GAP: A General Framework for Information Pooling in Two-Sample Sparse Inference

Yin Xia[1], T. Tony Cai[2], and Wenguang Sun[3]

Abstract

This paper develops a general framework for exploiting the sparsity information in two-sample multiple testing problems. We propose to first construct a covariate sequence, in addition to the usual primary test statistics, to capture the sparsity structure, and then incorporate the auxiliary covariates in inference via a three-step algorithm consisting of grouping, adjusting and pooling (GAP). The GAP procedure provides a simple and effective framework for information pooling. An important advantage of GAP is its capability of handling various dependence structures such as those arise from high-dimensional linear regression, differential correlation analysis, and differential network analysis. We establish general conditions under which GAP is asymptotically valid for false discovery rate control, and show that these conditions are fulfilled in a range of settings, including testing multivariate normal means, high-dimensional linear regression, differential covariance or correlation matrices, and Gaussian graphical models. Numerical results demonstrate that existing methods can be significantly improved by the proposed framework. The GAP procedure is illustrated using a breast cancer study for identifying gene-gene interactions.

Keywords: adjusted p-value; covariate-assisted inference; dependent tests; false discovery rate; multiple testing with groups; uncorrelated screening.

[1] Department of Statistics, School of Management, Fudan University. The research of Yin Xia was supported in part by NSFC Grants 11771094, 11690013 and "The Recruitment Program of Global Experts" Youth Project.

[2] Department of Statistics, The Wharton School, University of Pennsylvania. The research of Tony Cai was supported in part by NSF grant DMS-1712735 and NIH grants R01-GM129781 and R01-GM123056.

[3] Department of Data Sciences and Operations, University of Southern California. The research of Wenguang Sun was supported in part by NSF grant DMS-CAREER 1255406.

1 Introduction

Comparison of two high-dimensional objects that are measured under different conditions arises in a wide range of scientific fields such as genomics, neuroimaging, astrophysics and network analysis. Examples include identifying differences in the coordinates of two mean vectors, detecting changes in the entries of two correlation/covariance matrices, and comparing the connectivity between two networks. One phenomenon that arises particularly frequently in high dimensional data analysis is *sparsity*: out of a large number of features most of them are noise, and only a few features contain information of interest. This article focuses on large-scale multiple testing in a setting where both high-dimensional objects (vectors, matrices, networks, etc) are individually sparse.

Statistically, these problems can be formulated as follows. For $d = 1, 2$, let $\boldsymbol{Y}_d \sim P_{\boldsymbol{\beta}_d,\boldsymbol{\eta}_d}$ be p-dimensional random vectors, where $\boldsymbol{\beta}_d \in \mathbb{R}^m$ are the parameters of interest that are sparse, and $\boldsymbol{\eta}_d$ are nuisance parameters. Suppose we observe random samples $\{\boldsymbol{Y}_{1,\cdot,d}, \cdots, \boldsymbol{Y}_{n_d,\cdot,d}\}$ as independent copies of $\boldsymbol{Y}_d$, $d = 1, 2$, where $\boldsymbol{Y}_{k,\cdot,d} = \{Y_{k,i,d} : 1 \leq i \leq p\}$ and n_d are the sample sizes. The goal is to test simultaneously the hypotheses

$$H_{0,i} : \beta_{i,1} = \beta_{i,2} \quad \text{vs.} \quad H_{1,i} : \beta_{i,1} \neq \beta_{i,2}, \quad 1 \leq i \leq m. \tag{1.1}$$

We discuss a few examples before presenting the general framework.

Detection of gene-environment interactions. Recent research reveals that many complex diseases result from the interplay between genetic make-up and exposures to environmental risk factors (Hunter, 2005; Caspi and Moffitt, 2006). Identifying gene-environment interactions can improve the understanding of many disease phenotypes, say, how an external environmental factor interacts with internal genetic factors to generate disordered symptoms. When the environmental factor is binary such as smoking or alcohol status, the interaction effects can be captured by a two-sample high-dimensional regression model:

$$\boldsymbol{Y}_d^* = \boldsymbol{\mu}_d + \boldsymbol{X}_d\boldsymbol{\beta}_d + \boldsymbol{\epsilon}_d, \tag{1.2}$$

where $d = 1, 2$ denotes the environmental condition, $\boldsymbol{Y}_d^* = (Y_{1,d}^*, \ldots, Y_{n_d,d}^*)^\intercal$ are measurements of phenotypes, $\boldsymbol{\mu}_d = \mu_d \mathbf{1}^\intercal$ are the intercepts, with $\mathbf{1}^\intercal$ being a vector of ones, $\boldsymbol{\beta}_d = (\beta_{1,d}, \ldots, \beta_{m,d})^\intercal$ are the vectors of regression coefficients, $\boldsymbol{X}_d = (\boldsymbol{X}_{1,\cdot,d}^\intercal, \ldots, \boldsymbol{X}_{n_d,\cdot,d}^\intercal)^\intercal$ are the matrices of measurements of genomic markers, and $\boldsymbol{\epsilon}_d = (\epsilon_{1,d}, \ldots, \epsilon_{n_d,d})^\intercal$ are random errors. Interaction detection can be formulated as the two-sample multiple testing problem (1.1), where $\boldsymbol{Y}_{k,\cdot,d} = \{Y_{k,d}^*, \boldsymbol{X}_{k,\cdot,d}\}^\intercal \in \mathbb{R}^p$, and $\boldsymbol{\beta}_d \in \mathbb{R}^m$ with $m = p - 1$, $d = 1, 2$, are both individually sparse. The nuisance parameters $\boldsymbol{\eta}_d$ include the intercepts $\boldsymbol{\mu}_d$, the variance of $\boldsymbol{\epsilon}_d$ and the distributional parameters of $\boldsymbol{X}_d$.

Identification of sequentially activated genes. In microarray time-course experiments, the identification of genes that exhibit a specific temporal pattern of differential expression (DE) helps gain insights into the mechanisms of the underlying biological processes (Storey et al., 2005; Tai and Speed, 2006; Sun and Wei, 2011). The expression levels at the first time point usually serve as baseline levels and we expect that only a small proportion of genes would exhibit DE from the baseline. Among DE genes in response to treatment/intervention, some may be detected early while some cannot be detected until the change reaches its peak. A sequential perturbation pattern can be revealed by testing varied levels of DE genes between multiple consecutive time points, which can be formulated as a two-sample multiple testing problem (1.1). Here $\boldsymbol{Y}_d \in \mathbb{R}^p$ with $m = p$ are random vectors recording genes' measurements at times $d = 1, 2$, the parameters of primary interests are the mean expression level *after baseline removal*, i.e., $\boldsymbol{\beta}_d = \mathbb{E}(\boldsymbol{Y}_d) \in \mathbb{R}^m$, and the nuisance parameters $\boldsymbol{\eta}_d$ include the covariance matrix of $\boldsymbol{Y}_d$.

Analysis of differential networks. Detecting gene-gene interactions is a crucial step for understanding how groups of genes act together in different biological processes. A gene association network is a set of genes connected by edges representing their functional relationships. Recent research showed that it is of great importance to study how the association network structures change between two or more biological settings (Gill et al., 2010). To identify the set of genes whose connectivities have changed between two networks, a two-sample multiple testing problem (1.1) can be formulated to test the varied strengths of associations between gene pairs in the two networks, where $\boldsymbol{Y}_d \sim N(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) \in \mathbb{R}^p$ with the precision matrices $\boldsymbol{\Omega}_d = \boldsymbol{\Sigma}_d^{-1}$, $\boldsymbol{\beta}_d \in \mathbb{R}^m$ is the vectorized upper (or lower) off-diagonal elements of $\boldsymbol{\Omega}_d$ with $m = p(p-1)/2$, and the nuisance parameters $\boldsymbol{\eta}_d$ include the means $\boldsymbol{\mu}_d$. Both objects (networks) being tested tend to be very sparse. In Section 6 we illustrate the proposed method by analyzing a breast cancer study to identify gene-gene interactions.

1.1 Multiple comparisons with two sparse objects

In two-sample multiple testing, we are interested in making inference for $\theta_i = \mathbb{I}(\beta_{i,1} \neq \beta_{i,2})$, $1 \leq i \leq m$, where $\mathbb{I}(\cdot)$ is an indicator function. The conventional approach would begin with summarizing the data into a single vector of test statistics $\{T_1, \cdots, T_m\}$ for comparing the coordinates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and then choose a significance threshold to control the multiplicity. This approach ignores the important feature of the two-sample inference problem that both objects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are individually sparse. As a result, it suffers from substantial information loss. This can be intuitvely seen as follows. Let $\mathcal{I}_d = \{1 \leq i \leq m : \beta_{i,d} \neq 0\}$ denote the support of $\boldsymbol{\beta}_d$, $d = 1, 2$, and $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$ the union support. Note that the cardinality of $\mathcal{I}$ is small if both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are sparse; hence information about $\mathcal{I}$ can be potentially utilized to narrow down the focus in multiple testing via the following logical relationship

$$i \notin \mathcal{I} \text{ implies that } \theta_i = 0. \tag{1.3}$$

In the setting of testing sparse normal mean vectors, Cai et al. (2019) demonstrated in a recent paper that conventional practice leads to inefficient procedures. It is shown that an auxiliary covariate sequence can be constructed from the data to provide supplementary information and a data-driven procedure, which employs a covariate-assisted ranking and screening (CARS) approach, achieves substantial power gain over existing methods. However, CARS cannot be applied to dependent tests.

The goal of the present paper is to develop a new framework for two-sample multiple testing with auxiliary information. An important advantage of the proposed framework is its capability of handling a wide range of dependence structures such as those arise from high-dimensional linear regression, differential correlation analysis, and differential network analysis.

Our idea is to construct a covariate sequence $\{S_i : 1 \leq i \leq m\}$, in addition to the primary test statistics $\{T_i : 1 \leq i \leq m\}$, to capture the sparsity information, and then incorporate the information in the testing procedure to improve the efficiency. In contrast with conventional practice which only uses T_i to assess the significance of the difference, we aim to develop new methodologies that utilize m pairs of statistics $\{(T_i, S_i) : 1 \leq i \leq m\}$.

Denote a multiple testing procedure by a binary rule $\boldsymbol{\delta} = \{\delta_i : 1 \leq i \leq m\} \in \{0,1\}^m$, where $\delta_i = 1$ if we reject $H_{0,i}$ and $\delta_i = 0$ otherwise. In large-scale testing, the false discovery rate (FDR, Benjamini and Hochberg, 1995) has been widely used as a practical and powerful error criterion. For a given decision rule $\boldsymbol{\delta}$, the FDR is defined as

$$\mathrm{FDR}_{\boldsymbol{\delta}} = \mathbb{E}\left[\frac{\sum_{i=1}^m (1-\theta_i)\delta_i}{(\sum_{i=1}^m \delta_i) \vee 1}\right], \tag{1.4}$$

where $x \vee y = \max(x, y)$. To evaluate the efficiency of a testing procedure, we define the power of decision rule $\boldsymbol{\delta}$ as the expected proportion of correctly rejected non-null hypotheses,

$$\Psi_{\boldsymbol{\delta}} = \mathbb{E}\left[\frac{\sum_{i=1}^m \theta_i\delta_i}{\sum_{i=1}^m \theta_i}\right]. \tag{1.5}$$

The next section discusses a general framework for FDR control with pairs of observations.

1.2 GAP: An Integrative Framework for Two-sample Sparse Inference

There are two key issues in the methodological development: one is to construct the pair of test statistics (T_i, S_i) to capture the sample information accurately, and another is to integrate the information in T_i and S_i effectively.

To illustrate the proposed testing framework, we first discuss a simple example and then describe how the idea may be generalized to more complicated settings. Let $\boldsymbol{Y}_d \sim N(\boldsymbol{\beta}_d, \boldsymbol{I})$, where $\boldsymbol{\beta}_d = \mathbb{E}(\boldsymbol{Y}_d) = (\beta_{1,d}, \cdots, \beta_{m,d})$ are the population mean vectors, and $\boldsymbol{I}$ is an identity matrix, $d = 1, 2$. Denote $\{\boldsymbol{Y}_{1,\cdot,d}, \cdots, \boldsymbol{Y}_{n,\cdot,d}\}$ independent copies of $\boldsymbol{Y}_d$, $d = 1, 2$, where $\boldsymbol{Y}_{k,\cdot,d} = \{Y_{k,i,d} : 1 \leq i \leq m\}$, and n is the sample size. The population means are estimated as $\bar{Y}_{i,d} = n^{-1}\sum_{k=1}^{n} Y_{k,i,d}$, $1 \le i \le m$. To identify differential levels between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, our proposed framework suggests using the usual two-sample z statistic $T_i = \sqrt{\frac{n}{2}}(\bar{Y}_{i,1} - \bar{Y}_{i,2})$ as the primary statistic to assess the difference, and $S_i = \sqrt{\frac{n}{2}}(\bar{Y}_{i,1} + \bar{Y}_{i,2})$ as the auxiliary statistic to capture the information on $\mathcal{I}$ (since a large $|S_i|$ provides strong evidence that $i \in \mathcal{I}$). By construction, T_i and S_i are independent.

Consider a multiple testing problem with pairs of statistics $\{(T_i, S_i) : 1 \le i \le m\}$. The main idea is to exploit the information in S_i to construct more efficient procedures. Our proposed algorithm, detailed in Section 2.2, operates in three steps: grouping, adjusting and pooling (GAP). According to the logical relationship (1.3), the hypotheses become "unequal" in light of S_i. To reflect this heterogeneity, it is desirable to treat those more likely to be on the union support differently from the rest. The first *grouping* step divides all testing units into K groups based on S_i; this leads to heterogeneous groups with varied sparsity levels. The second *adjusting* step adjusts the p-values to incorporate the structural information revealed by grouping. The final *pooling* step combines the adjusted p-values from all groups and chooses a threshold to control the global FDR at the desired level.

The GAP algorithm provides a simple and effective framework for exploiting the auxiliary information in the covariate sequence. We establish in Section 3 the general conditions under which GAP is valid for FDR control, and show in Section 4 that these conditions are fulfilled by various dependency structures. Our numerical results demonstrate that the performance of existing methods can be greatly improved by GAP.

1.3 Our Contributions

Multiple testing under dependency is an important problem that has been extensively studied in the literature (Benjamini and Yekutieli, 2001; Sarkar, 2002; Efron, 2007; Sun and Cai, 2009). While recent progress has been made towards utilizing external covariates in multiple testing (Du and Zhang, 2014; Liu, 2014; Scott et al., 2015; Cai et al., 2019), most methods do not have a theoretical guarantee for FDR control under dependency. This important issue is addressed by our proposed framework. We show that, under mild conditions that are fulfilled by a class of models, GAP controls the FDR at the nominal level asymptotically.

Liu (2014) proposed the uncorrelated screening (US) method and showed that it controls the FDR and outperforms other methods. US first divides the hypotheses into two groups based on a screening statistic, and then applies the BH procedure to unadjusted p-values in both groups. Compared to US, GAP provides a more general and efficient framework for information pooling. In addition to its capability of handling dependency, the GAP procedure allows for more than two groups and hence captures the structural information more accurately. Moreover, GAP utilizes adjusted p-values so that the heterogeneity between groups can be exploited more efficiently. When pooling the testing results, the group-wise FDR levels are adaptively weighted among groups by GAP; the adaptivity leads to valid FDR control with much improved power. In contrast, US uses the same FDR across all groups. The simulation in Section 5 demonstrates that the efficiency gain of GAP over US via grouping and weighting can be substantial in many settings.

Our work also makes new contributions to multiple testing with groups. First, existing methods for testing with grouping structure [e.g. Efron (2008); Ferkingstad et al. (2008); Cai and Sun (2009); Hu et al. (2012)] have been mostly developed for the independent case that do not have guaranteed FDR control when the tests are dependent. Second, existing methods assume that the groups have been specified *a priori*. In contrast, GAP constructs the covariate sequence from the original data and determines the groups adaptively to maximize the power. Third, a major concern in Efron (2008) and Cai and Sun (2009) is that improper grouping may distort the null distribution of p-values and lead to invalid FDR analyses. This concern has been addressed by GAP, which employs the conditional independence principle to ensure proper grouping and validity in asymptotic FDR control. Finally, GAP utilizes a novel weighting strategy (via *normalizing*), which enables the development of a general theory for FDR control that can handle a wider class of dependency structures compared to existing works on weighted FDR.

1.4 Notation and Definitions

We summarize the notation and definitions that will be used throughout the paper. We follow the convention that v_i stands for the i^{th} entry of a vector v and $M_{i,j}$ for the entry in the i^{th} row and j^{th} column of a matrix M. For a vector $\boldsymbol{\beta}_d = (\beta_{1,d}, \ldots, \beta_{m,d})^\intercal \in \mathbb{R}^m$, define the ℓ_q norm by $|\boldsymbol{\beta}_d|_q = (\sum_{i=1}^m |\beta_{i,d}|^q)^{1/q}$ for $1 \leq q \leq \infty$. For a symmetric matrix $\mathbf{M}$, let $\lambda_{\max}(\mathbf{M})$ and $\lambda_{\min}(\mathbf{M})$ denote the largest and smallest eigenvalues of $\mathbf{M}$, respectively. For a set $\mathcal{H}$, denote $|\mathcal{H}|$ the cardinality of $\mathcal{H}$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ holds for all n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C such that $c \leq a_n/b_n \leq C$ for all n.

1.5 Organization of the Paper

The rest of the paper is organized as follows. Section 2 describes the GAP procedure and develops a general framework for information pooling in two-sample sparse inference. Theoretical properties of GAP are established in Section 3. This general framework is further illustrated in Section 4 under several specific settings. In Section 5, numerical comparisons with competitive methods demonstrate the merits of GAP. In Section 7, we apply the GAP procedure to a breast cancer study for identifying gene-gene interactions. The proofs and additional numerical results are given in the Appendix.

2 GAP: A General Framework for Two-Sample Inference

The GAP procedure for simultaneous comparisons of two high-dimensional sparse objects is based on m pairs of test statistics $\{(T_i, S_i) : 1 \leq i \leq m\}$. In Section 2.1, we discuss a few principles on how to construct the pairs. The GAP algorithm is described in detail in Section 2.2. Section 2.3 explains some key ideas to provide insights on why GAP works.

2.1 Constructing (T_i, S_i): Some Principles

Our proposed GAP procedure requires carefully constructed pairs (T_i, S_i). The roles of T_i and S_i are different: T_i is the primary test statistic to assess the significance of the difference, and S_i, which captures the sparsity information of the union support, is an auxiliary statistic to assist inference. A simple example of the pair is given in the introduction. However, the construction can be complicated in settings such as high-dimensional regression and Gaussian graphical models. We discuss here the important principles; related technical details are deferred until Section 4.

First, we construct (T_i, S_i) as standardized statistics so that they are faithful in reflecting the true data structure and comparable across tests. Second, T_i and S_i need to be asymptotically independent. The independence requirement guarantees that the null distribution of T_i would not be distorted by incorporating S_i in the inference. This is crucial for the validity of the proposed methodology. Specifically, we shall see that, in Steps 2 and 3 of the GAP algorithm, the Benjamini-Hochberg (BH, Benjamini and Hochberg, 1995) procedure is employed to control the FDR. BH assumes that the null distribution of the p-value is uniform. If T_i and S_i are correlated, then the grouping step would distort the null distribution of the p-values, which would lead to an invalid FDR control.

2.2 The GAP Procedure

We now give a precise description of the GAP algorithm and explain its merits as a general framework for information pooling. Let p_i be the p-value associated with T_i for testing $H_{0,i}$ vs. $H_{1,i}$. The GAP procedure consists of three steps: grouping, adjusting and pooling.

Step 1 (Grouping). Divide hypotheses into K groups to reflect the heterogeneity between testing units in light of S_i. Let $\lambda_0 = -\infty$, $-4\sqrt{\log m} \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{K-1} \leq 4\sqrt{\log m}$ and $\lambda_K = \infty$, where $\Lambda = \{\lambda_l : 1 \leq l \leq K-1\}$ is a subset of points from a regular grid $\mathcal{X} = \{(j/N)\sqrt{\log m} : j = -4N, -4N+1, \ldots, -1, 0, 1, \ldots, 4N-1, 4N\}$, with N being a large integer. The corresponding groups are $\mathcal{G}_l = \{1 \leq i \leq m : \lambda_{l-1} < S_i \leq \lambda_l\}$, for $1 \leq l \leq K$. The optimal choice of Λ will be determined in Step 2.

Step 2 (Adjusting). Define $m_l = |\mathcal{G}_l|$. Calculated adjusted p-values $p_i^w = \min\{p_i/w_l^o, 1\}$ if $i \in \mathcal{G}_l$, $1 \le l \le K$, where w_l^o will be calculated as follows.

- *Initial adjusting.* For a given grouping $\{\mathcal{G}_l : 1 \le l \le K\}$, let $\hat{\pi}_l$ be the estimated proportion of non-nulls in $\mathcal{G}_l$. The group-wise weights are computed as

$$w_l = \left\{ \sum_{l=1}^{K} \frac{m_l \hat{\pi}_l}{1 - \hat{\pi}_l} \right\}^{-1} \frac{m \hat{\pi}_l}{(1 - \hat{\pi}_l)}, \; 1 \le l \le K. \tag{2.6}$$

  Define adjusted p-values as $p_i^w = \min\{p_i/w_l, 1\}$ for $i \in \mathcal{G}_l$.

- *Further refining.* We search among all possible $\Lambda \subset \mathcal{X}$ to determine the optimal grouping (in the search we allow Λ to be an empty set, which means that we only have one group). Specifically, for each Λ, combine adjusted p-values from all groups and apply the BH procedure at level α to all adjusted p-values. Specifically, denote $p_{(1)}^w \le \cdots \le p_{(m)}^w$ the ordered adjusted p-values. The threshold is chosen as

$$k = \max\{i : p_{(i)}^w \le i\alpha/m\}. \tag{2.7}$$

  The weights w_l^o are computed using (2.6) based on the optimal grouping that yields the most rejections.

This step up-weights the hypotheses from groups with higher proportions of signals, and down-weight hypotheses from groups with lower proportions.

Step 3 (Pooling). Combine the adjusted p-values from all groups, where p_i^w are computed from Step 2 based on the optimal grouping. To control the FDR at a global level, apply BH (2.7) again to all adjusted p-values $\{p_i^w : 1 \le i \le m\}$.

The following remarks explain GAP in more detail and address some technical points.

Remark 1 There is a tradeoff in the choice of the number of groups K. Ideally, S_i should be modeled as a continuous variable as done in Cai et al. (2019) to maximize the power. However, it is difficult to achieve optimality under dependence. Our grouping step can be viewed as a discrete approximation to the ideal solution. Having more groups is helpful to reduce the approximation bias, whereas the algorithm becomes significantly slower and tends to be less stable with too many groups. In practice, we recommend $K = 3$ or 4.

Remark 2 In Step 2 we need to estimate the non-null proportion for each group. The sparsity estimation problem has been extensively studied in the literature; see Langaas et al. (2005); Meinshausen et al. (2006); Jin and Cai (2007) and Cai and Jin (2010) for recent developments. We use the method by Schweder and Spjøtvoll (1982) and Storey (2002) to estimate the non-null proportions, denoted by $\hat{\pi}_l^*$. The resulting weights are ad hoc but will be justified in the next section. We use $\hat{\pi}_l = (\epsilon \vee \hat{\pi}_l^*) \wedge (1 - \epsilon)$ with $\epsilon = 10^{-5}$ to restrict the estimated proportion in the range $[\epsilon, 1 - \epsilon]$; this would increase the stability of the algorithm. The procedure is robust to the choice of such ϵ. Theoretically, for any ϵ that is larger than m^{-C} for some constant $C > 0$, the asymptotic FDR control in Theorem 1 below will always hold, as shown in the Step 1 from the proof of Theorem 1. With $\epsilon = 10^{-5}$, such C can be any constant that is larger than $5 \log 10 / \log m$.

2.3 Some Insights on Why GAP Works

Before we rigorously establish the theoretical properties of GAP in Section 3, it is helpful to provide some important insights on the merit of the grouping strategy adopted by GAP as well as the weights used in GAP. The discussion here is informal as it is based on existing theory for independent tests. The theoretical results given in Section 3 are for the dependent case.

For multiple testing with known groups, the naive *pooled analysis* ignores the grouping information and applies the BH procedure to all the tests combined. The pooled analysis is inefficient and can even be invalid (Efron, 2008). Another natural approach is the *separate analysis*, which first applies BH to individual groups and then combine all the rejected hypotheses. This strategy is adopted by the US method proposed in Liu (2014). Although the separate analysis is always valid, it is inefficient because a common FDR level is used for all groups. To increase the power, one should adopt a more flexible strategy that allows the FDR levels to vary across groups (Cai and Sun, 2009; Hu et al., 2012).

The proposed GAP procedure adaptively chooses the group-wise FDR levels by utilizing adjusted p-values. This weighting approach in GAP is superior to both pooled and separate analyses as it incorporates group-wise information more effectively. Intuitively, GAP increases the overall power by allocating higher FDR levels to groups where signals are more common. Finally, it is important to emphasize that different from Efron (2008), Cai and Sun (2009) and Hu et al. (2012), GAP does not assume known groups. It constructs its own covariate sequence and searches the optimal grouping to maximize the power.

Genovese et al. (2006) and Basu et al. (2017) consider weighted multiple testing problems and show that multiple testing procedures with proper weights can control the FDR, but the power may be affected by the informativeness of the weights. A key step in our methodology is the standardization of the weights via (2.6), which ensures that after all groups are combined, the weights are always "proper" in the sense of Genovese et al. (2006). It follows that the inaccuracy of the estimates would not affect the validity of GAP for FDR control. Moreover, although proportion estimation is ad hoc, it should in general lead to informative weights; this point is further explained and confirmed by our simulation studies. We will consider the dependent case and show that GAP is valid for FDR control.

3 Theoretical Properties of GAP

In this section, we show that GAP guarantees FDR control asymptotically under regularity conditions on the pairs (T_i, S_i). We verify in Section 4 that these conditions are fulfilled by (T_i, S_i) that are carefully constructed in a range of important problems, including testing multivariate normal means, high-dimensional linear regression, differential covariance or correlation matrices, and Gaussian graphical models.

Denote $\{p_i^w : 1 \le i \le m\}$ the adjusted p-values determined by GAP and $\{p_{(i)}^w : 1 \le i \le m\}$ the ordered adjusted p-values. The false discovery proportion (FDP) of GAP is

$$\mathrm{FDP}_{\mathrm{GAP}} = \frac{\sum_{i \in \mathcal{H}_0} I(p_i^w \le p_{(\hat{k}^w)}^w)}{\sum_{i=1}^m I(p_i^w \le p_{(\hat{k}^w)}^w) \vee 1},$$

where $\hat{k}^w = \max\{i : p^w_{(i)} \leq \alpha i/m\}$. Then $\mathrm{FDR}_{\mathrm{GAP}} = \mathbb{E}\left(\mathrm{FDP}_{\mathrm{GAP}}\right)$. The following technical assumptions on T_i and S_i are needed in our theoretical development. Let A_τ be a subset of $\mathcal{H}_0$ with $|A_\tau| = o(m^\nu)$ for any $\nu > 0$. Define $\tilde{\mathcal{H}}_0 = \mathcal{H}_0 \setminus A_\tau$, and $n = n_1 + n_2$.

(A1) **Asymptotic Normality:** For the primary test statistics $\{T_i, i \in \tilde{\mathcal{H}}_0\}$, there exist two independent sets of i.i.d random variables $\{Z_{k,i}, k = 1, \ldots, n_1\}$ and $\{Z_{k,i}, k = n_1 + 1, \ldots, n\}$ satisfying $\mathbb{E}Z_{k,i} = 0$ and $\mathbb{E}\exp(KZ_{k,i}) < \infty$ for some $K > 0$, such that, for any constant $M > 0$, there exists some b_m satisfying $b_m = o\{(\log m)^{-1/2}\}$ that,

$$\mathbb{P}_{H_{0,i}}\left(\left|T_i - \frac{\sum_{k=1}^n Z_{k,i}}{\mathsf{Var}(\sum_{k=1}^n Z_{k,i})^{1/2}}\right| \geq b_m\right) = O(m^{-M}).$$

(A2) **Weak Dependency:** Define $(\rho_{i,j,1})_{m\times m} = \boldsymbol{R}_1 = \mathsf{Corr}(\boldsymbol{Z}_k)$ for $1 \leq k \leq n_1$ and $(\rho_{i,j,2})_{m\times m} = \boldsymbol{R}_2 = \mathsf{Corr}(\boldsymbol{Z}_k)$ for $n_1 + 1 \leq k \leq n$, where $\boldsymbol{Z}_k = (Z_{k,1}, \ldots, Z_{k,m})$. Then $\max_{1\leq i<j\leq m} |\rho_{i,j,d}| \leq \rho_d < 1$ for some constant $\rho_d > 0$ for $d = 1, 2$. Moreover, there exists $\gamma > 0$ such that $\max_{1\leq i\leq m} |\Gamma_i(\gamma)| \leq C$ for some constant $C > 0$, where $\Gamma_i(\gamma) = \{j : 1 \leq j \leq m, |\rho_{i,j,d}| \geq (\log m)^{-2-\gamma},\ \text{for } d = 1 \text{ or } 2\}$.

(A3) **Asymptotic Independency:** T_i and S_i are asymptotically independent under the null, i.e. for any constant $M > 0$,

$$\mathbb{P}_{H_{0,i}}(|T_i| \geq t, |S_i| \geq \lambda) = (1 + o(1))G(t)\mathbb{P}(|N(0, 1) + s_i| \geq \lambda) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$, $0 \leq \lambda \leq 4\sqrt{\log m}$ and $i \in \tilde{\mathcal{H}}_0$, where $s_i = \mathbb{E}(S_i)$, and for all $0 \leq j \leq 4N$ with fixed N,

$$\mathbb{P}_{H_{0,i}}(|T_i| \geq t, |S_i| < \lambda_j) = (1 + o(1))G(t)\mathbb{P}(|N(0, 1) + s_i| < \lambda_j) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$ and $i \in \tilde{\mathcal{H}}_0$, where $\lambda_j = (j/N)\sqrt{\log m}$.

Remark 3 Assumption (A1) is mild, as it only requires that T_i follows a standard normal distribution asymptotically. The assumption can be easily checked; see Section 4 for more details. Assumption (A2) indicates that not many primary statistics are strongly correlated with each other. Our testing framework is very different from that in conventional two-sample testing problems where one only needs to deal with the correlations between pairs of p-values. By contrast, due to the existence of a sequence of auxiliary statistics, we need to handle a more complicated correlation structure between pairs of (T_i, S_i). Thus the weak dependence condition is slightly stronger, which we speculate could be further relaxed with more sophisticated tools. Assumption (A3) is satisfied by our construction; see Propositions 1 and 3 in Section 4 for proofs.

Define $\mathcal{S}_\rho = \left\{i : 1 \leq i \leq m, |\beta_{i,1} - \beta_{i,2}| \geq \{(\log m)^{1+\rho}/n\}^{1/2}\right\}$. The next theorem shows that GAP controls both the FDP and FDR at the nominal level asymptotically.

Theorem 1 *Suppose for some $\rho > 0$ and some $\delta > 0$, $|\mathcal{S}_\rho| \geq [1/(\pi^{1/2}\alpha) + \delta](\log m)^{3/2}$. Assume that $n_1 \asymp n_2$ and $m_0 = |\mathcal{H}_0| \geq cm$ for some $c > 0$. Then under (A1) - (A3) with $\log m = o(n^{1/C})$ for some $C > 5$, we have*

$$\overline{\lim_{(n,m)\to\infty}} \mathit{FDR}_{\mathrm{GAP}} \leq \alpha, \text{ and } \lim_{(n,m)\to\infty} \mathbb{P}(\mathit{FDP}_{\mathrm{GAP}} \leq \alpha + \epsilon) = 1.$$

for any $\epsilon > 0$.

Remark 4 The condition on $|\mathcal{S}_\rho|$ is mild. It only requires that there are a few coordinates with differential effects exceeding $\{(\log m)^{1+\rho}/n\}^{1/2}$ for some constant $\rho > 0$ among m hypotheses in total. A more precise definition of $\mathcal{S}_\rho$ can be formulated by the standardized difference between $\beta_{i,1}$ and $\beta_{i,2}$, namely, $\mathcal{S}_\rho = \left\{i : 1 \leq i \leq m, \frac{|\beta_{i,1}-\beta_{i,2}|}{(\sigma^2_{w,i,1}+\sigma^2_{w,i,2})^{1/2}} \geq (\log m)^{1/2+\rho}\right\}$, where $\sigma^2_{w,i,1} + \sigma^2_{w,i,2} = \mathsf{Var}(\sum_{k=1}^n Z_{k,i})/n_1^2$ is defined in (A1), and will be discussed in detail in Section 4 under various settings.

We now turn to the power analysis. The next theorem shows that GAP dominates BH in power asymptotically. Our simulation results in Section 5 indicate that the power gain can be substantial in many settings. By applying the definition in (1.5), the powers of GAP and BH procedures can be calculated as follows

$$\Psi_{\mathrm{BH}} = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_1} I(p_i \leq p_{(\hat{k})})}{|\mathcal{H}_1|}\right],$$

where $\hat{k} = \max\{i, p_{(i)} \leq \alpha i/m\}$, and

$$\Psi_{\text{GAP}} = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{H}_1} I(p_i^w \leq p_{(\hat{k}^w)}^w)}{|\mathcal{H}_1|}\right],$$

where $\hat{k}^w = \max\{i, p_{(i)}^w \leq \alpha i/m\}$. The next theorem shows that the GAP procedure is more powerful than the BH procedure asymptotically.

Theorem 2 *Under the same conditions of Theorem 1, we have* $\Psi_{\text{GAP}} \geq \Psi_{\text{BH}} + o(1)$ *as* $m \to \infty$.

The previous theorem shows that GAP is more powerful than BH. To illustrate the power gain in a more explicit manner, we present an example based on theoretical calculations under a simple model; more details are given in Section **??** in the Supplement.

Example 1 Consider a two-point Gaussian mixture model: $\boldsymbol{Y}_d \sim N(\boldsymbol{\beta}_d, I)$, $d = 1, 2$, with $\beta_{i,1} = 0$ for $1 \leq i \leq m$, $\beta_{i,2} = \mu_0$ for $1 \leq i \leq m_1$, and $\beta_{i,2} = 0$ for $m_1 + 1 \leq i \leq m$. The primary and auxiliary statistics are respectively given by $T_i = \frac{1}{\sqrt{2}}(Y_{i,2} - Y_{i,1})$ and $S_i = \frac{1}{\sqrt{2}}(Y_{i,2} + Y_{i,1})$. Let t_{BH}, derived in Section **??**, denote the asymptotic threshold for the BH procedure. The asymptotic p-value threshold of GAP t_{GAP} is difficult to derive but a conservative threshold t^*_{BH}, defined in Section **??**, may be obtained. Specifically, we show that $t_{GAP} \geq t^*_{BH}$; hence t^*_{BH} may be used in place of t_{GAP} to characterize a lower bound on the power difference. The top and bottom rows of Figure 1 illustrate the powers of BH and GAP as functions of μ_0 and $\pi = m_1/m$, respectively. On the top row, we fix $\pi = 0.1$ and vary μ_0. On the bottom row, we fix $\mu_0 = 3.5$ and vary π. The nominal FDR level is 0.1. We can see that GAP with the conservative threshold t^*_{BH} controls the FDR below the nominal level and outperforms BH in power. The power ratios in the third column show that the auxiliary information is more helpful when signals are weak and sparse. We stress that due to the difficulty in obtaining an explicit formula for t_{GAP}, our result only provides a *lower bound*. In practice when t_{GAP} is used, the actual power gain may be even larger.

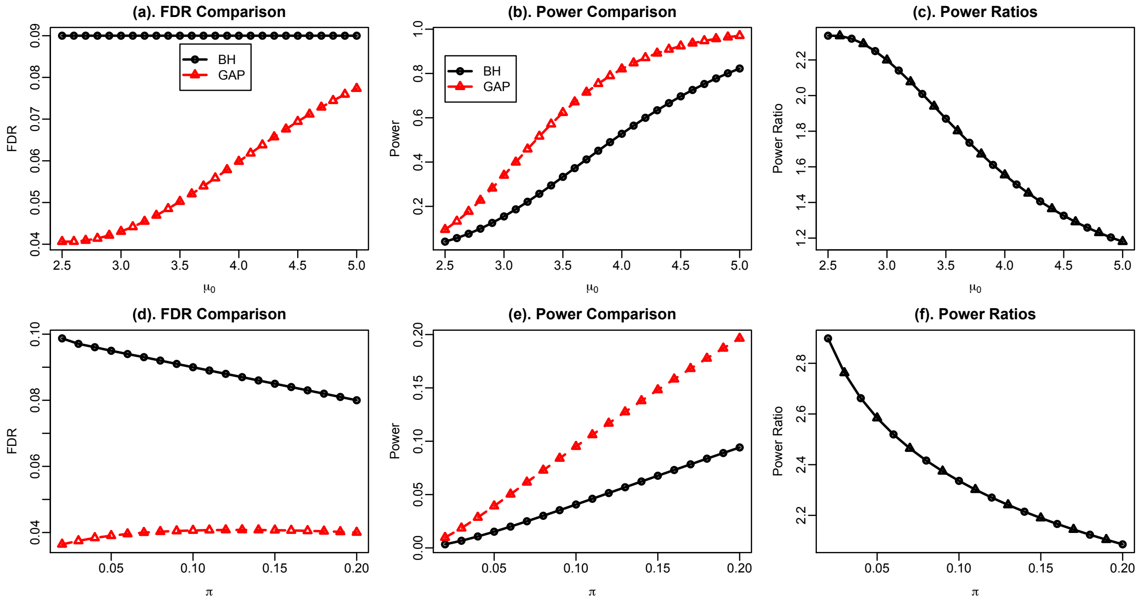

Figure 1: Theoretical calculations for the asymptotic powers of GAP vs. BH. The more conservative threshold t^*_{BH} has been used for calculating the FDR and power of GAP.

4 Construction of Primary and Auxiliary Statistics

The construction of (T_i, S_i) is a key step in our methodological development. We present the construction in detail for testing multivariate normal means and high-dimensional linear regression in Sections 4.1 and 4.2, respectively. The constructions for testing differential covariance or correlation matrices and Gaussian graphical models are similar and are summarized in Section 4.3. We show that the general conditions given in Section 3 are fulfilled by the constructed statistics and hence GAP is valid for FDR control in these settings.

4.1 Multivariate Normal Models

Let $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ be two random vectors recording the measurement levels of the same m features under two conditions, respectively. We assume that $\boldsymbol{Y}_d \sim N(\boldsymbol{\beta}_d, \boldsymbol{\Sigma}_d)$, where $\boldsymbol{\beta}_d = \mathbb{E}(\boldsymbol{Y}_d) = (\beta_{1,d}, \cdots, \beta_{m,d})$ denote the population mean vectors, and $\boldsymbol{\Sigma}_d = (\sigma_{i,j,d} : 1 \leq i, j \leq m)$ the covariance matrices, $d = 1, 2$. Suppose we have collected random samples $\{\boldsymbol{Y}_{1,\cdot,d}, \cdots, \boldsymbol{Y}_{n_d,\cdot,d}\}$ as independent copies of $\boldsymbol{Y}_d$, $d = 1, 2$, where $\boldsymbol{Y}_{k,\cdot,d} = \{Y_{k,i,d} : 1 \leq i \leq m\}$, n_d is the sample size in condition d.

We use T_i to capture the information on the difference. It is natural to start with the sample difference $\bar{\boldsymbol{Y}}_1 - \bar{\boldsymbol{Y}}_2 = (1/n_1)\sum_{k=1}^{n_1} \boldsymbol{Y}_{k,\cdot,1} - (1/n_2)\sum_{k=1}^{n_2} \boldsymbol{Y}_{k,\cdot,2}$. Let $\circ$ denote a Hadamard product and $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_m)$ a vector of weights. To extract information on the union support, we focus on a class of linear combinations of the form $\boldsymbol{\beta}_1 + \boldsymbol{\kappa} \circ \boldsymbol{\beta}_2$, which can be estimated as $\bar{\boldsymbol{Y}}_1 + \boldsymbol{\kappa} \circ \bar{\boldsymbol{Y}}_2$. The weights κ_i should be chosen carefully so that the pair T_i and S_i are asymptotically independent. If the true variances $\sigma_{i,i,d}$ are unknown, the weights can be estimated as $\hat{\kappa}_i = (n_2\hat{\sigma}_{i,1}^2)/(n_1\hat{\sigma}_{i,2}^2)$, where $\hat{\sigma}_{i,d}^2$ are the sample variances $(n_d)^{-1}\sum_{k=1}^{n_d}(Y_{k,i,d} - \bar{Y}_{i,d})^2$, $d = 1, 2$. Finally, T_i and S_i are standardized to ensure the comparability of the tests. Let $\hat{\sigma}_{w,i,d}^2 = \hat{\sigma}_{i,d}^2/n_d$, we propose the following pair of statistics:

$$(T_i, S_i) = \left(\frac{\bar{Y}_{i,1} - \bar{Y}_{i,2}}{(\hat{\sigma}_{w,i,1}^2 + \hat{\sigma}_{w,i,2}^2)^{1/2}}, \frac{\bar{Y}_{i,1} + \hat{\kappa}_i \bar{Y}_{i,2}}{(\hat{\sigma}_{w,i,1}^2 + \hat{\kappa}_i^2 \hat{\sigma}_{w,i,2}^2)^{1/2}} \right). \tag{4.8}$$

It is easy to see that T_i is asymptotically standard normal under the null. It follows that the two-sided (approximate) p-values is $p_i = 2\{1 - \Phi(|T_i|)\}$. Moreover, T_i and S_i are asymptotically independent as shown in the following proposition. Let $t_i = \frac{\beta_{i,1} - \beta_{i,2}}{(\sigma_{w,i,1}^2 + \sigma_{w,i,2}^2)^{1/2}}$ and $s_i = \frac{\beta_{i,1} + \kappa_i \beta_{i,2}}{(\sigma_{w,i,1}^2 + \kappa_i^2 \sigma_{w,i,2}^2)^{1/2}}$ with $\kappa_i = \sigma_{w,i,1}^2/\sigma_{w,i,2}^2$ and $\sigma_{w,i,d}^2 = \sigma_{i,d}^2/n_d$.

Proposition 1 *For any constant $M > 0$, we have*

$$\mathbb{P}(|T_i| \geq t, |S_i| \geq \lambda) = (1 + o(1))\mathbb{P}(|N(0,1) + t_i| \geq t)\mathbb{P}(|N(0,1) + s_i| \geq \lambda) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$, $0 \leq \lambda \leq 4\sqrt{\log m}$ and $i = 1, \ldots, m$. Furthermore, for all $0 \leq j \leq 4N$ with fixed N,

$$\mathbb{P}(|T_i| \geq t, |S_i| < \lambda_j) = (1 + o(1))\mathbb{P}(|N(0,1) + t_i| \geq t)\mathbb{P}(|N(0,1) + s_i| < \lambda_j) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$ and $i = 1, \ldots, m$, where $\lambda_j = (j/N)\sqrt{\log m}$.

4.2 High-dimensional Linear Regression

Consider the two-sample regression model (1.2). Let $\boldsymbol{X}_d = (\boldsymbol{X}_{1,\cdot,d}^\intercal, \ldots, \boldsymbol{X}_{n_d,\cdot,d}^\intercal)^\intercal$ be the $n_d \times m$ data matrix, and $\boldsymbol{Y}_d^* = (Y_{1,d}^*, \ldots, Y_{n_d,d}^*)^\intercal$ be the $n_d \times 1$ data matrix, for $d = 1, 2$.

Throughout, suppose that we have i.i.d random samples $\{Y^*_{k,d}, \boldsymbol{X}_{k,\cdot,d}, 1 \leq k \leq n_d\}$ with $\boldsymbol{X}_{k,\cdot,d} = (X_{k,1,d}, \ldots, X_{k,m,d})$ being a random vector with covariance matrix $\boldsymbol{\Sigma}_d$ for $d = 1, 2$. Define $\boldsymbol{\Sigma}_d^{-1} = \boldsymbol{\Omega}_d = (\omega_{i,j,d})$. For any vector $\boldsymbol{\mu}_d \in \mathbb{R}^m$, let $\boldsymbol{\mu}_{-i,d}$ denote the $(m-1)$-dimensional vector by removing the i^{th} entry from $\boldsymbol{\mu}_d$. For any $n \times m$ matrix $\boldsymbol{A}_d$, $\boldsymbol{A}_{i,-j,d}$ denotes the i^{th} row of $\boldsymbol{A}_d$ with its j^{th} entry removed and $\boldsymbol{A}_{-i,j,d}$ denotes the j^{th} column of $\boldsymbol{A}_d$ with its i^{th} entry removed. $\boldsymbol{A}_{-i,-j,d}$ denotes the $(n-1) \times (m-1)$ submatrix of $\boldsymbol{A}_d$ with its i^{th} row and j^{th} column removed. Let $\boldsymbol{A}_{\cdot,-j,d}$ denote the $n \times (m-1)$ submatrix of $\boldsymbol{A}_d$ with the j^{th} column removed, $\boldsymbol{A}_{i,\cdot,d}$ denote the i^{th} row of $\boldsymbol{A}_d$, $\boldsymbol{A}_{\cdot,j,d}$ denote the j^{th} column of $\boldsymbol{A}_d$ and $\bar{A}_{\cdot,j,d} = 1/n \sum_{i=1}^n A_{i,j,d}$. Let $\bar{\boldsymbol{A}}_{\cdot,-j,d} = 1/n \sum_{i=1}^n \boldsymbol{A}_{i,-j,d}$, $\bar{\boldsymbol{A}}_{\cdot,j,d} = (\bar{A}_{\cdot,j,d}, \ldots, \bar{A}_{\cdot,j,d})^{\intercal}_{n\times 1}$, and $\bar{\boldsymbol{A}}_{(\cdot,-j,d)} = (\bar{\boldsymbol{A}}^{\intercal}_{\cdot,-j,d}, \ldots, \bar{\boldsymbol{A}}^{\intercal}_{\cdot,-j,d})^{\intercal}_{n\times(m-1)}$. Let $\bar{\boldsymbol{A}}_d = 1/n \sum_{i=1}^n \boldsymbol{A}_{i,\cdot,d}$.

4.2.1 Construction of the primary statistic

We divide the process into four steps, which are described below.

Step 1. Reformulation via inverse regression. We first explain the idea of *inverse regression* (Liu and Luo, 2014; Xia et al., 2018). Suppose we swap the response vector with one of the columns in the design matrix, then we obtain the following model

$$X_{k,i,d} \quad = \quad \alpha_{i,d} + (Y^*_{k,d}, \boldsymbol{X}_{k,-i,d})\boldsymbol{\gamma}_{i,d} + \eta_{k,i,d}, d = 1, 2, \tag{4.9}$$

where $\boldsymbol{\gamma}_{i,d} = (\gamma_{i,1,d}, \ldots, \gamma_{i,m,d})^{\intercal}$, and $\eta_{k,i,d}$ has mean zero and variance $\sigma^2_{\eta_{i,d}}$, and is uncorrelated with $(Y^*_{k,d}, \boldsymbol{X}_{k,-i,d})$. The covariance between the old error term and new error term can be calculated as

$$r_{i,d} = \mathsf{Cov}(\epsilon_{k,d}, \eta_{k,i,d}) = -\sigma^2_{\eta_{i,d}}\beta_{i,d},$$

where $\sigma^2_{\eta_{i,d}} = (\beta^2_{i,d}/\sigma^2_{\epsilon_d} + \omega_{i,i,d})^{-1}$. Hence the problem (1.1) can be equivalently stated as

$$H_{0,i} : r_{i,1}/\sigma^2_{\eta_{i,1}} = r_{i,2}/\sigma^2_{\eta_{i,2}} \text{ versus } H_{1,i} : r_{i,1}/\sigma^2_{\eta_{i,1}} \neq r_{i,2}/\sigma^2_{\eta_{i,2}},\ 1 \leq i \leq m. \tag{4.10}$$

We shall see that the new formulation (4.10) is instrumental because the ratios can be easily estimated from data and enjoy good theoretical properties.

Step 2. Estimating the ratios. Let $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{1,d}, \ldots, \hat{\beta}_{m,d})$ and $\hat{\boldsymbol{\gamma}}_{i,d} = (\hat{\gamma}_{i,1,d}, \ldots, \hat{\gamma}_{i,m,d})$ be estimates of the coefficients using standard methods such as LASSO or Dantzig selector. Then the corresponding residuals can be calculated as

$$\begin{aligned}
\hat{\epsilon}_{k,d} &= Y_{k,d} - \bar{Y}_d - (\boldsymbol{X}_{k,\cdot,d} - \bar{\boldsymbol{X}}_d)\hat{\boldsymbol{\beta}}_d, \\
\hat{\eta}_{k,i,d} &= X_{k,i,d} - \bar{X}_{i,d} - \left\{Y_{k,d} - \bar{Y}_d, (\boldsymbol{X}_{k,-i,d} - \bar{\boldsymbol{X}}_{\cdot,-i,d})\right\} \hat{\boldsymbol{\gamma}}_{i,d}.
\end{aligned}$$

The sample covariance and variances are given by

$$\tilde{r}_{i,d} = n_d^{-1} \sum_{k=1}^{n_d} \hat{\epsilon}_{k,d} \hat{\eta}_{k,i,d}\ , \hat{\sigma}^2_{\epsilon_d} = n_d^{-1} \sum_{k=1}^{n_d} \hat{\epsilon}^2_{k,d}, \text{ and } \hat{\sigma}^2_{\eta_{i,d}} = n_d^{-1} \sum_{k=1}^{n_d} \hat{\eta}^2_{k,i,d}.$$

The ratios in (4.10) can thus be obtained correspondingly.

Step 3. Bias correction. The empirical estimates $\tilde{r}_{i,d}$ in the previous step are biased [this has been noted, for example, in Xia et al. (2018)]. Some calculations show that the following step can be used to remove the bias:

$$\hat{r}_{i,d} = \tilde{r}_{i,d} + \hat{\sigma}^2_{\epsilon_d} \hat{\gamma}_{i,1,d} + \hat{\sigma}^2_{\eta_{i,d}} \hat{\beta}_{i,d}. \tag{4.11}$$

Step 4. Standardization. The goal of this step is to make the estimated differences comparable across tests. Consider the estimated ratios $\hat{r}_{i,d}/\hat{\sigma}^2_{\eta_{i,d}}$. The corresponding variances $\sigma^2_{w,i,d} = (\sigma^2_{\epsilon_d}/\sigma^2_{\eta_{i,d}} + \beta^2_{i,d})/n_d$ can be estimated by $\hat{\sigma}^2_{w,i,d} = (\hat{\sigma}^2_{\epsilon_d}/\hat{\sigma}^2_{\eta_{i,d}} + \hat{\beta}^2_{i,d})/n_d$. The standardization step gives the following primary test statistic:

$$T_i = \frac{\hat{r}_{i,1}/\hat{\sigma}^2_{\eta_{i,1}} - \hat{r}_{i,2}/\hat{\sigma}^2_{\eta_{i,2}}}{(\hat{\sigma}^2_{w,i,1} + \hat{\sigma}^2_{w,i,2})^{1/2}}, \quad 1 \leq i \leq m. \tag{4.12}$$

4.2.2 Construction of the auxiliary statistic

Next we explain the main idea in constructing S_i. To capture the information on the union support effectively, we focus on $\beta_{i,1} + \kappa_i \cdot \beta_{i,2}$, or equivalently, a class of weighted sums $(r_{i,1}/\sigma^2_{\eta_{i,1}}) + \kappa_i(r_{i,2}/\sigma^2_{\eta_{i,2}})$. The inverse regression technique can be used to obtain $\hat{r}_{i,d}$ and $\hat{\sigma}^2_{\eta_i d}$. To make the pair T_i and S_i asymptotically independent, we choose the weights as

$\hat{\kappa}_i = \hat{\sigma}^2_{w,i,1}/\hat{\sigma}^2_{w,i,2}$. Similar as before, we need to standardize the estimated weighted sums to make the test statistics comparable. The variances of the weighted sums can be calculated similarly as Step 4 in the previous subsection. Therefore we propose the following auxiliary statistic

$$S_i = \frac{\hat{r}_{i,1}/\hat{\sigma}^2_{\eta_{i,1}} + (\hat{\sigma}^2_{w,i,1}/\hat{\sigma}^2_{w,i,2})(\hat{r}_{i,2}/\hat{\sigma}^2_{\eta_{i,2}})}{\{\hat{\sigma}^2_{w,i,1}(1+\hat{\sigma}^2_{w,i,1}/\hat{\sigma}^2_{w,i,2})\}^{1/2}}, \quad 1 \le i \le m. \tag{4.13}$$

4.2.3 Theoretical properties of T_i and S_i

This section establishes two important theoretical properties that are crucial for proving the validity of the GAP procedure in FDR control: (i) the asymptotic normality of T_i (Proposition 2) and (ii) the asymptotic independence between T_i and S_i (Proposition 3). We assume that the estimators of $\boldsymbol{\beta}_d$ and $\boldsymbol{\gamma}_{i,d}$ satisfy

$$\begin{aligned}
\mathbb{P}\left(\max\{|\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d|_1, \max_{1\le i\le m} |\hat{\boldsymbol{\gamma}}_{i,d} - \boldsymbol{\gamma}_{i,d}|_1\} \ge a_{n1}\right) &= O(m^{-M}),\\
\mathbb{P}\left(\max\{|\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d|_2, \max_{1\le i\le m} |\hat{\boldsymbol{\gamma}}_{i,d} - \boldsymbol{\gamma}_{i,d}|_2\} \ge a_{n2}\right) &= O(m^{-M}),
\end{aligned} \tag{4.14}$$

for any constant $M > 0$, where a_{n1} and a_{n2} satisfy

$$\max\{a_{n1}a_{n2}, a^2_{n2}\} = o\{(n\log m)^{-1/2}\}, \text{ and } a_{n1} = o(1/\log m). \tag{4.15}$$

Similar conditions are fulfilled by estimates obtained from standard high-dimensional regression methods such as the LASSO, SCAD or Dantzig Selector with mild sparsity assumptions (see, e.g., Zhang and Huang (2008), Candes and Tao (2007), Liu (2013) and Xia et al. (2018)). The next proposition shows that T_i follows a standard normal distribution asymptotically; according to this proposition we define two-sided p-values as $p_i = 2\{1 - \Phi(|T_i|)\}$.

Proposition 2 *Suppose (4.14) and (4.15), and the following two conditions hold:*

(C1) Assume that $\log m = o(n^{1/5})$, $n_1 \asymp n_2$, *and for some constants* $C_0, C_1, C_2 > 0$, $C_0^{-1} \le \lambda_{\min}(\boldsymbol{\Omega}_d) \le \lambda_{\max}(\boldsymbol{\Omega}_d) \le C_0$, $C_1^{-1} \le \sigma^2_{\epsilon_d} \le C_1$, *and* $|\boldsymbol{\beta}_d|_\infty \le C_2$ *for* $d = 1, 2$.

(C2) There exists some constant $K > 0$ *such that* $\max_{\mathsf{Var}(\boldsymbol{a}^\top \boldsymbol{X}^\top_{k,\cdot,d})=1} \mathbb{E}\exp(K(\boldsymbol{a}^\top \boldsymbol{X}^\top_{k,\cdot,d})^2)$

and $\mathbb{E}\exp(K\epsilon_{k,d}^2)$ *are finite.*

Then, as $n_1, n_2, m \to \infty$*,*

$$T_i - \frac{f_i}{(\sigma_{w,i,1}^2 + \sigma_{w,i,2}^2)^{1/2}} \Rightarrow N(0,1),$$

uniformly in $i = 1, \ldots, m$*, where* $f_i = (\tilde{\sigma}_{\epsilon_1}^2/\sigma_{\epsilon_1}^2 + \tilde{\sigma}_{\eta_{i,1}}^2/\sigma_{\eta_{i,1}}^2 - 1)\beta_{i,1} - (\tilde{\sigma}_{\epsilon_2}^2/\sigma_{\epsilon_2}^2 + \tilde{\sigma}_{\eta_{i,2}}^2/\sigma_{\eta_{i,2}}^2 - 1)\beta_{i,2}$ *and* $\tilde{\sigma}_{\epsilon_d}^2 = n_d^{-1}\sum_{k=1}^{n_d}(\epsilon_{k,d} - \bar{\epsilon}_{k,d})^2$ *and* $\tilde{\sigma}_{\eta_{i,d}}^2 = n_d^{-1}\sum_{k=1}^{n_d}(\eta_{k,i,d} - \bar{\eta}_{k,i,d})^2$ *with* $\bar{\epsilon}_{k,d} = n_d^{-1}\sum_{k=1}^{n_d}\epsilon_{k,d}$ *and* $\bar{\eta}_{k,i,d} = n_d^{-1}\sum_{k=1}^{n_d}\eta_{k,i,d}$*.*

Remark 5 *Note that under (C1),* $f_i = \{1+O_{\mathbb{P}}(\sqrt{\log m/n})\}\beta_{i,1} - \{1+O_{\mathbb{P}}(\sqrt{\log m/n})\}\beta_{i,2} = O_{\mathbb{P}}(\sqrt{\log m/n})\max\{|\beta_{i,1}|, |\beta_{i,2}|\}$ *under the null hypothesis* $H_{i,0} : \beta_{i,1} = \beta_{i,2}$*. Furthermore, the detailed convergence rate as required in (A1) is shown in the proof.*

Define $s_i = \frac{\beta_{i,1} + (\sigma_{w,i,1}^2/\sigma_{w,i,2}^2)\beta_{i,2}}{\sqrt{\sigma_{w,i,1}^2(1+\sigma_{w,i,1}^2/\sigma_{w,i,2}^2)}}$. Let $G(t) = 2(1 - \Phi(t))$ with $\Phi(t)$ to be the cumulative distribution function of standard normal random variable. The next proposition shows that T_i and S_i are asymptotically independent under the null $H_{i,0} : \beta_{i,1} = \beta_{i,2}$.

Proposition 3 *Suppose (C1), (C2), (4.14) and (4.15) hold. Then for any constant* $M > 0$*,*

$$\mathbb{P}\left(\left|T_i - \frac{f_i}{(\sigma_{w,i,1}^2 + \sigma_{w,i,2}^2)^{1/2}}\right| \geq t, |S_i| \geq \lambda\right) = (1+o(1))G(t)\mathbb{P}(|N(0,1)+s_i| \geq \lambda) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$*,* $0 \leq \lambda \leq 4\sqrt{\log m}$ *and* $i = 1, \ldots, m$*. Furthermore, for all* $0 \leq j \leq 4N$ *with fixed* N*,*

$$\mathbb{P}\left(\left|T_i - \frac{f_i}{(\sigma_{w,i,1}^2 + \sigma_{w,i,2}^2)^{1/2}}\right| \geq t, |S_i| < \lambda_j\right) = (1+o(1))G(t)\mathbb{P}(|N(0,1)+s_i| < \lambda_j) + O(m^{-M}),$$

uniformly for $0 \leq t \leq 4\sqrt{\log m}$ *and* $i = 1, \ldots, m$*, where* $\lambda_j = (j/N)\sqrt{\log m}$*.*

4.3 Covariance, Correlation and Precision Matrices

This section considers simultaneous inference with two sparse matrices. The ideas and techniques in the derivation of T_i and S_i in the regression context carry over to the settings for two-sample inference of covariance and precision matrices. Hence we omit the details and only outline the main steps in the derivation. Suppose we observe random samples $\{\boldsymbol{Y}_{1,\cdot,d}, \cdots, \boldsymbol{Y}_{n_d,\cdot,d}\}$ as independent copies of $\boldsymbol{Y}_d$, where we denote the covariance matrix of $\boldsymbol{Y}_d$ by $\boldsymbol{\Sigma}_d = (\beta_{i,j,d} : 1 \le i, j \le p)$, $d = 1, 2$. The goal is to make inference of $\theta_{i,j} = \mathbb{I}(\beta_{i,j,1} \neq \beta_{i,j,2})$. The two sparse objects are $\boldsymbol{B}_1 = (\beta_{i,j,1})_{p\times p}$ and $\boldsymbol{B}_2 = (\beta_{i,j,2})_{p\times p}$.

4.3.1 Covariance/Correlation Matrices

Suppose we are interested in detecting significant correlations/covariances changes between two populations. The problem can be formulated as a two-sample multiple testing problem (1.1) with $\boldsymbol{B}_d = \boldsymbol{\Sigma}_d$. Define the sample covariance matrices

$$(\hat{\beta}_{i,j,d})_{p\times p} := \hat{\boldsymbol{\Sigma}}_d = \frac{1}{n_d}\sum_{k=1}^{n_d}(\boldsymbol{Y}_{k,d} - \bar{\boldsymbol{Y}}_d)(\boldsymbol{Y}_{k,d} - \bar{\boldsymbol{Y}}_d)',$$

where $\bar{\boldsymbol{Y}}_d = \frac{1}{n_d}\sum_{k=1}^{n_d} Y_{k,d}$. We standardize $\hat{\beta}_{i,j,1} - \hat{\beta}_{i,j,2}$ by estimating the variances as introduced in Cai et al. (2013), namely,

$$\hat{\sigma}^2_{i,j,d} = \frac{1}{n_d^2}\sum_{k=1}^{n_d}\Big[(Y_{k,i,d} - \bar{Y}_{i,d})(Y_{k,j,d} - \bar{Y}_{j,d}) - \hat{\beta}_{i,j,1}\Big]^2, \quad \bar{Y}_{i,d} = \frac{1}{n_d}\sum_{k=1}^{n_d} Y_{k,i,d}.$$

Then we define the primary test statistics by

$$T_{i,j} = \frac{\hat{\beta}_{i,j,1} - \hat{\beta}_{i,j,2}}{(\hat{\sigma}^2_{i,j,1} + \hat{\sigma}^2_{i,j,2})^{1/2}}, \quad 1 \le i \le j \le p. \tag{4.16}$$

To capture the information on the union support, we focus on $\beta_{i,j,1} + \kappa_{i,j} \cdot \beta_{i,j,2}$. To make T_i and S_i asymptotically independent, we choose the weights as $\hat{\kappa}_{i,j} = \hat{\sigma}^2_{i,j,1}/\hat{\sigma}^2_{i,j,2}$, which leads to the following auxiliary statistic

$$S_{i,j} = \frac{\hat{\beta}_{i,j,1} + (\hat{\sigma}^2_{i,j,1}/\hat{\sigma}^2_{i,j,2})\hat{\beta}_{i,j,2}}{\{\hat{\sigma}^2_{i,j,1}(1 + \hat{\sigma}^2_{i,j,1}/\hat{\sigma}^2_{i,j,2})\}^{1/2}}.$$

For notational consistency, we rearrange the two-dimensional indices $\{(i,j) : 1 \le i \le j \le p\}$ as $\{(a_i, b_i) : 1 \le i \le m\}$, where $m = p(p+1)/2$. Then the primary and auxiliary statistics can be denoted

$$T_i = \frac{\hat{\beta}_{i,1} - \hat{\beta}_{i,2}}{(\hat{\sigma}^2_{w,i,1} + \hat{\sigma}^2_{w,i,2})^{1/2}}, \text{ and } S_i = \frac{\hat{\beta}_{i,1} + (\hat{\sigma}^2_{w,i,1}/\hat{\sigma}^2_{w,i,2})\hat{\beta}_{i,2}}{\{\hat{\sigma}^2_{w,i,1}(1 + \hat{\sigma}^2_{w,i,1}/\hat{\sigma}^2_{w,i,2})\}^{1/2}} \quad 1 \le i \le m, \tag{4.17}$$

where $\hat{\beta}_{i,d} = \hat{\beta}_{a_i,b_i,d}$ and $\hat{\sigma}^2_{w,i,d} = \hat{\sigma}^2_{a_i,b_i,d}$.

For testing the correlation matrices, we have $\boldsymbol{B}_d = \boldsymbol{D}_d^{-1/2}\boldsymbol{\Sigma}_d\boldsymbol{D}_d^{-1/2}$, with $\boldsymbol{D}_d$ being the diagonal matrix of $\boldsymbol{\Sigma}_d$. The primary and auxiliary statistics can be constructed based on

$$\hat{\beta}_{i,d} = \frac{\sum_{k=1}^{n_d}(Y_{k,a_i,d} - \bar{Y}_{a_i,d})(Y_{k,b_i,d} - \bar{Y}_{b_i,d})}{\{\sum_{k=1}^{n_d}(Y_{k,a_i,d} - \bar{Y}_{a_i,d})^2 \sum_{k=1}^{n_d}(Y_{k,b_i,d} - \bar{Y}_{b_i,d})^2\}^{1/2}},$$

where $\hat{\sigma}^2_{w,i,d}$ are the variance estimates of the above defined $\hat{\beta}_{i,d}$ as introduced in the denominator of equation (5) of Cai and Liu (2016). In the correlation matrix testing scenario, we have $m = p(p-1)/2$ because only off-diagonal elements are of primary interest. For both settings we can similarly show that $\{(T_i, S_i), 1 \le i \le m\}$ satisfy (A1) and (A3) in Section 3 under the regularity conditions as described in Cai et al. (2013), and the detailed proof is shown in Section **??**.

4.3.2 Gaussian Graphical Models

Suppose that $\boldsymbol{Y}_d \in \mathbb{R}^p \sim N(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$, then under the Gaussian Graphical Model (GGM) framework, we translate the problem of identifying changes of conditional dependency between variables of interest into testing the off-diagonal elements of two precision matrices $\boldsymbol{\Omega}_d = \boldsymbol{\Sigma}_d^{-1}$, namely, we have $\boldsymbol{B}_d = \boldsymbol{\Omega}_d$, and one wishes to test

$$H_{0,i,j} : \beta_{i,j,1} = \beta_{i,j,2} \text{ versus } H_{1,i,j} : \beta_{i,j,1} \ne \beta_{i,j,2}, \quad 1 \le i < j \le p.$$

We utilize the inverse regression models to estimate $\boldsymbol{\Omega}_d$ as studied in Xia et al. (2015), i.e.,

$$Y_{k,i,d} \;=\; \alpha_{i,d} + \boldsymbol{Y}_{k,-i,d}\gamma_{i,2} + \epsilon_{k,i,d}, \quad (i = 1, \dots, p;\ k = 1, \dots, n_d), \tag{4.18}$$

where $\epsilon_{k,i,d} \sim N(0, \sigma_{i,i,d} - \boldsymbol{\Sigma}_{i,-i,d}\boldsymbol{\Sigma}_{-i,-i,d}^{-1}\boldsymbol{\Sigma}_{-i,i,d})$ $(d = 1, 2)$ are independent of $\boldsymbol{Y}_{k,-i,d}$, and $\alpha_{i,d} = \mu_{i,d} - \boldsymbol{\Sigma}_{i,-i,d}\boldsymbol{\Sigma}_{-i,-i,d}^{-1}\boldsymbol{\mu}_{-i,d}$. The regression coefficient vectors $\boldsymbol{\gamma}_{i,d}$ and the error terms $\epsilon_{k,i,d}$ satisfy $\boldsymbol{\gamma}_{i,d} = -\omega_{i,i,d}^{-1}\boldsymbol{\Omega}_{-i,i,d}$ and $r_{i,j,d} = \mathsf{Cov}(\epsilon_{k,i,d}, \epsilon_{k,j,d}) = \frac{\omega_{i,j,d}}{\omega_{i,i,d}\omega_{j,j,d}}$. We construct the debiased estimators of $r_{i,j,d}$ by $\hat{r}_{i,j,d} = -(\tilde{r}_{i,j,d} + \tilde{r}_{i,i,d}\hat{\gamma}_{i,j,d} + \tilde{r}_{j,j,d}\hat{\gamma}_{j-1,i,d})$, for $1 \leq i < j \leq p$, and $\hat{r}_{i,j,d} = \tilde{r}_{i,j,d}$ when $i = j$, where $\tilde{r}_{i,j,d} = \frac{1}{n_d}\sum_{k=1}^{n_d}\hat{\epsilon}_{k,i,d}\hat{\epsilon}_{k,j,d}$, $\hat{\epsilon}_{k,i,d} = Y_{k,i,d} - \bar{Y}_{i,d} - (\boldsymbol{Y}_{k,-i,d} - \bar{\boldsymbol{Y}}_{\cdot,-i,d})\hat{\boldsymbol{\gamma}}_{i,d}$, and $\hat{\boldsymbol{\gamma}}_{i,d}$ are estimators of $\boldsymbol{\gamma}_{i,d}$ that can be obtained via Lasso and Dantzig selector. The primary test statistics can be constructed as

$$T_{i,j} = \frac{\hat{r}_{i,j,1}/(\hat{r}_{i,i,1}\hat{r}_{j,j,1}) - \hat{r}_{i,j,2}/(\hat{r}_{i,i,2}\hat{r}_{j,j,2})}{(\hat{\sigma}_{i,j,1}^2 + \hat{\sigma}_{i,j,2}^2)^{1/2}}, \quad 1 \leq i < j \leq p,$$

where $\hat{\sigma}_{i,j,d}^2 = (1 + \hat{\gamma}_{i,j,d}^2\hat{r}_{i,i,d}/\hat{r}_{j,j,d})/(n_d\hat{r}_{i,i,d}\hat{r}_{j,j,d})$ are the estimators of the variances. The auxiliary statistics are constructed as

$$S_{i,j} = \frac{\hat{r}_{i,j,1}/(\hat{r}_{i,i,1}\hat{r}_{j,j,1}) + (\hat{\sigma}_{i,j,1}^2/\hat{\sigma}_{i,j,2}^2)\hat{r}_{i,j,2}/(\hat{r}_{i,i,2}\hat{r}_{j,j,2})}{\{\hat{\sigma}_{i,j,1}^2(1 + \hat{\sigma}_{i,j,1}^2/\hat{\sigma}_{i,j,2}^2)\}^{1/2}}.$$

Rearranging the two-dimensional indices $\{(i,j) : 1 \leq i < j \leq p\}$ and setting $\{(a_i, b_i) : 1 \leq i \leq m\}$, the primary and auxiliary statistics can be denoted

$$T_i = \frac{\hat{\beta}_{i,1} - \hat{\beta}_{i,2}}{(\hat{\sigma}_{w,i,1}^2 + \hat{\sigma}_{w,i,2}^2)^{1/2}}, \text{ and } S_i = \frac{\hat{\beta}_{i,1} + (\hat{\sigma}_{w,i,1}^2/\hat{\sigma}_{w,i,2}^2)\hat{\beta}_{i,2}}{\{\hat{\sigma}_{w,i,1}^2(1 + \hat{\sigma}_{w,i,1}^2/\hat{\sigma}_{w,i,2}^2)\}^{1/2}} \quad 1 \leq i \leq m, \tag{4.19}$$

where $m = p(p-1)/2$, $\hat{\beta}_{i,d} = \hat{r}_{a_i,b_i,d}/(\hat{r}_{a_i,a_i}\hat{r}_{b_i,b_i})$ and $\hat{\sigma}_{w,i,d}^2 = \hat{\sigma}_{a_i,b_i,d}^2$. Again, it can be shown that $\{(T_i, S_i), 1 \leq i \leq m\}$ satisfy (A1) and (A3) in Section 3 under the regularity conditions described in Xia et al. (2015).

5 Simulation studies

We now turn to the numerical performance of the GAP algorithm. Simulation studies are carried out to compare the performance of the following methods: (a) The BH procedure (naive pooled analysis), denoted by BH. (b). Separate analysis (grouping without weighting) with 2 and 3 groups, denoted by 2G and 3G respectively. (c). The proposed

GAP procedure with 3 groups, denoted by GAP. We present the results for weakly dependent tests and high-dimensional linear regression in Sections 5.1 and 5.2, respectively. The results for Gaussian graphical models are provided in the Supplementary Material.

5.1 Weakly Dependent Tests

We simulate two vectors of correlated z-values of dimension $p = 2000$ from $\boldsymbol{Y}_d \sim N(\boldsymbol{\beta}_d, \boldsymbol{\Sigma})$, $d = 1, 2$, from the following three models, where three covariance matrices $\boldsymbol{\Sigma}^{(1)}$, $\boldsymbol{\Sigma}^{(2)}$ and $\boldsymbol{\Sigma}^{(3)}$ are considered, respectively.

- Model 1: $\boldsymbol{\Sigma}^{(1)} = (\sigma_{i,j}^{(1)})$, where $\sigma_{i,j}^{(1)} = 0.8^{|i-j|}$ for $1 \leq i, j \leq m$.
- Model 2: $\boldsymbol{\Sigma}^{(2)} = (\sigma_{i,j}^{(2)})$, where $\sigma_{i,i}^{(2)} = 1$, $\sigma_{i,j}^{(1)} = 0.5$ for $3(k-1)+1 \leq i \neq j \leq 3k$, $k = 1, ..., [m/3]$, and $\sigma_{ij}^{(2)} = 0$ otherwise.
- Model 3: $\boldsymbol{\Sigma}^{*(3)} = (\sigma_{i,j}^{*(3)})$ where $\sigma_{i,i}^{*(3)} = 1$, $\sigma_{i,j}^{*(3)} = 0.5 * \text{Bernoulli}(1, 0.05)$ for $i < j$ and $\sigma_{j,i}^{*(3)} = \sigma_{i,j}^{*(3)}$. For positive definiteness, further let $\boldsymbol{\Sigma}^{(3)} = (\boldsymbol{\Sigma}^{*(3)} + \delta \boldsymbol{I})/(1+\delta)$ with $\delta = |\lambda_{\min}(\boldsymbol{\Sigma}^{*(3)})| + 0.05$.

The mean vectors $\boldsymbol{\beta}_d$, $d = 1, 2$, are generated as follows. We first set $\beta_{i,1} = 3$, $\beta_{i,2} = \beta$ for $1 \leq i \leq 50$, $\beta_{i,1} = -3$, $\beta_{i,2} = -\beta$ for $51 \leq i \leq 100$, then vary β with values $6.5, 7.0, 7.5, 8.0$, and finally apply different methods at FDR level $\alpha = 0.05$. Empirical FDRs and powers are estimated based on 200 replications. The standard error of the estimated FDR for GAP is stable and is around 0.02 in all settings. Hence we feel that using 200 replications should to be sufficient for reaching a reliable conclusion. The FDR and power comparisons are illustrated in Figure 2. We make the following remarks based on the simulation results.

(a). The three plots in the left column show that all methods control the FDR reasonably well in all three settings.

(b). The power of BH can be greatly improved by 2G, which exploits the information in the auxiliary sequence.

(c). The power of 2G can be further increased by 3G and GAP.

(d). GAP has smaller FDR level and similar power compared to 3G.

(e). To further illustrate the difference between GAP and 3G, we adjust the FDR levels of GAP according to the ratios of the empirical FDRs for GAP and 3G, and then match the corresponding powers of GAP and 3G at roughly the same FDR level. The results are displayed in the three plots in the right column. We can see that GAP outperforms 3G in power under this new setting where the FDRs are matched at roughly the same level; this indicates that GAP has greater power than 3G at the same FDR level.

Remark 6 GAP utilizes a standardization step in its operation. This standardization step, which guarantees the FDR control, tends to lead to more conservative FDR levels as observed in our simulation studies. This normalizing step is desirable as it guarantees the validity of GAP for FDR control in more complicated situations such as high-dimensional regression models and large GGM. As we shall see in later simulation studies on GGM, 3G fails to control the FDR but GAP still works.

5.2 High-dimensional Linear Regression

Consider the two-sample regression model (1.2). The following three models considered in Xia et al. (2018) are used to generate the design matrices. Let $\boldsymbol{D} = (D_{i,j})$ be a diagonal matrix with $D_{i,i} = \text{Unif}(1,3)$ for $i = 1, \ldots, m$.

- Model 1: $\boldsymbol{\Omega}^{*(1)} = (\omega^{*(1)}_{i,j})$ where $\omega^{*(1)}_{i,i} = 1$, $\omega^{*(1)}_{i,i+1} = \omega^{*(1)}_{i+1,i} = 0.6$, $\omega^{*(1)}_{i,i+2} = \omega^{*(1)}_{i+2,i} = 0.3$ and $\omega^{*(1)}_{i,j} = 0$ otherwise. Let $\boldsymbol{\Omega}^{(1)} = \boldsymbol{D}^{1/2}\boldsymbol{\Omega}^{*(1)}\boldsymbol{D}^{1/2}$.

- Model 2: $\boldsymbol{\Omega}^{*(2)} = (\omega^{*(2)}_{i,j})$ where $\omega^{*(2)}_{i,j} = \omega^{*(2)}_{j,i} = 0.5$ for $i = 10(k-1)+1$ and $10(k-1)+2 \leq j \leq 10(k-1)+10$, $1 \leq k \leq m/10$. $\omega^{*(2)}_{i,j} = 0$ otherwise. Let $\boldsymbol{\Omega}^{(2)} = \boldsymbol{D}^{1/2}(\boldsymbol{\Omega}^{*(2)} + \delta \boldsymbol{I})/(1+\delta)\boldsymbol{D}^{1/2}$ with $\delta = |\lambda_{\min}(\boldsymbol{\Omega}^{*(2)})| + 0.05$.

- Model 3: $\boldsymbol{\Omega}^{*(3)} = (\omega^{*(3)}_{i,j})$ where $\omega^{*(3)}_{i,i} = 1$, $\omega^{*(3)}_{i,j} = 0.8 \times \text{Bernoulli}(1, 2/p)$ for $i < j$ and $\omega^{*(3)}_{j,i} = \omega^{*(3)}_{i,j}$. Let $\boldsymbol{\Omega}^{(3)} = \boldsymbol{D}^{1/2}(\boldsymbol{\Omega}^{*(3)} + \delta \boldsymbol{I})/(1+\delta)\boldsymbol{D}^{1/2}$ with $\delta = |\lambda_{\min}(\boldsymbol{\Omega}^{*(3)})| + 0.05$.

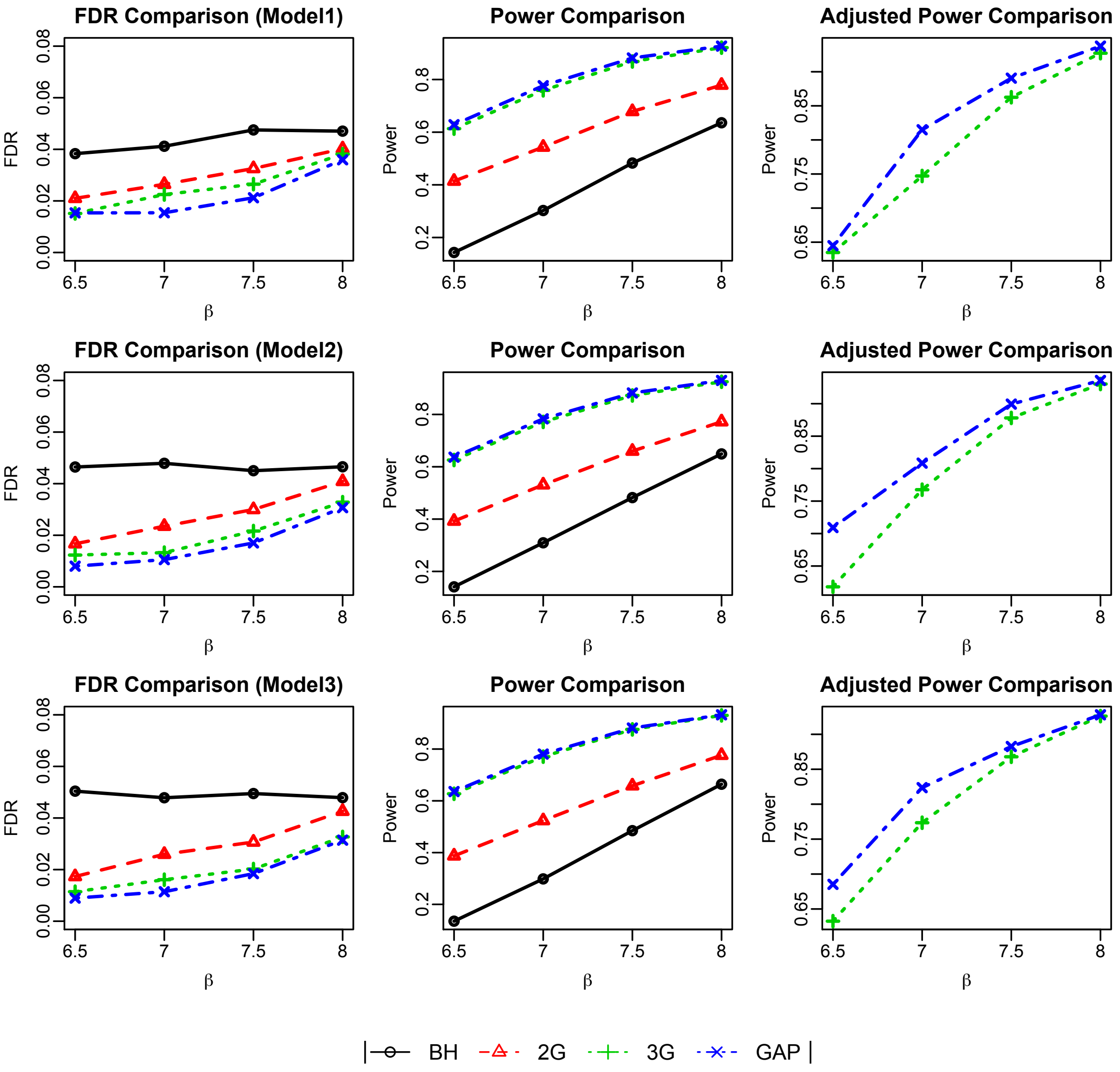

Figure 2: FDR, Power and adjusted Power comparisons on weakly dependent normal vectors between BH, 2G, 3G and GAP.

The design matrices are $\boldsymbol{X}_{k,\cdot,d}$, for $k = 1, \ldots, n_d$ and $d = 1, 2$, generated with some of the covariates being continuous and the others being discrete. We first obtain i.i.d samples $\boldsymbol{X}_{k,\cdot,d} \sim N(0, \boldsymbol{\Sigma}^{(f)})$ with $\boldsymbol{\Sigma}^{(f)} = (\boldsymbol{\Omega}^{(f)})^{-1}$, for $k = 1, \ldots, n_d$, with $f = 1, 2$ and 3, from three models above, and then replace l covariates of $\boldsymbol{X}_{k,\cdot,d}$ by one of three discrete values 0, 1 or 2, with probability 1/3 each, where l is a random integer between $\lfloor m/2 \rfloor$ and m.

Let $m = 200$ and $s = 15$. We randomly select s nonzero locations to form set $\Lambda_0 = \{k_1, \ldots, k_s\}$. Let $\beta_{k_i,1} = 2i^{0.5}n_1^{-a}$ and $\beta_{k_i,2} = 2.5i^{0.5}n_2^{-a}$ for $i = 1, \ldots, \lfloor s/2 \rfloor$, $\beta_{k_i,1} = -2i^{0.5}n_1^{-a}$, and $\beta_{k_i,2} = -2.5i^{0.5}n_2^{-a}$ for $i = \lfloor s/2 \rfloor + 1, \ldots, s$, with $a = 0.05, 0.1, 0.15$ and 0.2. Finally, we randomly select s nonzero locations respectively to form Λ_1 and Λ_2. Let $\beta_{k_i,1} = -2i^{0.5}n_1^{-a}$ for $k_i \in \Lambda_1 \setminus \Lambda_0$, and $\beta_{k_i,2} = 2.5i^{0.5}n_2^{-a}$ for $k_i \in \Lambda_2 \setminus \Lambda_0$. The sample sizes are taken to be $n = n_1 = n_2 = 200$. The reported FDR and power levels are calculated by averaging the results based on 50 replications. The regression coefficients $\boldsymbol{\beta}_d$ and $\boldsymbol{\gamma}_{i,d}$ are estimated by Lasso; see Section 5.1 of Xia et al. (2018) for a detailed description of the estimation procedure. We then construct the primary and auxiliary statistics based on estimated coefficients and apply different methods at the nominal FDR level $\alpha = 0.05$.

The FDR and power comparisons are illustrated in Figure 3. Similar conclusions can be drawn as before base on the simulation results: all the methods control the FDR at the pre-specified level; the power of BH is improved by 2G, which is further improved by 3G; and GAP is the most powerful method. It is important to note that GAP simultaneously has smaller FDR and larger power than 3G in all settings.

5.3 Simulations on Gaussian Graphical Models

We consider additional simulation comparisons on Gaussian Graphical Models in this section. The following four methods are studied: (a) The BH procedure, denoted BH. (b). Separate analysis (grouping without weighting) with 2 and 3 groups, denoted 2G, 3G. (c). The proposed GAP procedure with 3 groups, denoted by GAP.

Let $\boldsymbol{D} = (D_{i,j})$ be a diagonal matrix with $D_{i,i} = \text{Unif}(0.5, 2.5)$ for $i = 1, \ldots, m$. We considered the three graphical models as studied in Xia et al. (2015).

- Model 1: $\boldsymbol{\Omega}^{*(1)} = (\omega^{*(1)}_{i,j})$ where $\omega^{*(1)}_{i,i} = 1$, $\omega^{*(1)}_{i,i+1} = \omega^{*(1)}_{i+1,i} = 0.6$, $\omega^{*(1)}_{i,i+2} = \omega^{*(1)}_{i+2,i} = 0.3$

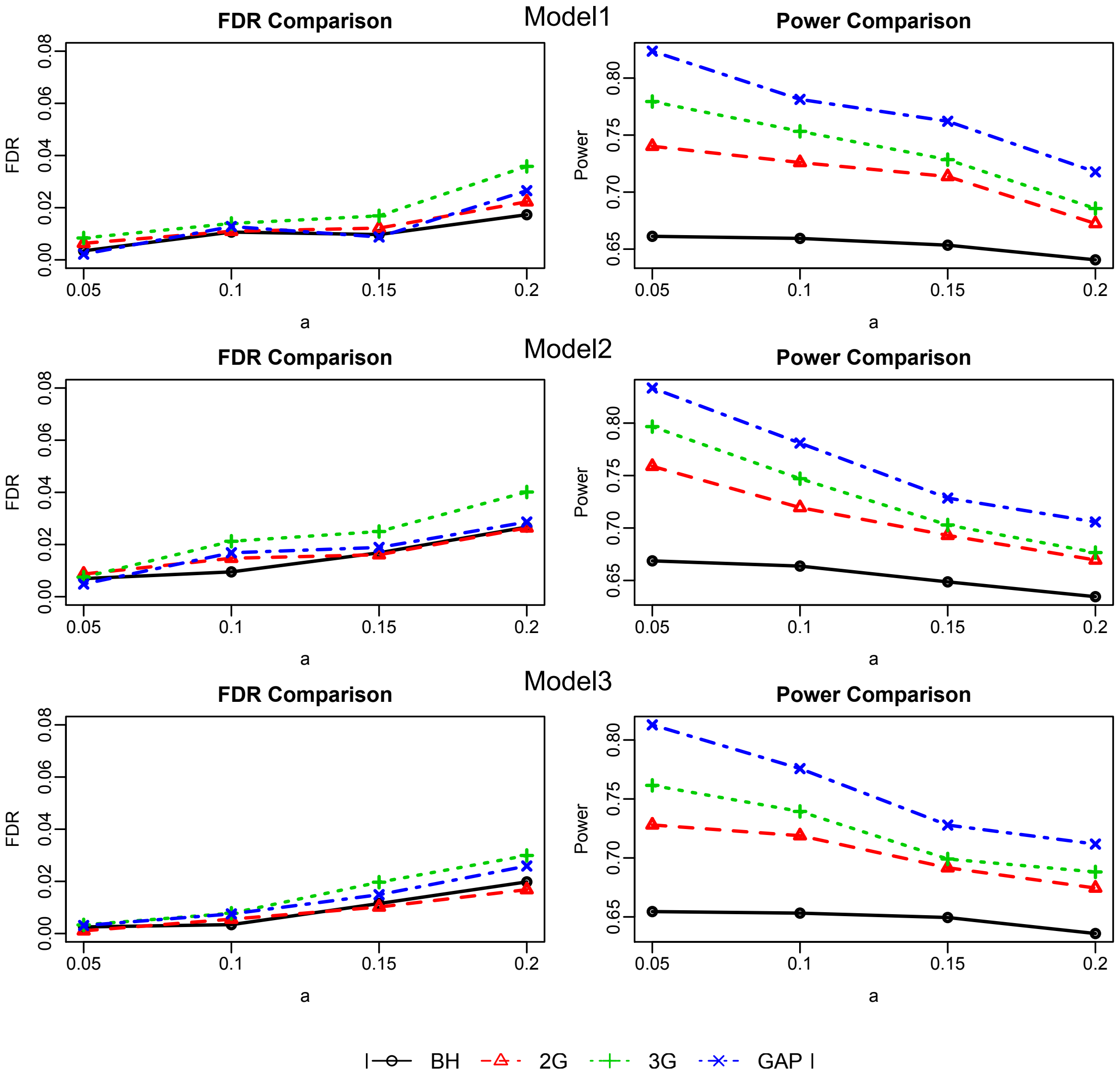

Figure 3: FDR and Power comparisons on regression models between BH, 2G, 3G and GAP.

and $\omega_{i,j}^{*(1)} = 0$ otherwise. $\boldsymbol{\Omega}^{(1)} = \boldsymbol{D}^{1/2}\boldsymbol{\Omega}^{*(1)}\boldsymbol{D}^{1/2}$.

- Model 2: $\boldsymbol{\Omega}^{*(2)} = (\omega_{i,j}^{*(2)})$ where $\omega_{i,j}^{*(2)} = \omega_{j,i}^{*(2)} = 0.5$ for $i = 10(k-1)+1$ and $10(k-1)+2 \leq j \leq 10(k-1)+10$, $1 \leq k \leq p/10$. $\omega_{i,j}^{*(2)} = 0$ otherwise. $\boldsymbol{\Omega}^{(2)} = \boldsymbol{D}^{1/2}(\boldsymbol{\Omega}^{*(2)} + \delta\boldsymbol{I})/(1+\delta)\boldsymbol{D}^{1/2}$ with $\delta = |\lambda_{\min}(\boldsymbol{\Omega}^{*(2)})| + 0.05$.

- Model 3: $\boldsymbol{\Omega}^{*(3)} = (\omega_{i,j}^{*(3)})$ where $\omega_{i,i}^{*(3)} = 1$, $\omega_{i,j}^{*(3)} = 0.8 \times \text{Bernoulli}(1, 0.05)$ for $i < j$ and $\omega_{j,i}^{*(3)} = \omega_{i,j}^{*(3)}$. $\boldsymbol{\Omega}^{(3)} = \boldsymbol{D}^{1/2}(\boldsymbol{\Omega}^{*(3)} + \delta\boldsymbol{I})/(1+\delta)\boldsymbol{D}^{1/2}$ with $\delta = |\lambda_{\min}(\boldsymbol{\Omega}^{*(3)})| + 0.05$.

We let $\boldsymbol{\Omega}_1^* = \boldsymbol{\Omega}^{(s)} = (\omega_{i,j}^{(s)})$ for $s = 1, 2, 3$, and construct $\boldsymbol{\Omega}_2^*$ by removing half of the nonzero entries in $\boldsymbol{\Omega}_1^*$ and setting the rest have magnitudes half of the original values. Let $\delta = |\lambda_{\min}(\boldsymbol{\Omega}_2^*)| + 0.05$, and set $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_1^* + \delta\boldsymbol{I}$ and $\boldsymbol{\Omega}_2 = \boldsymbol{\Omega}_2^* + \delta\boldsymbol{I}$. We select the dimension $p = 50, 100$ and 200, and set $n = n_1 = n_2 = 100$. The nominal level is chosen to be $\alpha = 0.1$. Empirical sizes and powers are estimated based on 50 replications.

The FDR and power comparisons are illustrated in Figure 4. We can see from the figure that most of the methods control the FDR at the pre-specified level well, while the 3G method has serious FDR distortions in Models 2 and 3. Figure 4 also shows that the power difference is very clear among these four procedures, and in all three models, the GAP procedure shows the clear advantage over BH, and 2G across all dimensions, and it has similar power performance as 3G. However, the power gain of 3G is due to the inflation of its FDR.

6 Analysis of Differential Gene Networks

This section applies the GAP algorithm for analyzing a breast cancer dataset to identify gene-gene interactions whose effect sizes have changed significantly between two groups of patients. In clinical practice, it has been discovered that many prominent genomic markers are useful predictors of breast cancer survival, and increasingly, pharmacogenomic endpoints are being incorporated into the design of clinical trials (Olopade et al., 2008). Empirical evidence from model organisms and human studies suggests that gene-gene interactions make an important contribution to total genetic variation of complex traits (Zerba et al.,

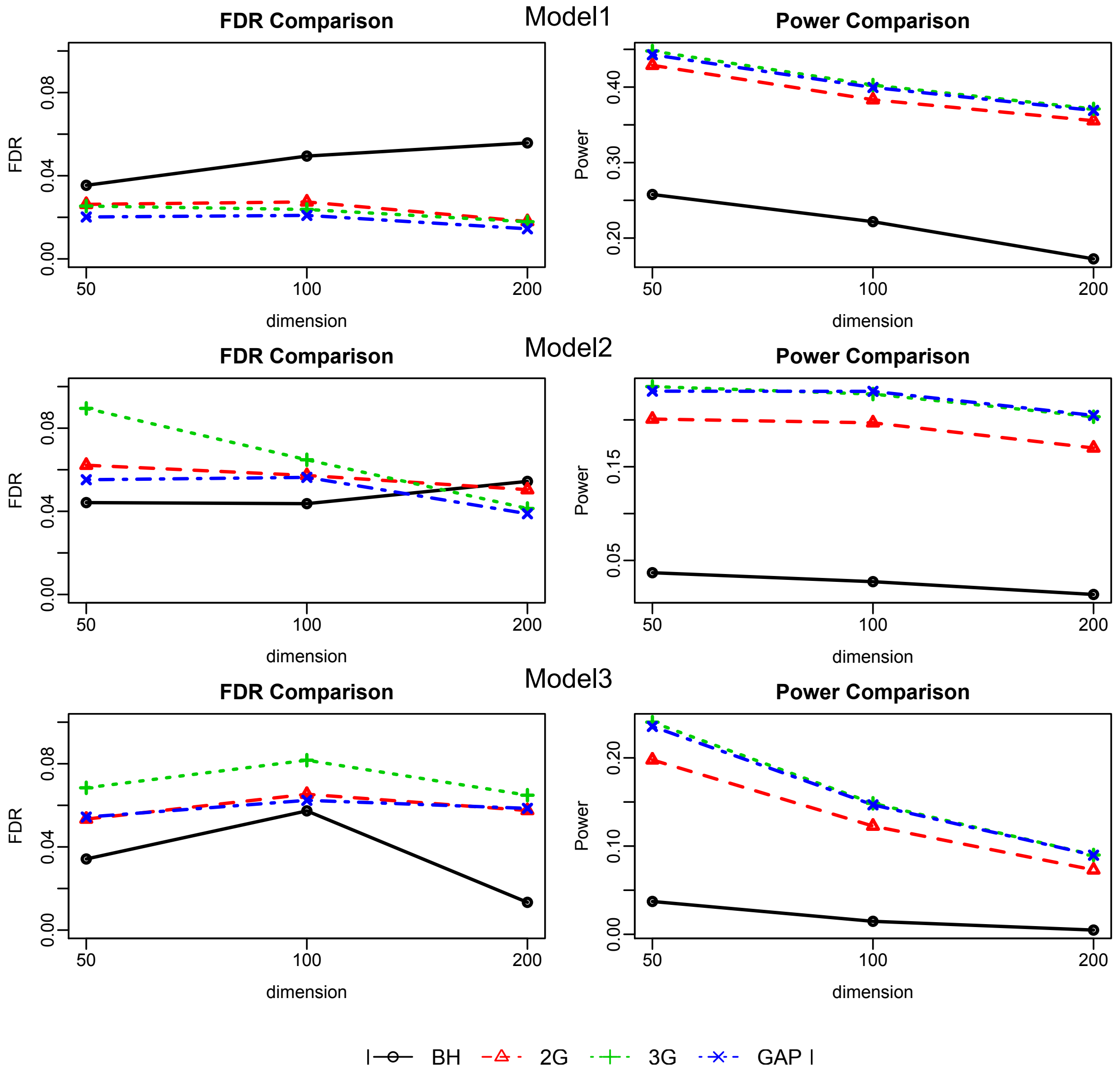

Figure 4: FDR and Power comparisons on Gaussian Graphical Models between BH, 2G, 3G and GAP.

2000; Marchini et al., 2005). However, most existing studies (Nathanson et al., 2001) have only established molecular pathways of pathogenesis for breast cancer, and few have investigated the interactions between genes, within and across pathways, that are associated with breast cancer survival.

Our analysis focuses on 32 pathways related to breast cancer survival (a total of 754 genes) based on the molecular signature database. This leads to $\binom{754}{2} = 283881$ pairs of potential gene-gene interactions. We consider two types of survivors in the study: 78 short term survivors who died within 5 years; and 69 long term survivors who have survived more than 10 years. Previous studies [Segal et al. (2003); Dobra et al. (2004)] revealed that transcriptional regulation of a single gene is generally defined by a small set of regulatory elements; hence we assume that the gene-gene interactions in the selected pathways are sparse, and propose to use an auxiliary sequence to capture the sparsity information in the data. Our goal is to identify gene-gene interactions that have significant changes of magnitude between the two types of survivors; this leads to a two-sample multiple testing problem as formulated in (1.1). We apply BH, 2G and GAP to carry out the analysis.

The BH procedure identifies 6 pairs of genes with significant changes in interaction at the FDR level of 0.1. For the 2G method, we first construct auxiliary statistics using the formulae in Section 4.3.2 and then divide the m pairs of genes into two groups. By setting the same FDR level for both groups, 2G identifies 15 pairs of genes. Finally we apply the GAP procedure by dividing the pairs into three groups and set up the FDR level for each group adaptively based on the non-null proportions. The data-driven cutoffs in Step 1 of GAP are $\lambda_1 = -3.9$ and $\lambda_2 = -1.3$, resulting in three groups with sizes $|\mathcal{G}_1| = 346$, $|\mathcal{G}_2| = 35086$ and $|\mathcal{G}_3| = 248449$, respectively. Group $\mathcal{G}_1$ has the highest non-null proportions and is assigned with the highest weight $w_1 = 176.17$. The GAP procedure selected 6 pairs of genes out of $\mathcal{G}_1$, whereas both BH procedure and 2G did not select any from this group. Group $\mathcal{G}_2$ has the second highest non-null proportions and is assigned the weight of $w_2 = 5.25$. GAP selected 13 pairs of genes from this group, again greater than the number of pairs selected by both 2G and BH. Finally, $\mathcal{G}_3$ has the lowest proportion of non-nulls, and is assigned with the weight of $w_3 = 0.16$. All three methods selected 3 pairs of genes from $\mathcal{G}_3$. In summary, the GAP procedure identifies 22 pairs of significant changes in interactions by combing all three groups.

If we set the FDR level at 0.05, then BH cannot identify any pairs of genes, 2G identifies 7 pairs, and GAP identifies 11 pairs. The above analysis has illustrated that the GAP procedure helps to discover more interactions. Nonetheless it is necessary to point out that more rejections do not always correspond to greater power of making true discoveries. The power gain should be corroborated by carefully designed new biological studies to replicate these findings.

7 Discussion

This paper develops a general framework for information pooling in two-sample sparse inference. The framework is illustrated and applied to different examples with various dependence structures, including testing multivariate normal means, high-dimensional linear regression, differential covariance or correlation matrices, and Gaussian graphical models. It is shown that the GAP procedure, which effectively exploits the auxiliary information on the sparsity structure of the data, controls the FDR at the nominal level and outperforms existing FDR methods in power.

Although the grouping and weighting strategy provides a powerful tool to capture the structural information in the data, the proposed GAP framework has several limitations. First, the grouping step involves the discretization of a continuous variable, which fails to fully utilize the auxiliary information and lead to some information loss. Creating more groups would reduce the information loss. However, the GAP framework cannot handle too many groups due to the increased computational burden (in searching for the optimal cutoffs) and the decreased accuracy of the proportion estimates. The study of the optimal tradeoffs between grouping, computation and estimation is an interesting but complicated problem. Finally, it remains an open issue regarding whether our proposed weights are optimal. Intuitively, the weights only encode the sparsity structure, but other structural or side information, such as prior knowledge, heteroscedasticity, clustering and hierarchical structures may also be helpful in improving the efficiency in large-scale statistical inference. Much research is still needed for developing new strategies that fully capture various auxiliary information alongside the primary data and optimally incorporate such information into existing multiple testing procedures.

References

Basu, P., T. T. Cai, K. Das, and W. Sun (2017). Weighted false discovery rate control in large-scale multiple testing. *J. Am. Statist. Assoc., to appear.*

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B 57*, 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 1165–1188.

Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist. 38*(1), 100–145.

Cai, T. T. and W. Liu (2016). Large-scale multiple testing of correlations. *J. Am. Statist. Assoc. 111*(513), 229–240.

Cai, T. T., W. Liu, and Y. Xia (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Statist. Assoc. 108*, 265–277.

Cai, T. T. and W. Sun (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc. 104*, 1467–1481.

Cai, T. T., W. Sun, and W. Wang (2019). CARS: Covariate assisted ranking and screening for large-scale two-sample inference (with discussion). *J. Roy. Statist. Soc. B*, to appear.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351.

Caspi, A. and T. E. Moffitt (2006). Gene–environment interactions in psychiatry: joining forces with neuroscience. *Nat. Rev. Neurosci. 7*(7), 583–590.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal. 90*(1), 196–212.

Du, L. and C. Zhang (2014). Single-index modulated multiple testing. *Ann. Statist. 42*(4), 1262–1311.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Assoc. 102*(477), 93–103.

Efron, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Stat. 2*, 197–223.

Ferkingstad, E., A. Frigessi, H. Rue, G. Thorleifsson, and A. Kong (2008). Unsupervised empirical bayesian multiple testing with external covariates. *Ann. Appl. Stat. 2*, 714–735.

Genovese, C. R., K. Roeder, and L. Wasserman (2006). False discovery control with p-value weighting. *Biometrika 93*(3), 509–524.

Gill, R., S. Datta, and S. Datta (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics 11*, Article 95.

Hu, J. X., H. Zhao, and H. H. Zhou (2012). False discovery rate control with groups. *J. Am. Statist. Assoc. 105*, 1215–1227.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nat. Rev. Genet. 6*(4), 287–298.

Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Am. Statist. Assoc. 102*(478), 495–506.

Langaas, M., B. H. Lindqvist, and E. Ferkingstad (2005). Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. Roy. Statist. Soc. B 67*(4), 555–572.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist. 41*, 2948–2978.

Liu, W. (2014). Incorporation of sparsity information in large-scale multiple two-sample t tests. *arXiv preprint arXiv:1410.4282*.

Liu, W. and S. Luo (2014). Hypothesis testing for high-dimensional regression models. *Technical Report*.

Marchini, J., P. Donnelly, and L. Cardon (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet. 37*(4), 413–417.

Meinshausen, N., J. Rice, et al. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist. 34*(1), 373–393.

Nathanson, K., R. Wooster, and B. Weber (2001). Breast cancer genetics: what we know and what we need. *Nat. Med. 7*, 552–556.

Olopade, O., T. Grushko, R. Nanda, and D. Huo (2008). Advances in Breast Cancer: Pathways to Personalized Medicine. *Clin. Cancer Res. 14*(24), 7988.

Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist. 30*, 239–257.

Schweder, T. and E. Spjøtvoll (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika 69*(3), 493–502.

Scott, J. G., R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass (2015). False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *J. Am. Statist. Assoc. 110*(510), 459–471.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet. 34*(2), 166–176.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B 64*(3), 479–498.

Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis (2005). Significance analysis of time course microarray experiments. *Proc Natl Acad Sci U S A 102*, 12837–12842.

Sun, W. and T. T. Cai (2009). Large-scale multiple testing under dependence. *J. Roy. Statist. Soc. B 71*(2), 393–424.

Sun, W. and Z. Wei (2011). Large-scale multiple testing for pattern identification, with applications to time-course microarray experiments. *J. Amer. Statist. Assoc. 106*, 73–88.

Tai, Y. C. and T. P. Speed (2006). A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Statist. 34*, 2387–2412.

Xia, Y., T. Cai, and T. T. Cai (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika 102*, 247–266.

Xia, Y., T. Cai, and T. T. Cai (2018). Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Stat. Sinica, 28*(1), 63–92.

Zaïtsev, A. Y. (1987). On the Gaussian approximation of convolutions under multidimensional analogues of SN Bernstein's inequality conditions. *Probab. Theory Rel. 74*, 535–566.

Zerba, K., R. Ferrell, and C. Sing (2000). Complex adaptive systems and human health: the influence of common genotypes of the apolipoprotein E (ApoE) gene polymorphism and age on the relational order within a field of lipid metabolism traits. *Hum. Genet. 107*(5), 466–475.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist. 36*(4), 1567–1594.