Quantitative Biology 2017, 5(4): 368–379
DOI 10.1007/s40484-017-0124-3

RESEARCH NOTE

On the use of kernel machines for Mendelian randomization

Weiming Zhang and Debashis Ghosh*

Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
* Correspondence: debashis.ghosh@ucdenver.edu

Received March 29, 2017; Revised June 3, 2017; Accepted June 5, 2017

***Background*: Properly adjusting for unmeasured confounders is critical for health studies in order to achieve valid testing and estimation of the exposure's causal effect on outcomes. The instrumental variable (IV) method has long been used in econometrics to estimate causal effects while accommodating the effect of unmeasured confounders. Mendelian randomization (MR), which uses genetic variants as the instrumental variables, is an application of the instrumental variable method to biomedical research fields, and has become popular in recent years. One often-used estimator of causal effects for instrumental variables and Mendelian randomization is the two-stage least square estimator (TSLS). The validity of TSLS relies on the accurate prediction of exposure based on IVs in its first stage.**
***Results*: In this note, we propose to model the link between exposure and genetic IVs using the least-squares kernel machine (LSKM). Some simulation studies are used to evaluate the feasibility of LSKM in TSLS setting.**
***Conclusions*: Our results show that LSKM based on genotype score or genotype can be used effectively in TSLS. It may provide higher power when the association between exposure and genetic IVs is nonlinear.**

Keywords: Mendelian randomization; kernel machine; instrumental variable; unmeasured confounder; casual inference

INTRODUCTION

In many epidemiological studies, investigators are often interested in understanding the relationship between environmental exposures or treatments and disease-related outcome variables in order to develop insights about disease etiology as well as to provide guidance for disease prevention. In such studies, a crucial factor to deal with is confounding, defined as the presence of variables that "confounds" the association between the exposure of interest with the outcome. One way to eliminate confounding is through a randomized controlled trial, in which the exposure of interest can be randomized. However, in the majority of epidemiological studies, this type of randomization is practically infeasible, due either to logistical or ethical reasons. Thus, the validity of the results from these studies rely on both a careful design to measure all potential confounders as well as statistical methods to properly control for the collected confounders in analyses. There are statistical methods for dealing with measured confounders, such as the propensity score [1–3] and the *E*-estimation [4] among others. However, it is difficult in general to control for unmeasured confounders.

The method of instrumental variables (IV) was developed by Wright [5] for simultaneous equations estimation. It has long been used in econometrics to estimate the causal effect of treatment while accommodating the effect of unmeasured confounders. The basic idea of IV is to extract unconfounded variation from the treatment variable using the association between IV and treatment, and to then use this extracted information to estimate and test the causal effect of treatment on outcomes.

Mendelian randomization (MR) is an application of IV methods using genetic data. The concept was first described by Dutch scientist Martjin Katan, although he did not use the term Mendelian randomization [6]. He pointed out that the different alleles of gene apolipoprotein E (apo E) were major determinants of plasma cholesterol levels in several populations, and the alleles were not affected by confounders and reverse causation from cancer since they were inherited from parents and had not changed since birth. Therefore, these genes could potentially be used to investigate the relationship between low-serum cholesterol and cancer. This example demonstrates the advantage of genetic variants as IVs. Genetic variants generally are not associated with potential social, economic, behavioral and environmental confounders. If they are associated with the exposures, but not directly associated with the outcomes, they will be valid instrumental variables for studying the link between exposures and the outcomes. However, instrumental variables have their weaknesses as well. A genetic variant can be correlated with an outcome through a causal variant due to linkage disequilibrium. A genetic variant also could have effect on multiple biological pathways, one of which may be causal to outcomes. These will invalidate variants as IV and complicate MR design and analyses.

There has been a greater use of MR methods in current biomedical research. One reason has been due to the accumulation of a large amount of genetic data from GWAS studies. Another is the advancement in new biotechnology leading to the emerging of the new types of data that benefits from MR design. For example, scientists have recently started to interrogate DNA methylation sites to see if DNA methylation works as a mediator to link between certain exposures and the outcomes. Unlike a person's genetic makeup, the DNA methylation in a person is determined by the complex interaction of genetic and environmental factors during the person's life. Factors such as smoking [7,8] and diet [9] have been shown to affect both global methylation and gene-specific methylation. Health conditions such as obesity can also change the DNA methylation [10], which raise the question of reverse causation when those health conditions are outcomes. New study designs have been proposed [11] to incorporate MR into study and analysis in order to hurdle the confounders and reverse causation problem.

One important step in MR studies is to accurately estimate the effect of genetic variants on the trait or exposure. Many new statistical methods have been proposed in the last decade for genetic association study to model the relationship between trait and genetic variants, which are usually single nucleotide polymorphism (SNP). Research on IV approaches has been more limited. We highlight two recent proposals. Lin *et al.* [12] developed two-stage regularization methods for high-dimensional IV regression. In its first stage, the exposures are regressed on potential IVs, and effects of optimal IVs are identified and estimated through a sparsity-inducing penalty function. In the second stage, the outcome is regressed on the first-stage prediction while variable selection is again performed through a sparsity-inducing penalty function. Kang *et al.* [13] proposed using regularization methods to handle the problem of invalid IVs. Their method, sisVIVE, applies the penalty procedure only in the first stage and estimates the causal effect of exposure on outcome when the proportion of invalid IVs is no higher than 50% while without knowing which IVs are invalid. The goal of the current paper is to evaluate the feasibility of least-squares kernel machines (LSKM) in MR studies. The LSKM is a semi-parametric kernel based method; we summarize the details of the approach in the Section of Least-Squares Kernel Machine. There have been non-parametric and kernel-based procedures for IV methods [14,15]. They estimated the non-parametric relationship between outcome and exposure in the presence of IVs. In this article, we focused on LSKM to model the link between exposure and IVs in order to achieve better estimate of exposure. The paper proceeds as follows. In the section of Background and Two-Stage Least Squares Estimation, we review the two-stage least squares approach to instrumental variables estimation. Least-squares kernel machines are reviewed in the Section of Least-Squares Kernel Machine; simulation studies evaluating the proposed approaches are given in Section of Simulation Studies. Section of Discussion concludes with some discussion.

BACKGROUND AND TWO-STAGE LEAST SQUARES ESTIMATION

We denote Y as a continuous outcome variable, X as a continuous exposure variable, U as the unmeasured variable that correlates with both Y and X, and Z as the instrumental variable. Assuming our data are on individuals indexed by $i = 1, \dots, n$, we usually regress Y on X to estimate the association between X and Y as in Equation (1).

$$y_i = \beta_0 + \beta_x x_i + \varepsilon_{y_i}. \tag{1}$$

Because of the presence of the unmeasured confounder U, the estimated effect β_x is biased in this simple regression. Another way to see is that the effect U is embodied in ε_{y_i} so that the error term in [11] is statistically correlated with x_i. In order to estimate the true effect of X on Y in a manner that is not confounded by U, the IV method uses the instrumental variable Z to extract the variation from X that is not affected by U. Then the confounder-free variation of X is used to estimate the effect of X on Y. This approach has three critical assumptions:

(i). The instrumental variable Z is associated with the exposure X.

(ii). The instrumental variable Z is independent of the unmeasured confounder U.

(iii). The instrumental variable affects outcome Y only through X, i.e., given X and U, Z is independent of Y. This is also called the no direct effect assumption.

Intuitively, the first assumption says that we can extract information about the variability of X using the instrumental variable Z, while the second assumption ensures that the extracted variability of X is free of U. The last assumption ensures that the estimated effects using the extracted information of X only comes from X.

The two-stage least squares is one of the most commonly used methods to estimate β_x using instrumental variables when the outcome is continuous. At the first stage we conduct a regression with model:

$$x_i=\alpha_0+\alpha_z z_i+\varepsilon_{x_i}, \tag{2}$$

where z_i is a vector of IVs. We obtain the fitted values $\hat{x}_i=\hat{\alpha}_0+\hat{\alpha}_z z_i$. At the second stage, we fit the model [1] using the fitted value of x_i:

$$y_i=\beta_0+\beta_x\hat{x}_i+\varepsilon_{y_i}. \tag{3}$$

In order to obtain the correct estimate of β_x, the following assumptions should hold at both stages: (i) independent and normally distributed error with homoscedastic variance; (ii) a linear relationship between X and Z; (iii) a linear relationship between Y and X. The first-stage regression essentially extracts the unmeasured-confounder-free information of exposure using instrumental variables. The second stage regression then uses this information to identify and estimate the association between outcome and exposure.

The least squares estimate of the first-stage linear regression coefficients is

$$\hat{\alpha}=(Z'Z)^{-1}Z'X.$$

The fitted values are

$$\hat{X}=Z(Z'Z)^{-1}Z'X.$$

The least squares estimate of the second-stage linear regression coefficients is given by

$$\hat{\beta}=\left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y.$$

Substituting $\hat{X}$ with the first-stage results, we have the TSLS estimates of β:

$$\hat{\beta}=(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y.$$

Both Z and X may include the constant term if the intercept terms are in the models. The TSLS estimator has been shown to be consistent and asymptotically normal distributed even in the presence of heteroscedasticity [16]. The direct estimate of the variance-covariance of $\hat{\beta}$ from model [33] is incorrect because it does not take the variability of $\hat{x}_i$ into account. The correct estimator when the errors are homoscedastic is

$$\hat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X)^{-1},$$

where $\hat{\sigma}^2$ is the residual variance and estimated using the observed value of exposure X, i.e., $\hat{\sigma}^2=\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)/(n-p)$, where n is the sample size and p is the number of estimated parameters in the second stage regression. Let $P_z=Z(Z'Z)^{-1}Z'$, we can see $P_z'P_z=P_z$. Therefore, $X'Z(Z'Z)^{-1}Z'X = X'P_zX = X'P_z'P_zX = (P_zX)'P_zX=(Z(Z'Z)^{-1}Z'X)'\left(Z(Z'Z)^{-1}Z'X=\hat{X}'\hat{X}\right.$. Hence, the correct estimator when the errors are homoscedastic can also be written as $\hat{\sigma}^2\left(\hat{X}'\hat{X}\right)^{-1}=\left[\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)/(n-p)\right]\left(\hat{X}'\hat{X}\right)^{-1}$.

If we do the two-stage least squares regressions manually without using any built-in special TSLS packages, the usual residual variance after the second stage regression from a statistical software is $\left[\left(Y-\hat{X}\hat{\beta}\right)'\left(Y-\hat{X}\hat{\beta}\right)/(n-p)\right]\left(\hat{X}'\hat{X}\right)^{-1}$ since $\hat{X}$ is used as the predictors in the second regression. In order to correct the variance estimates, we can multiply a factor

$$\frac{\left(Y-X\hat{\beta}\right)'\left(Y-X\hat{\beta}\right)/(n-p)}{\left(Y-\hat{X}\hat{\beta}\right)'\left(Y-\hat{X}\hat{\beta}\right)/(n-p)}$$

to the output variance $\left[\left(Y-\hat{X}\hat{\beta}\right)'\left(Y-\hat{X}\hat{\beta}\right)/(n-p)\right]\left(\hat{X}'\hat{X}\right)^{-1}$ which yields valid inference.

LEAST-SQUARES KERNEL MACHINES

In order to apply the two-stage least squares method successfully, we need to accurately estimate the exposure X in the first-stage regression. Just as with any regression problem, it may not be straightforward to model the relationship between exposure and the instrumental variables. For example, their relationship may be non-linear, and it may be difficult to identify the correct function form for their connections. Therefore, more flexible modeling methods are needed for the first-stage regression. Kernel machines are non-parametric methods that model non-linear or linear relations without specifying a rigid function form. We propose to use a semi-parametric kernel based method, least-squares kernel machine, in the first stage to obtain accurate fitted values when linear regressions cannot.

Kernel machines represent a class of methods that have been used in machine learning and have been recently studied in the biostatistical literature. They are based on the concept of a kernel function, defined as a symmetric positive definite function that provides a measure of similarity between pairs of observations. For example, the dot product of two vectors is a kernel function. The kernel function introduces non-linearity at the original space without specifying a functional form by implicitly transforming data to a new high-dimensional feature space. Although the kernel function is defined as the inner product in the new space, it works with the original data and avoids having to find the mapping to the complicated new feature space. Thus the kernel function provides both computational efficiency and flexibility in modeling.

The least-squares kernel machines (LSKM) was proposed by [17] to fit a semi-parametric model for genetic pathway data. It fits the model:

$$y_i = w_i' \beta + h(v_i) + \varepsilon_i, \tag{4}$$

where w_i is a vector of covariates including the constant term, and is a vector of gene expression measurements. $h(\cdot)$ is a centered smooth function. Under some regularity conditions, a kernel function $K(\cdot,\cdot)$ implicitly generate a unique function space spanned by a particular set of orthogonal basis functions for h such that h can be represented by the linear combination of the orthogonal basis functions. Equivalently, h can be represented using the kernel as

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot,\ v_i)$$

where α is a vector of unknown parameters and K the kernel function measuring the similarity between the a pair of subjects evaluated at their observed gene expression values. Liu *et al*. demonstrated the equivalency of LSKM with a certain linear mixed effects model, making parameter estimation accessible from commonly used statistical software such as R and SAS. LSKM provides a framework for a variance component-based score test for testing joint association between a set of explanatory variables and the outcome. The LSKM method was quickly adopted for genetic association study. The widely used sequence kernel association test (SKAT) for rare variant testing [17–19] and its extension including an optimal combination of SKAT tests [20,21], combining SKAT with common variant tests [22,23] and application of SKAT to meta-analysis of GWAS data [23] are all based on LSKM. It has been shown that LSKM variance component-based score test is related to the U statistics and some other proposed genetic association test statistics [24].

So far the typical applications of LSKM in genetics have all involved the variance component score test that tests the null hypothesis of no association between variants and traits such that without the need for estimating function $h(\cdot)$ in the Equation (4). However, in the MR setting, we will need to estimate the fitted values of the exposure using genetic IVs when doing TSLS. We propose replacing Equation (2) by the LSKM model

$$x_i = \alpha_0 + h(z_i) + \varepsilon_{X_i}. \tag{5}$$

Then using the predicted value from the linear mixed model $\hat{x}_i = \hat{\alpha_0} + \hat{h(z_i)}$ in fitting model 3 for estimating and testing the effect of exposure. This mixed model can be fitted using standard statistical software such as R. Since there are no statistical packages that incorporate the LSKM in the two-sage least squares method, we need to manually perform the two regressions. We propose to use the method described at the end of the Section of Background and Two-Stage Least Squares Estimation to correct estimates for variance and covariance, i.e., multiplying the factor

$$\frac{\left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)/(n-p)}{\left(Y - \hat{X}\hat{\beta}\right)' \left(Y - \hat{X}\hat{\beta}\right)/(n-p)}$$

to the output variance from the second regression. We provide a sample code in the appendix of this paper.

SIMULATION STUDIES

We conducted several simulation studies to evaluate the properties of testing and estimating the exposure causal effect when using LSKM in the first stage of TSLS. We organized our study designs based on the sample size, validity of the simulated instrumental variables, the instruments' effect size on the exposure and the form of their relationship with the exposure. The relationships between the outcome and the exposure were always simulated as linear.

Simulation methods

We simulated 1,000 datasets for each study. In all our studies described in this paper, we used 0.05 as the nominal type I error rate. Across all simulation studies, we simulated a set of either 5 or 6 independent SNPs that had the same minor allele frequency (MAF = 0.1) as the IVs. We chose to simulate multiple IVs because researchers often find more than one SNP to be associated with the exposure, and a previous study has shown that including multiple IVs in MR could reduce the variance of TSLS estimator [25]. However, the same study also showed the finite-sample bias of TSLS estimator with multiple IVs. Therefore, we did the first simulation study to evaluate the effect of small sample size on kernel based methods. The instruments were all valid and had the same effect on the exposure in this study.

1) Small sample, valid instruments with equal-sized linear effects

We simulated 500 independent subjects in the first simulation study. We simulated each subject's exposure using equation:

$$x_i = \left(\sum_{j=1}^{6} \alpha_j g_{ij}\right) + \alpha_u u_i + \varepsilon_{x_i}, \tag{6}$$

where $g_{i1} - g_{i6}$ are the genetic variables coded using the person's genotype in 6 SNPs. We used the dominant model for each SNP. We defined u_i to be the unmeasured confounder, drawn from a Normal (0, 1) distribution. The independent error term ε_{x_i} in Equation (6) was also drawn from a Normal (0,1) distribution. The α_{1-6} and α_u are predetermined coefficients with α_{1-6} set as 0.4 and α_u set as 1. Then we simulate the outcome as

$$y_i = \beta_x x_i + \beta_u u_i + \varepsilon_y, \tag{7}$$

where x_i and u_i were from the previous step, and the independent error ε_y was drawn from Normal(0,1). The causal effect β_x was 0 for type I error study and 0.5 for power study; the effect of the confounder β_u was 1. We analyzed simulated data with 6 different methods. For the traditional TSLS, we used either all 6 genetic variables or a summary genotype score as the IV in the model 2. The genotype score is defined as the unweighted sum of all genetic variables. For kernel based methods, we used Gaussian kernel with either genotype or genotype score, linear kernel with genotype and quadratic kernel with genotype. The Gaussian kernel is defined as $K(i,j) = e^{-d_{ij}/\rho^2}$, where d_{ij} is the Euclidian distance between ith and jth subjects calculated from either all 6 genotypes or the single genotype score, and ρ is a parameter that needs to be estimated using the data. The orthogonal basis functions for the Gaussian kernel-generated function space is the class of radial basis functions. The Gaussian kernel incorporates nonlinear relationships into the model. The linear kernel with genotype between ith and jth subjects is defined as $K(i,j) = g_i' g_j$, where g is a vector of six genetic variables. The linear kernel assumes the association between genetic variants and trait/exposure has a linear form. The quadratic kernel with genotype between ith and jth subjects is defined as $K(i,j) = (g_i' g_j + \rho)^2$, where ρ is a tuning parameter. We set ρ to be one in our studies. The orthogonal basis functions of quadratic kernel generated function space include main effects, the quadratic main effects and all two way interactions between SNPs.

2) Small sample, valid instruments with unequal-sized effects and small interaction between IVs

In order to investigate further the small samples with more complicated instruments' effect on exposure, we introduced unequal-sized effects of instruments and small interactions between the instruments in our second simulation study. We again simulated 500 subjects and used the dominant model for each SNP, but varied the effect of the different IVs on the exposure and included interactions between IVs. Our simulation equation for exposure is

$$x_i = \left(\sum_{j=1}^{6} \alpha_j g_{ij}\right) + \alpha_{12} g_{i1} g_{i2} + \alpha_{34} g_{i3} g_{i4}$$

$$+ \alpha_{56} g_{i5} g_{i6} + \alpha_u u_i + \varepsilon_{x_i}. \tag{8}$$

The coefficients were set as following: $\alpha_1 = 0.2, \alpha_2 = 0.4, \alpha_3 = 0.25, \alpha_4 = 0.3, \alpha_5 = 0.4, \alpha_6 = 0.1, \alpha_{12} = 0.05, \alpha_{34} = -0.2, \alpha_{56} = 0.2, \alpha_u = 1$. The simulation equation for outcome and its parameters remained the same as Equation (7) and its parameter settings in the first study. We analyzed the data using TSLS with genotype or genotype score. For kernel based methods, we focused on TSLS with genotype score Gaussian kernel.

3) Small sample, valid instruments with unequal-sized effects and larger interaction between IVs

Next, we increased the size of interactions between IVs in our third study. We used the dominant model and kept the Equation (8) but increased the size of interaction relative to the main effect to look at how a more extreme situation can affect each methods. The parameters were $\alpha_1 = 0, \alpha_2 = 0.2, \alpha_3 = 0.05, \alpha_4 = 0.1, \alpha_5 = 0, \alpha_6 = 0, \alpha_{12} = 0.5, \alpha_{34} = -0.5, \alpha_{56} = 0.6, \alpha_u = 1$. We analyzed the data using the genotype score linear kernel and genotype score quadratic kernel in addition to the three methods used in the last section, i.e., TSLS genotype, TSLS genotype score and the genotype score Gaussian kernel.

4) Larger sample, equal-sized linear effects without interaction, and possible irrelevant IVs

In this study, we increased the sample size to 2,500 subject per dataset and used an additive model for each SNP. We varied the number of IVs in different simulations and divided IVs as true IVs and irrelevant IVs. The effect of true IVs were all 0.4. The effect of irrelevant IVs were set at 0 although they were still independent of unmeasured confounder and independent of outcome given the exposure status, i.e., they satisfied the (ii) and (iii) assumptions of IVs described in the Section of Background and Two-Stage Least Squares Estimation, but had no correlation with exposure. We compared TSLS genotype, TSLS genotype score, genotype linear kernel and genotype score linear kernel in our analyses.

Simulation results

In the scenario of small sample, valid instruments with equal-sized linear effects (first simulation study), Table 1 shows that the type I error was correct for genotype scores at the nominal level of 5%, but inflated for the genotype. This is consistent with previous published results. The multiple SNPs divide the sample into multiple cells. For example, we were using 6 SNPs coded with a dominant model. The sample can be divided into $2^6 = 64$ cells based on the 6-loci genotype. When the sample size is not large, the small subsample size in each cell leads to a higher chance of imbalance in the unmeasured confounder leading to invalidation of IVs. This may be the reason for the finite-sample bias in MR [26]. The genotype score collapses the data into fewer cells. For 6 SNPs with a dominant model, the possible value of genotype score are 0 to 6, i.e., only 7 subgroups compares to the 64 subgroups from using genotypes directly. The Figure 1 shows the cell means of the simulated unmeasured confounder in small samples and large samples. We can see that the cell means of the unmeasured confounder in the sub-genetic groups of small sample have wider spread and more extreme values comparing to the large sample

Table 1. Type I error rate and power of small sample, valid instruments with equal-sized linear effects.

Methods	Hypothesis	Type I error rate or power[a]	Mean of causal effect estimates and its 95% confidence interval	Standard deviation of causal effect estimates
TSLS genotype	Null	8.9%	0.0598 (0.0504, 0.0692)	0.1517
TSLS genotype score	Null	4.8%	−0.0101 (−0.0210, 0.0008)	0.1761
Gaussian kernel TSLS genotype	Null	18.4%	0.1386 (0.1272, 0.1500)	0.1844
Gaussian kernel TSLS genotype score	Null	5.6%	0.0051 (−0.0062, 0.0164)	0.1817
Linear kernel TSLS genotype	Null	9.1%	0.0657 (0.0541, 0.0773)	0.1871
Quadratic kernel TSLS genotype	Null	17.8%	0.1382 (0.1266, 0.1498)	0.1871
TSLS genotype score	Alternative	76.4%	0.4815 (0.4709, 0.4921)	0.1703
Gaussian kernel TSLS genotype score	Alternative	80.4%	0.5197 (0.5091, 0.5303)	0.1703

a: type I error rate for null hypothesis and power for alternative hypothesis.

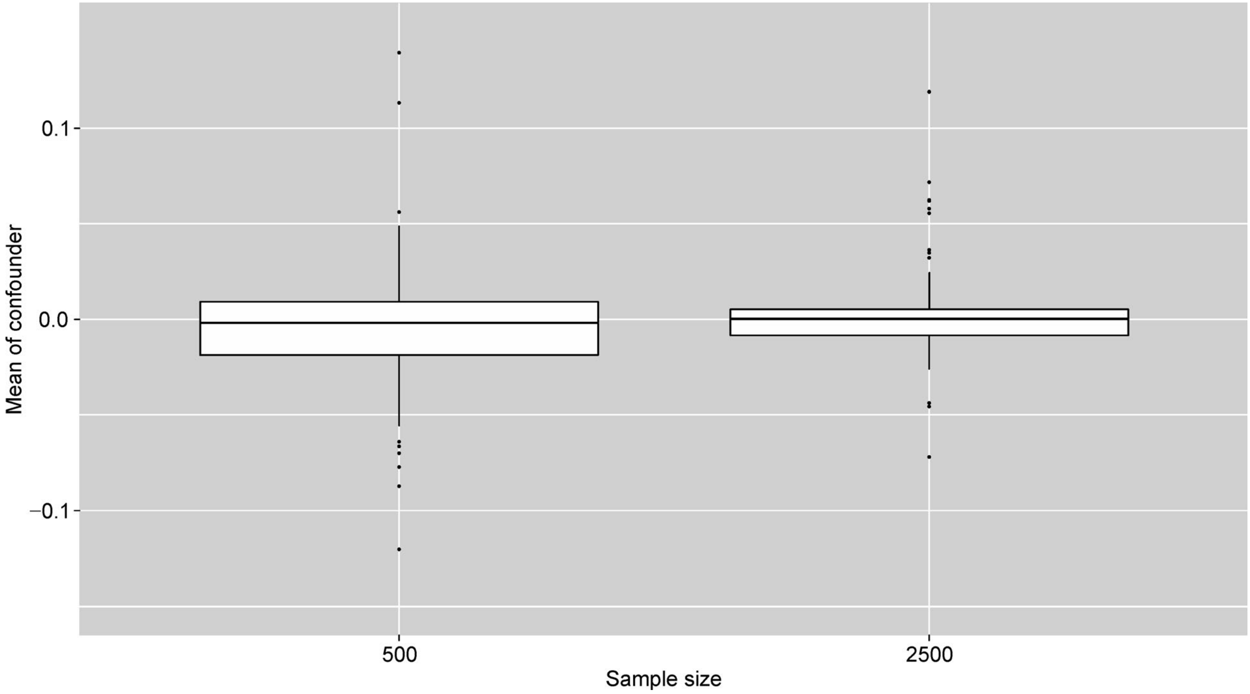

Figure 1. Cell mean of confounder. Mean of unmeasured confounder in each sub genotype group that was created by 6 SNPs whose MAF were all 0.1.

due to the sparsity of the cells that carry the minor alleles, thereby creating a spurious association between genetic IVs and the unmeasured confounder. The minor allele frequency also contributes to this bias (Table 2). When the sample size is fixed, the lower MAF of IVs causes a higher inflated type I error rate.

Table 1 indicates that the LSKM with genotype score also has the correct type I error rate and slightly higher power than using genotype score in linear regression while the genotype-based kernel methods have inflated type I error and the biased estimate of causal effect. The causal estimates from using genotype score was slightly biased downward.

For the small sample, valid instruments with unequal-sized linear effects and small interactions between IVs (simulation study 2), the Gaussian kernel had the best type I error and causal estimates under null hypothesis due to its non-linearity handling the interaction effects. The genotype-based approach again had inflated type I error rate and biased estimates (Table 3). By contrast, approaches using genotype score was conservative. The Gaussian kernel with genotype score had higher power than the genotype score (Table 3).

The quadratic kernel based on genotype score was the clear winner under the null hypothesis when there were larger interactions between IVs in small sample, valid instruments with unequal-sized effects (simulation study 3, Table 4). It also had better power than just using genotype score in the model. The genotype score was defined as

$$genotype\ score = SNP1 + SNP2 + SNP3 + SNP4$$

Table 2. The effects of minor allele frequency and sample size (N) on the finite sample bias of MR.

MAF	Type I error rate				
	N = 500	N = 1000	N = 1500	N = 2000	N = 2500
0.5	0.066	0.073	0.061	0.069	0.051
0.4	0.068	0.066	0.045	0.059	0.047
0.3	0.062	0.064	0.053	0.056	0.054
0.2	0.076	0.069	0.06	0.062	0.06
0.1	0.089	0.077	0.066	0.066	0.053
0.05	0.108	0.078	0.081	0.069	0.053
0.01	0.188	0.145	0.134	0.113	0.098

Table 3. Type I error and power of small sample, valid instruments with unequal-sized effects and small interaction between IVs.

Methods	Hypothesis	Type I error rate or power[a]	Mean of causal effect estimates and its 95% confidence interval	Standard deviation of causal effect estimates
TSLS genotype	Null	9.1%	0.0892 (0.0774, 0.1010)	0.1897
TSLS genotype score	Null	3.4%	−0.0384 (−0.0549, −0.0219)	0.2665
Gaussian kernel TSLS genotype score	Null	4.8%	−0.0040 (−0.0237, 0.0157)	0.3178
TSLS genotype score	Alternative	52.9%	0.4616 (0.4451, 0.4781)	0.2665
Gaussian kernel TSLS genotype score	Alternative	62.4%	0.5484 (0.5298, 0.5670)	0.3000

a: type I error rate for null hypothesis and power for alternative hypothesis.

Table 4. Type I error and power of small sample, valid instruments with unequal-sized effects and larger interaction between IVs.

Methods	Hypothesis	Type I error rate or power[a]	Mean of causal effect estimates and its 95% confidence interval	Standard deviation of causal effect estimates
TSLS genotype	Null	19.2%	0.2502 (0.2317, 0.2687)	0.2983
TSLS genotype score	Null	4.3%	0.1594 (−0.1223, 0.4411)	4.5453
Gaussian kernel TSLS genotype score	Null	7.1%	0.4693 (0.4105, 0.5281)	0.9487
Linear kernel TSLS genotype score	Null	4.1%	0.3134 (−0.0823, 0.7091)	6.3844
Quadratic kernel TSLS genotype score	Null	5.0%	0.0940 (−0.1600, 0.3480)	4.0976
TSLS genotype score	Alternative	32.6%	0.6594 (0.3777, 0.9411)	4.5453
Linear kernel TSLS genotype score	Alternative	33.9%	1.2500 (0.7509, 1.7491)	8.0529
Quadratic kernel TSLS genotype score	Alternative	37.1%	1.2700 (0.8841, 1.6559)	6.2266

a: type I error rate for null hypothesis and power for alternative hypothesis.

$$+SNP5+SNP6,$$

and the basis functions of the space generated by the quadratic kernel include genotype score and its quadratic. Thus, the quadratic kernel models main effect of each SNP, its quadratic and all interactions between two SNPs. It models the association between exposure and the variants closest to the simulated scenario.

Our results in Table 5 show that the genotype and genotype linear kernel have reduced finite-sample bias when all IVs are valid and the sample size is large. Liu *et al*. demonstrated that the kernel methods generally performed well when the majority of covariates input to the kernel functions were relevant with only a small number of irrelevant covariates, although some parameter estimates and the estimates of the function *h* exhibited slightly worse performance [17]. The irrelevant covariates in the LSKM paper corresponds to the situation of irrelevant IVs in our studies, which are the extreme example of weak IVs. Bound *et al*. [27] pointed out that weak IVs could lead to inconsistency of the IV causal estimator and the bias of the IV estimator could reach the level of the ordinary least squares estimator when the IVs are not correlated with exposure at all. They demonstrated that the relative inconsistency of IV relative to ordinary least squares (OLS) is $(\rho_{z,\varepsilon}/\rho_{x,\varepsilon})/\rho_{x,z}$, where X is the variable of interest, Z is the IV and ε is the residual error in the regression of outcome on variable X, which may

Table 5. Type I error rate of larger sample, equal-sized linear effects without interaction and possible irrelevant IVs

Methods	Number of (true IVs, irrelevant IVs)	Type I error rate	Mean of causal effect estimates and its 95% confidence interval	Standard deviation of causal effect estimates
Linear kernel from genotype	(5, 0)	4.7%	0.0076 (0.0029, 0.0123)	0.0762
Genotype		4.7%	0.0075 (0.0029, 0.0121)	0.0735
Genotype score		4.6%	−0.0037 (−0.0084, 0.001)	0.0755
Linear kernel from genotype score		4.6%	−0.0038 (−0.0085, 0.001)	0.0768
Linear kernel from genotype	(1, 9)	13.7%	0.1198 (0.1067, 0.1329)	0.2119
Genotype		13.6%	0.0974 (0.0884, 0.1064)	0.1456
Genotype score		3.2%	−0.4478 (−0.7589, −0.1367)	5.0195
Linear kernel from genotype score		3.1%	−0.1809 (−0.4248, 0.063)	3.9353
Linear kernel from genotype	(5, 5)	6.8%	0.0209 (0.0162, 0.0256)	0.0762
Genotype		6.7%	0.0201 (0.0156, 0.0246)	0.0721
Genotype score		4.5%	−0.0103 (−0.0171, −0.0035)	0.1095
Linear kernel from genotype score		4.4%	−0.0106 (−0.0177, −0.0035)	0.1140
Linear kernel from genotype	(10, 0)	6.0%	0.0091 (0.0058, 0.0124)	0.0539
Genotype		5.9%	0.0089 (0.0056, 0.0122)	0.0529
Genotype score		5.6%	−0.0036 (−0.007, −2e-04)	0.0548
Linear kernel from genotype score		5.6%	−0.0036 (−0.007, −2e-04)	0.0548
Linear kernel from genotype	(10, 5)	5.0%	0.017 (0.0137, 0.0203)	0.0529
Genotype		5.0%	0.016 (0.0128, 0.0192)	0.0510
Genotype score		5.2%	−0.0043 (−0.0083, −3e-04)	0.0640
Linear kernel from genotype score		5.2%	−0.0044 (−0.0084, −4e-04)	0.0640
Linear kernel from genotype	(15, 0)	5.7%	0.0106 (0.008, 0.0132)	0.0424
Genotype		5.7%	0.0104 (0.0078, 0.013)	0.0374
Genotype score		5.2%	−0.0024 (−0.005, 2e-04)	0.0424
Linear kernel from genotype score		5.2%	−0.0024 (−0.005, 2e-04)	0.0424
Linear kernel from genotype	(1, 19)	32.5%	0.3117 (0.2787, 0.3447)	0.5326
Genotype		31.0%	0.1780 (0.1702, 0.1858)	0.1253
Genotype score		2.8%	0.1486 (−1.0843, 1.3815)	19.8890
Linear kernel from genotype score		2.9%	0.2965 (−0.5759, 1.1689)	13.5082
Linear kernel from genotype	(5, 15)	12.2%	0.0529 (0.0482, 0.0576)	0.0755
Genotype		11.8%	0.0480 (0.0438, 0.0522)	0.0678
Genotype score		3.6%	−0.0068 (−0.016, 0.0024)	0.1483
Linear kernel from genotype score		3.6%	−0.0076 (−0.017, 0.0018)	0.1517

(*Continued*)

Methods	Number of (true IVs, irrelevant IVs)	Type I error rate	Mean of causal effect estimates and its 95% confidence interval	Standard deviation of causal effect estimates
Linear kernel from genotype	(10, 10)	8.5%	0.0276 (0.0244, 0.0308)	0.0510
Genotype		8.4%	0.0262 (0.0232, 0.0292)	0.0480
Genotype score		4.0%	−0.00068 (−0.0051, 0.0038)	0.0721
Linear kernel from genotype score		4.0%	−0.00072 (−0.0052, 0.0037)	0.0721
Linear kernel from genotype	(15, 5)	6.4%	0.0192 (0.0166, 0.0218)	0.0412
Genotype		6.4%	0.0186 (0.0161, 0.0211)	0.0400
Genotype score		3.8%	0.0001 (−0.0028, 0.0031)	0.0480
Linear kernel from genotype score		3.8%	0.0001 (−0.0029, 0.0031)	0.0480
Linear kernel from genotype	(20, 0)	5.7%	0.0138 (0.0116, 0.0160)	0.0361
Genotype		5.6%	0.0134 (0.0113, 0.0155)	0.0346
Genotype score		3.8%	0.0003 (−0.0019, 0.0025)	0.0361
Linear kernel from genotype score		3.8%	0.0003 (−0.0019, 0.0025)	0.0361

include unmeasured confounder, ρ_{ij} is the correlation between i, j. When the correlation between X and Z is very small, any correlation between Z and ε will be magnified. So when the IVs are weak or even irrelevant, any chance correlation between IVs and the unmeasured confounder can yield unstable estimates. Consistent with their findings, our results in Table 5 indicate that the genotype-based methods suffer the most when the majority of IVs were irrelevant. They had high inflated type I error and biased IV estimates, especially when both the number of IVs and the proportion or irrelevant IVs were large. However, in the MR setting, the genotype score only requires the SNP components to meet the criteria (ii) and (iii) describe in the section of Background and Two-Stage Least Squares Estimation as long as some SNP components are associated with exposure [26]. The genotype score-based methods had more deflated type I error when the proportion of irrelevant IVs increased. Its 95% confidence intervals were wider due to the increased variance of the parameter estimates, and the deviation between point estimates of the causal parameter and the true value increased with an increasing proportion of irrelevant IVs.

DISCUSSION

Mendelian randomization has been adopted as an important tool to combat unmeasured confounding in health-related studies. In this paper, we studied the application of the semi-parametric LSKM method in two-stage least squares (TSLS) procedures for Mendelian Randomization studies. We found that LSKM can be used effectively in the first stage of TSLS to estimate the exposure using genetic IVs. The flexibility of the kernel-based method in modeling the link between exposure and genetic variants was seen in our results. LSKM does not require a rigid functional form for modeling. It can model non-linear relationships between variants and exposure using different kernels, such as the identity by state (IBS) kernel, the Gaussian kernel and the polynomial kernel, among others. When there were complex interactions from genetic variants determining the exposure, kernel-based methods demonstrated the best power while keeping the nominal type I error rate.

Using multiple SNPs as IVs could potentially increase the power of MR studies. However, this practice also creates some problems for researchers. First, it can cause the finite-sample bias of the TSLS estimator through extensive subdivision of subjects into multi-loci genotypic subgroups. Our results indicate that kernels that directly use genotypes suffers from this pitfall as much as using genotypes directly in the model [2]. However, when the sample size increases, the bias subsides. In our case, when the sample size became 2500, even when the number of IVs was 20, the type I error rate of kernel-based method was still near the nominal levels as well as having very small bias in causal effect estimates. Also, kernels using genotype score were as robust as genotype score with respect to finite-sample bias while LSKM still have the advantage of modeling flexibility. Therefore, we suggest that genotype score should be preferred when using kernel methods with small samples. Second, using multiple genetic IVs can increase the chance for including irrelevant IVs. Our results shows that irrelevant IVs affect both genotype and genotype score based methods although using genotypes from multiple SNPs directly is worse, also both traditional methods and LSKM are affected similarly. A reviewer kindly pointed out the recent development by Wang *et al.* [28], in which haplotypes are used as instrumental variables as a way to combine multiple SNPs. Haplotypes carry the linkage disequilibrium information such that it may potentially increase the power of analyses, but it requires all the SNPs in the same region. However, the LSKM does not have such limitation. Wang *et al*. also observed an inflated type I error rate when the irrelevant SNPs were present. When they adopted a step-wise method to identify an optimal set of haplotypes by merging the haplotype subgroups, the type I error rates were close to the nominal level. The reviewer suggested methods such as training and testing, cross-validation for selecting valid IVs in LSKM setting. This idea is beyond the scope of the current paper and will be investigated in future work.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation under Grant (No. NSF ABI 1457935) and the National Institutes of Health under Grant (No. R01 GM117946).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Weiming Zhang and Debashis Ghosh declare they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

APPENDIX

Sample code:

```
library(EMMREML)
library(sem)

set.seed(2)
n < - 2500 #number of subjects
ng  < - 10 #number of SNPs
P  < - 0.1 #Minor allele frequency

a1  < - rep(0.4, ng) # effects of IV on X
a2  < - 1 # confounder’s effect on X
b1  < - 0.5 #causal effect
b2  < - 1 # confounder’s effect on outcome

ga  < - matrix(rbinom(n*ng, 1, P), nrow = n, ncol = ng)
gb  < - matrix(rbinom(n*ng, 1, P), nrow = n, ncol = ng)

gen < - ga + gb #additive model
#ga[1:3,]
#gb[1:3,]
#gen[1:3,]

u  < - rnorm(n, 0, 1) #unmeasured confounder
ex  < - rnorm(n, 0, 1) #error term in exposure x
ey  < - rnorm(n, 0, 1) #error term in outcome y
x  < - as.matrix(gen) %*% a1 + a2*u + ex #exposure
y  < - b1*x + b2*u + ey #outcome
gscore  < - rowSums(gen) #Genetic score
K  < - gscore %*% t(gscore) #linear kernel matrix
Z  < - diag(1, n, n) #random effect

#first stage, mixed model
emmfit  < - emmreml(x, matrix(rep(1, n)), Z, K,varbetahat = FALSE,varuhat = FALSE, PEVuhat = FALSE, test = FALSE)

fixed  < - emmfit$betahat[1,1]

#fitted values from first stage fixed effect + random effect
pred  < - emmfit$uhat + fixed

#second stage
fit2  < - lm(y ~ pred)

summary(fit2)

betas  < - coefficients(fit2)
df.resid  < - df.residual(fit2)

#sum of squre from wrong residuals
mse.wrong < -  sum(residuals(fit2)^2)/df.resid

#incorrect residual variance directly from 2nd regression
vcov.wrong = vcov(fit2)

#residual using observed exposure
resid.correct  < - y - betas[1] - betas[2] * x

#sum of square from correct residuals
mse.correct  < - sum(resid.correct^2) / df.resid

#correctd variance covariance
vcov.correct  < - (mse.correct / mse.wrong) * vcov.wrong

betas[2] #causal effect estimates

pval  < - 2*(1-pt(abs(betas[2]/sqrt(vcov.correct[2,2])), df.resid))

#####Using genetic score without kernel machine, manually correcting.

#First stage
fit1  < - lm(x ~ gscore)

#Second stage
fit2  < - lm(y ~ fit1$fitted)
summary(fit2)

betas  < - coefficients(fit2)
df.resid  < - df.residual(fit2)
```

```
#sum of squre from wrong residuals
mse.wrong < -  sum(residuals(fit2)^2)/df.resid

#incorrect residual variance directly from 2nd regression
vcov.wrong = vcov(fit2)

#residual using observed exposure
resid.correct  < - y - betas[1] - betas[2] * x

#sum of square from correct residuals
mse.correct  < - sum(resid.correct^2) / df.resid

#correctd variance covariance
vcov.correct  < - (mse.correct / mse.wrong) * vcov.wrong

betas[2] #causal effect estimates

pval  < - 2*(1-pt(abs(betas[2]/sqrt(vcov.correct[2,2])), df.resid))

#######Using genetic score without kernel machine, using tsls
function to crrect.
fit  < - tsls(y ~ x, ~ gscore, w = rep(1,length(x)))
summary(fit)
```

REFERENCES

1. Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. Biometrika. 70, 41–55
2. Rosenbaum, P. R. and Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc., 79, 516–524
3. Rosenbaum, P. R. and Rubin, D. B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am. Stat. 39, 33–38
4. Robins, J. M., Mark, S. D. and Newey, W. K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics, 48, 479–495
5. Wright, P. G. (1928) The Tariff on Animal and Vegetable Oils. New York: The Macmillan company
6. Katan, M. B. (2004) Apolipoprotein E isoforms, serum cholesterol, and cancer. Int. J. Epidemiol., 33, 9
7. Hillemacher, T., Frieling, H., Moskau, S., Muschler, M. A., Semmler, A., Kornhuber, J., Klockgether, T., Bleich, S. and Linnebank, M. (2008) Global DNA methylation is influenced by smoking behaviour. Eur. Neuropsychopharmacol., 18, 295–298
8. Bouwland-Both, M. I., van Mil, N. H., Tolhoek, C. P., Stolk, L., Eilers, P. H., Verbiest, M. M., Heijmans, B. T., Uitterlinden, A. G., Hofman, A., van Ijzendoorn, M. H., *et al.* (2015) Prenatal parental tobacco smoking, gene specific DNA methylation, and newborns size: the Generation R study. Clin. Epigenetics, 7, 83
9. Crider, K. S., Yang, T. P., Berry, R. J. and Bailey, L. B. (2012) Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. Adv. Nutr., 3, 21–38
10. Geach, T. (2017) Obesity: methylation a consequence not a cause. Nat. Rev. Endocrinol., 13, 127
11. Relton, C. L. and Davey Smith, G. (2012) Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int. J. Epidemiol., 41, 161–176
12. Lin, W., Feng, R., Li, H. (2015) Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. J. Am. Stat. Assoc., 110, 270–288
13. Kang, H., Zhang, A., Cai, T. and Small, D. (2016) Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. J. Am. Stat. Assoc., 111, 132–144
14. Hall, P., Horowitz, J. (2005) Nonparametric methods for inference in the presence of instrumental variables. Ann. Stat., 33, 2904–2929
15. Laurain, V., Toth, R., Piga, D. and Zheng, W. (2015) An instrumental least squares support vector machine for nonlinear system identification. Automatica. 54, 340–347
16. White, H. (1982) Instrumental variables regression with independent observations. Econometrica, 50, 483–99
17. Liu, D., Lin, X. and Ghosh, D. (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics, 63, 1079–1088
18. Kwee, L. C., Liu, D., Lin, X., Ghosh, D. and Epstein, M. P. (2008) A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet., 82, 386–397
19. Wu, M. C., Lee, S., Cai, T., Li,Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., 89, 82–93
20. Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M. Lin, X., and the NHLBI GO Exome Sequencing Project—ESP Lung Project Team. (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
21. Lee, S., Wu, M. C. and Lin, X. (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13, 762–775
22. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. and Lin, X. (2013) Sequence kernel association tests for the combined effect of rare and common variants. Am. J. Hum. Genet., 92, 841–853
23. Lee, S., Teslovich, T. M., Boehnke, M. and Lin, X. (2013) General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet., 93, 42–53
24. Zhang, W., Epstein, M. P., Fingerlin, T. E. and Ghosh, D. (2017) Links between the sequence kernel association and the kernel-based adaptive cluster tests. Stat. Biosci., 9, 246–258
25. Burgess, S. and Thompson, S. G. (2011) Bias in causal estimates from Mendelian randomization studies with weak instruments. Stat. Med., 30, 1312–1323
26. Burgess, S. and Thompson, S. G. (2015) Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation.

Boca Raton: CRC Press

27. Bound J., Jaeger, D. and Baker, R. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J. Am. Stat. Assoc., 90, 443–450
28. Wang, F., Meyer, N. J., Walley, K. R., Russell, J. A. and Feng, R. (2016) Causal genetic inference using haplotypes as instrumental variables. Genet. Epidemiol., 40, 35–44