OPEN ACCESS Freely available online

PLOS ONE

Rare Variants Detection with Kernel Machine Learning Based on Likelihood Ratio Test

Ping Zeng[1,2], Yang Zhao[1], Liwei Zhang[1], Shuiping Huang[2], Feng Chen[1]*

1 Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu, China, **2** Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical College, Xuzhou, Jiangsu, China

Abstract

This paper mainly utilizes likelihood-based tests to detect rare variants associated with a continuous phenotype under the framework of kernel machine learning. Both the likelihood ratio test (LRT) and the restricted likelihood ratio test (ReLRT) are investigated. The relationship between the kernel machine learning and the mixed effects model is discussed. By using the eigenvalue representation of LRT and ReLRT, their exact finite sample distributions are obtained in a simulation manner. Numerical studies are performed to evaluate the performance of the proposed approaches under the contexts of standard mixed effects model and kernel machine learning. The results have shown that the LRT and ReLRT can control the type I error correctly at the given α level. The LRT and ReLRT consistently outperform the SKAT, regardless of the sample size and the proportion of the negative causal rare variants, and suffer from fewer power reductions compared to the SKAT when both positive and negative effects of rare variants are present. The LRT and ReLRT performed under the context of kernel machine learning have slightly higher powers than those performed under the context of standard mixed effects model. We use the Genetic Analysis Workshop 17 exome sequencing SNP data as an illustrative example. Some interesting results are observed from the analysis. Finally, we give the discussion.

Citation: Zeng P, Zhao Y, Zhang L, Huang S, Chen F (2014) Rare Variants Detection with Kernel Machine Learning Based on Likelihood Ratio Test. PLoS ONE 9(3): e93355. doi:10.1371/journal.pone.0093355

Editor: Lin Chen, The University of Chicago, United States of America

Received November 16, 2013; **Accepted** March 3, 2014; **Published** March 27, 2014

Funding: This work was supported by National Natural Science Foundation of China (No. 81072389, 81373102), Research Found for the Doctoral Program of Higher Education of China (No. 20113234110002), Key Grant of Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 10KJA330034), College Philosophy and Social Science Foundation from Education Department of Jiangsu Province of China (No. 2013SJB790059, 2013SJD790032), Research Foundation from Xuzhou Medical College (No. 2012KJ02), Research and Innovation Project for College Graduates of Jiangsu Province of China (No. CXLX13_574), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fengchen@njmu.edu.cn

Introduction

For next-generation sequencing data identifying rare variants associated with phenotypes of interest is both practically and theoretically important [1–3]. Here the rare variant is typically defined as allele with minor allele frequency (MAF) less than 1%. The past few years have witnessed increasing evidence that the rare variants play an important role in many complex diseases and disorders [4–16]. There are also some other findings supporting the contributions of rare variants to the diseases. For example, according to the odds ratio (OR) distribution, it has been demonstrated that most rare variants have values above 2 and the mean OR is 3.74, while very few common variants (defined as MAF>1%) have values above 2 and the mean OR is 1.36 [17]. See also Box 1 in Cirulli and Goldstein [2].

However, it is a very challenging task to detect the casual rare variants due to their extremely low MAF. For rare variant association analyses the single locus methods designed for common variants are rather underpowered or not applicable [1,18–20], thus developing appropriate statistical approaches especially for rare variant has become an active research topic recently. A type of methods has been proposed by collapsing the rare variants within a functional region (e.g., gene and pathway) into one variant and then testing this collapsed variant [21–23]. In this paper, those tests are referred to as the burden test since they share the similar reasoning of collapsing. The burden test may be limited because it explicitly assumes that the variants within the collapsed region have the same direction of effect. However, in practice both protective and deleterious effects exist [1,18,19,24–26].

More recently, Wu et al. [18] proposed the sequence kernel association test (SKAT) for rare variant detection. The SKAT is a score based variance component test originally developed by Lin [27] under the framework of mixed effects model [28], and has been widely applied to pathway or gene set analyses [29–32]. Two very attractive features of the SKAT are that: (I) it avoids the directionality of effect and consequently can enhance the statistical power when both protective and deleterious effects are present; (II) it proceeds under the framework of kernel machine learning, and thus can capture more complicated nonlinear relationship among rare variants.

The SKAT, however, has itself shortcomings as argued by Zhan and Xu [16]. For SKAT, a large score value (i.e., a small p value) does not necessarily mean the effect of a group of rare variants is also great, it may be due to a lot of variants with very weak effects. Additionally, when examining a set of rare variants, geneticists and epidemiologists may need some metrics to measure their contribution together, like OR in logistic regression or estimated coefficient in linear regression in single locus association analysis. While the SKAT will not involve any parameter estimation, thus

cannot show effect differences across various sets of rare variants. Consequently, methods for rare variants with the capability to offer such information are desirable.

Motivated by the arguments above, in this paper we adopt the likelihood ratio test to detect the rare variants. Both the likelihood ratio test (LRT) and the restricted likelihood ratio test (ReLRT) are investigated and are performed under the same framework of mixed effects model of SKAT. A great advantage of LRT and ReLRT is that they not only examine the effect of a group of rare variants but also offer an effect measurement; this value in turn can be used to evaluate the relative importance of rare variants. To our best knowledge, the likelihood-based methods for rare variants have been not published before, nor are investigated under the framework of kernel machine learning, although the LRT and ReLRT are particularly popular in the literature.

In the rest of the paper, the SKAT and the burden test are first introduced, and then the LRT and ReLRT are discussed under the mixed effects model context and the kernel machine learning context, respectively. In this section, we will interpret how the kernel machine learning can be addressed with the mixed effects model and examine a group of rare variants via LRT and ReLRT. By using the eigenvalue representation of LRT and ReLRT, their exact finite sample distributions are obtained in a simulation manner. We perform extensive numerical studies to evaluate the performance of the proposed approaches and compare with the burden test and SKAT. The exome sequencing data from Genetic Analysis Workshop 17 (GAW17) is used as a practical application.

Methods

Notation

Let $X=[x_1, \ldots, x_p]$ denote the covariate vector of order p such as age, sex, smoking, and environmental exposure, and $G=[g_1, g_2, \ldots, g_m]$ the genotype vector of order m for rare variants within a functional region specified a priori. In the paper, we use the additive genetic model, so that $g=0$, 1, and 2 represent the number of minor alleles. For example, in the GAW17 data [33,34], there are 16 single nucleotide polymorphisms (SNPs) included within the gene *KDR*, then the genotype can be expressed as $G=[g_1, g_2, \ldots, g_{16}]$. Let Y denote the continuous phenotype of interest (e.g., weight, blood pressure, and triglyceride) and y_i, $i=1, 2, \ldots, n$ its realization values, here n is the sample size. Suppose further that the phenotype Y follows a normal distribution with variance σ^2 conditional on the covariates X and genotypes G.

Mixed effects model

First consider the linear mixed effects model [28,35]

$$\begin{aligned} \boldsymbol{Y} &= \beta_0 + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N\left(0, \sigma^2 \mathbf{I}_n\right), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}=[\beta_1, \ldots, \beta_p]$ are the fixed effects for covariates, β_0 is the intercept, and $\mathbf{I}_n$ is an identity matrix of order n; here $\boldsymbol{\gamma}=[\gamma_1, \ldots, \gamma_m]$ are the random effects for rare genotypes, each $\gamma_j, j=1, 2, \ldots, m$ is assumed to be normally distributed with mean zero and variance τw_j^2, where τ is a variance component and w is a prespecified weight related to MAF. For rare variant, $w=\text{Beta}(\text{MAF}; 1, 25)$ is recommended in Wu et al [18], which places more weight on rarer variant and less weight on common variant, where Beta is the beta density function. In the present paper we also follow this idea, but make a slight modification. That is, a scaled weight of $w_j=w_j/\max(\boldsymbol{w})$ is used, where the notation max indicates the maximum over all the w_js. In our experience, this modification is necessary to avoid numerical imprecisions encountered in the statistical software, such as the R statistical environment [36]. Greven et al. [37] gave a full description regarding this issue when performing the restricted likelihood ratio test for zero variance component in the linear mixed effects model.

Under these conditions, we can obtain

$$\text{Var}(\boldsymbol{Y}) = \tau \mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}' + \sigma^2 \mathbf{I}_n = \sigma^2 \mathbf{V}_\lambda, \tag{2}$$

where $\lambda=\tau/\sigma^2$, $\mathbf{V}_\lambda=\lambda\mathbf{G}\mathbf{W}\mathbf{W}\mathbf{G}'+\mathbf{I}_n$, $\mathbf{W}$ is a diagonal matrix of order m with elements being w. Clearly testing whether or not a group of rare variants are collectively associated with the phenotype is equivalent to testing the null hypothesis H_0: $\lambda=0$. Note that the classical definition of heritability is defined as $\tau/(\tau+\sigma^2)$, i.e., the proportion of phenotypic variance explained by a group of rare variants [38], then the heritability can be further expressed as $\lambda/(1+\lambda)$. Therefore the quantity λ is an analogue of the heritability and can be employed for measuring the relative impotence of different groups of rare variants.

Sequence kernel association test (SKAT)

According to Lee et al. [39] and Lee et al. [40], the original SKAT in Wu et al. [18] and the burden test can be studied within a unified framework if taking into account the correlation structure of the random effects. Suppose that the correlation structure among the m rare variants is $\mathbf{R}_\rho$, which is determined by the pairwise correlation coefficient $\text{corr}(g_j, g_l)=\rho$ between any variants j and l. The unified SKAT statistic is given as

$$\begin{aligned} Q &= \left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right)' \mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\mathbf{G}'\left(\boldsymbol{Y}-\hat{\boldsymbol{Y}}\right), \\ \mathbf{R}_\rho &= (1-\rho)\mathbf{I}_m + \rho 1_m 1'_m, \end{aligned} \tag{3}$$

where $\hat{\boldsymbol{Y}}$ is the predicted value under H_0. The test in Equation (3) is called the optimal SKAT (SKAT-O) since it can choose the correlation coefficient ρ adaptively to maximize the power when all the effects are in the same direction [39,40].

When $\rho=0$ (i.e., independent correlation), the SKAT-O reduces to the original SKAT in Wu et al. [18] and Lin [27], and when $\rho=1$ (i.e., perfect correlation), the optimal SKAT reduces to the burden test.

Under H_0, Q follows a mixture of chi-square distributions, the p values for the burden test and SKAT are obtained by the Davies method [41] or other methods [42,43]. The p value for the SKAT-O is obtained by using a grid search strategy [39,40].

Likelihood ratio test (LRT) and restricted likelihood ratio test (ReLRT)

When examining variance component in the mixed effects model, the LRT and ReLRT are a natural alternative. Note that the null hypothesis H_0: $\lambda=0$ is non-standard since under H_0 λ is on the boundary of the parameter space [44–47], and $\lambda=0$ if and only if $\tau=0$. The parameter space for λ is $\Omega=[0, \infty)$.

Replacing $\boldsymbol{\gamma}$ and σ^2 in model (1) with their maximum likelihood (ML) estimators [47], we obtain the profile log-likelihood function up to a constant independent of the parameters

$$L(\lambda) = -\frac{1}{2}\left\{n\log\left(\boldsymbol{Y}'\mathbf{P}'_\lambda\mathbf{V}_\lambda^{-1}\mathbf{P}_\lambda\boldsymbol{Y}\right) + \log|\mathbf{V}_\lambda|\right\}, \tag{4}$$

where

$$\mathbf{P}_\lambda = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{V}_\lambda^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}_\lambda^{-1}. \quad (5)$$

The LRT statistic is defined as

$$\begin{aligned}\mathrm{LRT}_n &= 2[\sup_{\lambda\in\Omega} L(\lambda) - L(\lambda=0)],\\ &= \sup_{\lambda\in\Omega}\left\{-n\log\left(\boldsymbol{Y}'\mathbf{P}'_\lambda\mathbf{V}_\lambda^{-1}\mathbf{P}_\lambda\boldsymbol{Y}\right) - \log|\mathbf{V}_\lambda| + n\log(\boldsymbol{Y}'\mathbf{P}_0\boldsymbol{Y})\right\},\end{aligned} \quad (6)$$

where

$$\mathbf{P}_0 = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'. \quad (7)$$

Using the spectral representation [37,47–49], it can be shown that LRT_n is equal to the following quantity in distribution

$$\begin{aligned}&f_n(\lambda)\\ &= \sup_{\lambda\in\Omega}\left\{n\log\left[1+\frac{N_n(\lambda)}{D_n(\lambda)}\right] - \sum_{j=1}^{m}\log(1+\lambda\xi_j) + n\log(\mathrm{Y}'\mathrm{P}_0\mathrm{Y})\right\},\end{aligned} \quad (8)$$

where ξ_j's are the eigenvalues of matrix $\mathbf{W}^{1/2}\mathbf{G}'\mathbf{G}\mathbf{W}^{1/2}$, and

$$\begin{aligned}N_n(\lambda) &= \sum_{j=1}^{m}\frac{\lambda\mu_j}{1+\lambda\mu_j}u_j^2,\\ D_n(\lambda) &= \sum_{j=1}^{m}\frac{1}{1+\lambda\mu_j}u_j^2 + \sum_{j=m+1}^{n-p}u_j^2,\end{aligned} \quad (9)$$

where μ_j's are the eigenvalues of matrix $\mathbf{W}^{1/2}\mathbf{G}'\mathbf{P}_0\mathbf{G}\mathbf{W}^{1/2}$, and u_j's are independently standard normal random variables.

The ML estimator of σ^2 is biased downward since it does not take into account the loss in degrees of freedom due to estimation of $\boldsymbol{\gamma}$. While the restricted maximum likelihood (REML) method provides an unbiased estimator for σ^2 by using a set of n - p linearly independent error contrasts [50–53]. The profile restricted log-likelihood function up to a constant independent of the parameters is given as

$$\begin{aligned}&L_{\mathrm{Re}}(\lambda)\\ &= \frac{1}{2}\left\{-\log|\mathrm{V}_\lambda| - (n-p)\log\left(\mathrm{Y}'\mathrm{P}'_\lambda\mathrm{V}_\lambda^{-1}\mathrm{P}_\lambda\mathrm{Y}\right) - \log\left|\mathrm{X}'\mathrm{V}_\lambda^{-1}\mathrm{X}\right|\right\}.\end{aligned} \quad (10)$$

The ReLRT statistic is defined as

$$\mathrm{ReLRT}_n = 2[\sup_{\lambda\in\Omega} L_{\mathrm{Re}}(\lambda) - L_{\mathrm{Re}}(\lambda=0)]. \quad (11)$$

Using the similar reasoning for LRT_n, it can be shown that ReLRT_n is equal to

$$\begin{aligned}&f_n(\lambda) =\\ &\sup_{\lambda\in\Omega}\left\{(n-p)\log\left[1+\frac{N_n(\lambda)}{D_n(\lambda)}\right] - \sum_{j=1}^{m}\log(1+\lambda\mu_j) + (n-p)\log(\mathrm{Y}'\mathrm{P}_0\mathrm{Y})\right\}\end{aligned} \quad (12)$$

in distribution.

By taking full advantage of the spectral representation used in Equations (8) and (12), Crainiceanu and Rupper [47] described a simulation-based algorithm for the finite sample distributions of LRT_n and ReLRT_n. This algorithm has been shown to be rather fast and accurate. The p values of the LRT and ReLRT are obtained by comparing the observed statistics to those simulated values.

Kernel machine learning

So far we have discussed how to detect the causal rare variants by using the LRT and ReLRT which are developed under the standard mixed effects model context. In this section, we turn to the recently popular kernel machine learning, explore its relationship with the mixed effects model, and demonstrate how to detect the causal rare variants in the kernel machine learning context via LRT and ReLRT. As we will see, there is a close connection between these two statistical theories, which provides a more flexible way for rare variant detection with kernel methods.

Using the same notation defined before, we describe the relationship between the phenotype Y and genotypes G and covariates X via a semi-parametric linear model [30,31]

$$y_i = \beta_0 + X_i\boldsymbol{\beta} + h(G_i) + \varepsilon_i, \quad (13)$$

where h is an unknown smooth function lying in a Hilbert space H_K generated by a positive definite kernel function K [31,54]. This space is called reproducing kernel Hilbert space (RKHS) under some regularity conditions [55–58]. The kernel function K essentially quantifies the genomic similarity or distance of two subjects and can be arbitrarily chosen as long as it satisfies the conditions of Mercer's theorem [55,57]. Model (13) is semi-parametric since the covariates X are fitted parametrically while the genotypes G are fitted non-parametrically.

To avoid over-fitting, estimation of h can be performed by maximizing the penalized log-likelihood function [31,59]

$$L_{H_K}(h|\zeta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[1y_{1i} - \beta_0 - 1X_{1i}\boldsymbol{\beta} - h(G_{1i})]^2 - \frac{1}{2}\zeta\|h\|^2_{H_K}, \quad (14)$$

where ζ is a penalization parameter controlling the balance between the goodness of fit and the complexity of the model [31,59], and the notation $\|\cdot\|$ is the norm in RKHS. The solution of h in Equation (14) is given in terms of the well-known representer theorem of Kimeldorf and Wahba [60] and Wahba [61]

$$h(G) = \sum_{l=1}^{n}\alpha_l K(G, G_l), \quad (15)$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_n]$ is an unknown vector of parameters and K is a reproducing kernel function [31,54].

We further rewrite h in the form of matrix as

$$h(G) = \mathbf{K}\boldsymbol{\alpha}, \quad (16)$$

where $\mathbf{K}$ is an $n\times n$ kernel matrix with its elements being $K(G_i, G_l)$. Various kernel functions have been designed in genetic statistics [59,62], such as the linear kernel, the polynomial kernel, the Gaussian kernel, and the identify by state (IBS) kernel. The explicit forms for these kernels can be found in Wu et al. [18], Wu et al. [32], Liu et al. [31], Kwee et al. [29], and Liu et al. [30]. If a kernel is weighted, then it is called a weighted kernel. In the paper the scaled weight described in Section 2.2 is used. Additionally,

once the kernel function is chosen, we assume that $\mathbf{K}$ is known completely. Consequently, inference about h in model (14) immediately reduces to inference about $\boldsymbol{\alpha}$.

Replacing h in Equation (14) with (16) yields

$$\begin{aligned} &L_{H_K}(h|\zeta) \\ &= -\frac{1}{2\sigma^2}(\mathrm{Y}-\beta_0-\mathrm{X}\beta-\mathrm{K}\alpha)'(\mathrm{Y}-\beta_0-\mathrm{X}\beta-\mathrm{K}\alpha)-\frac{1}{2}\zeta\|\mathrm{K}\alpha\|^2_{H_K}. \end{aligned} \tag{17}$$

Following the results of Gianola et al. [54], Wahba [61], and Wahba [63],

$$\|\mathbf{K}\boldsymbol{\alpha}\|^2_{H_K}=\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}, \tag{18}$$

Equation (17) is re-expressed as

$$\begin{aligned} &L_{H_K}(h|\zeta) \\ &= -\frac{1}{2\sigma^2}(\mathrm{Y}-\beta_0-\mathrm{X}\beta-\mathrm{K}\alpha)'(\mathrm{Y}-\beta_0-\mathrm{X}\beta-\mathrm{K}\alpha)-\frac{1}{2}\zeta\alpha'\mathrm{K}\alpha. \end{aligned} \tag{19}$$

From a Bayesian perspective [31,59], Equation (19) is the log-posterior distribution of β_0, $\boldsymbol{\beta}$ and α, thus can be described as the following hierarchical model

$$\begin{aligned} \boldsymbol{Y}|\beta_0,\boldsymbol{\beta},\boldsymbol{\alpha} &\sim N\left(\beta_0+\mathbf{X}\boldsymbol{\beta}+\mathbf{K}\boldsymbol{\alpha},\sigma^2\right), \\ \boldsymbol{\alpha} &\sim N\left(0,\tau\mathbf{K}^{-1}\right), \\ \beta_0,\boldsymbol{\beta} &\propto 1, \end{aligned} \tag{20}$$

where $\tau = 1/\zeta$. Since $h = \mathbf{K}\boldsymbol{\alpha}$, alternatively the hierarchical model is re-expressed as

$$\begin{aligned} \boldsymbol{Y}|\beta_0,\boldsymbol{\beta},\boldsymbol{h} &\sim N\left(\beta_0+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{h},\sigma^2\right), \\ \boldsymbol{h} &\sim N(0,\tau\mathbf{K}), \\ \beta_0,\boldsymbol{\beta} &\propto 1. \end{aligned} \tag{21}$$

In the paper we use the hierarchical model (21) since it avoids the calculation of inverse matrix and therefore reduces the computational cost.

Based on the arguments described above, we can construct the relationship between the semi-parametric linear model (13) and the mixed effects model (1). That is, model (13) is equivalent to the following mixed effects model

$$\begin{aligned} \boldsymbol{Y} &= \beta_0+\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{h}+\boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N\left(0,\sigma^2\mathbf{I}_n\right). \end{aligned} \tag{22}$$

The differences between model (22) and model (1) mainly lie in two aspects: (I) here $\mathbf{Z}$ is an identify matrix of order n, while $\mathbf{G}$ in model (1) is of dimension $n\times m$; (II) the unknown parameter $\boldsymbol{h}$ here is an n-dimensional vector with its covariance-variance matrix being $\tau\mathbf{K}$, while in model (1) the unknown parameter $\boldsymbol{\gamma}$ is an m-dimensional vector with its covariance-variance matrix being $\mathrm{diag}(\tau w_j^2)$, here the notation diag indicates a diagonal matrix.

Therefore, all the theories for the LRT and ReLRT developed under the context of mixed effects model can be also applicable in the context of kernel machine learning. The test of variance component in model (22) can proceed similarly in model (1). To distinct these two types of approaches, in the reminder of the paper, LRT.M and ReLRT.M are used to indicate the LRT and ReLRT for the mixed effects model, LRT.K and ReLRT.K are used to indicate the LRT and ReLRT for the kernel machine learning, and LRT and ReLRT are used to indicate both the two types.

Results

Simulation datasets

We generate genotypes based on the coalescent model for European population by using the package COSI [64]. A total of 100 kb gene region is simulated. Randomly selected continuous 30% subregions of the simulated genotypes are used. Variants with MAF less than 0.01 are defined as rare variants. Two covariates are considered, x_1 is a standard normal variable and x_2 is a binary variable with rate 0.5, and mutually independent. The sample size n is 300, 400, and 500.

For type I error simulations the phenotype is generated as

$$y\sim N(1.0+0.5x_1+0.5x_2,1),$$

and the number of runs is 2,000. In power simulations, 30% rare variants are causal variants, the effect size $|\gamma|$ is $0.3|\log_{10}\mathrm{MAF}|$, which leads to a size of 1.2 for MAF = 0.0001 and a size of 0.6 for MAF = 0.01. Among the causal rare variants, 0%, 30% or 50% have negative effects, i.e., in these settings their effects are $-0.3|\log_{10}\mathrm{MAF}|$. For power simulations the phenotype is generated as

$$y\sim N\left(1.0+0.5x_1+0.5x_2+\sum\nolimits_{j=1}^{q} g_j^c\gamma_j^c,1\right),$$

where q is the number of chosen causal rare variants, g^c's are the genotypes and γ^c's are the effect sizes given above. The number of runs is 1,000. The simulation characteristics under these specifications are displayed in **Table** 1.

In the present paper, seven methods including burden test, SKAT, SKAT-O, LRT.K, ReLRT.K, LRT.M, and ReLRT.M are compared. The first three tests are performed in the package SKAT [18], and the LRT and ReLRT are performed in the package RLRsim[65]. In practice the weighted kernel has been empirically shown to be more powerful compared to its unweighted counterpart [18,29], thus here we only consider the former. For comparison only the weighted linear kernel is used since under this situation both the mixed effects models in (1) and (22) are well specified, and the burden test and the SKAT-O are only able to be performed on the linear kernel.

Type I error and power

Table 2 displays the estimated Type I errors for all the tests. It can be seen from **Table** 2 that all the tests control the type I error correctly at the given α level. **Figure** 1 shows the estimated

Table 1. Simulation characteristics.

n	Total SNPs	Selected SNPs	Used rare variants	Causal rare variants
300	417	125	41	12
400	434	130	47	14
500	447	134	51	15

doi:10.1371/journal.pone.0093355.t001

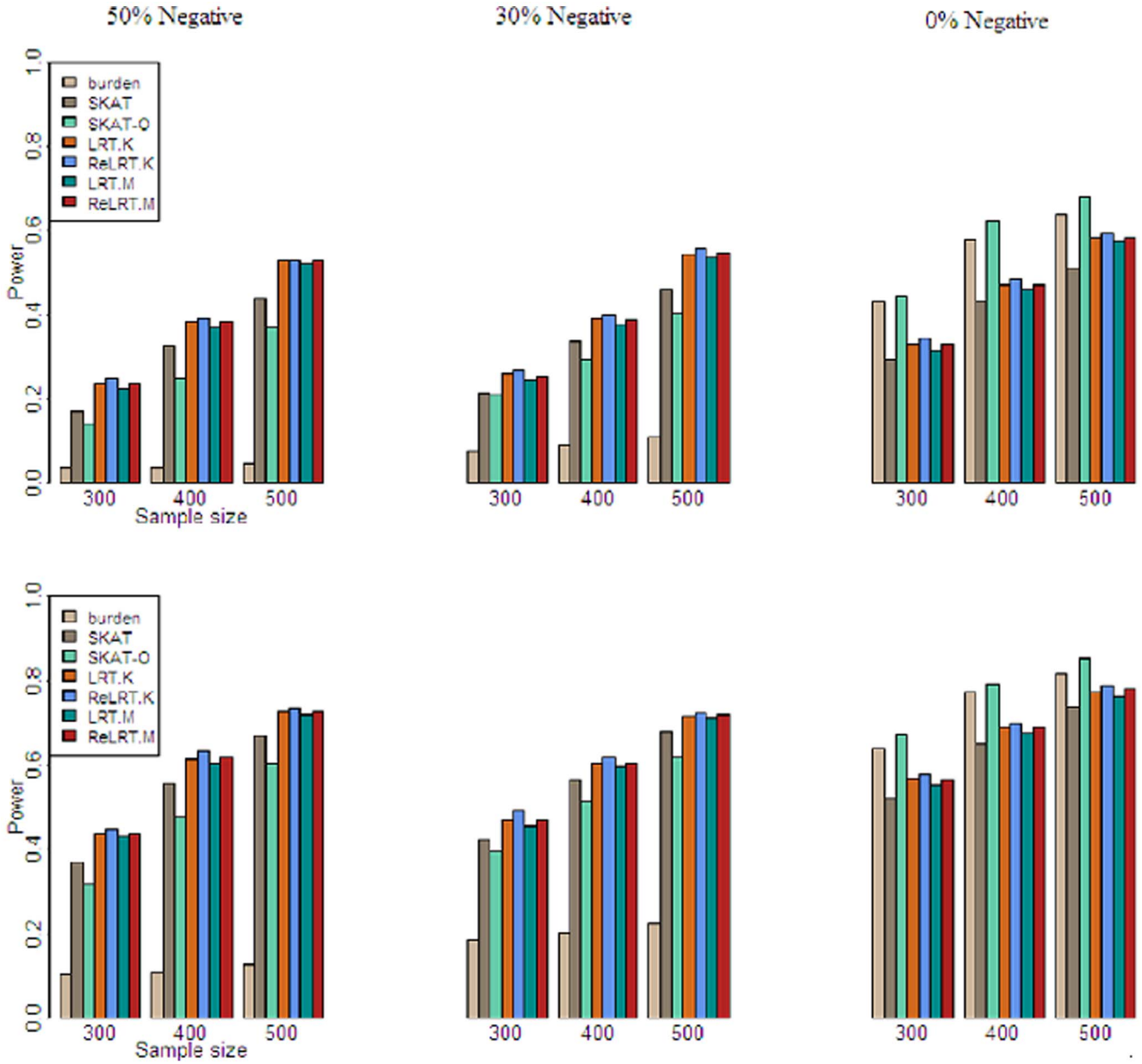

Figure 1. Estimated power for all the tests. The top panel is for $\alpha = 0.01$ and the bottom panel is for $\alpha = 0.05$. M% negative means that in these associated SNPs M% have effects $-0.3|\log_{10}\text{MAF}|$ and the rest (100-M)% are $0.3|\log_{10}\text{MAF}|$.
doi:10.1371/journal.pone.0093355.g001

Table 2. Estimated Type I error.

n	Burden	SKAT	SKAT-O	LRT.K	ReLRT.K	LRT.M	ReLRT.M
		$\alpha = 0.05$					
300	0.051	0.042	0.043	0.047	0.052	0.041	0.046
400	0.060	0.044	0.055	0.050	0.056	0.048	0.051
500	0.058	0.043	0.054	0.042	0.046	0.040	0.042
		$\alpha = 0.01$					
300	0.010	0.008	0.008	0.011	0.012	0.010	0.010
400	0.012	0.008	0.012	0.010	0.012	0.010	0.010
500	0.010	0.008	0.010	0.010	0.010	0.009	0.010

doi:10.1371/journal.pone.0093355.t002

Table 3. Losses of the power for $\alpha = 0.05$[&].

n	burden	SKAT	SKAT-O	LRT.K	ReLRT.K	LRT.M	ReLRT.M
		30% Negative[#]					
300	0.456	0.100	0.280	0.098	0.087	0.099	0.096
400	0.574	0.089	0.279	0.086	0.079	0.081	0.084
500	0.593	0.059	0.233	0.058	0.065	0.052	0.063
Average[$]	0.541	0.083	0.264	0.081	0.077	0.077	0.081
		50% Negative[#]					
300	0.537	0.151	0.356	0.128	0.132	0.124	0.127
400	0.667	0.096	0.313	0.073	0.066	0.073	0.071
500	0.692	0.068	0.247	0.050	0.052	0.043	0.053
Average[$]	0.632	0.105	0.305	0.084	0.083	0.080	0.084

[&]: The values are differences of power between the situation with none of the causal variants (i.e., 0%) being negative and the situation with 30% or 50% causal variants being negative.
[#]: It means that 30% or 50% causal variants are negatively related to phenotype with effects $-0.3|\log_{10}\text{MAF}|$ and the rest 70% or 50% are positively related to phenotype with effects $0.3|\log_{10}\text{MAF}|$.
[$]: The average is calculated across sample sizes.
doi:10.1371/journal.pone.0093355.t003

powers. **Table**s 3 and 4 present the losses of power for the situation that 30% or 50% causal rare variants have negative effects compared to the situation that none of the causal rare variants has negative effects. These values are obtained according to **Figure** 1. The average values in **Table**s 3 and 4 are calculated across sample sizes.

Some important observations from **Figure** 1, **Table**s 3 and 4 are listed as follows.

(I) When all the causal rare variants have the same direction of effect, the burden test and the SKAT-O are the most powerful, following by the LRT, ReLRT, and SKAT. These results are expected since both the burden test and the SKAT-O are designed especially for this situation.

(II) When both positive and negative effects are present, all the tests suffer from power decrease. Under this situation, the LRT and ReLRT have the highest powers, and the burden test suffers from the most reduction of power. For example, when $\alpha = 0.05$, $n = 500$ and all causal rare variants are in the same direction, for the burden test its power is 0.817, while its power decreases to 0.224 when 30% causal rare variants have negative effects and 0.125 when 50% causal rare variants have negative effects. The SKAT-O is no longer optimal and has a smaller power compared to the SKAT, suggesting that in practice using the SKAT rather than the SKAT-O may be safer since the situation that positive and negative effects occur simultaneously is more frequent than the situation that all the effects are in the same direction. Compared with the SKAT, the LRT and ReLRT reduce fewer powers, implying these two tests are relatively more robust to the mixture effects of rare variants.

Table 4. Losses of the power for $\alpha = 0.01$[&].

n	burden	SKAT	SKAT-O	LRT.K	ReLRT.K	LRT.M	ReLRT.M
		30% Negative[#]					
300	0.355	0.080	0.233	0.072	0.078	0.068	0.074
400	0.490	0.092	0.330	0.081	0.086	0.083	0.083
500	0.530	0.052	0.279	0.039	0.038	0.038	0.036
Average[$]	0.458	0.075	0.281	0.064	0.067	0.063	0.064
		50% Negative[#]					
300	0.393	0.124	0.301	0.094	0.094	0.090	0.091
400	0.545	0.105	0.373	0.089	0.092	0.090	0.087
500	0.592	0.074	0.310	0.054	0.065	0.054	0.056
Average[$]	0.510	0.101	0.328	0.079	0.084	0.078	0.078

[&]: The values are differences of power between the situation with none of the causal variants (i.e., 0%) being negative and the situation with 30% or 50% causal variants being negative.
[#]: It means that 30% or 50% causal variants are negatively related to phenotype with effects $-0.3|\log_{10}\text{MAF}|$ and the rest 70% or 50% are positively related to phenotype with effects $0.3|\log_{10}\text{MAF}|$.
[$]: The average is calculated across sample sizes.
doi:10.1371/journal.pone.0093355.t004

Table 5. Characteristics of the used GAW17 data[#].

Gene	Chr	Total	Rare	Causal	MAF	Causal Effects
HIF3A	19	21	15	3	7.17×10^{-3}~0.385	0.174668, 0.51468, 0.265181
FLT1	13	35	25	11	7.17×10^{-3}~0.291	0.18047, 0.457361, 0.732566, 0.839669, 0.38582, 0.549816, 0.623466, 0.653351, 0.59670, 0.549214, 0.090586
KDR	4	16	14	10	7.17×10^{-3}~0.165	0.598271, 0.715613, 0.503025, 1.17194, 0.149975, 0.610938, 0.318125, 0.312058, 1.171940, 0.417977

[#]: Chr indicates the chromosome, Total indicates the total number of SNPs contained in the gene, and Rare indicates the number of rare SNPs within the gene.
doi:10.1371/journal.pone.0093355.t005

(III) It can be seen that the LRT and ReLRT consistently outperform the SKAT regardless of the sample size and the proportion of the negative causal rare variants.

(IV) The ReLRT always has a higher power than the LRT, which may stem from the fact that the ReLRT gives the unbiased estimator of variance component.

(V) The LRT.K versus LRT.M and the ReLRT.K versus ReLRT.M behave comparably, but it is interesting that the ReLRT.K has a slightly larger power than the ReLRT.M, and the LRT.K also has a slightly larger power than the LRT.M.

Application

We apply these methods to the unrelated samples of the GAW17 [33,34]. The GAW17 data contains 24,487 SNPs across 3,205 autosomal genes on 697 individuals, three covariates (age, sex and smoke), three quantitative traits (Q1, Q2 and Q4), and a binary trait. Most of the SNPs are rare with MAF ranging from 0.07% to 25.8%, 74% have MAF less than 0.01 and 12.8% have MAF more than 0.05. This data was widely used on Genetic Analysis Workshop 17 to evaluate the newly developed methods for rare variant detection and compare to the existing ones.

Here we choose the quantitative trait Q1, and select the SNPs within genes *HIF3A*, *FLT1* and *KDR*. These selected genes are rather typical for our comparison of the methods. For *HIF3A*, 20% SNPs are causal rare variants with weak effects. For *FLT1*, 44% SNPs are causal rare variants with moderate effects. For *KDR*, 71.4% SNPs are causal rare variants with relatively strong effects. The characteristics of the selected data are depicted in **Table** 5. More detailed information regarding GAW17 data can be found in Almasy et al. [34].

We use the weighted linear kernel and define the rare variant as those with MAF less than 0.01, so the SNPs with MAF greater than such cut point are not included in the analysis. The results are listed in **Table** 6. The two types of LRT and ReLRT lead to the same results; to save space only one type is reported.

Some interesting results are observed form **Table** 6.

(I) Since all the causal rare SNPs within each gene are positively related to the phenotype Q1 [34], the burden test and SKAT-O have the smallest p values compared to other methods. The LRT and ReLRT obtain smaller p values than SKAT, and the ReLRT always has smaller p values compared to LRT.

(II) Due to the weak effects and small proportion of rare variants, the *HIF3A* cannot be discovered by all the methods; while the *FLT1* and *KDR* are successfully detected. But here it is noted that the p value of SKAT (1.29×10^{-3}) for *KDR* is much larger than those of LRT and ReLRT (with scale of 10^{-5}).

(III) The burden test, SKAT, and SKAT-O cannot give any evidence regarding the effect of the gene. For instance, *FLT1* and *KDR* can be viewed as moderate and strong signals, respectively, but instead the former has a much smaller p value than the latter. This may show a mistaken impression that the *FLT1* is more associated with the phenotype. Fortunately, the estimates of λ provided by LRT and ReLRT display the distinction, that is, the value of λ for *KDR* is larger than that for *FLT1*. From **Table** 6, it can be seen that the estimates of λ correctly reveal the effect strength of different genes. Here the result empirically documents that the LRT and ReLRT are preferred to the SKAT when comparing the contributions of various genes based on a set of rare variants.

Discussion

In this paper we have proposed the LRT and ReLRT to detect the rare variants associated with complex phenotypes from both the standard mixed effects model framework and the kernel

Table 6. Results of the used GAW17 data.

			p value			λ	
Gene	Burden	SKAT	SKAT-O	LRT	ReLRT	LRT	ReLRT
HIF3A	0.262	0.483	0.420	0.388	0.387	<0.001	<0.001
FLT1	6.12×10^{-8}	9.01×10^{-7}	1.03×10^{-9}	6.28×10^{-7}	5.44×10^{-7}	0.750	0.748
KDR	9.27×10^{-7}	1.29×10^{-3}	2.78×10^{-6}	4.99×10^{-5}	4.83×10^{-5}	1.778	1.767

doi:10.1371/journal.pone.0093355.t006

machine learning context. In the latter, the original space of genotypes is mapped to another higher dimensional space by the kernel function. Such a space may be potentially infinite dimensional and is referred to as a feature space in the machine learning literature where the model can proceed linearly [55,57,66]. An important advantage of kernel methods is that we do not have to construct the feature space explicitly since all the analyses can be finished directly over the kernel [66]. In fact the kernel function itself is frequently more efficient to compute than the map function or the inner product induced in H_K [55,57].

By using the representer theorem, the connection between the kernel machine learning and the mixed effects model is well established from the Bayesian point of view. This connection provides a convenient way to examine the rare variants under the context of kernel methods using the LRT and ReLRT. We can find that the kernel is actually the covariance structure for the random effects h, so it can be thought to be the prior correlation among subjects.

Our simulations have demonstrated that the performance of the LRT and ReLRT in the two contexts is comparable. However, it can be expected that the methods of LRT.K and ReLRT.K should be more flexible and attractive although only the linear kernel function is employed in the paper; but even then the LRT.K and ReLRT.K have displayed slightly larger powers than the LRT.M and ReLRT.M. Extending the proposed LRT and ReLRT to other kernel functions needs no any additional efforts, but more applications in practice are required to further understand the behaviors of various kernels. The choice of a kernel function is dependent on which feature space is used to approximate h [30,31]. Liu et al. [31] showed that in a simulation example the Gaussian kernel performed the best compared to other competing kernels.

In the paper, the exact finite sample distributions of LRT and ReLRT obtained by simulation are employed. One may attempt to use the 50:50 mixture distribution of χ_0^2 and χ_1^2 [44–46], where χ_0^2 is a point probability mass at zero and χ_1^2 is a chi-square distribution with 1 degree of freedom. However, it has been displayed that this mixture distribution is conservative [37,47]. It is obvious that the application of the exact finite sample distribution improves the powers of LRT and ReLRT. In addition, the LRT and ReLRT are required to estimate both the null and alternative models. By doing this more information especially from the rare variants is incorporated into the tests, accordingly the powers increase.

Our simulations have also demonstrated that the LRT and ReLRT (including LRT.K and ReLRT.K, and LRT.M and ReLRT.M) outperform the SKAT regardless the sample size and the proportion of negative effects of rare variants. Consequently our results here offer some empirical evidence that the LRT and ReLRT may be preferable to the score test (i.e., the SKAT) in the case of finite sample where the parameter of interest is constrained on the boundary. See also Kuo [48] and Verbeke and Molenberghs [67].

In this paper there are some other aspects concerning the kernel machine learning in rare variant detection that is warranted to be explored. For example, how to choose an optimal kernel function for real life sequencing data [31,68], how to select substantially important random effects (i.e., the true subset of rare variants associated with the phenotype) in a kernel function [69], and what are the exact finite sample distributions of the LRT and ReLRT if incorporating tuning parameters into the kernel function as done in Mallick et al. [58] and Liu et al. [31]. These problems are certainly interesting topics for further investigations.

Acknowledgments

We thank Stephen Schaffner and Michael Wu for their helpful suggestions about the simulation of rare variants. We are grateful to two reviewers for their comments and suggestions which substantially improve the manuscript and thanks also go to the editor and our academic editor for their support.

Author Contributions

Conceived and designed the experiments: PZ SH FC. Performed the experiments: PZ YZ. Analyzed the data: PZ LZ. Contributed reagents/materials/analysis tools: PZ LZ. Wrote the paper: PZ FC YZ.

References

1. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773–785.
2. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11: 415–425.
3. Pritchard JK (2001) Are rare variants responsible for susceptibility to common diseases? Am J Hum Genet 69: 124–137.
4. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, et al. (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet 44: 1326–1329.
5. Peng G, Fan Y, Palculict TB, Shen P, Ruteshouser EC, et al. (2013) Rare variant detection using family-based sequencing analysis. Proc Natl Acad Sci U S A 110: 3985–3990.
6. Smith G, Murray H, Brennan SO (2013) Identification of a rare variant haemoglobin (Hb Sinai-Baltimore) causing spuriously low haemoglobin A1c values on ion exchange chromatography. Ann Clin Biochem 50: 83–86.
7. Zuo L, Zhang X, Deng HW, Luo X (2013) Association of rare PTP4A1-PHF3-EYS variants with alcohol dependence. J Hum Genet 58: 178–179.
8. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513–516.
9. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70–79.
10. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. Science 324: 387–389.
11. Bowes J, Lawrence R, Eyre S, Panoutsopoulou K, Orozco G, et al. (2010) Rare variation at the TNFAIP3 locus and susceptibility to rheumatoid arthritis. Hum Genet 128: 627–633.
12. Feng T, Zhu X (2010) Genome-wide searching of rare genetic variants in WTCCC data. Hum Genet 128: 269–280.
13. Masson E, Chen JM, Scotet V, Le Marechal C, Ferec C (2008) Association of rare chymotrypsinogen C (CTRC) gene variations in patients with idiopathic chronic pancreatitis. Hum Genet 123: 83–91.
14. Yu L, Wynn J, Cheung Y, Shen Y, Mychaliska G, et al. (2013) Variants in GATA4 are a rare cause of familial and sporadic congenital diaphragmatic hernia. Hum Genet 132: 285–292.
15. Bacanu S-A, Nelson MR, Whittaker JC (2012) Comparison of Statistical Tests for Association between Rare Variants and Binary Traits. PLoS ONE 7: e42530.
16. Zhan H, Xu S (2012) Adaptive Ridge Regression for Rare Variant Detection. PLoS ONE 7: e44173.
17. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695–701.
18. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet 89: 82–93.
19. Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annu Rev Genet 44: 293–308.
20. Liu DJ, Leal SM (2010) A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. PLoS Genet 6: e1001156.
21. Li B, Leal SM (2009) Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. PLoS Genet 5: e1000481.
22. Madsen BE, Browning SR (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet 5: e1000384.

23. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutat Res 615: 28–56.
24. Basu S, Pan W (2011) Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35: 606–619.
25. Han F, Pan W (2010) A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70: 42–54.
26. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C (2011) A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease. PLoS Genet 7: e1001289.
27. Lin X (1997) Variance component testing in generalised linear models with random effects. Biometrika 84: 309–326.
28. Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 963–974.
29. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP (2008) A Powerful and Flexible Multilocus Association Test for Quantitative Traits. Am J Hum Genet 82: 386–397.
30. Liu D, Ghosh D, Lin X (2008) Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9: 292.
31. Liu D, Lin X, Ghosh D (2007) Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics 63: 1079–1088.
32. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, et al. (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. Am J Hum Genet 86: 929–942.
33. Ghosh S, Bickeboller H, Bailey J, Bailey-Wilson J, Cantor R, et al. (2011) Identifying rare variants from exome scans: the GAW17 experience. BMC Proc 5: S1.
34. Almasy L, Dyer T, Peralta J, Kent J, Charlesworth J, et al. (2011) Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc 5: S2.
35. Verbeke G, Molenberghs G (2009) Linear mixed models for longitudinal data. New York: Springer.
36. R Core Team (2013) R: A language and environment for statistica computing. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/.
37. Greven S, Crainiceanu CM, Küchenhoff H, Peters A (2008) Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. J Comput Graph Stat 17: 870–891.
38. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
39. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–775.
40. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, et al. (2012) Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet 91: 224–237.
41. Davies RB (1980) Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables. J R Stat Soc Series C 29: 323–333.
42. Liu H, Tang Y, Zhang HH (2009) A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal 53: 853–856.
43. Duchesne P, Lafaye De Micheaux P (2010) Computing the distribution of quadratic forms: further comparisons between the Liu-tang-zhang approximation and exact methods. Comput Stat Data Anal 54: 858–862.
44. Self SG, Liang KY (1987) Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. J Am Stat Assoc 82: 605–610.
45. Stram DO, Lee JW (1994) Variance Components Testing in the Longitudinal Mixed Effects Model. Biometrics 50: 1171–1177.
46. Liang KY, Self SG (1996) On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. J R Stat Soc Series B 58: 785–796.
47. Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc Series B 66: 165–185.
48. Kuo BS (1999) Asymptotics of ML estimator for regression models with a stochastic trend component. Econom Theor 15: 24–49.
49. Claeskens G (2004) Restricted likelihood ratio lack-of-fit tests using mixed spline models. J R Stat Soc Series B 66: 909–926.
50. Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.
51. Harville DA (1974) Bayesian inference for variance components using only error contrasts. Biometrika 61: 383–385.
52. Corbeil RR, Searle SR (1976) Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. Technometrics 18: 31–38.
53. Harville DA (1977) Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. J Am Stat Assoc 72: 320–338.
54. Gianola D, van Kaam JB (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289–2303.
55. Schölkopf B, Smola A (2001) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: The MIT Press.
56. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. New York: Cambridge University Press.
57. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. New York: Cambridge University Press.
58. Mallick BK, Ghosh D, Ghosh M (2005) Bayesian classification of tumours by using gene expression data. J R Stat Soc Series B 67: 219–234.
59. Schaid DJ (2010) Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 70: 109–131.
60. Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. J Math Anal Appl 33: 82–95.
61. Wahba G (1990) Spline models for observational data. SIAM: Society for Industrial and Applied Mathematics.
62. Schaid DJ (2010) Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Hum Hered 70: 132–140.
63. Wahba G (1999) Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In: Schölkopf B, Burges CJC, Smola AJ, editors. Advances in Kernel Methods-Support Vector Learning. The MIT Press, pp. 69–87.
64. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
65. Scheipl F, Greven S, Küchenhoff H (2008) Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Comput Stat Data Anal 52: 3283–3299.
66. Hofmann T, Schölkopf B, Smola AJ (2008) Kernel methods in machine learning. Ann Statist 36: 1171–1220.
67. Verbeke G, Molenberghs G (2003) The Use of Score Tests for Inference on Variance Components. Biometrics 59: 254–262.
68. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, et al. (2013) Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. Genet Epidemiol 37: 267–275.
69. Chen Z, Dunson DB (2003) Random effects selection in linear mixed models. Biometrics 59: 762–769.