2065195 \(2005e:62066\)](#)
- [17] EFRON, B., HASTIE, T., JOHNSTONE, I., AND TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32**, 2, 407–499. With discussion, and a rejoinder by the authors. [MR2060166 \(2005d:62116\)](#)
- [18] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* **9**, 3, 432–441.
- [19] GILL, R., DATTA, S., AND DATTA, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics* **11**.
- [20] GIRAUD, C., HUET, S., AND VERZELEN, N. (2012a). Graph selection with GGMselect. *Stat. Appl. Genet. Mol. Biol.* **11**, 3, Art. 3, 52. <http://dx.doi.org/10.1515/1544-6115.1625>. [MR2934907](#)
- [21] GIRAUD, C., HUET, S., AND VERZELEN, N. (2012b). Supplement to ‘High-dimensional regression with unknown variance’.
- [22] 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**.
- [23] 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.
- [24] INGSTER, Y. I., TSYBAKOV, A. B., AND VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4**, 1476–1526. <http://dx.doi.org/10.1214/10-EJS589>. [MR2747131 \(2012b:62231\)](#)
- [25] JAVANMARD, A. AND MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909. [MR3277152](#)

- [26] 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.
- [27] LAURENT, B. AND MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* **28**, 5, 1302–1338. [MR1805785 \(2002c:62052\)](#)
- [28] LAURITZEN, S. L. (1996). *Graphical models*. Oxford Statistical Science Series, Vol. **17**. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications. [MR1419991 \(98g:62001\)](#)
- [29] 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](#)
- [30] 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**, 27039–27047.
- [31] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J., AND TIBSHIRANI, R. (2014). Correction: Rejoinder to “A significance test for the lasso” [[MR3210977](#)]. *Ann. Statist.* **42**, 5, 2138–2139. <http://dx.doi.org/10.1214/14-AOS1261>. [MR3269975](#)
- [32] LOPES, M., JACOB, L., AND WAINWRIGHT, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, 1206–1214.
- [33] MEINSHAUSEN, N. (2013). Assumption-free confidence intervals for groups of variables in sparse high-dimensional regression. Available online at <http://arxiv.org/abs/1309.3489>.
- [34] MEINSHAUSEN, N. AND BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 3, 1436–1462. [MR2278363](#)
- [35] MEINSHAUSEN, N., MEIER, L., AND BÜHLMANN, P. (2009). p -values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104**, 488, 1671–1681. <http://dx.doi.org/10.1198/jasa.2009.tm08647>. [MR2750584](#)
- [36] MEINSHAUSEN, N. AND YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37**, 1, 246–270. <http://dx.doi.org/10.1214/07-AOS582>. [MR2488351 \(2010e:62176\)](#)
- [37] 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**.
- [38] RASKUTTI, G., WAINWRIGHT, M., AND YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research* **11**, 2241–2259. [MR2719855 \(2011h:62272\)](#)
- [39] 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**.
- [40] 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.
- [41] SHOJAIE, A. AND MICHAELIDIS, G. (2010). Network enrichment analysis in complex experiments. *Stat. Appl. Genet. Mol. Biol.* **9**, Art. 22, 36. <http://dx.doi.org/10.2202/1544-6115.1483>. MR2721702
- [42] SRIVASTAVA, M. S. AND DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99**, 3, 386–402. <http://dx.doi.org/10.1016/j.jmva.2006.11.002>. MR2396970 (2009e:62261)
- [43] STÄDLER, N. AND MUKHERJEE, S. (2012). Two-sample testing in high-dimensional models. Available online at <http://arxiv.org/abs/1210.4584v2>.
- [44] STÄDLER, N. AND MUKHERJEE, S. (2015). Multivariate gene-set testing based on graphical models. *Biostatistics* **16**, 1, 47–59.
- [45] SUN, T. AND ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 4, 879–898. <http://dx.doi.org/10.1093/biomet/ass043>. MR2999166
- [46] VAN DE GEER, S. A. AND BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3**, 1360–1392. <http://dx.doi.org/10.1214/09-EJS506>. MR2576316 (2011c:62231)
- [47] VERZELEN, N. (2012). Minimax risks for sparse regressions: ultra-high dimensional phenomena. *Electron. J. Stat.* **6**, 38–90. <http://dx.doi.org/10.1214/12-EJS666>. MR2879672
- [48] VERZELEN, N. AND VILLERS, F. (2009). Tests for Gaussian graphical models. *Comput. Statist. Data Anal.* **53**, 5, 1894–1905. <http://dx.doi.org/10.1016/j.csda.2008.09.022>. MR2649554
- [49] VERZELEN, N. AND VILLERS, F. (2010). Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.* **38**, 2, 704–752. <http://dx.doi.org/10.1214/08-AOS629>. MR2604699 (2011g:62193)
- [50] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 5, 2183–2202. <http://dx.doi.org/10.1109/TIT.2009.2016018>. MR2729873 (2011f:62084)
- [51] WASSERMAN, L. AND ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 5A, 2178–2201. <http://dx.doi.org/10.1214/08-AOS646>. MR2543689 (2010k:62287)
- [52] ZHANG, C.-H. AND ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76**, 1, 217–242. <http://dx.doi.org/10.1111/rssb.12026>. MR3153940