Statistica Sinica Preprint No: SS-2020-0361	
Title	Data Integration in High Dimension With Multiple Quantiles
Manuscript ID	SS-2020-0361
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0361
Complete List of Authors	Guorong Dai, Ursula Müller and Raymond James Carroll
Corresponding Author	Guorong Dai
E-mail	rondai@stat.tamu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

DATA INTEGRATION IN HIGH DIMENSION WITH MULTIPLE QUANTILES

Guorong Dai[a], Ursula U. Müller[b] and Raymond J. Carroll[b]

[a] *Fudan University*

[b] *Texas A&M University*

Abstract: In this study, we focus on the analysis of high-dimensional data that come from multiple sources ("experiments"), and thus have different, possibly correlated responses, but share the same set of predictors. The measurements of the predictors may be different across experiments. We introduce a new regression approach, using multiple quantiles to select those predictors that affect any of the responses at any quantile level and to estimate the nonzero parameters. Our approach differs from established methods by being able to handle heterogeneity in data sets and heavy-tailed error distributions, two difficulties that are often encountered in complex data scenarios. Our estimator minimizes a penalized objective function that aggregates the data from the different experiments. We establish the model selection consistency and asymptotic normality of the estimator. In addition, we present an information criterion that can be used for consistent model selection. Simulations and two data applications illustrate the advantages of our method in recovering the underlying regression

[a]Previously at Texas A&M University during the preparation of this work.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

models. These advantages come from taking the group structure induced by the predictors across experiments and the quantile levels into account.

Key words and phrases: Data integration; High dimensional data; Information criterion; Penalized quantile regression.

1. Introduction

To set the stage for this work on data integration (DI), consider K data sets from K different populations, where K is some fixed number, with linear regression models

$$Y_k = X_k^{\mathrm{T}} \alpha_k^* + U_k \quad (k = 1, \ldots, K). \tag{1.1}$$

Here, Y_k is a scalar response, X_k is a p-dimensional predictor, α_k^* is a p-dimensional parameter vector, and U_k is the error term. Zellner (1962) referred to this set of models as *seemingly unrelated regressions* and proposed the idea of estimating the regression parameters simultaneously using a generalized least squares method. The responses in model (1.1) are different, but dependent. The predictors are the same in the K data sets, but not their values. This can occur, for example, if individuals are assessed through various responses from different experiments and the predictor values are measured in different ways (Gao and Carroll, 2017).

Model (1.1), with the assumption that $E(U_k \mid X_k) = 0$, can also be

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

written as a heterogenous linear regression model,

$$E(Y_k - X_k^{\mathrm{T}}\alpha_k^* \mid X_k) = 0 \quad (k = 1, \ldots, K).$$

We consider the same scenario, but pursue a different approach. Instead of modeling the conditional mean of the response given the covariates, we assume heterogeneous linear regression models for the conditional quantiles $Q_{\tau_m}(X_k)$ at various quantile levels τ_m $(m = 1, \ldots, M)$; that is

$$E\{I(Y_k \leq X_k^{\mathrm{T}}\theta_{km}^*) - \tau_m \mid X_k\} = 0 \quad (k = 1, \ldots, K), \tag{1.2}$$

where $I(\cdot)$ is the indicator function and θ_{km}^* is a p-dimensional parameter vector. This is equivalent to

$$\mathrm{pr}(Y_k \leq X_k^{\mathrm{T}}\theta_{km}^* \mid X_k) = \mathrm{pr}\{Y_k \leq Q_{\tau_m}(X_k) \mid X_k\} = \tau_m$$

$(m = 1, \ldots, M;\ k = 1, \ldots, K)$. We are interested in the high-dimensional data situation and, therefore, let the dimension $p = p_n$ of the parameter vector tend to infinity as the sample size n increases. In addition, we assume that the data are sparse, that is, most of the parameters are zero, which means that only a fraction of the predictors affect the responses.

An important goal is to identify the relevant predictors. One possible approach is to aggregate each predictor's effect in all experiments by forming groups. In our scenario, all responses share the same set of predictors.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Hence, we have a natural group structure: the parameters of different quantiles and experiments that belong to the same predictor constitute a group; see Gao and Carroll (2017), who develop a group penalized estimation method using a pseudolikelihood. To handle the unspecified dependence between the responses in the K experiments, they pool the marginal likelihoods and impose an L_2-group penalization on the grouped parameters. The group penalty was introduced in a 1999 Australian National University Ph.D. thesis by S. Bakin, and then applied to group selection questions by Yuan and Lin (2006). Gao and Carroll (2017) use it to select predictors that are influential in any of the experiments. Their main tool is the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001). In addition, Gao and Carroll (2017) use the concept of the Bayesian information criterion to develop a pseudolikelihood information criterion that applies to high-dimensional scenarios. Their pseudolikelihood approach is an important advance, and useful when the distribution of the error can be modeled parametrically, which is not assumed in our case.

In this study, we use a linear quantile regression approach based on model (1.2); that is, we do not work with a likelihood, but with a different objective function. Quantile regression was introduced by Koenker and Bassett (1978); see also Koenker (2005). In contrast to classical regression, it provides a global picture of the predictors' effects on the distri-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

bution of the responses, and it is robust to heavy-tailed distributions. In high-dimensional settings, Belloni and Chernozhukov (2011) studied linear quantile regression with a Lasso penalty, Wang et al. (2012) proved the selection consistency of linear quantile regression with nonconvex penalty functions, and Sherwood and Wang (2016) derived the asymptotic properties of partially linear additive quantile regression with a nonconvex penalty. In addition to these works on single quantile regression, Zou and Yuan (2008a) introduced a composite quantile regression approach that considers multiple quantiles simultaneously. They assumed that the slopes are the same across quantiles, and used the adaptive Lasso penalty from Zou (2006). Their method shares the oracle properties proposed in Fan and Li (2001). In another paper, which focuses on computation and not on theoretical properties, Zou and Yuan (2008b) propose a related approach for the heterogeneous scenario, that is, when the covariates and errors are dependent so that the slopes vary across quantiles. They consider multiple responses, but model just one single quantile for each response. Their method is able to detect nonzero slopes simultaneously. The two 2008 studies by Zou and Yuan examine settings with a fixed number of parameters. Recently, the composite quantile approach of Zou and Yuan (2008a) was extended to high-dimensional scenarios by Gu and Zou (2020). These authors assume that the slopes are the same across quantiles, that is, homogene-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

ity. As such, their approach does not apply to the heterogeneous models investigated by our research. Fan et al. (2016) studied quantile regression with multiple responses under a "transnormal" assumption, which requires that the responses and predictors can be transformed into a multivariate normal variable using marginal monotone functions. This is not required in our model. In Table 3 of Section 4, we consider a simulation setting with some binary predictors, which violates this assumption. While we are interested in identifying relevant predictors, Fan et al. (2016) focus on predicting responses and estimating correlation matrices.

Our goal is simultaneous variable selection with multiple quantiles across K experiments. To take account of the unknown dependence structure between the responses in the different experiments, we integrate the data by summing their quantile loss functions. This is analogous to Gao and Carroll (2017), who pool the likelihood functions. In addition, similarly to Sherwood and Wang (2016), who also conduct variable selection with multiple quantiles, we apply a nonconvex penalty on the L_1-norm of the coefficients related to each predictor. This represents the overall strength of the predictor across multiple experiments and quantiles. The penalty function takes the group structure into account, and excludes covariates that have no impact on any of the responses at any of the quantile levels. Moreover, the L_1-norm is computationally convenient in quantile regression settings, ow-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

ing to the work of Peng and Wang (2015), who provide a "Quick Iterative Coordinate Descent" algorithm for solving nonconvex penalized quantile regression in high dimensions with no group structure. With modifications, their algorithm can be adapted to our approach; see Section 4.

Our work is largely motivated by the widespread existence of heterogeneity in complex data sets (Yu et al., 2003; Wang et al., 2012; Lee et al., 2014), such as the liver toxicity data set and the financial index data analyzed in Section 5. Classical regression focuses on the conditional mean or on one single conditional quantile of Y_k given X_k ($k = 1, \ldots, K$). In contrast, a major advantage of our approach is its ability to identify predictors in heterogeneous models that affect the responses at one or more quantile levels, but not necessarily globally. When the random errors in the data-generating mechanism have a heavy-tailed distribution, for example, a t-distribution with a small number of degrees of freedom, quantile based methods have a better estimation accuracy than that of competing approaches that use the quadratic loss function. Despite these clear advantages, multiple quantile regression for dependent data that originate from different sources has not, to the best of our knowledge, been studied in the literature. We also cover the high-dimensional data scenario by adding a nonconvex group penalty term. We establish the selection consistency and asymptotic normality of our estimator in this quite general setting under

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

mild assumptions. Additionally we propose a multiple quantile Bayesian information criterion (MQBIC) based on pooled check functions, which is an extension of the Bayesian information criterion for linear quantile regression (Lee et al., 2014) to the multiple-experiment scenario. Similarly to the pseudolikelihood information criterion in Gao and Carroll (2017), the MQBIC permits consistent model selection (see Section 3) and choice of the tuning parameter for the penalized estimator (see Section 4).

The main contribution of this study is the introduction of quantile-based methods to the high-dimensional scenario of DI. We propose a penalized estimation process and an information criterion, which identify the covariates that affect any of the responses at any of the quantile levels. Our method enjoys robustness, and can be applied to the complex scenario with heterogeneous data and dependent responses.

The rest of this article is organized as follows. In Section 2, we introduce our objective function, which involves a nonconvex group penalization term, and present the oracle properties of the estimator. The MQBIC is presented in Section 3, and its model selection consistency is established. In Section 4, we compare our method with other approaches using simulations. Our method is illustrated in Section 5 by means of empirical data examples. Section 6 gives a brief discussion of further questions. All proofs, as well as additional simulation results, are provided in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

For notational clarity, we assume in the following that the sample sizes and the quantile levels are the same in every experiment. The conclusions and methods are essentially the same if we drop these assumptions.

2. Penalized estimator

Throughout this article, we use the capital letter C to represent a generic constant, including C_1, C_2 and so on. We write I_m for the $m \times m$ identity matrix. The symbols $\|\cdot\|_1$ and $\|\cdot\|$ refer to the L_1- and L_2-norms of a vector, and $\otimes$ denotes the Kronecker product.

Our conditional quantile regression model is $Q_{\tau_m}(X_k) = X_k^{\mathrm{T}}\theta_{km}^*$, with ordered levels $0 < \tau_1 < \tau_2 < \cdots < \tau_M < 1$. We can set the first column of X_k to be $(1, \ldots, 1)^{\mathrm{T}}$ so that the model contains intercept terms. For notational convenience, we assume the intercepts all equal zero. The number of predictors $p = p_n$ tends to infinity as the sample size n increases.

For $k = 1, \ldots, K$ and $i = 1, \ldots, n$, we consider n independent copies $\{Y_{ki}, X_{ki}\}$, with $X_{ki} = (X_{ki1}, \ldots, X_{kip_n})^{\mathrm{T}}$ of the base observation $\{Y_k, X_k\}$ from model (1.1). Here, we use three subscripts to locate the predictors, that is, X_{kij} represents the jth component of the ith observation in the kth experiment. We write $X_{k\cdot j} = (X_{k1j}, \ldots, X_{knj})^{\mathrm{T}}$ for the vector. The data are summarized in Table 1.

The regression parameters θ_{km}^* $(k = 1, \ldots, K,\ m = 1, \ldots, M)$ are as-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

sumed to be sparse; that is, most of the components of θ^*_{km} are zero. Write $\theta^{*(j)}$ for the parameters related to the jth predictor ($j = 1, \ldots, p_n$) across the K experiments and the M quantile levels, that is,

$$\theta^{*(j)} = (\theta^*_{11j}, \ldots, \theta^*_{1Mj}, \ldots, \theta^*_{K1j}, \ldots, \theta^*_{KMj})^{\mathrm{T}}.$$

We want to select the predictors that have an effect on any of the responses, that is, we want to specify the set $\mathcal{A} = \{j : 1 \leq j \leq p_n, \|\theta^{*(j)}\| > 0\}$. Without loss of generality, let $\mathcal{A} = \{1, 2, \ldots, q_n\}$, that is, only the first q_n predictors have nonzero parameters. We assume that q_n tends to infinity as n and p_n increase. For convenience of notation, we use the letter a at the end of a subscript to refer to subvectors or submatrices that consist of components with subscripts in $\mathcal{A}$. For example, $X_{kia} = (X_{ki1}, \ldots, X_{kiq_n})^{\mathrm{T}}$, $X_{k\cdot a} = (X_{k1a}, \ldots, X_{kna})^{\mathrm{T}}$, and $\theta^*_{kma} = (\theta^*_{km1}, \ldots, \theta^*_{kmq_n})^{\mathrm{T}}$.

The dependence between the experiments is unspecified. To integrate the data, we therefore sum the quantile loss functions across the K experiments and the M quantiles,

$$\ell_n(\theta) = n^{-1} \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki}^{\mathrm{T}} \theta_{km}). \tag{2.1}$$

In the above, $\rho_m(x) = x\{\tau_m - I(x < 0)\}$ is the check function and

$$\theta = (\theta_{11}^{\mathrm{T}}, \ldots, \theta_{1M}^{\mathrm{T}}, \ldots, \theta_{K1}^{\mathrm{T}}, \ldots, \theta_{KM}^{\mathrm{T}})^{\mathrm{T}}$$

is a parameter vector. Loss functions analogous to (2.1) with $K = 1$ or

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 1: Data structure of multiple experiments

	Experiment 1	...	Experiment K
τ_1	$\theta^*_{11} = (\theta^*_{111}, \ldots, \theta^*_{11p_n})^{\mathrm{T}}$	...	$\theta^*_{K1} = (\theta^*_{K11}, \ldots, \theta^*_{K1p_n})^{\mathrm{T}}$
$\vdots$	$\vdots$		$\vdots$
τ_M	$\theta^*_{1M} = (\theta^*_{1M1}, \ldots, \theta^*_{1Mp_n})^{\mathrm{T}}$	...	$\theta^*_{KM} = (\theta^*_{KM1}, \ldots, \theta^*_{KMp_n})^{\mathrm{T}}$
$i = 1$	Y_{11}, $X_{11} = (X_{111}, \ldots, X_{11p_n})^{\mathrm{T}}$	...	Y_{K1}, $X_{K1} = (X_{K11}, \ldots, X_{K1p_n})^{\mathrm{T}}$
$\vdots$	$\vdots$		$\vdots$
$i = n$	Y_{1n}, $X_{1n} = (X_{1n1}, \ldots, X_{1np_n})^{\mathrm{T}}$	...	Y_{Kn}, $X_{Kn} = (X_{Kn1}, \ldots, X_{Knp_n})^{\mathrm{T}}$

Parameters related to $\tau_1, \ldots, \tau_M$ and observations $i = 1$ to n.

$M = 1$ were used for low-dimensional linear models by Zou and Yuan (2008b). The main difference between (2.1) and the composite quantile loss function considered in Zou and Yuan (2008a) and Gu and Zou (2020) is that we allow $\theta_{km} \neq \theta_{km'}$ $(1 \leq m \neq m' \leq M; k = 1, \ldots, K)$, that is, different slopes.

To select the predictors that affect any of the responses, a nonconvex penalty function $\Omega_{\lambda_n}(\cdot)$ with a tuning parameter λ_n is imposed on the overall impact of each predictor. That impact is represented by the L_1-norm of the vector $\theta^{(j)}$, which contains the parameters of the jth predictor in the K

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

experiments. This gives the overall objective function

$$\Gamma_{\lambda_n}(\theta) = \ell_n(\theta) + \sum_{j=1}^{p_n} \Omega_{\lambda_n}(\|\theta^{(j)}\|_1). \tag{2.2}$$

Our estimator is obtained by minimizing $\Gamma_{\lambda_n}(\theta)$. We use the SCAD penalty function (Fan and Li, 2001)

$$\begin{aligned}\Omega_{\lambda_n}(x) = &\lambda_n x I(0 \le x \le \lambda_n) + \\ &\frac{a\lambda_n x - (x^2 + \lambda_n^2)/2}{a-1} I(\lambda_n < x < a\lambda_n) + \frac{(a+1)\lambda_n^2}{2} I(x \ge a\lambda_n),\end{aligned}$$

where a is a constant that is usually set to 3.7 (Fan and Li, 2001). Before stating the asymptotic properties of our estimator, we make the following assumptions.

Assumption 1. *There is a constant $C > 0$ such that $|X_{kij}| \le C$ for every $k = 1, \ldots, K$, $i = 1, \ldots, n$, and $j = 1, \ldots, p_n$.*

Assumption 2. *For every $k = 1, \ldots, K$, there are positive constants C_1 and C_2 such that*

$$C_1 \le \lambda_{\min}(n^{-1} X_{k\cdot a}^{\mathrm{T}} X_{k\cdot a}) \le \lambda_{\max}(n^{-1} X_{k\cdot a}^{\mathrm{T}} X_{k\cdot a}) \le C_2,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ stand for the smallest and largest eigenvalues, respectively. In addition $X_{k\cdot a}$ and $(Y_{k1}, \ldots, Y_{kn})^{\mathrm{T}}$ are in "general positions," which is an identifiability condition that guarantees that a solution to the quantile regression problem exists (Koenker, 2005, Section 2.2.2).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Assumption 3. *For every $k = 1, \ldots, K$ and $m = 1, \ldots, M$, the conditional probability density $f_{km}(\cdot \mid x)$ of $\varepsilon_{km} = Y_k - X_k^{\mathrm{T}}\theta_{km}^*$ given $X_k = x$ is uniformly bounded and bounded away from zero in a neighborhood of zero, and has a derivative $f'_{km}(\cdot \mid x)$, which is uniformly bounded in a neighborhood of zero.*

Assumption 4. *The true model size satisfies $q_n = O(n^{c_1})$, for some $0 \leq c_1 < 1/2$.*

Assumption 5. *There are positive constants c_2 and C such that $2c_1 < c_2 \leq 1$, where c_1 is the constant introduced in Assumption 4, and*

$$n^{(1-c_2)/2}\min_{1\leq j\leq q_n}\|\theta^{*(j)}\|_1 \geq C.$$

Assumptions 1 and 2 guarantee the good behavior of the design matrices. Assumption 1 can be relaxed by letting the covariate values increase to infinity at a certain slow rate. We work with it mainly for reasons of simplicity and clarity; see, for example, the closely related articles by Wang et al. (2012), Lee et al. (2014), and Sherwood and Wang (2016), who also limit themselves to the case with bounded covariates. The conditions in Assumption 3 concern the unknown distribution of the random errors. They are considerably weaker than assuming a specific parametric model for the error distribution. Assumption 4 regulates the growth rate of the true model size. This is a standard assumption used to establish the asymptotic properties of sparse estimators in linear models with a diverging number of

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

parameters; see, for example, Wang et al. (2012) and Lee et al. (2014). In addition, it is weaker than the condition $q_n = o(n^{1/5})$ required by Gao and Carroll (2017). Assumption 5 excludes situations where the nonzero parameters decay too fast. Conditions similar to Assumptions 1–5 are required in Wang et al. (2012) for single experiments with a single quantile.

The oracle estimator $\widehat{\theta}$ is defined as the minimizer of $\ell_n(\theta)$ that knows that the first q_n components of θ are nonzero and that the others are zero, that is, $\|\widehat{\theta}^{(j)}\| = 0$, for $q_n < j \leq p_n$. The following theorem provides the model selection consistency of our estimator. More precisely, we show that, with probability tending to one, the oracle estimator can be obtained using our approach of minimizing the objective function $\Gamma_{\lambda_n}(\theta)$.

Theorem 1. *Let $S(\lambda_n)$ denote the set of local minimizers of $\Gamma_{\lambda_n}(\theta)$, and $\widehat{\theta}$denote the oracle estimator. Under Assumptions 1–5,* $\mathrm{pr}\{\widehat{\theta} \in S(\lambda_n)\} \to 1$ *as $n \to \infty$, if $\lambda_n = o\{n^{-(1-c_2)/2}\}$, $n^{-1/2} q_n = o(\lambda_n)$, and $n^{-1} log\, p_n = o(\lambda_n^2)$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Before stating Theorem 2, we introduce some notation. We write

$$\varepsilon_{kmi} = Y_{ki} - X_{ki}^{\mathrm{T}}\theta_{km}^{*}, \quad \varepsilon_{km} = (\varepsilon_{km1}, \ldots, \varepsilon_{kmn})^{\mathrm{T}},$$

$$\varepsilon = (\varepsilon_{11}^{\mathrm{T}}, \ldots, \varepsilon_{1M}^{\mathrm{T}}, \ldots, \varepsilon_{K1}^{\mathrm{T}}, \ldots, \varepsilon_{KM}^{\mathrm{T}})^{\mathrm{T}},$$

$$\psi_{kmi}(\varepsilon) = \tau_m - I(\varepsilon_{kmi} < 0), \quad \psi_{nkm}(\varepsilon) = \{\psi_{km1}(\varepsilon), \ldots, \psi_{kmn}(\varepsilon)\}^{\mathrm{T}},$$

$$\psi_{nk}(\varepsilon) = \{\psi_{nk1}(\varepsilon)^{\mathrm{T}}, \ldots, \psi_{nkM}(\varepsilon)^{\mathrm{T}}\}^{\mathrm{T}}, \quad \psi_n(\varepsilon) = \{\psi_{n1}(\varepsilon)^{\mathrm{T}}, \ldots, \psi_{nK}(\varepsilon)^{\mathrm{T}}\}^{\mathrm{T}},$$

$$H_n = E\{\psi_n(\varepsilon)\psi_n(\varepsilon)^{\mathrm{T}} \mid \mathcal{X}\} \text{ with } \mathcal{X} = \{X_{ki} : k = 1, \ldots, K, i = 1, \ldots, n\},$$

where $k = 1, \ldots, K$, $m = 1 \ldots, M$, and $i = 1, \ldots, n$. Further, we set

$$B_{nkm} = \mathrm{diag}\{f_{km}(0 \mid X_{k1}), \ldots, f_{km}(0 \mid X_{kn})\},$$

$$B_n = \mathrm{diag}(B_{n1}, \ldots, B_{nK}) \text{ with } B_{nk} = \mathrm{diag}(B_{nk1}, \ldots, B_{nkM}),$$

$$\theta_a^{*} = (\theta_{11a}^{*\mathrm{T}}, \ldots, \theta_{1Ma}^{*\mathrm{T}}, \ldots, \theta_{K1a}^{*\mathrm{T}}, \ldots, \theta_{KMa}^{*\mathrm{T}})^{\mathrm{T}},$$

$$\widehat{\theta}_{kma} = (\widehat{\theta}_{km1}, \ldots, \widehat{\theta}_{kmq_n})^{\mathrm{T}}, \ \widehat{\theta}_a = (\widehat{\theta}_{11a}^{\mathrm{T}}, \ldots, \widehat{\theta}_{1Ma}^{\mathrm{T}}, \ldots, \widehat{\theta}_{K1a}^{\mathrm{T}}, \ldots, \widehat{\theta}_{KMa}^{\mathrm{T}})^{\mathrm{T}},$$

$$X_a = \mathrm{diag}(I_M \otimes X_{1\cdot a}, \ldots, I_M \otimes X_{K\cdot a}), \ R_n = n^{-1}X_a^{\mathrm{T}}B_nX_a,$$

$$S_n = n^{-1}X_a^{\mathrm{T}}H_nX_a, \ \Sigma_n = R_n^{-1}S_nR_n^{-1}.$$

The next theorem gives the asymptotic normality of low-dimensional projections of the nonzero part $\widehat{\theta}_a$ of the oracle estimator $\widehat{\theta}$ from Theorem 1. An illustration of the result with histogram plots (for two simulation scenarios from Section 4) is provided in Section S3 of the Supplementary Material.

Theorem 2. *Let $q_n^{*} = q_n \times M \times K$. Consider an $s \times q_n^{*}$ matrix A_n with s*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

fixed and $A_n A_n^{\mathrm{T}} \to G$*, a positive-definite matrix. Then*

$$n^{1/2} A_n \Sigma_n^{-1/2} (\widehat{\theta}_a - \theta_a^*) \to N(\mathbf{0}, G) \quad (n \to \infty)$$

in distribution, provided Assumptions 1–4 are satisfied and $\lambda_{\min}(S_n)$ *is uniformly bounded away from zero.*

Theorems 1 and 2 establish the model selection consistency and asymptotic normality of our estimator when experiments are correlated. This shows that it is reasonable to aggregate information from multiple experiments, rather than ignoring the correlation and analyzing each experiment separately.

3. Multiple quantile Bayesian information criterion

To select the correct model, we use an information criterion that balances the goodness-of-fit and the complexity of a model. By applying this information criterion to a set of competing models, the true model can be identified with probability approaching one. In the context of quantile regression, Lee et al. (2014) develop a Bayesian information criterion with a diverging number of predictors. Their method considers one single quantile and deals with data from one single experiment. We use a generalized version of the criterion, now based on multiple quantiles and on data from several experiments, which improves its ability to select the correct model.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

The *multiple quantile* Bayesian information criterion of a submodel $\mathcal{D} \subset \{1, 2, \ldots, p_n\}$ is

$$\begin{aligned} \text{MQBIC}(\mathcal{D}) = &\log\{\textstyle\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}})\} \\ &+ (2n)^{-1}|\mathcal{D}|T_n \log n, \end{aligned} \tag{3.1}$$

where $\widehat{\theta}_{km\mathcal{D}} = \arg\min_{\theta\in\mathbb{R}^{|\mathcal{D}|}} \sum_{i=1}^{n} \rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}}\theta)$, for $k = 1, \ldots, K$ and $m = 1, \ldots, M$, $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$, and T_n is a sequence of positive constants diverging to infinity as n increases. The notation $X_{ki\mathcal{D}}$ refers to the subvectors of $X_{ki\cdot}$ that contain only components with subscripts in $\mathcal{D}$. We set an upper bound on the cardinality of competing models, say d_n, and search for the best of the submodels with a cardinality smaller or equal to d_n. Define $\mathcal{D}^* = \{1, 2, \ldots, q_n\}$ as the subset of $\{1, \ldots, p_n\}$ corresponding to the true model, and $\mathcal{M} = \{\mathcal{D} \subset \{1, \ldots, p_n\} : |\mathcal{D}| \leq d_n\}$ as the set of all competing models. The first part of the MQBIC represents the goodness-of-fit, and the second term is a penalty on the model complexity. To guarantee the model selection consistency of the MQBIC, we need the following assumptions, in addition to some of the assumptions from Section 2.

Assumption 6. *For every $k = 1, \ldots, K$, there are constants $0 < C_3 \leq C_4$ such that for any $\mathcal{D} \subset \{1, \ldots, p_n\}$, the matrix $X_{k\cdot\mathcal{D}} = (X_{k1\mathcal{D}}, \ldots, X_{kn\mathcal{D}})^{\mathrm{T}}$*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

satisfies

$$C_3 \leq \min_{|\mathcal{D}|\leq 2d_n} \lambda_{\min}(n^{-1}X_{k\cdot\mathcal{D}}^{\mathrm{T}}X_{k\cdot\mathcal{D}}) \leq \max_{|\mathcal{D}|\leq 2d_n} \lambda_{\max}(n^{-1}X_{k\cdot\mathcal{D}}^{\mathrm{T}}X_{k\cdot\mathcal{D}}) \leq C_4.$$

Assumption 7. *The full model size p_n is of order $p_n = O(n^{c_3})$, for some $c_3 > 0$; the true model size q_n is fixed, $q_n = q$, and satisfies $q \leq d_n = O(n^{c_4})$, for some $0 < c_4 < 1/2$.*

Assumption 8. *The sequence T_n in the definition (3.1) satisfies $T_n \to \infty$ and $n^{-1}T_n log\, n \to 0$.*

Assumption 9. *The average $n^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(\varepsilon_{kmi})$ of the check functions is bounded away from zero with probability tending to one.*

Assumption 6 extends Assumption 2 for the true model to all candidate models. This is common for scenarios with more regression parameters than observations, that is, $p_n > n$. In Assumption 7, the true model size is fixed because of a technical difficulty in handling the maximum of $|\mathcal{D}\backslash\mathcal{D}^*|^{-1}|n^{-1}\sum_{i=1}^{n}\{\rho_m(Y_{ki} - X_{ki\mathcal{D}}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}}) - \rho_m(Y_{ki} - X_{ki\mathcal{D}^*}^{\mathrm{T}}\widehat{\theta}_{km\mathcal{D}^*})\}|$ over the set of overfitted models $\{\mathcal{D} \in \mathcal{M} : \mathcal{D}^* \subset \mathcal{D}, \mathcal{D} \neq \mathcal{D}^*\}$ (Lee et al., 2014). Assumption 8 regulates the growth rate of the sequence T_n. Assumption 9 is made for convenience in the proofs because $n^{-1}\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(\varepsilon_{kmi})$ appears in denominators.

In the following theorem, we show that the true model has, with proba-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

bility tending to one, the smallest MQBIC value among all candidate models.

Theorem 3. *If Assumptions 1, 3, and 6–9 hold, then with probability tending to one, the true model can be selected by minimizing the MQBIC, that is,*

$$\lim_{n\to\infty}\mathrm{pr}\{\min_{\mathcal{D}\in(\mathcal{M}\setminus\{\mathcal{D}^*\})} MQBIC(\mathcal{D}) > MQBIC(\mathcal{D}^*)\} = 1.$$

Theorem 3 establishes the model selection consistency of the MQBIC for data from multiple dependent sources, which provides another approach to identifying the true underlying model. In the MQBIC approach, estimation and model selection are separate processes. This differs from minimizing the objective function in Section 2, which is a one-step procedure. The main advantage of the MQBIC is that we can use it to select the tuning parameter λ_n for the penalized estimation process in Section 2, which is computationally more efficient than cross-validation. The details are given in Section 4.

4. Simulations

In this section, we study the numerical performance of our estimators. We use the objective function (2.2) with $M = 5$ quantiles, $\tau_1 = 1/6, \tau_2 = 2/6, \ldots,$ and $\tau_5 = 5/6$, and study two different group structures, namely,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

complete and incomplete grouping. Complete grouping means that parameters of the same predictor can only be either all zero or all nonzero, while in the incomplete case, a group may contain both zero and nonzero parameters.

In both cases, the number of experiments is $K = 2$, and the sample size and the number of predictors are $(n, p) = (100, 100)$, $(100, 200)$, or $(200, 1000)$. The nonzero parameters are drawn independently from a uniform distribution on $[0.05, 1]$. For $K = 1, 2$, we generate independent random vectors X'_{ki}, for $i = 1, \ldots, n$, from a p-dimensional multivariate normal distribution with mean zero and a covariance matrix with (i, j)th component of $0.5^{|i-j|}$, for $1 \leq i, j \leq p$. The predictors X_{ki} for the different scenarios described below are transformations of X'_{ki}. For $i = 1, \ldots, n$, the error terms $(\xi_{1i}, \xi_{2i})^{\mathrm{T}}$ are drawn independently from a bivariate normal distribution with mean zero, or from a bivariate t-distribution with three degrees of freedom. The covariance matrix of (ξ_{1i}, ξ_{2i}) is Σ, with entries $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = 0.7$.

To minimize the objective function (2.2) with a fixed $\lambda_n = \lambda$, we use an algorithm developed by Peng and Wang (2015) for penalized quantile regression, modified for our scenario with multiple quantiles and experiments. We first apply the "Majorize-Minimization" algorithm with an initial value $\widehat{\theta}(0) = \mathbf{0}$. Let $\widehat{\theta}(r-1)$ denote the result from the $(r-1)$th iteration. Ac-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

cording to Section 3.1 of Peng and Wang (2015), the objective function (2.2) is majorized by

$$\ell_n(\theta) + \sum_{j=1}^{p} \Omega'_\lambda(\|\widehat{\theta}^{(j)}(r-1)\|_1+)\|\theta^{(j)}\|_1 = \sum_{k=1}^{K}\sum_{m=1}^{M} Q^{(k,m)}_{\widehat{\theta}(r-1)}(\theta_{km}) \quad (4.1)$$

at the rth iteration. Here, $\Omega'_\lambda(\cdot)$ is the derivative of $\Omega_\lambda(\cdot)$ with $\Omega'_\lambda(x_0+) = \lim_{x\downarrow x_0} \Omega'_\lambda(x)$, and

$$Q^{(k,m)}_{\widehat{\theta}(r-1)}(\theta_{km}) = n^{-1}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki}^{\mathrm{T}}\theta_{km}) + \sum_{j=1}^{p}\Omega'_\lambda(\|\widehat{\theta}^{(j)}(r-1)\|_1+)|\theta_{kmj}|.$$

The minimization of the majorizing function (4.1) can therefore be done by minimizing $Q^{(k,m)}_{\widehat{\theta}(r-1)}(\theta_{km})$ $(k = 1, \ldots, K; m = 1, \ldots, M)$ separately for each (k, m) using the "coordinate descent algorithm," which involves calculating weighted medians using the "quicksort" algorithm. A detailed description can be found in Section 3.2 of Peng and Wang (2015). We update $\widehat{\theta}(r-1)$ by

$$\widehat{\theta}(r) = \arg\min_\theta \sum_{k=1}^{K}\sum_{m=1}^{M} Q^{(k,m)}_{\widehat{\theta}(r-1)}(\theta_{km}),$$

and repeat this process until convergence. This yields the minimizer $\widehat{\theta}_{\lambda,km} = (\widehat{\theta}_{\lambda,km1}, \ldots, \widehat{\theta}_{\lambda,kmp})^{\mathrm{T}}$ $(k = 1, \ldots, K; m = 1, \ldots, M)$ of (2.2), with $\lambda_n = \lambda$. The tuning parameter is chosen from a grid Λ. For $\lambda \in \Lambda$, let $\mathcal{D}_\lambda = \{j :$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

$1 \le j \le p, \sum_{k=1}^{K}\sum_{m=1}^{M} |\widehat{\theta}_{\lambda,kmj}| > 0\}$. For the final estimator, we use

$$\widehat{\lambda} = \arg\min_{\lambda\in\Lambda}\big[\log\{\textstyle\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki}^{\mathrm{T}}\widehat{\theta}_{\lambda,km})\} + (2n)^{-1}|\mathcal{D}_\lambda|(\log n)T\big], \tag{4.2}$$

which minimizes the MQBIC. This approach adapts criterion (2.10) in Lee et al. (2014) to multiple quantile levels and experiments. They recommends $T = C\log p$ and their simulation results show that this type of information criterion tends to underfit models slightly. As such, we consider $T = (\log p)/3$ or $(\log p)/6$, and examine how this affects the performance of the method.

In each scenario, we record the following three indices:

1. Positive selection rate (PSR): the proportion of selected predictors that affect any quantile of any response. Then, formally, PSR $= |\widehat{\mathcal{A}} \cap \mathcal{A}|/|\mathcal{A}|$ with $\mathcal{A} = \{j : 1 \le j \le p, \|\theta^{*(j)}\| > 0\}$ and $\widehat{\mathcal{A}} = \{j : 1 \le j \le p, \|\widehat{\theta}^{(j)}\| > 0\}$.

2. False discovery rate (FDR): the proportion of selected predictors that affect no response, that is, $|\widehat{\mathcal{A}} \cap \mathcal{A}^c|/|\mathcal{A}^c|$.

3. Absolute error (AE): the absolute estimation error $(KM)^{-1}\|\widehat{\theta} - \theta^*\|_1$.

Our DI approach is compared with the following three methods:

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

(a) Combined analysis based on the τth quantile (CA-τ). This method considers only one quantile τ. The data from the K experiments are analyzed separately, then the K sets of selected predictors are merged. We find that in most cases, the CA-τ method selects more unimportant predictors than does our DI approach. Hence, the FDRs will rise even further when the results from different quantile levels are combined.

(b) Sparse canonical correlation analysis (SCCA) by Witten and Tibshirani (2009). To adapt this method for DI problems, we determine the sparse vectors

$$\{\widehat{w}_1, \widehat{w}_2\} = \arg\max_{\{w_1,w_2\}} \sum_{i=1}^{n} w_1^{\mathrm{T}} \widetilde{X}_i \widetilde{Y}_i^{\mathrm{T}} w_2$$

that satisfy

$$\|w_1\| \le 1, \|w_2\| \le 1, \|w_1\|_1 \le c_1, \|w_2\| \le c_2,$$

where c_1 and c_2 are some approriate tuning parameters and

$$\widetilde{X}_i = (X_{1i}^{\mathrm{T}}, X_{2i}^{\mathrm{T}}, \ldots, X_{Ki}^{\mathrm{T}})^{\mathrm{T}},\ \widetilde{Y}_i = (Y_{1i}, Y_{2i}, \ldots, Y_{Ki})^{\mathrm{T}} \quad (i = 1, \ldots, n).$$

Then, we view the predictors corresponding to the nonzero components of $\widehat{w}_1$ as the selected ones. Details of the implementation can be found in Witten and Tibshirani (2009). This approach does not generate estimators for the regression parameters, which explains why we have no values for the absolute errors (AE) in the tables below.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

(c) The DI method based on the least squares regression and the SCAD penalty (DI-LS), proposed by Gao and Carroll (2017), which we explained in the introduction. This method minimizes an objective function similar to (2.2), but with $M = 1$ and $\ell_n(\cdot)$ replaced by the quadratic loss function, and the SCAD penalty is applied to $\|\theta^{(j)}\|$ $(j = 1, \ldots, p_n)$. When calculating absolute errors, the target parameter is the vector $\widetilde{\theta}_k$ that satisfies $E(Y_k \mid X_k) = X_k^{\mathrm{T}}\widetilde{\theta}_k$ $(k = 1, \ldots, K)$.

The value of T only plays a role in minimizing the MQBIC. Hence, it only affects our DI and the CA-τ methods in the following tables. The tuning parameters in SCCA and DI-LS are selected using 10-fold cross-validation.

Tables 2 and 3 show the simulation results for a scenario with normal errors and a complete group structure. The nonzero parameters are α^*_{11}, α^*_{16}, $\alpha^*_{1(12)}$, $\alpha^*_{1(15)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{26}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$. Let $\Phi(\cdot)$ be the distribution function of a standard normal variable. For $k = 1, 2$ and $i = 1, \ldots, n$, the predictors in Table 2 are $X_{ki3} = \Phi(X'_{ki3})$ and $X_{kij} = X'_{kij}$ $(j \neq 3)$, while those in Table 3 are $X_{ki3} = \Phi(X'_{ki3})$ and $X_{kij} = I(X'_{kij} > 0)$ $(j = 20, \ldots, 25)$, and $X_{kij} = X'_{kij}$ otherwise. The binary predictors in Table 3 violate the transnormal assumption in Fan et al. (2016). The responses are $Y_{ki} = X_{ki}^{\mathrm{T}}\alpha^*_k + 0.7\xi_{ki}X_{ki3}$. Our DI method achieves the highest PSRs

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 2: Positive selection rates, false discovery rates, and absolute errors of our data-integration method and the competing approaches for models with normal errors and a complete group structure. Here, DI denotes the data-integration method based on multiple quantiles, CA-τ the combined analysis with one quantile $\tau = 2/6$ or $3/6$, SCCA the sparse canonical correlation analysis, DI-LS the data-integration method based on the least squares regression; PSR is the positive selection rate, FDR the false discovery rate, and AE the absolute error $(KM)^{-1}\|\widehat{\theta} - \theta^*\|_1$. The parameter T in criterion (4.2) equals $(\log p)/3$ or $(\log p)/6$. The sample and model sizes are (a) $(n, p) = (100, 100)$, (b) $(100, 200)$ or, (c) $(200, 1000)$.

	$T = (\log p)/3$			$T = (\log p)/6$		
(a)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	98.3 (5.0)	1.1 (1.5)	0.3 (0.1)	99.0 (4.0)	1.9 (2.4)	0.2 (0.1)
CA-(2/6)	83.3 (7.5)	2.4 (2.2)	0.6 (0.1)	92.3 (8.7)	19.2 (16.2)	0.8 (0.3)
CA-(3/6)	81.7 (5.0)	1.4 (1.4)	0.3 (0.1)	83.3 (4.1)	6.9 (8.7)	0.3 (0.2)
SCCA	53.5 (15.0)	4.4 (3.7)	–			
DI-LS	85.0 (6.5)	5.9 (5.0)	0.3 (0.1)			
(b)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	98.2 (5.2)	0.7 (0.7)	0.3 (0.1)	98.3 (5.0)	1.1 (1.3)	0.3 (0.1)
CA-(2/6)	78.0 (8.2)	0.8 (0.7)	0.7 (0.1)	89.3 (8.7)	28.1 (18.2)	1.5 (0.7)
CA-(3/6)	79.2 (7.3)	0.7 (0.7)	0.3 (0.1)	88.7 (7.1)	12.2 (15.4)	0.6 (0.5)
SCCA	56.3 (10.3)	2.3 (1.2)	–			
DI-LS	83.3 (5.8)	4.8 (3.9)	0.3 (0.1)			
(c)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	99.7 (2.3)	0.1 (0.1)	0.1 (0.1)	99.8 (1.7)	0.4 (0.8)	0.1 (0.1)
CA-(2/6)	82.3 (6.2)	0.7 (1.1)	0.5 (0.1)	84.5 (4.9)	3.0 (3.3)	0.5 (0.1)
CA-(3/6)	82.3 (4.0)	0.4 (0.7)	0.2 (0.1)	82.8 (2.9)	1.7 (1.6)	0.2 (0.1)
SCCA	80.0 (14.2)	6.6 (0.7)	–			
DI-LS	83.5 (2.9)	1.3 (1.4)	0.2 (0.1)			

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 3: We consider the same scenario as Table 2, but now the predictors $X_{kij} = I(X'_{kij} > 0)$ $(k = 1, 2;\ i = 1, \ldots, n;\ j = 20, \ldots, 25)$.

	$T = (\log p)/3$			$T = (\log p)/6$		
(a)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	93.0 (7.4)	1.8 (1.6)	0.3 (0.2)	98.0 (5.9)	2.5 (2.4)	0.3 (0.2)
CA-(2/6)	82.0 (5.7)	5.0 (3.4)	0.8 (0.2)	84.7 (5.6)	18.2 (13.9)	0.9 (0.3)
CA-(3/6)	81.2 (5.7)	2.4 (1.5)	0.4 (0.1)	83.0 (3.3)	6.1 (7.5)	0.3 (0.2)
SCCA	51.8 (19.2)	4.2 (4.9)	–			
DI-LS	84.3 (6.2)	6.2 (5.5)	0.3 (0.1)			
(b)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	87.8 (13.9)	0.9 (1.3)	0.4 (0.3)	94.5 (6.9)	1.9 (2.0)	0.4 (0.2)
CA-(2/6)	78.7 (9.2)	2.7 (1.7)	0.9 (0.3)	82.7 (4.7)	27.6 (16.1)	1.6 (0.7)
CA-(3/6)	80.2 (6.6)	1.1 (1.0)	0.4 (0.1)	83.0 (5.8)	11.2 (15.7)	0.6 (0.6)
SCCA	59.5 (13.2)	3.0 (2.1)	–			
DI-LS	83.0 (5.8)	4.2 (3.4)	0.4 (0.1)			
(c)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	97.5 (6.9)	0.2 (0.2)	0.2 (0.2)	98.8 (4.3)	0.4 (0.5)	0.2 (0.1)
CA-(2/6)	81.8 (4.8)	2.8 (2.4)	0.7 (0.1)	82.3 (4.0)	4.2 (3.7)	0.7 (0.1)
CA-(3/6)	82.0 (4.5)	0.9 (1.5)	0.2 (0.1)	83.0 (2.3)	1.5 (1.4)	0.2 (0.1)
SCCA	86.0 (12.9)	6.8 (0.7)	–			
DI-LS	83.5 (1.7)	1.3 (1.5)	0.3 (0.1)			

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

and the lowest FDRs. It also has the smallest AEs. Apparently, the CA-(3/6) and the DI-LS methods are fairly likely to miss predictors that are relevant at some quantile levels: for the first approach, only the conditional median is modeled, and for the second one, only the conditional mean. Our DI method, however, works well since it takes the heterogeneity into account and works with multiple quantile levels simultaneously. Another interesting observation in Table 3 is that $T = (\log p)/3$ tends to underfit models compared to $T = (\log p)/6$.

In Tables 4 and 5, we present the simulation results for the same scenario as in Table 2, but now the predictors have an *incomplete* group structure. The error variables in the two tables have a normal distribution (Table 4) and a t-distribution with three degrees of freedom (Table 5). The nonzero parameters are α^*_{14}, α^*_{16}, α^*_{19}, $\alpha^*_{1(12)}$, $\alpha^*_{1(15)}$, $\alpha^*_{1(20)}$ and α^*_{21}, α^*_{26}, $\alpha^*_{2(12)}$, $\alpha^*_{2(15)}$, $\alpha^*_{2(20)}$, $\alpha^*_{2(25)}$. For $i = 1, \ldots, n$, the predictors in the first experiment are $X_{1i1} = \Phi(X'_{1i1})$ and $X_{1ij} = X'_{1ij}$, for $j \neq 1$. The predictors in the second experiment are $X_{2i3} = \Phi(X'_{2i3})$ and $X_{2ij} = X'_{2ij}$, for $j \neq 3$. The responses are $Y_{1i} = X_{1i}^{\mathrm{T}}\alpha^*_1 + 0.7\xi_{1i}X_{1i1}$ and $Y_{2i} = X_{2i}^{\mathrm{T}}\alpha^*_2 + 0.7\xi_{2i}X_{2i3}$. Inspecting the quantities in the two tables, we see that our DI method again has higher PSRs and lower FDRs. Furthermore, it produces similar or smaller AEs to those of its competitors. We still observe that, in both tables, criterion (4.2) using $T = (\log p)/6$ selects larger models than those selected

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 4: We consider the same scenario as Table 2, but now the predictors have an incomplete group structure.

	$T = (\log p)/3$			$T = (\log p)/6$		
(a)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	97.2 (5.6)	1.8 (1.7)	0.4 (0.1)	98.0 (4.3)	2.4 (2.1)	0.3 (0.2)
CA-(2/6)	86.0 (6.8)	3.4 (3.0)	0.7 (0.1)	92.2 (7.3)	23.7 (16.5)	0.9 (0.3)
CA-(3/6)	84.6 (5.4)	2.2 (1.9)	0.4 (0.1)	87.2 (4.8)	7.6 (8.6)	0.4 (0.2)
SCCA	45.0 (12.5)	5.6 (4.4)	–			
DI-LS	88.6 (5.3)	7.1 (5.4)	0.4 (0.1)			
(b)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	91.3 (9.7)	0.8 (0.9)	0.4 (0.1)	96.6 (6.5)	2.0 (2.0)	0.4 (0.1)
CA-(2/6)	82.9 (6.0)	1.4 (1.2)	0.8 (0.1)	92.0 (7.1)	32.6 (18.3)	1.7 (0.7)
CA-(3/6)	83.8 (6.2)	1.1 (1.0)	0.4 (0.1)	87.1 (7.3)	13.7 (16.5)	0.8 (0.6)
SCCA	45.7 (12.1)	2.7 (2.6)	–			
DI-LS	87.4 (4.9)	5.0 (3.3)	0.4 (0.1)			
(c)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	98.2 (4.1)	0.2 (0.4)	0.2 (0.1)	98.2 (4.1)	0.4 (0.6)	0.2 (0.1)
CA-(2/6)	85.2 (5.3)	1.1 (1.5)	0.6 (0.1)	87.1 (5.2)	3.6 (4.1)	0.6 (0.1)
CA-(3/6)	85.8 (5.0)	0.9 (1.1)	0.3 (0.1)	87.7 (3.5)	2.4 (1.9)	0.2 (0.1)
SCCA	67.1 (10.2)	6.3 (0.8)	–			
DI-LS	87.9 (3.6)	1.2 (1.4)	0.3 (0.1)			

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 5: We consider the scenario from Table 4 with an incomplete group structure, but now the random errors follow a bivariate t-distribution with three degrees of freedom.

	$T = (\log p)/3$			$T = (\log p)/6$		
(a)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	93.7 (6.9)	1.4 (1.4)	0.5 (0.1)	94.9 (6.0)	2.0 (2.2)	0.4 (0.1)
CA-(2/6)	83.0 (6.8)	2.6 (2.6)	0.8 (0.1)	88.7 (8.0)	12.9 (13.0)	0.8 (0.3)
CA-(3/6)	81.2 (5.8)	1.7 (1.8)	0.5 (0.1)	84.8 (5.6)	5.4 (5.6)	0.4 (0.2)
SCCA	44.0 (13.9)	5.8 (6.3)	–			
DI-LS	83.3 (7.2)	9.9 (4.9)	0.7 (0.3)			
(b)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	89.7 (9.9)	0.7 (0.8)	0.5 (0.2)	94.1 (7.3)	1.7 (1.7)	0.5 (0.1)
CA-(2/6)	80.7 (6.6)	1.4 (1.5)	0.9 (0.2)	85.0 (8.4)	12.7 (15.7)	1.3 (0.8)
CA-(3/6)	81.3 (6.5)	0.9 (0.8)	0.5 (0.1)	83.7 (6.4)	4.8 (9.7)	0.6 (0.5)
SCCA	43.2 (10.5)	2.7 (3.3)	–			
DI-LS	83.6 (7.7)	7.7 (4.0)	0.9 (0.4)			
(c)	PSR(%)	FDR(%)	AE	PSR(%)	FDR(%)	AE
DI	96.2 (5.3)	0.1 (0.3)	0.3 (0.1)	96.3 (5.3)	0.3 (0.4)	0.3 (0.1)
CA-(2/6)	83.0 (6.0)	1.0 (1.5)	0.7 (0.1)	85.9 (6.3)	3.1 (3.4)	0.6 (0.1)
CA-(3/6)	83.4 (5.6)	0.6 (0.8)	0.3 (0.1)	85.4 (5.2)	1.6 (1.4)	0.3 (0.1)
SCCA	64.2 (9.3)	6.1 (0.8)	–			
DI-LS	84.7 (5.9)	2.6 (1.6)	0.7 (0.3)			

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

using $T = (\log p)/3$. In Table 5 with t-distributed error variables, the absolute errors of the DI-LS approach based on a least squares regression are significantly larger than those of our DI method, which corroborates the robustness of the quantile regression when the distribution of the errors is heavy tailed.

5. Examples

5.1 Multiple experiments

In this section, we apply our method to data from a liver toxicity study (Bushel et al., 2007), which are available in the `R` package `mixOmics` (Rohart et al., 2017). In the study, two groups of 32 male rats were exposed to non-toxic (50 or 150 mg/kg) and toxic (1,500 or 2,000 mg/kg) doses of acetaminophen (paracetamol), respectively. There is a data set for each group, which contains the rats' expression profiles of 3,116 genes and levels of cholesterol. Owing to the different experimental environments, the two data sets have different measurements. We want to identify the genes that significantly affect the response, namely, the level of cholesterol on a logarithmic scale, based on aggregating the two data sets. To preprocess the data, the genes are sorted by the absolute values of their correlation coefficients with the response in each set. The top 200 genes in each set are

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

included in the analysis as covariates. We observe that the absolute values of their realizations are all below 2.05, which indicates that Assumption 1 is satisfied.

To fit sparse models, we minimize the objective function (2.2) using all data. We consider quantiles $\tau_m = m/10$ for $m = 1, \ldots, 9$, and use two different penalties, the SCAD penalty and the minimax concave penalty (MCP). The tuning parameters of the penalties are chosen using formula (4.2), that is, as minimizers of the MQBIC, with $T = \log p/6$. In addition, we take an approach based on random partitions: we divide each data set randomly into two parts, a training set of size 24, and a validation set of size 8. This is repeated 50 times. The training set is used to select the parameters and obtain the parameter estimates, as before, by minimizing (2.2), with λ chosen using (4.2). The prediction errors

$$\sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki}^{\mathrm{T}}\widehat{\theta}_{km} - \widehat{b}_{km}) \tag{5.1}$$

defined by the loss function are calculated based on the estimates from the training sets and the data X, Y from the validation sets. Here, $\widehat{b}_{km}$ is the estimated intercept in the conditional quantile $Q_{\tau_m}(X_k)$.

For comparison, we also consider a combined analysis (CA), the SCCA method, and the DI-LS method described in Section 4. The CA method in this section now considers nine quantiles $\tau_1 = 1/10, \ldots, \tau_9 = 9/10$ instead

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

of one single quantile, as in Section 4. The data sets and the quantiles are treated separately, after which the results are combined. The prediction errors for the SCCA and DI-LS are calculated using

$$\sum_{k=1}^{K}\sum_{i=1}^{n} M|Y_{ki} - X_{ki}^{\mathrm{T}}\widehat{\theta}_k' - \widehat{b}_k'|/2. \tag{5.2}$$

Recall that the scale factor M is the number of quantiles used by both the DI and the CA. Here, it is used to make the prediction errors comparable with those of the DI and CA in (5.1), which sum $K \times M \times n$ quantile loss functions. For the SCCA method, $\widehat{\theta}_k'$ and $\widehat{b}_k'$ represent the slope vector and the intercept, respectively, obtained from the unpenalized least absolute deviations regression of Y_k on the selected subvector of X_k $(k = 1, \ldots, K)$. For the DI-LS method the estimates are generated directly from the penalized least squares regression. We record the sizes of the models that are fitted using the full data sets, as well as the simulated means and standard deviations of the model sizes and prediction errors otained from the 50 replications.

Table 6 shows the results of analyzing the liver toxicity data. When using the full data sets, our DI method with the SCAD penalty and the MCP penalty selects the same two covariates, which are also chosen by the combined analysis with either of the two penalties. Interestingly, the models fitted by SCCA or DI-LS do not include these two covariates. This difference

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 6: Analysis of the liver toxicity data. The sizes of the selected subset models (column 2) are based on all data, the average sizes and prediction errors (column 3 and 4) are based on the data using random partitions. The standard deviations are in parentheses. DI denotes the data-integration method based on multiple quantiles, CA the combined analysis, SCCA the sparse canonical correlation analysis, DI-LS the data-integration method based on the least squares regression, SCAD the smoothly clipped absolute deviation, and MCP the minimax concave penalty.

	All Data	Random Partition	
	Model Size	Model Size	Prediction error
DI with SCAD	2	3.30 (1.54)	1.19 (1.06)
DI with MCP	2	2.60 (1.09)	1.24 (1.33)
CA with SCAD	13	13.92 (3.84)	1.54 (1.26)
CA with MCP	11	14.86 (3.88)	1.40 (0.98)
SCCA	15	13.58 (1.75)	4.10 (1.37)
DI-LS	8	8.28 (3.77)	3.07 (1.59)

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

suggests heterogeneity in the data, because both the SCCA and the DI-LS method tend to ignore covariates that affect responses only at certain quantile levels, but not globally. Using the random partition approach, our DI method generates models that are, on average, more sparse than those obtained from the competitors, with lower prediction errors.

5.2 Multiple responses

As a second application, now with a multivariate response vector, we analyze data sets of financial market indices from the `R` package `FusionLearn` (Gao et al., 2019). These data contain three correlated indices: the VIX index, the S&P 500 index, and the Dow Jones index. The VIX and the S&P 500 are negatively correlated, and the S&P 500 and the Dow Jones are positively correlated (Gao and Carroll, 2017). The covariates are 46 major international equity indices, North American bond indices, and major commodity indices. In the analysis, the transformation $\log(V_t \,/\, V_y) \times 100$ of each index is used, where V_t and V_y denote the current and previous days' values, respectively. The training data set consists of 232 records of three years' market performances, with three-day spacing between the values. As shown in Gao and Carroll (2017), the values are not autocorrelated at a 5% significance level.

As before, we minimize the objective function (2.2) to select the covari-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

Table 7: Analysis of the financial market indices. The figures are the prediction errors and the sizes of the selected submodels. The full model size is $p = 46$. DI denotes the data-integration method based on multiple quantiles, CA the combined analysis, UR is unpenalized regression, SCCA the sparse canonical correlation analysis, DI-LS the data-integration method based on the least squares regression, SCAD denotes smoothly clipped absolute deviation, and MCP minimax concave penalty.

	Model Size	Prediction errors		
		VIX	Dow Jones	S&P 500
DI with SCAD	4	10052.8	523.4	307.2
DI with MCP	4	10020.6	522.4	309.2
CA with SCAD	20	10150.4	637.3	400.1
CA with MCP	20	10125.1	637.9	396.1
UR	46	13408.5	644.0	663.4
SCCA	6	12998.1	658.1	513.4
DI-LS	13	13996.3	720.5	499.5

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

ates and estimate the parameters. The quantiles in (2.2) are $\tau_m = m/20$, for $m = 1, 2, \ldots, 19$. We again use the SCAD penalty and the MCP, and determine their tuning parameters using criterion (4.2). The SCAD penalty selects four covariates, which are the same as those selected by the MCP penalty. The competing methods are the combined analysis with the two penalties and the unpenalized regression. The latter includes all 46 covariates in the model and generates estimators by minimizing the loss function (2.1) without a penalty term. We use the five fitted models for predictions based on a (different) validation data set with 464 records. The prediction errors for the three indices, that is, $\sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki}^{\mathrm{T}}\widehat{\theta}_{km} - \widehat{b}_{km})$, for $k = 1, 2, 3$, and the model sizes are recorded in Table 7. There, we also list the results for the SCCA and the DI-LS approach. The prediction errors for these two methods are $\sum_{i=1}^{n} M|Y_{ki} - X_{ki}^{\mathrm{T}}\widehat{\theta}'_k - \widehat{b}'_k|/2$, with $\widehat{\theta}'_k$ and $\widehat{b}'_k$ $(k = 1, 2, 3)$, as in (5.2).

Our DI method with the SCAD penalty and the MCP outperforms the other five approaches, whereas the DI with the SCAD penalty and the DI with the MCP yield similar prediction errors. Apart from that, our DI method selects models that are considerably smaller than those from the competitors, that is, it achieves more sparsity. As in Section 5.1, the two DI approaches and the two CA methods choose the same four predictors, whereas the SCCA selects only one, and the DI-LS selects none of them.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

This again indicates heterogeity in the data, that is, some predictors affect the responses only locally. The two empirical data examples in Sections 5.1 and 5.2 clearly demonstrate the advantages of our method, especially its ability to handle complex data.

6. Conclusion

To the best of our knowledge, this is the first time that a quantile regression approach has been applied to a DI scenario with high-dimensional data. By considering multiple quantiles simultaneously, we obtain a global picture of the relationship between the predictors and the responses. A penalized estimator and an information criterion, which aggregate information from multiple experiments, have been developed to select variables and to estimate the model parameters. Our method copes with heterogeneity in the data. It successfully exploits the group structure in the parameter set across quantiles and experiments so that influential predictors can be identified.

In practice, the quality and relevance of data may vary from one source to another. Therefore, a weighted version of the loss function (2.1),

$$\ell_n^{(w)}(\theta) = n^{-1}\sum_{k=1}^{K} w_k \sum_{m=1}^{M}\sum_{i=1}^{n}\rho_m(Y_{ki} - X_{ki}^{\mathrm{T}}\theta_{km}),$$

with weight vector $w = (w_1, \ldots, w_K)^{\mathrm{T}}$, may improve our estimator, which uses uniform weights. It would be worthwhile specifying and constructing

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0361

such weights for data from different experiments.

The nonconvex penalty function associated with the L_1-norm has different properties to those of the penalty function associated with the L_2-norm employed by Gao and Carroll (2017), which forces parameters in the same group to be all zero or all nonzero. When the least squares approach is used, Jiang and Huang (2015) show that the penalty associated with the L_1-norm can be applied if the group structure is incomplete, that is, both zero and nonzero parameters exist in the same group. This capacity is called a "bi-level selection" property. Here, we focus on groups of parameters to identify predictors that have an impact on one or more responses at some quantile levels. In the simulations of Section 4, we saw that the SCAD penalty with the L_1-norm actually performs well at the group level, even if the group structure is incomplete. The theoretical properties of the L_1-norm in the quantile regression setting still need to be investigated in greater detail.

Supplementary Material

- The proofs of the theoretical results and additional simulation results are provided in the online Supplementary Material.
- All the programs of Section 4 and 5 are available at `https://github.com/guorongdai/data_integration`.
- The data in Section 5.1 are from the `R` package `FusionLearn`, and the

data in Section 5.2 are from the `R` package `mixOmics`.

Acknowledgments

The authors thank the referees and the associate editor for their helpful comments. Dai and Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

References

Belloni, A. and V. Chernozhukov (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* **39**, 82–130.

Bushel, P. R., R. D. Wolfinger, and G. Gibson (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology* **1**, 15.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J., L. Xue, and H. Zou (2016). Multitask quantile regression under the transnormal model. *Journal of the American Statistical Association* **111**, 1726–1735.

Gao, X. and R. J. Carroll (2017). Data integration with high dimensionality. *Biometrika* **104**, 251–272.

Gao, X., Y. Zhong, and R. J. Carroll (2019). *FusionLearn: Fusion Learning*. R package version 0.1.1, available at `https://CRAN.R-project.org/package=FusionLearn`.

Gu, Y. and H. Zou (2020). Sparse composite quantile regression in ultrahigh dimensions with tuning parameter calibration. *IEEE Transactions on Information Theory* **66**, 7132–7154.

Jiang, D. and J. Huang (2015). Concave 1-norm group selection. *Biostatistics* **16**, 252–267.

Koenker, R. (2005). *Quantile Regression*. Cambridge, UK: Cambridge University Press.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Lee, E. R., H. Noh, and B. U. Park (2014). Model selection via Bayesian information criterion for quantile regression models. *Journal of the American Statistical Association* **109**, 216–229.

Peng, B. and L. Wang (2015). An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics* **24**, 676–694.

Rohart, F., B. Gautier, A. Singh, and K.-A. Le, Cao (2017). mixomics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology* **13**(11), e1005752. available at `http://www.mixOmics.org`.

Sherwood, B. and L. Wang (2016). Partially linear additive quantile regression in ultra-high dimension. *Annals of Statistics* **44**, 288–317.

Wang, L., Y. Wu, and R. Li (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association* **107**, 214–222.

Witten, D. M. and R. J. Tibshirani (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology* **8**, 1–27.

Yu, K., Z. Lu, and J. Stander (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**, 331–350.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and M. Yuan (2008a). Composite quantile regression and the oracle model selection theory. *Annals of Statistics* **36**, 1108–1126.

Zou, H. and M. Yuan (2008b). Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics & Data Analysis* **52**, 5296–5304.

Guorong Dai (corresponding author), Department of Statistics and Data Science, School of Management, Fudan University, Shanghai 200433, China

E-mail: rondai@stat.tamu.edu

Uschi U. Müller, Department of Statistics, Texas A&M University, College Station, TX 77843, USA

E-mail: uschi@stat.tamu.edu

Raymond J. Carroll, Department of Statistics, Texas A&M University, College Station, TX

77843, USA

E-mail: carroll@stat.tamu.edu