Published in final edited form as:

Biometrics. 2016 June ; 72(2): 364–371. doi:10.1111/biom.12438.

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

Dehan Kong,

Department of Biostatistics, University of North Carolina at Chapel Hill

Arnab Maity,

Department of Statistics, North Carolina State University

Fang-Chi Hsu, and

Department of Biostatistical Sciences, Wake Forest University

Jung-Ying Tzeng

Department of Statistics and Bioinformatics Research Center, North Carolina State University;

Department of Statistics, National Cheng-Kung University, Taiwan

Summary

We consider quantile regression for partially linear models where an outcome of interest is related to covariates and a marker set (e.g., gene or pathway). The covariate effects are modeled parametrically and the marker set effect of multiple loci is modeled using kernel machine. We propose an efficient algorithm to solve the corresponding optimization problem for estimating the effects of covariates and also introduce a powerful test for detecting the overall effect of the marker set. Our test is motivated by traditional score test, and borrows the idea of permutation test. Our estimation and testing procedures are evaluated numerically and applied to assess genetic association of change in fasting homocysteine level using the Vitamin Intervention for Stroke Prevention Trial data.

Keywords

Bootstrap; Genetic marker-set association; Kernel machines; Permutation; Quantile regression; Semiparametric; Smoothing parameter; Testing

1. Introduction

In this paper, we consider the problem of testing for association between a phenotype of interest and a marker-set (e.g. a gene or pathway), where the response values may contain outliers or have a skewed or heavy-tailed distribution, and the number of markers is typically large. We consider a semiparametric model where clinical and demographic covariates are modeled parametrically and the set of genetic variables is modeled jointly in a nonparametric way using statistical kernel machine. Instead of modeling the random error as

a mean zero random variable, as is usually done in least squares regression setting, we consider the case where the (pre-specified) quantile of the errors is assumed to be zero. We are interested in two aspects of the model, namely the estimation of the effects of both the genetic and clinical covariates, and testing for association between the genetic covariates and the response.

As a popular approach to semiparametrically model multi-dimensional covariates in non-parametric and semiparametric models, kernel machine (KM) methods (Vapnik, 1998; Scholkopf and Smola, 2001) have emerged as a powerful tool in association studies. KM methods can accommodate high number of covariates as well as complex relationships easily while allowing to perform hypothesis testing. For the semiparametric models with continuous response, Liu et al. (2007) and Kwee et al. (2008) developed least squares kernel machine (LSKM) regression models and showed its attractive connection with linear mixed models. They also proposed a score based testing procedure for the pathway effect. There has been a growing number of applications and extensions of such models in recent years (Wu et al., 2010, 2011; Maity and Lin, 2011; Monsees et al., 2011; Meyer et al., 2012; Maity et al., 2012).

While LSKM is a powerful regression tool, it has some serious limitations. First, the parameter estimation in LSKM is performed by minimizing a penalized least squares criterion. This criterion can be very sensitive if the data are generated from skewed/heavy tailed distributions. Second, LSKM only models the conditional mean of the response given the covariates. Thus one can only detect association in terms of mean. However, any association between the response and covariates in different quantile levels may not be detected. This issue is particularly important when the covariates only have effect on particular subsets of samples having higher or lower response levels rather than the overall mean. To this end, we propose a semiparametric quantile regression model using the kernel machine framework.

Quantile regression has been widely used in many areas, such as survival analysis (Yang, 1999), microarray study (Wang and He, 2007) and economics (Hendricks and Koenker, 1992). For an extensive review of the quantile regression, we refer to Koenker (2005). Compared to the ordinary least squares estimates, quantile regression, for example median regression, provides estimates that are robust to outliers. Quantile regression reveals how the covariates influence the location, scale and shape of the response distribution, and one can investigate the effect of the covariates onto different quantiles of response, which is more informative than the least squares. There is a rich literature on nonparametric and semiparametric quantile regression. Yu and Jones (1998) used kernel smoothing based local polynomial regression (Fan and Gijbels, 1996) to develop an estimation procedure for nonparametric quantile regression. Such techniques are later generalized for partially linear quantile regression models (Lee, 2003; Sun, 2005; Liang and Li, 2009; Song et al., 2012). Most of these procedures only consider the case where the nonparametrically modeled variable is univariate. While Lee (2003) and Sun (2005) discuss theoretical results for multivariate covariates, it is well known that the performance deteriorates sharply as dimension increases, see e.g., Bickel and Li (2007). Other nonparametric estimation procedures for quantile regression include smoothing splines models (Koenker et al., 1994;

Nychka et al., 1995; Yuan, 2006; Wu and Yu, 2014), elastic and plastic splines (Koenker and Mizera, 2002) for bivariate covariates, and estimation based on total variation regularization (Koenker and Mizera, 2004). These methods work well when the nonparametrically modeled covariate is scalar or low-dimensional. However, for moderate or high dimensional covariates, such methods are often computationally not feasible or very intensive. In addition, none of the above mentioned techniques provide any procedure to perform hypothesis testing for the effect of the nonparametrically modeled covariate. In the context of quantile regression using Reproducing Kernel Hilbert Space approach, Takeuchi et al. (2006) and Li, Liu, and Zhu (2007) considered a purely nonparametric quantile regression model and developed a solution path algorithm with respect to the tuning parameter to estimate the model components. Liu and Wu (2011) considered the simultaneous non-crossing quantile regression for the same model. However they neither develop any testing procedure to test for association nor propose any estimator of standard errors of the resulting estimates. In this article, we will develop estimation procedure for the model components along with their standard errors, and propose a testing procedure to assess association between the genetic covariate and the response.

Our research is motivated by the Vitamin Intervention for Stroke Prevention (VISP) trial (Toole et al., 2004), a multicenter, double-blind, randomized, controlled clinical trial that aimed to study the effect of vitamins on preventing recurrent stroke. Details about the study are provided in Section 5. We are interested in evaluating the effect of 9 candidate genes involved in the Hcy metabolic pathway on the Hcy level (Hsu et al., 2011) as it has been suggested that Hcy level can be used to predict risk of recurrent stroke (Toole et al., 2004; Pettigrew et al., 2008), and genetic variations might be attributed to mild to moderate hyperhomocystinemia (Sharma et al., 2006; Fredriksen et al., 2007). When we used LSKM, no genes in the Hcy pathway are found significant after multiple testing adjustment. This motivated us to propose a quantile regression kernel machine (QRKM) to test for gene effects on one of the quantiles. When using our QRKM to explore further, we found that Hcy level is significantly associated with gene *CBS* at quantile levels of 0.5 and 0.8, and with gene *TCN1* at quantile level of 0.8, after adjusting for multiple tests performed at different genes and quantiles.

We make three major contributions in this article. First, we develop a simple and fast algorithm to solve the semiparametric model for a fixed tuning parameter. Second, we introduce a bootstrap based tuning method, which provides stable selection results and can provide the standard errors of the estimates of the model components with no extra computation cost. Finally, we develop a procedure for testing the joint effect of genetic variables under the semiparametric quantile regression framework. Since the loss function of the quantile regression model is nonsmooth, we can not use the score test in kernel machine literature. Instead, we propose a test statistic based on the subgradient of the check function, and develop a permutation method to compute p-values. To the best of our knowledge, this is first such methodology in the quantile regression kernel machine literature.

2. Penalized Quantile Regression Estimation using Kernel Machines

Suppose we observe n independent triples $(y_i, \mathbf{X}_i, \mathbf{Z}_i)$ where $\mathbf{X}_i$ is a vector of p covariates, y_i is a continuous response and $\mathbf{Z}_i$ is a vector of q covariates. In our motivating data, y_i denotes change in Hcy level, $\mathbf{X}_i$ denotes genotype of a set of SNPs and $\mathbf{Z}_i$ is a vector of age and sex of the individual. We consider a partially linear model to relate the response to the clinical covariates and the genetic covariates:

$$y_i = f(\mathbf{X}_i) + \mathbf{Z}_i^\top \boldsymbol{\beta} + \beta_0 + \varepsilon_i, \quad (2.1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^\top$ denotes the covariate effect, β_0 denotes the intercept effect, $f(\cdot)$ is an unknown centered function quantifying the genetic effect, and ε_i is the random error.

We consider a quantile regression model, where for a fixed value τ we assume the τ th quantile of ε_i conditional on $\mathbf{X}_i$ and $\mathbf{Z}_i$ is assumed to be zero. As we have an intercept term in the model, we assume that $E(f(\cdot)) = 0$ for identifiability. The parameters $\boldsymbol{\beta}$, β_0 and $f(\cdot)$ can be estimated from the following minimization problem

$$\min_{\beta_0, \boldsymbol{\beta}, f} \sum_{i=1}^n \rho_\tau(y_i - f(\mathbf{X}_i) - \mathbf{Z}_i^\top \boldsymbol{\beta} - \beta_0), \quad (2.2)$$

where $\rho_\tau(r) = \tau 1(r \geq 0) + (\tau - 1) 1(r < 0)$ is the check function and $1(\cdot)$ denotes the indicator function. Typically, one assumes a parametric form for $f(\cdot)$ and model the SNP effects. For example, $f(\mathbf{X}_i) = \mathbf{X}_i^\top \boldsymbol{\eta}$ for some unknown parameter vector $\boldsymbol{\eta}$ corresponds to a linear model with main SNP effects only. Such parametric assumptions may be too strong and may not work well if the true underlying effect is nonlinear. To allow for more flexibility, we assume $f(\cdot)$ has a nonparametric form. However, the solution of (2.2) is not unique if we do not put any constraints on the unknown function $f(\cdot)$.

To solve the problem of over-fitting, we consider the penalized version of (2.2)

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

where $P_\lambda(f)$ is a penalty function which reflects the smoothness of $f(\cdot)$. One commonly used penalty function is $\lambda \int f''(x)^2 dx$, which is analogous to classical least squares smoothing spline model pioneered by Wahba (1990) and Gu (2002). Here λ is a penalty parameter controlling smoothness of $f(\cdot)$. In this article, we assume the function $f(\cdot)$ resides in a functional space H_K generated by a positive definite kernel function $K(\cdot, \cdot)$. From Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), there is a one-to-one correspondence between a positive definite kernel function and H_K under some regularity conditions. We can expand $f(\cdot)$ using the basis functions in H_K , where the basis functions can be represented

using the kernel function. By Representer Theorem (Kimeldorf and Wahba, 1971), the solution for the nonparametric function $f(\cdot)$ can be written using the dual representation

$$f(\mathbf{X}) = \lambda^{-1} \sum_{i=1}^n \theta_i K(\mathbf{X}, \mathbf{X}_i), \quad (2.3)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^\top$ are unknown parameters.

For the τ th quantile, our estimation of $f(\cdot)$, $\boldsymbol{\beta}$ and β_0 is based on minimizing

$$\min_{\beta_0, \boldsymbol{\beta}, f \in H_K} \sum_{i=1}^n \rho_\tau(y_i - f(\mathbf{X}_i) - \mathbf{Z}_i^\top \boldsymbol{\beta} - \beta_0) + \lambda \|f\|_{H_K}^2 / 2, \quad (2.4)$$

where the penalty function $\frac{\lambda}{2} \|f\|_{H_K}^2$ is determined by the kernel function to control the roughness of the function. Combining (2.3) and (2.4), the optimization problem becomes

$$\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta}} \sum_{i=1}^n \rho_\tau(y_i - \beta_0 - \lambda^{-1} \sum_{i'=1}^n \theta_{i'} K(\mathbf{X}_i, \mathbf{X}_{i'}) - \mathbf{Z}_i^\top \boldsymbol{\beta}) + (2\lambda)^{-1} \sum_{i,i'=1}^n \theta_i \theta_{i'} K(\mathbf{X}_i, \mathbf{X}_{i'}). \quad (2.5)$$

With a slight abuse of notation, define a matrix $\mathbf{K}$ such that the (i, i') -th element is $K(\mathbf{X}_i, \mathbf{X}_{i'})$. We show that the solution of (2.5) can be obtained from a quadratic programming problem.

Theorem 1: For a fixed λ , the minimizer of (2.5) can be found by solving

$$\max_{\boldsymbol{\theta}} \{-(2\lambda)^{-1} \boldsymbol{\theta}^\top \mathbf{K} \boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbf{y}\}, \quad (2.6)$$

subject to $-(1 - \tau) \leq \theta_i \leq \tau$, for $1 \leq i \leq n$, $\sum_{i=1}^n \theta_i = 0$, and $\sum_{i=1}^n Z_{il} \theta_i = 0$, for $1 \leq l \leq q$.

The proof of the theorem is presented in the Supplementary Materials. After we obtain the solution $\hat{\boldsymbol{\theta}}$ from (2.6), we plug the solution into (2.5) and solve for $\boldsymbol{\beta}$ and β_0 in

$$\min_{\boldsymbol{\beta}, \beta_0} \sum_{i=1}^n \rho_\tau(y_i - \beta_0 - \mathbf{Z}_i^\top \boldsymbol{\beta}),$$

where $y'_i = y_i - \frac{1}{\lambda} \sum_{i'=1}^n \hat{\theta}_{i'} K(\mathbf{X}_i, \mathbf{X}_{i'})$. For the nonparametric function $f(\cdot)$, we plug in the estimate $\hat{\theta}$ into (2.3) and get the estimate $\hat{f}(\cdot)$. We used `quantreg` package in R to solve the above regression problem and the quadratic problem in (2.6).

The regularization parameter λ plays an important role in controlling the smoothness of the function $f(\cdot)$. We discuss the selection of the tuning parameter in Web Appendix C.

3. Testing for the Joint Effect

In biomedical studies, evidence of association between a gene and a response is as valuable as estimation of the actual effect. Our goal is to test the hypothesis $H_0 : f(\cdot) = 0$. Note that we have assumed that $E(f(\cdot)) = 0$ for identifiability, this hypothesis is equivalent to testing whether $\mathbf{X}_j$ has a constant effect or not. Using LSKM, Liu et al. (2007) tested the whole genetic effects using the score test, where they assume $\varepsilon_i \sim N(0, \sigma^2)$. Since the check function is nonsmooth, we can not directly apply the score test under the least squares case. Instead, we propose a score type of test statistic using the subgradient of the check function. Recall that we have n independent triples $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$. We fit a linear quantile regression using only $(Y_i, \mathbf{Z}_i)$, that is to fit the null model. Let $\hat{u}_i = Y_i - \hat{\beta}_0 - \mathbf{Z}_i^\top \hat{\beta}$ ($1 \leq i \leq n$), and define $w_i = \tau$ if $\hat{u}_i > 0$, and $w_i = \tau - 1$ if $\hat{u}_i < 0$. For those $\hat{u}_i = 0$, we assign the corresponding $w_i = \tau$ with probability $1 - \tau$, and $w_i = \tau - 1$ with probability τ . Our proposed test statistic is

$$T = \mathbf{w}^\top \mathbf{K} \mathbf{w},$$

where $\mathbf{K}$ is the aforementioned kernel matrix and $\mathbf{w} = (w_1, \dots, w_n)^\top$.

To get the p-value and perform hypothesis testing, we need to get the null distribution of the test statistic T . As our test statistic T depends on the binary random variable w_i , the distribution of T is no longer mixture of chi-square distribution as the least squares case. We apply a permutation based procedure to empirically obtain the distribution of T . We first fit our quantile regression kernel machine procedure on the data $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i), i = 1, \dots, n\}$ and get the residuals $\hat{v}_i = Y_i - \hat{f}(\mathbf{X}_i) - \mathbf{Z}_i^\top \hat{\beta} - \hat{\beta}_0$ ($1 \leq i \leq n$).

Step 1: Permute v_i ($i = 1, \dots, n$) without replacement and a new permutation $\hat{v}_i^*$.

Step 2: Add $\hat{\beta}_0 + \mathbf{Z}_i^\top \hat{\beta}$ to $\hat{v}_i^*$ and get the mimic data Y_i^* .

Step 3: Use $(Y_i^*, \mathbf{Z}_i)$ to do a linear quantile regression and get the new residuals $\hat{u}_i^*$. Obtain the w_i^* using the same rule as w_i , where we replace $\hat{u}_i$ by $\hat{u}_i^*$. Denote $\mathbf{w}^* = (w_1^*, \dots, w_n^*)^\top$.

Step 4: Calculate the mimic test statistic, $T^* = \mathbf{w}^{*\top} \mathbf{K} \mathbf{w}^*$.

Repeat the above steps for N times, and we obtain the $T_1^*, \dots, T_N^*$, which mimics the null distribution of the test statistic. Finally, the p -value can be calculated as

$$p\text{-value} = \sum_{i=1}^N 1(T_i^* \geq T) / N, \text{ where } 1(\cdot) \text{ denotes the indicator function.}$$

4. Simulation Study

We conduct a simulation study to evaluate the performance of our proposed procedure. We investigate two aspects of the model, namely the estimation of the effects of both $\mathbf{X}_j$ and $\mathbf{Z}_j$, and testing for the effect of $\mathbf{X}_j$.

4.1 Simulation for Estimation

We generate data from the model in (2.1) where $\mathbf{Z}_j = (Z_{j1}, Z_{j2})^\top$ were generated from a standard bivariate normal distribution. We generate $\mathbf{X}_j$ using the same frequency distribution of the SNPs as found on the *MTR* gene in the real data application ($p = 20$ SNPs). We set the true value of $\boldsymbol{\beta} = (1, 1)^\top$, and $\beta_0 = 0$. Define $g(\mathbf{X}_i) = 1 + \sum_{k=1}^p X_{ik} \eta_k + \sum_{k=2}^p X_{i1} X_{ik} \gamma_k$, with $\eta_1 = 0.4$, $\eta_2 = \dots = \eta_p = 0.7$, $\gamma_2 = \dots = \gamma_p = 0.2$. We consider $f_1(\cdot) = g(\cdot)$ and $f_2(\cdot) = \text{sign}\{g(\cdot)\} |g(\cdot)|^{1/2}$ as our true functions. For ε_j , we consider the standard normal, t (with degrees of freedom 3) and $\chi_1^2 - 1$ distributions. We consider the sample size $n = 200$. For the quantile, we use $\tau = 0.1, 0.5$, and 0.8 . We use the identity-by-state (IBS) kernel (Wessel and Schork, 2006) in our simulation. We use LSKM as a benchmark approach with five fold cross validation to tune the regularization parameter.

We run 1000 Monte Carlo repetitions and report the mean and standard deviation of the $\boldsymbol{\beta}$ estimates, which are vectors of length 2. We also record the bootstrap standard deviation, which is a byproduct of the tuning process to compare with the numerical study. For LSKM, since we do not use bootstrap tuning, we do not report this quantity, and we present the result using “NA”. We also record the mean absolute deviation (MAD) as

$$\frac{1}{n} \sum_{i=1}^n |\bar{f}(\mathbf{X}_i) - \hat{\bar{f}}(\mathbf{X}_i)|, \text{ where } \bar{f} \text{ is the centered function for } f \text{ and } \hat{\bar{f}} \text{ is the centered estimated function.}$$

The results are in Table 1. We can see that our method performs well in estimating the parameters for different quantiles and different functions. The bootstrap estimate of the standard deviation is close to the standard deviation obtained from 1000 Monte Carlo estimates, which suggests that we can obtain reasonable variance estimates of the parameter by our bootstrap tuning method. We can also see that MADs are quite different for $f_1(\cdot)$ and $f_2(\cdot)$ because the true $f_1(\cdot)$ and $f_2(\cdot)$ are not in the same magnitude. For normal error, we have found that LSKM can achieve smaller MAD as well as smaller standard deviation of $\boldsymbol{\beta}$ than QRKM. This is expected because the error ε_j 's are generated from normal distribution, and least squares estimates are more efficient than quantile regression method under this scenario. For different error distributions, i.e. t -distribution with degrees of freedom 3 and $\chi_1^2 - 1$, the findings are similar as the normal error case, except that sometimes our QRKM may have smaller MAD and standard deviation of $\boldsymbol{\beta}$ at certain quantiles than LSKM. The

possible reason is that when the error deviates from normal, least squares estimates may be less efficient than quantile regression under certain scenarios.

We have considered two other cases, namely, (1) a different gene, *CBS* ($p = 10$) is used to simulate $\mathbf{X}_i$ and IBS kernel was used, and (2) $\mathbf{X}_i$ are generated from a continuous distribution and gaussian kernel was used. Detailed descriptions and the results are in Section B in the Supplementary Materials. The findings of both cases are similar as the *MTR* case.

4.2 Simulation for Testing

The simulation settings are the same as in Section 4.1 with the modification that the data are generated from the model

$$Y_i = c * h(\mathbf{X}_i) + \mathbf{Z}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $h(\cdot)$ is either $f_1(\cdot)$ or $f_2(\cdot)$, and c quantifies departure from H_0 . For each generated data set, we fit our quantile regression model in (2.1) and test for $f(\cdot) = 0$. We investigate our testing procedure for different quantile levels $\tau = 0.1, 0.5, 0.8$. Note that when $c = 0$, it corresponds to the null hypothesis of no effect of $\mathbf{X}$. We set different critical values $\alpha = 0.001, 0.01, 0.05$ to check the type-I error using our permutation type of test. When $c = 0$, the true model does not involve $\mathbf{X}$, so the type-I error would be the same for f_1 and f_2 functions. To examine power, we consider different c values for f_1 ($c = 0.1, 0.2, 0.3, 0.4, 0.5$) and f_2 ($c = 0.6, 1.2, 1.8, 2.4, 3$). We set different c values for f_1 and f_2 respectively because the magnitudes of f_1 and f_2 are different. We compare our test with the score test in LSKM.

Table 2 shows the type I error analysis results based on 10000 replicates. It is clear that our testing procedure has reasonable type-I error rate under different settings. This is true when the error distribution changes from standard normal to t and chi-squared as well. However, for the LSKM test, we can see that the type-I error rate is at the nominal level when the error distribution is normal, however, the type-I error rates are inflated when the errors deviate from normal distribution, especially for nominal level 0.001 and 0.01. This indicates that we should use LSKM with caution when the error is not normal. For our QRKM test, it is robust to the error distributions.

We have also plotted the power curves for nominal level 0.05 in Figure 1. From the results, we can see that for symmetric error distributions (normal and t), quantile 0.5 has the largest power, followed by quantile 0.8. For quantile 0.1 and LSKM, their powers are comparable. For chi-square error distribution, since it is right skewed, the quantile 0.1 and quantile 0.5 have comparable larger power, followed by quantile 0.8, and the LSKM has the lowest power. From the results, we can see that our testing procedure is generally more powerful than the LSKM test for the categorical covariates.

We repeated the above testing simulations using a different gene, *CBS* gene ($p = 10$), to simulate the categorical covariates $\mathbf{X}_i$. The results are in Section B.1 in the Supplementary Materials, and the findings are similar as the *MTR* case. We have also considered the

continuous covariates $\mathbf{X}_i$ for our testing procedure. The detailed description as well as the results are in Section B.2 in the Supplementary Materials. We found that, for continuous covariates, our testing procedure has reasonable type-I error rates as expected; however, the type-I error rates for LSKM are more stable for continuous covariates (using gaussian kernel) compared the categorical covariates (using IBS kernel). From the power results, our QRKM test may achieve larger or smaller power than LSKM test for continuous covariates under different scenarios. Since LSKM is testing the association in terms of mean, while our method is testing association at lower or higher quantiles, sometimes LSKM may have larger power because the two methods are testing different hypothesis.

5. Application to the Vitamin Intervention for Stroke Prevention Trial

To illustrate the proposed methodology, we applied the QRKM on samples collected from the VISP trial. The trial enrolled patients who were 35 or older with a nondisabling cerebral infarction within 120 days of randomization and Hcy levels in the top quartile for the U.S. population. Subjects were randomly assigned to receive daily doses of either a high-dose formulation (containing 25 mg vitamin B6, 0.4 mg vitamin B12, and 2.5 mg folic acid) or a low-dose formulation (containing 200 mg vitamin B6, 6 mg vitamin B12, and 20 mg folic acid). The patients were followed up for a maximum of 2 years. The average follow-up time was 1.7 years. A total of 2100 VISP participants were enrolled in VISP genetic study and genotype information was collected from 9 candidate genes that are involved in homocysteine metabolism. Our analysis here focused on the genetic influence on the Hcy level obtained from a 2 hr methionine load test measured at baseline. After deleting the subjects with missingness, we obtain 1587 subjects. We assess the joint effect of each of the 9 genes, adjusting for the age, sex and the population stratification (the first 10 PCs of all genes) as did in the original study (Hsu et al., 2011). We used the IBS kernel to summarize the multi-locus information. In QRKM, we separately considered the Hcy level at the quantiles of 0.1, 0.5 and 0.8 as our response variable in Table 3.

From the results, we see that the gene *CBS* is significant at the quantiles 0.5 and 0.8, and gene *TCN1* is significant at the quantile 0.8 after Bonferroni correction (i.e., $0.05/(9 \times 3) = 0.00185$) using our QRKM method. For the LSKM method, after Bonferroni correction (i.e., $0.05/9 = 0.0056$), none of the genes are significant. Gene *TCN1* encodes a member of the vitamin B12-binding protein family. It is associated with vitamin B12 deficiency. Several genome-wide association studies have also shown that SNPs in gene *TCN1* were associated with vitamin B12 (Grarup et al., 2013; Tanaka et al., 2009). Low intake of B12 level may be a risk factor for elevated Hcy level (Kalita et al., 2009). Additionally, gene *TCN1* may influence transcobalamin blood level and is a weak determinant of Hcy level (Namour et al., 2001). Further, gene *CBS* is a member of the folate one-carbon metabolism pathway. This pathway may mediate many biological processes in the cell, such as methionine metabolism and Hcy synthesis. Gene *CBS* provides instructions for making an enzyme which is responsible for using vitamin B6 to convert Hcy to cystathionine (Yi et al., 2000). This gene was also shown to be associated with Hcy level (Liewers et al., 2001; Zinck et al., 2015). As the LSKM test is built based on the conditional mean while our QRKM focuses on different quantiles, in practical situations, it might happen that the association is not through the mean but through the quantiles. Our QRKM test can detect such associations while the mean

based LSKM test might not. Ideally, one could employ all these tests to cover various types of associations.

For estimation, we pick up the *TCNI* gene. Out of the 3 SNPs, there are 14 distinct multi-locus genotypes. We plot the estimates of the effects across different quantiles 0.1, 0.5 and 0.8 for different genotypes, see Figure 2(a). We then proceed with effect estimation of the multi-locus genotypes on the Hcy quantile of 0.8, which is found to be significant in our testing procedure. In Figure 2(b), we report the effect estimates of each genotype and their 95% confidence intervals constructed using the standard error obtained by the bootstrap.

We have included the frequencies of all the genotypes in Web Table 6 in the Supplementary Materials, and they maintain the same order as those in Figure 2. The genotypes are displayed by their order of appearance in the VISP data. We observe several genotypes with the effect confidence intervals that do not intersect with 0 for quantile 0.8, and we mark them bold. The genotype specific information on the effect sizes may facilitate the comprehension of the possible mechanisms that lead to the global significance of *TCNI*.

6. Discussion

We develop estimation and testing procedures for semiparametric quantile regression under the kernel machine framework. The usual least squares based LSKM test provides association testing based on conditional mean response. In contrast, our proposed procedure investigates association at pre-specified quantile level(s). In practical situations, it is possible that the marker set has no significant association in terms of mean response, and the actual association is at a lower or higher quantile of the response. Thus these two tests (LSKM and quantile based test) are detecting two different types of association. Ideally, in practical situations, one could employ both these tests to capture a wide variety of association rather than just focusing on mean based tests.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Michele M. Sale and Dr. Bradford B. Worrall for providing the VISP data (NIH grant U01 HG005160). We also thank the Editor, the Associate Editor and the referees for their helpful and constructive comments. Maity's research was partially supported by NIH grant R00 ES017744 and a NCSU Faculty Research and Professional Development (FRPD) grant. Hsu's research was partially supported by NIH grant U01 HG005160. Tzeng's research was partially supported by NIH grants R01 MH084022 and P01 CA142538.

References

- Bickel, P.J.; Li, B. Complex datasets and inverse problems, volume 54 of IMS Lecture Notes Monogr Ser. Inst. Math. Statist.; Beachwood, OH: 2007. Local polynomial regression on unknown manifolds; p. 177-186.
- Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 1. Cambridge University Press; 2000.
- Fan, J.; Gijbels, I. Local polynomial modelling and its applications, volume 66 of Monographs on Statistics and Applied Probability. Chapman & Hall; London: 1996.

- Fredriksen Å, Meyer K, Ueland PM, Vollset SE, Grotmol T, Schneede J. Large-scale population-based metabolic phenotyping of thirteen genetic polymorphisms related to one-carbon metabolism. *Human Mutation*. 2007; 28:856–865. [PubMed: 17436311]
- Grarup N, Sulem P, Sandholt CH, Thorleifsson G, Ahluwalia TS, Steinthorsdottir V, Bjarnason H, Gudbjartsson DF, Magnusson OT, Sparsø T, et al. Genetic architecture of vitamin b12 and folate levels uncovered applying deeply sequenced large datasets. *PLoS genetics*. 2013; 9:e1003530. [PubMed: 23754956]
- Gu, C. Smoothing spline ANOVA models. Springer Series in Statistics; Springer-Verlag, New York: 2002.
- Hendricks W, Koenker R. Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*. 1992; 87:58–68.
- Hsu FC, Sides E, Mychaleckyj J, Worrall B, Elias G, Liu Y, Chen WM, Coull B, Toole J, Rich S, et al. Transcobalamin 2 variant associated with poststroke homocysteine modifies recurrent stroke risk. *Neurology*. 2011; 77:1543–1550. [PubMed: 21975197]
- Kalita J, Kumar G, Bansal V, Misra UK. Relationship of homocysteine with other risk factors and outcome of ischemic stroke. *Clinical Neurology and Neurosurgery*. 2009; 111:364–367. [PubMed: 19185985]
- Kimeldorf GS, Wahba G. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*. 1971; 33:82–95.
- Koenker, R. Quantile regression, volume 38 of *Econometric Society Monographs*. Cambridge University Press; Cambridge: 2005.
- Koenker, R.; Mizera, I. Statistical data analysis based on the L_1 -norm and related methods (Neuchâtel, 2002), *Stat Ind Technol*. Birkhäuser; Basel: 2002. Elastic and plastic splines: some experimental comparisons; p. 405-414.
- Koenker R, Mizera I. Penalized triograms: total variation regularization for bivariate smoothing. *J R Stat Soc Ser B Stat Methodol*. 2004; 66:145–163.
- Koenker R, Ng P, Portnoy S. Quantile smoothing splines. *Biometrika*. 1994; 81:673–680.
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *American Journal of Human Genetics*. 2008; 82:386–397. [PubMed: 18252219]
- Lee S. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric Theory*. 2003; 19:1–31.
- Li Y, Liu Y, Zhu J. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*. 2007; 102:255–268.
- Liang H, Li R. Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*. 2009; 104:234–248. [PubMed: 20046976]
- Lievers K, Kluijtmans L, Heil SG, Boers G, Verhoef P, van Oepenraay-Emmerzaal D, den Heijer M, Trijbels FJ, Blom HJ. A 31 bp vnt in the cystathionine beta-synthase (cbs) gene is associated with reduced cbs activity and elevated post-load homocysteine levels. *European Journal of Human Genetics*. 2001; 9:583–589. [PubMed: 11528503]
- Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*. 2007; 63:1079–1088. 1311. [PubMed: 18078480]
- Liu Y, Wu Y. Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *Journal of Nonparametric Statistics*. 2011; 23:415–437. [PubMed: 22190842]
- Maity A, Lin X. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics*. 2011; 67:1271–1284. [PubMed: 21504419]
- Maity A, Sullivan PF, Tzeng J. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*. 2012; 36:686–695. [PubMed: 22899176]
- Meyer NJ, Daye ZJ, Rushefski M, Aplenc R, Lanken PN, Shashaty MGS, et al. Snp-set analysis replicates acute lung injury genetic risk factors. *BMC Medical Genetics*. 2012; 13:52. [PubMed: 22742663]

- Monsees GM, Kraft P, Chanock SJ, Hunter DJ, Han J. Comprehensive screen of genetic variation in dna repair pathway genes and postmenopausal breast cancer risk. *Breast Cancer Research and Treatment*. 2011; 125:207–214. [PubMed: 20496165]
- Namour F, Olivier JL, Abdelmouttaleb I, Adjalla C, Debard R, Salvat C, Guéant JL. Transcobalamin codon 259 polymorphism in ht-29 and caco-2 cells and in caucasians: relation to transcobalamin and homocysteine concentration in blood. *Blood*. 2001; 97:1092–1098. [PubMed: 11159542]
- Nychka D, Gray G, Haaland P, Martin D, O'Connell M. A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*. 1995; 90:1171–1178.
- Pettigrew LC, Bang H, Chambless LE, Howard VJ, Toole JF, Investigators V, et al. Assessment of pre- and post-methionine load homocysteine for prediction of recurrent stroke and coronary artery disease in the vitamin intervention for stroke prevention trial. *Atherosclerosis*. 2008; 200:345–349. [PubMed: 18166191]
- Scholkopf, B.; Smola, AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press; Cambridge, MA, USA: 2001.
- Sharma P, Senthilkumar R, Brahmachari V, Sundaramoorthy E, Mahajan A, Sharma A, Sengupta S. Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. *Lipids in Health and Disease*. 2006; 5:1–19. [PubMed: 16430779]
- Song S, Ritov Y, Härdle WK. Bootstrap confidence bands and partial linear quantile regression. *Journal of Multivariate Analysis*. 2012; 107:244–262.
- Sun Y. Semiparametric efficient estimation of partially linear quantile regression models. *Annals of Economics and Finance, Society for AEF*. 2005; 6:105–127.
- Takeuchi I, Le QV, Sears TD, Smola AJ. Nonparametric quantile estimation. *The Journal of Machine Learning Research*. 2006; 7:1231–1264.
- Tanaka T, Scheet P, Giusti B, Bandinelli S, Piras MG, Usala G, Lai S, Mulas A, Corsi AM, Vestri A, et al. Genome-wide association study of vitamin b6, vitamin b12, folate, and homocysteine blood concentrations. *The American Journal of Human Genetics*. 2009; 84:477–482. [PubMed: 19303062]
- Toole JF, Malinow MR, Chambless LE, Spence JD, Pettigrew LC, Howard VJ, Sides EG, Wang CH, Stampfer M. Lowering homocysteine in patients with ischemic stroke to prevent recurrent stroke, myocardial infarction, and death: the vitamin intervention for stroke prevention (visp) randomized controlled trial. *Journal of the American Medical Association*. 2004; 291:565–575. [PubMed: 14762035]
- Vapnik, VN. *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. John Wiley & Sons Inc., New York: A Wiley-Interscience Publication; 1998. Statistical learning theory.
- Wahba, G. *Spline models for observational data*, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM); Philadelphia, PA: 1990.
- Wang H, He X. Detecting differential expressions in GeneChip microarray studies: a quantile approach. *Journal of the American Statistical Association*. 2007; 102:104–112.
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *American Journal of Human Genetics*. 2006; 79:792–806. [PubMed: 17033957]
- Wu C, Yu Y. Partially linear modeling of conditional quantiles using penalized splines. *Computational Statistics & Data Analysis*. 2014; 77:170–187.
- Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (skat). *American Journal of Human Genetics*. 2011; 89:82–93. [PubMed: 21737059]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful snp set analysis for case-control genomewide association studies. *American Journal of Human Genetics*. 2010; 86:929–942. [PubMed: 20560208]

- Yang S. Censored median regression using weighted empirical survival and hazard functions. *Journal of the American Statistical Association*. 1999; 94:137–145.
- Yi P, Melnyk S, Pogribna M, Pogribny IP, Hine RJ, James SJ. Increase in plasma homocysteine associated with parallel increases in plasma s-adenosylhomocysteine and lymphocyte dna hypomethylation. *Journal of Biological Chemistry*. 2000; 275:29318–29323. [PubMed: 10884384]
- Yu K, Jones MC. Local linear quantile regression. *Journal of the American Statistical Association*. 1998; 93:228–237.
- Yuan M. GACV for quantile smoothing splines. *Computational Statistics & Data Analysis*. 2006; 50:813–829.
- Zinck JW, de Groh M, MacFarlane AJ. Genetic modifiers of folate, vitamin b-12, and homocysteine status in a cross-sectional study of the canadian population. *The American Journal of Clinical Nutrition*. 2015 page to appear.

Figure 1.

Results from simulation study: the categorical covariates using the same frequency distribution as that on the *MTR* gene. Displayed are the power curves of QRKM for quantiles $\tau = 0.1, 0.5, 0.8$ and the LSKM comparison test for different functions f_1 and f_2 and different error distributions (normal, t and chi-squares) based on 500 Monte Carlo repetitions. The first row corresponds to function f_1 , the second row corresponds to function f_2 . The three columns from left to right correspond to normal, t and chi-squared random errors. The dashed, dotted, dash dotted and solid lines correspond to the power curves of QRKM for quantiles $\tau = 0.1, \tau = 0.5, \tau = 0.8$ and LSKM test respectively.

Figure 2.

Effect estimation of *TCNI* gene in VISP data analysis. Panel (a) plots the estimated effects for each of the 3-SNP genotypes across different quantiles 0.1, 0.5 and 0.8, represented by dashed, dotted and dash dotted lines respectively. Panel (b) plots the estimated effects and confidence intervals for different genotypes when quantile level is 0.8.

Table 1

Simulation results of our estimation procedure compared with LSKM for categorical covariates using the same frequency distribution as that on the MTR gene. Displayed are average estimates for β (column 4), empirical and bootstrap estimated standard error (columns 5-6), mean absolute deviation (MAD) (column 7) for different error distributions (normal, t and chi-squares), different functions $f_1(\cdot)$ and $f_2(\cdot)$ and different quantiles $\tau = 0.1, 0.5, 0.8$. The true value of β is $(1,1)^T$, and the simulation is based on 1000 Monte Carlo repetitions.

Method	Distribution	f	mean of β	sd of β	bootstrap sd	MAD
QRKM $\tau = 0.1$	normal	f1	(1.01,1.00)	(0.18,0.17)	(0.17,0.17)	1.03
QRKM $\tau = 0.5$	normal	f1	(1.00,0.99)	(0.11,0.12)	(0.14,0.14)	0.57
QRKM $\tau = 0.8$	normal	f1	(1.00,1.00)	(0.15,0.16)	(0.16,0.16)	0.60
LSKM	normal	f1	(1.00,1.00)	(0.08,0.08)	NA	0.35
QRKM $\tau = 0.1$	normal	f2	(1.01,1.00)	(0.13,0.13)	(0.12,0.12)	0.26
QRKM $\tau = 0.5$	normal	f2	(1.00,1.00)	(0.09,0.09)	(0.09,0.10)	0.22
QRKM $\tau = 0.8$	normal	f2	(1.00,1.00)	(0.11,0.11)	(0.10,0.10)	0.24
LSKM	normal	f2	(1.00,1.00)	(0.08,0.07)	NA	0.21
QRKM $\tau = 0.1$	t	f1	(1.01,1.01)	(0.26,0.25)	(0.24,0.24)	1.33
QRKM $\tau = 0.5$	t	f1	(1.00,1.00)	(0.14,0.15)	(0.16,0.16)	0.68
QRKM $\tau = 0.8$	t	f1	(1.00,1.00)	(0.21,0.22)	(0.22,0.22)	0.82
LSKM	t	f1	(1.01,1.00)	(0.15,0.14)	NA	0.49
QRKM $\tau = 0.1$	t	f2	(1.01,1.00)	(0.20,0.20)	(0.19,0.19)	0.30
QRKM $\tau = 0.5$	t	f2	(1.00,1.00)	(0.10,0.10)	(0.11,0.11)	0.23
QRKM $\tau = 0.8$	t	f2	(1.00,1.00)	(0.14,0.14)	(0.14,0.14)	0.26
LSKM	t	f2	(1.00,1.00)	(0.14,0.14)	NA	0.35
QRKM $\tau = 0.1$	$\chi_1^2 - 1$	f1	(1.00,1.00)	(0.09,0.10)	(0.10,0.10)	0.61
QRKM $\tau = 0.5$	$\chi_1^2 - 1$	f1	(1.00,1.00)	(0.11,0.11)	(0.13,0.13)	0.46
QRKM $\tau = 0.8$	$\chi_1^2 - 1$	f1	(1.00,1.01)	(0.25,0.24)	(0.25,0.24)	0.89
LSKM	$\chi_1^2 - 1$	f1	(1.01,1.00)	(0.11,0.12)	NA	0.43

Method	Distribution	f	mean of β	sd of β	bootstrap sd	MAD
QRKM $\tau = 0.1$	$\chi^2_1 - 1$	t_2	(1.00,1.00)	(0.03,0.03)	(0.04,0.04)	0.09
QRKM $\tau = 0.5$	$\chi^2_1 - 1$	t_2	(1.00,1.00)	(0.07,0.07)	(0.08,0.08)	0.19
QRKM $\tau = 0.8$	$\chi^2_1 - 1$	t_2	(1.01,1.01)	(0.20,0.19)	(0.18,0.18)	0.30
LSKM	$\chi^2_1 - 1$	t_2	(1.00,1.00)	(0.10,0.11)	NA	0.29

Table 2

Displayed are the Type-I error of our proposed test for different quantiles compared with LSKM test for different nominal levels and different error distributions (normal, t and chi-squares). We consider categorical covariates using the same frequency distribution as that on the MTR gene. The simulation is based on 10000 Monte Carlo repetitions.

Method	distribution	0.001	0.01	0.05
QRKM $\tau = 0.1$	normal	0.0003	0.0089	0.0498
QRKM $\tau = 0.5$	normal	0.0008	0.0113	0.0496
QRKM $\tau = 0.8$	normal	0.0007	0.0096	0.0473
LSKM	normal	0.0015	0.0103	0.0514
QRKM $\tau = 0.1$	t	0.0005	0.0098	0.0499
QRKM $\tau = 0.5$	t	0.0005	0.0102	0.0501
QRKM $\tau = 0.8$	t	0.0009	0.0093	0.0471
LSKM	t	0.0056	0.0175	0.0582
QRKM $\tau = 0.1$	$\chi_1^2 - 1$	0.0009	0.0107	0.0492
QRKM $\tau = 0.5$	$\chi_1^2 - 1$	0.0010	0.0096	0.0475
QRKM $\tau = 0.8$	$\chi_1^2 - 1$	0.0007	0.0112	0.0483
LSKM	$\chi_1^2 - 1$	0.0058	0.0208	0.0546

Results from the real data application. The number of SNPs on each gene is reported in column 2. Columns 3-5 display the p-values of our proposed test for 9 genes on different quantiles 0.1, 0.5, 0.8. The last column report the p-values of the LSKM test.

Table 3

Gene name	SNP number	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.8$	LSKM
<i>BHMT</i>	5	0.072	0.108	0.263	0.413
<i>BHMT2</i>	4	0.222	0.371	0.748	0.514
<i>CBS</i>	10	0.013	0.0004	0.0004	0.046
<i>CTH</i>	10	0.559	0.432	0.820	0.039
<i>MTHFR</i>	9	0.745	0.478	0.429	0.509
<i>MTR</i>	20	0.785	0.864	0.572	0.446
<i>MTRR</i>	5	0.020	0.070	0.860	0.170
<i>TCN1</i>	3	0.808	0.218	0.0006	0.073
<i>TCN2</i>	20	0.455	0.622	0.835	0.314