Published in final edited form as:

Biometrics. 2018 June ; 74(2): 764–766. doi:10.1111/biom.12785.

A Note on Testing and Estimation in Marker-Set Association Study Using Semiparametric Quantile Regression Kernel Machine

Xiang Zhan^{1,2,*} and Michael C. Wu^{1,**}

¹Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

²Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, U.S.A.

Summary

Kong et al. (2016, *Biometrics* **72**, 364–371) presented a quantile regression kernel machine (QRKM) test for robust analysis of genetic marker-set association studies. A potential limitation of QRKM is the permutation-based test design may be unscalable for the massive sizes of modern datasets. In this article, we present an alternative strategy for p-value calculation of QRKM, which is capable of speeding up the QRKM testing procedure dramatically while maintaining the same testing performance as QRKM. The effectiveness of our approach is demonstrated via simulation studies.

Keywords

Fast permutation test; Genetic marker-set association; Kernel machines; Quantile regression

1. Introduction

In an excellent article on testing of marker-set association analysis (Kong et al., 2016), the authors developed a semiparametric quantile regression kernel machine (QRKM) test to examine the association between the conditional quantiles of a response variable and a group of genetic markers. In particular, the QRKM framework models the effect of covariates on a given conditional quantile of the response parametrically and the genetic marker-set effect on the conditional quantile of the response nonparametrically. Then a score-type test statistic is derived from the subgradient of the check function and the null distribution of this score-type test statistic is obtained via permutations. There are some potential computational limitations of the QRKM test. First, many permutations are needed when a stringent p-value is required, as in many genome-wide association studies. Second, the permutation-based test

* xzhan@fhcrc.org. ** mcwu@fhcrc.org.

Supplementary Materials

Web Appendix referenced in Sections 3, which includes additional numerical results; and R codes implementing FQRKM, are available with this article at the *Biometrics* website on Wiley Online Library.

design is not scalable to big datasets, in which sample sizes are in the thousands, that are frequently encountered in modern statistical sciences. The purpose of this article is to present a computationally-motivated fast QRKM testing strategy that can address the computational limitations.

2. Methods

2.1 Review of QRKM

Let $(y, \mathbf{X}, \mathbf{Z})$ denote the triplet of response, genetic markers and covariates. The following quantile regression model was considered to relate the response to genetic markers and covariates:

$$Q_y(\tau|\mathbf{X}, \mathbf{Z}) = f(\mathbf{X}) + \mathbf{Z}\boldsymbol{\beta} + \beta_0 + Q_\varepsilon(\tau), \quad (1)$$

where $f(\cdot)$ is some centered unknown function quantifying effect of the marker-set, $\boldsymbol{\beta}$ is the regression coefficient for the covariates, and $Q_\varepsilon(\tau)$ is the τ -th quantile of ε for $\tau \in (0, 1)$. By assuming f in some function space H_K spanned by a certain kernel machine, the QRKM fits the quantile regression model (1) via the minimization problem

$$\min_{\beta_0, \boldsymbol{\beta}, f \in H_K} \sum_{i=1}^n \rho_\tau(y_i - f(\mathbf{X}_i) - \mathbf{Z}_i'\boldsymbol{\beta} - \beta_0) + P_\lambda(f), \quad (2)$$

where $\rho_\tau(t) = t(\tau - I_t)$, I is the indicator function, and $P_\lambda(f)$ is a penalty function which regulates the smoothness of $f(\cdot)$. Then both an estimating and a testing procedure for $\boldsymbol{\beta}$ and $f(\cdot)$ was proposed (Kong et al., 2016), while we focus on the testing part in this article.

The joint effect of the marker-set is tested as $H_0: f(\cdot) = 0$ via a score-type statistic $T = \mathbf{w}'\mathbf{K}\mathbf{w}$, where $\mathbf{w} = (w_1, \dots, w_n)'$ are binary random variables taking values τ or $\tau - 1$ depending on whether the residuals of the null quantile regression model $Q_y(\tau|\mathbf{Z}) = \mathbf{Z}\boldsymbol{\beta} + \beta_0 + Q_\varepsilon(\tau)$ are positive or not, and $\mathbf{K}$ is an $n \times n$ kernel matrix calculated from the marker-set. Due to the non-smooth check loss, the null distribution of $T = \mathbf{w}'\mathbf{K}\mathbf{w}$ is obtained via permutation. In particular, QRKM first fits model (2) for the full data $\{(y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \dots, n\}$ and obtains the full model residuals. Next, a permutation of the full model residuals are used to generate pseudo response y_i^* . Using the pseudo response, null quantile regression model $(y_i^*, \mathbf{Z}_i)$ is fitted and used to calculate $\mathbf{w}^*$ and $T^* = \mathbf{w}^{*'}\mathbf{K}\mathbf{w}^*$. The permutation procedure is repeated N times to calculate $\{T_1^*, \dots, T_N^*\}$. Finally, the p-value is obtained by comparing the observed QRKM statistic $T = \mathbf{w}'\mathbf{K}\mathbf{w}$ to the multiple permuted statistics $\{T_1^*, \dots, T_N^*\}$.

2.2 Fast Permutation

The computational burden of the original QRKM test comes from the following aspects: 1) fitting the full QRKM model once; 2) fitting null quantile regression model (y_i^*, Z_i) for N times; and 3) calculating permuted statistics T^* for N times. To reduce the computational cost by omitting the need for explicit permutations, we present the fast QRKM testing procedure in the following.

According to the dual representation of Kong et al. (2016), $f(X) = \lambda^{-1} \sum_{i=1}^n \theta_i K(X, X_i)$, where $K(\cdot, \cdot)$ represents the kernel function. Since $f(\cdot)$ is assumed to be a centered function in QRKM, hence $K(\cdot, \cdot)$ should be centered. In terms of matrix, we assume $\mathbf{K}$ to be a centered kernel matrix. In other words, we consider a slightly different QRKM statistic $T = \mathbf{w}' \mathbf{H} \mathbf{K} \mathbf{H} \mathbf{w} = \text{tr}(\mathbf{H} \mathbf{K} \mathbf{H} \mathbf{w} \mathbf{w}')$, where $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'/n$ is the n -th order centering matrix and $\text{tr}(\cdot)$ denotes the trace of a matrix. With this centered version statistic, we are able to implement some fast permutation testing strategies (Josse, Pagè, and Husson, 2008) developed for RV-type statistic of the form $\text{tr}(\mathbf{H} \mathbf{A} \mathbf{A}' \mathbf{H} \mathbf{B} \mathbf{B}')$, where $\mathbf{A}$ and $\mathbf{B}$ are two arbitrary input matrices. To see this, we are able to decompose kernel matrix $\mathbf{K}$ as $\mathbf{K} = \Phi \Phi'$ (such a Φ exists since $\mathbf{K}$ is symmetric and positive semi-definite). If we set $\mathbf{A} = \Phi$ and $\mathbf{B} = \mathbf{w}$, then the centered version of QRKM statistic $T = \text{tr}(\mathbf{H} \mathbf{K} \mathbf{H} \mathbf{w} \mathbf{w}')$ reduces to a RV-type statistic, and thus the existing fast permutation RV test designs (Josse et al., 2008) can be directly applied to our new QRKM statistic $T = \text{tr}(\mathbf{H} \mathbf{K} \mathbf{H} \mathbf{w} \mathbf{w}')$.

Specifically, instead of explicitly drawing permutations and recalculating the test statistic, the fast RV permutation strategies approximate the null permutation distribution to a simpler and known distribution. Let $\mathbf{K}^*$ denote a permuted kernel matrix obtained by shuffling rows and columns of $\mathbf{K}$ simultaneously. Then the null empirical distribution of all permutations $\{T_b^{*} = \text{tr}(\mathbf{H} \mathbf{K}_b^* \mathbf{H} \mathbf{w} \mathbf{w}'), b = 1, \dots, n!\}$ is approximated to a simpler distribution by moment matching. Since $\mathbf{w}$ is calculated under the null model without $\mathbf{X}$, shuffling rows and columns of $\mathbf{K}$ (which only depends on $\mathbf{X}$) does not affect the null distribution of $\mathbf{w}$. In other words, T_b^{*} 's have the same null distribution as the test statistic T , and therefore, it is reasonable to calculate the test p-value using the empirical distribution of $\{T_b^{*} = \text{tr}(\mathbf{H} \mathbf{K}_b^* \mathbf{H} \mathbf{w} \mathbf{w}'), b = 1, \dots, n!\}$. To approximate this empirical permutation distribution, several simpler distributions (normal, log-normal, Edgeworth and Pearson type III) have been used (Josse et al., 2008). Among these approximations, Pearson type III distribution provides the most accurate and efficient performance and hence is adopted to approximate the empirical distribution of T_b^{*} 's. Since analytical forms of sample moments of $\{T_b^{*} = \text{tr}(\mathbf{H} \mathbf{K}_b^* \mathbf{H} \mathbf{w} \mathbf{w}'), b = 1, \dots, n!\}$ are available (see Josse et al. (2008) and references therein), closed-form Pearson type III density can be calculated. Finally, based on this closed-form Pearson type III density, fast and analytical p-value calculation is allowed in our new fast QRKM method.

3. Numerical Studies

We follow the same simulation setup and set the number of permutations $N = 50,000$ for QRKM test as in Kong et al. (2016). For ease of presenting, we term our new testing strategy

as fast QRKM (FQRKM) in this section. The type I error is presented in Table 1. Based on the table, FQRKM $\tau = 0.1$ is slightly inflated when $n = 200$. It has been shown that FQRKM $\tau = 0.1$ test tends to have the correct type I error with a larger sample size (see additional numerical studies in Web Appendix A). On the other hand, both FQRKM $\tau = 0.5$ and FQRKM $\tau = 0.8$ have reasonable type I errors. Finally, the average computing time reported in Table 1 shows that, FQRKM is about 6,000 times faster than QRKM under $n = 200$ (the computing time of QRKM and FQRKM was counted after both w and K were calculated from the observed samples).

The power of QRKM and FQRKM under the significance level of $\alpha = 0.05$ are presented in Figure 1. Because the type I error of FQRKM $\tau = 0.1$ might be slightly inflated under $\alpha = 0.05$, and also for ease of presenting, we omitted the power curves with $\tau = 0.1$. Based on Figure 1, it can be seen that the power difference between the QRKM test and the FQRKM test with same quantile is negligible under each scenario. To conclude, FQRKM can serve as a useful QRKM-equivalent test when computational expedience is desired.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIH Grants U10 CA180819 and R01 HG007508 and the Hope Foundation. The authors thank the associated editor and reviewers for thoughtful and constructive comments that have helped improve the article.

References

- Josse J, Pagès J, Husson F. Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis*. 2008; 53:82–91.
- Kong D, Maity A, Hsu FC, Tzeng JY. Testing and estimation in marker–set association study using semiparametric quantile regression kernel machine. *Biometrics*. 2016; 72:364–371. [PubMed: 26575303]

Figure 1.

Power curves of QRKM and FQRKM for quantiles $\tau = 0.5$ and 0.8 based on 500 Monte Carlo replications. The left, middle and right columns corresponds to normal, t and chi-squares error distribution respectively. The first row corresponds to function f_1 , and the second row corresponds to f_2 , where f_1 and f_2 are the same notations specified in Kong et al. 2016. The solid, dashed lines correspond to $\tau = 0.5$ and 0.8 respectively. $\circ$ represents QRKM and $\triangle$ represents FQRKM.

Table 1

Type I errors of QRKM test and FQRKM test with $n = 200$. The simulation is based 10,000 Monte Carlo replicates. The empirical type I error was evaluated under nominal levels $\alpha = 0.001, 0.01$ and 0.05 respectively. Type I errors outside the 95% CI [$\alpha - 1.96\sqrt{\alpha(1-\alpha)/10000}, \alpha + 1.96\sqrt{\alpha(1-\alpha)/10000}$] are typed in bold. The unit of average computing time (ACT) is in seconds.

Distribution	Method	0.001	0.01	0.05	ACT
Normal	QRKM $\tau = 0.1$	0.0012	0.0095	0.0482	293.6
	QRKM $\tau = 0.5$	0.0009	0.0113	0.0489	257.2
	QRKM $\tau = 0.8$	0.0006	0.0084	0.0467	280.5
	FQRKM $\tau = 0.1$	0.0015	0.0132	0.0501	0.051
	FQRKM $\tau = 0.5$	0.0009	0.0100	0.0489	0.051
	FQRKM $\tau = 0.8$	0.0008	0.0095	0.0470	0.051
t	QRKM $\tau = 0.1$	0.0012	0.0092	0.0483	297.4
	QRKM $\tau = 0.5$	0.0007	0.0085	0.0500	262.8
	QRKM $\tau = 0.8$	0.0010	0.0099	0.0497	284.4
	FQRKM $\tau = 0.1$	0.0016	0.0138	0.0537	0.052
	FQRKM $\tau = 0.5$	0.0011	0.0092	0.0499	0.052
	FQRKM $\tau = 0.8$	0.0008	0.0110	0.0524	0.051
$\chi^2_1 - 1$	QRKM $\tau = 0.1$	0.0009	0.0082	0.0461	298.7
	QRKM $\tau = 0.5$	0.0012	0.0106	0.0532	258.3
	QRKM $\tau = 0.8$	0.0012	0.0107	0.0501	280.7
	FQRKM $\tau = 0.1$	0.0012	0.0143	0.0549	0.051
	FQRKM $\tau = 0.5$	0.0011	0.0112	0.0531	0.051
	FQRKM $\tau = 0.8$	0.0011	0.0115	0.0500	0.051