*mathematics*

MDPI

*Article*

# Data-Adaptive Multivariate Test for Genomic Studies Using Fused Lasso

**Masao Ueki** [1,2]

[1] School of Information and Data Sciences, Nagasaki University, 1-14 Bunkyo-machi, Nagasaki 852-8521, Japan; uekimrsd@nifty.com
[2] RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

**Abstract:** In genomic studies, univariate analysis is commonly used to discover susceptible variants. It applies univariate regression for each variant and tests the significance of the regression coefficient or slope parameter. This strategy, however, may miss signals that are jointly detectable with other variants. Multivariate analysis is another popular approach, which tests grouped variants with a predefined group, e.g., based on a gene, pathway, or physical location. However, the power will be diminished if the modeling assumption is not suited to the data. Therefore, data-adaptive testing that relies on fewer modeling assumptions is preferable. Possible approaches include a data-adaptive test proposed by Ueki (2021), which applies to various data-adaptive regression models using a generalization of Yanai's generalized coefficient of determination. While several regression models are possible choices for the data-adaptive test, this paper focuses on the fused lasso that can count for the effect of adjacent variants and investigates its performance through comparison with other existing tests. Simulation studies demonstrate that the test using fused lasso has a high power compared to the existing tests including the univariate regression test, saturated regression test, SKAT (sequence kernel association test), burden test, SKAT-O (optimized sequence kernel association test), and the tests using lasso, ridge, and elastic net when assuming a similar effect of adjacent variants.

**Keywords:** fused lasso; genomic studies; multivariate test; Yanai's generalized coefficient of determination

**MSC:** 62P10; 62H15; 62-08



**Citation:** Ueki, M. Data-Adaptive Multivariate Test for Genomic Studies Using Fused Lasso. *Mathematics* **2024**, *12*, 1422. https://doi.org/10.3390/math12101422

Academic Editors: Wei Wu, Chang-Xing Ma and Li-Pang Chen

Received: 16 February 2024
Revised: 15 April 2024
Accepted: 3 May 2024
Published: 7 May 2024



## 1. Introduction

Univariate analysis is often used in genome-wide association studies [1,2]. Univariate analysis is based on univariate regression for each variant with a target phenotype and tests the significance of the regression coefficient or slope parameter. The use of the univariate regression model, however, restricts the alternative model to being overly simple, meaning it is unable to capture complex phenomena as the model consists of only a few parameters, such as situations where single genetic variants are independently associated with the disease. The strategy may miss signals jointly detectable with other variants.

Multivariate analysis is another popular approach. It tests grouped variants simultaneously. The groups are predefined by users based on gene, pathway, or physical location, etc. [3–7]. The typical statistical methods include burden test [8] and SKAT (sequence kernel association test) [6]. Burden test collapses rare variants in a region into a single variable, and subsequently, this variable is tested for association with the phenotype. SKAT aggregates the associations between variants and the phenotype through a kernel matrix, which is derived as a variance-component test in the mixed models where regression coefficients are assumed to be independent and follow a distribution with the variance component or a random effect. Burden test is powerful when a large proportion of variants are causal and associated with a trait with the same direction of effect. SKAT (optimized sequence kernel association test) is powerful in the presence of both positive and negative effects

of variants in a genomic region. SKAT-O is a omnibus test that combines burden test and SKAT data-adaptively by a linear combination of SKAT and burden test statistics and the optimal combination is found by minimizing the $p$-value [9]. A detailed review of burden test, SKAT and SKAT-O is given in Lee et al. [10]. Since the disease-development mechanism is unknown a priori for many complex diseases, it is often difficult to specify an appropriate method or model in exploring susceptibility genes or variants. There exist many data-adaptive approaches, such as Sham and Curtis [11], Hirotsu et al. [12], Freidlin et al. [13], González et al. [14], Li et al. [15], Hothorn and Hothorn [16], Joo et al. [17], Zang and Fung [18], Ueki [19], but they are not versatile because null distribution specific to each test is often required, which differs from familiar tests whose null distribution is normal or chi-squared. It is necessary to develop special algorithms to compute the null distributions or expensive numerical procedures such as permutation tests.

Recently, Ueki [20] developed a data-adaptive test based on Yanai's generalized coefficient of determination [21,22]. The test is based on the regression model that maximizes Yanai's generalized coefficient of determination [21,22], generalizing it to any modeling procedure. Yanai's generalized coefficient of determination is proportional to the covariance between a response variable and its predicted value divided by the square root of the generalized degrees of freedom [23], and the dimension of the selected model tends to be large under the null hypothesis of no effect. The above characteristic under the null hypothesis enables the type I error rate to be controlled approximately with the significance threshold for the saturated model, without having test-specific null distributions. Since it is simple and simulation-free in computing $p$-value, the data-adaptive test is readily applicable to genome-wide scans as a multivariate test for assessing grouped variants. Actually, Ueki [20] applied it to lasso [24], ridge [25], and elastic net [26] for flexible data-adaptive variant discovery in real genomic study data. Among them, variable selection approaches, i.e., lasso and elastic net, can automatically remove variants irrelevant to predicting the phenotype by setting the corresponding regression coefficients to zero while accounting for correlations between variants. This data-adaptive filtering of variants helps to interpret the results.

Groups of variants are sometimes made based on physical location. This in turn implies that the adjacency between variants could be useful information to further enhance the power of detecting variant sets associated with phenotype. Fused lasso [27] is a popular penalized regression model that allows explicitly incorporating adjacency information in grouped variants in the variable selection scheme. The assumption that the adjacent variants may have a similar effect is sometimes considered in the existing literature [10,28,29]. Bao and Wang [29] proposed genome-wide association studies using a penalized moving-window regression with a fused lasso-like penalty [30] to incorporate the adjacency of variants and linkage disequilibrium.

This paper considers the fused lasso for the data-adaptive test of Ueki [20] and investigates the performances as a multivariate test for genomic studies through simulation studies. The fused lasso test is compared with the univariate regression test, saturated regression test, SKAT, burden test, SKAT-O [6,9], and the data-adaptive tests of Ueki [20] using lasso, ridge, and elastic net. Simulation studies compare the above group tests. Real genotype data from the 1000 Genomes Project is used for simulating the phenotype in genomic analyses.

The rest of this paper is organized as follows. Section 2 describes the methods including descriptions of the testing procedure of Ueki [20], its application to penalized regression including the fused lasso, and the description of simulation studies. Section 3 describes the results of the simulation studies. Section 4 concludes the paper.

## 2. Methods

### 2.1. Test Based on Yanai's Generalized Coefficient of Determination

This section presents the test developed by Ueki [20]. Suppose that $n$ samples are observed where a response variable $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $d$ explanatory variables

$\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ are collected for each sample, in which $\boldsymbol{X}_j = (x_{1j}, \ldots, x_{nj})^T$ for $j = 1, \ldots, d$. For application to genomic studies, the response variable $\boldsymbol{y}$ corresponds to quantitative phenotype and $\boldsymbol{X}$ is the grouped variants to be tested for association with $\boldsymbol{y}$. Consider a set of regression models indexed by a tuning parameter $\lambda$, $g_\lambda(\boldsymbol{y})$ that models the conditional expectation $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{X}) = E(\boldsymbol{y}|\boldsymbol{X})$ given $\boldsymbol{X}$. It contains models typically ordered by the extent of complexity controlled by $\lambda$, which eventually tends to the saturated model as $\lambda \to 0$. The saturated model is given at $\lambda = 0$, i.e., $g_0(\boldsymbol{y}) = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$, where $\boldsymbol{P}_{\boldsymbol{X}}$ is the projection matrix onto $\boldsymbol{X}$. The model sequence considered includes the lasso, ridge, elastic net, fused lasso, generalized lasso, and many other regression models in statistics or machine learning. To be specific, the testing procedure applies to the model sequence $g_\lambda(\boldsymbol{y})$. The procedure includes a model selection step based on a generalization of Yanai's generalized coefficient of determination [21]. Yanai's generalized coefficient of determination is a measure of similarity between two linear spaces and has been used for variable selection in principal component analysis [31]. A description of the original Yanai's generalized coefficient of determination and its generalized version are presented in Appendix A.

Instead of $\boldsymbol{y}$ and $\boldsymbol{X}$, centered variables $\widetilde{\boldsymbol{y}} = \boldsymbol{Q}_{\boldsymbol{1}_n}\boldsymbol{y}$, $\widetilde{\boldsymbol{\mu}} = \boldsymbol{Q}_{\boldsymbol{1}_n}\boldsymbol{\mu}$, and $\widetilde{\boldsymbol{X}} = \boldsymbol{Q}_{\boldsymbol{1}_n}\boldsymbol{X}$ are considered. Here, $\boldsymbol{Q}_{\boldsymbol{1}_n} = \boldsymbol{I}_n - n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^T$, $\boldsymbol{I}_n$ is the $n$th identity matrix, and $\boldsymbol{1}_n$ is the $n$-vector of ones. Let $\{g_\lambda(\widetilde{\boldsymbol{y}}) : \lambda \geq 0\}$ be a model sequence indexed by a tuning parameter $\lambda \geq 0$, where the saturated model at $\lambda = 0$ is given by $g_0(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}$. Then, Yanai's generalized coefficient of determination for a modeling procedure $g_\lambda$ is given by

$$r(\widetilde{\boldsymbol{y}}, g_\lambda) = \frac{||\widetilde{\boldsymbol{y}}||^{-2}\widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})}{\mathrm{gdf}_0(g_\lambda)^{1/2}}, \tag{1}$$

where

$$\mathrm{gdf}_0(g_\lambda) = E_{\widetilde{\boldsymbol{\mu}}=\boldsymbol{0}}\{\widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})\}, \tag{2}$$

and $E_{\widetilde{\boldsymbol{\mu}}=\boldsymbol{0}}$ indicates the expectation under the assumption of $\widetilde{\boldsymbol{\mu}} = \boldsymbol{0}$.

The quantity $\mathrm{gdf}_0(g_\lambda)$ coincides with the generalized degrees of freedom of $g_\lambda$ defined by $\mathrm{cov}\{\widetilde{\boldsymbol{y}}, g_\lambda(\widetilde{\boldsymbol{y}})\} = E\{(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^T g_\lambda(\widetilde{\boldsymbol{y}})\}$ [23,32] under the null hypothesis $\boldsymbol{\mu} = \alpha_0\boldsymbol{1}_n$ because of $\widetilde{\boldsymbol{\mu}} = \boldsymbol{Q}_{\boldsymbol{1}_n}\boldsymbol{\mu} = \boldsymbol{0}$, where $\alpha_0$ is an intercept parameter. For least-squares regression with explanatory variables $\widetilde{\boldsymbol{X}}_s$, the sub-matrix of $\widetilde{\boldsymbol{X}}$ consisting of the column vectors $(\widetilde{\boldsymbol{X}}_j)_{j\in s}$ in a given index set $s \subset \{1, \ldots, d\}$, the generalized degrees of freedom is given by $\mathrm{tr}(\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s}) = |s|$ [23], and consequently, (1) reduces to the original Yanai's generalized coefficient of determination (A1) presented in Appendix A. The quantity (1) possesses a property that the expectation of $r(\widetilde{\boldsymbol{y}}, g_\lambda)$ under the null hypothesis $\boldsymbol{\mu} = \alpha_0\boldsymbol{1}_n$ is approximately proportional to $\mathrm{gdf}_0(g_\lambda)^{1/2}$ if $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma_0^2\boldsymbol{I}_n)$. Therefore, by assuming that $\mathrm{gdf}_0(g_\lambda) \leq d$, because of the assumption $g_0(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}$, the model that achieves the maximum, $\max_\lambda r(\widetilde{\boldsymbol{y}}, g_\lambda)$, may have a large dimensionality and is close to that of the saturated model with a high probability. In contrast, under the alternative hypothesis of $\boldsymbol{\mu} \neq \alpha_0\boldsymbol{1}_n$, assuming that $g_\lambda(\widetilde{\boldsymbol{y}}) \approx \widetilde{\boldsymbol{\mu}}$, the expectation of $r(\widetilde{\boldsymbol{y}}, g_\lambda)$ is approximately proportional to $||\widetilde{\boldsymbol{\mu}}||^2/\mathrm{gdf}_0(g_\lambda)^{1/2}$, and the model with the smallest $\mathrm{gdf}_0(g_\lambda)^{1/2}$ is chosen.

Let the significance threshold for the hypothesis test be $\alpha \in (0, 1)$. For a given model sequence $\{g_\lambda(\widetilde{\boldsymbol{y}}) : \lambda \geq 0\}$, the selected model is the model at the tuning parameter that maximizes the Yanai's generalized coefficient of determination, $\widehat{\lambda}^*$, i.e.,

$$\widehat{\lambda}^* = \mathrm{argmax}_\lambda r(\widetilde{\boldsymbol{y}}, g_\lambda).$$

Exploiting the property that the selected model by the Yanai's generalized coefficient of determination tends to be the saturated model, or $\widehat{\lambda}^* \approx 0$ under the null hypothesis $\widetilde{\boldsymbol{\mu}} = \mathbf{0}$, the proposed test procedure is to reject the null hypothesis when

$$\frac{||\widetilde{\boldsymbol{y}}||^2 d^{(1-\delta)/2} r(\widetilde{\boldsymbol{y}}, g_{\widehat{\lambda}^*})/d}{\hat{\sigma}_y^2} > \bar{F}_\alpha^{-1}(d) \quad \text{if} \quad \mathrm{gdf}_0(g_{\widehat{\lambda}^*}) < d^{1-\gamma}, \tag{3}$$

$$\frac{||\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}||^2/d}{\hat{\sigma}_y^2} > \bar{F}_\alpha^{-1}(d) \quad \text{if} \quad \mathrm{gdf}_0(g_{\widehat{\lambda}^*}) \geq d^{1-\gamma}, \tag{4}$$

where $\bar{F}_\alpha^{-1}(d)$ is the $(1-\alpha)$th quantile of the $F$-distribution with $(d, n-d-1)$ degrees of freedoms, $\delta$ is a small constant set as $1/d$, and $\hat{\sigma}_y^2 = ||\boldsymbol{Q}_{(\mathbf{1}_n, \boldsymbol{X})}\boldsymbol{y}||^2/(n-d-1)$.

The rationale of the above test procedure, (3) and (4), is described in what follows. First, note that, at $\lambda = 0$,

$$\frac{||\widetilde{\boldsymbol{y}}||^2 d^{1/2} r(\widetilde{\boldsymbol{y}}, g_\lambda)/d}{\hat{\sigma}_y^2} = \frac{d^{1/2}\widetilde{\boldsymbol{y}}^T g_0(\widetilde{\boldsymbol{y}})/d}{\mathrm{gdf}_0(g_0)^{1/2}\hat{\sigma}_y^2} = \frac{||\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}||^2/d}{\hat{\sigma}_y^2}, \tag{5}$$

which is also the left-hand side of (3) in the case when $\widehat{\lambda}^* = 0$ and $\delta = 0$. The quantity $\frac{||\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}||^2/d}{\hat{\sigma}_y^2}$ above is the usual $F$-statistic under the saturated model, that is, it follows an $F$-distribution with $(d, n-d-1)$ degrees of freedoms if $\boldsymbol{y} \sim N(\alpha_0 \mathbf{1}_n, \sigma_0^2 \boldsymbol{I}_n)$. Because, under the null hypothesis, the generalized degrees of freedom corresponding to $\max_\lambda r(\widetilde{\boldsymbol{y}}, g_\lambda)$ are close to $d$ with a high probability if $d$ is large, the saturated model $F$-statistic can be used if the generalized degrees of freedom of the selected model is close to $d$, which is the second case (4). In other cases where the generalized degrees of freedom of the selected model are not close to $d$, which rarely occurs when the null hypothesis is true, the first case (3) is used for the test.

The parameter $\gamma$ is a given constant in $(0, 1)$, which plays a role to judge whether the selected model is close to the saturated model, and is set as $\gamma = 0.01$ according to Ueki [20]. The parameter $\delta$ is introduced to alleviate slight inflation in test statistic due to finite $d$ observed in preliminary numerical experiments, which is arbitrarily set as $\delta = 1/d$ to satisfy $\delta \to 0$ as $d \to \infty$. Under certain regularity conditions, the above test procedure with a significance level $\alpha$ for the null hypothesis $\boldsymbol{\mu} = \alpha_0 \mathbf{1}_n$ may approximately maintain the type I error rate at $\alpha$ as $d/n$ tends to a fixed constant in $(0, 1)$ when $n \to \infty$. See Section 2.4 of Ueki [20] for the theoretical results.

### *2.2. Data-Adaptive Test Using Penalized Regression*

Penalized regression models such as lasso, ridge, elastic net and fused lasso are regarded as a model sequence, $\{g_{(\lambda_1,\ldots,\lambda_l)}(\widetilde{\boldsymbol{y}}) : \lambda_1 \geq 0, \ldots, \lambda_l \geq 0\}$, where $\lambda_1, \ldots, \lambda_l$ are $l$ non-negative tuning parameters to control model complexity. The penalized regression models are written in a unified manner as the following minimization problem:

$$\mathrm{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} ||\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}||^2 + \mathrm{pen}_{(\lambda_1,\ldots,\lambda_l)}(\boldsymbol{\beta}) \right\}, \tag{6}$$

where $\mathrm{pen}_{(\lambda_1,\ldots,\lambda_l)}(\boldsymbol{\beta})$ is a penalty function of $\boldsymbol{\beta}$ with tuning parameters $\lambda_1, \ldots, \lambda_l$. Various procedures are obtained by altering the penalty function as follows:

- Lasso: $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^d |\beta_j|$.
- Ridge: $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^d |\beta_j|^2$.
- Elastic net: $\mathrm{pen}_{(\lambda_1,\lambda_2)}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^d |\beta_j| + \lambda_2 \sum_{j=1}^d |\beta_j|^2$.
- Fused lasso: $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=2}^d |\beta_j - \beta_{j-1}|$.

The lasso and elastic net produce exactly zero regression coefficients, while the ridge does not do so. Thus, the lasso and elastic net are more suitable than the ridge regression

when there exist redundant explanatory variables in the $d$ variables that give zero regression coefficient. The fused lasso is suited to the data where indexes of explanatory variables have an order, such as physical location. The fused lasso penalizes successive differences of regression coefficients in absolute value and makes the regression coefficients identical for some of the adjacent variants depending on the value of $\lambda$. This feature is absent in lasso, ridge, and elastic net, and there exist several applications of the fused lasso in different fields [33–35]. The fact that the fused lasso can explicitly handle the adjacency between variants suggests an application to multivariate tests for genome-wide association studies. For example, it would be useful when investigators attempt to assume a similar variant effect in unknown subregions in the studied gene region.

To conduct the data-adaptive test with the penalized regression, the generalized degrees of freedom method (2) is required for computing the Yanai's generalized coefficient of determination (1). Fortunately, estimates for the generalized degree of freedom are available for the above penalized regression models. For the ridge regression, $g_\lambda(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_\lambda \widetilde{\boldsymbol{y}}$ where $\boldsymbol{P}_\lambda = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + n\lambda \boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{X}}^T$, the generalized degrees of freedom are given explicitly as $\mathrm{gdf}_0(g_\lambda) = \mathrm{tr}(\boldsymbol{P}_\lambda)$ if $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma_0^2 \boldsymbol{I}_n)$, which can be used as an estimate. Analogously, for the lasso, $\mathrm{gdf}_0(g_\lambda) = E(|A_\lambda|)$ holds [36–38], where $A_\lambda$ is the active set at a given tuning parameter $\lambda$, hence, the cardinality $|A_\lambda|$ can be used as an estimate. More generally, for the elastic net with a tuning parameter vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ (the first and second elements are for $L_1$- and $L_2$-norms), $\mathrm{tr}\{\widetilde{\boldsymbol{X}}_{A_\lambda}(\widetilde{\boldsymbol{X}}_{A_\lambda}^T\widetilde{\boldsymbol{X}}_{A_\lambda} + n\lambda_2 \boldsymbol{I}_{|A_\lambda|})^{-1}\widetilde{\boldsymbol{X}}_{A_\lambda}^T\}$ can be used as an estimate, where $A_{\boldsymbol{\lambda}}$ is the active set at a given tuning parameter $\boldsymbol{\lambda}$. The form of the generalized degrees of freedom for the generalized lasso including the fused lasso is given by Tibshirani and Taylor [37], where the generalized lasso is the penalized regression (6) with a penalty function $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \lambda||\boldsymbol{D}\boldsymbol{\beta}||_1$ for a given specified penalty matrix $\boldsymbol{D}$. It includes lasso and fused lasso as a special case. The generalized degrees of freedom for other models are given in Chen et al. [39]. If no closed-form estimate is available, a simulation-based method is a possible approach [23]. For the fused lasso, the generalized degrees of freedom is equal to the expected number of fused groups [37]. This paper uses the number of fused groups of the estimated regression coefficients as an estimate of the generalized degrees of freedom.

### *2.3. Other Multivariate Tests*

Other methods for the multivariate tests for a given group comprising of $d$ variants considered in this paper are as follows.

- Univariate regression test: Minimum of the $d$ Bonferroni adjusted $p$-values from univariate $F$-test for each variant, i.e., $\min\{d\min(p_1, \ldots, p_d), 1\}$, where $p_j$ denotes the $p$-value for testing if the regression coefficient is zero in the univariate normal linear regression model with the phenotype $y$ and the $j$th variant as the explanatory variable ($j = 1, \ldots, d$). The Bonferroni correction is the method that adjusts the significance level of individual tests to level $\alpha/d$, where $\alpha$ is the desired family-wise error rate [40], which gives the Bonferroni adjusted $p$-value as defined in Wright [41]. The univariate regression test does not take the correlation between $d$ variants into account.
- Saturated regression test: $F$-test for the analysis of variance under the saturated linear regression model with normal error using all $d$ variants simultaneously, i.e., test for the null hypothesis $H_0 : \beta_1 = \cdots = \beta_d = 0$ with the saturated normal linear regression model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d + \epsilon$ for the phenotype $y$ and the $d$ variants in a given group, where $\beta_j$ is the regression coefficient for $x_j$, $\beta_0$ is the intercept, and $\epsilon$ is the normal error with mean zero and nonzero variance. The saturated regression test is susceptible to the "curse of dimensionality", i.e., when $d$ is large relative to $n$, and cannot be used when $d > n$.
- SKAT: SKAT is developed for analysis of association between variants in a region and a phenotype [6]. It can be seen as a variance component test in the induced mixed models where regression coefficients are assumed to be independent and follow a distribution with the variance component or a random effect. The statistic of the SKAT forms $(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0)$, where $\widehat{\boldsymbol{\mu}}_0$ is the estimated mean under the null

hypothesis, and $\boldsymbol{K}$ is an $n \times n$ kernel matrix with genotype data in the region, and a given prespecified weight to give a higher weight for rarer variant [6]. It is known to be robust when variants in a genomic region have both positive and negative effects. `SKAT` function in R package `SKAT` is used with default option which conducts the SKAT test of Wu et al. [6].

- Burden test: Burden tests collapse rare variants in a genetic region into a single burden variable, and then the burden variable is tested for an association with the phenotype in the region. It is a score test for an aggregated effect of $d$ variables [8,42], which is made by combining minor allele counts in the region into a single variable. The burden test is powerful if a large proportion of the rare variants in a region are truly causal and influence the phenotype in the same direction. `SKAT` function in R package `SKAT` is used with option `r.corr=1`.
- SKAT-O: A combination of SKAT and burden test [9]. SKAT-O considers optimal test of the form $(1-\rho)Q_1 + \rho Q_2$, where $Q_1$ and $Q_2$ are the SKAT and burden test statistics, respectively, and $\rho$ is a parameter between 0 and 1 to optimally combine $Q_1$ and $Q_2$. An optimal $\rho$ is found by minimizing the $p$-value computed based on $(1-\rho)Q_1 + \rho Q_2$ with respect to $\rho$ [9]. `SKAT` function in R package `SKAT` is used with option `method="SKATO"`.

### *2.4. Description of Simulation Studies*

The simulation studies aim to investigate the power of the data-adaptive test using fused lasso as a multivariate test for multiple single nucleotide variants under various settings through comparison with the above multivariate tests, i.e., univariate regression test, saturated regression test, SKAT, burden test, SKAT-O, and the data-adaptive tests using lasso, ridge, and elastic net. The investigation includes assessing the type I error rates of the multivariate tests.

The simulations consider two different scenarios to generate $d$ single nucleotide variants $\boldsymbol{X} = (x_{ij})_{i=1,\ldots,n;j=1,\ldots,d}$ for $n$ individuals. The two scenarios are as follows.

- Genotypes using 1000 Genomes Project data: For the simulation, whole genome sequencing data from the 1000 Genomes Project, phase 3, is used [43]. A total of 493 individuals from the European population (Utah Residents (CEPH) with Northern and Western European ancestry, i.e., Toscani from Italy, Finnish from Finland, British from England and Scotland, and Iberian from Spain) are extracted. In total, 3,837,178 single nucleotide variants of chromosome 10 are used. Chromosome 10 is often used to evaluate statistical methods that account for linkage disequilibrium in human genetics, e.g., [44,45], and is therefore suitable to evaluate the methods under a practical correlation structure in genotypes. The following quality control is applied: excluding loci with missing rates $> 0.99$, Hardy–Weinberg equilibrium test $p$-value $< 10^{-5}$, or minor allele frequency $< 0.05$. Then, the pruning based on linkage disequilibrium is applied by the PLINK software version 1.9 using the `--indep-pairwise 50 5 0.99` option, resulting in 143,222 variants. From those variants, we randomly choose a set of $d$ contiguous variants as $\boldsymbol{X} = (x_{ij})_{i=1,\ldots,493;j=1,\ldots,d}$, in which $x_{ij}$ denotes the number of minor alleles, i.e., $x_{ij} \in \{0, 1, 2\}$. Missing genotypes are replaced by the mean of each locus. Fixed sample size $n = 493$ is used, and two scenarios for the number of variants $d = 50$ or 100 are considered. To see the effect of correlation between variants, an additional simulation is carried out with genotypes that are randomly shuffled for each locus, eliminating the correlation between the variants.
- Simulated genotypes under exchangeable correlation structure: First, $d$ minor allele frequencies are randomly generated from a uniform distribution in $[0.05, 0.5]$. Then, $d$ correlated binary variables (i.e., 0 or 1) are generated using `bindata` package for R with variance-covariance matrix $\boldsymbol{S}$ independently for $i = 1, \ldots, 2n$, in which $\boldsymbol{S}$ is the $d \times d$ matrix with off-diagonal and diagonal elements of $\rho$ and 1, respectively, giving a $2n \times d$ binary matrix. Let its rows 1 to $n$, and rows $(n+1)$ to $2n$ be $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, respectively. Then, an $n \times d$ genotype matrix is made by $\boldsymbol{X} = \boldsymbol{X}^{(1)} + \boldsymbol{X}^{(2)}$,

whose elements take values in $\{0,1,2\}$. It is equivalent to the situation where the genotypes are under the Hardy–Weinberg equilibrium. Three scenarios for the pairs of sample size and number of variants are considered, $(n,d) = (400,50)$, $(800,100)$, and $(1200,150)$, the aim of which is to confirm type I error control as $n \to \infty$ while $d/n$ is kept constant. Two scenarios $\rho = 0.3$ and 0.7 are also considered.

Given the genotype matrix $\boldsymbol{X}$ with $d$ variants made in each of the above scenarios, quantitative phenotype is generated by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, where $\epsilon_1, \ldots, \epsilon_n$ are generated independently and identically from standard normal distribution. Three scenarios for the regression coefficients $\boldsymbol{\beta}_0$ are considered, as follows:

- Random: This scenario is considered for comparing the power of the multivariate tests. Given $r_\beta \in (0,1)$, let $d_0 = \lfloor dr_\beta \rfloor$ variants have nonzero regression coefficients and remaining $d - d_0$ variables have zero regression coefficients, and the $d_0$ variants are randomly selected from the $d$ variants. Now, $100 \times r_\beta$% have nonzero regression coefficients among $d$ variants. Then, $d_0$ nonzero regression coefficients are independently generated from normal distribution $N(0, \sigma_\beta^2)$, with standard deviation $\sigma_\beta = 0.005/r_\beta$, multiplied by $1/\{2\text{MAF}(1-\text{MAF})\}^{1/2}$ where MAF denotes the minor allele frequency of the corresponding variant. The above scheme gives larger variance of nonzero regression coefficients for smaller nonzero proportion, $r_\beta$, and also results in rarer variants having larger effects as commonly considered in polygenic models [46]. For the proportion of nonzero effect variants, three values are considered; $r_\beta = 0.2, 0.5$, and 0.8.
- Ordered: This scenario is considered for comparing power of the multivariate tests. Given $r_\beta \in (0,1)$, let $d_0 = \lfloor dr_\beta \rfloor$ variants have nonzero regression coefficients and remaining $d - d_0$ variables have zero regression coefficients, and the $d_0$ variants are randomly selected from the $d$ variants. Now, $100 \times r_\beta$% have nonzero regression coefficients among $d$ variants. Then, $d_0$ normal random variables are independently and identically drawn from $N(0, \sigma_\beta^2)$ with standard deviation $\sigma_\beta = 0.005/r_\beta$. Next, negative and positive values simulated above are placed on the lower and upper index sides, respectively. Zero regression coefficients are placed at the remaining $d - d_0$ indexes in the middle, which are between the indexes of negative and positive values. For example, if $d = 10$, $d_0 = 5$ and simulated $d_0 = 5$ nonzero values of regression coefficients are $(-0.5, 2.2, -3.4, 1.0, -0.7)$, then, the ordered regression coefficients result in $(-0.5, -3.4, -0.7, 0, 0, 0, 0, 2.2, 1.0)$. Similar to the above scenario, the simulated values from normal distribution are multiplied by $1/\{2\text{MAF}(1-\text{MAF})\}^{1/2}$ where MAF denotes the minor allele frequency of the corresponding variant. For the proportion of nonzero effect variants, three values are considered, $r_\beta = 0.2, 0.5$, and 0.8. This scenario considers a situation where there exist positive-effect, non-effect, and negative-effect blocks. The indexes separating each block are unknown. Unlike the above "Random" scenario, the "Ordered" scenario corresponds to the situation where the modeling by the fused lasso is suitable because the regression coefficients have a block structure. It is thus expected that the data-adaptive test using the fused lasso exhibits a higher power than the other methods because competing tests do not explicitly account for the ordering information of regression coefficients.
- Null: This scenario is considered for assessing type I error rates of the multivariate tests. All of the $d$ regression coefficients are set to zero.

The following multivariate tests are investigated: data-adaptive test for the elastic net, lasso, ridge regression, fused lasso, univariate regression test, saturated regression test, SKAT, burden test, and SKAT-O. For the scenarios "Random" and "Ordered", power of the nine competing multivariate tests is evaluated at three nominal levels, $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$ based on the $p$-value outputted from each test. The above three levels correspond to the situation where test is carried out using 50, 5000, and 500,000 groups of variants at the family-wide rate of 5% with the Bonferroni correction, respectively. 1000 replicates are used for each simulation scenario. For the scenario "Null", the type

I error rate of the nine competing multivariate tests is evaluated at four nominal levels, $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. In total, 50,000 replicates are used for simulations.

## 3. Results

Simulation results under scenarios "Random" and "Ordered" are given in Figures 1–4, which summarize the power of the nine multivariate tests, data-adaptive test for the elastic net, lasso, ridge regression, fused lasso, univariate regression test, saturated regression test, SKAT, burden test, and SKAT-O. Scenarios "Random" and "Ordered" allow comparison between the absence and presence of block structure on the regression coefficients in terms of the power of the compared tests. It is expected that the test based on the fused lasso performs better in the presence of the block structure.

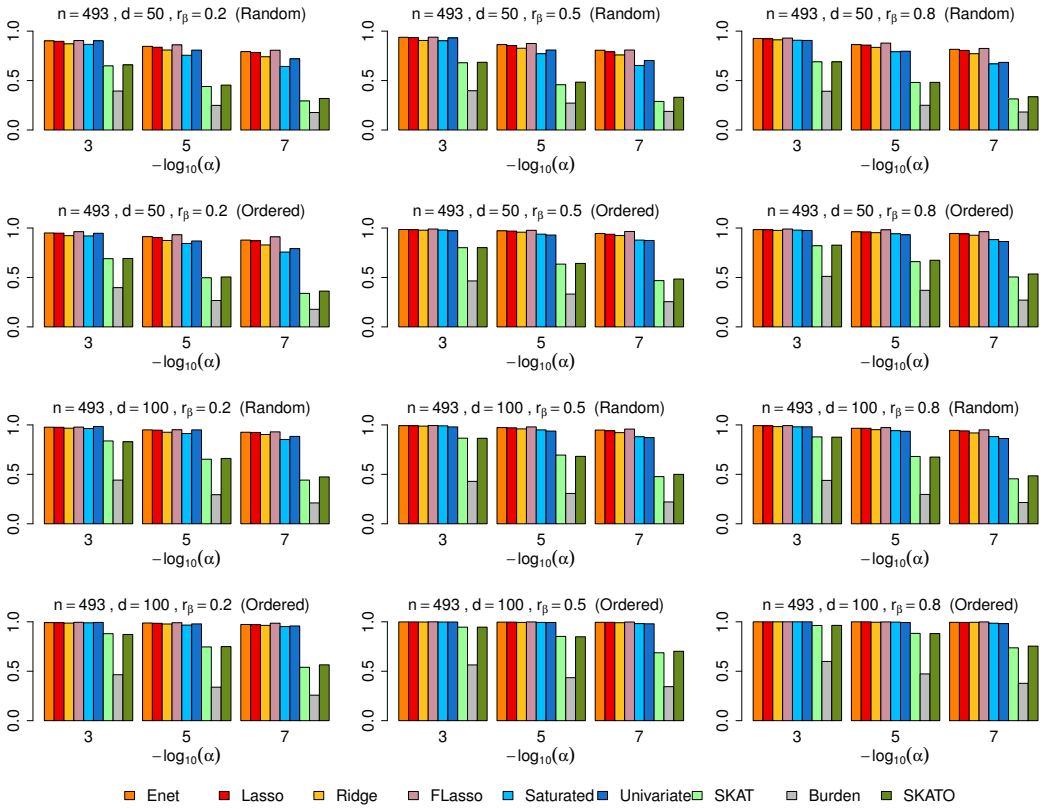


**Figure 1.** Power in simulation studies (1000 replicates) using 1000 Genomes Project data. Three scenarios for the proportion of nonzero regression coefficients, $r_\beta \in \{0.2, 0.5, 0.8\}$. Regression coefficients are randomly placed ("Random") or placed in order ("Ordered"); power is evaluated at three significance levels $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Number of variants to be tested is $d = 50$ or 100. Enet, data-adaptive test for elastic net; Lasso, data-adaptive test for lasso; Ridge, data-adaptive test for ridge regression; FLasso, data-adaptive test for fused lasso; Saturated, saturated regression test; Univariate, univariate regression test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

In Figures 1 and 2, the power of the nine tests from the simulation using 1000 Genomes Project data is given. Simulations in Figures 1 and 2 are the same except for that the variants are randomly shuffled in the latter. In Figure 1, the fused lasso test tends to give a higher or comparable power compared with the other eight competing tests throughout the scenarios. SKAT, burden test, and SKAT-O tend to give a lower power than the other tests. It can be seen from Figures 1 and 2 that SKAT, burden test, and SKAT-O have lower power than the other tests (i.e., the data-adaptive tests with the elastic net, lasso, ridge regression, and fused lasso, and univariate regression test, and saturated regression test). The observed differences in power between competing tests appear to be roughly

unchanged with varying the number of variants $d$ and the proportion of nonzero regression coefficients $r_\beta$.

In Figure 1, there is no clear difference on the power of the fused lasso test relative to other eight tests by comparison between "Random" and "Ordered" scenarios. In Figure 2, especially comparing the power of the fused lasso test in the top three panels with that in the three panels in the second row, it exhibits higher power in "Ordered" scenario than in "Random" scenario relative to other eight tests. For example, the power of the fused lasso test in $n = 493, d = 50, r_\beta = 0.2$ ("Random") is lower than saturated regression test at $-\log_{10}(\alpha) = 7$, but the power of the fused lasso test in $n = 493, d = 50, r_\beta = 0.2$ ("Ordered") is the best among the compared nine tests. This result implies that the fused lasso test performs better in the presence of block structure on the regression coefficients in terms of power as expected. Since the chief difference between Figures 1 and 2 is the magnitude of the correlation between variants, it could be considered that the correlation could influence on power of the fused lasso test, although the actual mechanism is unknown.

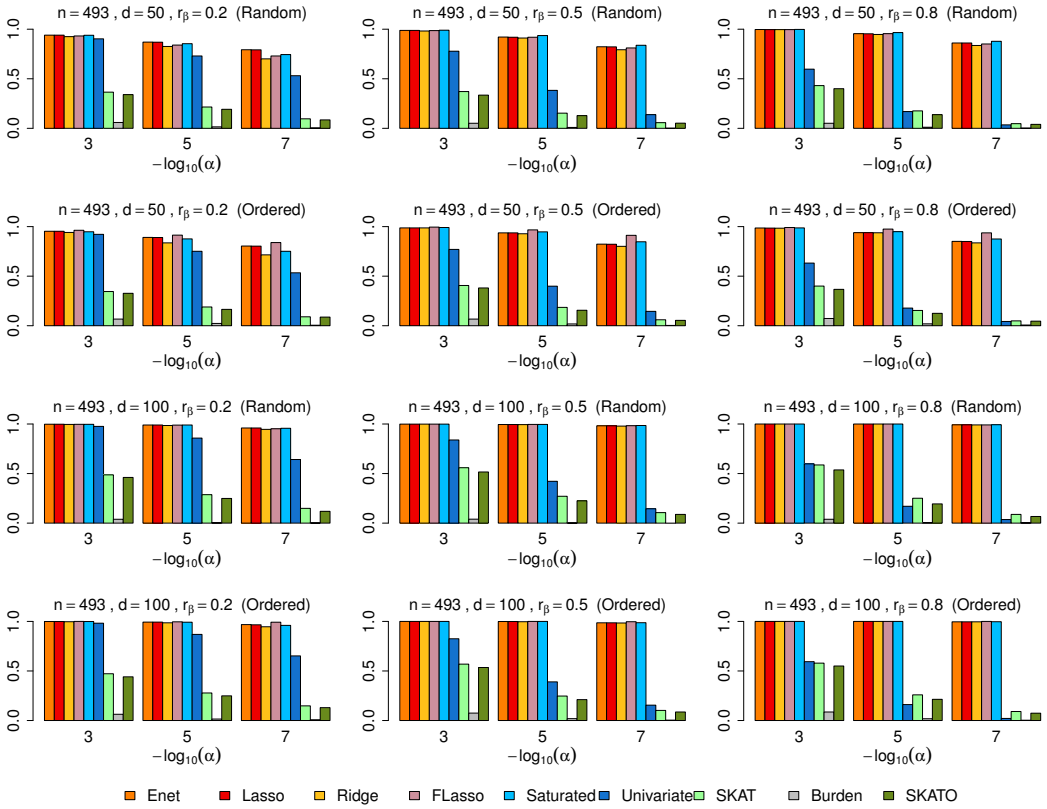


**Figure 2.** Power in simulation studies (1000 replicates) using 1000 Genomes Project data where variants are randomly shuffled to eliminate correlation structure between variants. Three scenarios for the proportion of nonzero regression coefficients, $r_\beta \in \{0.2, 0.5, 0.8\}$. Regression coefficients are randomly placed ("Random") or placed in order ("Ordered"); power is evaluated at three significance levels $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Number of variants to be tested is $d = 50$ or 100. Enet, data-adaptive test for elastic net; Lasso, data-adaptive test for lasso; Ridge, data-adaptive test for ridge regression; FLasso, data-adaptive test for fused lasso; Saturated, saturated regression test; Univariate, univariate regression test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

Figures 3 and 4 give the results of the simulation of genotypes under exchangeable correlation structure with $\rho = 0.3$ and 0.7, respectively. Overall, the power results in Figures 3 and 4 show a similar tendency to that observed in Figures 1 and 2, and the fused lasso test gives a higher or comparable power compared with the other eight tests. As in Figures 1 and 2, it can be seen from Figures 3 and 4 that SKAT, burden test, and

SKAT-O have lower power than the other tests. The observed differences in power between competing tests appear to be roughly unchanged with varying the number of variants $d$ and the proportion of nonzero regression coefficients $r_\beta$. Similar to Figure 2, comparing the power of the fused lasso test in the top three panels with the three panels in the second row of Figure 3 or 4, it exhibits higher power in the "Ordered" scenario than in the "Random" scenario relative to other eight tests, implying that the fused lasso test performs better in the presence of block structure on the regression coefficients in terms of power as expected.

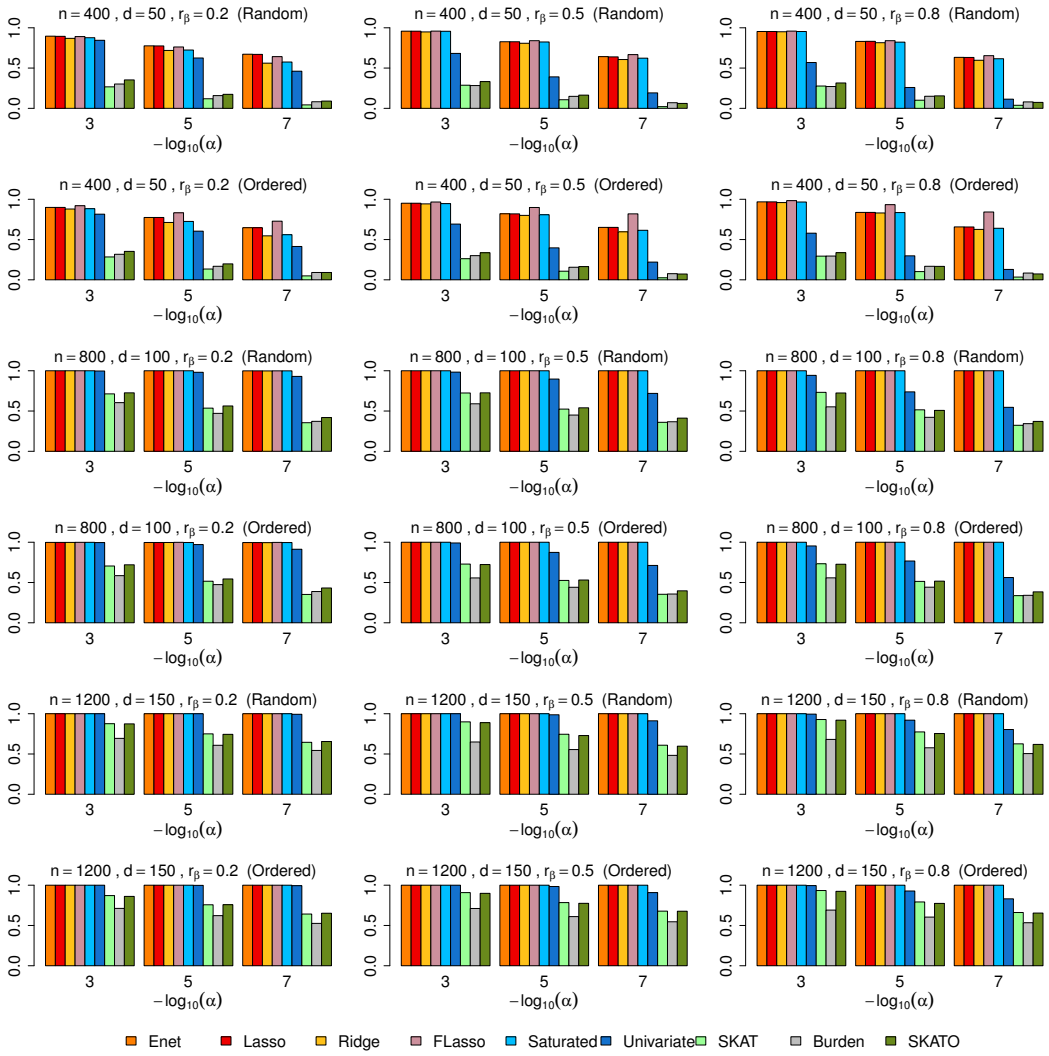


**Figure 3.** Power in simulation studies (1000 replicates) under exchangeable correlation structure with $\rho = 0.3$. Three scenarios for the proportion of nonzero regression coefficients, $r_\beta \in \{0.2, 0.5, 0.8\}$. Regression coefficients are randomly placed ("Random") or placed in order ("Ordered"); power is evaluated at three significance levels $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Sample size and number of variants to be tested are $(n, d) = (400, 50)$, $(800, 100)$, or $(1200, 150)$. Enet, data-adaptive test for elastic net; Lasso, data-adaptive test for lasso; Ridge, data-adaptive test for ridge regression; FLasso, data-adaptive test for fused lasso; Saturated, saturated regression test; Univariate, univariate regression test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

The results of type I error rate in scenario "Null" are given in Table 1. Compared with the nominal levels, type I error rates for the data-adaptive tests using elastic net, lasso, ridge regression, and fused lasso are controlled, especially when both $n$ and $d$ are large. In simulations using 1000 Genomes Project data, type I error rates for the data-adaptive tests using elastic net, lasso, ridge regression, and fused lasso are conservatively controlled as

expected from its theoretical result [20]; that is, the type I error rate of the data-adaptive test is asymptotically lower than the given nominal level as both $n$ and $d$ tends to $\infty$ while $n/d$ is kept constant. Type I error rates tend to be slightly larger for the shuffled genotypes than for the unshuffled genotypes for the tests with elastic net, lasso, and ridge regression except for the test with fused lasso. In simulations under an exchangeable correlation structure, type I error rates for the data-adaptive tests using elastic net, lasso, ridge regression, and fused lasso are slightly larger than the nominal level when $n = 400$, but turn out to be lower than the nominal level when $n = 800$. Type I error rates for the other tests, i.e., saturated regression test, univariate regression test, SKAT, burden test, and SKAT-O, are well controlled. Univariate regression test results in lower type I error rate than the nominal level due to the Bonferroni correction.

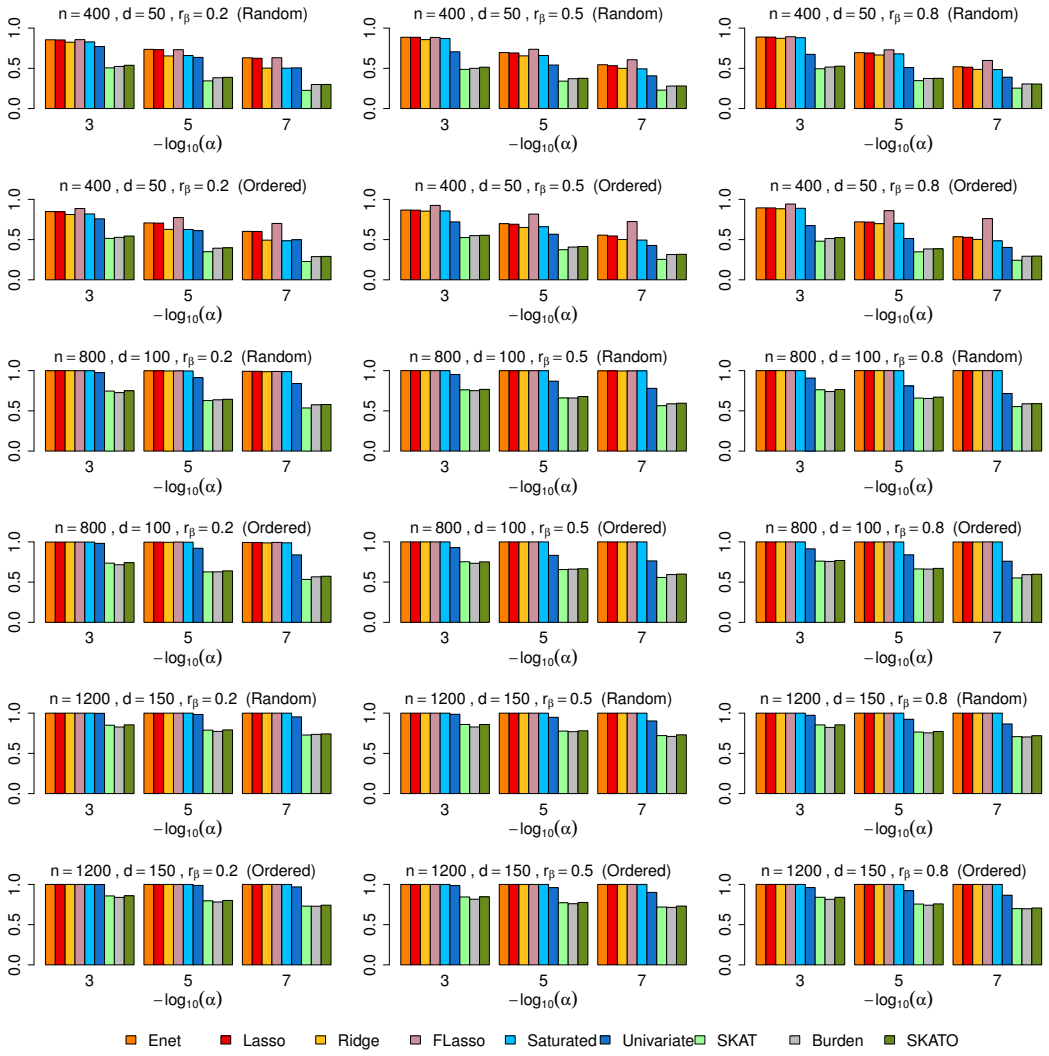


**Figure 4.** Power in simulation studies (1000 replicates) under exchangeable correlation structure with $\rho = 0.7$. Three scenarios for the proportion of nonzero regression coefficients, $r_\beta \in \{0.2, 0.5, 0.8\}$. Regression coefficients are randomly placed ("Random") or placed in order ("Ordered"); power is evaluated at three significance levels $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Sample size and number of variants to be tested are $(n, d) = (400, 50)$, $(800, 100)$, or $(1200, 150)$. Enet, data-adaptive test for elastic net; Lasso, data-adaptive test for lasso; Ridge, data-adaptive test for ridge regression; FLasso, data-adaptive test for fused lasso; Saturated, saturated regression test; Univariate, univariate regression test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

**Table 1.** Results of "Null" simulations (50,000 replicates). Type I error rate is evaluated at four nominal significance levels, $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. The first column gives sample size $n$, number of variants $d$, and correlation structure between the $d$ variants. Enet, test using elastic net; Lasso, test using lasso; Ridge, test using ridge regression; FLasso, test using fused lasso; Saturated, test under saturated regression test; Univariate, univariate regression test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

| $n$, $d$, Correlation | $\alpha$ (Nominal Level) | Enet | Lasso | Ridge | FLasso | Saturated | Univariate | SKAT | Burden | SKATO |
|---|---|---|---|---|---|---|---|---|---|---|
| Simulation using 1000 Genomes Project data | | | | | | | | | | |
| $n = 493$, $d = 50$, Shuffled | 0.1000 | 0.0728 | 0.0724 | 0.0654 | 0.0732 | 0.1019 | 0.0946 | 0.0960 | 0.0977 | 0.0975 |
| | 0.0100 | 0.0062 | 0.0061 | 0.0051 | 0.0069 | 0.0103 | 0.0104 | 0.0093 | 0.0094 | 0.0095 |
| | 0.0010 | 0.0005 | 0.0005 | 0.0004 | 0.0008 | 0.0009 | 0.0012 | 0.0010 | 0.0009 | 0.0010 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0003 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| $n = 493$, $d = 100$, Shuffled | 0.1000 | 0.0761 | 0.0756 | 0.0679 | 0.0748 | 0.1007 | 0.0967 | 0.0993 | 0.0984 | 0.1000 |
| | 0.0100 | 0.0068 | 0.0067 | 0.0057 | 0.0067 | 0.0100 | 0.0099 | 0.0089 | 0.0101 | 0.0098 |
| | 0.0010 | 0.0004 | 0.0004 | 0.0004 | 0.0005 | 0.0007 | 0.0010 | 0.0007 | 0.0012 | 0.0009 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 |
| $n = 493$, $d = 50$ | 0.1000 | 0.0671 | 0.0643 | 0.0131 | 0.0775 | 0.1001 | 0.0454 | 0.0963 | 0.0974 | 0.0978 |
| | 0.0100 | 0.0056 | 0.0052 | 0.0006 | 0.0072 | 0.0097 | 0.0051 | 0.0092 | 0.0092 | 0.0100 |
| | 0.0010 | 0.0004 | 0.0004 | 0.0000 | 0.0009 | 0.0010 | 0.0006 | 0.0008 | 0.0009 | 0.0009 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 |
| $n = 493$, $d = 100$ | 0.1000 | 0.0646 | 0.0611 | 0.0052 | 0.0771 | 0.0985 | 0.0434 | 0.0947 | 0.0993 | 0.0994 |
| | 0.0100 | 0.0053 | 0.0048 | 0.0002 | 0.0070 | 0.0102 | 0.0050 | 0.0089 | 0.0091 | 0.0104 |
| | 0.0010 | 0.0003 | 0.0003 | 0.0000 | 0.0006 | 0.0009 | 0.0005 | 0.0008 | 0.0007 | 0.0008 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0000 | 0.0001 |
| Simulation under exchangeable correlation structure | | | | | | | | | | |
| $n = 400$, $d = 50$, $\rho = 0.3$ | 0.1000 | 0.1020 | 0.1015 | 0.0907 | 0.1017 | 0.0988 | 0.0877 | 0.0995 | 0.1003 | 0.1019 |
| | 0.0100 | 0.0113 | 0.0112 | 0.0096 | 0.0119 | 0.0105 | 0.0102 | 0.0099 | 0.0097 | 0.0107 |
| | 0.0010 | 0.0012 | 0.0012 | 0.0008 | 0.0016 | 0.0011 | 0.0011 | 0.0012 | 0.0007 | 0.0011 |
| | 0.0001 | 0.0002 | 0.0002 | 0.0001 | 0.0004 | 0.0002 | 0.0001 | 0.0000 | 0.0002 | 0.0001 |
| $n = 400$, $d = 50$, $\rho = 0.7$ | 0.1000 | 0.1014 | 0.1006 | 0.0881 | 0.1018 | 0.0986 | 0.0645 | 0.1006 | 0.1014 | 0.1035 |
| | 0.0100 | 0.0103 | 0.0102 | 0.0083 | 0.0108 | 0.0097 | 0.0077 | 0.0095 | 0.0098 | 0.0107 |
| | 0.0010 | 0.0012 | 0.0012 | 0.0008 | 0.0015 | 0.0011 | 0.0008 | 0.0010 | 0.0009 | 0.0012 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0004 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 |
| $n = 800$, $d = 100$, $\rho = 0.3$ | 0.1000 | 0.0978 | 0.0974 | 0.0905 | 0.0968 | 0.0991 | 0.0860 | 0.0980 | 0.1025 | 0.1031 |
| | 0.0100 | 0.0095 | 0.0095 | 0.0086 | 0.0100 | 0.0098 | 0.0094 | 0.0091 | 0.0093 | 0.0101 |
| | 0.0010 | 0.0011 | 0.0011 | 0.0010 | 0.0012 | 0.0011 | 0.0010 | 0.0010 | 0.0008 | 0.0010 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 |
| $n = 800$, $d = 100$, $\rho = 0.7$ | 0.1000 | 0.1002 | 0.0996 | 0.0893 | 0.0999 | 0.1013 | 0.0594 | 0.0996 | 0.0997 | 0.1041 |
| | 0.0100 | 0.0098 | 0.0097 | 0.0081 | 0.0100 | 0.0100 | 0.0077 | 0.0095 | 0.0100 | 0.0107 |
| | 0.0010 | 0.0010 | 0.0009 | 0.0008 | 0.0011 | 0.0010 | 0.0010 | 0.0011 | 0.0010 | 0.0012 |
| | 0.0001 | 0.0001 | 0.0001 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| $n = 1200$, $d = 150$, $\rho = 0.3$ | 0.1000 | 0.0782 | 0.0777 | 0.0722 | 0.0769 | 0.1012 | 0.0884 | 0.1017 | 0.0987 | 0.1037 |
| | 0.0100 | 0.0066 | 0.0065 | 0.0057 | 0.0064 | 0.0098 | 0.0108 | 0.0095 | 0.0103 | 0.0105 |
| | 0.0010 | 0.0004 | 0.0004 | 0.0003 | 0.0004 | 0.0008 | 0.0012 | 0.0011 | 0.0010 | 0.0012 |
| | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0000 |
| $n = 1200$, $d = 150$, $\rho = 0.7$ | 0.1000 | 0.0773 | 0.0767 | 0.0684 | 0.0761 | 0.0997 | 0.0572 | 0.1005 | 0.1004 | 0.1050 |
| | 0.0100 | 0.0066 | 0.0066 | 0.0053 | 0.0065 | 0.0100 | 0.0069 | 0.0106 | 0.0101 | 0.0115 |
| | 0.0010 | 0.0006 | 0.0005 | 0.0004 | 0.0005 | 0.0010 | 0.0009 | 0.0010 | 0.0010 | 0.0011 |
| | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0002 | 0.0001 | 0.0001 |

## 4. Conclusions

This paper investigates the performances of the recently developed test by Ueki [20] applied to the fused lasso as a multivariate test for genomic studies through simulation studies. The existing multivariate tests compared in this paper are univariate regression tests, saturated regression tests, SKAT, and burden tests. Those tests ignore the ordering of variants in physical position. In some cases, researchers may exploit the location of the variants for effect size modeling [10,28]. In genome-wide scans, sets of grouped variants are often given based on genes, pathways, or physical position. The grouping strategy typically includes ambiguity and may differ depending on criteria such as the choice of database. It is unknown whether all grouped variants can be considered to have the same or similar effects.

Fused lasso modeling can be useful when the grouped effect is unknown, because it estimates the grouped effect from data. Indeed, in the simulation studies assuming block structure of regression coefficients, i.e., scenario "Ordered", the case where ordering

between adjacent variants is important, the test using fused lasso exhibits a high power compared to the existing tests including univariate regression test, saturated regression test, SKAT, SKAT-O, and the tests using lasso, ridge, and elastic net. Furthermore, the test using fused lasso still shows high or comparable power compared with other tests even without assuming the block structure (i.e., "Random" scenario). In most simulation scenarios, tests using the elastic net, lasso, and fused lasso perform well compared with other competing tests considered in this paper. In some scenarios, the saturated regression test shows a comparable power with the data-adaptive tests using the elastic net, lasso, and fused lasso, but results in lower performance than the data-adaptive tests, e.g., nine panels in the first to third rows of Figure 1 (simulations with unshuffled 1000 Genomes Project data). For the other tests, there are cases where the univariate regression test performs poorly as observed in Figures 2–4. SKAT, burden test, and SKAT-O show lower power than univariate regression tests. The possible reasons for this include that these tests are designed for the study of rare variant association and do not perform well for the common variants considered in this paper.

The test procedure of Ueki [20] is a general method applicable to various regression procedures. The model sequence contains low- to high-complexity models, where the high-complexity model is the saturated regression model with a large number of explanatory variables. A low-complexity model would be selected if this kind of model can adequately capture the data-generating process. Furthermore, the type I error is controlled without normality assumption on error distribution, if the number of explanatory variables is sufficiently large [20]. Although, in this paper, the testing framework is demonstrated for linear or additive model for variants, it is in principle applicable to nonlinear models such as the genetic models involving interaction terms, e.g., gene–gene interaction and gene–environment interaction as in [19,47]. Other potential applications include association studies with many phenotypes called PheWAS [48] and those with high-dimensional nuisance parameters [49].

Through simulation studies, this paper evaluates the performance of the data-adaptive test using fused lasso and shows its potential applicability to the test for grouped variants in genomic studies. It is interesting to apply it to real genomic data with actual phenotypes and to find a case where the fused lasso has a practical advantage. Furthermore, computational feasibility in large-scale data is desirable for current genomic studies as in the UK Biobank [50]. However, penalized regression models require higher computational cost than the standard univariate regression, which may lead to computational burden in the data-adaptive tests for biobank-scale genomic data. Also, while quantitative phenotypes are investigated, extension to generalized linear models to deal with binary phenotypes is another direction worthy of research. For applications to real genomic studies, the above topics need be addressed in future.

**Funding:** This research was funded by JSPS KAKENHI Grant Numbers 23K11009.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Publicly available datasets were analyzed in this study. These data can be found here: https://www.internationalgenome.org/(accessed on 28 August 2022).

**Acknowledgments:** The computation in this work was completed using the facilities of the Institute of Statistical Mathematics. The author would like to thank the four anonymous reviewers for their insightful comments and suggestions that led to the considerable improvement of the paper.

**Conflicts of Interest:** The author declares no conflicts of interest.

## Appendix A. Yanai's Generalized Coefficient of Determination and Its Generalization

The following is a brief description of Yanai's generalized coefficient of determination. Yanai's generalized coefficient of determination can measure a similarity between two linear

spaces, which has been used for variable selection in principal component analysis [31]. For two linear subspaces spanned by an $n \times c$ matrix $\boldsymbol{Y}$ and an $n \times d$ matrix $\boldsymbol{X}$, let the corresponding projection matrixes be $\boldsymbol{P}_{\boldsymbol{Y}}$ and $\boldsymbol{P}_{\boldsymbol{X}}$. Then, Yanai's generalized coefficient of determination $r(\boldsymbol{Y}, \boldsymbol{X})$ is given as follows:

$$r(\boldsymbol{Y}, \boldsymbol{X}) = \frac{\mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Y}} \boldsymbol{P}_{\boldsymbol{X}})}{c^{1/2} d^{1/2}}.$$

Since $c = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Y}}) = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{Y}}^2)$ and $d = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}}) = \mathrm{tr}(\boldsymbol{P}_{\boldsymbol{X}}^2)$, by the Cauchy–Schwarz inequality, $r(\boldsymbol{P}_{\boldsymbol{Y}}, \boldsymbol{P}_{\boldsymbol{X}}) \leq 1$ and the equality holds if and only if $\boldsymbol{P}_{\boldsymbol{Y}} = \boldsymbol{P}_{\boldsymbol{X}}$. Thus, its value being close to 1 indicates that the two linear spaces are similar. It is noteworthy that $r(\boldsymbol{Y}, \boldsymbol{X})$ can be used even when $c \neq d$, i.e., the number of dimensions differs.

Considering the special case with $c = 1$ for $\boldsymbol{Y}$, the Yanai's generalized coefficient of determination is also applicable to model selection in least-squares regression by expressing it through projection matrix. Consider a variable selection problem with response variable $\boldsymbol{y}$ and candidate $d$ explanatory variables $\boldsymbol{X}$. For a given subset of $d$ variables, $s \subset \{1, \ldots, d\}$, the Yanai's generalized coefficient of determination can be written as

$$r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s) = \frac{\mathrm{tr}(\boldsymbol{P}_{\widetilde{\boldsymbol{y}}} \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s})}{|s|^{1/2}} = \frac{||\widetilde{\boldsymbol{y}}||^{-2} \widetilde{\boldsymbol{y}}^T \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}}{|s|^{1/2}} = \frac{||\widetilde{\boldsymbol{y}}||^{-2} \widetilde{\boldsymbol{y}}^T \widetilde{\boldsymbol{X}}_s \widetilde{\boldsymbol{\beta}}_s}{|s|^{1/2}}, \tag{A1}$$

where $\widetilde{\boldsymbol{X}}_s$ denotes the sub-column matrix of $\widetilde{\boldsymbol{X}}$ corresponding to the index set $s$, $|s|$ denotes the cardinality of $s$, and $\widetilde{\boldsymbol{\beta}}_s$ is the least-squares estimate of regression of $\widetilde{\boldsymbol{y}}$ onto $\widetilde{\boldsymbol{X}}_s$. The value $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ being close to 1 means that $\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}$ is a good modeling procedure. The quantity $\widetilde{\boldsymbol{y}}^T \widetilde{\boldsymbol{X}}_s \widetilde{\boldsymbol{\beta}}_s$ in the numerator is proportional to the sample covariance between the observation $\widetilde{\boldsymbol{y}}$ and the fitted value $\widetilde{\boldsymbol{X}}_s \widetilde{\boldsymbol{\beta}}_s$, and is optimistic if it is used as a measure of model fit. The denominator, $|s|^{1/2}$, penalizes the apparent goodness of the model, allowing the model to be evaluated by adjusting for model complexity. The metric is invariant by replacing $\widetilde{\boldsymbol{X}}$ by $\widetilde{\boldsymbol{X}}\boldsymbol{B}$ with a $d \times d$ regular matrix $\boldsymbol{B}$, that is, $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}) = r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}\boldsymbol{B})$.

Ueki [20] generalized Yanai's generalized coefficient of determination for a modeling procedure $g_\lambda$ as given by (1), i.e.,

$$r(\widetilde{\boldsymbol{y}}, g_\lambda) = \frac{||\widetilde{\boldsymbol{y}}||^{-2} \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})}{\mathrm{gdf}_0(g_\lambda)^{1/2}},$$

with $\mathrm{gdf}_0(g_\lambda) = E_{\widetilde{\boldsymbol{\mu}}=\boldsymbol{0}}\{\widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})\}$, and $E_{\widetilde{\boldsymbol{\mu}}=\boldsymbol{0}}$ indicates the expectation under the assumption of $\widetilde{\boldsymbol{\mu}} = \boldsymbol{0}$. If $g_\lambda(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}$, the $r(\widetilde{\boldsymbol{y}}, g_\lambda)$ reduces to the original Yanai's generalized coefficient of determination in (A1). The denominator, $\mathrm{gdf}_0(g_\lambda)$, is a key quantity for the use in hypothesis testing.

Consider the null hypothesis $H_{0,n} : \boldsymbol{\mu} = \alpha_0 \boldsymbol{1}_n$, in the regression model, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, in which $\alpha_0$ is some constant and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I}_n)$, and $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\mu}$. Then, since $E(\widetilde{\boldsymbol{y}}^T \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}) = \sigma_0^2 |s|$, the expectation of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ is approximately proportional to $|s|^{1/2}$. It is monotonically increasing as the model dimensionality $|s|$ increases. Noting that $||\widetilde{\boldsymbol{y}}||^2$ does not depend on $s$, the Yanai's generalized coefficient of determination tends to select a model with large dimensionality under the null hypothesis of no effect $\boldsymbol{\mu} = \alpha_0 \boldsymbol{1}_n$. Specifically, for a given model sequence with large $d$, $\mathcal{M} = \{\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}, |s| = 1, \ldots, d\}$, the model that achieves the maximum of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ among the model sequence tends to be close to the saturated model. Equivalently, the dimensionality of the selected model is close to $d$ with high probability. On the other hand, under the alternative hypothesis of $\boldsymbol{\mu} \neq \alpha_0 \boldsymbol{1}_n$, the expectation of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ does not necessarily increase monotonically, in contrast to when the null hypothesis is true. For example, if $\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}}$, or the model completely recovers $\widetilde{\boldsymbol{\mu}}$, it holds that $E(\widetilde{\boldsymbol{y}}^T \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s} \widetilde{\boldsymbol{y}}) = \sigma_0^2 |s| + ||\widetilde{\boldsymbol{\mu}}||^2$. Then, the expectation of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ is approximately proportional to $\sigma_0^2 |s|^{1/2} + ||\widetilde{\boldsymbol{\mu}}||^2 / |s|^{1/2}$. If $||\widetilde{\boldsymbol{\mu}}||^2$ is sufficiently large, the second term dominates the first term, and the model with the smallest $|s|$ is chosen, which

differs from the case of null hypothesis where the saturated model tends to be chosen. The different behavior on dimensionality of the selected model by Yanai's generalized coefficient of determination under the alternative and under the null hypotheses suggests using it for hypothesis test. The saturated model is used to set the significance level because it tends to be selected by Yanai's generalized coefficient of determination under the null hypothesis. More rigorous arguments are found in Ueki [20].

## References


1. Risch, N.; Merikangas, K. The future of genetic studies of complex human diseases. *Science* **1996**, *273*, 1516–1517. [CrossRef]
2. Burton, P.R.; Clayton, D.G.; Cardon, L.R.; Craddock, N.; Deloukas, P.; Duncanson, A.; Kwiatkowski, D.P.; McCarthy, M.I.; Ouwehand, W.H.; Samani, N.J.; et al. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **2007**, *447*, 661–678.
3. Schaid, D.J.; Rowland, C.M.; Tines, D.E.; Jacobson, R.M.; Poland, G.A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **2002**, *70*, 425–434. [CrossRef]
4. Dudbridge, F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* **2008**, *66*, 87–98. [CrossRef]
5. Madsen, B.E.; Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **2009**, *5*, e1000384. [CrossRef]
6. Wu, M.C.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **2011**, *89*, 82–93. [CrossRef]
7. Ueki, M.; Kawasaki, Y.; Tamiya, G. Detecting genetic association through shortest paths in a bidirected graph. *Genet. Epidemiol.* **2017**, *41*, 481–497. [CrossRef]
8. Li, B.; Leal, S.M. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **2008**, *83*, 311–321. [CrossRef]
9. Lee, S.; Wu, M.C.; Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **2012**, *13*, 762–775. [CrossRef]
10. Lee, S.; Abecasis, G.R.; Boehnke, M.; Lin, X. Rare-variant association analysis: Study designs and statistical tests. *Am. J. Hum. Genet.* **2014**, *95*, 5–23. [CrossRef]
11. Sham, P.C.; Curtis, D. Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. *Ann. Hum. Genet.* **1995**, *59*, 97–105. [CrossRef]
12. Hirotsu, C.; Aoki, S.; Inada, T.; Kitao, Y. An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis. *Biometrics* **2001**, *57*, 769–778. [CrossRef]
13. Freidlin, B.; Zheng, G.; Li, Z.; Gastwirth, J.L. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Hum. Hered.* **2002**, *53*, 146–152. [CrossRef] [PubMed]
14. González, J.R.; Carrasco, J.L.; Dudbridge, F.; Armengol, L.; Estivill, X.; Moreno, V. Maximizing association statistics over genetic models. *Genet. Epidemiol.* **2008**, *32*, 246–254. [CrossRef]
15. Li, Q.; Zheng, G.; Li, Z.; Yu, K. Efficient approximation of *p*-value of the maximum of correlated tests, with applications to genome-wide association studies. *Ann. Hum. Genet.* **2008**, *72*, 397–406. [CrossRef]
16. Hothorn, L.A.; Hothorn, T. Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. *Biom. J.* **2009**, *51*, 659–669. [CrossRef]
17. Joo, J.; Kwak, M.; Chen, Z.; Zheng, G. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. *Stat. Med.* **2010**, *29*, 158–180. [CrossRef]
18. Zang, Y.; Fung, W.K. Robust Mantel–Haenszel test under genetic model uncertainty allowing for covariates in case-control association studies. *Genet. Epidemiol.* **2011**, *35*, 695–705. [CrossRef]
19. Ueki, M. On the choice of degrees of freedom for testing gene-gene interactions. *Stat. Med.* **2014**, *33*, 4934–4948. [CrossRef]
20. Ueki, M. Testing conditional mean through regression model sequence using Yanai's generalized coefficient of determination. *Comput. Stat. Data Anal.* **2021**, *158*, 107168. [CrossRef]
21. Yanai, H. A proposition of generalized method for forward selection of variables. *Behaviormetrika* **1980**, *7*, 95–107. [CrossRef]
22. Cadima, J.F.C.L.; Jolliffe, I.T. Variable selection and the interpretation of principal subspaces. *J. Agric. Biol. Environ. Stat.* **2001**, *6*, 62–79. [CrossRef]
23. Ye, J. On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* **1998**, *93*, 120–131. [CrossRef]
24. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
25. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
26. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
27. Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 91–108. [CrossRef]

28. Cheng, Y.; Dai, J.Y.; Kooperberg, C. Group association test using a hidden Markov model. *Biostatistics* **2016**, *17*, 221–234. [CrossRef] [PubMed]
29. Bao, M.; Wang, K. Genome-wide association studies using a penalized moving-window regression. *Bioinformatics* **2017**, *33*, 3887–3894. [CrossRef] [PubMed]
30. Huang, J.; Liu, J.; Ma, S.; Wang, K. Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. *Stat. Its Interface* **2013**, *6*, 99–115. [CrossRef]
31. Jolliffe, I. *Principal Component Analysis*; Springer: New York, NY, USA, 2002.
32. Efron, B. The estimation of prediction error. *J. Am. Stat. Assoc.* **2004**, *99*, 619–632. [CrossRef]
33. Friedman, J.; Hastie, T.; Hofling, H.; Tibshirani, R. Pathwise coordinate optimization. *Ann. Appl. Stat.* **2007**, *1*, 302–332. [CrossRef]
34. Tibshirani, R.; Wang, P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **2008**, *9*, 18–29. [CrossRef]
35. Shimamura, K.; Ueki, M.; Kawano, S.; Konishi, S. Bayesian generalized fused lasso modeling via NEG distribution. *Commun. Stat.–Theory Methods* **2018**, *48*, 4132–4153. [CrossRef]
36. Zou, H.; Hastie, T.; Tibshirani, R. On the "degrees of freedom" of the lasso. *Ann. Stat.* **2007**, *35*, 2173–2192. [CrossRef]
37. Tibshirani, R.J.; Taylor, J. Degrees of freedom in lasso problems. *Ann. Stat.* **2012**, *40*, 1198–1232. [CrossRef]
38. Dossal, C.; Kachour, M.; Fadili, M.; Peyré, G.; Chesneau, C. The degrees of freedom of the Lasso for general design matrix. *Stat. Sin.* **2013**, *23*, 809–828. [CrossRef]
39. Chen, X.; Lin, Q.; Sen, B. On degrees of freedom of projection estimators with applications to multivariate nonparametric regression. *J. Am. Stat. Assoc.* **2019**, *115*, 173–186. [CrossRef]
40. Bland, J.M.; Altman, D.G. Statistics notes: Multiple significance tests: The Bonferroni method. *BMJ* **1995**, *310*, 170. [CrossRef]
41. Wright, S.P. Adjusted *p*-values for simultaneous inference. *Biometrics* **1992**, *48*, 1005–1013. [CrossRef]
42. Lin, D.Y.; Tang, Z.Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **2011**, *89*, 354–367. [CrossRef]
43. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef]
44. Howie, B.; Fuchsberger, C.; Stephens, M.; Marchini, J.; Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **2012**, *44*, 955–959. [CrossRef]
45. O'Connell, J.; Gurdasani, D.; Delaneau, O.; Pirastu, N.; Ulivi, S.; Cocca, M.; Traglia, M.; Huang, J.; Huffman, J.E.; Rudan, I.; et al. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **2014**, *10*, e1004234. [CrossRef]
46. Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D.R.; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **2010**, *42*, 565–569. [CrossRef]
47. Ueki, M.; Tamiya, G.; for Alzheimer's Disease Neuroimaging Initiative. Smooth-threshold multivariate genetic prediction incorporating gene–environment interactions. *G3* **2021**, *11*, jkab278. [CrossRef]
48. Bush, W.S.; Oetjens, M.T.; Crawford, D.C. Unravelling the human genome–phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **2016**, *17*, 129–145. [CrossRef]
49. Chen, J.; Li, Q.; Chen, H.Y. Testing generalized linear models with high-dimensional nuisance parameters. *Biometrika* **2023**, *110*, 83–99. [CrossRef]
50. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **2015**, *12*, e1001779. [CrossRef]