Computational Statistics and Data Analysis 158 (2021) 107168

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Testing conditional mean through regression model sequence using Yanai's generalized coefficient of determination

Masao Ueki, and for the Alzheimer's Disease Neuroimaging Initiative[1]

School of Information and Data Sciences, Nagasaki University, 1-14 Bunkyo, Nagasaki 852-8521, Japan[2]

ARTICLE INFO

Article history:
Received 3 February 2020
Received in revised form 28 December 2020
Accepted 28 December 2020
Available online 19 January 2021

Keywords:
Generalized degrees of freedom
Test for conditional mean
Model selection
Yanai's generalized coefficient of determination

ABSTRACT

In high-dimensional data analysis such as in genomics, repeated univariate regression for each variable is utilized to screen useful variables. However, signals jointly detectable with other variables may be overlooked. While the saturated model using all variables may not work in high-dimensional data, based on prior knowledge, group-wise analysis for a pre-defined group is often developed, but the power will be limited if the knowledge is insufficient. A flexible test procedure is thus proposed for conditional mean applicable to a variety of model sequences that bridge between low and high complexity models as in penalized regression. The test is based on the model that maximizes a generalization of the Yanai's generalized coefficient of determination by exploiting the tendency for the dimensionality to be large under the null hypothesis. The test does not require complicated null distribution computation, thereby enabling large-scale testing application. Numerical studies demonstrated that the proposed test applied to the lasso and elastic net had a high power regardless of the simulation scenarios. Applied to a group-wise analysis in real genome-wide association study data from Alzheimer's Disease Neuroimaging Initiative, the proposal gave a higher association signal than the existing methods.

1. Introduction

Currently, large-scale high-dimensional data are being collected and analyzed in almost all fields including genomics, medicine, biology, agriculture, ecology, neuroscience, marketing, social science, and economics. It is necessary to extract only useful information for hypothesis generation. However, ubiquitous statistical tool applicable to such huge data is limitedly available. Univariate analysis for each variable with a target response variable is frequently used for screening purposes, e.g. in genome-wide, epigenome-wide and phenome-wide association studies (Risch and Merikangas, 1996; Rakyan et al., 2011; Bush et al., 2016) and in a voxel-wise test for functional magnetic resonance imaging data analysis (Friston et al., 1994). Similarly, a genome-wide environment interaction study explores the interaction effect of each genetic variant and environment factor pair (Kraft et al., 2007). Those tests examine the effects under a given alternative model represented by a few parameters (e.g. a single genetic variant is independently associated with disease),

E-mail address: uekimrsd@nifty.com.

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[2] Supplementary material including Supplementary Tables and Figures is attached.

https://doi.org/10.1016/j.csda.2021.107168
0167-9473/

but they might be insufficient to describe complex phenomenon. To extract new findings with minimal prior knowledge, more complex models would be needed.

Alternative approaches include group-wise analysis for pre-defined groups of variables. For example, in genome-wide association studies, nearby variants are grouped and tested separately for each group (Schaid et al., 2002; Dudbridge, 2008; Madsen and Browning, 2009; Wu et al., 2011; Ueki et al., 2017). Similarly, a multi-voxel test called a searchlight mapping is proposed in functional magnetic resonance imaging data analysis (Kriegeskorte et al., 2006). However, test is underpowered if the saturated model is excessively redundant compared to the genuine structure. The power can be gained by custom-made tests using prior knowledge but is limited if the knowledge is incomplete. The incompleteness is often unavoidable when exploring various candidate variables. Data-adaptive approaches have been proposed in genomic studies (Sham and Curtis, 1995; Hirotsu et al., 2001; Freidlin et al., 2002; González et al., 2008; Li et al., 2008; Hothorn and Hothorn, 2009; Joo et al., 2010; Zang and Fung, 2011; Lee et al., 2012; Ueki, 2014). However, these approaches often require complicated null distribution calculation either analytically or computationally, or otherwise are only applicable to low-complexity models. It is helpful if there is a framework that fits existing highly data-adaptive procedures to a hypothesis test without both custom-made modification and complicated null distribution computation.

This paper develops a flexible data-driven test procedure for conditional mean, directly applicable to various existing statistical models that bridge between low and high complexity models via a tuning parameter as in penalized regression. The test is based on the model that maximizes the Yanai's generalized coefficient of determination (Yanai, 1980; Cadima and Jolliffe, 2001) generalized to any modeling procedure. It is proportional to the covariance between a response variable and its predicted value divided by the square root of the generalized degrees of freedom (Ye, 1998). Under the null hypothesis of no effect, the selected model tends to have a large dimensionality unlike the familiar model selection criteria (Akaike, 1974; Schwarz, 1978; Craven and Wahba, 1978; Nishii, 1984; Foster and George, 1994; Shao, 1997; Chen and Chen, 2008). Exploiting the behavior under the null hypothesis, type I error is approximately controlled based on an asymptotic result of Wang and Cui (2013) without complicated null distribution computation using a significance threshold for the saturated model (or the largest model in the sequence). Since it is simple and simulation-free in computing p-value, the proposed method is suitable for effect discovery in high-dimensional data which requires a large number of tests.

Through simulation studies for group-wise test problems, the proposed test adapted to the lasso (Tibshirani, 1996), ridge (Hoerl and Kennard, 1970) and elastic net (Zou and Hastie, 2005) showed higher power across a variety of scenarios in comparison with the existing methods including univariate regression test, saturated model test and tests assuming random effects (Wu et al., 2011; Lee et al., 2012). Applied to a group-wise analysis for a real genome-wide association study data from Alzheimer's Disease Neuroimaging Initiative (ADNI), the proposed test showed a higher association signal at the known risk variant than the existing methods.

2. Methods

2.1. Preliminary

Consider a situation where n observations are obtained together with a response variable $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and d explanatory variables $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)$ where $\boldsymbol{X}_j = (x_{1j}, \ldots, x_{nj})^T$. Then, consider a set of regression models indexed by a tuning parameter λ, $g_\lambda(\boldsymbol{y})$ which 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 λ, 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 regression, elastic net, generalized lasso, and many other regression models in statistics or machine learning.

2.2. Yanai's generalized coefficient of determination and its potential use for hypothesis testing

The 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 (Jolliffe, 2002). For two linear subspaces spanned by $\boldsymbol{Y}$ (a matrix with size $n \times c$) and $\boldsymbol{X}$ (a matrix with size $n \times d$), let the corresponding projection matrixes be $\boldsymbol{P}_{\boldsymbol{Y}}$ and $\boldsymbol{P}_{\boldsymbol{X}}$. Then, the Yanai's generalized coefficient of determination, $r(\boldsymbol{Y}, \boldsymbol{X})$ say, is given by

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

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 $\boldsymbol{P}_{\boldsymbol{Y}} = \boldsymbol{P}_{\boldsymbol{X}}$. Hence, the value close to 1 indicates a similarity between the two linear spaces. Notably, $r(\boldsymbol{Y}, \boldsymbol{X})$ can be used even if the number of dimensions differs, i.e. $c \neq d$.

Next, consider the special case with $c = 1$ for $\boldsymbol{Y}$. The idea is in principle applicable to model selection in least-squares regression through the projection matrix representation. To this end, consider a variable selection problem with response variable $\boldsymbol{y}$ and candidate d explanatory variables $\boldsymbol{X}$. 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

$\widetilde{\boldsymbol{X}} = \boldsymbol{Q}_{\mathbf{1}_n}\boldsymbol{X}$ are considered. Here, $\boldsymbol{Q}_{\mathbf{1}_n} = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$, $\boldsymbol{I}_n$ is the nth identity matrix, and $\mathbf{1}_n$ is the n-vector of ones. For a given subset of d variables, $s \subset \{1, \ldots, d\}$, the Yanai's generalized coefficient of determination can be written as

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

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 $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ 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, allowing to evaluate the model by accounting for model complexity. The metric is a geometric quantity in the sense that it is invariant by replacing $\widetilde{\boldsymbol{X}}$ by $\widetilde{\boldsymbol{X}}\boldsymbol{B}$ with a $d \times d$ regular matrix $\boldsymbol{B}$, i.e. $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}) = r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}\boldsymbol{B})$.

From now on, the Yanai's generalized coefficient of determination is explored from a different perspective, i.e. application in hypothesis testing. Consider the null hypothesis $H_{0,n} : \boldsymbol{\mu} = \alpha\mathbf{1}_n$, in the regression model, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, in which α is some constant and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_0^2\boldsymbol{I}_n)$, 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. Thus, noting that $\|\widetilde{\boldsymbol{y}}\|^2$ does not depend on s, it is expected that the Yanai's generalized coefficient of determination tends to select a much larger model under the null hypothesis of no effect $\boldsymbol{\mu} = \alpha\mathbf{1}_n$. 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, i.e. its dimensionality is close to d with high probability. On the other hand, under the alternative hypothesis of $\boldsymbol{\mu} \neq \alpha\mathbf{1}_n$, the expectation of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ does not necessarily increase monotonically, unlike the case where 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 is contrasted from the case of null hypothesis where the saturated model tends to be chosen. Hence, the behavior under the alternative differs from that under the null hypothesis. The above perspective suggests the potential use for hypothesis testing of the null hypothesis $\boldsymbol{\mu} = \alpha\mathbf{1}_n$. That is, the model that gives the maximum of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ is tested by referring to the saturated model, which is the model selected with probability tending to 1 under the null hypothesis. The following section details the testing procedure for more general modeling procedures.

2.3. Testing procedure

To apply the above arguments for least-squares regression model sequence to more general modeling procedures, a new testing procedure is proposed by generalizing the Yanai's generalized coefficient of determination (1) while inheriting the property described in the last paragraph of the previous section. The generalization is simply replacing the degrees of freedom in the denominator by the generalized degrees of freedom. 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, the generalized version for a modeling procedure g_λ is given by

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

where $\mathrm{gdf}_0(g_\lambda) = E_{\widetilde{\boldsymbol{\mu}}=\mathbf{0}}\{\widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})\}$, and $E_{\widetilde{\boldsymbol{\mu}}=\mathbf{0}}$ indicates the expectation under the assumption of $\widetilde{\boldsymbol{\mu}} = \mathbf{0}$. The quantity $\mathrm{gdf}_0(g_\lambda)$ coincides with the generalized degrees of freedom of g_λ defined by $\mathrm{cov}\{\widetilde{\boldsymbol{y}}, g_\lambda(\widetilde{\boldsymbol{y}})\} = E\{(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^T g_\lambda(\widetilde{\boldsymbol{y}})\}$ (Ye, 1998; Efron, 2004) under the null hypothesis $\boldsymbol{\mu} = \alpha\mathbf{1}_n$ because of $\widetilde{\boldsymbol{\mu}} = \boldsymbol{Q}_{\mathbf{1}_n}\boldsymbol{\mu} = \mathbf{0}$. For least-squares regression with explanatory variables $\widetilde{\boldsymbol{X}}_s$, the generalized degrees of freedom is given by $\mathrm{tr}(\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s}) = |s|$ (Ye, 1998), and hence, (2) reduces to the original Yanai's generalized coefficient of determination (1). Although (2) is no longer interpreted as a model-fit measure, it possesses a property that the expectation of $r(\widetilde{\boldsymbol{y}}, g_\lambda)$ under the null hypothesis $\boldsymbol{\mu} = \alpha\mathbf{1}_n$ is approximately proportional to $\mathrm{gdf}_0(g_\lambda)^{1/2}$. 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 high probability. (The assumption that $\mathrm{gdf}_0(g_\lambda) \leq d$ may hold for constraint linear regressions including ridge regression and lasso (Kaufman and Rosset, 2014, Theorem 2).) Similarly, for the alternative hypothesis of $\boldsymbol{\mu} \neq \alpha\mathbf{1}_n$, assuming that $g_\lambda(\widetilde{\boldsymbol{y}}) \approx \widetilde{\boldsymbol{\mu}}$, the expectation of $r(\widetilde{\boldsymbol{y}}, \widetilde{\boldsymbol{X}}_s)$ 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. The above heuristic arguments are made rigorous in the next subsections.

The proposed testing procedure is described in detail. First note that, at $\lambda = 0$,

$$\frac{\|\widetilde{\boldsymbol{y}}\|^2 d^{1/2} r(\widetilde{\boldsymbol{y}}, g_\lambda)/d}{\hat{\sigma}_y^2} = \frac{d^{1/2}\widetilde{\boldsymbol{y}}^T g_0(\widetilde{\boldsymbol{y}})/d}{\mathrm{gdf}_0(g_0)^{1/2}\hat{\sigma}_y^2} = \frac{\|\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}\|^2/d}{\hat{\sigma}_y^2}, \tag{3}$$

where $\hat{\sigma}_y^2 = \|\boldsymbol{Q}_{(\mathbf{1}_n, \boldsymbol{X})}\boldsymbol{y}\|^2/(n - d - 1)$. This is the usual F-statistic. If $\boldsymbol{y} \sim N(\alpha\mathbf{1}_n, \sigma_0^2\boldsymbol{I}_n)$, (3) follows an F-distribution with $(d, n - d - 1)$ degrees of freedoms. It is also expected that the generalized degrees of freedom that achieves its maximum,

$\max_\lambda r(\widetilde{\boldsymbol{y}}, g_\lambda)$, is close to d with high probability under the null hypothesis if d is large. Then, the proposed test procedure is to reject the null hypothesis $\boldsymbol{\mu} = \boldsymbol{0}$ if

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

where $\bar{F}_\alpha^{-1}(d)$ is the $(1-\alpha)$th quantile of the F-distribution with $(d, n-d-1)$ degrees of freedoms at a given significance threshold α, and $\widehat{\lambda}^* = \mathrm{argmax}_\lambda r(\widetilde{\boldsymbol{y}}, g_\lambda)$.

Here, γ is a given fixed constant in $(0, 1)$, which plays a role to judge whether the selected model is close to the saturated model. If the selected model is regarded as the saturated model, the test statistic switches to that under the saturated model. Type I error control and power are considered in the following subsections.

In practice, generalized degrees of freedom are not always available in an exact form, but instead an estimate is available. For the ridge regression, $g_\lambda(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_\lambda \widetilde{\boldsymbol{y}}$ where $\boldsymbol{P}_\lambda = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + n\lambda \boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{X}}^T$, $\mathrm{gdf}_0(g_\lambda) = \mathrm{tr}(\boldsymbol{P}_\lambda)$, which can be used explicitly. For the lasso, $\mathrm{gdf}_0(g_\lambda) = E(|A_\lambda|)$ holds (Zou et al., 2007; Tibshirani and Taylor, 2012; Dossal et al., 2013), where A_λ is the active set at a given tuning parameter λ, 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_λ is the active set at a given tuning parameter $\boldsymbol{\lambda}$ as before. The generalized degrees of freedom for other models are given in Chen et al. (2019). If no closed-form estimate is available, simulation-based method is a possible approach (Ye, 1998).

2.4. Type I error control

Here, type I error control for the proposed procedure is given under a high-dimensional regression model with the set up being a special case of Wang and Cui (2013) page 136 in which regression coefficients are zero. Specifically, they consider a linear regression model, $y_i = \alpha + \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ for $i = 1, \ldots, n$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent and identically distributed with $\boldsymbol{\Sigma}_x = \mathrm{var}(\boldsymbol{x}_i)$ assumed to be positive definite, and $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed error with $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma_0^2$, $\boldsymbol{\beta}$ is the d-dimensional vector of regression coefficients, and α is a nuisance intercept parameter. Note that the error distribution is not necessarily normal. Without normality assumption of ϵ_is, Theorem of Wang and Cui (2013) implies that the statistic (3) with the F distribution as the null distribution for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ still gives a valid type I error control asymptotically as $n \to \infty$ with d/n tending to a constant in $(0, 1)$. The test statistic (3) is termed as the generalized F-statistic, cf. Eq. (2.3) or Eq. (3.1) of Wang and Cui (2013). (Note that the name "generalized F statistic" comes from the non-normal error distribution and penalized regression is not considered in Wang and Cui (2013).) To be specific, the following conditions taken from Wang and Cui (2013) are assumed.

(C1) $\boldsymbol{x}_i$ is linearly generated by a m-variate random vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})^T$ so that $\boldsymbol{x}_i = \boldsymbol{\Gamma}\boldsymbol{z}_i + \boldsymbol{\mu}_x$, where $\boldsymbol{\Gamma}$ is a $d \times m$ matrix for some $m \geq d$ such that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}_x$, each z_{il} has finite 8-th moment, $E(\boldsymbol{z}_i) = \boldsymbol{0}$, $\mathrm{var}(\boldsymbol{z}_i) = \boldsymbol{I}_m$, $E(z_{ik}^4) = 3 + \Delta$ and for any $\sum_{v=1}^k l_v \leq 8$, $E(z_{1i_1}^{l_1} z_{1i_2}^{l_2} \cdots z_{1i_k}^{l_k}) = E(z_{1i_1}^{l_1})E(z_{1i_2}^{l_2}) \cdots E(z_{1i_k}^{l_k})$, where Δ is some finite constant.

(C2) $\mu_4 = E(\epsilon_i^4) < \infty$;

(C3) $\rho_n = d/n \to \rho \in (0, 1)$ as $n \to \infty$.

Under the null hypothesis $\boldsymbol{\beta} = \boldsymbol{0}$ and the assumptions (C1)–(C3), the following statement holds, which is a special case of Theorem of Wang and Cui (2013) (i.e. $\delta_{\beta_2} = 0$ and $p_1 = 0$ in their notation):

$$P\left\{\frac{\|\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}\|^2/d}{\hat{\sigma}_y^2} > \bar{F}_\alpha^{-1}(d)\right\} = \alpha + o(1). \tag{5}$$

In the case that the above type I error control of the generalized F-test holds, type I error control of the proposed test is investigated. Consider a model sequence indexed by a tuning parameter $\lambda \geq 0$, $g_\lambda(\widetilde{\boldsymbol{y}})$, which gives the saturated model by $g_0(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}$ at $\lambda = 0$. To investigate the type I error, the null hypothesis $\boldsymbol{\mu} = \alpha\boldsymbol{1}_n$ is assumed. Candidate models are considered at J fixed discrete points $0 = \lambda_1 < \cdots < \lambda_J$. Let $g_{(j)} = g_{\lambda_j}$ for $j = 1, \ldots, J$, and the corresponding generalized degrees of freedom are denoted by $\mathrm{gdf}_{(1)}, \ldots, \mathrm{gdf}_{(J)}$ with $\mathrm{gdf}_{(1)} = \mathrm{gdf}_0(g_0) = \mathrm{tr}(\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}) = d$, the degrees of freedom for the saturated model. The following theorem describes the asymptotic control of type I error rate.

Theorem 1. *For the model sequence $\{g_{(j)} : j = 1, \ldots, J\}$ such that $g_{(1)}(\widetilde{\boldsymbol{y}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}$, and, for each j, $\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}}) \geq 0$ almost surely. Let γ be a given constant in $(0, 1)$ and assume that $q_\alpha(d) \geq d$. Then, under conditions (C1)–(C3) with $n \to \infty$, the type I error of the test procedure (4) is asymptotically no greater than α.*

Note that the assumption that $q_\alpha(d) \geq d$ is true if α is sufficiently small.

2.5. Power consideration

In this section, power analysis of the proposed test is given under which the effect size increases as $n \to \infty$, where d also increases at the same rate of n as in the above type I error analysis. Consider a model sequence, $g_\lambda : \widetilde{\boldsymbol{y}} \mapsto g_\lambda(\widetilde{\boldsymbol{y}})$, indexed by a tuning parameter $\lambda \geq 0$, in which g_0 gives the fit under the saturated model. Here, it is assumed that $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed error with $E(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma_0^2$. In the following argument, as in the previous sections, the notation $\widetilde{\cdot}$ denotes the centered quantity. In this case, for instance, $\widetilde{\boldsymbol{\mu}}^T \mathbf{1}_n = 0$ holds. Let $\lambda^* = \mathrm{argmin}_\lambda \|\widetilde{\boldsymbol{\mu}} - g_\lambda(\widetilde{\boldsymbol{\mu}})\|^2$. The following technical conditions are imposed throughout:

(D1) for any $\lambda \geq 0$, g_λ is Lipschitz, i.e. there exists a constant $K_\lambda > 0$ such that $\|g_\lambda(\widetilde{\boldsymbol{y}}) - g_\lambda(\widetilde{\boldsymbol{y}} + \boldsymbol{\Delta})\| \leq K_\lambda \|\boldsymbol{\Delta}\|$ for any $\widetilde{\boldsymbol{y}}$ and $\boldsymbol{\Delta}$.
(D2) for any $\lambda \leq \lambda^*$, there exist a constant $C_\lambda > 0$ such that $\widetilde{\boldsymbol{\mu}}^T g_\lambda(\widetilde{\boldsymbol{\mu}}) = C_\lambda \|\widetilde{\boldsymbol{\mu}}\|^2$, and a constant $D_\lambda > 0$ such that $\|g_\lambda(\widetilde{\boldsymbol{\mu}})\|^2 = D_\lambda \|\widetilde{\boldsymbol{\mu}}\|^2$.
(D3) $\mathrm{gdf}_0(g_{\lambda^*}) = O(1)$, and $\mathrm{gdf}_0(g_\lambda) \geq \mathrm{gdf}_0(g_{\lambda^*})$ for any $\lambda \leq \lambda^*$.
(D4) $\|\widetilde{\boldsymbol{\mu}}\|^2 = n\nu_n$ where $\nu_n \to \infty$ as $n \to \infty$.

(D1) assumes that the modeling procedure is a Lipschitz function. (D2) implies that covariance between $\widetilde{\boldsymbol{\mu}}$ and the fitted result from the modeling procedure applied to $\widetilde{\boldsymbol{\mu}}$ can grow at the same rate of $\|\widetilde{\boldsymbol{\mu}}\|^2$ for $\lambda \leq \lambda^*$ (i.e. the optimal model or the models larger than the optimum), which is the rate for g_λ such that $g_\lambda(\widetilde{\boldsymbol{\mu}}) = \widetilde{\boldsymbol{\mu}}$, and similarly for $\|g_\lambda(\widetilde{\boldsymbol{\mu}})\|^2$. (D3) assumes that the generalized degrees of freedom of the modeling procedure does not depend on d at the optimum (i.e. the model sequence can capture the underlying data structure in a parsimonious way), and also it is no greater than the generalized degrees of freedom at $\lambda \leq \lambda^*$. (D4) assumes that the average effect size, $\|\widetilde{\boldsymbol{\mu}}\|^2/n$, increases as a function of n. Then, the following statement holds.

Theorem 2. *Under conditions (D1)–(D4), for a given constant γ in $(0, 1)$, the proposed test is asymptotically more powerful than the generalized F-test under the saturated model.*

Remark. Condition (D2) is illustrated with two examples. First example is the least-squares regression on the sub-column matrix $\widetilde{\boldsymbol{X}}_s$ of $\widetilde{\boldsymbol{X}}$, say $g_s(\widetilde{\boldsymbol{\mu}}) = \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s}\widetilde{\boldsymbol{\mu}}$. If $\widetilde{\boldsymbol{\mu}}$ is represented by a linear combination of $\widetilde{\boldsymbol{X}}_s$, then $\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_s}\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\mu}}$. In this case, $\widetilde{\boldsymbol{\mu}}^T g_s(\widetilde{\boldsymbol{\mu}}) = \|\widetilde{\boldsymbol{\mu}}\|^2$, and $\|g_s(\widetilde{\boldsymbol{\mu}})\|^2 = \|\widetilde{\boldsymbol{\mu}}\|^2$, which implies that condition (D2) is fulfilled. Second example is the ridge regression, $g_\lambda(\widetilde{\boldsymbol{\mu}}) = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \lambda \boldsymbol{I}_d)^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{\mu}}$. Consider the singular value decomposition $\widetilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrixes of sizes $n \times n$ and $d \times d$, and $\boldsymbol{\Xi}$ is the $n \times d$ rectangular diagonal matrix with diagonal entries being singular values of $\widetilde{\boldsymbol{X}}$. Let $\widetilde{\boldsymbol{u}} = \boldsymbol{U}^T\widetilde{\boldsymbol{\mu}}$. Assume that $\widetilde{\boldsymbol{X}}$ is of full rank and that $d < n$. Let the minimum and maximum singular values of $\widetilde{\boldsymbol{X}}$ be $\xi_{\min}$ and $\xi_{\max}$, respectively. Then, $g_\lambda(\widetilde{\boldsymbol{\mu}}) = \boldsymbol{U}\boldsymbol{\Xi}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Xi}^T\boldsymbol{\Xi}\boldsymbol{V}^T + \lambda \boldsymbol{V}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Xi}^T\widetilde{\boldsymbol{u}}$, and $\widetilde{\boldsymbol{\mu}}^T g_\lambda(\widetilde{\boldsymbol{\mu}}) = \sum_{j=1}^d \frac{\xi_j^2}{\xi_j^2+\lambda}(\widetilde{\boldsymbol{u}})_j^2 \in [\frac{\xi_{\min}^2}{\xi_{\min}^2+\lambda}\|\widetilde{\boldsymbol{\mu}}\|^2, \frac{\xi_{\max}^2}{\xi_{\max}^2+\lambda}\|\widetilde{\boldsymbol{\mu}}\|^2]$. Similarly, $\|g_\lambda(\widetilde{\boldsymbol{\mu}})\|^2 = \sum_{j=1}^d \left(\frac{\xi_j^2}{\xi_j^2+\lambda}\right)^2 (\widetilde{\boldsymbol{u}})_j^2 \in [\left(\frac{\xi_{\min}^2}{\xi_{\min}^2+\lambda}\right)^2 \|\widetilde{\boldsymbol{\mu}}\|^2, \left(\frac{\xi_{\max}^2}{\xi_{\max}^2+\lambda}\right)^2 \|\widetilde{\boldsymbol{\mu}}\|^2]$. Consequently, if $\xi_{\min} > 0$, condition (D2) is fulfilled.

3. Simulation studies

Simulation studies for testing association between a set of d variables $\boldsymbol{X}$ and a response variable $\boldsymbol{y}$ were conducted. Sample size n and number of variables d were set as $n = 400\kappa$ and $d = 50\kappa$ for three cases of $\kappa = 1, 2$, and 3. The proposed testing procedure was adapted to the lasso, ridge, and elastic net by `glmnet` package for R, where generalized degrees of freedoms were calculated as described earlier. For γ in (4), which judges the closeness between the selected model and the saturated model, $\gamma = 0.01$ was considered throughout. (The results with $\gamma = 0.1$ were provided in Supplementary material.) Competing methods were as follows.

Univariate regression test minimum of d p-values from univariate (generalized) F-test for each variable adjusted by the Bonferroni correction (i.e. raw p-value multiplied by d) were used as the representative p-value.

Saturated regression test (generalized) F-test for the analysis of variance under the normal model using all d variables simultaneously.

Sequence kernel association test a score test under a random effect model for d regression coefficients developed for group-wise test of genotype data implemented in `SKAT` package for R.

Burden test a score test for an aggregated effect of d variables implemented in `SKAT` package for R.

Optimized sequence kernel association test an optimal combination of the sequence kernel association test and burden test implemented in `SKAT` package for R.

Table 1
Type I error rates from simulation (10 000 replicates) for scenario 1 with normal error at each nominal significance level $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. Column with (κ, ρ) describes the scenarios, where κ gives $n = 400\kappa$ and $d = 50\kappa$, and ρ specifies the correlation structure of $\boldsymbol{X}$. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, test under saturated model; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

(κ, ρ)	α	Enet	Lasso	Ridge	Saturated	Univariate	SKAT	Burden	SKATO
(1,0.3)	0.1	0.1068	0.1063	0.0945	0.1045	0.0797	0.1007	0.1019	0.1034
	0.01	0.0124	0.0123	0.0101	0.0117	0.0092	0.0099	0.01	0.0102
	0.001	0.0020	0.0019	0.0013	0.0018	0.0005	0.0007	0.0008	0.0007
	0.0001	0.0004	0.0003	0.0001	0.0004	0.0001	0.0001	0.0001	0.0001
(1,0.7)	0.1	0.0966	0.0961	0.0790	0.0947	0.0338	0.1016	0.1018	0.1019
	0.01	0.0112	0.0113	0.0081	0.0105	0.0044	0.0095	0.0097	0.0097
	0.001	0.0016	0.0016	0.0011	0.0014	0.0005	0.001	0.001	0.001
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001
(2,0.3)	0.1	0.0972	0.0964	0.0890	0.0982	0.0733	0.1012	0.1060	0.1040
	0.01	0.01	0.0098	0.0083	0.0101	0.0076	0.0106	0.0103	0.01
	0.001	0.0003	0.0003	0.0002	0.0003	0.0005	0.001	0.0011	0.0011
	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
(2,0.7)	0.1	0.0942	0.0939	0.0794	0.0956	0.0277	0.0972	0.0976	0.0971
	0.01	0.0097	0.0097	0.0071	0.0103	0.0043	0.0099	0.0101	0.01
	0.001	0.0013	0.0013	0.001	0.0013	0.0003	0.0009	0.0009	0.0009
	0.0001	0.0002	0.0002	0.0002	0.0002	0.0000	0.0000	0.0000	0.0000
(3,0.3)	0.1	0.0990	0.0980	0.0921	0.1007	0.0728	0.1048	0.1029	0.1043
	0.01	0.0091	0.0091	0.0081	0.0096	0.0099	0.0111	0.0114	0.0113
	0.001	0.0011	0.0011	0.0009	0.0011	0.0015	0.0009	0.0008	0.0008
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.0001
(3,0.7)	0.1	0.0957	0.0948	0.0828	0.0992	0.0247	0.0979	0.0997	0.0991
	0.01	0.0089	0.0087	0.0068	0.0094	0.0031	0.0101	0.0101	0.01
	0.001	0.0009	0.0009	0.0009	0.0009	0.0006	0.001	0.001	0.001
	0.0001	0.0004	0.0004	0.0001	0.0004	0.0002	0.0002	0.0002	0.0002

Two scenarios for correlation structures of $\boldsymbol{X}$ were considered.

Scenario 1 d explanatory variables $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ were generated as $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{S})$ with zero mean vector and variance–covariance matrix $\boldsymbol{S}$ independently for $i = 1, \ldots, n$, in which $\boldsymbol{S}$ is the $d \times d$ matrix with off-diagonal and diagonal elements are ρ and 1, respectively. Then, $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the error, and $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ denotes the true regression coefficients. Three kinds of distributions for $\boldsymbol{\epsilon}$ were considered: Each element of $\boldsymbol{\epsilon}$ was independent and identically generated from (i) standard normal distribution, (ii) normal distribution with mean zero and standard deviation independent and identically generated from Exp(1)/1.5, and (iii) log-normal distribution, $(e^Z - e^{1/2})/2$ where Z is a standard normal random variate generated independently and identically.

Scenario 2 The data-generation model is the same as that in scenario 1 except that autocorrelation structure for the variance–covariance matrix of $\boldsymbol{X}$ was used instead. Specifically, (j, k)-entry of the $d \times d$ matrix $\boldsymbol{S}$ is given by $\rho^{|j-k|}$.

3.1. Type I error rate

For type I error simulation under scenarios 1 and 2, two correlation structures for $\boldsymbol{S}$ were considered, namely, $\rho \in \{0.3, 0.7\}$. 10000 replicates were used for simulations and type I error rates were evaluated at each of four nominal levels, $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. The results are given in Tables 1–2 for normal error with $\gamma = 0.01$, Supplementary Tables S1–S2 for normal–exponential standard deviation error with $\gamma = 0.01$, Supplementary Tables S3–S4 for exponential error with $\gamma = 0.01$, Supplementary Tables S5–S6 for normal error with $\gamma = 0.1$, Supplementary Tables S7–S8 for normal–exponential standard deviation error with $\gamma = 0.1$, and Supplementary Tables S9–S10 for exponential error with $\gamma = 0.1$. It can be seen that type I error rates were well controlled for all tests, including the proposed test applied to elastic net, lasso, and ridge regression, regardless of error distribution as stated in Theorem 1. $\gamma = 0.1$ gave slightly lower type I error rate than when $\gamma = 0.01$, but the difference was small in particular at lower nominal levels. As κ increases, type I error rate of the proposed test tended to that of the saturated model as expected.

3.2. Power

For power simulation under scenarios 1 and 2, among d, only $d_0 = \lfloor dr_b \rfloor$ variables had nonzero regression coefficients and remaining $d - d_0$ variables had zero regression coefficients, where the d_0 variables were randomly selected in each simulation replicate. Two effect size distributions for nonzero coefficients were considered: (a) d_0 nonzero regression coefficients were independently generated from $N(0, \sigma_b^2)$, with standard deviation $\sigma_b = 0.005/r_b$; (b) d_0 nonzero

Table 2
Type I error rates from simulation (10 000 replicates) for scenario 2 with normal error at each nominal significance level $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$. Column with (κ, ρ) describes the scenarios, where κ gives $n = 400\kappa$ and $d = 50\kappa$, and ρ specifies the correlation structure of $\boldsymbol{X}$. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, test under saturated model; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

(κ, ρ)	α	Enet	Lasso	Ridge	Saturated	Univariate	SKAT	Burden	SKATO
(1,0.3)	0.1	0.1028	0.1025	0.0914	0.0998	0.0994	0.0884	0.0913	0.0888
	0.01	0.0089	0.0088	0.0075	0.0085	0.0105	0.0068	0.0089	0.0065
	0.001	0.001	0.001	0.0009	0.001	0.0011	0.0002	0.001	0.0002
	0.0001	0.0001	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
(1,0.7)	0.1	0.1024	0.1019	0.0812	0.0980	0.0769	0.0983	0.1103	0.1037
	0.01	0.0101	0.0101	0.0065	0.0091	0.0084	0.0074	0.0112	0.0099
	0.001	0.0009	0.0009	0.0006	0.0008	0.001	0.0007	0.0013	0.0007
	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
(2,0.3)	0.1	0.0945	0.0944	0.0874	0.0955	0.0944	0.0879	0.1	0.0957
	0.01	0.0092	0.0092	0.0080	0.0096	0.0109	0.0070	0.01	0.0098
	0.001	0.001	0.001	0.0009	0.0009	0.0006	0.0007	0.0013	0.0011
	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0001
(2,0.7)	0.1	0.0970	0.0968	0.0782	0.0987	0.0802	0.0929	0.1031	0.0968
	0.01	0.0094	0.0093	0.0075	0.0096	0.0107	0.0095	0.0115	0.0094
	0.001	0.0006	0.0006	0.0004	0.0006	0.0014	0.0003	0.0014	0.0011
	0.0001	0.0002	0.0002	0.0001	0.0002	0.0001	0.0001	0.0002	0.0002
(3,0.3)	0.1	0.0980	0.0972	0.0920	0.1008	0.0965	0.0878	0.1012	0.0968
	0.01	0.0103	0.0101	0.0091	0.0106	0.0094	0.0071	0.0095	0.0086
	0.001	0.0009	0.0009	0.0008	0.0009	0.0009	0.0005	0.0009	0.0007
	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000
(3,0.7)	0.1	0.1005	0.0999	0.0816	0.1029	0.0842	0.0971	0.0998	0.0980
	0.01	0.0105	0.0104	0.0068	0.0112	0.0097	0.0080	0.0113	0.0102
	0.001	0.0007	0.0007	0.0001	0.0007	0.0012	0.001	0.0006	0.0012
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000	0.0003	0.0000	0.0001

regression coefficients were randomly chosen from $\{-0.4/d_0^{1/2}, 0.4/d_0^{1/2}\}$. In addition, the following scenarios were considered: four sparseness proportions, $r_b \in \{0.05, 0.2, 0.5, 1\}$; two correlation structures for $\boldsymbol{S}$, $\rho \in \{0.3, 0.7\}$. Power was evaluated at three nominal levels, $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. 500 replicates were used for all simulation runs.

The resulting power is given in Figs. 1–4 for normal error with $\gamma = 0.01$, Supplementary Figures S1–S4 for normal–exponential standard deviation error with $\gamma = 0.01$, Supplementary Figures S5–S8 for exponential error with $\gamma = 0.01$, Supplementary Figures S9–S12 for normal error with $\gamma = 0.1$, Supplementary Figures S13–S16 for normal–exponential standard deviation error with $\gamma = 0.1$, and Supplementary Figures S17–S20 for exponential error with $\gamma = 0.1$. Overall, the proposed test applied to the elastic net or lasso gave a high power uniformly for all simulation scenarios. Test applied to the ridge regression showed a high power in some scenarios but resulted in low power in other scenarios. Univariate and saturated tests gave high power in some scenarios but resulted in low power in other scenarios, showing both strength and weakness depending on alternative hypotheses and hence lack of flexibility. Sequence kernel association, burden, and optimized sequence kernel association tests gave lower power than the elastic net and lasso tests for many scenarios except for some cases. In some scenarios (e.g. panel (2, 0.05, 0.7) in Fig. 2), the elastic net and lasso tests did not show the best performance but the power was comparable to the test under the saturated model. Influence of γ and error distribution was negligible. To summarize, the proposed test applied to the elastic net and lasso showed a high power in a variety of alternative hypotheses, while the power of other tests highly depended on simulation scenarios.

4. Real data application

Performance of the proposed test was examined through a real genome-wide association study data publicly available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org. ADNI is an ongoing, longitudinal study with primary purpose being to explore the genetic and neuroimaging information associated with late-onset Alzheimer's disease (LOAD). The study investigators recruited elderly subjects older than 65 years of age comprising about 400 subjects with mild cognitive impairment (MCI), about 200 subjects with Alzheimer's disease (AD), and about 200 healthy controls. Each subject was followed for at least 3 years. During the study period, the subjects were assessed with magnetic resonance imaging (MRI) measures and psychiatric evaluation to determine the diagnosis status at each time point. After applying a standard quality-control, the dataset included 528984 SNPs in total. There were 166

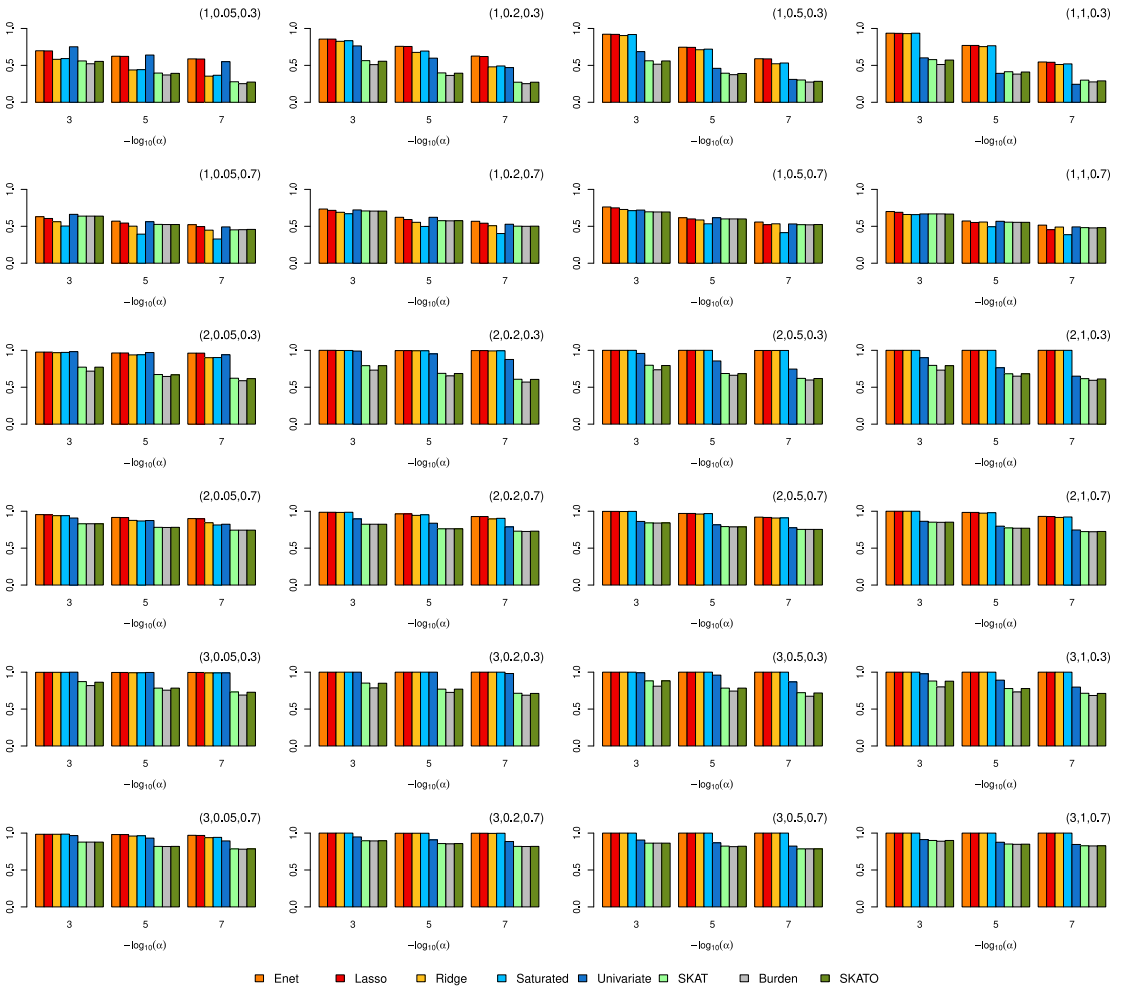

Fig. 1. Power in simulation studies (500 replicates) for scenario 1 with normal error under effect size scenario (a), evaluated at significance level $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Triplet (κ, r_b, ρ) on the top right in each panel denotes different setups: κ gives $n = 400\kappa$ and $d = 50\kappa$, ρ specifies the correlation structure of $\boldsymbol{X}$ and r_b denotes the proportion of nonzero variables. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

controls and numbers of cases with early MCI (EMCI), late MCI (LMCI), and AD were 67, 110, and 341, respectively, and were scored as 0, 2, 3, and 4. The score was considered as a continuous response variable.

Correlation is often present in genotype data due to linkage disequilibrium. To account for the correlation structure, group-wise analysis was considered in a sliding-window approach as frequently used in genetic studies (e.g. The UK10K Consortium, 2015). Each of 22 autosomes was divided in small regions of 1×10^6 base-pair interval with 2×10^5 base-pair overlap, which resulted in 2331 regions to be tested for association. For each region, eight tests considered in the simulation studies were compared. Figs. 5 and 6 give the manhattan and quantile–quantile plots where $\gamma = 0.01$ was used for the proposed test. Result under $\gamma = 0.1$ is given in Supplementary Figures S21 and S22.

The APOE4 gene located on chromosome 19 is one of known risk factors for Alzheimer's disease. The corresponding region was between 49056021 and 50456021 in base-pair position and included 205 variants. At the nominal family-wise error rate of 5% using Bonferroni correction (i.e. raw p-value threshold before the correction being $0.05/2331 \approx 2.1 \times 10^{-5}$), the proposed test applied to elastic net and lasso detected this gene region, in which both tests gave an identical p-value of 3.4×10^{-49}. (The p-value was identical both for $\gamma = 0.01$ and 0.1.) The univariate test also detected this region with the minimum p-value 1.9×10^{-20} (after Bonferroni correction by the number of variants, 205, in the region) and the other tests failed to detect at the nominal level. The p-values were 0.0003, 8.9×10^{-5}, 0.00097, 0.0076, and 0.001 for the ridge ($\gamma = 0.01$ and 0.1), saturated model test, sequence kernel association test, burden test, and optimized sequence kernel association test, respectively. The elastic net and lasso tests produced much small p-value than the univariate test. The tendency of giving the small p-value was also observed in the simulation studies.

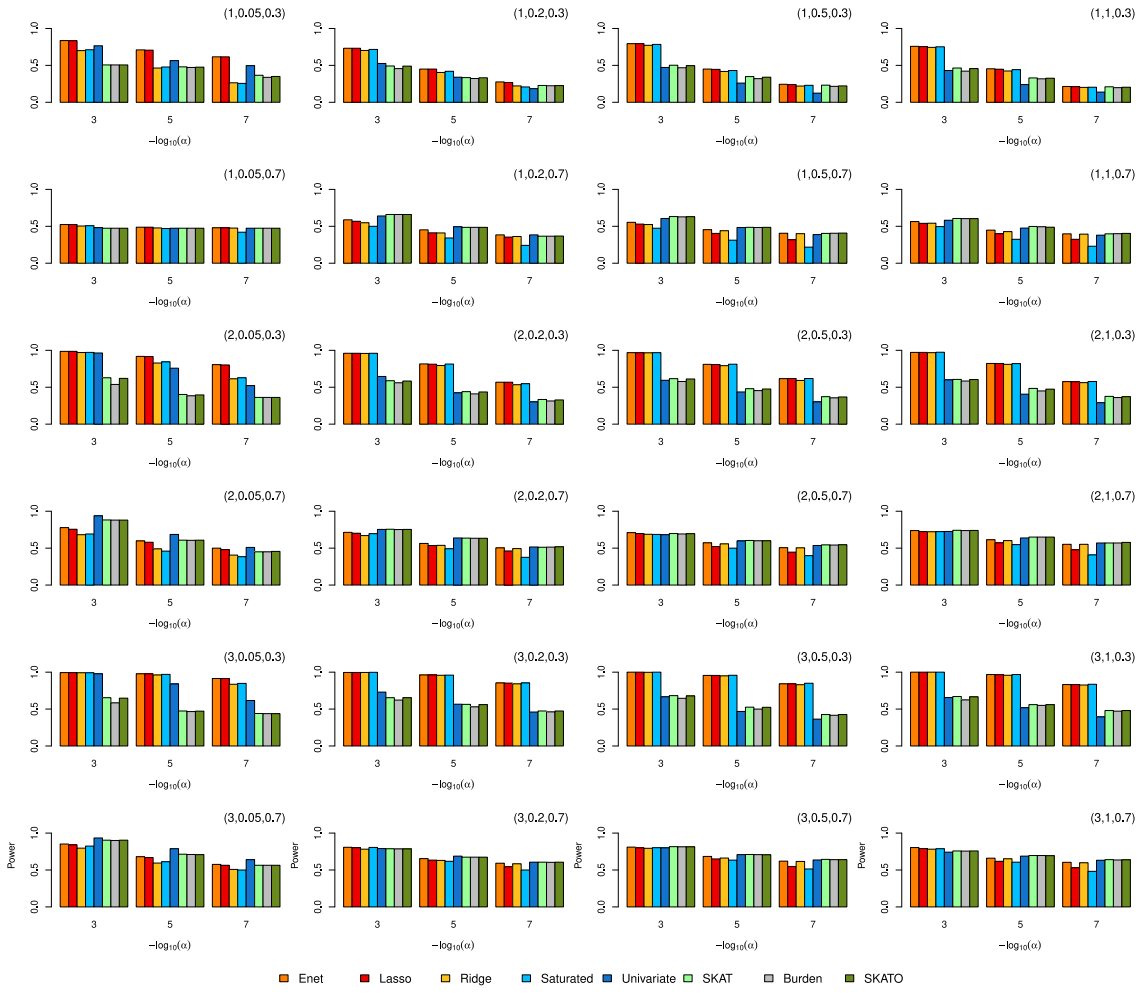

Fig. 2. Power in simulation studies (500 replicates) for scenario 1 with normal error under effect size scenario (b), evaluated at significance level $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Triplet (κ, r_b, ρ) on the top right in each panel denotes different setups: κ gives $n = 400\kappa$ and $d = 50\kappa$, ρ specifies the correlation structure of $\boldsymbol{X}$ and r_b denotes the proportion of nonzero variables. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

For further insight, the variant rs429358 located at the base-pair position 50103781 in the gene region was the only variant that gave nonzero regression coefficient from the variable selection by the elastic net and lasso. On the other hand, the univariate test identified two variants which passed the genome-wide significance level: rs2075650 at base-pair position 50087459 (p-value was 1.9×10^{-11}) and rs429358 (p-value was 9.5×10^{-23}). Pearson's correlation coefficient between two variants was 0.74. The high correlation suggested that one of variants was redundant and unnecessary. For further investigation, by applying multiple regression simultaneously using both rs2075650 and rs429358, the p-value for the former regression coefficient was 0.53 while that for the latter was 6.6×10^{-13}, implying that the former was less useful and the latter was the main contributing factor. The above result coincides with the result from variable selection by the elastic net or the lasso. Schaid et al. (2018) argue that univariate regression test cannot locate causal variants and penalized regression may be one of possible approaches for fine mapping, which in turn suggests that the proposed framework is also useful in this purpose.

In many typical genome-wide association studies, most variants are considered to follow the null hypothesis of no effect. In this real data application, the type I error rates of all tests were controlled as expected under the null hypothesis, as shown in Fig. 6 and Supplementary Figure S22. The proposed test gave slightly deflated p-values and the proposed type I error control worked well in this real data application.

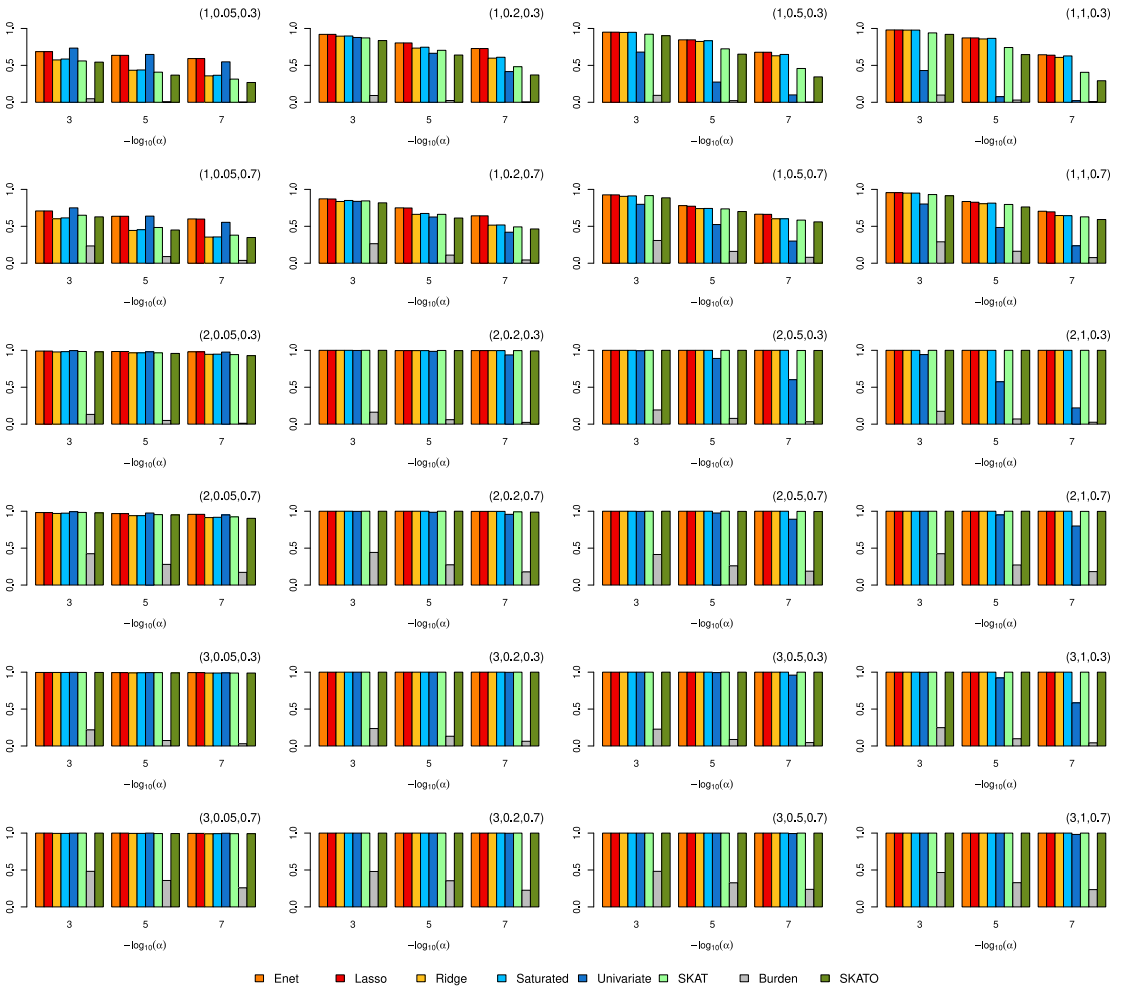

Fig. 3. Power in simulation studies (500 replicates) for scenario 2 with normal error under effect size scenario (a), evaluated at significance level $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Triplet (κ, r_b, ρ) on the top right in each panel denotes different setups: κ gives $n = 400\kappa$ and $d = 50\kappa$, ρ specifies the correlation structure of $\boldsymbol{X}$ and r_b denotes the proportion of nonzero variables. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

5. Concluding remarks

This paper presented a test procedure of conditional mean of zero through model sequence that connects models with different dimensionality from low to high complexity. Such model sequences commonly appear in practice, including penalized regression (e.g. L_1-, L_2-, and nonconvex penalties) or even more complicated machine learning models. The proposed procedure directly fits existing procedures to the proposed test without custom-made modification, and is expected to work well if the model sequence given by users contains models that can adequately capture the underlying data structure such as sparse regressions. Test combined with data-adaptive model search is attractive, particularly when there is great uncertainty in the alternative hypothesis. Hypothesis test with data-adaptive modeling usually requires complicated null distribution, which sometimes turns out to be analytically intractable, and then computer-intensive method, such as resampling, is required for computation. It is advantageous that the proposed procedure does not require computationally intensive method, where the generalized degrees of freedom are the only ingredient. The computational efficiency enables high-dimensional data application as exemplified in a group-wise test problem for genome-wide association studies.

While the proposed test was demonstrated for linear regression models in this paper, the framework is in principle applicable to nonlinear models such as splines. Nonlinear models are worthy to apply to data which encounter low effect size and lack of replicability issues. In such cases, it is possible that typically used models are too simple to capture the data generating process, trying more sophisticated models rather than linear models may facilitate effect discovery.

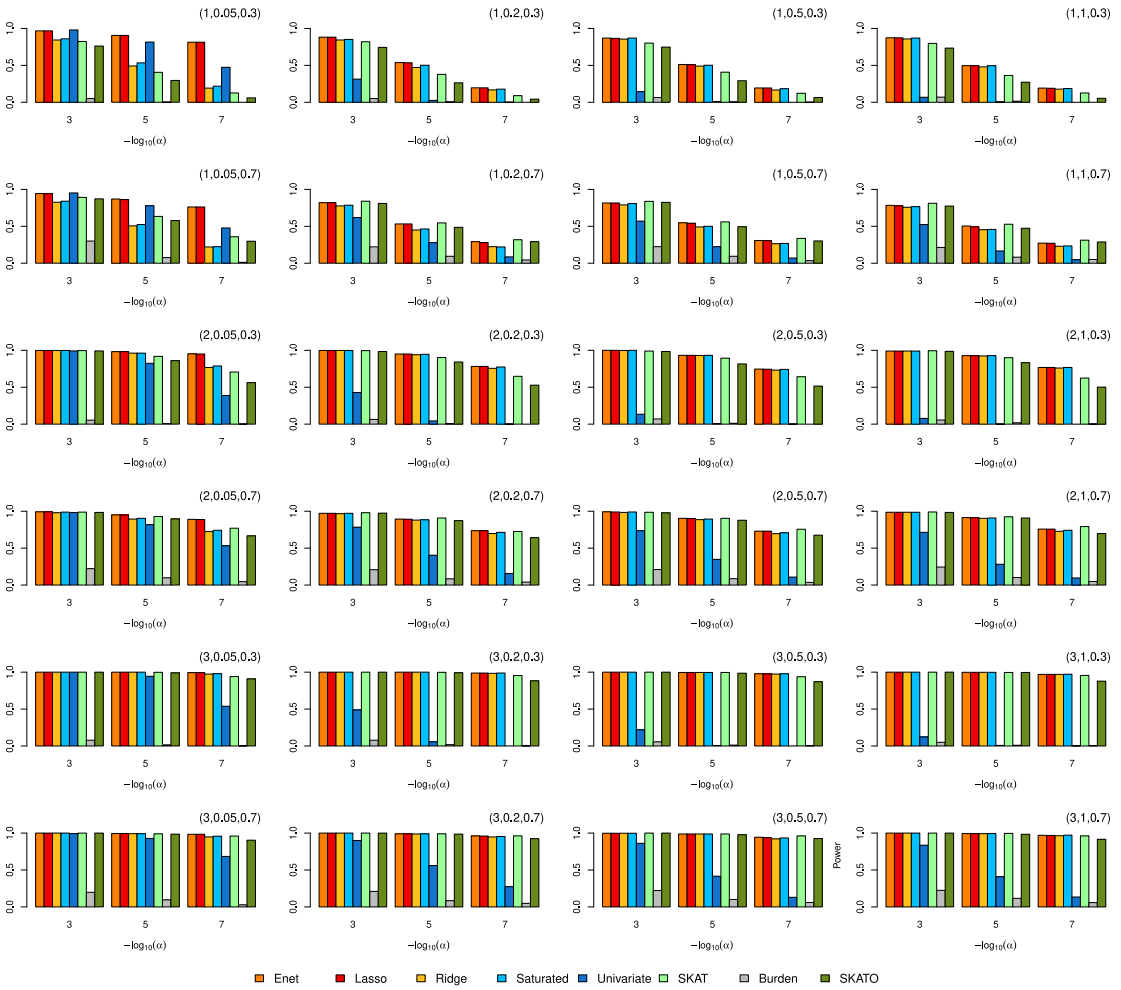

Fig. 4. Power in simulation studies (500 replicates) for scenario 2 with normal error under effect size scenario (b), evaluated at significance level $\alpha \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Triplet (κ, r_b, ρ) on the top right in each panel denotes different setups: κ gives $n = 400\kappa$ and $d = 50\kappa$, ρ specifies the correlation structure of $\boldsymbol{X}$ and r_b denotes the proportion of nonzero variables. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

There remain many topics regarding the optimality of the form (2). The proposed generalization of the Yanai's generalized coefficient of determination was designed for inheriting the monotonically increasing property with dimension under $\boldsymbol{\mu} = \alpha \mathbf{1}_n$ for hypothesis testing. It would no longer be regarded as a model goodness measure unlike the original Yanai's generalized coefficient of determination. Also, it might be possible that other function of $\mathrm{gdf}_0(g_\lambda)$ is suitable in the denominator instead of $\mathrm{gdf}_0(g_\lambda)^{1/2}$. Study on the form of the test statistic is an interesting research direction and shall be considered in future.

Acknowledgments

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data collection and sharing for this project was funded by theAlzheimer's Disease Neuroimaging Initiative (ADNI), United States (National Institutes of Health Grant U01 AG024904) and DOD ADNI, United States (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, United States, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; EliLilly and Company; EuroImmun; F. Hoffmann-La Roche

Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The computation in this work has been done using the facilities of the Institute of Statistical Mathematics. The author is grateful for the helpful comments by two referees, associate editor and Prof. Yoshinori Kawasaki. This work was supported by JSPS, Japan KAKENHI Grant No. 20K11723.

Appendix

Proof of Theorem 1. To prove the theorem, for a given $\gamma \in (0,1)$, model indexes are partitioned into two index sets, $U = \{j : \mathrm{gdf}_{(j)} < d^{1-\gamma}\}$ and $O = \{j : \mathrm{gdf}_{(j)} \geq d^{1-\gamma}\}$. The corresponding Yanai's generalized coefficient of determination is denoted as

$$r_{(j)} = \|\widetilde{\boldsymbol{y}}\|^{-2}\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_{(j)}^{1/2},$$

for $j = 1, \ldots, J$.

Let $\widehat{j} = \mathrm{argmax}_j r_{(j)}$. Then, the type I error of the proposed test procedure is written by

$$P(E_U \cup E_O) \leq P(E_U) + P(E_O), \tag{6}$$

where

$$\begin{aligned} E_U &= \{\|\widetilde{\boldsymbol{y}}\|^2 d^{1/2} r_{\widehat{(j)}} > q_\alpha(d)\} \cap \{\mathrm{gdf}_{\widehat{(j)}} < d^{1-\gamma}\}, \\ E_O &= \{\|\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}\|^2 > q_\alpha(d)\} \cap \{\mathrm{gdf}_{\widehat{(j)}} \geq d^{1-\gamma}\}. \end{aligned} \tag{7}$$

Then, by the Bonferroni inequality,

$$\begin{aligned} P(E_U) &\leq P[\cup_{j\in U}\{\|\widetilde{\boldsymbol{y}}\|^2 d^{1/2} r_{(j)} > q_\alpha(d)\}] \\ &\leq \sum_{j\in U} P\{\|\widetilde{\boldsymbol{y}}\|^2 d^{1/2} r_{(j)} > q_\alpha(d)\} \\ &= \sum_{j\in U} P\{d^{1/2}\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_{(j)}^{1/2} > q_\alpha(d)\}. \end{aligned}$$

Since $\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}}) \geq 0$ almost surely, by the Markov inequality,

$$\begin{aligned} P\{d^{1/2}\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_{(j)}^{1/2} > q_\alpha(d)\} &\leq d^{1/2} E\{\widetilde{\boldsymbol{y}}^T g_{(j)}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_{(j)}^{1/2}\}/q_\alpha(d) \\ &= d^{1/2}\mathrm{gdf}_{(j)}^{1/2}/q_\alpha(d). \end{aligned}$$

For any $j \in U$, the right-hand side is further bounded above by $d^{1/2}d^{(1-\gamma)/2}/q_\alpha(d)$, which converges to zero as $d \to \infty$ because $q_\alpha(d) \geq d$ by assumption. Therefore, $P(E_U) \to 0$ as $d \to \infty$. Next, $P(E_O)$ is bounded above by $P\{\|\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}\|^2 > q_\alpha(d)\}$, which is α due to (5). Consequently, the type I error rate is approximately bounded above by α as $d \to \infty$.

Proof of Theorem 2. To prove the theorem, note that $\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{\epsilon}}$. Since ϵ_i are independently and identically distributed random variables with mean 0 and variance σ_0^2, $E(\widetilde{\boldsymbol{\epsilon}}) = E(\boldsymbol{Q}_{\mathbf{1}_n}\boldsymbol{\epsilon}) = \mathbf{0}$, and $E\|\widetilde{\boldsymbol{\epsilon}}\|^2 = E(\boldsymbol{\epsilon}^T\boldsymbol{Q}_{\mathbf{1}_n}\boldsymbol{\epsilon}) = \sigma_0^2\mathrm{tr}(\boldsymbol{Q}_{\mathbf{1}_n}) = O(n)$.

First, note that

$$\max_\lambda \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2} \geq \widetilde{\boldsymbol{y}}^T g_{\lambda^*}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_{\lambda^*})^{1/2} = \{\widetilde{\boldsymbol{\mu}}^T g_{\lambda^*}(\widetilde{\boldsymbol{y}}) + \widetilde{\boldsymbol{\epsilon}}^T g_{\lambda^*}(\widetilde{\boldsymbol{y}})\}/\mathrm{gdf}_0(g_{\lambda^*})^{1/2}. \tag{8}$$

For the second term in the parenthesis on the most right-hand side of (8), consider the decomposition:

$$\widetilde{\boldsymbol{\epsilon}}^T g_{\lambda^*}(\widetilde{\boldsymbol{y}}) = \widetilde{\boldsymbol{\epsilon}}^T g_{\lambda^*}(\widetilde{\boldsymbol{\mu}}) + \widetilde{\boldsymbol{\epsilon}}^T\{g_{\lambda^*}(\widetilde{\boldsymbol{y}}) - g_{\lambda^*}(\widetilde{\boldsymbol{\mu}})\}.$$

For the first term, by the Cauchy–Schwarz inequality and (D2),

$$|\widetilde{\boldsymbol{\epsilon}}^T g_{\lambda^*}(\widetilde{\boldsymbol{\mu}})| \leq \|\widetilde{\boldsymbol{\epsilon}}\|\|g_{\lambda^*}(\widetilde{\boldsymbol{\mu}})\| = O_p(n^{1/2}\|\widetilde{\boldsymbol{\mu}}\|),$$

and similarly, for the second term, by (D1) and (D2),

$$|\widetilde{\boldsymbol{\epsilon}}^T\{g_{\lambda^*}(\widetilde{\boldsymbol{y}}) - g_{\lambda^*}(\widetilde{\boldsymbol{\mu}})\}| \leq \|\widetilde{\boldsymbol{\epsilon}}\|\|g_{\lambda^*}(\widetilde{\boldsymbol{y}}) - g_{\lambda^*}(\widetilde{\boldsymbol{\mu}})\| \leq K_{\lambda^*}\|\widetilde{\boldsymbol{\epsilon}}\|^2 = O_p(n),$$

both of them are of order $o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)$ by (D4). Therefore, $\widetilde{\boldsymbol{\epsilon}}^T g_{\lambda^*}(\widetilde{\boldsymbol{y}}) = o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)$.

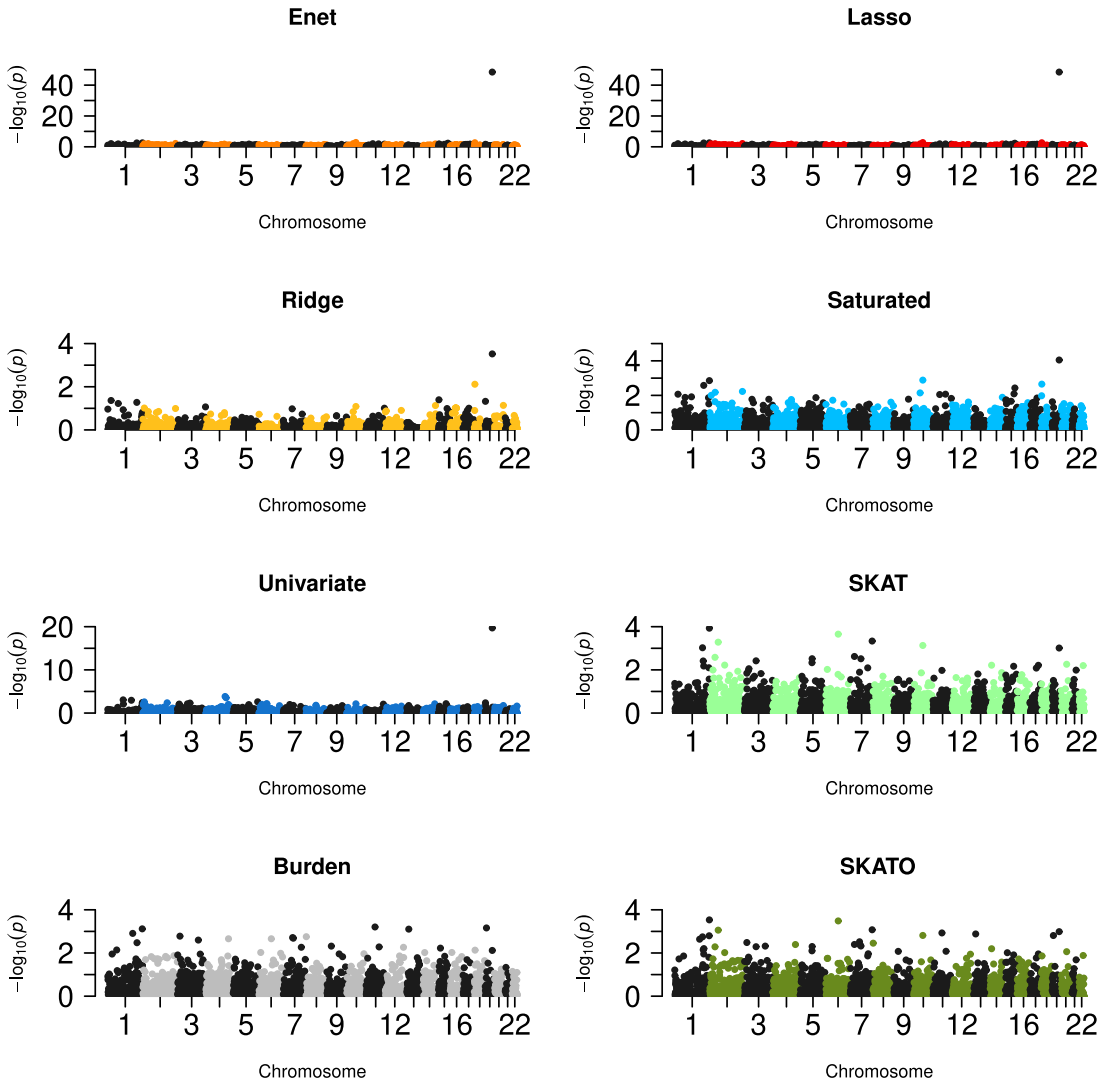

Fig. 5. Manhattan plot from the result in real data application. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test.

Similarly, for the first term in the parenthesis on the most right-hand side of (8), consider the decomposition:

$$\widetilde{\boldsymbol{\mu}}^T g_{\lambda*}(\widetilde{\boldsymbol{y}}) = \widetilde{\boldsymbol{\mu}}^T g_{\lambda*}(\widetilde{\boldsymbol{\mu}}) + \widetilde{\boldsymbol{\mu}}^T \{g_{\lambda*}(\widetilde{\boldsymbol{y}}) - g_{\lambda*}(\widetilde{\boldsymbol{\mu}})\}.$$

For the second term, by (D1) and (D2),

$$|\widetilde{\boldsymbol{\mu}}^T \{g_{\lambda*}(\widetilde{\boldsymbol{y}}) - g_{\lambda*}(\widetilde{\boldsymbol{\mu}})\}| \leq \|\widetilde{\boldsymbol{\mu}}\| \|g_{\lambda*}(\widetilde{\boldsymbol{y}}) - g_{\lambda*}(\widetilde{\boldsymbol{\mu}})\| \leq \|\widetilde{\boldsymbol{\mu}}\| K_{\lambda*} \|\widetilde{\boldsymbol{\epsilon}}\| = O_p(n^{1/2}\|\widetilde{\boldsymbol{\mu}}\|),$$

which is of order $o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)$ by (D4). Recalling (D2),

$$\widetilde{\boldsymbol{y}}^T g_{\lambda*}(\widetilde{\boldsymbol{y}}) = C_{\lambda*}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2). \tag{9}$$

Therefore, by (D3), the left-hand side of (8) becomes

$$\max_{\lambda} \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2} \geq C_{\lambda*}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2). \tag{10}$$

Hence, the test statistic increases at the same or faster rate of $\|\boldsymbol{\mu}\|^2$.

It needs to exclude the situation where the generalized degrees of freedom at the optimal λ, i.e. $\mathrm{argmax}_\lambda \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2}$ are greater than $d^{1-\gamma}$; otherwise, the test statistic reduces to that on the saturated model by definition of the

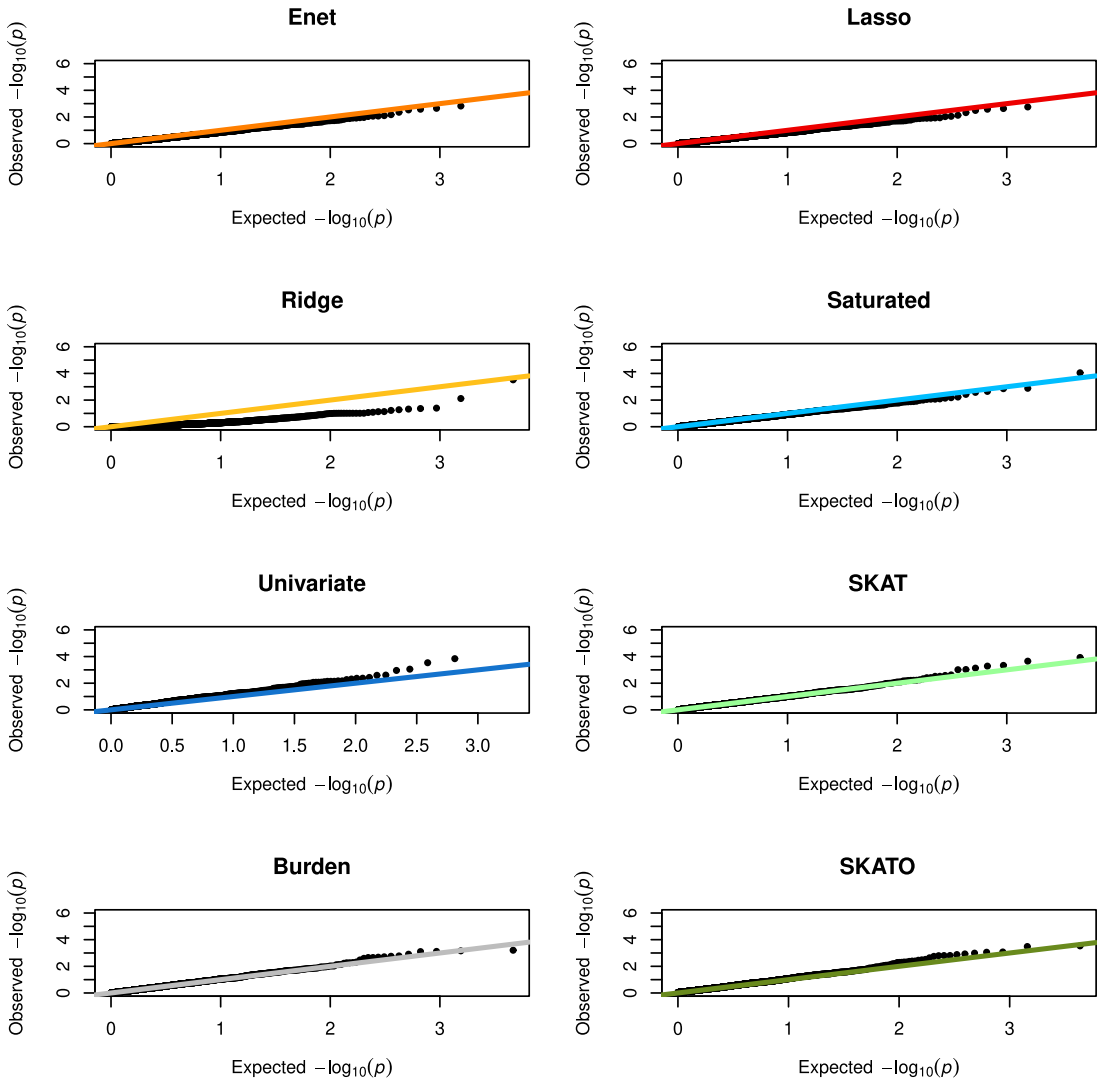

Fig. 6. Quantile–quantile plot from the result in real data application. Enet, proposed test for elastic net ($\gamma = 0.01$); Lasso, proposed test for lasso ($\gamma = 0.01$); Ridge, proposed test for ridge regression ($\gamma = 0.01$); Saturated, saturated model test; Univariate, univariate test; SKAT, sequence kernel association test; Burden, burden test; SKATO, optimized sequence kernel association test. For elastic net and lasso tests, plotted area on y-axis is restricted up to 6 for visibility.

procedure. To this end, for any λ_O such that $\mathrm{gdf}_0(g_{\lambda_O}) \geq d^{1-\gamma}$, it suffices that the following probability converges to 0:

$$P\{\max_{\lambda} \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2} \leq \widetilde{\boldsymbol{y}}^T g_{\lambda_O}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_{\lambda_O})^{1/2}\}. \tag{11}$$

Due to (10), (11) is bounded above by

$$P\{C_{\lambda^*}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2) \leq \widetilde{\boldsymbol{y}}^T g_{\lambda_O}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_{\lambda_O})^{1/2}\}. \tag{12}$$

Note that, by (D3), $\lambda_O \leq \lambda^*$, and $\widetilde{\boldsymbol{\mu}}^T g_{\lambda_O}(\widetilde{\boldsymbol{\mu}}) = C_{\lambda_O}\|\widetilde{\boldsymbol{\mu}}\|^2$ by (D2). Thus, an analogous argument in deriving (9) gives

$$\widetilde{\boldsymbol{y}}^T g_{\lambda_O}(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_{\lambda_O})^{1/2} = \{C_{\lambda_O}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)\}/\mathrm{gdf}_0(g_{\lambda_O})^{1/2} \leq \{C_{\lambda_O}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)\}/d^{(1-\gamma)/2},$$

in which the last inequality is due to $\mathrm{gdf}_0(g_{\lambda_O}) \geq d^{1-\gamma}$. Therefore, (12) is bounded above by

$$P[C_{\lambda^*}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2) \leq \{C_{\lambda_O}\|\widetilde{\boldsymbol{\mu}}\|^2 + o_p(\|\widetilde{\boldsymbol{\mu}}\|^2)\}/d^{(1-\gamma)/2}], \tag{13}$$

which converges to 0 as $d \to \infty$, and so as for (11). Consequently, (10) holds with probability tending to 1, implying that the test statistic increases at the same or faster rate of $\|\widetilde{\boldsymbol{\mu}}\|^2$.

Now, it is ready to compare with the test under the saturated model. Note that g_λ at $\lambda = 0$ gives the saturated model. Therefore, if

$$d^{1/2} \max_\lambda \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2} > d^{1/2}\widetilde{\boldsymbol{y}}^T g_0(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_0)^{1/2} = \|\boldsymbol{P}_{\widetilde{\boldsymbol{X}}}\widetilde{\boldsymbol{y}}\|^2 \tag{14}$$

holds with probability tending to 1 as $d \to \infty$, the proposed test is more powerful than the test under the saturated model because the significance threshold $q_\alpha(d)$ for the former is common with that for the latter. Since $\mathrm{gdf}_0(g_0) = d$, due to (14), it suffices to show that

$$P\{\max_\lambda \widetilde{\boldsymbol{y}}^T g_\lambda(\widetilde{\boldsymbol{y}})/\mathrm{gdf}_0(g_\lambda)^{1/2} \le \widetilde{\boldsymbol{y}}^T g_0(\widetilde{\boldsymbol{y}})/d^{1/2}\} \to 0, \tag{15}$$

as $d \to \infty$ due to $\mathrm{gdf}_0(g_0) = d \ge d^{1-\gamma}$. By an analogous argument in (11)–(13) in which γ is replaced by δ, (15) holds as $d \to \infty$, showing the claim in (14).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107168. Supplementary Tables S1-S10 and Figures S1-S22 are included.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Bush, W.S., Oetjens, M.T., Crawford, D.C., 2016. Unravelling the human genome–phenome relationship using phenome-wide association studies. Nature Rev. Genet. 17 (3), 129–145.

Cadima, J.F.C.L., Jolliffe, I.T., 2001. Variable selection and the interpretation of principal subspaces. J. Agric. Biol. Environ. Stat. 6 (1), 62–79.

Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95 (3), 759–771.

Chen, X., Lin, Q., Sen, B., 2019. On degrees of freedom of projection estimators with applications to multivariate nonparametric regression. J. Amer. Statist. Assoc. 115 (529), 173–186.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31 (4), 377–403.

Dossal, C., Kachour, M., Fadili, M., Peyré, G., Chesneau, C., 2013. The degrees of freedom of the lasso for general design matrix. Statist. Sinica 23 (2), 809–828.

Dudbridge, F., 2008. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum. Hered. 66 (2), 87–98.

Efron, B., 2004. The estimation of prediction error. J. Amer. Statist. Assoc. 99 (467), 619–632.

Foster, D.P., George, E.I., 1994. The risk inflation criterion for multiple regression. Ann. Statist. 22 (4), 1947–1975.

Freidlin, B., Zheng, G., Li, Z., Gastwirth, J.L., 2002. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. Hum. Hered. 53 (3), 146–152.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2 (4), 189–210.

González, J.R., Carrasco, J.L., Dudbridge, F., Armengol, L., Estivill, X., Moreno, V., 2008. Maximizing association statistics over genetic models. Genet. Epidemiol. 32 (3), 246–254.

Hirotsu, C., Aoki, S., Inada, T., Kitao, Y., 2001. An exact test for the association between the disease and alleles at highly polymorphic loci with particular interest in the haplotype analysis. Biometrics 57 (3), 769–778.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Hothorn, L.A., Hothorn, T., 2009. Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. Biom. J. 51 (4), 659–669.

Jolliffe, I., 2002. Principal Component Analysis. Springer Verlag, New York.

Joo, J., Kwak, M., Chen, Z., Zheng, G., 2010. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. Stat. Med. 29 (1), 158–180.

Kaufman, S., Rosset, S., 2014. When does more regularization imply fewer degrees of freedom? sufficient conditions and counterexamples. Biometrika 101 (4), 771–784.

Kraft, P., Yen, Y.-C., Stram, D.O., Morrison, J., Gauderman, W.J., 2007. Exploiting gene-environment interaction to detect genetic associations. Hum. Hered. 63 (2), 111–119.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103 (10), 3863–3868.

Lee, S., Wu, M.C., Lin, X., 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13 (4), 762–775.

Li, Q., Zheng, G., Li, Z., Yu, K., 2008. Efficient approximation of P-value of the maximum of correlated tests, with applications to genome-wide association studies. Ann. Hum. Genet. 72 (3), 397–406.

Madsen, B.E., Browning, S.R., 2009. A groupwise association test for rare mutations using a weighted sum statistic. In: Schork, N.J. (Ed.), PLoS Genet. 5 (2), e1000384.

Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Statist. 12 (2), 758–765.

Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S., 2011. Epigenome-wide association studies for common human diseases. Nature Rev. Genet. 12 (8), 529–541.

Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. Science 273 (5281), 1516–1517.

Schaid, D.J., Chen, W., Larson, N.B., 2018. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Rev. Genet. 19 (8), 491–504.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., Poland, G.A., 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. 70 (2), 425–434.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.

Sham, P.C., Curtis, D., 1995. Monte Carlo Tests for associations between disease and alleles at highly polymorphic loci. Ann. Hum. Genet. 59 (1), 97–105.
Shao, J., 1997. An asymptotic theory for linear model selection (with discussion). Statist. Sinica 7 (2), 221–264.
The UK10K Consortium, 2015. The UK10k project identifies rare variants in health and disease. Nature 526 (7571), 82–90.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Tibshirani, R., Taylor, J., 2012. Degrees of freedom in lasso problems. Ann. Statist. 40 (2), 1198–1232.
Ueki, M., 2014. On the choice of degrees of freedom for testing gene-gene interactions. Stat. Med. 33 (28), 4934–4948.
Ueki, M., Kawasaki, Y., Tamiya, G., 2017. Detecting genetic association through shortest paths in a bidirected graph. Genet. Epidemiol. 41 (6), 481–497.
Wang, S., Cui, H., 2013. Generalized f test for high dimensional linear regression coefficients. J. Multivariate Anal. 117, 134–149.
Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89 (1), 82–93.
Yanai, H., 1980. A proposition of generalized method for forward selection of variables. Behaviormetrika 7 (7), 95–107.
Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. J. Amer. Statist. Assoc. 93 (441), 120–131.
Zang, Y., Fung, W.K., 2011. Robust mantel–haenszel test under genetic model uncertainty allowing for covariates in case-control association studies. Genet. Epidemiol. 35 (7), 695–705.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2), 301–320.
Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the lasso. Ann. Statist. 35 (5), 2173–2192.