Statistica Sinica Preprint No: SS-2022-0292	
Title	Integrative Quantile Regression Analysis of Heterogeneous Multisource Data with Privacy Preserving
Manuscript ID	SS-2022-0292
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0292
Complete List of Authors	Senlin Yuan, Xuerong Chen, Yu Wu and Jianguo Sun
Corresponding Authors	Xuerong Chen
E-mails	chenxuerong@swufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Integrative Quantile Regression Analysis of Heterogeneous Multisource Data with Privacy Preserving

Senlin Yuan[1], Xuerong Chen[1], Yu Wu[2] and Jianguo Sun[3]

[1]*Southwestern University of Finance and Economics*,

[2]*Nanjing Agricultural University and* [3]*University of Missouri*

Abstract: Researchers have used and discussed multisource data integrative analysis in many fields. In this paper, we focus on quantile regression for an analysis that has not been investigated in the literature. Specifically, we consider quantile integrative analysis of multisource and high-dimensional data where both homogeneity and heterogeneity may exist in covariate effects among different data sets. We aim to detect the homogenous and heterogenous effects, obtain the estimators of corresponding parameters, and improve the statistical efficiency of the potential homogeneous covariate effects by integrating the information contained in different data sources, while the raw data are unavailable. For the problem, we propose an objective function based on a composite penalty. In particular, we propose the composite penalty term to pursue the homogeneous and nonzero covariate effects when the dimension of covariates is high; the main term of the objective function can aggregate the quantile regression estimators from the various data sources and hence improve the statistical efficiency of potential homogeneous covariate effects. Meanwhile, it relies only on the summary

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

statistics from each data source and thus can protect privacy to a great extent. The proposed privacy protection estimators of the homogeneous effects achieve the same statistical efficiency as the benchmark estimators obtained based on individual-level data. We establish the selection consistency and asymptotic normality of the proposed estimators for homogeneous effects, and the numerical results suggest the performance of the proposed estimators is good. Finally, we apply the proposed method to the Chinese Annual Survey of Industrial Firms data set.

Keywords and phrases: Privacy preservation; Quantile regression; Heterogeneous data; Integrative analysis; Distributed learning; High dimensional.

1. Introduction

Multisource data integrative analysis or meta-analysis has become increasingly important because of the amount of data generated in various application fields. Examples include the PM2.5 data from different cities in China (Liang et al., 2016), the social media data from different social platforms (Moniz and Torgo, 2018), and the electronic health records data from different hospitals (Cai et al., 2022). Among others, multisource data integrative analysis can improve statistical efficiency by increasing the effective sample size, and provide more comprehensive information for decision makers.

One area very close to integrative analysis is distributed learning be-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

cause both involve data fusion. Distributed learning can be categorized into two types depending on whether all data is stored together. The first type involves all data being stored in same place, also often referred to as "divide and conquer". It divides the whole dataset into small data blocks and collaboratively trains the model by using parallel computing systems (Zhang et al., 2015). Hence, it requires data to be homogeneous because the data is usually randomly divided. The second type is where the data has been naturally collected and stored in different locations, and hence there may be heterogeneity among different data blocks (Duan et al., 2021). It is also known as federated learning (Ghosh et al., 2019). It is worth noting that distributed learning focuses on the trade-off between computational efficiency and estimation efficiency, while integrative analysis prioritizes estimation efficiency. Integrative analysis and the second type of distributed learning can be regarded as the same if the differences in focus are ignored. In the following context, we will no longer distinguish the literature from the two fields, since many methods are common to both fields, although we still focus more on integrative analysis.

Multisource data integrative analysis faces several challenges. The first challenge is that individual-level data for integrative analysis are often unavailable because of privacy considerations. Hence authors usually focus

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

on data fusion based on summary statistics from different data sources. In particular, meta-analysis (Lin and Zeng, 2010) and confidence distribution methods (Liu et al., 2015) based on summary statistics alone are frequently applied. The second challenge is that data is usually high-dimensional. High-dimensional inference problems have captured attention within statistics over the past decade (Tibshirani et al., 2015), and led to the investigation of high-dimensional distributed learning as a means of tackling large datasets with high-dimensional features, see, e.g., Lee et al. (2017). The third challenge is that there exists potential heterogeneity among different data sources. The identification of homogeneity can reduce model complexity and improve statistical efficiency and power. Note that "homogeneity" and "heterogeneity" may have different meanings or interpretations in various applications. In our paper, homogeneity means that the effects on the same covariates are the same across all data sources, and heterogeneity allows for the effects on the same covariates to be different, see, e.g., Cai et al. (2022). It is worth mentioning that although Cai et al. (2022) have considered integrative regression analysis of high-dimensional heterogeneous data and gave a method to identify the effects of the homogeneous covariates, their approach relies on smooth objective functions or estimation equations, and thus cannot be generalized easily to some powerful but

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

unsmooth approaches such as the quantile regression.

Quantile regression (QR), first introduced by Koenker and Bassett (1978), has been studied extensively because of its fascinating features, such as its robustness to data with outliers, and its ability to provide a comprehensive analysis of the impact of covariates on response variables, etc. QR integrative analysis or QR distributed learning faces challenges both from computation and theoretical derivation due to the non-smoothness of the objective function involved. Some authors have discussed QR for distributed learning while there are few studies of QR integrative analysis. For example, Volgushev et al. (2019) developed the inference procedure for quantile process based on a naive divide and conquer quantile regression estimator at fixed quantile levels. Chen and Zhou (2020) presented a divide-and-conquer based QR analysis method for big data. The proposed estimator, which is a weighted average of local estimators obtained using small data blocks and computed on separate machines, has a closed form. However, their model fails to address the challenges posed by high-dimensional sparse problems. To overcome the computational difficulty in QR and the memory constraint, Chen et al. (2019) and Chen et al. (2020) proposed divide and conquer QR methods based on kernel smoothing loss function and least-square-type loss function respectively for massive data

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

with high covariate dimensionality. However, both methods involve kernel technique, and hence bandwidth choice caused some trouble in their application. Wang and Lian (2020) extended the surrogate-likelihood idea of Jordan et al. (2019) to high-dimensional QR and established the related theoretical properties of the non-smooth loss function for the situation where the number of covariates is greater than the sample size. Notably, all of the QR methods mentioned above do not take into account the potential heterogeneity of the data. To our best knowledge, there is currently no published paper on integrative QR analysis of high-dimensional heterogeneous multisource data, despite its crucial importance in various real-world applications. For instance, when examining the economic development of Chinese enterprises, it is intriguing to investigate how variables such as corporate debt ratio and corporate size influence the total factor productivity (TFP) of enterprises in four industrial cities in China. An enterprise's TFP is a significant metric for gauging its economic development accomplishments, and a comprehensive exploration of it at various levels can provide a more accurate reflection of the enterprise's economic development status. In Section 6, our QR integrative analysis revealed heterogeneity in the impact of covariates on the TFP of enterprises in the four industrial cities in China. For a more detailed understanding, please refer to Section 6.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

One area closely to QR integrative analysis is QR transfer learning, detail please see Huang et al. (2022) , Jin et al. (2022) and Zhang and Zhu (2022). It's important to note that QR transfer learning and QR integrative analysis also have different focuses. The former aims to transfer knowledge from source data to improve the learning performance of the target problem, while the latter aims to improve the estimation efficiency of homogeneous effects across all datasets. However, one common challenge encountered in both fields is how to perform transfer learning or data integrative analysis while data sets can not be shared due to privacy protection. Actually, individual-level data is unavailable in many real applications.

It motivates us to consider the integrative QR analysis of high-dimensional heterogeneous multisource data with privacy preservation. Specifically, the contribution of this article is multi-fold. First, the proposed quantile integrative framework accommodates the situation that all data sources may be heterogeneous while some of the covariates effects may be homogeneous. The procedure utilizes composite penalty to detect homogeneous and heterogeneous effects, estimates the corresponding parameters, and enhances the statistical efficiency of potential homogeneous covariate effects by integrating information from diverse data sources. Second, we establish the consistency and asymptotic normality of the estimators for homogeneous

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

effects under regular conditions, when the number of covariates goes to infinity as the smallest sample size among different data sources goes to infinity. The proposed privacy protection estimators of the homogeneous effects can achieve the same statistical efficiency as the benchmark estimators obtained based on individual-level data. Third, the proposed approach can address the high-dimensional data and obtain a sparse consistent estimator, which relies on only the summary statistics from each data source and hence can protect privacy to a great extent. Finally, the proposed procedure is highly efficient and robust when the data are skewed, contain outliers, or suffer heavy tail noise, and thus it has wide application prospects. It is worth noting that the proposed method can be used as a communication-efficient distributed learning approach for heterogeneous data.

The remainder of this paper is organized as follows. In Section 2, we present the proposed estimation procedure, and in Section 3 we establish the asymptotic properties of the proposed estimators. We discuss the implementation of the proposed approach in Section 4, particularly the adoption of the minorization-maximization algorithm. In Section 5, we present simulation results obtained to assess the proposed method's performance; these results suggest it works well for practical situations. In Section 6, we apply the method to the Chinese Annual Survey of Industrial Firms data set.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Finally, in Section 7, we offer discussion and concluding remarks. Technical details, additional simulations and the variable information in practical application are given in the Supplementary Material.

2. Integrative QR analysis

Suppose that there are K independent data sources and let n_k denote the number of subjects in the kth data source, $k = 1, \ldots, K$. Also, let $Y_i^{(k)}, \mathbf{X}_i^{(k)}$ denote the response variable and the p_n-dimensional vector of covariates associated with the ith subject in the k-th data source, respectively. Without loss of generality, we assume the observations within each data source are independent and the covariate $\mathbf{X}_i^{(k)}$ includes 1 as the first component. We assume some covariates have homogeneous effects across different data sources, but we do not know which covariates they are. We will develop a homogeneous detection method to identify them later. Beforehand, we cannot and hence do not distinguish between homogeneous variables and heterogeneous variables in the model, and we write all covariates effects in heterogeneous form. For the k-th data source and a specified quantile level $\tau \in (0, 1)$, define the population QR parameter of interest as

$$\boldsymbol{\beta}_{\tau 0}^{(k)} = \arg\min_{\boldsymbol{\beta}_\tau^{(k)}} L^{(k)}(\boldsymbol{\beta}_\tau^{(k)}), \text{where } L^{(k)}(\boldsymbol{\beta}_\tau^{(k)}) = E\rho_\tau(Y_i^{(k)} - \mathbf{X}_i^{(k)T}\boldsymbol{\beta}_\tau^{(k)}),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

where $\rho_\tau(\epsilon) = \epsilon(\tau - I(\epsilon < 0))$. For simplicity, throughout the paper we will omit the subscript τ of $\boldsymbol{\beta}_\tau^{(k)}$ when there is no confusion, or write the true parameter $\boldsymbol{\beta}_{\tau 0}^{(k)}$ as $\boldsymbol{\beta}_0^{(k)}$ and further define $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{(1)T}, \ldots, \boldsymbol{\beta}_0^{(K)T})^T$.

To estimate the QR parameters of interest, we first will discuss the situation where raw data are available. For $j = 1, \ldots, p_n$ and $k = 1, \ldots, K$, let $\boldsymbol{\beta}_j = (\beta_j^{(1)}, \ldots, \beta_j^{(K)})^T$, where $\beta_j^{(k)}$ is the j-th component of $\boldsymbol{\beta}^{(k)}$. Define $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)T}, \ldots, \boldsymbol{\beta}^{(K)T})^T$. Define $\boldsymbol{\beta}_{0,-1}^{(k)}$ as a vector obtained by omitting the 1th component of $\boldsymbol{\beta}_0^{(k)}$ for $k = 1, \ldots, K$. Denote $N = \sum_{k=1}^K n_k$ as the total sample size of all data sources. If all data sources have the same covariate effect, that is, when all data sources are homogeneous, to improve the estimation efficiency, it is natural to consider the following integrative loss function

$$L_N(\boldsymbol{\beta}) = \frac{1}{N} \sum_{k=1}^{K} n_k L_n^{(k)}(\boldsymbol{\beta}^{(k)}), \text{where } L_n^{(k)}(\boldsymbol{\beta}^{(k)}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \rho_\tau(Y_i^{(k)} - \mathbf{X}_i^{(k)T} \boldsymbol{\beta}^{(k)}).$$

However, it is not easy to distinguish the covariates with and without homogeneous effects. To detect the potential homogeneous effects, we will borrow the random effect modeling idea by treating the $\boldsymbol{\beta}^{(k)}, k = 1, \ldots, K$, as random effects for different data sources. Specifically, if $\boldsymbol{\beta}_{0,-1}^{(1)}, \ldots, \boldsymbol{\beta}_{0,-1}^{(K)}$ share similar supports and magnitudes, for $j = 2, \ldots, p_n, k = 1, \ldots, K$, we can decompose $\beta_j^{(k)}$ as $\alpha_j + \gamma_j^{(k)}$, where α_j denotes the average effects and $\gamma_j^{(k)}$ represents the deviance of effects in study k from the average effects α_j

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

with $\sum_{k=1}^{K} \gamma_j^{(k)} = 0$ for identifiability. Cheng et al. (2015), and Cai et al. (2022) among others, have used similar techniques.

Define $\boldsymbol{\gamma}_j = (\gamma_j^{(1)}, \ldots, \gamma_j^{(K)})^T$. In the following, we will consider the composite penalty function

$$\phi(\boldsymbol{\beta}) = \sum_{j=2}^{p_n} P_{\lambda_1}(|\alpha_j|) + \sum_{j=2}^{p_n} P_{\lambda_2}(\|\boldsymbol{\gamma}_j\|)\,,$$

which is the mixture of a penalty for a single parameter and a group penalty for a parameter vector. In the above, $\|\cdot\|$ denotes the L_2 norm for a vector, λ_1 and λ_2 are the tuning parameters, and $P_\lambda(\cdot)$ is a penalty function. In the following, we will focus on the SCAD penalty (Fan and Li, 2001) because of its oracle properties, and its first derivative is defined as

$$P'_\lambda(b) = \lambda\left\{ I(b \le \lambda) + \frac{(a\lambda - b)_+}{(a-1)\lambda} I(b > \lambda) \right\},$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. Note the use of the penalty function divides covariate effects into three mutually exclusive groups as follows (1) Homogeneous effects: $\alpha_j \neq 0$ and $\|\boldsymbol{\gamma}_j\| = 0$; (2) Heterogeneous effects: $\|\boldsymbol{\gamma}_j\| \neq 0$; (3) Null effects: $\alpha_j = 0$ and $\|\boldsymbol{\gamma}_j\| = 0$. It follows that a natural objective function based on individual-level data is given by

$$G_N(\boldsymbol{\beta}) = L_N(\boldsymbol{\beta}) + \phi(\boldsymbol{\beta}). \tag{2.1}$$

If individual-level data are available, then we can estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}_{ILD} = \arg\min_{\boldsymbol{\beta}} G_N(\boldsymbol{\beta})$ and reference it as the benchmark estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Of course, as discussed earlier, raw data are often unavailable because of privacy preservation methods. Then when only summary statistics are available, we write the integrative loss function $L_N(\boldsymbol{\beta})$ as

$$L_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{k=1}^{K} n_k(L_n^{(k)}(\boldsymbol{\beta}^{(k)}) - L_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)})) + \frac{1}{N}\sum_{k=1}^{K} n_k L_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)}),$$

where $\tilde{\boldsymbol{\beta}}^{(k)} = \arg\min_{\boldsymbol{\beta}^{(k)}} L_n^{(k)}(\boldsymbol{\beta}^{(k)})$, and we reference it as the local estimator. Note the second term of the right-hand side of the equation above does not contain unknown parameters. Thus, the minimization of $L_N(\boldsymbol{\beta})$ is equivalent to the minimization of $N^{-1}\sum_{k=1}^{K} n_k(L_n^{(k)}(\boldsymbol{\beta}^{(k)}) - L_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)}))$. In addition, note that under some regularity conditions, we can derive (see Lemma 2 for further details)

$$\begin{aligned} & N^{-1}\sum_{k=1}^{K} n_k(L_n^{(k)}(\boldsymbol{\beta}^{(k)}) - L_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)})) \\ = \; & (2N)^{-1}\sum_{k=1}^{K} n_k(\boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)})^T V_k(\boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)}) + o_p(1), \end{aligned} \tag{2.2}$$

where $V_k = n_k^{-1}\sum_{i=1}^{n_k} E[\mathbf{X}_i^{(k)}\mathbf{X}_i^{(k)T} f_{Y|X}(\mathbf{X}_i^{(k)T}\boldsymbol{\beta}_0^{(k)}|\mathbf{X}_i^{(k)})]$ and $f_{Y|X}(\cdot|\mathbf{X}_i^{(k)})$ denotes the conditional probability density function of Y given X. Motivated by these, we propose the following objective function

$$Q_N(\boldsymbol{\beta}) = \frac{1}{2N}\sum_{k=1}^{K} n_k(\boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)})^T \tilde{V}_k(\boldsymbol{\beta}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)}) + \phi(\boldsymbol{\beta}), \tag{2.3}$$

where $\tilde{V}_k$ denotes an estimator of V_k, and one way to obtain it is to use the bootstrap procedure (discussed later). Clearly, the integrative loss function

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

$Q_N(\boldsymbol{\beta})$ depends only on the summary statistics $\{n_k, \tilde{\boldsymbol{\beta}}^{(k)}, \tilde{V}_k,\ k = 1, \ldots, K\}$ provided by individual data sources. Thus, a natural estimator of the regression parameter $\boldsymbol{\beta}$ with privacy preservation is given by $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q_N(\boldsymbol{\beta})$.

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_n})^T$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{(1)T}, \ldots, \boldsymbol{\gamma}^{(K)T})^T$, where $\boldsymbol{\gamma}^{(k)} = (\gamma_1^{(k)}, \ldots, \gamma_{p_n}^{(k)})^T$. To identify the structure of covariates, the objective function (2.3) can be rewritten as

$$Q_N(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2N} \sum_{k=1}^{K} n_k (\boldsymbol{\alpha} + \boldsymbol{\gamma}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)})^T \tilde{V}_k (\boldsymbol{\alpha} + \boldsymbol{\gamma}^{(k)} - \tilde{\boldsymbol{\beta}}^{(k)}) + \phi(\boldsymbol{\beta}). \quad (2.4)$$

Based on this new objective function $Q_N(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, we can obtain $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) = \arg\min_{\boldsymbol{\alpha}, \boldsymbol{\gamma}} Q_N(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, and further obtain $\widehat{\beta}_j^{(k)} = \widehat{\alpha}_j + \widehat{\gamma}_j^{(k)}$.

Remark 1. The objective function $Q_N(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ (or $Q_N(\boldsymbol{\beta})$) has some nice features. On the one hand, it protects the privacy of the data to some extent since it only relies on summary statistics from different data sources instead of raw data. On the other hand, the composite penalty function can identify homogeneous, heterogeneous, and null effects, which results in a sparse solution for high-dimensional issues.

3. Asymptotic properties

We will now establish the asymptotic properties of the proposed estimator $\widehat{\boldsymbol{\beta}}$. To do this, we first define some notation. Let $L_N^*(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ denote the

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

first term on the right side of (2.4), and $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0p_n})^T$ and $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_0^{(1)T}, \ldots, \boldsymbol{\gamma}_0^{(K)T})^T$ be the true values of corresponding parameters. Let $q_n = Kp_n$ be the total number of parameters, $F_{Y|X}(\cdot|\mathbf{X}_i^{(k)})$ denote the conditional distribution function of response Y given X. Denote $A_\alpha = \{j : \alpha_j \neq 0, j = 2, \ldots, p_n\}$, $A_\gamma = \{j : \|\boldsymbol{\gamma}_j\| \neq 0, j = 2, \ldots, p_n\}$. Let $A = A_\alpha \cup A_\gamma$, $A^* = A_\alpha \cap A_\gamma^c$, where A_γ^c denotes the complement of set A_γ. Obviously, the index set A^c represents the covariates producing null effect, A^* denotes the covariates producing non null homogeneous effect. Throughout the paper, $\Lambda(\cdot)$ are eigenvalues of a matrix, and $\Lambda_{\min}(\cdot) \leq \cdots \leq \Lambda_{\max}(\cdot)$. Furthermore, let $|\cdot|$ represent its cardinality for a set, define $\|\cdot\|_1$ as the L_1 norm of the vector, and for the matrix, $\|\cdot\|$ is defined by $\|\cdot\| = \Lambda_{\max}^{1/2}(\cdot)$. Also define $U_k = n_k^{-1} \sum_{i=1}^{n_k} E(\mathbf{X}_i^{(k)} \mathbf{X}_i^{(k)T})$

In the following, we assume that the number of parameters p_n goes to infinity as the sample size n_k increases but $p_n < n_k$, $k = 1, \ldots, K$. We also assume that the sample size n_k for all data sources diverges in the same order $O(N/K)$, and $K = O(n^\iota)$ $(0 \leq \iota \leq 1/3)$, where $n = \min_{1 \leq k \leq K}\{n_k\}$. It is common for K to diverge at some rate that depends on the minimum sample size in meta-analysis and distributed learning (Lin and Xi (2011), Chen and Zhou (2020)). In addition, we need the following regularity conditions.

(C1) The conditional probability density function of Y given covariates

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

$f_{Y|X}(\cdot|\mathbf{X}_i^{(k)})$ and its first-order derivative is bounded for any $k \in \{1, \ldots, K\}$.

(C2) There exist constants b_1 and b_2 such that $0 < b_1 \leq \min_{1\leq k\leq K} \Lambda_{\min}(U_k) \leq \max_{1\leq k\leq K} \Lambda_{\max}(U_k) \leq b_2 < \infty$ and $\max_{1\leq i\leq n_k, 1\leq j\leq p_n} |\mathbf{X}_{ij}^{(k)}| = O_p(1)$ for $k = 1, \ldots, K$.

(C3) $\|\tilde{V}_k - V_k\| = O_p(\sqrt{\frac{p_n}{n}})$ for $k = 1, \ldots, K$.

(C4) $K(|A_\alpha|\lambda_1^2 + |A_\gamma|\lambda_2^2) = o(1)$ as $n \to \infty$, $\lambda_1 \to 0$, $\lambda_2 \to 0$.

(C5) $\sqrt{\frac{p_n}{n}}/\lambda_1 \to 0$ as $n \to \infty$, $\lambda_1 \to 0$, and $\sqrt{\frac{p_n}{n}}/\lambda_2 \to 0$ as $n \to \infty$, $\lambda_2 \to 0$.

(C6) $\liminf_{n\to\infty} \liminf_{\theta\to 0} \frac{P'_\lambda(\theta)}{\lambda} > 0$.

(C7) There are positive constants C and D, when $\alpha_1, \alpha_2 > C\lambda$, $|P''_\lambda(\alpha_1) - P''_\lambda(\alpha_2)| \leq D|\alpha_1 - \alpha_2|$.

Note that Condition (C1) is common in the QR literature (Chen et al. (2015)), and Condition (C2) is about the behavior of the covariate matrix. Condition (C3) is essential to obtain the asymptotic normality of the proposed estimator (Chen and Zhou (2020)), and Condition (C4) guarantees the consistency of $\widehat{\boldsymbol{\beta}}$ (Huang and Xie (2007)). Conditions (C5) and (C6) are required to obtain the proposed estimator's the selection consistency, and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Condition (C7) is imposed for the smoothness of the nonconcave penalty functions and is essential for the asymptotic normality of the proposed estimator. Conditions (C5)-(C7) can be found in Fan and Peng (2004).

Now we are ready to establish the consistency of the proposed estimators and their selection consistency.

Theorem 1 (Estimation Consistency)**.** *Assume that Conditions (C1), (C2), and (C4) hold, if* $p_n(\log p_n)^3/n \to 0$ *as* $n \to \infty$*, then* $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| = O_p\big((q_nK/N)^{1/2}\big)$.

Theorem 2 (Selection Consistency)**.** *Assume that Conditions (C1)-(C6) hold, if* $\lambda_1 \to 0$*,* $\lambda_2 \to 0$*, and* $p_n(\log p_n)^3/n \to 0$ *as* $n \to \infty$*, then* $Pr(\widehat{\boldsymbol{\beta}}_{A^c} = 0) \to 1$*,* $\Pr(\widehat{\alpha}_{A_\alpha^c} = 0) \to 1$*, and* $\Pr(\|\widehat{\boldsymbol{\gamma}}_{A_\gamma^c}\| = 0) \to 1$.

Remark 2. Theorem 2 guarantees the sparsity remains valid when the number of parameters diverges, and it provides two results about consistency. The first is the variable selection consistency, meaning for the case of null effect (i.e. $j \in A^c$), the corresponding estimators equal to zero with probability approaching 1. The second is the consistency of homogeneous and heterogeneous detection. Note that $\widehat{\boldsymbol{\beta}}_{A_\gamma^c} = \widehat{\boldsymbol{\alpha}}_{A_\gamma^c}$ are the homogeneous covariate effects because $\|\widehat{\boldsymbol{\gamma}}_{A_\gamma^c}\| = 0$ with probability approaching 1. This means the proposed method can detect the homogeneous covariate effect correctly with probability approaching 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

To establish the asymptotic normality, divide $\boldsymbol{\beta}^{(k)}$ as $\boldsymbol{\beta}^{(k)} = (\boldsymbol{\beta}_{A^*}^{(k)T}, \boldsymbol{\beta}_{A^{*c}}^{(k)T})^T$, where $\boldsymbol{\beta}_{A^*}^{(k)}$ denotes the part of non null homogeneous effects and $\boldsymbol{\beta}_{A^{*c}}^{(k)}$ denotes the others. For simplicity, write $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{A^*}^T, \boldsymbol{\alpha}_{A^{*c}}^T)$ and do the same for $\boldsymbol{\alpha}_0$. Note that for $j \in A^*$, $\beta_j^{(1)} = \cdots = \beta_j^{(K)} = \alpha_j$. It thus follows that $\boldsymbol{\beta}_{A^*}^{(1)} = \cdots = \boldsymbol{\beta}_{A^*}^{(K)} = \boldsymbol{\alpha}_{A^*}$. Furthermore, decompose the matrix V_k as $(V_k^{11}, V_k^{12}; V_k^{21}, V_k^{22})$, where V_k^{11} is the $|A^*| \times |A^*|$ submatrix of V_k and does the same for $\tilde{V}_k$. Let $\boldsymbol{b}$ represent the gradient vector that is the derivative of the first penalty term concerning $\boldsymbol{\alpha}_{A^*}$ at $\boldsymbol{\alpha}_{0A^*}$ with its jth component being $P'_{\lambda_1}(|\alpha_{0j}|)\text{sign}(\alpha_{0j}), j = 2, \ldots, |A^*|$. Let Σ_{λ_1} denote the diagonal hessian matrix that is the second derivative of the first penalty term concerning $\boldsymbol{\alpha}_{A^*}$ at $\boldsymbol{\alpha}_{0A^*}$ with its jth diagonal element being $P''_{\lambda_1}(|\alpha_{0j}|), j = 2, \ldots, |A^*|$.

Theorem 3 (Asymptotic Normality)**.** *Assume that Conditions (C1)-(C7) hold, if* $\lambda_1 \to 0$, $\lambda_2 \to 0$, *and* $p_n^3(\log p_n)^2/n \to 0$ *as* $n \to \infty$, *then for any* $j \in A^*$, *we have that* $\sqrt{N}\boldsymbol{e}^T\big[\big(N^{-1}\sum_{k=1}^K n_k V_k^{11} + \Sigma_{\lambda_1}\big)(\widehat{\boldsymbol{\alpha}}_{A^*} - \boldsymbol{\alpha}_{0A^*}) + \boldsymbol{b}\big]/\sigma \xrightarrow{D} N(0,1)$ *for any* $|A^*|$*-dimensional vector* $\boldsymbol{e}$ *such that* $\|\boldsymbol{e}\| = 1$, *where* $\sigma^2 = \boldsymbol{e}^T\Sigma\boldsymbol{e}$ *and* $\Sigma = N^{-1}\sum_{k=1}^K n_k(V_k^{11}, V_k^{12})\tau(1-\tau)V_k^{-1}U_kV_k^{-1}(V_k^{11}, V_k^{12})^T$.

Remark 3. Theorem 3 presents the asymptotic normality of homogeneous effects. The conclusion of this theorem and Lemma 5 suggests that $\widehat{\boldsymbol{\alpha}}_{A^*}$ and $\widehat{\boldsymbol{\alpha}}_{ILDA^*}$, the proposed estimator and the benchmark estimator of the homogeneous effects given by minimizing (2.4) and (2.1), can achieve same

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

asymptotic efficiency.

Remark 4. In this paper, we focus on improving the asymptotic efficiency of the homogeneous effects by integrative analysis of multiple data sources. For the heterogeneous effects $\widehat{\boldsymbol{\beta}}_{A_\gamma}$, compared to the naive estimator based on only its own data source, the integrative estimator does not have significant efficiency gain, which was confirmed by extended experimental exploration.

Asymptotic covariance estimation

To estimate the asymptotic covariance matrix of $\widehat{\boldsymbol{\alpha}}_{A^*}$, one approach is to apply the plug-in method to obtain $N^{-1}\tilde{B}^{-1}\tilde{\Sigma}\tilde{B}^{-1}$. In the above, $\tilde{B} = N^{-1}\sum_{k=1}^{K} n_k \tilde{V}_k^{11} + \widehat{\Sigma}_{\lambda_1}$, where $\widehat{\Sigma}_{\lambda_1}$ denotes the estimator of Σ_{λ_1} with its jth diagonal element given by $P'_{\lambda_1}(|\widehat{\alpha}_j|)$, $\tilde{\Sigma} = N^{-1}\sum_{k=1}^{K} n_k(\tilde{V}_k^{11}, \tilde{V}_k^{12})\tau(1-\tau)\tilde{V}_k^{-1}\tilde{U}_k\tilde{V}_k^{-1}(\tilde{V}_k^{11}, \tilde{V}_k^{12})^T$. Here $\tilde{V}_k = n_k^{-1}\sum_{i=1}^{n_k} \mathbf{X}_i^{(k)}\mathbf{X}_i^{(k)T}\widehat{f}_{Y|X}(\mathbf{X}_i^{(k)T}\tilde{\boldsymbol{\beta}}^{(k)}|\mathbf{X}_i^{(k)})$ and $\tilde{U}_k = n_k^{-1}\sum_{i=1}^{n_k} \mathbf{X}_i^{(k)}\mathbf{X}_i^{(k)T}$. This involves the estimation of the conditional density $f_{Y|X}(\mathbf{X}_i^{(k)T}\tilde{\boldsymbol{\beta}}^{(k)}|\mathbf{X}_i^{(k)})$, which is usually done by the kernel method in general, then the involved choice of the bandwidth is quite unstable, especially in high-dimensional situations.

To avoid the nonparametric density estimation, we suggest employing the efficient resampling method discussed in Zeng and Lin (2008) to directly estimate $\tilde{V}_k$. Specifically, let $\Psi_n^{(k)}(\boldsymbol{\beta}^{(k)}) \triangleq n_k^{-1}\sum_{i=1}^{n_k} \mathbf{X}_i^{(k)}(I(Y_i^{(k)} - \mathbf{X}_i^{(k)T}\boldsymbol{\beta}^{(k)} < 0) - \tau) = 0$ be the estimating equation corresponding to mini-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

mizing $L_n^{(k)}(\boldsymbol{\beta}^{(k)})$. One can verify that $n_k\Psi_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)})$ satisfies the asymptotic expansion (2.1) of Zeng and Lin (2008) (see Lemma 3 for further details). Then, the least squares (LS) resampling method proposed by Zeng and Lin (2008) can be used to obtain the estimator $\tilde{V}_k$. More specifically, let $\check{\boldsymbol{\beta}}^{(k)} = \tilde{\boldsymbol{\beta}}^{(k)} + n_k^{-1/2}\mathbf{Z}$, where $\mathbf{Z}$ is a zero-mean p_n -dimensional random vector independent of the data and can be generated from the multivariate normal distribution, as discussed later. Then, we have

$$\sqrt{n_k}\Psi_n^{(k)}(\check{\boldsymbol{\beta}}^{(k)}) = V_k\mathbf{Z} + o_p(1). \tag{3.1}$$

Note each row of V_k can be regarded as the unknown parameter of a linear regression by (3.1). To implement this, we can first generate R realizations of $\mathbf{Z}$, denoted by $\mathbf{Z}_1, \ldots, \mathbf{Z}_R$, calculate $\sqrt{n_k}\Psi_n^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)} + n_k^{-1/2}\mathbf{Z}_r)(r = 1, \ldots, R)$ and the least squares estimator of $\sqrt{n_k}\Psi_{nj}^{(k)}(\tilde{\boldsymbol{\beta}}^{(k)} + n_k^{-1/2}\mathbf{Z}_r)(r = 1, \ldots, R)$ on $\mathbf{Z}_r(r = 1, \ldots, R)$, and then set the jth least squares estimate to be the jth row of $\tilde{V}_k$ for $j = 1, \ldots, p_n$, where $\Psi_{nj}^{(k)}$ denotes the jth component of $\Psi_n^{(k)}$. Note the estimators $\tilde{V}_k$ in (2.3) and (2.4) can also be obtained by the same method.

4. MM algorithm

In this section, we discuss the determination of the proposed estimator $\widehat{\boldsymbol{\beta}}$ or the minimization of the objective function (2.4). This is not straightforward

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

because the penalty function term $\phi(\boldsymbol{\beta})$ is nondifferentiable at the origin and does not have continuous second-order derivatives. To overcome this, we employ the minorization-maximization (MM) algorithm (Hunter and Li, 2005). Then the penalty function $P_\lambda(|b|)$ can be represented by

$$P_\lambda(|b|) = P_\lambda(|b_0|) + \frac{(b^2 - b_0^2)P'_\lambda(|b_0|)}{2(\eta + |b_0|)}, \tag{4.1}$$

where η is a prespecified perturbation to prevent the denominator on the right side of (4.1) from being 0. Then, the $m+1$ step estimates can be updated by

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}^{\{m+1\}}, \widehat{\boldsymbol{\gamma}}^{\{m+1\}}) \;=\; \arg\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \Bigg\{ & L_N^*(\boldsymbol{\alpha},\boldsymbol{\gamma}) + \sum_{j=2}^{p_n} \frac{(\alpha_j^2 - \widehat{\alpha}_j^{\{m\}2})P'_{\lambda_1}(|\widehat{\alpha}_j^{\{m\}}|)}{2(\eta_1 + |\widehat{\alpha}_j^{\{m\}}|)} \\ & + \sum_{j=2}^{p_n} \frac{(\|\boldsymbol{\gamma}_j\|^2 - \|\widehat{\boldsymbol{\gamma}}_j^{\{m\}}\|^2)P'_{\lambda_2}(\|\widehat{\boldsymbol{\gamma}}_j^{\{m\}}\|)}{2(\eta_2 + \|\widehat{\boldsymbol{\gamma}}_j^{\{m\}}\|)} \Bigg\}.\end{aligned}$$

In the above, following Hunter and Li (2005), we propose to fix $\eta_1 = \frac{\varepsilon}{2N\lambda_1}\min\{|\widehat{\alpha}_j^{\{0\}}| : \widehat{\alpha}_j^{\{0\}} \neq 0\}$, and $\eta_2 = \frac{\varepsilon}{2N\lambda_2}\min\{\|\widehat{\boldsymbol{\gamma}}_j^{\{0\}}\| : \|\widehat{\boldsymbol{\gamma}}_j^{\{0\}}\| \neq 0\}$ for a given tolerance ε and then apply the Newton-Raphson algorithm. For the initial estimates, a good choice is $(\widehat{\boldsymbol{\alpha}}^{\{0\}}, \widehat{\boldsymbol{\gamma}}^{\{0\}}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\gamma}} L_N^*(\boldsymbol{\alpha},\boldsymbol{\gamma})$.

For the selection of the tuning parameters λ_1 and λ_2, which control the sparsity of the model and are essential in the penalized procedure, a commonly used approach is the cross-validation criterion. However, it is not feasible here because individual-level data are unavailable. We propose to minimize the generalized information criterion (GIC), defined as $GIC(\lambda_1, \lambda_2) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

$L_N^*(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) + \vartheta_n df(\lambda_1, \lambda_2)$ (Zhang et al., 2010), where ϑ_n is a positive number that controls the properties of variable selection and $df(\lambda_1, \lambda_2)$ is the degrees of freedom of the model. Following Zhang et al. (2010), we employ $df(\lambda_1, \lambda_2) = \text{tr}[(\nabla^2_{A_{\widehat{\alpha}}, A_{\widehat{\gamma}}} Q_N(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}))^{-1} \nabla^2_{A_{\widehat{\alpha}}, A_{\widehat{\gamma}}} L_N^*(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}))]$, where tr(.) represents the trace of a matrix and ∇^2 denotes the second order partial derivative with respect to $(\boldsymbol{\alpha}^T_{A_{\widehat{\alpha}}}, \boldsymbol{\gamma}^{(1)T}_{A_{\widehat{\gamma}}}, \ldots, \boldsymbol{\gamma}^{(k-1)T}_{A_{\widehat{\gamma}}}, \boldsymbol{\gamma}^{(k+1)T}_{A_{\widehat{\gamma}}}, \ldots, \boldsymbol{\gamma}^{(K)T}_{A_{\widehat{\gamma}}})$ after reparameterizing by plugging $\boldsymbol{\gamma}^{(k)} = -(\boldsymbol{\gamma}^{(1)} + \ldots + \boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k+1)} + \boldsymbol{\gamma}^{(K)})$ into $Q_N(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ or $L_N^*(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. Further, let $\vartheta_n = \log N/N$, which leads to GIC being the Bayesian information criterion (BIC). Meanwhile, based on the conditions of Theorem 3, we provide the tuning parameters (λ_1, λ_2) with a rough reference range $\left(\log(n)^{c_1}\sqrt{p_n/n}, \log(n)^{c_2}\sqrt{p_n/n}\right)$ for some $c_1, c_2 > 0$. And the tuning parameters λ_1, λ_2 are selected by searching on a two-dimensional grid to minimize the BIC. For the parameter a above, following Fan and Li (2001), we will set $a = 3.7$. In the numerical studies below, we evaluate the number of zero coefficients such that an estimate is treated as zero if its absolute value is smaller than 10^{-6}.

5. A simulation study

We conduct a simulation study to evaluate the proposed method's empirical performance. Consider the situation $K = 4$, $p_n = 50$, and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

$[n_1, n_2, n_3, n_4] = [1000, 1200, 1400, 1600]$. In addition, we set $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ as

$$\begin{pmatrix} \boldsymbol{\alpha}^T \\ \boldsymbol{\gamma}^{(1)T} \\ \boldsymbol{\gamma}^{(2)T} \\ \boldsymbol{\gamma}^{(3)T} \\ \boldsymbol{\gamma}^{(4)T} \end{pmatrix} = \begin{pmatrix} 0 & 2.5_{1\times 5} & 2.5_{1\times 5} & 0_{1\times 5} & 0_{p_n-16} \\ 0 & 2.5_{1\times 5} & 0_{1\times 5} & 2.5_{1\times 5} & 0_{p_n-16} \\ 0 & -2.5_{1\times 5} & 0_{1\times 5} & -2.5_{1\times 5} & 0_{p_n-16} \\ 0 & -2.5_{1\times 5} & 0_{1\times 5} & 2.5_{1\times 5} & 0_{p_n-16} \\ 0 & 2.5_{1\times 5} & 0_{1\times 5} & -2.5_{1\times 5} & 0_{p_n-16} \end{pmatrix},$$

and generate the data from $Y_i^{(k)} = \mathbf{X}_i^{(k)T}\boldsymbol{\beta}^{(k)} + \epsilon_i^{(k)}$, where $\boldsymbol{\beta}^{(k)} = \boldsymbol{\alpha} + \boldsymbol{\gamma}^{(k)}$. In the above, we assume that $\epsilon_i^{(k)} \sim N(0,1)$ or $\epsilon_i^{(k)} \sim N(0, (1+0.3X_{i7}^{(k)})^2)$, corresponding to the location-shift model or the location-scale-shift model, respectively; $X_i^{(k)}$ follows the multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\rho})$ with $\boldsymbol{\rho}_{uv} = 0.5^{|u-v|}$ and contains an intercept. Thus, the τth conditional quantile of $Y_i^{(k)}$ is given by $\mathbb{Q}_\tau(Y_i^{(k)}|\mathbf{X}_i^{(k)T}) = \mathbf{X}_i^{(k)T}\boldsymbol{\beta}_\tau^{(k)}$. Here, $\boldsymbol{\beta}_\tau^{(k)} = \boldsymbol{\beta}^{(k)} + \Phi_\tau^{-1}\boldsymbol{\xi}_1$ or $\boldsymbol{\beta}_\tau^{(k)} = \boldsymbol{\beta}^{(k)} + \Phi_\tau^{-1}\boldsymbol{\xi}_1 + 0.3\Phi_\tau^{-1}\boldsymbol{\xi}_7$, corresponding to the location-shift or location-scale-shift model, respectively; Φ_τ^{-1} denotes the τ-th quantile of the standard normal distribution, and $\boldsymbol{\xi}_j (j = 1, 7)$ is a p_n-dimensional vector with the jth element being one and all other elements being zero.

To assess the performance of the proposed method (denoted as PPD below), we considered two types of quantities. (1) Identification accuracy: the true zero rate (TZR) and the error rate (ER) for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ to identify the homogeneity and heterogeneity of covariates, which are de-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

fined as $TZR(\boldsymbol{\gamma}) = \frac{1}{M}\sum_{m=1}^{M} \frac{\sum_{j=2}^{p_n} I(\|\boldsymbol{\gamma}_j\|=0) I(\|\widehat{\boldsymbol{\gamma}}_j^{\{m\}}\|=0)}{\sum_{j=2}^{p_n} I(\|\boldsymbol{\gamma}_j\|=0)} \times 100\%$ and $ER(\boldsymbol{\gamma}) = \big\{1 - TZR(\boldsymbol{\gamma}) + \frac{1}{M}\sum_{m=1}^{M} \frac{\sum_{j=2}^{p_n} I(\|\boldsymbol{\gamma}_j\|\neq 0) I(\|\widehat{\boldsymbol{\gamma}}_j^{\{m\}}\|=0)}{\sum_{j=2}^{p_n} I(\|\boldsymbol{\gamma}_j\|\neq 0)}\big\} \times 100\%$ with respect to $\boldsymbol{\gamma}$, where M denotes the number of replications. The definition of TZR and ER with respect to $\boldsymbol{\alpha}$ are the same. (2) Estimation accuracy: the absolute estimation error (AE) $\| \boldsymbol{\beta}_\tau - \widehat{\boldsymbol{\beta}}_\tau \|_1$.

In addition, we calculated the empirical bias (BIAS) of the estimates of the homogeneous effect $\alpha_7 \sim \alpha_{11}$, the standard deviation (SD), the standard errors of the estimators (SE) which is obtained by the resampling based variance estimation method given in Section 3, and the 95% empirical coverage probability (COV). For comparison, we also considered the benchmark estimator given by the method based on individual-level data, and refer to it as ILD. The results given below are on three quantile levels $\tau = \{0.25, 0.5, 0.75\}$ with 500 replications. Additional simulations involving more residual distributions and comparisons with integrative linear regressions are included in the Supplementary Material, due to space limitation.

Tables 1 and 2 present the simulation results obtained under the location-shift model, and the results on TZR and ER in Table 1 indicate both the proposed and benchmark estimators can well identify the correct covariate structure. In terms of estimation accuracy, the proposed estimates have slightly larger AEs than the benchmark estimator, but the difference is

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 1: Simulation results based on the location-shift model.

τ	Method	TZR		ER		AE($\times 10^{-2}$)
		$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	
0.25	PPD	99.78	99.94	0.22	0.06	14.35
	ILD	100.00	100.00	0.00	0.00	10.45
0.5	PPD	100.00	100.00	0.00	0.00	13.65
	ILD	100.00	100.00	0.00	0.00	6.39
0.75	PPD	99.75	99.93	0.25	0.07	15.29
	ILD	100.00	100.00	0.00	0.00	10.94

quite small. Table 2 provides the results of the homogeneous effects. We can see the proposed estimator seems to be unbiased and the proposed variance estimation also seems to perform well. In particular, the variance estimates for the two estimators are close to each other. In addition, the coverage probability of the PPD under different quantile levels is almost all around the nominal level of 95%, which ensures the accuracy of the statistical inference on the covariates with homogeneous effects. Overall the PPD is competitive with the ILD.

Simulation results under the location-scale-shift model are presented in Tables 3 and 4 and show similar trends to those obtained under the location-shift model. The high TZR and low ER in Table 3 demonstrate the PPD's accuracy in identifying covariate effects, with its AE asymptotically equal

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 2: Simulation results about the homogeneous effects based on the location-shift model (all entries are multiplied by 100).

τ	Method		α_7	α_8	α_9	α_{10}	α_{11}
0.25	PPD	BIAS	0.04	-0.03	0.11	0.14	0.19
		SD	2.41	2.37	2.39	2.42	2.28
		SE	2.59	2.58	2.61	2.33	2.31
		COV	97.40	96.20	96.20	93.40	95.20
	ILD	BIAS	-0.07	0.13	0.03	-0.07	0.00
		SD	2.60	2.55	2.54	2.39	2.19
		SE	2.32	2.58	2.58	2.59	2.31
		COV	93.60	96.00	96.20	93.80	95.60
0.5	PPD	BIAS	0.04	0.08	0.13	0.30	0.26
		SD	2.39	2.51	2.50	2.69	2.27
		SE	2.33	2.61	2.60	2.60	2.33
		COV	96.80	95.80	95.40	92.40	95.40
	ILD	BIAS	-0.06	0.04	0.03	-0.05	0.05
		SD	2.29	2.44	2.44	2.27	2.07
		SE	2.25	2.51	2.52	2.51	2.25
		COV	96.20	96.00	96.20	94.20	95.40
0.75	PPD	BIAS	0.21	0.17	0.16	0.03	0.26
		SD	2.52	2.53	2.66	2.64	2.19
		SE	2.68	2.69	2.67	2.40	2.39
		COV	95.99	95.99	95.39	93.39	95.99
	ILD	BIAS	0.00	-0.04	0.23	-0.09	0.02
		SD	2.57	2.59	2.50	2.50	2.35
		SE	2.39	2.67	2.68	2.68	2.40
		COV	95.20	95.60	95.40	93.60	95.40

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

to those of the ILD. Furthermore, the PPD's four statistical performance evaluation indicators in Table 4 are comparable to those of the ILD on covariates with homogeneous effects. In summary, our proposed estimator with privacy preservation and the benchmark estimator based on the ILD are asymptotically equivalent, as shown by the simulation results.

Table 3: Simulation results based on location-scale-shift model.

τ	Method	TZR		ER		AE($\times 10^{-2}$)
		$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	
0.25	PPD	99.79	99.85	0.21	0.15	13.78
	ILD	100.00	100.00	0.00	0.00	11.68
0.5	PPD	99.93	99.98	0.07	0.02	12.32
	ILD	97.60	100.00	2.40	0.00	7.28
0.75	PPD	98.90	99.56	1.10	0.44	14.32
	ILD	100.00	100.00	0.00	0.00	12.87

6. An application

We apply the proposed method to the data set of the Chinese Annual Survey of Industrial Firms (ASIF), which is conducted for all industrial firms whose annual sales more than RMB 5 million. This dataset collects the firm's detailed information including name, identification, ownership, balance sheet, profit and loss, and cash flow. With approximately

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 4: Simulation results about the homogeneous effects based on the location-scale-shift model (all entries are multiplied by 100).

τ	Method		α_7	α_8	α_9	α_{10}	α_{11}
0.25	PPD	BIAS	0.98	-0.09	0.06	0.02	0.21
		SD	2.12	2.23	2.18	2.25	1.88
		SE	2.48	2.49	2.48	2.23	2.06
		COV	95.59	97.39	97.60	94.39	96.39
	ILD	BIAS	0.67	0.02	0.04	-0.04	-0.11
		SD	2.22	2.24	2.15	2.25	2.07
		SE	2.01	2.39	2.38	2.38	2.13
		COV	95.40	96.00	97.00	92.60	93.60
0.5	PPD	BIAS	0.26	0.00	0.15	0.19	0.16
		SD	2.15	2.13	2.42	2.29	1.96
		SE	2.31	2.31	2.32	2.07	1.94
		COV	96.20	97.20	94.40	92.60	95.80
	ILD	BIAS	0.10	-0.03	0.04	0.04	-0.07
		SD	2.01	2.24	2.11	2.10	1.95
		SE	1.95	2.32	2.32	2.32	2.07
		COV	96.80	95.60	97.00	94.80	95.60
0.75	PPD	BIAS	-0.94	0.12	0.09	0.08	0.16
		SD	2.17	2.25	2.40	2.19	1.90
		SE	2.46	2.45	2.44	2.20	1.98
		COV	95.80	96.60	94.80	94.00	96.40
	ILD	BIAS	-0.51	-0.08	0.09	0.06	-0.11
		SD	2.25	2.26	2.34	2.20	2.07
		SE	1.93	2.36	2.36	2.36	2.12
		COV	95.80	96.20	96.60	94.40	92.40

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

400,000 firms surveyed annually, we have more than 1 million observations for the years 2001-2007. An interesting note is how variables affect the company's total factor productivity(TFP), which computed via Olley and Pakes (1996)'s method.

In this paper, we focus on the TFP performance of China's four major industrial cities Shanghai, Ningbo, Hangzhou, and Suzhou in 2006. Specifically, one goal of the study is to assess the effects of various variables on TFP in the four major industrial cities and to determine the homogeneous effects, heterogeneous effects and null effects. The variable considered are company age, company scale, fixed asset ratio, etc. Due to space limitation, the complete variable information is described in Section S3 of the Supplementary Material. In addition, we also introduce an intercept term (Int). In summary, $p_n = 25$, $K = 4$ and $[n_1, n_2, n_3, n_4] = [5735, 4608, 3463, 4170]$.

Tables 5 - 7 present the analysis produced by the proposed method based on the quantile levels $\tau = 0.25, 0.5, 0.75$, respectively, including the estimated $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ along with the estimated standard errors for the identified covariates with homogeneous effects (in parentheses). From Table 5, we can find that under the 25th quantile level, the AgeS , Ind24, Ind34 and Ind36 are detected as the null effects, while Age, DebtR, Worker, Ind33, Ind39 and Ind40 are detected as homogeneous effects. Using our method,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

the remaining 14 covariates are detected as heterogeneous effects. From Tables 6 and 7, we find that for median and 75th quantile level, Age and Ind39 also have homogeneous effects while DebtR and Worker become heterogeneous. Moreover, the homogeneous properties for other industry codes are different under three different quantile levels.

Among the heterogeneous effects, the scale of enterprises, which has been claimed to have a direct effect on TFP possibly due to the scale of economies (Geroski, 1998), exhibits obvious heterogeneity in different cities because enterprises of different sizes in different cities have considerable differences in industry layout, industry segmentation, and innovation. Regarding the homogeneous covariates, most of homogeneous effects on TFP is significant under the significance level of 5%. First, the coefficient of Age indicates the TFP will decrease as a company's age increasing by one year. A possible explanation is that a company's TFP development exhibits some regularity in its life cycle, with younger firms often having more advanced technology than older firms, resulting in higher TFP (Demir et al., 2022). Hence, the effects of age on TFP are homogeneous for firms in different areas. The influence of debt ratio on firm TFP is homogeneous in the 25th quantile but heterogeneous in the other quantiles. This might occur because firms in lower quantiles rely more on bank credit, and bank credit

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

costs vary little in different regions; bank credit costs are mainly affected by the benchmark interest rate set by the Peoples' Bank of China. As a result, the same debt burden has a consistent impact on firms in different regions. However, this phenomenon does not exist in high-quantile estimates because firms with higher TFP have greater leverage and more diverse financing channels and finance a greater share of investments through equity (Zhang and Liu, 2017), and the cost of debt varies greatly among firms in different regions. The influence of the number of employees on TFP is also homogeneous in the 25th quantile, but heterogeneous in the other quantiles. This may occur because productivity exerts a positive effect on employment (Mollick and Cabral, 2009) and thus firms with lower TFP have lower employee requirements. Although the average salary level of Shanghai, Hangzhou, Ningbo and Suzhou differ to some extent, the gap in the minimum wage is quite small. In 2006, the average salary in Shanghai, Hangzhou, Ningbo, and Suzhou is 3232 RMB, 2703 RMB, 2408 RMB, and 2334 RMB, respectively, whereas the minimum wage are the same (i.e, 750 RMB per month). Therefore, the labor costs faced by firms in the lower TFP quantile are almost homogeneous. However, in firms with high TFP scores, firms' employee requirements differ considerably in different regions, and the wage level gap in different cities is sizable as well since

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

workers' wages also increase with TFP (Chan et al., 2020). This leads to the heterogeneous effect of the number of employees on TFP in the higher quantiles. Finally, let's look at the homogeneity performance of different industries in the four industrial cities. The coefficient estimates of Ind33, Ind39 and Ind40 all indicate they have more TFP than the textile industry, potentially because the textile industry is more labor-intensive and its TFP is generally low (Li and Lv, 2021). Specifically, Ind33, Ind39, and Ind40 have 10.5, 6.551 and 2.574 more TFP than textile industry, respectively. Overall, when there may be heterogeneity and high-dimensional covariates in different industrial cities, we do our best to identify the homogeneity effect; further improve the effectiveness of the homogeneity coefficient, and via the identification of null effects, concurrently process high-dimensional issues.

7. Discussion and concluding remarks

In this paper, we discussed an integrative QR analysis of high-dimensional and heterogeneous data from multiple sources with privacy preservation. We propose a composite penalized objective function based only on the summary statistics from each data source, and we establish the asymptotic properties of the proposed estimators for the problem, including both es-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 5: Analysis results for the ASIF data with $\tau = 0.25$.

Variable	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}^{(Shanghai)}$	$\boldsymbol{\beta}^{(Ningbo)}$	$\boldsymbol{\beta}^{(Hangzhou)}$	$\boldsymbol{\beta}^{(Suzhou)}$
Int	22.374	2.227	47.503	17.946	21.822
Age	-0.070(0.016)	-0.070	-0.070	-0.070	-0.070
AgeS	0.000	0.000	0.000	0.000	0.000
Asset	2.460	5.571	1.079	2.896	0.294
DebtR	-7.413(0.717)	-7.413	-7.413	-7.413	-7.413
FixR	-33.417	-63.038	-24.310	-32.381	-13.939
ExpR	0.000	-6.600	5.983	-0.664	1.281
Worker	-3.131(0.084)	-3.131	-3.131	-3.131	-3.131
ScaleL	25.246	6.706	22.038	26.791	45.448
ScaleS	-2.119	-5.880	1.016	-14.543	10.930
Col	0.002	10.740	-8.666	15.118	-17.186
LLC	8.459	9.343	-17.056	3.884	37.663
LBS	9.621	23.470	0.944	2.442	11.626
Pri	10.077	15.727	-11.090	17.792	17.877
HMT	4.128	18.005	-22.829	21.775	-0.439
Fore	6.642	17.665	-20.633	9.044	20.493
Ind24	0.000	0.000	0.000	0.000	0.000
Ind26	6.962	8.165	-0.568	1.988	18.263
Ind29	2.028	4.553	-1.803	-12.600	17.964
Ind33	10.500(0.128)	10.500	10.500	10.500	10.500
Ind34	0.000	0.000	0.000	0.000	0.000
Ind36	0.000	0.000	0.000	0.000	0.000
Ind39	6.551(0.100)	6.551	6.551	6.551	6.551
Ind40	2.574(0.324)	2.574	2.574	2.574	2.574
Ind41	6.688	-4.302	2.947	18.950	9.154

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 6: Analysis results for the ASIF data with $\tau = 0.5$.

Variable	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}^{(Shanghai)}$	$\boldsymbol{\beta}^{(Ningbo)}$	$\boldsymbol{\beta}^{(Hangzhou)}$	$\boldsymbol{\beta}^{(Suzhou)}$
Int	19.088	-25.042	60.240	9.462	31.693
Age	-0.128(0.018)	-0.128	-0.128	-0.128	-0.128
AgeS	0.000	0.000	0.000	0.000	0.000
Asset	8.591	15.826	2.274	5.124	11.141
DebtR	-13.695	-10.784	-9.049	-15.877	-19.069
FixR	-59.591	-99.291	-45.809	-34.212	-59.050
ExpR	0.000	-5.011	-1.883	5.514	1.380
Worker	-7.531	-12.538	-0.244	-2.732	-14.609
ScaleL	27.956	7.341	30.133	35.030	39.319
ScaleS	-6.716	-7.655	-2.962	-2.723	-13.523
Col	4.189	11.915	-27.613	3.025	29.427
LLC	2.667	11.029	-26.606	13.388	12.855
LBS	5.200	-0.023	-23.311	24.856	19.277
Pri	2.216	21.229	-32.957	7.108	13.485
HMT	0.000	11.394	-31.936	12.251	8.291
Fore	3.366	22.550	-31.857	2.294	20.479
Ind24	1.969	0.927	5.996	8.502	-7.550
Ind26	7.341	4.704	4.882	-3.807	23.586
Ind29	-1.267	-5.725	20.198	-15.747	-3.792
Ind33	11.374	15.456	-0.148	11.909	18.281
Ind34	0.298(0.023)	0.298	0.298	0.298	0.298
Ind36	0.761(0.566)	0.761	0.761	0.761	0.761
Ind39	1.609(0.533)	1.609	1.609	1.609	1.609
Ind40	-0.466(0.590)	-0.466	-0.466	-0.466	-0.466
Ind41	1.752	8.494	-9.569	2.131	5.952

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Table 7: Analysis results for the ASIF data with $\tau = 0.75$.

Variable	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}^{(Shanghai)}$	$\boldsymbol{\beta}^{(Ningbo)}$	$\boldsymbol{\beta}^{(Hangzhou)}$	$\boldsymbol{\beta}^{(Suzhou)}$
Int	55.579	-52.291	184.849	-1.360	91.117
Age	-0.058(0.009)	-0.058	-0.058	-0.058	-0.058
AgeS	0.000	0.000	0.000	0.000	0.000
Asset	15.207	31.770	6.048	8.785	14.223
DebtR	-13.208	-8.103	-19.590	-17.629	-7.511
FixR	-91.445	-146.906	-49.679	-61.937	-107.260
ExpR	0.000	-16.327	0.037	10.983	5.308
Worker	-14.559	-28.075	-4.229	-4.496	-21.435
ScaleL	38.587	15.748	22.302	56.141	60.157
ScaleS	-7.707	-3.576	-1.788	-1.972	-23.491
Col	-37.077	9.481	-151.095	30.686	-37.381
LLC	-45.105	-1.404	-159.265	3.999	-23.750
LBS	-32.042	-15.469	-152.932	60.758	-20.527
Pri	-42.413	-7.107	-156.563	14.674	-20.655
HMT	-45.448	-4.561	-160.118	14.798	-31.913
Fore	-30.138	14.178	-147.399	19.368	-6.698
Ind24	-0.294(0.021)	-0.294	-0.294	-0.294	-0.294
Ind26	21.117	4.595	23.463	18.434	37.976
Ind29	-9.651	-14.275	-2.465	-2.155	-19.709
Ind33	35.852	49.082	11.380	52.635	30.310
Ind34	-6.137	-18.847	-3.804	-7.311	5.413
Ind36	-2.157	-22.432	0.607	2.660	10.538
Ind39	1.849(0.881)	1.849	1.849	1.849	1.849
Ind40	0.000	-7.541	1.986	2.207	3.347
Ind41	-1.451(2.038)	-1.451	-1.451	-1.451	-1.451

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

timation and selection consistency and the asymptotic normality. For the implementation of the proposed approach, we develop an MM algorithm. The proposed method allows us to detect the homogeneous, heterogeneous and nonzero covariate effects, and it enables the number of covariates to go to infinity as the smallest sample size among different data sources goes to infinity. The numerical results indicate it works well in practical situations.

In this article, we have assumed the number of covariates is smaller than the sample size for each data source. However, it would be useful to generalize the proposed method to the situation where the number of covariates is larger than these sample sizes. For this, we will need to address the local variables selection issue, and the establishment of relevant theories is also challenging because the objective function is non-smooth. As discussed earlier, the proposed approach preserves privacy in the sense that it uses only summary statistics instead of individual, original data. Thus, we may investigate the same problem under different privacy mechanisms (Dwork and Roth, 2014).

Supplementary Material

We provide the proof details of our main results and additional simulations. In addition, we provide variable information in practical application.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Acknowledgments

The authors are grateful to the editor, associate editor and two referees for very helpful comments. Chen's work was supported by National Key R&D Program of China (2022YFA1003702), the National Natural Science Foundation of China (NSFC) (12371296, 11931014), and Guanghua Talent Project of Southwestern University of Finance and Economics.

References

Cai, T., M. Liu, and Y. Xia (2022). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association 117*(540), 2105–2119.

Chan, M., S. Salgado, and M. Xu (2020). Heterogeneous passthrough from tfp to wages. *Working paper*.

Chen, L. and Y. Zhou (2020). Quantile regression in big data: A divide and conquer based strategy. *Computational Stats & Data Analysis 144*, 106892.

Chen, X., W. Liu, X. Mao, and Z. Yang (2020). Distributed high-dimensional regression under a quantile loss function. *Journal of Machine Learning Research 21*(182), 1–43.

Chen, X., W. Liu, and Y. Zhang (2019). Quantile regression under memory constraint. *The Annals of Statistics 47*(6), 3244–3273.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

REFERENCES

Chen, X., A. T. K. Wan, and Y. Zhou (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association 110*(510), 723–741.

Cheng, X., W. Lu, and M. Liu (2015). Identification of homogeneous and heterogeneous variables in pooled cohort studies. *Biometrics 71*(2), 397–403.

Demir, F., C. Hu, J. Liu, and H. Shen (2022). Local corruption, total factor productivity and firm heterogeneity: Empirical evidence from chinese manufacturing firms. *World Development 151*, 105770.

Duan, R., Y. Ning, and Y. Chen (2021). Heterogeneity-aware and communication-efficient distributed statistical inference. *Biometrika 109*(1), 67–83.

Dwork, C. and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*(3-4), 211–407.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and H. Peng (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics 32*(3), 928–961.

Geroski, P. A. (1998). An applied econometrician's view of large company performance. *Review of Industrial Organization 13*, 271–294.

Ghosh, A., J. Hong, D. Yin, and K. Ramchandran (2019). Robust federated learning in a heterogeneous environment. *ArXiv abs/1906.06629*.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

REFERENCES

Huang, J., M. Wang, and Y. Wu (2022). Transfer learning with high-dimensional quantile regression. *arXiv preprint arXiv:2211.14578*.

Huang, J. and H. Xie (2007). Asymptotic oracle properties of scad-penalized least squares estimators. *Lecture Notes-Monograph Series 55*(2), 149–166.

Hunter, D. R. and R. Li (2005). Variable selection using mm algorithms. *The Annals of Stats 33*(4), 1617–1642.

Jin, J., J. Yan, and K. Chen (2022). Transfer learning with quantile regression. *arXiv preprint arXiv:2212.06693*.

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research 18*(1), 115–144.

Li, Z. and B. Lv (2021). Total factor productivity of chinese industrial firms: evidence from 2007 to 2017. *Applied Economics 53*(60), 6910–6926.

Liang, X., S. Li, S. Zhang, H. Huang, and S. X. Chen (2016). $PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities. *Journal of Geophysical Research (Atmospheres) 121*(17), 10220–10236.

Lin, D. Y. and D. Zeng (2010). On the relative efficiency of using summary statistics versus

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

individual-level data in meta-analysis. *Biometrika 97*(2), 321–332.

Lin, N. and R. Xi (2011). Aggregated estimating equation estimation. *Statistics and its Interface 4*(1), 73–83.

Liu, D., R. Y. Liu, and M. Xie (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *Journal of the American Statistical Association 110*(509), 326–340.

Mollick, A. V. and R. Cabral (2009). Productivity effects on mexican manufacturing employment. *The North American Journal of Economics and Finance 20*(1), 66–81.

Moniz, N. and L. Torgo (2018). Multi-source social feedback of online news feeds. *arXiv preprint arXiv:1801.07055*.

Olley, G. S. and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica 64*(6), 1263–1297.

Tibshirani, R., M. Wainwright, and T. Hastie (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Volgushev, S., S. Chao, and G. Cheng (2019). Distributed inference for quantile regression processes. *Annals of Statistics 47*(3), 1634–1662.

Wang, L. and H. Lian (2020). Communication-efficient estimation of high-dimensional quantile regression. *Analysis and Applications 18*(06), 1057–1075.

Zeng, D. and D. Lin (2008). Efficient resampling methods for nonsmooth estimating functions.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0292

Biostatistics 9(2), 355–363.

Zhang, D. and D. Liu (2017). Determinants of the capital structure of chinese non-listed enterprises: Is tfp efficient? *Economic Systems 41*(2), 179–202.

Zhang, Y., J. Duchi, and M. Wainwright (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research 16*(1), 3299–3340.

Zhang, Y., R. Li, and C. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association 105*(489), 312–323.

Zhang, Y. and Z. Zhu (2022). Transfer learning for high-dimensional quantile regression via convolution smoothing. *arXiv preprint arXiv:2212.00428*.

Center of Statistical Research, Southwestern University of Finance and Economics

E-mail: yuansenlin26@163.com

Center of Statistical Research, Southwestern University of Finance and Economics

E-mail: chenxuerong@swufe.edu.cn, chenxr522@foxmail.com

School of Finance, Nanjing Agricultural University

E-mail: wuyu@njau.edu.cn

Department of Statistics, University of Missouri

E-mail: sunj@missouri.edu