Meta-Analysis models with group structure for pleiotropy detection at gene and variant level by using summary statistics from multiple datasets

Pierre-emmanuel Sugier, Yazdan Asgari, Mohammed Sedki, Thérèse Truong, Benoit Liquet

Meta-Analysis models with group structure for pleiotropy detection at gene and variant level by using summary statistics from multiple datasets

Pierre-Emmanuel Sugier[1,2,3,*], Yazdan Asgari[2], Mohammed Sedki[4], Thérèse Truong[2], and Benoit Liquet[1,5]

[1]Laboratoire de Mathématiques et de leurs Applications de Pau, Université de Pau et des Pays de l'Adour, UMR CNRS 5142, E2S-UPPA, France
[2]Team "Exposome and Heredity", Paris-Saclay University, UVSQ, Gustave Roussy, Inserm, CESP, Villejuif, France
[3]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France
[4]Team Psychiatrie du développement et trajectoires, Paris-Saclay University, UVSQ, Gustave Roussy, Inserm, CESP, Villejuif, France
[5]School of Mathematical and Physical Sciences, Macquarie University, Sydney, Australia
[*]Corresponding author: pierre-emmanuel.sugier@u-paris.fr

Abstract

Genome-wide association studies (GWAS) have highlighted the importance of pleiotropy in human diseases, where one gene can impact two or more unrelated traits. Examining shared genetic risk factors across multiple diseases can enhance our understanding of these conditions by pinpointing new genes and biological pathways involved. Furthermore, with an increasing wealth of GWAS summary statistics available to the scientific community, leveraging these findings across multiple phenotypes could unveil novel pleiotropic associations. Existing selection methods examine pleiotropy association one by one at a scale of either the genetic variant or the gene, and though cannot consider all the genetic information at the same time. To address this limitation, we propose a new approach called MPSG (Meta-analysis model adapted for Pleiotropy Selection with Group structure). This method performs a penalised multivariate meta-analysis method adapted for pleiotropy and takes into account the group structure information nested in the data to select relevant variants and genes (or pathways) from all the genetic information. To do so, we implemented an alternating direction method of multipliers (ADMM) algorithm. We compared the performance of the method with other benchmark meta-analysis approaches as GCPBayes, PLACO, and ASSET by considering as inputs different kinds of summary statistics. We provide an application of our method to the identification of potential pleiotropic genes between breast and thyroid cancers.

1 Introduction

Genome-wide association studies (GWASs) that aim to detect associations between observable traits and independently-tested genetic variation across the genome have highlighted that the phenomena of pleiotropy is widely spread in human diseases [Solovieff et al., 2013]. Pleiotropy refers to the phenomenon where one genetic locus affects multiple different or possibly unrelated phenotypes. Pleiotropy can be considered at the scale of the genetic variant, when the variant usually called single nucleotide polymorphism (SNP) is influencing different phenotypes, or at the scale of the gene, when the same gene affects different traits through the same or possibly different SNPs. Identification of these pleiotropic effects may inform common biological mechanisms between different diseases, helping to disentangle their relationships, have implication for personal genomics in disease prevention as genetics tests for one disease could have implications for another disease, and lead to new opportunities for drug development.
Analysing several datasets on multiple phenotypes all together can boost the statistical power for detecting true cross-phenotype associations that pleiotropic effects can cause. Though, this requires having access to individual-level data from multiple samples on different phenotypes that can be particularly challenging in practice. In the other hand, GWASs are providing an increasing amount of summary data on genotype-

phenotype associations, and most of these results are made publicly available to the scientific community. Meta-analysis methods that combine summary statistics and corresponding inferential results across studies offer flexibility to jointly study the effect of genes on multiple phenotypes by exploiting both individual-level data and the large number of accessible summary statistics from GWAS results. These methods improve the statistical power to detect a true association by exploiting more data for which only summary-level data are available. In the classical meta-analysis approach, p-values or effect sizes are combined across multiple studies of the same trait [Evangelou and Ioannidis, 2013]. Under these approaches, only effects in the same direction can be identified. However, and a pleiotropic variant might even have an opposite effect on two traits. Some approaches have been proposed to adapt meta-analysis models for the analysis of pleiotropy at the SNP-level. Several methods have been proposed to extend meta-analysis models to the analysis of pleiotropy at the SNP-level [Majumdar et al., 2018, Bhattacharjee et al., 2012, Ray and Chatterjee, 2020]. Especially, the methods ASSET [Bhattacharjee et al., 2012] and more recently PLACO [Ray and Chatterjee, 2020] have been proposed to answer this problematic in frequentist frameworks. Though, these methods only consider pleiotropy at the SNP-level and test every SNP independently. Some pleiotropy effects involving different SNPs of the gene between studies can be missed.

Incorporating prior biological information in GWAS such as group structure information (gene or pathway) has shown some success in classical GWAS approaches. Recent efforts have been made to develop methods allowing to explore cross-phenotype associations both at the SNP and gene-level [Asgari et al., 2023, Baghfalaki et al., 2021, Sutton et al., 2022]. To our knowledge, only the GCPBayes methods have been proposed to handle summary statistics [Baghfalaki et al., 2021]. GCPBayes approaches allow for pleiotropy selection at both variable (SNPs) and group (genes) levels with multiple diseases in each group. The method performs meta-analysis for each group within a Bayesian meta-analysis framework, adapted for pleiotropy. Such methods can also consider the correlation between genetic variants of the same gene in the analysis. Though, this method only allows each gene to be analysed one by one independently, and the high computational cost of the Bayesian framework can make the analysis of very long genes difficult.

Here, we propose a novel penalised multivariate meta-analysis method adapted for pleiotropy that takes into account the group structure information nested in the data by considering all genetic variables simultaneously. We implement an alternating direction method of multipliers (ADMM) algorithm to perform both regularisation at the variable and group levels. The performance of the novel approach is compared to benchmark gene-level and SNP-level meta-analysis approaches on simulated data by considering different kinds of summary data as inputs. Our method is then applied to the identification of potential pleiotropic genes between breast and thyroid cancer by using summary statistics from pathway candidate genes.

2 Method

2.1 Notation and terminology

The term "summary statistics" refers to the key information about the genetic associations identified in a GWAS without sharing individual-level data. Suppose that we have access to summary statistics information from S independent GWAS. For each study $s = 1, ..., S$, the available information pertains to the estimators $\hat{\beta}_s$ of $\beta_s \in \mathbb{R}^p$, where $\hat{\beta}_s = (\hat{\beta}_{1s}, ..., \hat{\beta}_{ps})^\top$ is typically the minimizer of some loss function, with $\widehat{V}_s$ their corresponding estimated covariance matrices, and p is the number of variables in each dataset. In such a context, the set of available variables can slightly differ between studies. Then, for more generality, we will note p the union of all available variables. For $j = 1, ..., p$, let S_j denote the set of studies which contain the jth covariate in the summary statistics. Let the genetic information be structured into G groups. A group is a set of variables with a biological meaning. For example, this can be a set of SNPs belonging to a gene, or SNPs belonging to a set of genes acting together in the same biological pathway. For $g = 1, ..., G$, let π_g denote the set of m_g covariates belonging to the group g. Then, $p = \sum_{g=1}^{G} m_g$. Considering the genetic information being structured into groups, we can model $\widehat{V}_s$ as a block diagonal matrix of $\widehat{V}_s^{(g)} \equiv \widehat{Cov}(\hat{\beta}_s^{(g)})$, with $g = 1, ..., G$ and $\hat{\beta}_s^{(g)}$ is a subset of the vector $\hat{\beta}_s$ containing only the entries corresponding to variables within group g. When raw data are available, it is possible to obtain $\widehat{V}_s$. In the context of GWAS analysis, each genetic variant is generally tested independently for its association with a disease with some regression techniques by using some adjustment variables. Then, in the context of post-GWAS analysis, summary statistics are usually in the form of $\widehat{\beta}_s$ vectors of regression coefficients of multiple univariate analyses and the corresponding vectors of variance estimators $\widehat{v}_s = (\hat{v}_{1s}, ..., \hat{v}_{ps})^\top$.

In this case, $\widehat{V}_s$ is in the form of a diagonal matrix.

2.2 Inverse-variance estimator for meta-analysis design

Traditional meta-analysis methods propose to analyse each covariate separately. Though, a multivariate meta-analysis version called the inverse-variance estimator $(\sum_{s=1}^{S} \widehat{V}_s^{-1})^{-1}(\sum_{s=1}^{S} \widehat{V}_s^{-1}\hat{\beta}_s)$ has been suggested, which is the minimizer of $\sum_{s=1}^{S}(\hat{\beta}_s - \beta_s)^T \widehat{V}_s^{-1}(\hat{\beta}_s - \beta_s)$ under the constraint that $\beta_1 = \ldots = \beta_S$ [Lin and Zeng, 2010]. Motivated by this estimator, a sparse meta-analysis (SMA) method has been developed to perform variable selection in meta-analysis when raw data are not available [He et al., 2016].

In the context of pleiotropy, different studies are performed on possibly multiple different phenotypes. In contrast to traditional meta-analysis, the true underlying effects for a given genetic variant can vary significantly across studies in terms of magnitude and direction. Hence, we propose an approach to perform group and variable selection in the context of pleiotropy detection by using multivariate meta-analysis that can take into account group information structures in the data.

2.3 Model for group and variable selection

We consider the multivariate inverse-variance estimator without constraint being our objective function as follows:

$$\ell(\boldsymbol{B}) = \sum_{g=1}^{G}\sum_{s=1}^{S}(\hat{\beta}_s^{(g)} - \beta_s^{(g)})^T \widehat{V}_s^{(g)^{-1}}(\hat{\beta}_s^{(g)} - \beta_s^{(g)}), \tag{1}$$

where the matrix $\boldsymbol{B} \in \mathbb{R}^{p\times S}$ consists of the vectors $\beta_s \in \mathbb{R}^p$ stacked column-wise for $s = 1, \ldots, S$. Each β_s is itself partitioned into G groups as $\beta_s = ({\beta_s^{(1)}}^\top, \ldots, {\beta_s^{(G)}}^\top)^\top$.

Our Meta-analysis approach, adapted for Pleiotropy by integrating Sparsity into Groups of variables (MPSG), is based on penalised likelihood maximisation. Using the likelihood form for independent datasets (1), we propose the penalised likelihood estimate:

$$\widehat{\boldsymbol{B}}_{\alpha,\lambda} = \underset{\boldsymbol{B}\in\mathbb{R}^{p\times S}}{\operatorname{argmin}} \left\{\ell(\boldsymbol{B}) + \lambda(1-\alpha)\|\boldsymbol{B}\|_{G_{2,1}} + \lambda\alpha\|\boldsymbol{B}\|_{l_{2,1}}\right\} \tag{2}$$

$$\text{where } \|\boldsymbol{B}\|_{G_{2,1}} = \sum_{g=1}^{G}\sqrt{m_g}\sqrt{\sum_{i\in\pi_g}\sum_{s\in S_i}\beta_{is}^2}$$

$$\text{and } \|\boldsymbol{B}\|_{l_{2,1}} = \sum_{g=1}^{G}\sum_{i\in\pi_g}\sqrt{\sum_{s\in S_i}\beta_{is}^2}$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$ are regularisation parameters weighting a $G_{2,1}$-norm penalty $\|\boldsymbol{B}\|_{G_{2,1}}$ and $l_{2,1}$-norm penalty $\|\boldsymbol{B}\|_{l_{2,1}}$. The parameter λ controls an overall amount of penalisation, while α determines how much penalisation is used for each penalty. The $G_{2,1}$-norm (as in [Wang et al., 2012]) fixes the group structure across studies and allows group selection by encouraging sparsity at the group-level. The $l_{2,1}$-norm allows to select relevant pleiotropic variables within groups. The penalisation matches the penalisation proposed in previous work [Sutton et al., 2022, Wang et al., 2012] but is adapted to a meta-analysis context.

2.4 ADMM algorithm

We propose to fit this model defined in (2) using the alternating direction method of multipliers (ADMM) algorithm [Boyd et al., 2011]. To simplify the notation we define $\lambda_1 = (1-\alpha)\lambda$ and $\lambda_2 = \lambda\alpha$. The ADMM formulation of our optimisation problem is given by

$$\min_{\boldsymbol{B},\boldsymbol{Z}} \left\{\ell(\boldsymbol{B}) \quad + \quad \lambda_1\|\boldsymbol{Z}\|_{G_{2,1}} \quad + \quad \lambda_2\|\boldsymbol{Z}\|_{l_{2,1}}\right\}$$

$$\text{subject to } \boldsymbol{Z} = \boldsymbol{B}$$

Algorithm 1 Calculate $\boldsymbol{Z}^{(t+1)}$

Require: $\lambda > 0$ and $\alpha \in]0, 1[$

$\lambda_1 \Leftarrow (1-\alpha)\lambda$ and $\lambda_2 \Leftarrow \lambda\alpha$
$\boldsymbol{Y}^{(t+1)} \Leftarrow \boldsymbol{B}^{(t+1)} + \boldsymbol{U}^{(t)}$
for $j = 1, \ldots, p$ **do**
  if $[\boldsymbol{Y}^{(t+1)}]_{(j,\cdot)} \leq \lambda_1$ **then**
    $[\boldsymbol{Z}^{(t+1)}]_{(j,\cdot)} \Leftarrow 0$
  else
    $[\boldsymbol{Z}^{(t+1)}]_{(j,\cdot)} \Leftarrow \frac{\|[\boldsymbol{Y}^{(t+1)}]_{(j,\cdot)}\|_F - \lambda_1}{\|[\boldsymbol{Y}^{(t+1)}]_{(j,\cdot)}\|_F}[\boldsymbol{Y}(t+1)]_{(j,\cdot)}$
  end if
end for
for $g = 1, \ldots, G$ **do**
  if $\|[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)}\|_F \leq \lambda_2$ **then**
    $[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)} \Leftarrow 0$
  else
    $[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)} \Leftarrow \frac{\|[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)}\|_F - \lambda_2}{\|[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)}\|_F}[\boldsymbol{Z}^{(t+1)}]_{(\pi_g,\cdot)}$
  end if
end for

where $\boldsymbol{Z} \in \mathbb{R}^{p\times K}$. The augmented Lagrangian introduces an auxiliary variable $\boldsymbol{U}$ with Lagrange multiplier ρ and is given by the following:

$$\mathcal{L}_\rho(\boldsymbol{B}, \boldsymbol{Z}, \mathbf{U}) = \ell(\boldsymbol{B}) + \lambda_1\|\boldsymbol{Z}\|_{G_{2,1}} + \lambda_2\|\boldsymbol{Z}\|_{l_{2,1}} + \frac{\rho}{2}\|\boldsymbol{B} - \boldsymbol{Z} + \boldsymbol{U}\|_F^2 + \frac{\rho}{2}\|\boldsymbol{U}\|_F^2$$

The ADMM algorithm makes the following set of updates:

$$\begin{aligned}
\boldsymbol{B}^{t+1} &= \underset{\boldsymbol{B}\in\mathbb{R}^{p\times K}}{\text{argmin}} \quad \mathcal{L}_\rho(\boldsymbol{B}, \boldsymbol{Z}^{(t)}, \boldsymbol{U}^{(t)}) \\
\boldsymbol{Z}^{t+1} &= \underset{\boldsymbol{Z}\in\mathbb{R}^{p\times K}}{\text{argmin}} \quad \mathcal{L}_\rho(\boldsymbol{B}^{(t+1)}, \boldsymbol{Z}, \boldsymbol{U}^{(t)}) \\
\boldsymbol{U}^{t+1} &= \boldsymbol{U}^{(t)} + \boldsymbol{B}^{(t+1)} - \boldsymbol{Z}^{(t+1)}.
\end{aligned}$$

Each iteration of the algorithm consist of three sub-problems. The first update $\boldsymbol{B}^{(t+1)}$ can be solved analytically. The update $\boldsymbol{U}^{(t+1)}$ is simply a dual variable update. The update $\boldsymbol{Z}^{(t+1)}$ is a convex optimisation problem. Following [Jenatton et al., 2011], the update $\boldsymbol{Z}^{(t+1)}$ can be seen as a conjugation of proximal functions that can be solved sequentially with problems separable in every group. Let $[\boldsymbol{A}]_{(\pi_g,\cdot)}$ denote the rows of a matrix $\boldsymbol{A}$ corresponding to the SNP indices in π_g. the update $\boldsymbol{Z}^{(t+1)}$ consists of the following two loops proposed in algorithm 1.

2.5 Adaptive weights

While penalised approaches allow for the shrinkage of coefficients to zero, they come at the cost of possibly excessive shrinkage of non-zero coefficients. This has motivated several approaches that aim to reduce the effect of shrinkage on non-zero coefficients. Based on our previous work [Sutton et al., 2022], we have incorporated an adaptive lasso approach that assigns weights to each coefficient of lasso and group lasso penalties. Using an appropriate choice for the weights, penalisation for non-zero coefficients can be reduced and these coefficients will suffer less shrinkage. Hence, we proposed weighted versions both for $G_{2,1}$ and ℓ_1-norm and as follows (respectively),

$$\sum_{g=1}^{G}\gamma_g\sqrt{\sum_{i\in\pi_g}\sum_{s\in S_i}\beta_{is}^2}, \quad \text{and} \quad \sum_{g=1}^{G}\sum_{i\in\pi_g}w_i\sqrt{\sum_{s\in S_i}\beta_{is}^2}.$$

Analogous to the similar adaptive group and sparse-group lasso material in the literature [Zou, 2006, Münch et al., 2021], the $G_{2,1}$-norm weights γ_g are taken to be the inverse of the $G_{2,1}$-norm of the OLS

coefficients for $g = 1, ..., G$. Similarly, we take the weights for the $\ell_{2,1}$-norm to be w_i where w_i is chosen as the inverse of the $\ell_{2,1}$-norm applied to the OLS coefficients for $i = 1, ..., p$. That is, we set the weights to be:

$$\gamma_g = \frac{1}{\sqrt{\sum_{i \in \pi_g} \sum_{s \in S_i} \hat{\beta}_{is}^2}}, \quad \text{and} \quad w_i = \frac{1}{\sqrt{\sum_{s \in S_i} \hat{\beta}_{is}^2}}.$$

2.6 Hyper-parameters tuning

The tuning parameters α and λ control the trade-off between model sparsity and model fit. Motivated by the work of Wang and Leng [He et al., 2016], we determine the tuning parameter by a modified information criterion (MIC) as following:

$$MIC_{\alpha,\lambda} = SSE_{\alpha,\lambda} + \sum_{s=1}^{S} q_{s,\alpha,\lambda} \frac{log(n_s)}{n_s}$$

$$\text{where } SSE_{\alpha,\lambda} = \sum_{g=1}^{G} \sum_{s=1}^{S} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\alpha,\lambda}^{(g)})^\top \widehat{V}_s^{(g)^{-1}} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\alpha,\lambda}^{(g)})$$

where $\hat{\beta}_{s,\alpha,\lambda}^{(g)}$ is an element of the solution of (2) for the study s and the group g, under α and λ, and $q_{\alpha,\lambda,s}$ is the total number of nonzero components of $\hat{\beta}_{s,\alpha,\lambda}$. The $SSE_{\alpha,\lambda}$ measures the overall model fit for the S studies, whereas the second part of the $MIC_{\alpha,\lambda}$ measures the model complexity among the S studies.The appendix section A provides consistency results for model selection of the MIC criterion.

3 Simulation studies

In this section, we compare the performance of our proposed method with other state-of-the-art methods GCPBayes, ASSET, and PLACO to detect pleiotropic effects using GWAS summary statistics (Baghfalaki et al. [2021], Bhattacharjee et al. [2012], Ray and Chatterjee [2020]). The main aim of our methods is to detect pleiotropic signals at both variable and group levels. Hence, we want to investigate the performance of our method to detect pleiotropic groups comprising relevant variables, and to select the relevant variables inside pleiotropic groups.

3.1 Design of the simulations

For these simulations, we considered $p = 250$ variables divided into $G = 5$ groups of $m_g = 50$ variables for $S = 2$ studies. We generated summary statistics for both studies *i.e.* beta estimates and their covariance matrix. Performances are evaluated over 50 replications. In each replication, we generated beta estimates for each study β_s, where $s = 1, 2$, considering 60% group sparsity, two out of five groups have a non-zero effect. For groups with non-zero effect, we considered intra-group sparsity (IGS) using the form of the true effects in study 1, $\beta_1^{(g)}$ following the simulation settings in [Baghfalaki et al., 2021] as below:

$$\beta_1^{(g)} = (\beta_{true}.\zeta_{(1-\text{IGS}).\text{m}_\text{g}}^\top, 0_{\text{IGS}.\text{m}_\text{g}}^\top)^\top \tag{3}$$

where $\beta_{true} \in Re$ is the magnitude of the effects, $\zeta_q = (\zeta_1, \cdots, \zeta_q)^\top$ a q-dimensional vector of a discrete random variable which takes the value 1 or -1 with equal probabilities, and 0_q denotes a q-dimensional vector with 0 elements. Then, we set $\beta_2^{(1)} = \beta_1^{(1)}$ and $\beta_2^{(2)} = -\beta_1^{(2)}$. We set IGS = 30% and $\beta_{true} = 0.1$ and 0.15. For each group, the regression coefficients $\hat{\beta}_s^{(g)}$ were simulated using a distributional assumption as follows:

$$\hat{\beta}_s^{(g)} \sim N_{m_g}(\beta_s^{(g)}, V^{(g)}),\ s = 1, 2,\ \ g = 1, \cdots, G, \tag{4}$$

where $V^{(g)}$ is defined as $V^{(g)} = \frac{\sigma^2}{N}(\Sigma^{(g)})^{-1}$ with $\Sigma^{(g)}$ a compound-symmetry matrix of correlation between variables in the group with non-diagonal components ρ, σ^2 a noise calculated such as the signal-to-noise ratio (SNR) in each β_{true} value set up is fixed (SNR $= 0.3$ for $\beta_{true} = 0.1$, and SNR $= 0.6$ for

$\beta_{true} = 0.15$), and N is the number of individuals from which the $\Sigma^{(g)}$ is supposed to come from. SNR is defined by $\frac{{\beta_s}^T \Sigma_s \beta_s}{\sigma^2}$, with Σ_s a block diagonal matrix of $\Sigma_s^{(g)}, g = 1, ..., G$.

We set N to be the same for both studies s. We test different scenarios with $\rho = 0.25, 0.5$ and $N = 1500, 2000, 2500, 3000$. For the subsequent simulation analyses, we considered $\widehat{V}^{(g)} = V^{(g)}, \forall g = 1, \ldots, G$.

3.1.1 Simulation study by varying the distribution of effects in groups

To look at the behaviour of each approach when the distribution of the global effect among variables is different, we performed another simulation study for IGS = 70%, by keeping other parameters unchanged, and doing this only for $\beta_{true} = 0.1$ and SNR $= 0.3$. This scenario corresponds to the same global amount of effect per group, but distributed into fewer variables within the group.

3.1.2 Simulation study with G = 50

As a final simulation study, we wanted to test the performances of each approach with a smaller $\frac{n}{p}$ ratio by adding more non-pleiotropic groups in the data. Hence, we performed a simulation study with G $= 50$ but by keeping the number of pleiotropic groups to 2, resulting in 96% of group sparsity. The other parameters stayed unchanged.

3.1.3 GWAS summary statistics

The summary statistics required for gene-level analysis include the estimated coefficients of regression and their covariance matrix. However, in the GWAS context, the covariance matrix between estimates may not be available but only include the regression coefficients and their standard errors. Although the complete summary data can be approximated by using some reference panels in this context, this is not always possible. Thus, we considered two important scenarios in order to investigate the possible loss of power by not considering the intra-group covariance between beta estimates.

In the first scenario, we considered the general form of multivariate summary statistics with the corresponding non-diagonal covariance matrices that we refer to as (ND), generated as presented in the 3.1 section. In the second scenario, we considered the situation when only the diagonal of the covariance matrix is known. Hence, the covariance matrix considered as inputs for the methods is the diagonal of V, that we refer to as (D).

3.1.4 Selection of pleiotropic signals

We first applied the MPSG method to the simulated data over different scenarios. The main aim of our method is to detect pleiotropic signals at both variable and group levels. A variable is selected as pleiotropic by MPSG if its components for both studies are non-zero in the final model determined through the MIC. A group is selected as pleiotropic if at least one variable in the group is selected as pleiotropic. In this simulation study, we decided to perform every analyses by using both MPSG with (MPSG-W) and without adaptive weights (MPSG), to see how the use of these weights can affect the method.

We used three other methods to compare the performance of MPSG to detect pleiotropy signals with benchmark approaches considering summary statistics. We considered the Bayesian meta-analysis method GCPBayes [Baghfalaki et al., 2021] also designed for the detection of pleiotropy at group and variable levels. We considered the DS function as this is the best scalable GCPBayes method and the one that is used in practice for the analysis of genome-wide datasets [Asgari et al., 2023]. We evaluated variable-level pleiotropy detection by comparing our method to widely used single-SNP meta-analysis approaches such as ASSET [Bhattacharjee et al., 2012] and PLACO [Ray and Chatterjee, 2020].

ASSET is a frequentist method that extends standard fixed effects meta-analysis by considering the effects in each analysis to be either in the same direction or possibly opposite directions allowing the detection of opposite pleiotropic effects. This approach uses summary statistics and does not take into account group structure. The outputs of this approach are a p-value of global association and an optimal subset of non-null studies that are associated with each SNP. To consider a variable as pleiotropic, we selected only variables that remained significant at the 5% level after an FDR correction for both studies.

PLACO is a frequentist method that differs from ASSET by testing the null hypothesis based on the product of the Z-statistics of the genetic variants across two studies and derive a null distribution of the test statistic

in the form of a mixture distribution that allows for fractions of variants to be associated with none or only one of the traits. Similarly to ASSET, we choose to consider a variable as pleiotropic by selecting only variables that remained significant at the 5% level after a FDR correction.

GCPBayes is a Bayesian statistical framework based on Monte Carlo Markov chain (MCMC) through Gibbs sampling. The DS method models a multivariate Dirac spike and slab priors for each group. Groups with both $\theta \geq 0.5$ and a local Bayesian false discovery rate lower than 0.05 were selected as pleiotropic (see [Baghfalaki et al., 2021] for more details). For selected groups, we considered non-zero variables by using a 95% credible interval (CI). Variables with non-zero effects in both studies were selected as pleiotropic.

3.2 Simulation results

We evaluated the performance of all the approaches of interest by considering the Matthews correlation coefficient (MCC) [Chicco and Jurman, 2020] as the first measure. The detail of true positive rate (TPR) and true negative rate (TNR) is also considered.

3.2.1 Group-level pleiotropy

The results of the MCC for pleiotropy at group-level are reported in Tables 1 and 3. The detail of TPR and TNR are also reported in Appendix B (Supplementary tables 7, 9, 10, and 13).

As expected, both three approaches show better results in the context of full consideration of the covariance matrix of the summary statistics (ND) than with diagonal matrix (D) only. Both three methods show similar general tendencies in terms of performances when the intra-group correlation is moderated ($\rho = 0.25$) or high ($\rho = 0.50$), but all approaches are more affected by lack of power when only the diagonal of the covariance matrix is available (D).

When IGS $= 30\%$, DS shows better performance than MPSG in every scenario (see Table 1). In particular, MPSG can lack of power when N is small as TPR can be low. The addition of adaptive weights to our approach brings more power to select pleiotropic groups without lowering the TNR, hence the method shows better performances with than without adaptive weights.

When IGS $= 70\%$, MPSG-W reaches similar level of performances than DS. In particular, when N is small, MPSG-W shows slightly better results than DS in the (D) context, and slightly worse in the (ND) context. This gap slightly increases both ways when IGS is higher.

The simulation study with $G = 50$ by adding groups with no effects showed similar results (see supplementary tables 14, 17, 19).

3.2.2 Variable-level pleiotropy

The results of the MCC for pleiotropy at variable-level are reported in Tables 2 and 4. The details of TPR and TNR are also reported in Appendix B (Supplementary tables 8, 11, 12, and 15).

As for group-level pleiotropy, both MPSG approaches show better results for (ND) summary statistics than with (D) summary statistics only. Though, DS (ND) and DS(D) have really similar performances when IGS is high.

DS (ND) and MPSG-W (ND) performs always better than SNP-level methods ASSET and PLACO. DS (D) also performs better than ASSET and PLACO when IGS $= 30\%$, but PLACO performs better or slightly the same when IGS $= 70\%$.

MPSG-W (ND) shows the best performances for variable-level pleiotropy selection in always all scenarios, or is very similar to DS (ND). MPSG-W (D) can though struggle in terms of TPR, and so in MCC when N is small, that can result from the fact the method is not capable of selecting any true groups in such scenarios. In scenarios where MPSG-W (D) selects groups the right way, it performs similarly to PLACO and DS (D). Hence, MPSG-W has the best performance when (ND) summary statistics are available and still seems to perform well in selecting the right pleiotropic variables inside selected groups when only (D) summary statistics can be used. It is worth noting that if most of the time MPSG-W outperforms MPSG, MPSG(ND) shows better results than MPSG-W(ND) in the two scenarios with IGS $= 30\%$, $\rho = 0.5$, and N $= 1.5$k. MPSG(ND) still outperforms DS(ND) most of the time.

The simulation study with $G = 50$ by adding groups with no effects showed similar results (see supplementary tables 16, 18, 20).

Table 1: Results of simulations for group pleiotropy with $S = 2$ for IGS=30%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages. DS: Dirac Spike function of GCPBayes; MPSG: MPSG without adaptive weights; MPSG-W: MPSG with adaptive weights.

N		DS	MPSG	MPSG-W
$\rho = 0.25.\ \beta = 0.1$				
1.5k	ND	1 (0)	0.37(0.48)	0.82(0.35)
	D	0.96(0.17)	0(0)	0.19(0.34)
2k	ND	1(0)	0.65(0.47)	0.99(0.05)
	D	0.99(0.05)	0.05(0.22)	0.40(0.48)
2.5k	ND	1(0)	1(0)	0.99(0.05)
	D	0.99(0.05)	0.35(0.48)	0.63(0.47)
3k	ND	1(0)	0.98(0.08)	0.99(0.05)
	D	1(0)	0.51(0.5)	0.77(0.42)
$\rho = 0.25.\ \beta = 0.15$				
1.5k	ND	1(0)	0.99(0.07)	1(0)
	D	1(0)	0.61(0.49)	0.77(0.42)
2k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.77(0.42)	0.90(0.3)
2.5k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.87(0.33)	0.94(0.24)
3k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	0.92(0.27)	0.97(0.15)
$\rho = 0.50.\ \beta = 0.1$				
1.5k	ND	1(0)	0.69(0.46)	0.94(0.21)
	D	0.83(0.34)	0(0)	0.15(0.32)
2k	ND	1(0)	0.82(0.39)	1(0)
	D	0.93(0.19)	0.01(0.09)	0.3(0.45)
2.5k	ND	1(0)	0.98(0.14)	1(0)
	D	0.96(0.16)	0.23(0.42)	0.35(0.48)
3k	ND	1(0)	0.99(0.05)	1(0)
	D	0.96(0.16)	0.33(0.47)	0.42(0.48)
$\rho = 0.50.\ \beta = 0.15$				
1.5k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	0.33(0.47)	0.54(0.49)
2k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	0.49(0.5)	0.66(0.46)
2.5k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.58(0.5)	0.75(0.43)
3k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.74(0.44)	0.77(0.42)

4 Real data analysis of breast and thyroid cancers

4.1 Study population

To first validate its utility, we applied MPSG to the analysis of the pleiotropy between breast and thyroid cancers. These two cancers are both more frequent in women and are influenced by hormonal and reproductive factors.

Individuals diagnosed with breast cancer are more likely to develop thyroid cancer as a secondary malignancy than patients diagnosed with other cancer types, and vice-versa [Nielsen et al., 2016]. This association does not seem to be explained by the consequences of cancer treatment or increased surveillance, but rather suggests common etiologies or common genetic mechanisms for the two diseases. Thyroid and

Table 2: Results of simulations for variable pleiotropy with $S = 2$ for IGS=30%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages. DS: Dirac Spike function of GCPBayes; MPSG: MPSG without adaptive weights; MPSG-W: MPSG with adaptive weights.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	0.46(0.15)	0.22(0.29)	0.39(0.23)	0.38(0.16)	0.28(0.17)
	D	0.44(0.17)	0(0)	0.08(0.2)		
2k	ND	0.59(0.16)	0.44(0.34)	0.69(0.24)	0.50(0.17)	0.44(0.19)
	D	0.58(0.18)	0.06(0.22)	0.34(0.45)		
2.5k	ND	0.69(0.17)	0.83(0.13)	0.82(0.21)	0.57(0.18)	0.56(0.21)
	D	0.69(0.18)	0.34(0.46)	0.57(0.47)		
3k	ND	0.77(0.16)	0.89(0.09)	0.89(0.16)	0.63(0.15)	0.67(0.22)
	D	0.77(0.16)	0.49(0.48)	0.74(0.41)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	0.77(0.16)	0.89(0.09)	0.9(0.14)	0.63(0.15)	0.67(0.22)
	D	0.76(0.17)	0.58(0.46)	0.74(0.41)		
2k	ND	0.87(0.13)	0.91(0.07)	0.95(0.08)	0.71(0.11)	0.82(0.18)
	D	0.87(0.13)	0.74(0.4)	0.87(0.3)		
2.5k	ND	0.93(0.09)	0.92(0.06)	0.97(0.07)	0.75(0.09)	0.9(0.14)
	D	0.93(0.1)	0.83(0.31)	0.92(0.24)		
3k	ND	0.96(0.07)	0.94(0.04)	0.98(0.02)	0.77(0.07)	0.95(0.1)
	D	0.96(0.08)	0.87(0.26)	0.96(0.14)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	0.32(0.21)	0.38(0.26)	0.35(0.17)	0.26(0.21)	0.19(0.19)
	D	0.31(0.23)	0(0)	0.07(0.2)		
2k	ND	0.43(0.26)	0.51(0.27)	0.56(0.25)	0.36(0.23)	0.29(0.25)
	D	0.42(0.28)	0.02(0.13)	0.27(0.42)		
2.5k	ND	0.53(0.26)	0.7(0.18)	0.62(0.26)	0.45(0.23)	0.38(0.29)
	D	0.52(0.27)	0.22(0.4)	0.33(0.46)		
3k	ND	0.61(0.26)	0.75(0.15)	0.68(0.25)	0.50(0.23)	0.47(0.32)
	D	0.60(0.28)	0.32(0.45)	0.38(0.47)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	0.60(0.27)	0.77(0.15)	0.71(0.25)	0.50(0.23)	0.47(0.32)
	D	0.59(0.28)	0.32(0.45)	0.49(0.48)		
2k	ND	0.71(0.25)	0.81(0.14)	0.80(0.23)	0.58(0.22)	0.61(0.32)
	D	0.70(0.27)	0.47(0.47)	0.61(0.47)		
2.5k	ND	0.78(0.23)	0.88(0.11)	0.87(0.19)	0.63(0.22)	0.70(0.3)
	D	0.77(0.24)	0.55(0.47)	0.72(0.43)		
3k	ND	0.83(0.21)	0.89(0.09)	0.89(0.17)	0.67(0.19)	0.77(0.28)
	D	0.83(0.22)	0.70(0.42)	0.75(0.42)		

breast cancers share some similarities in their biology: both are more frequent in women, and are influenced by hormonal and reproductive factors.

To investigate pleiotropy between these two distinct cancers, we used individual level data from the CECILE study, a French population-based case-control study on breast cancer (1,125 cases and 1,172 controls), and from the French studies included in the EPITHYR consortium on thyroid cancer (CATHY, Young-thyr, and E3N studies totalizing 1,129 women cases and 1,174 women controls) (Truong et al. [2021]). The study designs, quality control, and specific data cleaning were described in detail previously [Clavel-Chapelon et al., 2010, Baghfalaki et al., 2021, Cordina-Duverger et al., 2017, Xhaard et al., 2015]. Only women of European ancestry were kept for the analyses. To illustrate our methods, we selected variants from genes included in 10 candidate pathways with only few overlapping SNPs. For each pair of SNPs belonging to the same genes in extremely high correlation ($r^2 > 0.98$), one SNP was removed. Then, only SNPs belonging to non overlapping groups (genes and pathways) were selected. At the end of the quality filtering process, the two datasets included the same panel of 3,766 SNPs within 331 genes (see table 5).

Table 3: Results of simulations for group pleiotropy with $S = 2$ for IGS=70%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages. DS: Dirac Spike function of GCPBayes; MPSG: MPSG without adaptive weights; MPSG-W: MPSG with adaptive weights.

N		DS	MPSG	MPSG-W
$\rho = 0.25.\ \beta = 0.1$				
1k	ND	0.96(0.12)	0.8(0.36)	0.92(0.15)
	D	0.69(0.39)	0.27(0.43)	0.72(0.4)
1.5k	ND	1(0)	0.93(0.15)	0.94(0.13)
	D	0.86(0.33)	0.64(0.45)	0.84(0.32)
2k	ND	1(0)	0.92(0.15)	0.97(0.1)
	D	0.93(0.22)	0.75(0.38)	0.9(0.23)
2.5k	ND	1(0)	0.92(0.15)	0.98(0.08)
	D	0.99(0.05)	0.87(0.27)	0.97(0.09)
3k	ND	1(0)	0.94(0.15)	0.98(0.08)
	D	1(0)	0.9(0.17)	0.99(0.07)
$\rho = 0.50.\ \beta = 0.1$				
1k	ND	0.98(0.08)	0.89(0.26)	0.92(0.17)
	D	0.44(0.48)	0.21(0.4)	0.52(0.47)
1.5k	ND	1(0)	0.95(0.13)	0.96(0.12)
	D	0.66(0.44)	0.4(0.47)	0.67(0.44)
2k	ND	1(0)	0.94(0.15)	0.99(0.07)
	D	0.77(0.39)	0.54(0.47)	0.78(0.39)
2.5k	ND	1(0)	0.93(0.14)	0.97(0.09)
	D	0.82(0.35)	0.55(0.46)	0.83(0.33)
3k	ND	1(0)	0.93(0.15)	0.98(0.08)
	D	0.85(0.35)	0.69(0.42)	0.84(0.33)

Table 4: Results of simulations for variable pleiotropy with $S = 2$ for IGS=70%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages. DS: Dirac Spike function of GCPBayes; MPSG: MPSG without adaptive weights; MPSG-W: MPSG with adaptive weights.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1k	ND	0.67(0.18)	0.71(0.31)	0.85(0.12)	0.55(0.18)	0.67(0.21)
	D	0.54(0.34)	0.28(0.43)	0.63(0.4)		
1.5k	ND	0.86(0.13)	0.91(0.07)	0.94(0.07)	0.69(0.14)	0.85(0.15)
	D	0.76(0.31)	0.67(0.44)	0.82(0.34)		
2k	ND	0.94(0.08)	0.94(0.05)	0.97(0.03)	0.75(0.11)	0.93(0.1)
	D	0.88(0.23)	0.78(0.37)	0.92(0.22)		
2.5k	ND	0.97(0.05)	0.94(0.04)	0.98(0.02)	0.79(0.08)	0.97(0.05)
	D	0.96(0.08)	0.89(0.23)	0.98(0.02)		
3k	ND	0.98(0.03)	0.94(0.04)	0.99(0.02)	0.81(0.07)	0.98(0.03)
	D	0.98(0.03)	0.95(0.03)	0.98(0.02)		
$\rho = 0.50, \beta = 0.1$						
1k	ND	0.53(0.28)	0.67(0.24)	0.73(0.2)	0.42(0.23)	0.5(0.31)
	D	0.35(0.4)	0.21(0.4)	0.45(0.44)		
1.5k	ND	0.72(0.26)	0.83(0.13)	0.81(0.19)	0.57(0.24)	0.67(0.32)
	D	0.57(0.41)	0.42(0.48)	0.61(0.45)		
2k	ND	0.79(0.24)	0.87(0.11)	0.88(0.15)	0.64(0.23)	0.77(0.28)
	D	0.68(0.38)	0.57(0.47)	0.7(0.41)		
2.5k	ND	0.85(0.2)	0.89(0.1)	0.91(0.12)	0.7(0.19)	0.85(0.22)
	D	0.75(0.36)	0.59(0.47)	0.8(0.36)		
3k	ND	0.9(0.16)	0.91(0.08)	0.93(0.11)	0.73(0.17)	0.88(0.18)
	D	0.81(0.35)	0.72(0.41)	0.81(0.36)		

Table 5: Non-overlapping pathway chosen for the study

Pathway	Description	#Gene	#SNP
F_$obesity$	Obesity and obesity-related phenotypes	48	857
F_DNA	DNA repair	88	610
F_$circadian$	Circadian Rhythm	23	559
F_$xeno$	Xenobiotics metabolism	68	531
F_pub_he2010_4	Precocious or delayed puberty	16	329
F_$cell_cycle$	Cell cycle	19	249
F_$tobacco_hsa00760$	Nicotinate and nicotinamide metabolism	23	229
F_$inflammatory$	Inflammatory response	26	182
F_$oglyc_hsa00511$	Other glycan degradation	15	111
F_$folate$	Folate metabolism	5	50

4.2 Statistical analysis

To perform pleiotropy analysis, we used the same summary statistics for any meta-analysis methods. These summary statistics have been calculated previously in each dataset and for each gene (*i.e.* the regression coefficients and their covariance matrix for each gene) from a Bayesian hierarchical GLM by using Gaussian priors from the `BhGLM` package [Baghfalaki et al., 2021].

We run a pleiotropy analysis both at gene and SNP levels by using MPSG with and without adaptive-weights, for comparison. We also compared the results from our methods with previously published pleiotropy analysis results on the same data using the Bayesian meta-analysis model called GCPBayes at SNP and gene level [Baghfalaki et al., 2021]. We preferentially considered the DS function as it is the one used in the pipeline for genome-wide analysis of pleiotropy [Asgari et al., 2023]. A gene was then considered as pleiotropic with GCPBayes if $\theta > 0.5$. We also considered SNP-level pleiotropy analysis results by using the method ASSET. These results used an FDR correction for multiple testing. As only pleiotropic effects were explored, we only considered SNPs detected in both datasets. For better SNP-level results comparison, we also performed the most up-to-date SNP-level analysis method PLACO on these data. To get comparative results with ASSET, we applied an FDR to correct for multiple testing to consider pleiotropic SNPs by using PLACO.

4.3 Application results

Previous SNP-level analysis with ASSET did not detect any significant SNPs after a multiple testing correction. As for ASSET, SNP-level pleiotropy analysis using PLACO did not detect any SNP after a multiple testing correction.

Gene-level analysis using MPSG with adaptive weights selected 15 genes as pleiotropic. These results have been reported in Table 6. MPSG was able to retrieve the 9 previously selected genes by GCPBayes [Baghfalaki et al., 2021]. The 9 common genes were also detected with MPSG without the use of adaptive weights.

MPSG also selected 6 additional genes as pleiotropic. As the previously reported *RORA* (RAR-related orphan receptor A) gene located in chromosome 15, MPSG retrieved another gene of the ROR-Family Receptor Tyrosine Kinases, *RORB* (RAR-related orphan receptor B) located on chromosome 9. Both proteins encoded by these genes help to regulate some genes involved in circadian rhythm. *IL18RAP* (interleukin 18 receptor accessory protein), and *UGT1A8* (UDP glucuronosyltransferase family 1 member A8) , are both located in chromosome 2. *GLB1* (GLB1 galactosidase beta 1), *RPA3* (replication protein A3), and *CASC1*, an alias for *DNAI7* (dynein axonemal intermediate chain 7), are respectively located in the chromosomes 3, 7, and 12. *CASC1* is in the same genomic region as the previously reported gene *BCAT1* (branched-chain amino acid transaminase 1).

One additional gene was selected as pleiotropic by using this strategy, *PARP2* (poly(ADP-ribose) polymerase 2) located on chromosome 14 which has been previously identified as potentially pleiotropic gene on the same datasets by using raw data ([Sutton et al., 2022]). *PARP2* also obtained a substantial probability of being pleiotropic with GCPBayes ($\theta = 0.24$) but did not reach the selection threshold.

MPSG was able to select genes with different numbers of SNPs, from short genes as *IL18RAP* with 13 SNPs

to longer genes as *RORA* with 281 SNPs. The method was also able to select numerous SNPs as potentially pleiotropic inside the selected genes, with also different proportions of selected SNPs in selected genes: from 8% of selected SNPs for RPA3 to almost 67% for *RORA*.

In particular, MPSG selected many more SNPs than GCPBayes in genes that were selected in common. GCPBayes results from the DS approach selected only one pleiotropic SNP for each of *NEGR1*, *NPAS2*, and *EGFR*, were MPSG selected respectively 33, 13, and 28 SNPs. The HS approach of GCBayes, that is most suitable for SNP selection than DS, only selected 2 more pleiotropic SNPs than DS in *NEGR1*, and also found an additional pleiotropic SNP in *EBF2*, for which MPSG selected 11 SNPs.

Table 6: **Pleiotropic genes detected by MPSG with adaptive-weights, and corresponding θ value for GCPBayes**. The number of selected SNPs for each gene is represented in parentheses. Genes in bold were detected by every gene-level meta-analysis approaches. The first column indicates the genetic locus, starting with the chromosome number, and followed by the arm of the chromosome (q: long, p: short) and the exact position of the locus from the centromere.

Locus	#SNPs	Gene (#SNPs)	GCPBayes θ (#SNPs)
1p31.1	93	***NEGR1*** (33)	1.00 (1)
1p22.1	77	***TGFBR3*** (19)	1.00
2q11.2	60	***NPAS2*** (13)	1.00 (1)
2q12.1	13	*IL18RAP* (3)	8.00×10^{-4}
2q37.1	58	*UGT1A8* (8)	0.0024
3p22.3	22	*GLB1* (6)	0.0043
7p21.3	25	*RPA3* (2)	4.05×10^{-8}
7p11.2	103	***EGFR*** (28)	1.00 (1)
8p21.2	60	***EBF2*** (11)	1.00
9p21.2-p21.1	197	***LRRN6C*** (101)	1.00
9q21.13	34	*RORB* (5)	5.56×10^{-14}
12p12.1	50	***BCAT1*** (13)	1.00
12p12.1	17	*CASC1* (3)	0.0023
15q22.2	281	***RORA*** (188)	0.99
16q12.2	122	***FTO*** (43)	1.00

5 Computational time

We carry out a short simulation study using the current implementation of the MPSG R package to explore the computational time of our method and compare it with the other state-of-the-art gene-level pleiotropy method GCBPayes. To have comparative results, we used the DS function of the GCPBayes package that is more scalable in terms of computation than the HS method and so more suitable to a pleiotropy analysis at a genome-wide scale.

This simulation study has been conducted on a server with Intel® Xeon® Processor E7-8860 v4 2.20 GHz, 516 GB RAM, CentOS 7.9.2009. Parallelisations have been implemented to perform the calculations of the MPSG method on a 2D-grid of hyperparameters and can be used. We choose to simulate data by considering different sets with 9 groups and to run MPSG and GCPBayes for these 9 groups by using parallelizations over 9 cores (to match with the number of α values tested).

We considered S = 2 studies for all simulations. As the calculation times of the methods can be affected by the size of the groups considered, we chose to vary the size of the genes for each simulation by considering m_g = 50, 100, 200, and 500. Groups were generated by using the same procedure used in the simulation section, with the following parameters: GS = $\frac{2}{3}$, IGS = $\frac{1}{2}$, N = 2500, ρ = 0.25, β_{true} = 0.1, SNR = 0.3. The computational time is also dependant on the number of hyperparameters tested for MPSG, and the number of iterations of the Gibbs sampler for DS. For DS, we set the number of iterations to $20,000$ with $10,000$ a burn-in, as performed in [Baghfalaki et al., 2021]. For MPSG, we used a 2D-grid of 9 α values and 30 λ values to calculate the best model with the MIC. We performed these simulations by using both MPSG with and without adaptive weights to compare both performances.

We used the microbenchmark function from the microbenchmark R package to compare the performances of both methods, by performing each simulation 10 times. Figure 1 presents the boxplots of the calcula-

tion times for each of the three approaches on batches of 10 simulations according to the different group lengths.

As expected, the computational time increases when the number of variables within a group increases. The computational time is shorter for MPSG without the use of adaptive weights in every scenario. GCPBayes by using the DS function has the longest computational time in every scenario, and the gap with both MPSG approaches is increasing greatly with the gene lengths, ranging from twice as long as MPSG without adaptive weights for m_g = 50 to more than 20 times for m_g = 500. It can also be noted that the incorporation of adaptive weights into the MPSG method increases the average calculation time by 10 to 40% depending on the scenarios.

Results of these simulations show that MPSG can allow to compute fast gene-level pleiotropy analysis, with a significant improvement over GCPBayes, especially in genome-wide contexts where some genes with a very high number of SNPs are present. This should remain true even by increasing significantly the number of hyper-parameters tested for the recovery of the best model by using MPSG.

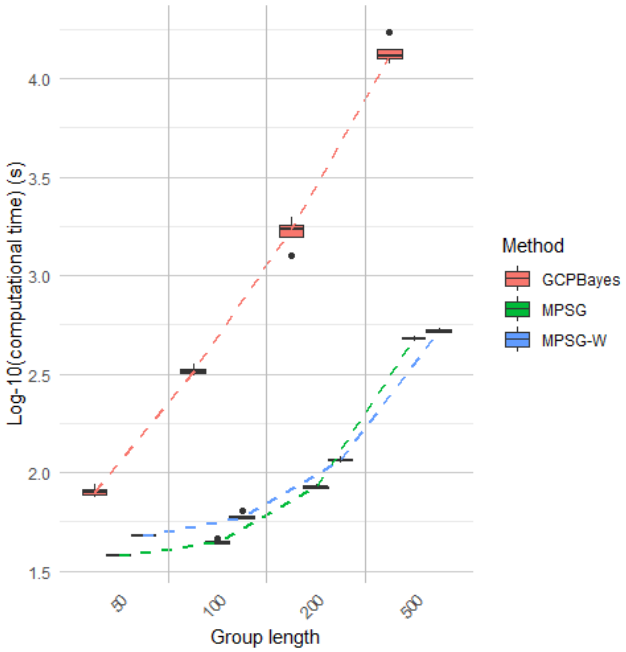

Figure 1: Computational times over group length for both GCPBayes, MPSG, and MPSG with adaptive-weights methods. Computational time is represented in the $\log_{10}$ scale. Dash lines show median values in every scenario for each method.

6 Discussion

Meta-analysis methods adapted for pleiotropy that combine summary statistics and corresponding inferential results across studies offer flexibility to jointly study the effect of genetic variants on multiple phenotypes by exploiting both individual level data and a large number of accessible summary statistics from GWAS results. Then, with the high number of covariates in GWAS data it is crucial to select the most relevant variables to enhance model accuracy and to provide better interpretability of the results. Furthermore, considering known biological structures of the data as genes or pathways, can provide more statistical power to identify relevant pleiotropic signals and help in the interpretation of the results. We present a novel feature selection meta-analysis method adapted for pleiotropy detection at the gene (or pathway) and the SNP-level in GWAS data. MPSG uses penalised likelihood, exploiting different penalties, to induce structured sparsity at a the gene and the SNP-level. This method takes into account the effect of linkage disequilibrium in groups by incorporating known group structures, considering all the pool of structured data together in a single comprehensive analysis of multiple traits. It takes into account heterogeneity in the size and direction of the genetic effects across traits. The method uses an ADMM algorithm to reach convergence and a MIC is then computed to select the best model according to two hyperparameters for group and variable regularization along studies.

We applied our method to the identification of potential pleiotropic genes between breast and thyroid cancer by using summary statistics from candidate pathways data. MPSG retrieved the 9 genes previously identified by GCPBayes and allowed the identification of 6 additional potentially pleiotropic genes between breast and thyroid cancers.

We have compared the performance of the novel approach with benchmark gene-level and SNP-level pleiotropy-based meta-analysis approaches through simulated studies by considering different kinds of summary data as inputs. Simulation study shown that the use of adaptive weights increased the performances of the method in almost all scenarios. It has been also shown that MPSG has slightly lower performances in detecting pleiotropy at the gene-level than the benchmark approach GCPBayes. Though, these comparative results were dependent on the underlying distribution of the effects in groups, being more favorable to MPSG in scenarios with less causal SNPs and a higher correlation in relevant groups, which is very likely in a real context. MPSG also has better performances than GCPBayes in detecting pleiotropy at SNP-level in selected groups. Finally, group-level methods MPSG and GCPBayes have shown better performances than SNP-level only methods, especially when full summary statistics were available. On the other hand, this was not always true when only summary statistics from GWAS results (diagonal) were available. Notably, MPSG's performance collapses more than that of GCPBayes in this case. This shows the importance of summary statistics considerations. When individual data are available, the choice of the first stage analysis has to be made carefully to build proper multivariate summary statistics. When only GWAS summary statistics are available, some extra work can be performed to better approximate full summary statistics considering the linkage disequilibrium structure from a correct reference panel (https://pan.ukbb.broadinstitute.org/). Recent works have been made to impute full summary statistics when the empirical covariance of the covariate-response pair is not available, to detect biologically causal variants [Chen et al., 2024, He et al., 2024].

It should be noted that taking into account the group structure requires labeling variables into groups. We choose to annotate SNPs to genes according to their genetic coordinates as it is commonly used in such a context. Though, it is well known that most of the identified SNPs by GWAS on complex diseases such as cancers are located in non-coding regions [Liu et al., 2020]. Hence, mapping trait-associated SNPs to their nearest gene can fail to identify the functional gene [Smemo et al., 2014]. As a future work, it could be of interest to rather consider functional SNPs-to-genes annotations by integrating external others "omics" data through the analysis with MPSG, as it has been proposed for several methods in other contexts [Gerring et al., 2021]. For example, this could be handled by working on some extension of adaptive weights.

We had access to two candidate pathway datasets and then considered the analysis of pleiotropic genes and SNPs between two studies, but MPSG can handle more than two studies in the same analysis. Though, the selection of a pleiotropic gene in such a context could have different meanings and then lead to interpretation difficulties. For example, only two out of three studies could be relevant for a signal, but cannot be highlighted. Hence, future work could also consider an additional regularization structure to allow for study selection in selected groups.

7 Conclusion

We present a novel feature selection meta-analysis method at group and variable levels adapted for pleiotropy detection in genome-wide data. Our method is way faster than the state-of-the-art GCPBayes approach with almost as good performances in detecting pleiotropy at the gene-level when adaptive weights and proper inputs are considered, and shows better performances at the SNP-level inside selected genes. The methods have been implemented in a user-friendly R statistical package called "MPSG", available at https://github.com/PESugier/MPSG.

As a meta-analysis method, it offers the flexibility to be used on any kind of response variable in GWAS data, and can also be useful for the detection of pleiotropic signals in other kinds of structured data as for example expression data by considering pathways as a group of genes. This method could be extended to study selection in pleiotropic genes, and the integration of multiple external omics data for a better consideration of functional biology in SNPs-to-genes annotations.

8 Competing interests

No competing interest is declared.

9 Author contributions statement

P-E.S. and B.L. designed the novel method. P-E.S. implemented the methods, ran the simulation study and performed the analysis on real data. P-E.S. and B.L. interpreted the simulation results. T.T. provided the real datasets and participated in their interpretation with Y.A. All authors wrote the manuscript. All authors read and approved the final manuscript. P-E.S. developed the code and maintains the package.

10 Acknowledgments

The authors acknowledge Pascal Guénel (PI of CECILE and CATHY studies) and Florent de Vathaire (PI of Young-thyr study) for providing the datasets for the application. The INSERM and Aviesan ITMO cancer are acknowledged for their support of "Advanced Machine Learning Algorithms for leveraging Pleiotropy effect project". This work is also part of the Inserm Cross-Cutting Project GOLD.

References

Y. Asgari, P.-E. Sugier, T. Baghfalaki, E. Lucotte, M. Karimi, M. Sedki, A. Ngo, B. Liquet, and T. Truong. Gcp-bayes pipeline: a tool for exploring pleiotropy at the gene level. *NAR Genomics and Bioinformatics*, 5(3): lqad065, 2023.

T. Baghfalaki, P.-E. Sugier, T. Truong, A. N. Pettitt, K. Mengersen, and B. Liquet. Bayesian meta-analysis models for cross cancer genomic investigation of pleiotropic effects using group structure. *Statistics in Medicine*, 40(6):1498–1518, 2021.

S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, P. Hartge, M. Yeager, C. C. Chung, S. J. Chanock, and N. Chatterjee. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835, 2012.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011. ISSN 1935-8237. doi: 10.1561/2200000016.

Z. Chen, Z. He, B. B. Chu, J. Gu, T. Morrison, C. Sabatti, and E. Candès. Controlled variable selection from summary statistics only? a solution via ghostknockoffs and penalized regression. *ArXiv*, 2024.

D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

F. Clavel-Chapelon, G. Guillas, L. Tondeur, C. Kernaleguen, and M.-C. Boutron-Ruault. Risk of differentiated thyroid cancer in relation to adult weight, height and body shape over life: the french e3n cohort. *International journal of cancer*, 126(12):2984–2990, 2010.

E. Cordina-Duverger, C. Leux, M. Neri, C. Tcheandjieu, A.-V. Guizard, C. Schvartz, T. Truong, and P. Guénel. Hormonal and reproductive risk factors of papillary thyroid cancer: A population-based case-control study in france. *Cancer epidemiology*, 48:78–84, 2017.

E. Evangelou and J. P. Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*, 14(6):379–389, 2013.

Z. F. Gerring, A. Mina-Vargas, E. R. Gamazon, and E. M. Derks. E-magma: an eqtl-informed method to identify risk genes using genome-wide association study summary statistics. *Bioinformatics*, 37(16):2245–2249, 2021.

Q. He, H. H. Zhang, C. L. Avery, and D. Lin. Sparse meta-analysis with high-dimensional data. *Biostatistics*, 17(2):205–220, 2016.

Z. He, B. Chu, J. Yang, J. Gu, Z. Chen, L. Liu, T. Morrison, M. E. Belloy, X. Qi, N. Hejazi, et al. Beyond guilty by association at scale: searching for causal variants on the basis of genome-wide summary statistics. *bioRxiv*, 2024.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal Methods for Hierarchical Sparse Coding. *J. Mach. Learn. Res.*, 12:2297–2334, July 2011. ISSN 1532-4435.

D.-Y. Lin and D. Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

W. Liu, M. Li, W. Zhang, G. Zhou, X. Wu, J. Wang, Q. Lu, and H. Zhao. Leveraging functional annotation to identify genes associated with complex diseases. *PLOS Computational Biology*, 16(11):e1008315, 2020.

A. Majumdar, T. Haldar, S. Bhattacharya, and J. S. Witte. An efficient bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS genetics*, 14(2):e1007139, 2018.

M. M. Münch, C. F. Peeters, A. W. Van Der Vaart, and M. A. Van De Wiel. Adaptive group-regularized logistic elastic net regression. *Biostatistics*, 22(4):723–737, 2021.

S. M. Nielsen, M. G. White, S. Hong, B. Aschebrook-Kilfoy, E. L. Kaplan, P. Angelos, S. A. Kulkarni, O. I. Olopade, and R. H. Grogan. The breast–thyroid cancer link: a systematic review and meta-analysis. *Cancer Epidemiology and Prevention Biomarkers*, 25(2):231–238, 2016.

D. Ray and N. Chatterjee. A powerful method for pleiotropic analysis under composite null hypothesis identifies novel shared loci between type 2 diabetes and prostate cancer. *PLoS genetics*, 16(12):e1009218, 2020.

S. Smemo, J. J. Tena, K.-H. Kim, E. R. Gamazon, N. J. Sakabe, C. Gómez-Marín, I. Aneas, F. L. Credidio, D. R. Sobreira, N. F. Wasserman, et al. Obesity-associated variants within fto form long-range functional connections with irx3. *Nature*, 507(7492):371–375, 2014.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics*, 14(7):483–495, July 2013.

M. Sutton, P.-E. Sugier, T. Truong, and B. Liquet. Leveraging pleiotropic association using sparse group variable selection in genomics data. *BMC medical research methodology*, 22(1):1–12, 2022.

T. Truong, F. Lesueur, P.-E. Sugier, J. Guibon, C. Xhaard, M. Karimi, O. Kulkarni, E. A. Lucotte, D. Bacq-Daian, A. Boland-Auge, et al. Multiethnic genome-wide association study of differentiated thyroid cancer in the epithyr consortium. *International journal of cancer*, 148(12):2935–2946, 2021.

H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, 28(2):229–237, Jan. 2012. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btr649.

C. Xhaard, F. de Vathaire, E. Cléro, S. Maillard, Y. Ren, F. Borson-Chazot, G. Sassolas, C. Schvartz, M. Colonna, B. Lacour, et al. Anthropometric risk factors for differentiated thyroid cancer in young men and women from eastern france: a case-control study. *American journal of epidemiology*, 182(3):202–214, 2015.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476):1418–1429, 2006.

A Appendix A: Consistency of the MIC criterion

The proof follows the framework presented in [He et al., 2016]. We assume throughout that for each study s, $n_s\widehat{V}_s \to_p \Sigma_s$ and assuming block diagonal matrix we have $n_s\widehat{V}_s^{(g)} \to_p \Sigma_s^{(g)}, \forall g = 1, \ldots, G$. For each study s, we also assume that $n_s/n \to \nu_s$ as $n \to \infty$. Let $\nu_{\min} = \min_{1 \leqslant s \leqslant S} \{\nu_s\}$ and $\nu_{\max} = \max_{1 \leqslant s \leqslant S} \{\nu_s\}$.

Let $\mathcal{M}$ denote an arbitrary model, $\mathcal{M}_\eta$ denotes the model under $\eta = (\alpha, \lambda)$, and $\mathcal{M}_T$ denote the true model. We say that $\mathcal{M}$ is an under-fitted model if $\mathcal{M}_T \not\subset \mathcal{M}$ and an over-fitted model if $\mathcal{M}_T \subset M$ and $\mathcal{M} \neq \mathcal{M}_T$. Define

$$\hat{\beta}_{s,\mathcal{M}} = \underset{\beta_s \in R^p : \beta_{js}=0, \forall j \notin \mathcal{M}}{\operatorname{argmin}} \left(\beta_s - \hat{\beta}_s\right)^\top \widehat{V}_s^{-1} \left(\beta_s - \hat{\beta}_s\right).$$

Further, we assume $\widehat{V}_s$ as a block diagonal matrix of $\widehat{V}_s^{(g)} \equiv \widehat{Cov}(\hat{\beta}_s^{(g)})$, with $g = 1, ..., G$. Then, by re-arranging the variables with the variable group structure we denote $\hat{\beta}_{s,\mathcal{M}} = (\hat{\beta}_{s,\mathcal{M}}^{(1)^\top}, \ldots, \hat{\beta}_{s,\mathcal{M}}^{(G)^\top})^\top$. Due to the assumption of the independence between group, we also have

$$\hat{\beta}_{s,\mathcal{M}}^{(g)} = \underset{\beta_s^{(g)} \in R^{m_g} : \beta_{is}^{(g)}=0, \forall i \in (\pi_g \cap \bar{\mathcal{M}})}{\operatorname{argmin}} \left(\beta_s^{(g)} - \hat{\beta}_s^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\beta_s^{(g)} - \hat{\beta}_s^{(g)}\right)$$

Generally, remark that $\hat{\beta}_{s,\mathcal{M}} \neq \hat{\beta}_s$ because of the constraint that $\beta_{js}^{(g)} = 0, \forall j \notin \mathcal{M}$ and futher for some g we have $\hat{\beta}_{s,\mathcal{M}}^{(g)} \neq \hat{\beta}_s^{(g)}$.

Suppose that $\eta_* = (\alpha_*, \lambda_*)$ yields the true model, $\eta_u = (\alpha_u, \lambda_u)$ yields an under-fitted model, and $\eta_o = (\alpha_o, \lambda_o)$ yields an overfitted model. We wish to prove that, with high probability, $MIC_{\eta_u} > MIC_{\eta_t}$ and $MIC_{\eta_o} > MIC_{\eta_*}$.

A.1 Proof of $MIC_{\eta_u} > MIC_{\eta_*}$

Both $\hat{\beta}_{s,\eta_*}^{(g)}$ and $\hat{\beta}_s^{(g)}$ are consistent for $\beta_s^{(g)}$. Further, we have $\forall g = 1, \ldots, G$ $n_s\widehat{V}_s^{(g)} \to_p \Sigma_s^{(g)}$ and $\sum_{s=1}^S q_{s,\eta_*} \frac{log(n_s)}{n_s} \to 0$ then it is easy to verify that $MIC_{\eta_*} = o_p(1)$.

Now, for η_u:

$$MIC_{\eta_u} = \sum_{g=1}^G \sum_{s=1}^S (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_u}^{(g)})^T \widehat{V}_s^{(g)^{-1}} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_u}^{(g)}) + \sum_{s=1}^S q_{s,\eta_u} \frac{log(n_s)}{n_s}$$

$$MIC_{\eta_u} \geq \sum_{g=1}^G \sum_{s=1}^S (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_u}^{(g)})^T \widehat{V}_s^{(g)^{-1}} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_u}^{(g)})$$

$$MIC_{\eta_u} \geq \sum_{g=1}^G \sum_{s=1}^S (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_u}}^{(g)})^T \widehat{V}_s^{(g)^{-1}} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_u}}^{(g)})$$

$$MIC_{\eta_u} \geq \min_{\mathcal{M}_T \not\subset M} \sum_{g=1}^G \sum_{s=1}^S (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}}^{(g)})^T \widehat{V}_s^{(g)^{-1}} (\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}}^{(g)})$$
$$\to \sum_{g=1}^G \sum_{s=1}^S (\beta_s^{(g)} - \beta_{s,\mathcal{M}}^{(g)})^T \widehat{V}_s^{(g)^{-1}} (\beta_s^{(g)} - \beta_{s,\mathcal{M}}^{(g)}) > 0$$

As $\mathcal{M}_T \not\subset \mathcal{M}$ (i.e., $\mathcal{M}$ is an underfitting model) at least one component of $\beta_{s,\mathcal{M}}$ is 0 when the corresponding component of β_s is not, and so the difference between the two terms is different from 0. Then, since $\forall s, \forall g, \widehat{V}_s^{(g)^{-1}}$ is definite positive, the limit of the last element is strictly positive.

A.2 Proof of $MIC_{\eta_o} > MIC_{\eta_*}$

Let find an lower bound for $\mathcal{A} = n\left(MIC_{\eta_o} - MIC_{\eta_*}\right)$:

$$
\begin{aligned}
\mathcal{A} \geqslant & \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_o}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_o}^{(g)}\right) - \\
& \nu_{\min}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right) + \\
& \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} \left(q_{s,\eta_o} - q_{s,\eta_*}\right) \log\left(n_s\right) \\
\geqslant & \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_o}}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_o}}^{(g)}\right) - \\
& \nu_{\min}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right) + \\
& \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} \left(q_{s,\eta_o} - q_{s,\eta_*}\right) \log\left(n_s\right) \\
\geqslant & \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_o}}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}_{\eta_o}}^{(g)}\right) - \\
& \nu_{\min}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right) + \\
& \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} \log\left(n_s\right) \\
\geqslant & \nu_{\max}^{-1} \inf_{\mathcal{M}\supset\mathcal{M}_T} \sum_{k=1}^{K}\sum_{g=1}^{G} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\mathcal{M}}^{(g)}\right) - \\
& \nu_{\min}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} n_s \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right)^\top \widehat{V}_s^{(g)^{-1}} \left(\hat{\beta}_s^{(g)} - \hat{\beta}_{s,\eta_*}^{(g)}\right) + \\
& \nu_{\max}^{-1} \sum_{g=1}^{G}\sum_{s=1}^{S} \log\left(n_s\right).
\end{aligned}
$$

The second inequality follows from the definition of $\hat{\beta}_{s,\mathcal{M}_{\eta_o}}^{(g)}$. The third inequality holds because the considered model is an overfitted model. In the last inegality, the first term is $O_p(1)$ because for any $\mathcal{M} \supset \mathcal{M}_T, \hat{\beta}_{s,\mathcal{M}}^{(g)}$ is $\sqrt{n_s}$ consistent; the second term is also $O_p(1)$; and the third term goes to infinity. Hence, $MIC_{\eta_o} > MIC_{\eta_*}$ with probability tending to 1 .

B Appendix B: Simulation study

Table 7: Results of simulations for group pleiotropy with $S = 2$ and $G = 5$ for IGS=30%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25$. $\beta = 0.1$				
1.5k	ND	1 (0)	0.37(0.48)	0.84(0.36)
	D	0.95(0.18)	0(0)	0.17(0.31)
2k	ND	1(0)	0.66(0.48)	1(0)
	D	0.99(0.07)	0.06(0.24)	0.39(0.48)
2.5k	ND	1(0)	1(0)	1(0)
	D	0.99(0.07)	0.36(0.48)	0.63(0.47)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.52(0.5)	0.77(0.42)
$\rho = 0.25$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.62(0.49)	0.77(0.42)
2k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.78(0.42)	0.9(0.3)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.88(0.33)	0.94(0.24)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.92(0.27)	0.98(0.14)
$\rho = 0.50$. $\beta = 0.1$				
1.5k	ND	1(0)	0.69(0.46)	0.94(0.22)
	D	0.81(0.35)	0(0)	0.14(0.3)
2k	ND	1(0)	0.82(0.39)	1(0)
	D	0.91(0.22)	0.02(0.14)	0.3(0.45)
2.5k	ND	1(0)	0.98(0.14)	1(0)
	D	0.96(0.17)	0.24(0.43)	0.35(0.48)
3k	ND	1(0)	1(0)	1(0)
	D	0.96(0.17)	0.34(0.48)	0.41(0.48)
$\rho = 0.50$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.34(0.48)	0.53(0.49)
2k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.5(0.51)	0.66(0.47)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.58(0.5)	0.75(0.43)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.74(0.44)	0.77(0.42)

Table 8: Results of simulations for variable pleiotropy with $S = 2$ and $G = 5$ for IGS=30%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	0.28(0.15)	0.17(0.24)	0.25(0.22)	0.14(0.12)	0.14(0.12)
	D	0.28(0.16)	0(0)	0.05(0.19)		
2k	ND	0.44(0.2)	0.37(0.3)	0.6(0.34)	0.36(0.17)	0.28(0.19)
	D	0.44(0.2)	0.06(0.24)	0.34(0.46)		
2.5k	ND	0.58(0.21)	0.8(0.22)	0.77(0.3)	0.44(0.19)	0.42(0.24)
	D	0.57(0.22)	0.36(0.48)	0.57(0.49)		
3k	ND	0.68(0.21)	0.91(0.15)	0.87(0.23)	0.51(0.17)	0.55(0.27)
	D	0.68(0.22)	0.52(0.5)	0.75(0.43)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	0.97(0.05)	0.99(0.04)	1(0.01)	0.71(0.1)	0.98(0.04)
	D	0.97(0.06)	1(0)	1(0)		
2k	ND	0.99(0.02)	1(0)	1(0)	0.73(0.08)	1(0.01)
	D	0.99(0.03)	1(0)	1(0)		
2.5k	ND	1(0.01)	1(0)	1(0)	0.74(0.07)	1(0.01)
	D	1(0.01)	1(0)	1(0)		
3k	ND	1(0.01)	1(0)	1(0)	0.74(0.07)	1(0)
	D	1(0.01)	1(0)	1(0)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	0.19(0.18)	0.28(0.2)	0.2(0.17)	0.15(0.15)	0.09(0.13)
	D	0.18(0.18)	0(0)	0.05(0.18)		
2k	ND	0.3(0.25)	0.41(0.26)	0.44(0.34)	0.23(0.21)	0.18(0.21)
	D	0.29(0.26)	0.02(0.14)	0.26(0.43)		
2.5k	ND	0.4(0.29)	0.64(0.26)	0.51(0.36)	0.31(0.23)	0.27(0.27)
	D	0.4(0.29)	0.24(0.43)	0.34(0.48)		
3k	ND	0.49(0.31)	0.69(0.24)	0.58(0.35)	0.37(0.24)	0.36(0.33)
	D	0.49(0.32)	0.34(0.48)	0.38(0.49)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	0.84(0.24)	0.96(0.09)	0.93(0.15)	0.6(0.22)	0.83(0.26)
	D	0.82(0.27)	0.78(0.42)	0.82(0.38)		
2k	ND	0.91(0.16)	0.97(0.07)	0.97(0.09)	0.65(0.18)	0.9(0.18)
	D	0.9(0.18)	0.82(0.39)	0.93(0.24)		
2.5k	ND	0.78(0.28)	0.93(0.11)	0.89(0.2)	0.57(0.23)	0.78(0.3)
	D	0.71(0.38)	0.6(0.49)	0.8(0.39)		
3k	ND	0.97(0.07)	0.99(0.05)	0.99(0.03)	0.71(0.11)	0.97(0.07)
	D	0.97(0.08)	0.96(0.2)	1(0.01)		

Table 9: Results of simulations for group pleiotropy with $S = 2$ and $G = 5$ for IGS=30%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25.\ \beta = 0.1$				
1.5k	ND	1(0)	0.99(0.05)	0.98(0.08)
	D	1(0)	1(0)	1(0)
2k	ND	1(0)	0.99(0.07)	0.99(0.05)
	D	1(0)	0.99(0.05)	1(0)
2.5k	ND	1(0)	1(0)	0.99(0.05)
	D	1(0)	0.99(0.05)	0.99(0.05)
3k	ND	1(0)	0.98(0.08)	0.99(0.05)
	D	1(0)	0.99(0.05)	0.99(0.05)
$\rho = 0.25.\ \beta = 0.15$				
1.5k	ND	1(0)	0.99(0.07)	1(0)
	D	1(0)	0.99(0.05)	0.99(0.05)
2k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.99(0.05)	1(0)
2.5k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.99(0.07)	1(0)
3k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	1(0)	0.99(0.05)
$\rho = 0.50.\ \beta = 0.1$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
2k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.99(0.05)	1(0)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.99(0.05)	1(0)
3k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.99(0.05)	1(0)
$\rho = 0.50.\ \beta = 0.15$				
1.5k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	0.99(0.05)	1(0)
2k	ND	1(0)	0.99(0.05)	0.99(0.05)
	D	1(0)	0.99(0.05)	0.99(0.05)
2.5k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	1(0)	1(0)
3k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	1(0)	1(0)

Table 10: Results of simulations for group pleiotropy with $S = 2$ and $G = 5$ for IGS=70%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25$. $\beta = 0.1$				
1.5k	ND	1(0)	1(0)	1(0)
	D	0.85(0.34)	0.7(0.46)	0.87(0.32)
2k	ND	1(0)	1(0)	1(0)
	D	0.92(0.23)	0.82(0.39)	0.95(0.21)
2.5k	ND	1(0)	1(0)	1(0)
	D	0.99(0.07)	0.94(0.24)	1(0)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
$\rho = 0.50$. $\beta = 0.1$				
1.5k	ND	1(0)	1(0)	1(0)
	D	0.64(0.44)	0.44(0.5)	0.68(0.45)
2k	ND	1(0)	1(0)	1(0)
	D	0.75(0.39)	0.6(0.49)	0.78(0.39)
2.5k	ND	1(0)	1(0)	1(0)
	D	0.81(0.36)	0.62(0.49)	0.85(0.34)
3k	ND	1(0)	1(0)	1(0)
	D	0.85(0.35)	0.76(0.43)	0.85(0.34)

Table 11: Results of simulations for variable pleiotropy with $S = 2$ and $G = 5$ for IGS=30%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	1(0)	1(0.01)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	1(0.01)	0.99(0.01)	1(0)	1(0)
	D	1(0)	1(0.01)	1(0.01)		
2.5k	ND	1(0)	0.98(0.02)	0.99(0.01)	1(0.01)	1(0)
	D	1(0)	0.99(0.02)	0.99(0.01)		
3k	ND	1(0)	0.97(0.02)	0.99(0.01)	0.99(0.01)	1(0)
	D	1(0)	0.98(0.02)	0.99(0.01)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	1(0)	0.97(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		
2k	ND	1(0)	0.97(0.02)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		
2.5k	ND	1(0)	0.97(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		
3k	ND	1(0)	0.97(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	1(0)	0.99(0.01)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	0.99(0.01)	0.99(0.01)	1(0)	1(0)
	D	1(0)	1(0)	1(0.01)		
2.5k	ND	1(0)	0.98(0.02)	0.99(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.02)	0.99(0.01)		
3k	ND	1(0)	0.98(0.02)	0.99(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.02)	1(0.01)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	1(0)	0.97(0.02)	1(0.01)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0.01)		
2k	ND	1(0)	0.97(0.02)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		
2.5k	ND	1(0)	0.97(0.02)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		
3k	ND	1(0)	0.97(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.98(0.01)	1(0)		

Table 12: Results of simulations for variable pleiotropy with $S = 2$ and $G = 5$ for IGS=70%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	0.77(0.2)	0.94(0.1)	0.94(0.12)	0.55(0.18)	0.77(0.23)
	D	0.7(0.32)	0.7(0.46)	0.83(0.36)		
2k	ND	0.9(0.13)	0.97(0.07)	0.98(0.05)	0.63(0.16)	0.89(0.17)
	D	0.84(0.27)	0.82(0.39)	0.94(0.24)		
2.5k	ND	0.95(0.09)	0.99(0.04)	1(0.01)	0.68(0.12)	0.96(0.08)
	D	0.94(0.12)	0.94(0.24)	1(0.01)		
3k	ND	0.97(0.05)	0.99(0.03)	1(0.01)	0.71(0.1)	0.98(0.04)
	D	0.98(0.05)	1(0.01)	1(0.01)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	0.6(0.32)	0.82(0.18)	0.74(0.3)	0.42(0.25)	0.57(0.36)
	D	0.51(0.41)	0.44(0.5)	0.59(0.47)		
2k	ND	0.69(0.32)	0.87(0.16)	0.84(0.24)	0.51(0.24)	0.68(0.35)
	D	0.62(0.4)	0.6(0.49)	0.68(0.45)		
2.5k	ND	0.78(0.28)	0.89(0.14)	0.89(0.2)	0.57(0.23)	0.78(0.3)
	D	0.71(0.38)	0.62(0.49)	0.8(0.39)		
3k	ND	0.84(0.24)	0.93(0.11)	0.92(0.18)	0.6(0.22)	0.83(0.26)
	D	0.79(0.35)	0.76(0.43)	0.81(0.38)		

Table 13: Results of simulations for group pleiotropy with $S = 2$ and $G = 5$ for IGS=70%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25. \beta = 0.1$				
1.5k	ND	1(0)	0.93(0.15)	0.94(0.13)
	D	1(0)	0.93(0.16)	0.97(0.12)
2k	ND	1(0)	0.92(0.16)	0.97(0.1)
	D	1(0)	0.93(0.15)	0.95(0.14)
2.5k	ND	1(0)	0.92(0.16)	0.98(0.08)
	D	1(0)	0.93(0.15)	0.97(0.09)
3k	ND	1(0)	0.93(0.16)	0.98(0.08)
	D	1(0)	0.89(0.18)	0.99(0.07)
$\rho = 0.50. \beta = 0.1$				
1.5k	ND	1(0)	0.95(0.14)	0.96(0.13)
	D	1(0)	0.95(0.15)	0.98(0.08)
2k	ND	1(0)	0.93(0.16)	0.99(0.07)
	D	1(0)	0.94(0.16)	0.99(0.07)
2.5k	ND	1(0)	0.93(0.15)	0.97(0.09)
	D	1(0)	0.93(0.17)	0.97(0.09)
3k	ND	1(0)	0.93(0.15)	0.98(0.08)
	D	1(0)	0.93(0.18)	0.98(0.08)

Table 14: Results of simulations for group pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25$. $\beta = 0.1$				
1.5k	ND	1(0)	0.39(0.48)	0.69(0.44)
	D	0.97(0.15)	0(0)	0.17(0.35)
2k	ND	1(0)	0.59(0.49)	0.98(0.06)
	D	0.98(0.07)	0(0)	0.34(0.47)
2.5k	ND	1(0)	0.93(0.24)	0.99(0.06)
	D	0.99(0.04)	0.26(0.44)	0.58(0.5)
3k	ND	1(0)	0.99(0.04)	0.98(0.06)
	D	0.99(0.04)	0.5(0.51)	0.72(0.45)
$\rho = 0.25$. $\beta = 0.15$				
1.5k	ND	1(0)	0.98(0.05)	0.99(0.05)
	D	1(0)	0.58(0.5)	0.74(0.44)
2k	ND	1(0)	0.99(0.04)	0.98(0.05)
	D	1(0)	0.76(0.43)	0.86(0.35)
2.5k	ND	1(0)	1(0.03)	1(0)
	D	1(0)	0.84(0.37)	0.96(0.2)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.92(0.27)	0.96(0.2)
$\rho = 0.50$. $\beta = 0.1$				
1.5k	ND	1(0)	0.71(0.45)	0.8(0.39)
	D	0.83(0.35)	0(0)	0.14(0.32)
2k	ND	1(0)	0.77(0.41)	1(0.03)
	D	0.89(0.28)	0(0)	0.29(0.45)
2.5k	ND	1(0)	0.97(0.15)	1(0.03)
	D	0.93(0.21)	0.2(0.4)	0.36(0.48)
3k	ND	1(0)	1(0.03)	0.99(0.04)
	D	0.95(0.2)	0.3(0.46)	0.4(0.49)
$\rho = 0.50$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0.03)	1(0)
	D	0.99(0.06)	0.32(0.47)	0.49(0.5)
2k	ND	1(0)	0.99(0.05)	1(0)
	D	1(0)	0.48(0.5)	0.58(0.5)
2.5k	ND	1(0)	1(0.03)	1(0.03)
	D	1(0)	0.52(0.5)	0.74(0.44)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.72(0.45)	0.74(0.44)

Table 15: Results of simulations for variable pleiotropy with $S = 2$ and $G = 5$ for IGS=70%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	1(0)	0.99(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.99(0.01)	0.99(0.01)		
2k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	0.99(0.01)		
2.5k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		
3k	ND	1(0)	0.99(0.01)	1(0)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0.01)		
2k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		
2.5k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		
3k	ND	1(0)	0.99(0.01)	1(0.01)	1(0)	1(0)
	D	1(0)	0.99(0.01)	1(0)		

Table 16: Results of simulations for variable pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of Matthews Correlation Coefficient (MCC) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	0.51(0.14)	0.26(0.32)	0.33(0.23)	0.23(0.14)	0.4(0.19)
	D	0.5(0.16)	0(0)	0.04(0.11)		
2k	ND	0.64(0.15)	0.39(0.33)	0.69(0.24)	0.35(0.17)	0.56(0.21)
	D	0.63(0.18)	0(0)	0.27(0.42)		
2.5k	ND	0.74(0.15)	0.78(0.23)	0.84(0.19)	0.46(0.18)	0.69(0.19)
	D	0.74(0.16)	0.25(0.43)	0.54(0.49)		
3k	ND	0.82(0.14)	0.89(0.1)	0.9(0.14)	0.54(0.18)	0.78(0.18)
	D	0.81(0.15)	0.48(0.49)	0.71(0.44)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	0.81(0.14)	0.9(0.09)	0.91(0.14)	0.54(0.18)	0.78(0.18)
	D	0.8(0.15)	0.55(0.48)	0.73(0.43)		
2k	ND	0.9(0.11)	0.93(0.06)	0.95(0.09)	0.65(0.15)	0.89(0.13)
	D	0.9(0.11)	(0.73)0.41	0.85(0.35)		
2.5k	ND	0.94(0.08)	0.94(0.05)	0.97(0.07)	0.72(0.13)	0.94(0.09)
	D	0.94(0.09)	0.81(0.36)	0.93(0.22)		
3k	ND	0.97(0.06)	0.95(0.04)	0.98(0.05)	0.76(0.11)	0.96(0.07)
	D	0.97(0.06)	0.88(0.26)	0.95(0.2)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	0.37(0.22)	0.42(0.28)	0.32(0.17)	0.16(0.16)	0.26(0.24)
	D	0.34(0.24)	0(0)	0.04(0.12)		
2k	ND	0.49(0.25)	0.48(0.27)	0.57(0.23)	0.24(0.21)	0.38(0.3)
	D	0.47(0.27)	0(0)	0.23(0.4)		
2.5k	ND	0.59(0.24)	0.71(0.17)	0.64(0.25)	0.32(0.24)	0.49(0.32)
	D	0.56(0.28)	0.19(0.39)	0.32(0.46)		
3k	ND	0.65(0.25)	0.78(0.13)	0.66(0.25)	0.39(0.26)	0.57(0.32)
	D	0.63(0.27)	0.29(0.44)	0.36(0.47)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	0.64(0.25)	0.79(0.13)	0.71(0.24)	0.39(0.26)	0.57(0.32)
	D	0.63(0.26)	0.3(0.45)	0.44(0.48)		
2k	ND	0.74(0.24)	0.85(0.12)	0.78(0.23)	0.49(0.26)	0.7(0.3)
	D	0.74(0.25)	0.46(0.48)	0.54(0.49)		
2.5k	ND	0.81(0.21)	0.89(0.1)	0.88(0.18)	0.57(0.25)	0.78(0.26)
	D	0.81(0.22)	0.5(0.49)	0.73(0.44)		
3k	ND	0.86(0.18)	0.91(0.08)	0.89(0.17)	0.62(0.24)	0.83(0.22)
	D	0.86(0.19)	0.69(0.44)	0.73(0.44)		

Table 17: Results of simulations for group pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25$. $\beta = 0.1$				
1.5k	ND	1(0)	0.39(0.49)	0.72(0.45)
	D	0.96(0.17)	0(0)	0.16(0.34)
2k	ND	1(0)	0.59(0.49)	1(0)
	D	0.97(0.12)	0(0)	0.34(0.47)
2.5k	ND	1(0)	0.94(0.24)	1(0)
	D	0.99(0.07)	0.26(0.44)	0.58(0.5)
3k	ND	1(0)	1(0)	1(0)
	D	0.99(0.07)	0.5(0.51)	0.72(0.45)
$\rho = 0.25$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.58(0.5)	0.74(0.44)
2k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.76(0.43)	0.86(0.35)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.84(0.37)	0.96(0.2)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.92(0.27)	0.96(0.2)
$\rho = 0.50$. $\beta = 0.1$				
1.5k	ND	1(0)	0.72(0.45)	0.8(0.39)
	D	0.81(0.36)	0(0)	0.12(0.3)
2k	ND	1(0)	0.78(0.42)	1(0)
	D	0.87(0.3)	0(0)	0.28(0.44)
2.5k	ND	1(0)	0.98(0.14)	1(0)
	D	0.91(0.24)	0.2(0.4)	0.36(0.48)
3k	ND	1(0)	1(0)	1(0)
	D	0.94(0.22)	0.3(0.46)	0.4(0.49)
$\rho = 0.50$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.34(0.48)	0.53(0.49)
2k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.5(0.51)	0.66(0.47)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.58(0.5)	0.75(0.43)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	0.74(0.44)	0.77(0.42)

Table 18: Results of simulations for variable pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of True Positive Rate (TPR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	0.29(0.14)	0.17(0.22)	0.17(0.16)	0.08(0.07)	0.2(0.14)
	D	0.28(0.14)	0(0)	0.02(0.06)		
2k	ND	0.45(0.18)	0.27(0.25)	0.55(0.34)	0.16(0.12)	0.37(0.21)
	D	0.44(0.19)	0(0)	0.25(0.42)		
2.5k	ND	0.58(0.2)	0.7(0.29)	0.76(0.3)	0.25(0.15)	0.53(0.24)
	D	0.58(0.21)	0.26(0.44)	0.53(0.5)		
3k	ND	0.69(0.2)	0.85(0.2)	0.86(0.24)	0.34(0.18)	0.66(0.25)
	D	0.68(0.21)	0.5(0.51)	0.72(0.45)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	0.68(0.2)	0.89(0.17)	0.87(0.23)	0.34(0.18)	0.66(0.25)
	D	0.67(0.21)	0.58(0.5)	0.74(0.44)		
2k	ND	0.83(0.17)	0.94(0.12)	0.94(0.15)	0.47(0.18)	0.83(0.21)
	D	0.82(0.18)	0.76(0.43)	0.86(0.35)		
2.5k	ND	0.9(0.14)	0.96(0.11)	0.97(0.12)	0.57(0.16)	0.91(0.16)
	D	0.9(0.14)	0.84(0.37)	0.94(0.24)		
3k	ND	0.95(0.1)	0.98(0.08)	0.98(0.09)	0.62(0.15)	0.95(0.12)
	D	0.94(0.11)	0.92(0.27)	0.95(0.2)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	0.19(0.17)	0.27(0.19)	0.13(0.08)	0.05(0.07)	0.13(0.15)
	D	0.18(0.17)	0(0)	0.01(0.07)		
2k	ND	0.31(0.24)	0.32(0.19)	0.39(0.33)	0.11(0.12)	0.24(0.24)
	D	0.3(0.25)	0(0)	0.22(0.41)		
2.5k	ND	0.41(0.28)	0.58(0.26)	0.49(0.37)	0.17(0.17)	0.35(0.31)
	D	0.4(0.29)	0.2(0.4)	0.32(0.47)		
3k	ND	0.49(0.3)	0.67(0.24)	0.51(0.37)	0.23(0.21)	0.44(0.34)
	D	0.48(0.31)	0.3(0.46)	0.35(0.47)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	0.48(0.31)	0.69(0.24)	0.59(0.37)	0.23(0.21)	0.44(0.34)
	D	0.47(0.31)	0.32(0.47)	0.44(0.5)		
2k	ND	0.61(0.31)	0.78(0.22)	0.68(0.36)	0.32(0.24)	0.59(0.35)
	D	0.61(0.32)	0.48(0.5)	0.54(0.5)		
2.5k	ND	0.71(0.29)	0.87(0.19)	0.83(0.29)	0.4(0.25)	0.69(0.33)
	D	0.7(0.3)	0.52(0.5)	0.74(0.44)		
3k	ND	0.78(0.27)	0.9(0.16)	0.84(0.27)	0.46(0.25)	0.76(0.31)
	D	0.77(0.27)	0.72(0.45)	0.74(0.44)		

Table 19: Results of simulations for group pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W
$\rho = 0.25$. $\beta = 0.1$				
1.5k	ND	1(0)	1(0)	1(0.01)
	D	1(0)	1(0)	1(0.01)
2k	ND	1(0)	1(0)	1(0.01)
	D	1(0)	1(0)	1(0.01)
2.5k	ND	1(0)	1(0)	1(0.01)
	D	1(0)	1(0)	1(0)
3k	ND	1(0)	1(0)	1(0.01)
	D	1(0)	1(0)	1(0)
$\rho = 0.25$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0.01)	1(0)
	D	1(0)	1(0)	1(0)
2k	ND	1(0)	1(0.01)	1(0.01)
	D	1(0)	1(0)	1(0)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
$\rho = 0.50$. $\beta = 0.1$				
1.5k	ND	1(0)	1(0)	1(0.01)
	D	1(0)	1(0)	1(0)
2k	ND	1(0)	1(0.01)	1(0)
	D	1(0)	1(0)	1(0)
2.5k	ND	1(0)	1(0.01)	1(0)
	D	1(0)	1(0)	1(0)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
$\rho = 0.50$. $\beta = 0.15$				
1.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
2k	ND	1(0)	1(0.01)	1(0)
	D	1(0)	1(0)	1(0)
2.5k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)
3k	ND	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)

Table 20: Results of simulations for variable pleiotropy with $S = 2$ and $G = 50$ for IGS=30%. Mean (Standard Deviation) of True Negative Rate (TNR) over 50 replications, expressed as percentages.

N		DS	MPSG	MPSG-W	ASSET	PLACO
$\rho = 0.25, \beta = 0.1$						
1.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
3k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
$\rho = 0.25, \beta = 0.15$						
1.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
3k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
$\rho = 0.50, \beta = 0.1$						
1.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
3k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
$\rho = 0.50, \beta = 0.15$						
1.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
2.5k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		
3k	ND	1(0)	1(0)	1(0)	1(0)	1(0)
	D	1(0)	1(0)	1(0)		