Adjustment for the uncertainty of predicted expression in transcriptome-wide association study: a fusion of measurement error theory and bootstrapping

Arunabha Majumdar[1] and Tanushree Haldar[2]

[1]Department of Mathematics, Indian Institute of Technology Hyderabad, Kandi, Telangana 502285, India

[2]School of Pharmacy, University of California, San Francisco, California 94143, USA

Abstract

Transcriptome-wide association study (TWAS) is a promising approach to identifying novel genes associated with complex traits. The method commonly used in a standard TWAS approach combines two subsequent regressions. A significant criticism of the method is that the uncertainty of the predicted expression is ignored in the second-step regression, which can lead the final inference on the gene trait association subject to an uncontrolled rate of false positives, reducing the reliability of the detected association signals. We propose a novel approach to adjust for the uncertainty of predicted expression in TWAS. We improvise techniques from measurement error theory to derive the adjustment factor that needs to be incorporated in the test statistic obtained from standard TWAS. Next, we apply bootstrapping techniques for penalized regression to estimate the adjustment factor. We use simulations to show that the traditional TWAS approach inflates the type 1 error rate, whereas our adjusted TWAS approach adequately controls it. Due to an inflated false positive rate, the standard TWAS produces a higher power than the adjusted TWAS approach. For a multi-trait TWAS, ASSET has an excessive type 1 error rate based on the traditional TWAS statistics, contrary to the adjusted TWAS statistics. Overall, the proposed TWAS approach is a valid test for the gene-trait association while adequately controlling the false positive rate.

Introduction

Transcriptome-wide association study (TWAS) is a promising approach to identifying novel genes associated with complex traits [1, 2]. It provides better biological insights into the genetic architecture of a complex trait than a genome-wide association study (GWAS) [3]. The method commonly used in a standard TWAS approach combines two subsequent regressions. In the first-step regression based on an external panel of expression and genotype data, we train a prediction model to estimate the genetic component of gene expression. Next, we consider separate GWAS data in which no expression measurements are available. We implement the prediction model from the first-step regression to estimate the genetic component of expression in the GWAS data [1, 2]. Finally, in the second step, we regress a trait on the predicted genetic component of expression (predicted expression) in the GWAS data to test for an association between the gene and the trait. A significant criticism of the standard TWAS approach is that the uncertainty of the predicted expression is ignored in the second-step regression. The uncertainty arises due to the variance of the prediction equation for expression estimated in the external panel data. Due to the unadjusted uncertainty, the final inference on the gene trait association can be subject to an uncontrolled rate of false positives, which can reduce the reliability of the detected association signals.

This article investigates this vital question under a frequentist framework. We modify techniques from measurement error theory [4] to derive the adjustment factor that needs to be multiplied by the test statistic obtained from a standard TWAS to account for the uncertainty of predicted expression. Then we apply bootstrapping techniques for penalized regression [5] to estimate the adjustment factor. We refer to this approach as adjTWAS (adjusted TWAS). We focus on the summary statistics version of TWAS, which requires summary-level association data for GWAS instead of individual-level data [1] and can easily be applied to various complex traits using publicly available summary-level GWAS data.

We use extensive simulations to show that the standard TWAS inflates the type 1 error rate, whereas adjTWAS adequately controls it. Due to an inflated false positive rate, the standard TWAS produces a higher power than adjTWAS. In a multi-trait TWAS setting, we compare the

performance of the subset-based meta-analysis method ASSET [6], using the test statistics obtained by standard TWAS and adjTWAS. We find that ASSET produces an inflated type 1 error rate based on traditional TWAS statistics but sufficiently controls it based on adjTWAS statistics. Analogous to the univariate scenario, due to an excessive type 1 error rate, the multi-trait TWAS power obtained by ASSET is higher for standard TWAS than adjTWAS.

Materials and methods

Regression of gene expression on local SNPs in external panel

Let us consider a given gene and its expression in a tissue of interest. In the reference panel data, we regress the expression of a gene on the genotypes of its local SNPs. Consider the following linear model:

$$E = \mathbf{u}'\boldsymbol{\lambda} + v \tag{1}$$

Let E denote the mean-centered expression of the gene, and $\mathbf{u}$ denotes the normalized (centered for mean and then divided by standard deviation) genotype vector for p local SNPs of the gene. $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ denotes the effect size vector, and v denotes the random error. Since the number of local SNPs can be close to or larger than the sample size (e.g., GTEx data [7, 8]), a penalized regression is implemented in practice to estimate $\boldsymbol{\lambda}$. Lasso [9] and Elastic net [10] are two commonly used methods. In Lasso, $\hat{\boldsymbol{\lambda}}$ is obtained by minimizing the penalized loss function:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \sum_{i=1}^{n} (E_i - \mathbf{u}_i'\boldsymbol{\lambda})^2 + w \sum_{j=1}^{p} |\lambda_j|$$

Here, $w \geq 0$ denotes the weight associated with the penalty term, which determines the amount of shrinkage. In Elastic net, $\hat{\boldsymbol{\lambda}}$ is given by:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \sum_{i=1}^{n} (E_i - \mathbf{u}_i'\boldsymbol{\lambda})^2 + w\{\alpha \sum_{j=1}^{p} |\lambda_j| + (1-\alpha) \sum_{j=1}^{p} \lambda_j^2\}$$

Here, $\alpha \in (0,1)$ determines the relative contribution of the Lasso and ridge regression penalty. Let $\hat{\boldsymbol{\lambda}}$ denote the estimate of $\boldsymbol{\lambda}$ obtained by the preferred penalized regression. Let $\mathrm{E}(\hat{\boldsymbol{\lambda}})$ and $\mathrm{cov}(\hat{\boldsymbol{\lambda}})$ denote the expectation vector and covariance matrix of $\hat{\boldsymbol{\lambda}}$, respectively.

Regression of trait on the genetic component of expression in GWAS data

In general, expression measurements are not available in GWAS data. Assume that the genotype data for the set of p local SNPs considered for the gene in the expression panel is also available in GWAS data. The genetic component of expression is predicted using $\hat{\boldsymbol{\lambda}}$, obtained in the first-step regression. We assume that the expression and GWAS data are from the same population such that the effect of local SNPs on the expression is the same between the two datasets. If $\mathbf{g}$ denotes the genotype vector for the local SNPs in GWAS, the true genetic component of expression is given by: $x = \mathbf{g}^{'}\boldsymbol{\lambda}$. Since $\boldsymbol{\lambda}$ is not known, we predict x by $X = \mathbf{g}^{'}\hat{\boldsymbol{\lambda}}$. The main goal of TWAS is to test for an association between the trait and the true genetic component of expression. Let Y denote the trait of interest and $(y_1, \ldots, y_n)$ be the trait values for the GWAS individuals. For i^{th} individual, let $\mathbf{g}_i$ denote the genotype vector for the local SNPs, $x_i = \mathbf{g}_i^{'}\boldsymbol{\lambda}$, and $X_i = \mathbf{g}_i^{'}\hat{\boldsymbol{\lambda}}$, $1 \leq i \leq n$. Thus, X_i is an estimate of x_i subject to an estimation/measurement error.

Ideal and working regressions

We consider two regressions, one considering x as the regressor and the other considering X as the regressor, where x is unobserved, and X is observed. We specify the ideal regression as the following:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, 1 \leq i \leq n \tag{2}$$

β_1 is the effect of the actual genetic component of expression on the trait. Assume that $\{\epsilon_i, 1 \leq i \leq n\}$ are i.i.d. $N(0, \sigma_\epsilon^2)$. The working regression is specified as follows:

$$y_i = b_0 + b_1 X_i + e_i, 1 \leq i \leq n \tag{3}$$

Here b_1 is the effect of the noisy predicted expression on the trait, and $\{e_i, i = 1, \ldots, n\}$ are i.i.d. $N(0, \sigma_e^2)$. In standard TWAS, we implement the working regression ignoring the uncertainty or measurement error in X observations. Note that $\mathrm{E}(X_i) = \mathbf{g}_i^{'}\mathrm{E}(\hat{\boldsymbol{\lambda}})$ and $\mathrm{var}(X_i) = \mathbf{g}_i^{'}\mathrm{cov}(\hat{\boldsymbol{\lambda}})\mathbf{g}_i$. Our main goal is to test whether the true genetic component of expression affects the trait, i.e., $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. However, in standard TWAS, we test $H_0 : b_1 = 0$ vs $b_1 \neq 0$ using the least square estimator of $b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$.

From the ideal regression model (equation 2), $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$ and $\mathrm{var}(y_i) = \mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$. We assume that x_i is not random. $\mathrm{cov}(X_i, y_i) = \mathrm{cov}(\mathbf{g}_i^{'}\hat{\boldsymbol{\lambda}}, \beta_0 + \beta_1 x_i + \epsilon_i) = 0$. We make reasonable assumptions that $\boldsymbol{\lambda}$, hence x_i, is an unknown fixed quantity, and the random error in ideal regression ϵ_i is independent of $\hat{\boldsymbol{\lambda}}$. Thus, $\mathrm{cov}(X_i, y_i) = 0$, and if the expression panel data sample size is sufficiently large such that $\hat{\boldsymbol{\lambda}}$ asymptotically follows multivariate normal [11], we can assume that X_i and y_i are independently distributed. We consider $\mathbf{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ to be independently distributed in the following sections.

Unbiased estimator of β_1 and related test

The least squares estimator of b_1 obtained from the working regression is as follows: $\hat{b}_1 = \frac{S_{Xy}}{S_{XX}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$. Note that $\sum_{i=1}^n X_i = \sum_{i=1}^n \mathbf{g}_i^{'}\hat{\boldsymbol{\lambda}} = (\sum_{i=1}^n \mathbf{g}_i)^{'}\hat{\boldsymbol{\lambda}} = 0$. Here, $\sum_{i=1}^n \mathbf{g}_i = \mathbf{0}$, because genotype data of local SNPs are normalized, i.e., zero mean and unity variance for each local SNP. Thus, $\bar{X} = \mathbf{0}$ and $\hat{b}_1 = \frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{i=1}^n X_i^2}$. We note that $\hat{b}_1$ is an unbiased estimator of b_1. Our main goal is to obtain an unbiased estimator of β_1 and test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. A natural question is: can $\hat{b}_1$ be adjusted to formulate an unbiased estimator of β_1?

$\mathrm{E}(\hat{b}_1) = \mathrm{E}\left(\frac{\sum_{i=1}^n X_i(y_i - \bar{y})}{\sum_{j=1}^n X_j^2}\right) = \sum_{i=1}^n \mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}(y_i - \bar{y})\right) = \sum_{i=1}^n \mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right)\mathrm{E}(y_i - \bar{y})$.

Under the ideal regression model (equation 2),

$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$. $\mathrm{E}(\bar{y}) = \frac{1}{n}\sum_{i=1}^n E(y_i) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) = \beta_0 + \frac{\beta_1}{n}\sum_{i=1}^n x_i = \beta_0$. Because, $\sum_{i=1}^n x_i = \sum_{i=1}^n \mathbf{g}_i^{'}\boldsymbol{\lambda} = (\sum_{i=1}^n \mathbf{g}_i)^{'}\boldsymbol{\lambda} = 0$. Again, $\sum_{i=1}^n \mathbf{g}_i = \mathbf{0}$ because the genotype data of local SNPs are normalized. So, $E(y_i - \bar{y}) = \beta_1 x_i$.

Next, $X_i = \mathbf{g}_i^{'}\hat{\boldsymbol{\lambda}}$. $\sum_{j=1}^n X_j^2 = \sum_{j=1}^n X_j^{'}X_j = \sum_{j=1}^n \hat{\boldsymbol{\lambda}}^{'}\mathbf{g}_j\mathbf{g}_j^{'}\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}^{'}\left(\sum_{j=1}^n \mathbf{g}_j\mathbf{g}_j^{'}\right)\hat{\boldsymbol{\lambda}} = n\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}$. Here, R denotes the linkage disequilibrium (LD) matrix of the p local SNPs based on the GWAS

data. Thus, $\mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right) = \mathrm{E}\left(\frac{\mathbf{g}_i'\hat{\boldsymbol{\lambda}}}{n\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right) = \frac{1}{n}\mathbf{g}_i'\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$.

Thus, $\mathrm{E}(\hat{b}_1) = \sum_{i=1}^n \mathrm{E}(y_i - \bar{y})\mathrm{E}\left(\frac{X_i}{\sum_{j=1}^n X_j^2}\right) = \frac{\beta_1}{n}\sum_{i=1}^n x_i\mathbf{g}_i'\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right) = \frac{\beta_1}{n}\sum_{i=1}^n(\boldsymbol{\lambda}'\mathbf{g}_i)\mathbf{g}_i'\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$ $= \frac{\beta_1}{n}\boldsymbol{\lambda}'(\sum_{i=1}^n \mathbf{g}_i\mathbf{g}_i')\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right) = \beta_1\boldsymbol{\lambda}'R\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right) = B\beta_1$, where $B = \boldsymbol{\lambda}'R\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$ is a constant. We have $\mathrm{E}\left(\frac{\hat{b}_1}{B}\right) = \beta_1$. Note that B involves unknown terms.

Next, we derive the variance of $\hat{\beta}_1 = \frac{\hat{b}_1}{B}$. $\mathrm{var}(\hat{\beta}_1) = \frac{\mathrm{var}(\hat{b}_1)}{B^2}$. From the above, $\hat{b}_1 = \frac{\sum_{i=1}^n(X_i-\bar{X})(y_i-\bar{y})}{\sum_{j=1}^n(X_j-\bar{X})^2}$. Here, both X and y are random variables. Since $X_i = \mathbf{g}_i'\hat{\boldsymbol{\lambda}}$, the randomness of X_i comes through $\hat{\boldsymbol{\lambda}}$ Denote $\mathbf{X} = (X_1, \ldots, X_n)$. Using the law of total variance, $\mathrm{var}(\hat{b}_1) = \mathrm{E}(\mathrm{var}(\hat{b}_1|\mathbf{X})) + \mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X}))$. $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \mathrm{var}\left(\frac{\sum_{i=1}^n(X_i-\bar{X})(y_i-\bar{y})}{\sum_{j=1}^n(X_j-\bar{X})^2}\middle|\mathbf{X}\right)$. We note that $\mathbf{X}$ is fixed only when $\hat{\boldsymbol{\lambda}}$ is fixed. Hence, $\mathrm{var}\left(\frac{\sum_{i=1}^n(X_i-\bar{X})(y_i-\bar{y})}{\sum_{j=1}^n(X_j-\bar{X})^2}\middle|\mathbf{X}\right) = \frac{1}{\{\sum_{j=1}^n(X_j-\bar{X})^2\}^2}\mathrm{var}\{\sum_{i=1}^n(X_i-\bar{X})(y_i-\bar{y})|\mathbf{X}\} = \frac{1}{\{\sum_{j=1}^n(X_j-\bar{X})^2\}^2}\mathrm{var}\{\sum_{i=1}^n(X_i-\bar{X})y_i|\mathbf{X}\}$, since $\sum_{i=1}^n(X_i-\bar{X})|\mathbf{X} = 0$. $\mathrm{var}\{\sum_{i=1}^n(X_i-\bar{X})y_i|\mathbf{X}\} = \sum_{i=1}^n(X_i-\bar{X})^2\mathrm{var}(y_i|X_i)$ $= \sigma_e^2\sum_{i=1}^n(X_i-\bar{X})^2$. The covariance terms become zero because, based on the working regression model (3), $y_i|X_i$ are independently distributed for $i = 1, \ldots, n$ and $\mathrm{var}(y_i|X_i) = \sigma_e^2$. Thus, $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{\sum_{i=1}^n(X_i-\bar{X})^2}$. Furthermore, $\bar{X} = \mathbf{0}$, conditioned on $\mathbf{X}$. Hence, $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{\sum_{i=1}^n X_i^2} = \frac{\sigma_e^2}{\sum_{i=1}^n X_i^2} = \frac{\sigma_e^2}{n\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}$. Thus, $\mathrm{E}(\mathrm{var}(\hat{b}_1|\mathbf{X})) = \frac{\sigma_e^2}{n}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$. Here, the expectation concerns $\hat{\boldsymbol{\lambda}}$ based on the expression panel data. The other part of $\mathrm{var}(\hat{b}_1)$ is $\mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X}))$. Again, we use that $\bar{X}|\mathbf{X} = 0$. Thus, $\mathrm{E}(\hat{b}_1|\mathbf{X}) = \mathrm{E}\left(\frac{\sum_{i=1}^n X_i(y_i-\bar{y})}{\sum_{j=1}^n X_j^2}\middle|\mathbf{X}\right) = \sum_{i=1}^n \frac{X_i}{\sum_{j=1}^n X_j^2}\mathrm{E}\left((y_i-\bar{y})|\mathbf{X}\right)$. From the working regression model, $\mathrm{E}(y_i|X_i) = b_0 + b_1X_i$. So $\mathrm{E}(\bar{y}|\mathbf{X}) = b_0$ and $\mathrm{E}\left((y_i-\bar{y})|\mathbf{X}\right) = b_1X_i$. Thus, $\mathrm{E}(\hat{b}_1|\mathbf{X}) = b_1$ and $\mathrm{var}(\mathrm{E}(\hat{b}_1|\mathbf{X})) = 0$. We also note that conditioned on $\mathbf{X}$, $\hat{b}_1$ is an unbiased estimator of b_1. Thus, putting together all the parts, $\mathrm{var}(\hat{b}_1) = \frac{\sigma_e^2}{n}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$.

Given that B is known, a plausible unbiased estimator of β_1 is $\frac{\hat{b}_1}{B}$ with the corresponding variance $\mathrm{var}\left(\frac{\hat{b}_1}{B}\right) = \frac{\sigma_e^2}{nB^2}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$. But B is unknown: $B = \boldsymbol{\lambda}'R\mathrm{E}\left(\frac{\hat{\boldsymbol{\lambda}}}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)$. To test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, denote the Z statistic of adjusted TWAS as $Z_{adj} = \frac{\frac{\hat{b}_1}{B}}{\frac{\hat{\sigma}_e}{\sqrt{n}B}\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)}} = \frac{\sqrt{n}\hat{b}_1}{\hat{\sigma}_e\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)}}$. Next, we derive the Z statistic for the traditional TWAS based on the working regression model is $\frac{\hat{b}_1}{se(\hat{b}_1)}$, where $\hat{b}_1 = \frac{\sum_{i=1}^n X_i(y_i-\bar{y})}{\sum_{j=1}^n X_j^2}$ and $\mathrm{var}(\hat{b}_1|\mathbf{X}) = \frac{\sigma_e^2}{n\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}$. Thus, the Z statistic for traditional TWAS is $Z_{stand} = \frac{\sqrt{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\sqrt{n}\hat{b}_1}{\hat{\sigma}_e}$. Thus, we obtain the relation: $Z_{adj} = Z_{stand}\frac{1}{\sqrt{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)}}$. Hence, $Z_{adj} = Z_{stand} \times A$, where the adjustment factor is $A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}}}{\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}'R\hat{\boldsymbol{\lambda}}}\right)}}$. To adjust for the uncertainty of

the predicted expression, we need to multiply the Z statistic used in the traditional TWAS estimate by the adjustment factor A. In one of the following sections, we discuss an approach to estimating the adjustment factor.

Estimation of the adjustment factor based on bootstrapping

Since the adjustment factor is unknown for a given gene, we need to estimate it. The adjustment factor is $A = \frac{\frac{1}{\sqrt{\hat{\boldsymbol{\lambda}}' R \hat{\boldsymbol{\lambda}}}}}{\sqrt{\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}' R \hat{\boldsymbol{\lambda}}}\right)}}$, where $\hat{\boldsymbol{\lambda}}$ is the estimated QTL effect based on the expression data, and R is the estimated LD-matrix of the local SNPs based on the GWAS data. Deriving a closed form formula of $\mathrm{E}\left(\frac{1}{\hat{\boldsymbol{\lambda}}' R \hat{\boldsymbol{\lambda}}}\right)$ is extremely challenging. Hence, we propose an estimation approach based on bootstrapping. In the analytical form of A, $\hat{\boldsymbol{\lambda}}$ but not R involves variability. We perform bootstrapping for the expression on local SNPs regression in the expression panel data. This regression (equation 1) is generally a penalized regression. The bootstrapped estimates of $\boldsymbol{\lambda}$ should correctly approximate the distribution of the full-sample estimate $\hat{\boldsymbol{\lambda}}$. We use the bootstrap expectation to estimate the denominator in A since the numerator is already calculated. There is a considerable amount of literature on bootstrapping for penalized regressions. We mainly focus on the residual bootstrap technique instead of a pair-wise bootstrap, assuming that the linear model assumption in equation 1 is correct. Chatterjee and Lahiri [12] showed that the estimates of $\boldsymbol{\lambda}$ based on standard residual bootstrap for Lasso regression might not converge in distribution to $\hat{\boldsymbol{\lambda}}$. Chatterjee and Lahiri [5] proved an important result that the usual residual bootstrap for adaptive Lasso [11] converges in distribution to the full sample estimate. In the same paper, they proposed a version of residual bootstrap for Lasso, which converges. However, since it is computationally very intensive, we implement residual bootstrap under the framework of adaptive Lasso [11]. The adaptive Lasso (ALasso) estimator of the local SNPs' effects on the expression is given by:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^p} \sum_{i=1}^{n} (E_i - \mathbf{u}_i' \boldsymbol{\lambda})^2 + w \sum_{j=1}^{p} \frac{|\lambda_j|}{|\hat{\lambda}_{j,ols}|^{\gamma}},$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ and $\hat{\boldsymbol{\lambda}}_{ols} = \{\hat{\lambda}_{j,ols}, 1 \leq l \leq n\}$ denotes the ordinary least squares estimate (OLS) of $\boldsymbol{\lambda}$ based on the full sample. Here, p denotes the number of local SNPs, w denotes the

weight associated with the penalty, and $\gamma > 0$ allows for varying the influence of the OLS estimate on the penalty term. We consider $\gamma = 1$ for simplicity in this work. The adaptive Lasso allows differential shrinkage of the effects of the local SNPs on the expression. We use five-fold cross-validation to choose the optimal w. Adaptive Lasso enjoys the oracle property, i.e., the probability of identifying the true subset of local SNPs having non-zero effects converges to one as the sample size increases, and the ALasso estimator asymptotically follows multivariate normal under some mild regularity conditions [11]. Next, we outline the residual bootstrap for ALasso.

Residual bootstrap for adaptive Lasso

As defined above, let $\hat{\boldsymbol{\lambda}}$ denote the full sample ALasso estimate of $\boldsymbol{\lambda}$. Define the ALasso residuals as: $\hat{\epsilon}_i = E_i - \mathbf{u}_i^{'}\hat{\boldsymbol{\lambda}}$, and the mean-centered residuals as $r_i = \hat{\epsilon}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{\epsilon}_j,\ 1 \leq i \leq n$. Let $(r_1^+, r_2^+, \ldots, r_n^+)$ denotes a random sample of size n drawn from $(r_1, r_2, \ldots, r_n)$ with replacement. Define $E_i^+ = \mathbf{u}_i^{'}\hat{\boldsymbol{\lambda}} + r_i^+,\ 1 \leq i \leq n$. Consider $E_i^+, 1 \leq i \leq n$ and the original $\mathbf{u}_i, 1 \leq i \leq n$ to implement ALasso regression:

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^p} \sum_{i=1}^{n}(E_i^+ - \mathbf{u}_i^{'}\boldsymbol{\lambda})^2 + w\sum_{j=1}^{p}\frac{|\lambda_j|}{|\tilde{\lambda}_{j,ols}|},$$

where $\tilde{\boldsymbol{\lambda}}_{ols} = \{\tilde{\lambda}_{j,ols}, 1 \leq l \leq n\}$ is the OLS estimate of $\boldsymbol{\lambda}$ based on the residual bootstrap data $(E_i^+, \mathbf{u}_i), 1 \leq i \leq n$. Thus, for a given residual bootstrap sample, $\tilde{\boldsymbol{\lambda}}$ provides the bootstrapped ALasso estimate of $\boldsymbol{\lambda}$. We use five-fold cross-validation to choose optimal w. We consider 300 residual bootstrap samples and obtain 300 bootstrap estimates of $\boldsymbol{\lambda}$: $\tilde{\boldsymbol{\lambda}}_1, \ldots, \tilde{\boldsymbol{\lambda}}_{300}$. We compute $\frac{1}{\tilde{\boldsymbol{\lambda}}_j^{'}R\tilde{\boldsymbol{\lambda}}_j}, 1 \leq j \leq 300$, where R is the LD matrix of local SNPs estimated based on the GWAS data. We use the sample mean of $\{\frac{1}{\tilde{\boldsymbol{\lambda}}_j^{'}R\tilde{\boldsymbol{\lambda}}_j}, 1 \leq j \leq 300\}$ to estimate the denominator in the adjustment factor. Since the proportion of the variance of expression explained by the local SNPs is sometimes limited, the true value of $\boldsymbol{\lambda}^{'}R\boldsymbol{\lambda}$ can be small. Hence, we also consider trimmed mean of $\{\frac{1}{\tilde{\boldsymbol{\lambda}}_j^{'}R\tilde{\boldsymbol{\lambda}}_j}, 1 \leq j \leq 300\}$ as a robust estimate of $\mathrm{E}\left(\frac{1}{\tilde{\boldsymbol{\lambda}}^{'}R\tilde{\boldsymbol{\lambda}}}\right)$.

Summary statistics version of standard TWAS

It may often be challenging to get access to individual-level GWAS data. Gusev et al. [1] proposed a summary statistics version of TWAS (summary TWAS) which only uses GWAS summary statistics. As a result, it can easily be implemented for various traits based on publicly available GWAS summary statistics. We focus on adjusting the Z statistic obtained from the summary TWAS. In the following, we outline the derivation of the standard summary TWAS Z statistic. Consider the estimated vector of local SNPs' effects on expression, $\hat{\boldsymbol{\lambda}}$, obtained from the regression in equation 1. Next, consider the trait regression on the estimated genetic component in GWAS (equation 3).

$$\hat{b} = (X^{'}X)^{-1}X^{'}Y = \frac{1}{n}(\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}})^{-1}\hat{\boldsymbol{\lambda}}^{'}G^{'}Y = \frac{\hat{\boldsymbol{\lambda}}^{'}G^{'}Y}{n\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}} \tag{4}$$

Suppose the trait Y is also individually regressed on the genotype of jth local SNP: $\mathrm{E}(Y) = \alpha_j g_j$, where α_j is the marginal effect of the SNP on the trait. If $\mathbf{g}_j$ denotes the normalized genotype vector of jth SNP, $\hat{\alpha}_j = (\mathbf{g}_j^{'}\mathbf{g}_j)^{-1}\mathbf{g}_j^{'}Y = \frac{1}{n}\mathbf{g}_j^{'}Y$. Let $\boldsymbol{\alpha}$ denote the vector of marginal effects of the local SNPs on the trait. Since, $\mathbf{g}_j^{'}Y = n\hat{\alpha}_j$, $G^{'}Y = n\hat{\boldsymbol{\alpha}}$. Thus, $\hat{b} = \frac{\hat{\boldsymbol{\lambda}}^{'}\hat{\boldsymbol{\alpha}}}{\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}}$. Note that $\hat{\boldsymbol{\lambda}}$ is considered fixed in the standard TWAS. $\mathrm{var}(\hat{b}) = \frac{\hat{\boldsymbol{\lambda}}^{'}\mathrm{cov}(\hat{\boldsymbol{\alpha}})\hat{\boldsymbol{\lambda}}}{(\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}})^2}$. Here, $\mathrm{cov}(\hat{\boldsymbol{\alpha}}) = \frac{\sigma_y^2}{n}R$, R is the LD matrix of the SNPs. Thus, $\mathrm{var}(\hat{b}) = \frac{\sigma_y^2}{n\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}}$ and $Z = \frac{\sqrt{n}\hat{\boldsymbol{\lambda}}^{'}\hat{\boldsymbol{\alpha}}}{\hat{\sigma}_y\sqrt{\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}}}$. Note that the Z statistic for testing $\alpha_j = 0$ in the GWAS is given by: $\frac{\hat{\alpha}_j}{s.e.(\hat{\alpha}_j)} = \frac{\sqrt{n}\hat{\alpha}_j}{\hat{\sigma}_y}$. Thus, the Z statistic for testing $b = 0$ versus $b \neq 0$ is given by:

$$Z_{\text{twas}} = \frac{\hat{\boldsymbol{\lambda}}^{'}Z_{\text{gwas}}}{\sqrt{\hat{\boldsymbol{\lambda}}^{'}R\hat{\boldsymbol{\lambda}}}} \tag{5}$$

Here Z_{gwas} denotes the vector of Z statistics for testing marginal effect vector $\boldsymbol{\alpha}$, obtained from the GWAS. We can use 1000 genome data to estimate R. To adjust for the uncertainty of the predicted expression, we propose to use $Z_{\text{twas}} \times \hat{A}$, where $\hat{A}$ is the estimated adjustment factor.

Multi-trait TWAS

We can implement a multi-trait TWAS approach based on the adjusted TWAS statistics. Suppose $(y_1, \ldots, y_k)^{'}$ denotes a multivariate trait, each of which can have a null or non-null effect due to the

true genetic component of expression: $\boldsymbol{y} = x \times \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = (y_1, \ldots, y_k)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)'$. The error term $(\epsilon_1, \ldots, \epsilon_k)'$ follows multivariate normal with zero mean vector and a dispersion matrix. We adopt the standard approach to compute the trait-specific TWAS Z statistic for $y_j = \beta_j x + \epsilon_j$, for each j at a time, $1 \leq j \leq k$, while adjusting for the uncertainty of predicted expression. The covariance matrix of the adjusted Z statistics across the traits remains the same as in the standard TWAS [13] because the trait-specific adjusted Z statistic is obtained by scaling the Z statistic obtained from the standard TWAS. To demonstrate our approach's usefulness in multi-trait TWAS, we implement the subset-based meta-analysis method ASSET [6], which requires the vector of Z statistics and its standard error (unity vector for Z statistics) and correlation matrix as the input. ASSET tests for the global null hypothesis: $H_0 : \beta_1 = \ldots = \beta_k = 0$ versus $H_1 : \cup_{j=1}^{k} \{\beta_j = 0\}$.

Simulation study

We perform extensive simulations to evaluate the proposed approach both for TWAS for univariate and multivariate traits. We contrast the type 1 error rates and powers estimated by the standard TWAS and the adjusted TWAS approach that accounts for the uncertainty of predicted expression. We implemented the hapgen software to simulate genotype data and use simulation models based on the genotype data to generate the expression and GWAS trait data.

Simulation design

The software package Fusion [1] analyzed the expression-genotype data in the Young Finish Sequencing (YFS) study to identify all the locally heritable genes in the whole blood tissue. For simplicity in computation, we selected one locally heritable gene, DGCR8, on chromosome 22. To run hapgen, we considered an African population's LD structure for the genomic region surrounding the gene. We consider 300 local SNPs, with MAF > 0.2, randomly selected from the 1MB local region surrounding the gene. To mimic the TWAS setting, we generate two independent samples of the expression panel and GWAS genotype data. We choose the sample size of the expression panel data as 500 and the sample size of the GWAS data as 10,000.

We simulate expressions based on the genotype data in the expression panel, where we normalize the genotype data of each local SNP. For the selected gene, we consider a linear model to simulate the expression: $E = \mathbf{u}'\boldsymbol{\lambda} + v$, where E denotes the expression, and $\mathbf{u}$ denotes the normalized genotype data of the local SNPs. Assuming that $\text{var}(E) = 1$, we consider $v \sim N(0, 1 - h_e^2)$, where h_e^2 is the local heritability of the expression due to the SNPs considered. We assume that 5% of the local SNPs affect the expression. If m_c denotes the number of such SNPs, each element of $\boldsymbol{\lambda}$ follows $N(0, \frac{h_e^2}{m_c})$. We consider $h_e^2 = 20\%$.

We next simulate the GWAS trait using the linear model $y = \beta_1 x + \epsilon$ based on genotype data in the GWAS. Here, y denotes the GWAS trait, $x = \mathbf{g}'\boldsymbol{\lambda}$ denotes the true genetic component of expression, and $\mathbf{g}$ is the normalized genotyped data in GWAS for the same set of local SNPs, which were also considered in the expression panel. Since we consider one gene, the genetic component of its expression is expected to have a small to moderate effect on the complex trait. We assume that $\text{var}(Y) = \text{var}(\epsilon) = 1$. We consider $\beta_1 = 0$ for estimating the type 1 error rate, and $\beta_1 = 0.02, 0.04, 0.06, 0.08, 0.1$ while estimating the power. In a given simulation setting, we consider the 0.05 level of significance to estimate the type 1 error rate and power based on 1000 iterations.

We use a similar setup for a multi-trait TWAS. We consider a linear model for four traits: $\boldsymbol{y} = x \times \boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} = (y_1, y_2, y_3, y_4)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$. We assume that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_4)'$ follows multivariate normal with the correlation matrix in which off-diagonal elements are generated at random in the range $(-0.4, 0.4)$. For simplicity, we assume that the same tissue type is causal for the traits. We test for the global null hypothesis $H_0 : \boldsymbol{\beta} = (0, 0, 0, 0)'$ versus the alternative hypothesis that the $\boldsymbol{\beta}$ vector is non-zero. To estimate the power, we consider $\boldsymbol{\beta} = (0, a, 0, a)'$, where $a = 0.02, 0.04, 0.06, 0.08, 0.1$.

Analysis pipeline

For a given iteration under a specific simulation scenario, we use the Fusion software package [1] to test whether the gene is locally heritable based on a p-value threshold $\frac{0.05}{20000}$. The simulated data in a specific iteration is analyzed only if the gene is locally heritable. We apply Lasso and Elastic net to the expression panel data to obtain a prediction model for the genetic component

of expression in traditional TWAS. Next, in the simulated GWAS data, summary statistics for the local SNPs are computed. Using the in-sample LD for the GWAS data, we calculate the Z statistic for traditional TWAS, one based on the predicted expression obtained by the Lasso-based estimated model and another based on Elastic net. To implement the adjusted TWAS, we first apply ALasso to the whole sample expression panel data and then use residual bootstrap to generate 300 residual estimates of $\boldsymbol{\lambda}$. Finally, we compute the adjusted TWAS statistic combining the whole sample and bootstrap estimates of $\boldsymbol{\lambda}$ and the in-sample GWAS LD matrix. For multi-trait TWAS, we calculate both the standard and adjusted TWAS statistics for each trait and run ASSET using the GWAS trait correlation matrix as an estimate of the correlation matrix of Z statistics [13].

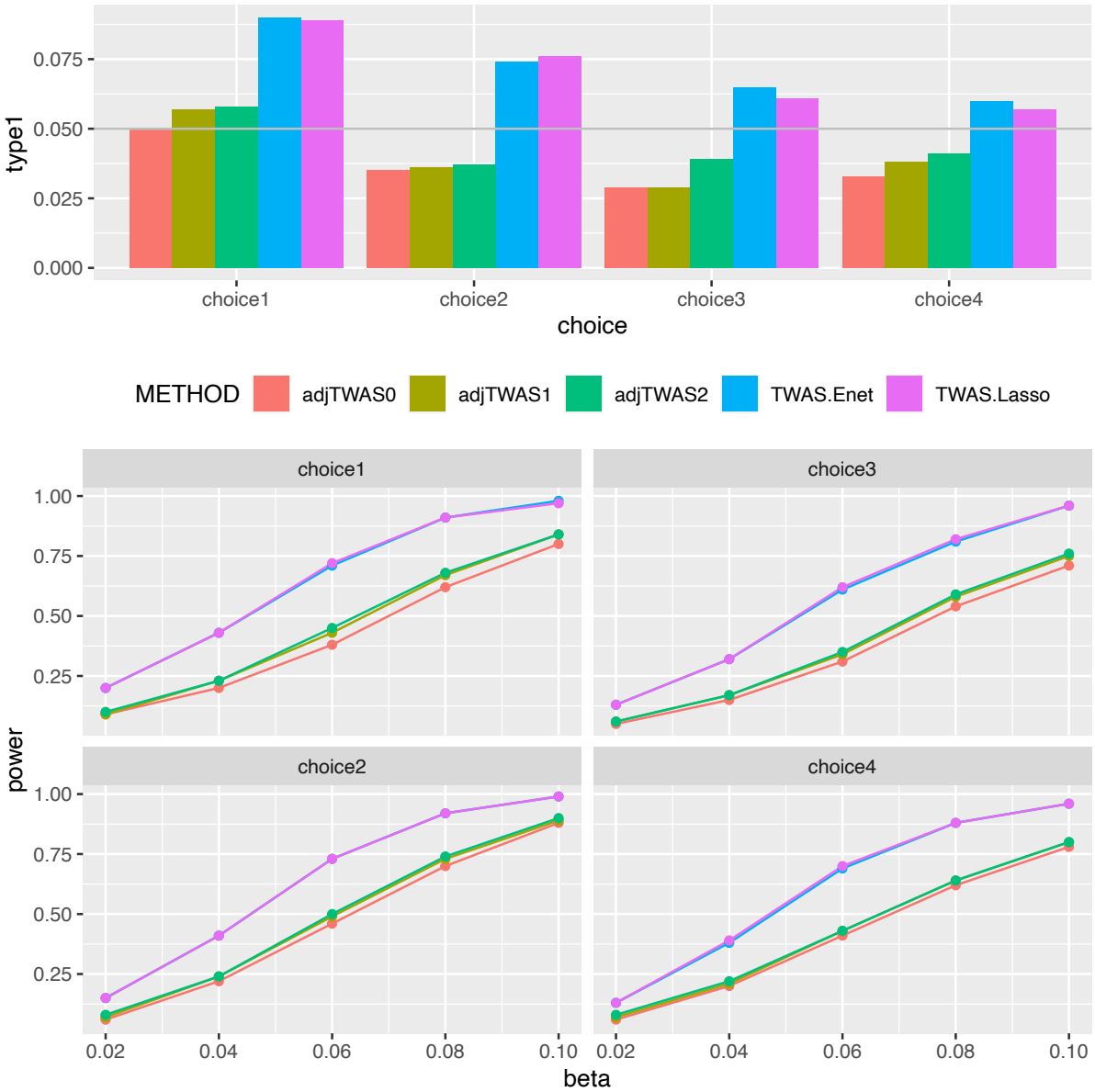
Simulation results

We observe that the standard TWAS approach produces an inflated type 1 error rate for univariate TWAS (Figure 1). We attribute the inflation to the lack of adjustment for the uncertainty of predicted expression. On the contrary, adjTWAS controls the type 1 error rate substantially better. Among the different choices for the trimming percentage to calculate the trimmed mean considered while estimating the adjustment factor for the adjusted TWAS statistic, the estimated type 1 error rate becomes less conservative for a more significant percentage of trimming. Thus, we can choose a higher trimming percentage to make the adjusted TWAS less conservative. However, excessive trimming may result in inadequate control of the type 1 error rate.

For example, in the first simulation scenario, the standard TWAS estimated the type 1 error rate as 0.09, but for adjTWAS, the estimated type 1 error rate is 0.05 for adjTWAS0 (0% trimmed mean), 0.057 for adjTWAS1 (5% trimmed mean), and 0.058 for adjTWAS2 (10% trimmed mean). The standard TWAS estimated the type 1 error rate in the second simulation scenario as 0.076 (Lasso-based) and 0.074 (Elastic net-based). But for adjTWAS, the estimated type 1 error rate is 0.035 for adjTWAS0, 0.036 for adjTWAS1, and 0.037 for adjTWAS2. Concerning power, the standard TWAS approach, based on Lasso and Elastic net, produces higher power than the adjTWAS. The power gain is mainly due to an inflated false positive rate and is not statistically justified. As expected, the power loss decreases for adjTWAS, which uses a trimmed mean with a higher percentage. In

first simulation scenario, for $\beta_1 = 0.04$, adjTWAS0, adjTWAS1, adjTWAS2 produced around 20% lower power compared to the standard TWAS using Lasso or Elastic net.

Figure 1: Comparison of type 1 error rate and power between the standard TWAS and adjusted TWAS for the univariate trait. adjTWAS0, adjTWAS1, adjTWAS2 denote the adjusted TWAS approaches that use $0\%, 5\%, 10\%$ trimmed means when estimating the adjustment factor.



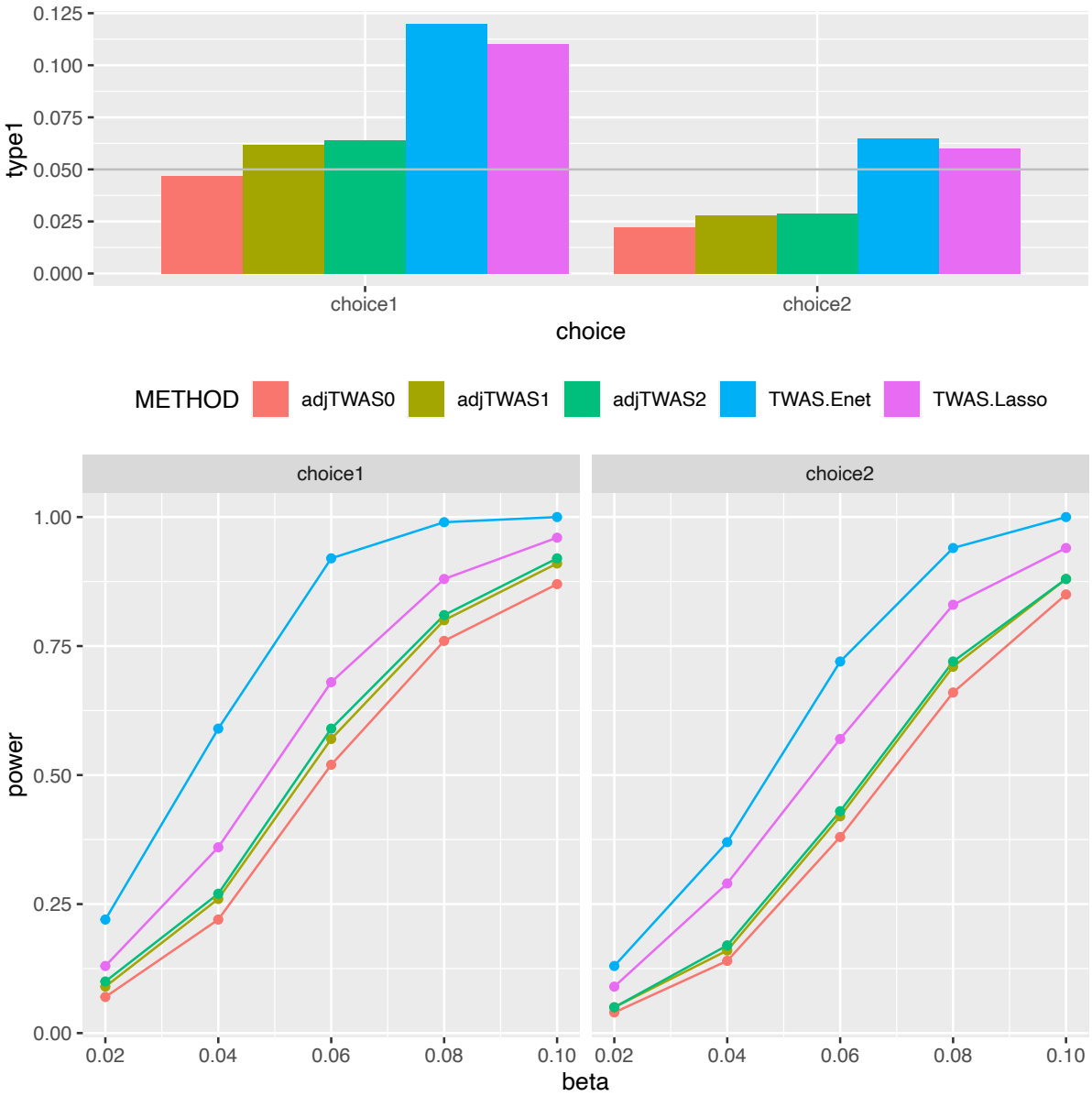
For multi-trait TWAS, we observe a similar comparative performance of the methods. Using traditional TWAS statistics, ASSET produced an inflated type 1 error rate (Figure 2). But it

controls the type 1 error rate substantially better using the adjTWAS statistics. Again, a more significant trimming percentage used in adjTWAS produces a less conservative estimate of the type 1 error rate. For example, in the first simulation scenario considered for the multi-trait TWAS, ASSET estimated a type 1 error rate of 0.11 using the standard TWAS statistic (Lasso-based). On the contrary, it produced a type 1 error rate of 0.047, 0.062, and 0.064 using the statistics obtained by adjTWAS0, adjTWAS1, and adjTWAS2. Concerning power, it has a higher power based on standard TWAS statistics than adjTWAS statistics, mainly due to an inflated false positive rate. For example, in the first simulation scenario ($\beta_1 = \beta_3 = 0, \beta_2 = \beta_4 = 0.04$), ASSET based on standard TWAS statistics produces $32-37\%$ higher power than adjTWAS at the expense of severely inflated type 1 error rate. As expected, multi-trait TWAS has higher power than univariate TWAS when using adjusted TWAS statistics in both cases.

Discussion

The standard TWAS approach suffers from the inherent methodological limitation that the uncertainty or variability of predicted expression is not adjusted. We propose a novel approach to adjust for the uncertainty of predicted expression. Our method uses techniques from measurement error theory to derive the adjustment factor that should be multiplied by the standard TWAS statistic. We propose an approach based on bootstrapping for high-dimensional regression to estimate the adjustment factor. We also extend the approach to a multi-trait TWAS setting. Using simulations, we show that the standard TWAS approach can produce an inflated false positive rate. As a result, a substantial proportion of the gene-trait associations identified by standard TWAS can be false positive signals. Our approach adequately controls the false positive rate both for a univariate and multi-trait TWAS. Due to better control of the type 1 error rate, our method produces lower power than the standard TWAS approach. In summary, we propose a TWAS approach that adjusts for the predicted expression's uncertainty and is a statistically valid test for gene-trait association.

Figure 2: Comparison of type 1 error rate and power produced by ASSET using the standard TWAS and adjusted TWAS statistics for a multivariate trait. adjTWAS0, adjTWAS1, adjTWAS2 denote the adjusted TWAS approaches that use $0\%, 5\%, 10\%$ trimmed means when estimating the adjustment factor.



References

[1] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245, 2016.

[2] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[3] Nicholas Mancuso, Simon Gayther, Alexander Gusev, Wei Zheng, Kathryn L Penney, Zsofia Kote-Jarai, Rosalind Eeles, Matthew Freedman, Christopher Haiman, and Bogdan Pasaniuc. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nature communications*, 9(1):1–11, 2018.

[4] Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

[5] Arindam Chatterjee and Soumendra Nath Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

[6] Samsiddhi Bhattacharjee, Preetha Rajaraman, Kevin B Jacobs, William A Wheeler, Beatrice S Melin, Patricia Hartge, Meredith Yeager, Charles C Chung, Stephen J Chanock, Nilanjan Chatterjee, et al. 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.

[7] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[8] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[10] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[11] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[12] Arindam Chatterjee and S Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, 2010.

[13] Helian Feng, Nicholas Mancuso, Bogdan Pasaniuc, and Peter Kraft. Multitrait transcriptome-wide association study (twas) tests. *Genetic Epidemiology*, 45(6):563–576, 2021.