Statistics Innovation

Article

Open Access

https://doi.org/10.48130/stati-0026-0015

*Statistics Innovation* **2026**, 3: e014

# Quantile varying multi-index coefficient model for synergistic gene–environment interactions

Shunjie Guan[1], Yusen Zhang[2] and Yuehua Cui[1*]

[1] *Department of Statistics and Probability, Michigan State University, East Lansing, MI 48823, USA*
[2] *Xi'an Jiaotong-Liverpool University, Suzhou 215123, China*
* Correspondence: cuiy@msu.edu (Cui Y)

## Abstract

Gene–environment (G × E) interaction plays an important role in our understanding of complex traits, particularly when multiple environmental exposures act jointly as a mixture. Existing methods for studying continuous traits primarily focus on modeling conditional mean effects, whereas scientific and clinical interest often lies in identifying specific quantiles of the outcome distribution. We propose a penalized quantile regression framework for the varying multi-index coefficient model to investigate genetic effects and gene-by-mixture interactions across different quantiles. The method simultaneously identifies genetic variants with varying effects (interaction), constant effects (main effects), or no effect, while estimating environmental loading parameters within the mixture. By modeling multiple quantiles, the framework captures heterogeneous genetic and environmental effects along the response distribution. Simulation studies demonstrate accurate variable selection and robust estimation performance. In a real-data application to birth weight, the proposed approach identifies quantile-specific genetic main effects and heterogeneous environmental mixture effects, with limited evidence of nonlinear gene-by-mixture interaction. The proposed method provides a flexible and computationally efficient tool for high-dimensional G × E studies.

**Keywords:** Quantile regression, Gene–environment interaction, Environmental mixtures, High-dimensional variable selection, Varying multi-index coefficient model

**Citation:** Guan S, Zhang Y, Cui Y. 2026. Quantile varying multi-index coefficient model for synergistic gene–environment interactions. *Statistics Innovation* 3: e014 https://doi.org/10.48130/stati-0026-0015

## Introduction

Identification of gene–environment (G × E) interactions has long been a central focus in genetic epidemiology. G × E interactions cause variation in the effect of a genotype on disease risk across different environmental exposure conditions[1,2]. Traditionally, G × E interactions have been studied using models that consider a single environmental exposure, with subsequent extensions allowing for nonlinear G × E effects[3]. However, an increasing number of epidemiological studies suggest that disease risk may be influenced by simultaneous exposure to multiple environmental factors[4,5]. When both multiple genetic and environmental variables are incorporated, the model dimension increases substantially due to the inclusion of numerous interaction terms, leading to estimation instability and inflated standard errors, a phenomenon commonly referred to as the curse of dimensionality. To alleviate this burden, the varying multi-index coefficient model (VMICM) proposed by Liu et al.[6] has been applied to model the interaction between a mixture of environmental variables $\boldsymbol{X}$ and genetic factors $\boldsymbol{G}$ as

$$\boldsymbol{Y} = \sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta})G_k + \boldsymbol{\epsilon}, \tag{1}$$

where $m_k(\cdot)$, $k = 0, 1, \cdots, p$, are continuous smooth functions, $\boldsymbol{\beta}$ is a $q$-dimensional loading parameter for $q$ environmental variables, and $\boldsymbol{\epsilon}$ denotes the random error. For cases where both $p$ and $q$ are large, Guan et al.[7] proposed an iterative three-step variable selection procedure for model (1), which classifies the nonparametric function $m_k(\cdot)$ into three categories: varying, constant, and zero, corresponding to gene-by-mixture interaction effects (varying), genetic main effects without interaction (constant), and no genetic effect (zero), respectively.

Model (1) is developed for continuous traits, where the primary objective is to model the conditional mean effect. For a binary trait, Guan et al.[8] extended model (1) and proposed a variable selection framework to identify important genetic and environmental variables. In practical applications, however, researchers are often interested in assessing genetic effects at specific quantiles of a continuous trait distribution rather than at the mean. For example, in studies on birth weight, the focus is typically not on how genes influence the average birth weight, but rather on the upper or lower quantiles of the distribution, since extremely high or low birth weight may lead to long-term health complications[9,10]. Even when interest lies in the central tendency of the conditional distribution, median regression (quantile regression with $\tau = 0.5$) yields more robust estimators than mean regression, especially in the presence of outliers or heavy-tailed trait distributions. However, extending methods developed under mean or logistic regression to the quantile regression setting is non-trivial. Therefore, it is necessary to develop a quantile varying multi-index coefficient model within a high-dimensional penalized regression framework to identify genetic and environmental variables associated with specific quantiles of a disease trait distribution.

With the widespread availability of hundreds of thousands, or even millions, of single-nucleotide polymorphisms (SNPs), genetic studies now operate in inherently high-dimensional settings. In such contexts, traditional model selection procedures, including forward or backward stepwise selection and information-criterion-based approaches such as AIC or Bayesian information criterion (BIC), become computationally impractical and statistically unreliable. When the number of predictors is large relative to the sample size, classical selection methods often exhibit instability, inflated variance, and poor reproducibility. To overcome these limitations, penalized regression methods have become a powerful and widely adopted framework for high-dimensional variable selection. The key idea is to augment the loss (or likelihood) function with a penalty

https://doi.org/10.48130/stati-0026-0015

term, thereby enabling simultaneous parameter estimation and variable selection. Different penalty functions yield estimators with distinct theoretical and practical properties, such as sparsity, reduced estimation bias, and computational scalability, making penalized approaches particularly well suited for large-scale genetic analyses. Fan & Li[11] established three desirable properties for penalized estimators: sparsity, unbiasedness, and continuity. They further introduced the concept of the oracle property, according to which an estimator performs asymptotically as well as if the true underlying model were known in advance. Since then, a rich class of penalization methods has been developed. For example, the smoothly clipped absolute deviation (SCAD) penalty[11] and the minimax concave penalty (MCP)[12] are nonconvex penalties designed to reduce estimation bias while maintaining sparsity. The adaptive LASSO[13] method achieves the oracle property through data-driven weights applied to the $\ell_1$ penalty. These methods have become foundational tools for high-dimensional inference in genetic and genomic studies.

Building upon this rich body of literature, we propose a quantile regression-based variable selection framework under model (1) to investigate how the modification of genetic effects by simultaneous exposure to multiple environmental factors influences a disease trait under different quantiles. The proposed approach incorporates the MCP penalty within a high-dimensional penalized quantile regression framework to achieve effective variable selection while reducing the estimation bias. We assess the performance of the proposed method through comprehensive simulation studies and real-data analysis. By extending penalized variable selection techniques to the quantile varying multi-index coefficient model, our work contributes to and enriches the existing literature on high-dimensional variable selection for quantile regression.

Although the proposed framework is motivated by existing varying multi-index coefficient models for mean and binary outcomes, extending these approaches to a high-dimensional quantile regression setting leads to substantial additional challenges due to the non-smooth quantile loss function, nonconvex penalization, spline approximation, and iterative optimization procedure. The primary objective of the current work is therefore to develop a practical estimation and variable selection framework together with empirical evaluation through simulations and real-data analysis. A formal theoretical investigation of identifiability, estimation consistency, oracle properties, and convergence behavior under the proposed quantile framework remains an important topic for future research. The rest of the paper is organized as follows. The following section introduces the proposed variable selection framework, including the penalized quantile loss function, the iterative estimation algorithm, and the selection of tuning parameters and initial values for $\boldsymbol{\beta}$. The section "Simulation studies" examines the finite-sample performance of the proposed method through Monte Carlo simulations. The section "Real-data application" describes the application of the proposed method to a birth-weight data set. The last section concludes with discussion.

# Materials and methods

Throughout the paper, superscript $T$ denotes matrix transpose, $\|\cdot\|_p$ denotes the $L_p$ norm, and log($a$) denotes the natural logarithm of $a$. For simplicity, we use "constant" to refer to a nonzero constant effect unless otherwise specified.

## Model setup

For a random sample of size $n$, let $\boldsymbol{Y}_{n\times 1}$ denote the response vector, $\boldsymbol{X}_{n\times q}$ denote the matrix of environmental variables, and $\boldsymbol{G}_{n\times(p+1)}$ denote the matrix of genetic variables. We propose the following quantile varying multi-index coefficient model:

$$\boldsymbol{Y}(\tau)=\sum_{k=0}^{p} m_k(\boldsymbol{X}\boldsymbol{\beta}(\tau),\tau)\boldsymbol{G}_k+\boldsymbol{\epsilon}(\tau), \tag{2}$$

where $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^T$ is the continuous trait of interest, and $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_q)$ contains $q$ continuous environmental variables. The genetic matrix is defined as $\boldsymbol{G}_{n\times(p+1)}=(\boldsymbol{G}_0,\boldsymbol{G}_1,\cdots,\boldsymbol{G}_p)$, where $\boldsymbol{G}_0=(1,\cdots,1)^T$ is the intercept term and $\boldsymbol{G}_k$ is the length-$n$ vector for the $k$-th genetic variable, $k=1,2,\cdots,p$. The functions $\{m_k(\cdot,\tau)\}_{k=0,1,\cdots,p}$ are unknown nonparametric functions at quantile level $\tau$, and $\boldsymbol{\beta}(\tau)=(\beta_1(\tau),\cdots,\beta_q(\tau))^T$ is the corresponding vector of environmental loading parameters. The error term $\boldsymbol{\epsilon}(\tau)$ satisfies $P\{\boldsymbol{\epsilon}(\tau)<0\mid\boldsymbol{X},\boldsymbol{G}\}=\tau$, for a given quantile level $0<\tau<1$. The special case $\tau=0.5$ corresponds to the median regression. For notational simplicity, all parameters are regarded as quantile-specific; hence, we use the notations $\boldsymbol{\beta}$ and $m_k(\cdot)$ in place of $\boldsymbol{\beta}(\tau)$ and $m_k(\cdot,\tau)$, respectively.

## Parameter estimation

Our goal is to estimate and select unknown functions $\{m_k(\cdot)\}_{k=0,1,\cdots,p}$ and the unknown loading parameter $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_q)^T$. To ensure identifiability, we assume $\|\boldsymbol{\beta}\|_2=1$ and $\beta_1>0$, and that $m_k(\cdot)$ cannot have the form $m_k(\boldsymbol{u})=\boldsymbol{\alpha}^T\boldsymbol{u}\boldsymbol{\beta}^T\boldsymbol{u}+\boldsymbol{\gamma}^T\boldsymbol{u}+c$. Following Schumaker[14], we construct B-spline basis functions for $m_k(\cdot)$. We adopt B-spline basis functions because of their flexibility, local support property, and computational efficiency in high-dimensional settings[15–17]. Compared with alternative-basis expansions, B-splines yield sparse design matrices and provide a stable approximation for smooth nonlinear functions with relatively low computational complexity. Other basis functions, such as wavelet or Fourier bases, could also be incorporated into the proposed framework, although their comparative performance is beyond the scope of the current work. Let $u=\boldsymbol{X}\boldsymbol{\beta}$. To represent the unknown smooth functions $m_k(\cdot)$, we use a B-spline expansion with $K$ interior knots and degree $h$. Specifically, each coefficient function is approximated by

$$m_k(u)\approx\gamma_{k1}+\bar{\boldsymbol{B}}(u)\boldsymbol{\gamma}_{k*}, \tag{3}$$

where $\bar{\boldsymbol{B}}(u)$ denotes the centered B-spline basis vector, $\boldsymbol{\gamma}_{k*}=(\gamma_{k2},\gamma_{k3},\cdots,\gamma_{kL})^T$, $\boldsymbol{\gamma}_k=(\gamma_{k1},\boldsymbol{\gamma}_{k*}^T)^T$, and $L=K+h+1$. Under this approximation, Model (2) can be expressed as

$$\boldsymbol{Y}=\sum_{k=0}^{p}\left\{\gamma_{k1}+\bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}\right\}\boldsymbol{G}_k+\boldsymbol{\epsilon}. \tag{4}$$

Thus, estimating the unknown functions $\{m_k(\cdot)\}_{k=0}^p$ is reduced to estimating the spline coefficient vectors $\{\gamma_{k1},\boldsymbol{\gamma}_{k*}\}_{k=0}^p$ together with the loading parameter $\boldsymbol{\beta}$. This representation also provides a natural way to distinguish different types of genetic effects. If $\|\boldsymbol{\gamma}_{k*}\|_2\neq 0$, then the effect of $\boldsymbol{G}_k$ varies with the environmental index $\boldsymbol{X}\boldsymbol{\beta}$, indicating a gene-by-mixture interaction. If $\|\boldsymbol{\gamma}_{k*}\|_2=0$ but $|\gamma_{k1}|\neq 0$, then $\boldsymbol{G}_k$ has a constant genetic effect that does not depend on the environmental mixture. If both $\|\boldsymbol{\gamma}_{k*}\|_2=0$ and $|\gamma_{k1}|=0$, then $\boldsymbol{G}_k$ has no detectable effect on $\boldsymbol{Y}$[18,19].

Motivated by Guan et al.[7], we estimate the model parameters by minimizing the penalized quantile loss. Let $\rho_\tau(u)=u\{\tau-I(u<0)\}$ denote the check loss function. We define the objective function as

$$\begin{aligned}Q_\tau(\boldsymbol{\beta},\boldsymbol{\gamma})=&\sum_{i=1}^{n}\rho_\tau\left(Y_i-\sum_{k=0}^{p}\left\{\gamma_{k1}+\bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\boldsymbol{\gamma}_{k*}\right\}G_{ik}\right)\\&+n\sum_{k=1}^{p}p_{\lambda_1}(\|\boldsymbol{\gamma}_{k*}\|_2)+n\sum_{k=1}^{p}p_{\lambda_2}(|\gamma_{k1}|)I(\|\boldsymbol{\gamma}_{k*}\|_2=0)+n\sum_{d=2}^{q}p_{\lambda_3}(|\beta_d|).\end{aligned} \tag{5}$$

The first term measures the lack of fit under quantile regression, while the remaining terms impose sparsity on the varying components, constant genetic effects, and environmental loading parameters, respectively. The penalty $p_{\lambda_1}(\cdot)$ is applied to the group norm $\|\boldsymbol{\gamma}_{k*}\|_2$ to determine whether the $k$-th genetic effect varies with the environmental index. Conditional on the absence of a varying effect, the penalty $p_{\lambda_2}(\cdot)$ is then used to determine whether the corresponding constant coefficient $\gamma_{k1}$ is nonzero. The last penalty, $p_{\lambda_3}(\cdot)$, selects important components of the loading vector $\boldsymbol{\beta}$. The intercept function $m_0(\cdot)$ is not penalized, so neither $\boldsymbol{\gamma}_{0*}$ nor $\gamma_{01}$ appears in the penalty terms. In addition, $\beta_1$ is left unpenalized because the identifiability constraint requires $\beta_1 > 0$. Throughout this work, we use the MCP penalty[12,20], defined by

$$p(x,\lambda) = \lambda \int_0^x \left(1 - \frac{s}{\delta\lambda}\right)_+ \mathrm{d}s,$$

where $\lambda > 0$ is the tuning parameter and $\delta > 0$ controls the concavity of the penalty.

## Estimation algorithm

Guan et al.[7,8] developed three-step iterative procedures for Model (1) in the conditional mean and binary response settings. These procedures classify each nonparametric function $m_k(\cdot)$ as varying, constant, or zero, corresponding to the sets V, C, and Z, respectively. We extend this estimation strategy to the quantile regression framework.

Step 1: For a given $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^{(0)}$, the step 1 estimator of $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}^{(1)} = \{\hat{\gamma}_{k1}^{(1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(1)T}\}_{k=0,1,\cdots,p}^T$, can be obtained by optimizing the following grouped penalized regression:

$$\hat{\boldsymbol{\gamma}}^{(1)} = \min_{\boldsymbol{\gamma}} Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)}),$$

where

$$Q_1(\boldsymbol{\gamma}|\lambda_1, \hat{\boldsymbol{\beta}}^{(0)}) = \sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{k=0}^p [\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik}\right) + n\sum_{k=1}^p p_{\lambda_1}(\|\boldsymbol{\gamma}_{k*}\|_2).$$

Rather than penalizing the individual components of $\boldsymbol{\gamma}_{k*} = (\gamma_{k2}, \cdots, \gamma_{kL})^T$, we apply the penalty to its $L_2$ norm. This group-wise penalty is used because the varying effect of $\boldsymbol{G}_k$ is determined based on whether the spline coefficient vector $\boldsymbol{\gamma}_{k*}$ is nonzero. Thus, step 1 separates $m_k(\cdot)$, $k = 1, \cdots, p$, into two groups: varying (V) and non-varying (NV). Specifically, $m_k(\cdot)$ is classified as varying if $\|\hat{\boldsymbol{\gamma}}_{k*}^{(1)}\|_2 > 0$, and as non-varying if $\|\hat{\boldsymbol{\gamma}}_{k*}^{(1)}\|_2 = 0$.

Step 2: Based on the step 1 estimators of the B-spline coefficients $\boldsymbol{\gamma}$, the step 2 estimators $\hat{\boldsymbol{\gamma}}^{(2)} = \{(\hat{\gamma}_{k1}^{(2)}, \hat{\boldsymbol{\gamma}}_{k*}^{(2)})_{k\in V}, (\hat{\gamma}_{k1}^{(2)})_{k\in C}\}$ can be obtained via the penalized quantile regression. Note that $\hat{\boldsymbol{\gamma}}_{k*}^{(2)} = 0$ automatically if $\hat{\boldsymbol{\gamma}}_{k*}^{(1)} = 0$. We obtained the estimator as

$$\hat{\boldsymbol{\gamma}}^{(2)} = \min_{\boldsymbol{\gamma}} Q_2(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)})$$

where

$$Q_2\left(\boldsymbol{\gamma}|\lambda_2, \boldsymbol{\beta}^{(0)}, \hat{\boldsymbol{\gamma}}^{(1)}\right) = \sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{k\in V}[\gamma_{k1} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta}^{(0)})\boldsymbol{\gamma}_{k*}]G_{ik} - \sum_{k\in C}\gamma_{k1}^{(2)}G_{ik}\right) + n\sum_{k=1}^p p_{\lambda_2}(|\gamma_{k1}^{(2)}|)I(\|\hat{\boldsymbol{\gamma}}_{k*}^{(1)}\|_2 = 0). \tag{6}$$

Based on the initial estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}^{(0)}$, we can obtain the estimators of the B-spline coefficients $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\gamma}}^{(2)}$, and classify $m_k(\cdot)$, $k = 1, \cdots, p$, into V, C, or Z.

*Step 3*: Update $\hat{\boldsymbol{\beta}}$ via the penalized regression using the expression

$$\hat{\boldsymbol{\beta}} = \min_{\|\boldsymbol{\beta}\|_2=1} Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}),$$

where

$$Q_3(\boldsymbol{\beta}|\lambda_3, \hat{\boldsymbol{\gamma}}^{(2)}) = \sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{k=0}^p [\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\boldsymbol{\beta})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}]G_{ik}\right) + n\sum_{d=2}^q p_{\lambda_3}(|\beta_d|).$$

*Step 4*: Set $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}$; then, iterate steps 1 to 3 until convergence. Denote $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$ as the converged estimators. In practice, the iterative procedure is terminated when the relative change in the loading parameter satisfies $\frac{\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\|_2}{\|\hat{\boldsymbol{\beta}}^{(t)}\|_2} < 10^{-4}$, or when the maximum number of iterations is reached. Empirically, the proposed procedure performed well under the selected spline specifications in our simulation studies.

Due to the discontinuous penalty structure involving the indicator function $I(\|\gamma_{k*}\|_2 = 0)$, direct joint optimization of Eq. (5) is computationally difficult. Following Guan et al.[7,8], we therefore adopt a sequential classification strategy in which step 1 first separates varying and non-varying components, and step 2 subsequently distinguishes constant and zero effects among the non-varying components. Consequently, the resulting estimator should be interpreted as the solution obtained from a blockwise iterative approximation procedure rather than as an exact global minimizer of Eq. (5).

From a computational perspective, the main cost of the proposed procedure arises from repeated penalized quantile regression fitting during the grid search for the tuning parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$, as well as the selection of the spline order $h$ and number of knots $K$. In our numerical studies, the iterative algorithm typically converged within a moderate number of iterations and was computationally feasible for the simulation and real-data settings considered. The use of B-spline basis functions with compact support leads to relatively sparse design matrices, and the coordinate descent algorithm further improves the computational efficiency and numerical stability.

Additional details on the estimation algorithm are provided in the Supplementary File 1. The implementation of the iterative procedure requires selecting the tuning parameters $\lambda_1$, $\lambda_2$, and $\lambda_3$, the spline degree $h$, the number of interior knots $K$, and an appropriate initial value for $\boldsymbol{\beta}$.

## Selection of parameters

The BIC[21] is widely used for tuning parameter selection in linear mean regression models. Its use in quantile regression has also been theoretically justified[22,23]. Accordingly, we use a BIC-type criterion to select the tuning parameters, together with the spline degree $h$ and the number of interior knots $K$ for the B-spline approximation.

### *Selection of the tuning parameters* $\lambda_1, \lambda_2, \lambda_3$

Step 1: We select $\lambda_1$ as the minimizer of

$$\mathrm{BIC}(\lambda_1) = \log\left\{\sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{k=0}^p \left[\hat{\gamma}_{k1}^{(\lambda_1)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}\right]G_{ik}\right)\right\} + \frac{\log(n)}{2n} df_{\lambda_1},$$

where $\{\hat{\gamma}_{k1}^{(\lambda_1)}, \hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_1)}\}_{k=0,1,\cdots,p}$ are the minimizers of $Q_1(\boldsymbol{\gamma} \mid \lambda_1, \hat{\boldsymbol{\beta}}^{(0)})$ defined above, $\hat{\boldsymbol{\beta}}^{(0)}$ is the estimator obtained from the previous iteration, and $df_{\lambda_1}$ denotes the total number of nonzero coefficients corresponding to the penalized parameters under $\lambda_1$.

Step 2: We select $\lambda_2$ as the minimizer of

$$\mathrm{BIC}(\lambda_2) = \log\left\{\sum_{i=1}^n \rho_\tau\left(Y_i - \sum_{k=0}^p \left[\hat{\gamma}_{k1}^{(\lambda_2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})\hat{\boldsymbol{\gamma}}_{k*}^{(\lambda_2)}\right]G_{ik}\right)\right\} + \frac{\log(n)}{2n} df_{\lambda_2},$$

https://doi.org/10.48130/stati-0026-0015

where $\{\hat{\gamma}_{k1}^{(\lambda_2)}, \hat{\gamma}_{k*}^{(\lambda_2)}\}_{k=0,1,\cdots,p}$ are the minimizers of $Q_2(\boldsymbol{\gamma} \mid \lambda_2, \hat{\boldsymbol{\beta}}^{(0)})$ defined above, and $df_{\lambda_2}$ denotes the total number of nonzero coefficients associated with the penalized parameters under $\lambda_2$.

*Step 3*: We select $\lambda_3$ as the minimizer of

$$\mathrm{BIC}(\lambda_3) = \log\left\{\sum_{i=1}^{n}\rho_\tau\left(Y_i - \sum_{k=0}^{p}\left[\hat{\gamma}_{k1}^{(2)} + \bar{\boldsymbol{B}}(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(\lambda_3)})\hat{\boldsymbol{\gamma}}_{k*}^{(2)}\right]G_{ik}\right)\right\} + \frac{\log(n)}{2n}df_{\lambda_3},$$

where $\hat{\boldsymbol{\beta}}^{(\lambda_3)}$ is the minimizer of $Q_3(\boldsymbol{\beta} \mid \lambda_3, \hat{\boldsymbol{\gamma}}^{(2)})$ defined above, and $df_{\lambda_3}$ denotes the number of nonzero components in $\boldsymbol{\beta}$ when $\lambda_3$ is used as the penalty parameter. In practice, the degrees of freedom associated with the penalized estimators are approximated by the number of selected nonzero parameters, following the common practice in high-dimensional penalized regression. Although the proposed framework involves grouped MCP penalties and iterative updating of $\beta$, this approximation provides a computationally efficient criterion for tuning parameter selection in quantile regression settings.

To determine the optimal tuning parameters, $\lambda_1$, $\lambda_2$, and $\lambda_3$ are determined over a grid of exponentially decreasing values, with the minimum set to $10^{-3}$. The maximum value for each tuning parameter is chosen as the smallest value at which all corresponding penalized estimates shrink to zero. For computational efficiency, the number of grid points is set to 100.

### *Selection of the order h and the number of interior knots K*

As discussed in Guan et al.[7], a higher order of the B-spline basis functions allows for greater flexibility but results in more complex functional forms and reduced interpretability. From a practical perspective, G × E effects are unlikely to exhibit highly nonlinear patterns. To balance model flexibility and interpretability, we look for the optimal spline order over the set $h \in \{2,3,4\}$.

For the number of interior knots $K$, we consider the candidate set $\mathcal{K} = \{2,3,4,5\}$. For each combination of $K$ and $h$, we fit the following intercept-only model:

$$Y = m_0(\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon}. \tag{7}$$

As noted in Guan et al.[7], this strategy substantially reduces the computational burden compared to fitting the full model for every $(K,h)$ combination. The optimal values of $K$ and $h$ are selected by minimizing

$$\log\left\{\sum_{i=1}^{n}\rho_\tau(Y_i - \hat{Y}_i)\right\} + \frac{\log(n)}{2n}(K+h+1),$$

where $\hat{Y}_i$ denotes the fitted value for the $i$-th subject under Model (7). We acknowledge that selecting the spline order $h$ and the number of knots $K$ using the intercept-only model may not yield the optimal smoothness level for all nonparametric functions $m_k(\cdot)$. Our primary motivation for this strategy is to substantially reduce the computational burden, since repeated fitting of the full high-dimensional model over all candidate spline specifications is computationally intensive. Empirically, the proposed procedure demonstrated reasonably stable performance under the considered spline settings. Developing more adaptive spline-selection procedures and conducting formal sensitivity analyses are important directions for future research.

## Selection of the initial values

For single-index models, the initial value of $\boldsymbol{\beta}$ is commonly set to $(1,0,\cdots,0)^T$ or $(1/\sqrt{q},\cdots,1/\sqrt{q})^T$. However, these choices did not provide stable performance in our simulation studies. Since $\boldsymbol{\beta}$ enters the model through the nonlinear index $\boldsymbol{X}\boldsymbol{\beta}$, the optimization problem is nonconvex and may be sensitive to initialization. We therefore obtain the initial value of $\boldsymbol{\beta}$ by fitting the intercept-only model (7), which provides a data-adaptive starting point using information from the observed response. A formal multiple-random-start sensitivity analysis was not conducted and is left for future investigation.

# Simulation studies

We performed simulation studies to examine the finite-sample performance of the proposed variable selection procedure. Performance was evaluated using four criteria: the oracle percentage for classifying the nonparametric functions $m_k(u)$, the integrated mean squared error (IMSE) of $\hat{m}_k(u)$, the oracle percentage for selecting the loading parameter $\boldsymbol{\beta}$, and the mean squared error (MSE) of $\boldsymbol{\beta}$. All summaries were computed over 1,000 simulation replications.

The oracle percentage for $m_k(u)$ is defined as the proportion of correct classifications of $m_k(u)$. For example, if the true function $m_k(u)$ belongs to the varying category and it is correctly classified as varying in $g$ out of 1,000 replications, then the oracle percentage for $m_k(u)$ is computed as $\frac{g}{1,000}\times 100\%$.

The IMSE of $m_k(\cdot)$ is defined as

$$\frac{1}{1,000}\sum_{r=1}^{1,000}\left[\frac{1}{100}\sum_{j=1}^{100}\left(\hat{m}_k^{(r)}(u_j) - m_k(u_j)\right)^2\right],$$

where $\hat{m}_k^{(r)}(u_j)$ denotes the estimated value of $m_k(u_j)$ in the $r$-th replication. The grid points $u_j$ are chosen as the $j$-th percentiles over the range of $\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)}$. We report both the IMSE under the selected model (Model IMSE) and the IMSE under the oracle model (Oracle IMSE), with the oracle model assuming prior knowledge of the true classification of $m_k(\cdot)$.

The oracle percentage for $\boldsymbol{\beta}$ is defined as the proportion of correct selections of its components. Specifically, if $\beta_d \neq 0$ and it is correctly identified as nonzero in $g$ out of 1,000 replications, then the oracle percentage for $\beta_d$ is calculated as $\frac{g}{1,000}\times 100\%$.

The MSE of $\beta_d$ is computed as

$$\frac{1}{1,000}\sum_{r=1}^{1,000}\left(\hat{\beta}_d^{(r)} - \beta_d\right)^2,$$

where $\hat{\beta}_d^{(r)}$ denotes the estimate of $\beta_d$ obtained in the $r$-th simulation replication.

## Simulation setting

The data were generated according to Model (2). We generated $q = 5$ independent environmental variables, each sampled from a $\mathrm{Unif}(0,1)$ distribution. For the loading parameter $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots,\beta_q)^T$, we set $\beta_1 = \beta_2 = 1/\sqrt{2}$ and $\beta_j = 0$ for $j = 3,4,5$, so that only the first two environmental variables contributed to the environmental mixture. The error term $\boldsymbol{\epsilon}$ was generated from a $N(0,1)$ distribution. To ensure that the $\tau$-th conditional quantile of the error term equals zero, we defined

$$\boldsymbol{\epsilon}(\tau) = \boldsymbol{\epsilon} - F^{-1}(\tau),$$

where $F$ denotes the cumulative distribution function of $\boldsymbol{\epsilon}$. Subtracting $F^{-1}(\tau)$ ensures that the $\tau$-th quantile of $\boldsymbol{\epsilon}(\tau)$ is equal to zero. For the genetic factors $\boldsymbol{G}$, we evaluated the performance of the proposed method under both continuous and discrete settings.

## The continuous case

We first evaluated the performance of the proposed method when the genetic predictors $\boldsymbol{G}$ were continuous and independently generated from a $N(0,1)$ distribution. The nonparametric functions

were specified as follows: $m_0(u) = 2\sin(2\pi u)$, $m_1(u) = 2\cos(\pi u) + 2$, $m_2(u) = \sin(2\pi u) + \cos(\pi u) + 1$, $m_3(u) = 2$, $m_4(u) = 2.5$, and $m_k(u) = 0$ for $k = 5, \cdots, p$. Thus, $m_0$, $m_1$, and $m_2$ were varying functions, $m_3$ and $m_4$ were nonzero constants, and the remaining components corresponded to null effects. We considered $p$ = 50 and 100, quantile levels $\tau$ = 0.25, 0.5, 0.75, and sample size $n$ = 2,000.

Table 1 presents the selection and estimation results for the nonparametric functions $m_k(\cdot)$. For the truly nonzero components ($m_0$ to $m_4$), the oracle percentages were essentially 100% across all settings, indicating that the proposed method consistently identified both varying and constant effects at all quantiles. For the zero components, the oracle percentages were highest at the median ($\tau = 0.5$), ranging from 98.7% to 99.1%, corresponding to very low false-positive rates (approximately 1%). In contrast, for $\tau = 0.25$ and $\tau = 0.75$, the oracle percentages for the null effects ranged from about 88% to 91%, implying moderately higher false-positive rates at the tail quantiles.

Regarding estimation accuracy, the IMSE values for $m_1$ to $m_4$ were generally on the order of $10^{-2}$ to $10^{-3}$. The IMSE values were consistently smaller at $\tau = 0.5$ than at $\tau = 0.25$ and $\tau = 0.75$. For $m_0$, the IMSE was substantially larger at the tail quantiles than at the median under the selected model, reflecting greater variability when estimating nonlinear effects in the extremes of the distribution. Comparing $p$ = 50 and $p$ = 100, we did not observe a meaningful difference in selection or estimation performance.

Table 2 summarizes the selection and estimation results for the loading parameters $\boldsymbol{\beta}$. For the truly nonzero loadings ($\beta_1$ and $\beta_2$), the oracle percentages were essentially 100% across all settings, indicating that the proposed method reliably identifies the important loading parameters. For the zero loadings ($\beta_3$, $\beta_4$, and $\beta_5$), the oracle percentage at the median ($\tau = 0.5$) was approximately 98.5%, reflecting a low false-positive rate. Comparing the lower and upper quantiles, the oracle percentage at $\tau = 0.25$ (around 99%) was slightly higher than that at $\tau = 0.75$ (around 96%). In contrast, the MSE at $\tau = 0.25$ was larger than that at $\tau = 0.75$, indicating a slightly greater estimation variability at the lower quantile. Across $p = 50$ and $p = 100$, we did not observe any meaningful difference in model performance. Overall, the proposed method accurately

**Table 1.** Selection and estimation accuracy of $m_k(\cdot)$ for continuous $\boldsymbol{G}$.

| $\tau$ | $m(\cdot)$ | $p$ = 50 | | | $p$ = 100 | | |
|---|---|---|---|---|---|---|---|
| | | Oracle/% | IMSE (Model) | IMSE (Oracle) | Oracle/% | IMSE (Model) | IMSE (Oracle) |
| 0.25 | $m_0(.)$ | 100.0% | 2.78E+00 | 2.96E−02 | 100.0% | 2.67E+00 | 2.97E−02 |
| | $m_1(.)$ | 100.0% | 8.16E−02 | 6.16E−03 | 100.0% | 7.07E−02 | 6.16E−03 |
| | $m_2(.)$ | 100.0% | 9.59E−02 | 1.30E−02 | 100.0% | 7.80E−02 | 1.31E−02 |
| | $m_3(.)$ | 100.0% | 2.58E−02 | 8.77E−04 | 100.0% | 2.19E−02 | 9.71E−04 |
| | $m_4(.)$ | 100.0% | 2.56E−02 | 1.02E−03 | 99.8% | 2.12E−02 | 9.80E−04 |
| | Zero | 88.8% | 1.03E−02 | 0 | 90.6% | 7.27E−03 | 0 |
| 0.5 | $m_0(.)$ | 100.0% | 7.49E−02 | 2.88E−02 | 100.0% | 7.70E−02 | 2.91E−02 |
| | $m_1(.)$ | 100.0% | 4.98E−02 | 5.07E−03 | 100.0% | 4.89E−02 | 5.31E−03 |
| | $m_2(.)$ | 100.0% | 5.72E−02 | 1.22E−02 | 100.0% | 5.77E−02 | 1.23E−02 |
| | $m_3(.)$ | 99.7% | 1.25E−03 | 7.90E−04 | 99.9% | 1.30E−03 | 7.90E−04 |
| | $m_4(.)$ | 99.9% | 1.50E−03 | 7.91E−04 | 99.8% | 1.48E−03 | 8.29E−04 |
| | Zero | 98.7% | 1.00E−04 | 0 | 99.1% | 6.26E−05 | 0 |
| 0.75 | $m_0(.)$ | 100.0% | 2.60E+00 | 3.01E−02 | 100.0% | 2.60E+00 | 3.13E−02 |
| | $m_1(.)$ | 100.0% | 6.25E−02 | 6.08E−03 | 100.0% | 6.17E−02 | 6.18E−03 |
| | $m_2(.)$ | 100.0% | 7.12E−02 | 1.35E−02 | 100.0% | 7.14E−02 | 1.37E−02 |
| | $m_3(.)$ | 99.9% | 2.42E−02 | 8.58E−04 | 100.0% | 2.18E−02 | 9.03E−04 |
| | $m_4(.)$ | 99.9% | 2.62E−02 | 9.93E−04 | 100.0% | 2.20E−02 | 9.63E−04 |
| | Zero | 88.5% | 1.05E−02 | 0 | 90.7% | 7.19E−03 | 0 |

**Table 2.** Selection and estimation accuracy of $\boldsymbol{\beta}$.

| $\tau$ | $\beta$ | $p$ = 50 | | | $p$ = 100 | | |
|---|---|---|---|---|---|---|---|
| | | Oracle/% | MSE (Model) | MSE (Oracle) | Oracle/% | MSE (Model) | MSE (Oracle) |
| 0.25 | $\beta_1$ | 100.0% | 1.20E−02 | 4.76E−05 | 100.0% | 6.71E−03 | 4.36E−05 |
| | $\beta_2$ | 99.8% | 1.47E−02 | 4.75E−05 | 100.0% | 7.69E−03 | 4.37E−05 |
| | $\beta_3$ | 99.9% | 7.15E−05 | 0 | 99.9% | 5.81E−05 | 0 |
| | $\beta_4$ | 99.9% | 2.09E−04 | 0 | 99.9% | 2.93E−04 | 0 |
| | $\beta_5$ | 99.9% | 9.60E−05 | 0 | 99.9% | 9.53E−05 | 0 |
| 0.5 | $\beta_1$ | 100.0% | 5.29E−05 | 3.78E−05 | 100.0% | 5.33E−05 | 3.54E−05 |
| | $\beta_2$ | 100.0% | 5.29E−05 | 3.78E−05 | 100.0% | 5.34E−05 | 3.54E−05 |
| | $\beta_3$ | 98.6% | 1.52E−06 | 0 | 98.3% | 1.23E−06 | 0 |
| | $\beta_4$ | 98.3% | 3.33E−06 | 0 | 98.8% | 3.24E−06 | 0 |
| | $\beta_5$ | 98.7% | 1.85E−06 | 0 | 98.9% | 1.45E−06 | 0 |
| 0.75 | $\beta_1$ | 100.0% | 4.40E−03 | 4.81E−05 | 100.0% | 3.60E−03 | 4.89E−05 |
| | $\beta_2$ | 100.0% | 4.51E−03 | 4.80E−05 | 100.0% | 3.49E−03 | 4.88E−05 |
| | $\beta_3$ | 95.3% | 4.14E−05 | 0 | 96.3% | 1.87E−05 | 0 |
| | $\beta_4$ | 95.9% | 5.31E−05 | 0 | 96.2% | 4.23E−05 | 0 |
| | $\beta_5$ | 95.8% | 3.85E−05 | 0 | 95.8% | 2.32E−05 | 0 |

https://doi.org/10.48130/stati-0026-0015

selects and estimates the loading parameters with high reliability across all quantiles.

## The discrete case

We further evaluated the performance of the proposed method with discrete genetic predictors $\boldsymbol{G}$. One important application of our framework is the identification of significant SNPs within a gene or pathway. Under an additive genetic model, SNPs take values 0, 1, and 2 corresponding to genotypes aa, Aa, and AA, respectively. We generated $\boldsymbol{G}$ according to

$$P(G_{ij}=0)=\mathrm{MAF}^2,\ P(G_{ij}=1)=2\mathrm{MAF}(1-\mathrm{MAF}),\ P(G_{ij}=2)=(1-\mathrm{MAF})^2,$$

where $G_{ij}$ denotes the $j$-th SNP variant for the $i$-th subject, $i=1,\cdots,n$ and $j=1,\cdots,p$. Simulations were conducted under $p=50,100$, $\tau=0.25,0.5,0.75$, and $n=2,000$. The nonparametric functions $m_k(\cdot)$ and their corresponding minor allele frequencies (MAFs) are summarized in Table 3. This setup allows us to examine the model performance across SNPs with a wide range of allele frequencies.

Table 4 summarizes the selection and estimation performance for the nonparametric functions $m_k(\cdot)$ when the genetic predictors are discrete. The median regression setting ($\tau=0.5$) generally yielded the strongest performance. In particular, the oracle percentage for varying effects was about 99% at $\tau=0.5$, whereas it ranged from 90% to 97% at the tail quantiles when $p=50$ and from 84% to 91% when $p=100$. The model IMSE was also smaller at $\tau=0.5$ than at $\tau=0.25$ or $\tau=0.75$. When comparing the two model dimensions, the $p=50$ setting showed slightly better performance than $p=100$ in terms of oracle percentage and oracle IMSE, although the differences were relatively small.

**Table 3.** Function $m_k(\cdot)$ and the corresponding MAF for $\boldsymbol{G}_k$.

| $m(\cdot)$ function | MAF of $\boldsymbol{G}_k$ |
|---|---|
| $m_0(u)=2sin(2\pi u)$ | – |
| $m_1(u)=2cos(\pi u)+2$ | 0.5 |
| $m_2(u)=sin(2\pi u)+cos(\pi u)+1$ | 0.5 |
| $m_3(u)=2cos(\pi u)+2$ | 0.3 |
| $m_4(u)=sin(2\pi u)+cos(\pi u)+1$ | 0.3 |
| $m_5(u)=2cos(\pi u)+2$ | 0.1 |
| $m_6(u)=sin(2\pi u)+cos(\pi u)+1$ | 0.1 |
| $m_7(u)=2$ | 0.5 |
| $m_8(u)=2$ | 0.3 |
| $m_9(u)=2$ | 0.1 |
| $m_k(u)=0, k>9$ | Unif (0.05, 0.5) |

Figure 1 further illustrates the relationship between MAF and performance. As MAF increases from 0.1 to 0.5, the oracle percentage increases and the IMSE decreases. This confirms that the model performs better for variants with larger MAF than for those with lower MAF, which is expected because variants with lower MAFs contain less information. This behavior is consistent with the reduced Fisher information for relatively rare variants.

Table 5 presents the selection and estimation results for the loading parameters $\boldsymbol{\beta}$ with discrete $\boldsymbol{G}$. The proposed method identified the nonzero loadings ($\beta_1$ and $\beta_2$) with nearly 100% oracle percentage across all settings. For the zero loadings ($\beta_3,\beta_4,\beta_5$), the false-positive rate was very low at $\tau=0.25$ and $\tau=0.5$, and moderately higher at $\tau=0.75$. The MSE values were on the order of $10^{-5}$ to $10^{-7}$, indicating effective shrinkage toward zero.

In summary, the proposed method accurately selects and estimates $m(\cdot)$ and $\boldsymbol{\beta}$ under discrete genetic predictors. The performance improves with increasing MAF, nonzero loadings are consistently identified, and false-positive rates remain low across quantiles, with slightly higher rates at the upper quantile.

The current simulation design primarily evaluates the numerical performance of the proposed penalized quantile estimation and variable selection procedure. Specifically, the same true functions $m_k(\cdot)$ and loading vector $\beta$ are used across quantile levels, while the error term is shifted to ensure that the $\tau$-th conditional quantile of the error equals zero. This setup allows us to isolate the effect of the quantile loss function and assess whether the proposed method can correctly classify varying, constant, and zero effects at different quantiles. However, it does not fully represent data-generating mechanisms in which $m_k(\cdot,\tau)$ or $\beta(\tau)$ genuinely differ across quantiles. More comprehensive simulation studies involving quantile-dependent nonlinear functions, quantile-specific loading parameters, correlated genetic and environmental variables, and skewed or heavy-tailed error distributions would provide further insight into the potential of the method to recover distribution-specific heterogeneity and represent an important direction for future work.

Overall, the simulation settings were designed to provide a controlled environment for evaluating whether the proposed method can distinguish varying, constant, and null effect structures. Although some scenarios yielded high selection accuracy, performance decreased in more challenging settings, including tail quantiles, larger model dimensions, and lower MAFs. These patterns are consistent with the increased estimation difficulty in case of more limited information. From a computational perspective,

**Table 4.** Selection and estimation accuracy for $m_k(\cdot)$ with discrete $\boldsymbol{G}$.

| $\tau$ | Type | $p=50$ | | | $p=100$ | | |
|---|---|---|---|---|---|---|---|
| | | Oracle/% | IMSE (Model) | IMSE (Oracle) | Oracle/% | IMSE (Model) | IMSE (Oracle) |
| 0.25 | $m_0(\cdot)$ | 100.0% | 7.54E−01 | 2.47E−02 | 100.0% | 7.98E−01 | 2.44E−02 |
| | V | 97.5% | 1.44E−01 | 2.30E−02 | 91.4% | 2.23E−01 | 2.32E−02 |
| | C | 94.2% | 1.66E−02 | 3.19E−03 | 95.5% | 1.66E−02 | 3.22E−03 |
| | Z | 93.9% | 4.71E−03 | 0 | 96.0% | 3.05E−03 | 0 |
| 0.5 | $m_0(\cdot)$ | 100.0% | 4.84E−02 | 2.35E−02 | 100.0% | 4.97E−02 | 2.30E−02 |
| | V | 99.9% | 9.14E−02 | 2.00E−02 | 99.3% | 9.84E−02 | 1.98E−02 |
| | C | 95.4% | 7.52E−03 | 2.68E−03 | 95.5% | 9.30E−03 | 2.76E−03 |
| | Z | 93.6% | 2.87E−03 | 0 | 93.8% | 2.87E−03 | 0 |
| 0.75 | $m_0(\cdot)$ | 100.0% | 1.02E+00 | 2.53E−02 | 100.0% | 1.09E+00 | 2.48E−02 |
| | V | 90.8% | 2.29E−01 | 2.33E−02 | 84.1% | 3.35E−01 | 2.30E−02 |
| | C | 96.7% | 1.35E−02 | 3.19E−03 | 98.6% | 1.26E−02 | 3.14E−03 |
| | Z | 96.8% | 2.09E−03 | 0 | 98.5% | 9.24E−04 | 0 |

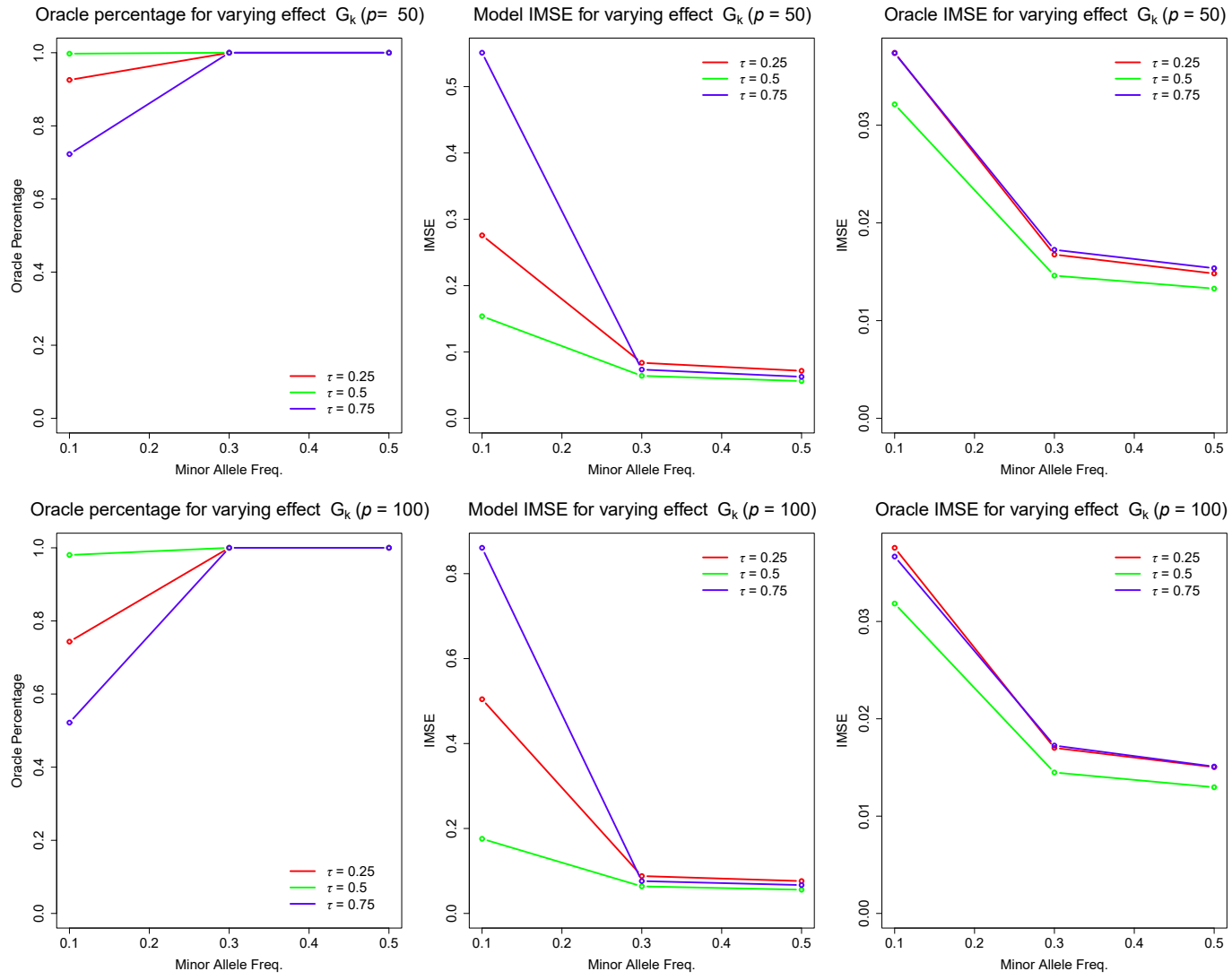


**Fig. 1** Selection and estimation accuracy of $m_k(\cdot)$ for discrete $\boldsymbol{G}$.

**Table 5.** Selection and estimation accuracy for $\boldsymbol{\beta}$ with discrete $\boldsymbol{G}$.

| $\tau$ | $\beta$ | $p = 50$ | | | $p = 100$ | | |
|---|---|---|---|---|---|---|---|
| | | Oracle/% | MSE (Model) | MSE (Oracle) | Oracle/% | MSE (Model) | MSE (Oracle) |
| 0.25 | $\beta_1$ | 100.0% | 3.98E−04 | 4.42E−05 | 100.0% | 4.63E−04 | 4.75E−05 |
| | $\beta_2$ | 100.0% | 4.04E−04 | 4.43E−05 | 100.0% | 4.69E−04 | 4.74E−05 |
| | $\beta_3$ | 100.0% | 0 | 0 | 100.0% | 0 | 0 |
| | $\beta_4$ | 100.0% | 0 | 0 | 100.0% | 0 | 0 |
| | $\beta_5$ | 100.0% | 0 | 0 | 100.0% | 0 | 0 |
| 0.5 | $\beta_1$ | 100.0% | 5.20E−05 | 3.96E−05 | 100.0% | 5.30E−05 | 3.97E−05 |
| | $\beta_2$ | 100.0% | 5.21E−05 | 3.97E−05 | 100.0% | 5.30E−05 | 3.97E−05 |
| | $\beta_3$ | 99.1% | 2.27E−07 | 0 | 99.1% | 1.00E−06 | 0 |
| | $\beta_4$ | 98.9% | 3.43E−07 | 0 | 98.4% | 8.12E−07 | 0 |
| | $\beta_5$ | 98.9% | 5.45E−07 | 0 | 98.7% | 5.87E−07 | 0 |
| 0.75 | $\beta_1$ | 100.0% | 2.75E−04 | 5.13E−05 | 100.0% | 3.70E−04 | 4.69E−05 |
| | $\beta_2$ | 100.0% | 2.84E−04 | 5.15E−05 | 100.0% | 3.59E−04 | 4.69E−05 |
| | $\beta_3$ | 93.6% | 2.52E−05 | 0 | 93.0% | 2.26E−05 | 0 |
| | $\beta_4$ | 94.3% | 2.23E−05 | 0 | 93.9% | 1.80E−05 | 0 |
| | $\beta_5$ | 93.5% | 3.46E−05 | 0 | 93.3% | 3.08E−05 | 0 |

the proposed procedure remains feasible for moderate-scale studies, although repeated penalized quantile regression fitting during tuning parameter selection constitutes the primary computational cost.

https://doi.org/10.48130/stati-0026-0015

## Real-data application

We applied the proposed variable selection method to a birth-weight data set from the Gene Environment Association Studies initiative (GENEVA), which was supported by the Genes, Environment and Health Initiative (GEI). Birth weight has been widely recognized as an important early-life health indicator, which has established associations with morbidity and mortality during infancy and with risks of multiple diseases later in life[9,10]. Birth weight is influenced by both fetal genetic factors and maternal environmental exposures. After standard quality control procedures, that is, removing SNPs with more than $5\%$ missing values, with MAF less than 0.05, or with significant deviation from Hardy–Weinberg equilibrium ($p < 0.001$), the final data set included 1,126 subjects and 590,913 SNPs.

For the environmental variables $\boldsymbol{X}$, we performed marginal regression analyses of birth weight on each environmental factor and selected those with $p < 0.05$. Three environmental variables were retained: mother's mean OGTT (oral glucose tolerance test), diastolic blood pressure ($X_1$), mother's 1-h OGTT glucose level ($X_2$), and mother's mean OGTT systolic blood pressure ($X_3$). This marginal screening step was used as a practical dimension-reduction procedure to reduce the computational burden and focus on environmental variables that have detectable marginal associations with birth weight. However, we acknowledge that marginal screening may exclude environmental variables with weak marginal effects that contribute jointly through the environmental mixture index. Therefore, the selected environmental variables should be interpreted as a working set for illustrating the proposed method rather than as an exhaustive set of potentially relevant maternal exposures.

All SNPs were mapped to known genes based on genomic location. We retained genes containing at least 30 SNPs, resulting in 2,076 genes for analysis. The proposed model was fitted to each gene at $\tau$ = 0.25, 0.5, and 0.75. Since the first component of the loading parameter $\boldsymbol{\beta}$ is constrained to be positive for identifiability, we fitted the model three times by permuting the order of the environmental variables within the index function. A SNP was considered to have a reliable signal only if it was consistently classified as varying or constant across all three model fits. Thus, we report results for only stable classifications across permutations.

The proposed model identified 122 genes with constant genetic effects and no genes with varying effects, suggesting limited evidence of nonlinear G × E interaction with the selected environmental factors in this data set. Given the large number of genes analyzed and the absence of formal multiple-testing adjustment, these findings should be interpreted as exploratory rather than confirmatory. Nevertheless, the absence of varying effects suggests limited evidence of G × E interaction with the selected environmental factors in this data set. As an illustration, we focus on gene *ST3GAL1*[24], located on chromosome 8, which contains 39 SNPs in the cleaned data set. This example demonstrates how the proposed framework can identify quantile-specific genetic effects and heterogeneous environmental mixture effects. Table 6 presents the estimated SNP effects for *ST3GAL1* at different quantiles. The left, middle, and right columns correspond to $\tau = 0.25, 0.5,$ and $0.75$, respectively. Several SNPs exhibit quantile-specific effects. For example, rs13267049, rs6986303, and rs6990329 are associated with lower birth weight at $\tau = 0.25$, while rs2142306 shows an effect only at the median. At the upper quantile ($\tau = 0.75$), rs2736860, rs9643299, and rs7460764 are selected. Notably, SNP rs7831227 shows increasingly negative effects as the quantile increases, suggesting heterogeneous genetic influence across the birth weight distribution. A genome-wide association study by Horikoshi et al.[25] reported that approximately 15% of the variance in birth weight is explained by common genetic variants. Therefore, it is not surprising that only a limited number of SNPs with relatively small effects were identified.

Table 7 presents the estimated loading parameters for the environmental index within gene *ST3GAL1*. The first, second, and third rows correspond to $\tau$ = 0.25, 0.5, and 0.75, respectively.

Figure 2 illustrates the estimated environmental index function $\hat{m}_0(\cdot)$ across quantiles. The red, blue, and black curves correspond to $\tau$ = 0.25, 0.5, and 0.75, respectively. Because the estimated $\boldsymbol{\beta}$ differs across quantiles, the index ranges vary accordingly. At the upper quantile, the fitted birth weight initially decreases with increasing index $\boldsymbol{X\beta}$ and then stabilizes. For the median, the fitted curve decreases and subsequently increases, indicating a nonlinear environmental effect. At the lower quantile, the fitted curve shows an overall increasing trend. These patterns highlight heterogeneous environmental influences across different parts of the birth-weight distribution, demonstrating the advantage of the proposed quantile-based modeling framework. Although 122 genes were identified with constant effects, *ST3GAL1* is presented only as an illustrative example. A more comprehensive biological investigation

**Table 6.** Effect of SNPs in gene *ST3GAL1*.

| SNP ID | $\tau = 0.25$ | $\tau = 0.50$ | $\tau = 0.75$ |
|---|---|---|---|
| rs13267049 | 0.0260 | 0 | 0 |
| rs2736860 | 0 | 0 | −0.0290 |
| rs2142306 | 0 | 0.0720 | 0 |
| rs6986303 | 0.1129 | 0 | 0 |
| rs6990329 | 0.1411 | 0 | 0 |
| rs9643299 | 0 | 0 | −0.0489 |
| rs7460764 | 0 | 0 | −0.1077 |
| rs7831227 | −0.0294 | −0.0801 | −0.1007 |

**Table 7.** Estimated loading parameters corresponding to gene *ST3GAL1*.

| $\tau$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|---|---|---|---|
| 0.25 | 0.272 | 0.707 | 0.653 |
| 0.5 | 0.895 | 0.000 | −0.445 |
| 0.75 | 0.637 | −0.288 | −0.715 |

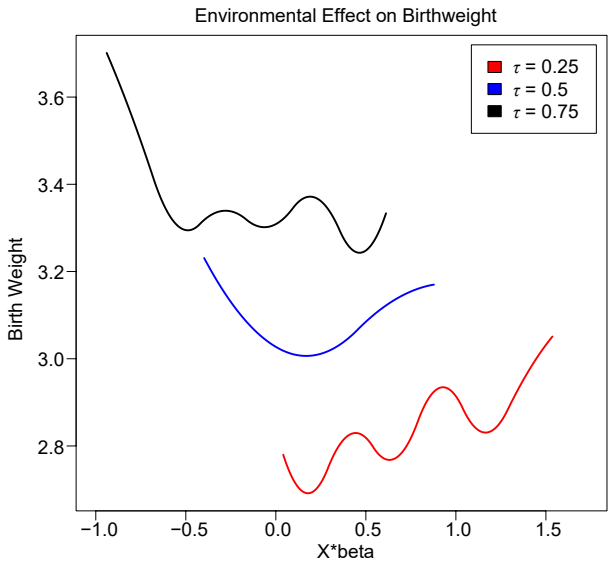


**Fig. 2** Plot of environmental mixture index effect on birth weight.

would require ranked summaries of detected genes and SNPs, resampling-based selection frequencies, uncertainty assessment, and validation against prior birth-weight literature. Because no nonlinear varying effects were detected in this application, future work should also compare the proposed framework with simpler alternatives, such as a main-effect-only quantile regression model or the previously proposed mean-based VMICM.

Because the real-data analysis involved screening environmental variables, fitting the proposed model across 2,076 genes at three quantile levels, and repeating the analysis under different orderings of the environmental variables, multiplicity and selection stability are important considerations. In this study, the real-data analysis is intended as an illustrative application of the proposed method rather than a confirmatory genetic association analysis. Therefore, the identified genes and SNPs should be interpreted cautiously as exploratory findings. Future applications should incorporate formal multiple-testing adjustment, control of the false discovery rate, or resampling-based stability assessment to improve the reliability of biological conclusions.

# Discussion

VMICM provides a flexible framework for modeling nonlinear interactions between genetic variants $\boldsymbol{G}$ and a mixture of environmental factors $\boldsymbol{X}$[6,26,27]. Guan et al.[7,8] developed a three-step variable selection procedure for mean and binary outcomes. In this paper, we extend that framework to conditional quantile regression, allowing investigation of genetic and environmental effects across different parts of the response distribution. Compared with conditional mean regression, quantile regression offers several important advantages. First, modeling multiple quantiles provides a more comprehensive characterization of the outcome distribution. Genetic or environmental effects that are weak at the mean may be pronounced in the lower or upper quantiles. Second, quantile regression is more robust to heavy-tailed errors and outliers, which are common in biomedical traits. By examining heterogeneous effects across quantiles, the proposed framework can reveal distribution-specific patterns that would be masked under mean-based models.

The proposed method accommodates high-dimensional genetic predictors through MCP penalization, enabling simultaneous classification of genetic variants into varying (interaction), constant (main effect), or zero effect categories. Simulation studies demonstrate strong finite-sample performance in both continuous and discrete genetic settings. The performance improves with increasing MAF, consistent with the greater statistical information available for common variants. The method also demonstrates stable performance under the simulation settings considered.

In the real-data application to birth weight, we identified genetic variants with quantile-specific main effects but found no evidence of nonlinear gene-by-mixture interaction with respect to the selected maternal environmental factors. This suggests that, within the studied population, genetic effects may act independently of the considered environmental mixture. However, the absence of detected interaction does not preclude interaction effects in other populations or under different environmental exposures. Moreover, nonlinear interaction effects generally require larger sample sizes to achieve adequate power, and future studies with larger cohorts may reveal additional structure.

The proposed framework involves multiple tuning parameters, including penalty parameters and spline approximation parameters. Although BIC-based selection demonstrated stable empirical performance in our simulation studies, the final model may still be sensitive to tuning parameter specification in certain settings. Developing formal sensitivity analysis procedures and investigating alternative tuning criteria, such as cross-validation or extended BIC, represent important directions for future research. In addition, although the proposed algorithm demonstrated stable empirical convergence across the simulation and real-data analyses, a formal theoretical investigation of identifiability, estimation consistency, oracle properties, and convergence guarantees is yet to be done for the proposed nonconvex penalized quantile optimization framework.

Several limitations of our study merit discussion. First, the identifiability constraints $\|\boldsymbol{\beta}\|_2 = 1$ and $\beta_1 > 0$ may introduce sensitivity to the ordering of environmental variables. Although we mitigate this by refitting models under different orderings and retaining stable results, alternative parameterizations could be explored. Second, the normalization constraint limits direct interpretation of individual loading coefficients. If marginal environmental effects are of primary interest, complementary modeling approaches may be appropriate. Third, although the proposed method provides effective estimation and variable selection, a formal statistical inference for the loading parameter $\beta$ is not yet available. Constructing confidence intervals and hypothesis-testing procedures is challenging due to the combination of nonconvex penalization, spline approximation, identifiability constraints, and nonsmooth quantile loss functions. Developing valid post-selection inference procedures for the proposed quantile VMICM framework represents an important direction for future research[20,28]. In addition, the simulation settings assume independent environmental variables and independently generated genetic predictors. This design provides a controlled environment for evaluating the proposed method, but it is more favorable than many real genomic applications. In practice, SNPs within a gene may be correlated due to linkage disequilibrium, and environmental exposures may also exhibit correlation. Such dependence structures may affect the estimation accuracy and selection stability. Future simulation studies incorporating correlated SNPs and correlated environmental mixtures would provide a more comprehensive evaluation of the proposed framework.

Overall, the proposed quantile varying multi-index coefficient model provides a powerful and flexible tool for investigating heterogeneous genetic and environmental effects in complex trait studies, particularly when environmental exposures act as a mixture.

## Author contributions

The authors confirm their contribution to the paper as follows: study conception and design: Cui Y; methodological development: Guan S, Cui Y; data collection and preparation: Guan S, Cui Y; analysis and interpretation of results: Guan S, Zhang Y, Cui Y; draft manuscript preparation: Guan S; manuscript revision: Guan S, Cui Y. All authors reviewed the results and approved the final version of the manuscript.

## Data availability

The birth-weight data analyzed in this study were obtained from the Gene Environment Association Studies initiative (GENEVA), funded by the Genes, Environment and Health Initiative (GEI). Data access is subject to the policies and approval procedures of the original data repository. The data are not publicly distributed by the authors.

## Acknowledgments

The authors thank the reviewers and editors for their constructive comments, which helped improve the quality and clarity of the manuscript.

## Conflict of interest

The authors declare that they have no conflict of interest.

**Supplementary information** accompanies this paper online at: https://doi.org/10.48130/stati-0026-0015.

## Dates

Received 5 March 2026; Revised 6 June 2026; Accepted 1 July 2026; Published online 31 August 2026

## References

[1] Falconer DS. 1952. The problem of environment and selection. *The American Naturalist* 86(830):293–298
[2] Ottman R. 1996. Gene–environment interaction: definitions and study design. *Preventive Medicine* 25(6):764–770
[3] Ma S, Yang L, Romero R, Cui Y. 2011. Varying coefficient model for gene–environment interaction: a non-linear look. *Bioinformatics* 27(15):2119–2126
[4] Carpenter DO, Arcaro K, Spink DC. 2002. Understanding the human health effects of chemical mixtures. *Environmental Health Perspectives* 110:25–42
[5] Sexton K, Hattis D. 2007. Assessing cumulative health risks from exposure to environmental mixtures-three fundamental questions. *Environmental Health Perspectives* 115(5):825–832
[6] Liu X, Cui Y, Li R. 2016. Partial linear varying multi-index coefficient model for integrative gene-environment interactions. *Statistica Sinica* 26:1037–1060
[7] Guan S, Zhao M, Cui Y. 2023. Variable selection for single-index varying-coefficients models with applications to synergistic G × E interactions. *Electronic Journal of Statistics* 17(1):823–857
[8] Guan S, Liu X, Cui Y. 2025. Variable selection for generalized single-index varying-coefficient models with applications to synergistic G × E interactions. *Mathematics* 13(3):469
[9] Reyes L, Mañalich R. 2005. Long-term consequences of low birth weight. *Kidney International* 68:S107–S111
[10] Magnusson Å, Laivuori H, Loft A, Oldereid NB, Pinborg A, et al. 2021. The association between high birth weight and long-term outcomes—implications for assisted reproductive technologies: a systematic review and meta-analysis. *Frontiers in Pediatrics* 9:675775
[11] Fan J, Li R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360
[12] Zhang CH. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38(2):894–942
[13] Zou H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429
[14] Schumaker L. 2007. *Spline functions: basic theory*. 3rd ed. Cambridge: Cambridge University Press doi: 10.1017/CBO9780511618994
[15] de Boor C. 2001. *A practical guide to splines*. Rev. ed. New York: Springer www.researchgate.net/profile/Carl-De-Boor/publication/200744645 (accessed on 24 August 2026)
[16] Eilers PHC, Marx BD. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2):89–121
[17] Ruppert D, Wand MP, Carroll RJ. 2003. *Semiparametric regression*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511755453
[18] Tang Y, Wang HJ, Zhu Z, Song X. 2012. A unified variable selection approach for varying coefficient models. *Statistica Sinica* 22(2):601–628
[19] Wu C, Zhong PS, Cui Y. 2018. Additive varying-coefficient model for nonlinear gene-environment interactions. *Statistical Applications in Genetics and Molecular Biology* 17(2):20170008
[20] Wang L, Wu Y, Li R. 2012. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* 107(497):214–222
[21] Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6(2):461–464
[22] Galvao AF, Montes-Rojas GV. 2010. Penalized quantile regression for dynamic panel data. *Journal of Statistical Planning and Inference* 140(11):3476–3497
[23] Lee ER, Noh H, Park BU. 2014. Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* 109(505):216–229
[24] Fan TC, Yeo HL, Hung TH, Chang NC, Tang YH, et al. 2025. ST3GAL1 regulates cancer cell migration through crosstalk between EGFR and neuropilin-1 signaling. *The Journal of Biological Chemistry* 301(4):108368
[25] Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, et al. 2016. Genome-wide associations for birth weight and correlations with adult disease. *Nature* 538(7624):248–252
[26] Ma S, Song PXK. 2015. Varying index coefficient models. *Journal of the American Statistical Association* 110(509):341–356
[27] Ma S, Xu S. 2015. Semiparametric nonlinear regression for detecting gene and environment interactions. *Journal of Statistical Planning and Inference* 156:31–47
[28] Belloni A, Chernozhukov V, Kato K. 2019. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association* 114(526):749–758