| Statistica Sinica Preprint No: SS-2021-0276 | |
|---|---|
| **Title** | Integrative Analysis for High-Dimensional Stratified Models |
| **Manuscript ID** | SS-2021-0276 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202021.0276 |
| **Complete List of Authors** | Jian Huang,<br>Yuling Jiao,<br>Wei Wang,<br>Xiaodong Yan and<br>Liping Zhu |
| **Corresponding Author** | Xiaodong Yan |
| **E-mail** | yanxiaodong128@163.com |
| Notice: Accepted version subject to English editing. | |

Statistica Sinica

# Integrative analysis for high-dimensional stratified models

Jian Huang[a], Yuling Jiao[b], Wei Wang[c], Xiaodong Yan[c*], Liping Zhu[d]

[a] *Department of Statistics and Actuarial Science, University of Iowa*

[b] *School of Mathematics and Statistics, Wuhan University*

[c] *Zhongtai Securities Institute for Financial Studies, Shandong University*

[d] *Institute of Statistics and Big Data, Renmin University of China*

*Abstract:* In modern economic studies, the population heterogeneity of multiple strata and the high dimensionality of the predictors pose a major challenge. In this study, we introduce an integrative procedure that can be used to explore the information regarding group and sparsity structures for high-dimensional and heterogeneous stratified models. Further, we propose $K$-regression modeling as a hybrid of complex and simple models exhibiting arbitrary dependence on the stratum features, but linear dependence on other variables. $K$-regression models preeminently exhibit the following features:(i) they are essentially non-parametric with respect to the stratified feature, and parametric linearly effects in other variables with potentially integrative pattern because the effects and the corresponding sparsity structures can be the same for the strata in common groups but vary across different groups; (ii) the devised $K$-regression algorithm can automatically integrate the strata pertaining to common regression model and simultaneously estimate the corresponding effects simultaneously; (iii) the proposal quickly

recovers the subpopulation and sparsity structure of the $K$-regression models within massive and high-dimensional strata; (iv) the resulting estimators exhibit two-layer oracle properties, i.e., the oracle estimator obtained using the known group and sparsity structures is the local minimizer of the objective function with high probability. The stratum-specific bootstrap (SSB) sampling scheme was developed to improve the integration accuracy. Furthermore, the simulation studies provide supportive evidence that the newly proposed method performs appropriately in case of finite samples; a real data example has been provided for illustration.

*Key words and phrases:* $K$-regression; Integrative analysis; Heterogeneity; Group fixed effect; Massive data; High-dimensionality; Stratum-specific bootstrap.

## 1. Introduction

Stratum is usually considered to be a method that fits stratified models for the value of each categorical feature and is routinely applied to various econometric issues. For example, financial shares are often classified based on their prices or trading volumes, and the firms are classified into subgroups based on the pairwise interaction of their credit ratings and industry attributes. Currently, stratum is transformed into a universal state with explosive growth of raw data being observed in many disciplines across the social and econometric sciences (Varian, 2014; Einav and Levin, 2014).

This is because big data generally characterize the variety of sources of massive datasets, enabling multiple strata from different backgrounds, such as different experimental methods, distinct geographic locations (census, tract, county, state, etc.), external classification, observable explanatory categories, or nested (hierarchical) or non-nested datasets, pooled cross-sectional datasets and panel datasets.

Although the homogeneous assumption (Phillips and Sul, 2007; Browning and Carro, 2007; and Su and Chen, 2013) considerably facilitates the estimation and inference procedures for certain specified common parameters, it could be misleading if multiple strata exhibit a heterogeneous structure. The stratum phenomenon in a big data scenario can be usually elucidated so that the populations could be heterogeneous across strata due to the different dataset sources (Zhao *et al.*, 2016). Several studies have investigated the crucial importance of stratified models and controlling latent heterogeneity with respect to the panel data models by regarding an individual as a stratum (Pesaran and Tosetti, 2011; Su and Jin, 2012; Song, 2013; Li and Lu, 2014; Chudik and Pesaran, 2015). However, the development of an alternative that only concentrates on modeling in each stratum was inadvisable because of little observations in each stratum causing "incidental parameter" issues such as panel data models (Hsiao and Pesaran, 2008;

Lu *et al.*, 2016). Therefore, recent literatures depicted the heterogeneity by assuming that multiple strata belong to several homogeneous groups within a broadly heterogeneous population(Lin and Ng, 2011; Bester and Hansen, 2016; Bonhomme and Manresa, 2015), i.e., the regression parameters exhibiting group patterns are identical within each group but different across groups and the observations in each stratum are obviously associated with a common population (Su *et al.*, 2016; Su and Ju, 2018; Sarafidis and Weber, 2015). Additionally, time-specified stratum (Bai, 2010; Kim, 2011) can be generated in the panel data, where multiple strata are obtained in different times and one stratum includes observations of some subjects in each time point (Qian and Su, 2016; Li *et al.*, 2017). Ma and Huang (2017, 2018) proposed a pairwise-fusion penalized approach to conduct subgroup analysis for heterogeneous intercepts and coefficients.

Variety or heterogeneity of big data is ubiquitous, usually coupled with high dimensionality. However, the existing methods cannot be directly used to analyze such stratified models that are subject to different sparsity structures across different latent groups to simultaneously achieve high dimensionality and heterogeneity. Herein, we aim to explore the common features among multiple strata that exhibit high dimensionality by conducting an integrative procedure that accurately combines the strata that originally

belong to a common group into one group and estimate the sparsity structure of group-specific parameters. Our devised penalty-based $K$-regression demonstrates the followings: (i) the $K$ stratified regression models serve as a hybrid of complex and simple models, implying that it is arbitrarily dependent on the stratum feature, but that it is simply (typically linear) dependent on other covariates, i.e., it is essentially non-parametric with respect to the stratified features, and parametric with a simple form in case of other variables; (ii) the numerical algorithm denotes the computational ease and speed with respect to the integration of the common structure and recovery of the sparsity information in each stratum; (iii) the resulting oracle estimator with *a priori* knowledge of the group direction and sparsity information in each stratum is a local minimizer of the proposed objective function with high probability.

The rest of this paper is organized as follows. Section 2 describes the proposed $K$-regression model and a fast iterative algorithm. Section 3 establishes the theoretical properties. An SSB sampling scheme is proposed to reduce the integrating error in Section 4. The finite-sample performance of the proposed method are evaluated in Sections 5 and 6. The concluding remarks and the technical proof is provided in Section 7 and Appendix.

## 2. Models and estimators

### 2.1 The Kregression method

Let us assume that we observe data items or records of the form $(Z_i, X_i, Y_i)$, $i = 1, \cdots, n$ with a triple population form $(Z, X, Y)$. Here, $Z_i$ is the ordered or unordered categorical variable with $M$ classes $\mathcal{Z} = \{z_1, z_2, \cdots, z_M\}$, based on which we will conduct stratum, and the corresponding sample size with respect to the $z_m$ stratum can be given as $n_m = \sum_{i=1}^{n} I(Z_i = z_m)$, where $I(\cdot)$ denotes the indicator function; further, $n_1 + \cdots + n_M = n$. $X_i$ is the other type of $p$-vector covariate with support $\mathcal{R}^p$, and $Y_i$ is the response or dependent variable. The stratified models are characterized based on their dependence on a set of arbitrarily selected categorical features, and linear dependence on the remaining features, which implies that

$$\boldsymbol{Y}_{z_m} = \boldsymbol{X}_{z_m}\boldsymbol{\beta}_{z_m} + \boldsymbol{\epsilon}_{z_m}, m = 1, \ldots, M, \tag{2.1}$$

where $\boldsymbol{Y}_{z_m} = \{Y_i, i = 1, \cdots, n, Z_i = z_m\}$, which induces the notations of $\boldsymbol{X}_{z_m}$ and $\boldsymbol{\epsilon}_{z_m}$ in a similar manner. $\boldsymbol{\beta}_{z_m} = (\beta_{z_m 1}, \cdots, \beta_{z_m p})^\top$ denotes the stratum-specific coefficient vector. The stratified models (2.1) elucidates homogeneity within every stratum and the heterogeneity across them.

Big data in econometrics typically comprise multiple datasets obtained from various sources (Varian, 2014; Einav and Levin, 2014). For instance,

it is gathered from several locations, during different time periods, or under distinct data collection procedures, which correspond to the generation of the stratum set $\mathcal{Z}$. Other classical instances of big data in econometrics include the fact that big financial data are often collected from more than one thousand daily transactions from ten thousands stocks, resulting in the generation of $M$=10000 strata; the real-estate big data consists of ten thousand daily accumulated observations obtained across 10 years from 344 communities, and the number of strata (3440) is calculated using the product of the numbers of communities and years. This example implies the flexible the generation of the stratum set $\mathcal{Z}$ by interacting different values of different categorical variables. Futhermore, the stratum set $\mathcal{Z}$ can be formed based on the values of continuous variable by adopting slicing techniques; by slicing the confidence [10,50] of the stock prices into four partitions [10,20), [20,30), [30,40), [40,50], we can obtain four strata.

The stratified models are more flexible than single or average models in denoting the heterogenous stratum-specified characteristics. However, with increasing number of strata, similarity or generality may exist across them due to their homogeneous characteristics, which implies that the distinct stratified models may belong to one common model. Furthermore, the observations in some strata are extremely rare; even when $n_m = 1$ or when

the stratum-specified number of covariates is considerably larger than that of the observations, i.e., $p \gg n_m$, the strength should be borrowed from the neighborhood models. In reality, we do not know that which strata arise from the same regression model or which strata can be borrowed to improve their own power. Therefore, we assume that the $M$ strata arise from $K$ (i.e., $1 \leq K \leq M$) regression models and introduce another group set $\mathcal{G} = \{\mathcal{G}_1, \cdots, \mathcal{G}_K\}$ as a mutually exclusive partition of $\{z_1, \cdots, z_M\}$, for $z_m \in \mathcal{G}_k$, $\boldsymbol{\beta}_{z_m} = \boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k$ is the common value of $\boldsymbol{\beta}_{z_m}$'s. Further, we propose $K$-regression models using

$$\boldsymbol{Y}_{z_m} = \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_{z_m}, z_m \in \mathcal{G}_k. \tag{2.2}$$

The $K$ stratified regression models, which are simplified as $K$-regression models in (2.2), inherit the advantages of stratified models (2.1); for instance, the $K$-regression models exhibit arbitrary dependence on the stratum categorical variable, $Z$, but simple (typically linear) dependence on other variables $X$, which elucidates that they are essentially non-parametric with respect to the stratified feature, and parametric with a simple form in relation to other variables. Compared with clasical mixture model, $K$ regression model (2.2) inherits the semiparametric superity and more robustness than likelihood. And different from the considerations of Li et al. (2022), $K$ regression charaterize the group-specific heterogenerity by

introducing the latent group pattern parameters $\mathcal{G}_k$'s.

The $K$-regression modeling in (2.2) is considerably flexible to accommodate heterogeneity for various types of multiple strata with econometric analysis including the following: (i) pooled cross-sectional datasets, where strata are collected from $M$ different time periods by observing different subjects during each temporal interval; and the $m$th one contains $n_m$ individuals. $K$-regression modeling is designed to detect $K$ populations across the $M$ strata and specify which strata belong to common groups; (ii) a panel/longitudinal dataset, where $M$ strata are generated from $M$ subjects (such as individuals, firms, countries, or regions) over $n_m$ repeated measurements on the $m$th one. Then, these datasets denote balanced panel/longitudinal data if $n_1 = \cdots = n_m$; further, we design $K$ regression modeling to specify the $K$ subgroup structures and integrate individuals belonging to common subgroup.

*Remark 1.* Another interesting discovery of integrative analysis with respect to balanced panel datasets was obtained using another method in which $M$ strata were formed, because the $M$ strata can be assumed to be generated from $M$ time points (not subjects), where the $m$th stratum covers $n_m$ observations on $n_m$ respective subjects. Subsequently, the indexes $\{z_1, \cdots, z_M\}$ do not exhibit a qualitative nature, whereas they exhibit or-

dered categorical values. Therefore, an integrative procedure searches the structural break jump location (Qian and Su, 2016; Li *et al.*, 2017; Wang *et al.*, 2019). Specifically, $K$ subgroup divisions imply the determination of $K$-1 structural breaks, and the temporal intervals of structural break can be obtained as $\{(\max\{\mathcal{G}_{k-1}\},\ \min\{\mathcal{G}_k\}) : k = 2, \ldots, K\}$, where $\max\{A\}$ $\min\{A\}$ denote the maximum and minimum values in set $A$, respectively.

*Remark 2.* Apart from the fact that strata can be generated across multiple time periods, such as pooled cross-sectional or panel/longitudinal datasets, $M$ strata can also be collected from different geographic locations or using experimental methods. We can also divide $n$ observations into $M$ strata based on some observable categorical variables such as gender, race, and even their pairwise interactions, etc., pertaining to the economical objective of integrative analysis.

### 2.2 Estimator and Computation

$K$-regression models in (2.2) can be used to achieve our major objective of statistical estimation and inference with respect to the $Kp$ coefficient vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \cdots, \boldsymbol{\alpha}_K^\top)^\top$ and group parameter $\mathcal{G}$. Given the value $K$ and tuning parameter $\lambda$, the estimators of group parameter $\mathcal{G}$ and coefficient $\boldsymbol{\alpha}$ in model (2.2) can be defined as the minimizer of the following objective

function:

$$\ell_p(\boldsymbol{\alpha}, \mathcal{G}; K, \lambda) = \frac{1}{2}\sum_{k=1}^{K}\sum_{z_m \in \mathcal{G}_k} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k\|^2 + \frac{n}{K}\sum_{k=1}^{K}\sum_{j=1}^{p} p_\lambda(|\alpha_{kj}|). \quad (2.3)$$

Although only sparsity structure of $\alpha_k$'s are considered, it should be emphasized that each potential $K$ strata could also share information in $\alpha_k$'s, such as group penalty (Yuan and Lin, 2006) can detect the group structures and strata-specific coefficients, fused penalty (Tibshirani, 2005) on the pairwise difference between $\alpha_{kj}$ and $\alpha_{kl}$ can be used to check their order values.

The penalized objective function (2.3) is nonconvex; for the given values of $\boldsymbol{\alpha}$, the $k$th group set can be obtained as

$$\mathcal{G}_k(\boldsymbol{\alpha}) = \Big\{m : \{\text{argmin}_{k\in\{1,\cdots,K\}}\|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k\|^2\} = k\Big\}. \quad (2.4)$$

Further, we perform a plug-in procedure to update the estimator $\widehat{\boldsymbol{\alpha}}$ using the following profiled objective function:

$$\widehat{\boldsymbol{\alpha}} = \text{argmin}_{\boldsymbol{\alpha}\in\mathcal{R}^{Kp}}\Big\{\frac{1}{2}\sum_{k=1}^{K}\sum_{z_m \in \mathcal{G}_k(\boldsymbol{\alpha})} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k\|^2 + \frac{n}{K}\sum_{k=1}^{K}\sum_{j=1}^{p} p_\lambda(|\alpha_{kj}|)\Big\} \quad (2.5)$$

Subsequently, the $\mathcal{G}$ can be eventually estimated as $\widehat{\mathcal{G}} = (\widehat{\mathcal{G}}_1(\widehat{\boldsymbol{\alpha}}), \cdots, \widehat{\mathcal{G}}_K(\widehat{\boldsymbol{\alpha}}))^\top$.

Further, the close connection between (2.3) and the well-known *kmeans* clustering algorithm is explored to obtain a fast and efficient computing procedure; the simple and fast iterative algorithm presented in Algorithm A.1 in the appendix of supplemental materials generates a group estimator

$\widehat{\mathcal{G}}$ and coefficient estimator $\widehat{\boldsymbol{\alpha}}$ in (2.3) based on the optimization in (2.4) and (2.5), respectively; this algorithm is repeated until some convergence criterion is obtained as the input.

The computation of this algorithm under fixed $K$ and tuning parameter $\lambda$ is fast because (i) it quickly alternates between the integrative and updated steps. The "integrative" step minimizes the objective function with respect to the membership assignment given fixed $\alpha_k$'s and determines the integration of $m$ stratum into the $k$th subpopulation with respect to the minimum quadratic loss, resulting in rapid computation during this step. In the "updated" step, we separately update the estimator $\boldsymbol{\alpha}_k^{(s+1)}$ separately for $k = 1, \cdots, K$ by

$$\boldsymbol{\alpha}_k^{(s+1)} = \text{argmin}_{\boldsymbol{\alpha}_k} \left\{ \frac{1}{2} \sum_{z_m \in \mathcal{G}_k^{(s+1)}} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m} \boldsymbol{\alpha}_k\|^2 + \frac{n}{K} \sum_{j=1}^{p} p_\lambda(|\alpha_{kj}|) \right\}, \quad (2.6)$$

and the fast coordinate descent algorithm is utilized for calculating $\boldsymbol{\alpha}_k^{(s+1)}$. (ii) Furthermore, the objective function is observed to become non-increasing during the iterative procedure, causing rapid numerical convergence. However, Algorithm A.1 in the appendix of supplemental materials is sensitive to the starting point $\boldsymbol{\alpha}^{(0)}$. Therefore, we use a computational strategy of *kmeans* and generate several initial values to obtain stable estimators. Subsequently, we consider a more efficient alternative by extending the

variable neighborhood search method as the state-of-the-art heuristic to solve the minimum sum-of-squares partitioning problem Bonhomme and Manresa (2015), allowing for high-dimensional covariates to our proposed $K$-regression modeling. The concrete computing procedure has been presented in Algorithm A.2.

## 3. Theoretical results

### 3.1 Notations

We assume the *prior* information that the true number $K$ was known and characterize the asymptotic properties of the estimators to study the theoretical results of the proposed K regression estimator.

First, we introduce some notations and regularity conditions. Under the sparsity assumption of every subpopulation in high dimensionality, we can obtain $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1}^\top, \boldsymbol{\alpha}_{k2}^\top)^\top$, where $\boldsymbol{\alpha}_{k1} \in \mathcal{R}^{q_k}$ and $\boldsymbol{\alpha}_{k2} \in \mathcal{R}^{p-q_k}$ correspond to the nonzero and zero components of $\boldsymbol{\alpha}_k$, respectively. Under the above notation, $\boldsymbol{\alpha}_{0k}$ can be written as $\boldsymbol{\alpha}_{0k} = (\boldsymbol{\alpha}_{0k1}^\top, \mathbf{0}^\top)^\top$, where $\boldsymbol{\alpha}_{0k1}$ is the true value of $\boldsymbol{\alpha}_{k1}$. Then through ranking the nonzero part of parameters ahead of zeros, $\boldsymbol{\alpha}$ can be rewritten as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\mathcal{K}1}^\top, \boldsymbol{\alpha}_{\mathcal{K}2}^\top)^\top$, where $\boldsymbol{\alpha}_{\mathcal{K}1} = (\boldsymbol{\alpha}_{11}^\top, \cdots, \boldsymbol{\alpha}_{K1}^\top)^\top$ and $\boldsymbol{\alpha}_{\mathcal{K}2} = (\boldsymbol{\alpha}_{12}^\top, \cdots, \boldsymbol{\alpha}_{K2}^\top)^\top$, then the true coefficient vector $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{0\mathcal{K}1}^\top, \mathbf{0}^\top)^\top$, $\text{supp}(\boldsymbol{\alpha}_{0\mathcal{K}1}) = \sum_{k=1}^K q_k = q_{\mathcal{K}}$, and estimator

$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_{\mathcal{K}1}^\top, \widehat{\boldsymbol{\alpha}}_{\mathcal{K}2}^\top)^\top$. $\mathcal{G}_0 = \{\mathcal{G}_{01}, \cdots, \mathcal{G}_{0K}\}$ and $\widehat{\mathcal{G}} = \{\widehat{\mathcal{G}}_1, \cdots, \widehat{\mathcal{G}}_K\}$ denote the true and estimated group parameter values respectively.

Let $\widetilde{\Pi} = \{\pi_{mk}\}$ denote an $M \times K$ matrix with $\pi_{mk} = 1$ for $z_m \in \mathcal{G}_{0k}$ and $\pi_{mk} = 0$ for $m \notin \mathcal{G}_{0k}$. Let $\Pi = \widetilde{\Pi} \otimes I_p$, $\boldsymbol{Y} = \text{diag}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M)$, $\mathbb{Y} = (\boldsymbol{Y}\widetilde{\Pi})^+$, where $A^+$ denotes a vector obtained from row sums of matrix $A$. $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{z_1}^\top, \ldots, \boldsymbol{\epsilon}_{z_M}^\top)^\top = (\epsilon_1, \ldots, \epsilon_n)^\top$, $\boldsymbol{X} = \text{diag}(\boldsymbol{X}_{z_1}, \ldots, \boldsymbol{X}_{z_M})$, $\mathbb{X} = (\boldsymbol{X}\Pi)_{n\times(Kp)}$, $\mathbb{X}_1$ and $\mathbb{X}_2$ are $n \times q_{\mathcal{K}}$ and $n \times (Kp - q_{\mathcal{K}})$ submatrixes of $\mathbb{X}$ corresponding to the decomposition of $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\mathcal{K}1}^\top, \boldsymbol{\alpha}_{\mathcal{K}2}^\top)^\top$. Note that $\Pi^\top\Pi = \text{diag}(|\mathcal{G}_{01}|, \ldots, |\mathcal{G}_{0K}|) \otimes I_p$. For a given vector $b = (b_1, \ldots, b_t) \in \mathcal{R}^t$ and a symmetric matrix $A_{t\times t}$, define $\|b\|_\infty = \max_{1\le s\le t} |b_s|$, $\|A\|_\infty = \max_{1\le i\le t} \sum_{j=1}^t |A_{ij}|$, $\|A\| = \|A\|_2 = \max_{b\in\mathcal{R}^t, \|b\|=1} \|Ab\|$ and $\|A\|_{2,\infty} = \max_{1\le i\le t} \|A_{i,}\|$, where $A_{i,}$ denotes vector of $i$th row of $A$. Let $\gamma_{\min}(A)$ and $\gamma_{\max}(A)$ be the smallest and largest eigenvalues of $A$ respectively, and further Let

$$b_n = \min_{k\neq k'} \|\boldsymbol{\alpha}_{0k} - \boldsymbol{\alpha}_{0k'}\|$$

be the minimum difference of the coefficient vector between every two populations.

Denote $d_n = \frac{1}{2}\min_{1\le j\le q_{\mathcal{K}}} |\alpha_{0\mathcal{K}1j}|$ as be half of the minimum signal. Let $N_k = \sum_{z_m\in\mathcal{G}_{0k}} n_m$, $N_{\min} = \min_{1\le k\le K} N_k$ and $N_{\max} = \max_{1\le k\le K} N_k$, which represent the true minimum and maximum sample sizes among all popula-

tions. Meanwhile, $n_{\min} = \min_{1\le m\le M} n_m$ and $n_{\max} = \max_{1\le m\le M} n_m$ denote the minimum and maximum sample sizes among the whole strata, respectively. $p'_\lambda(a)$ and $p''_\lambda(a)$ denote the first and second derivations of penalty $p_\lambda(a)$ about $a$. Let $c$ and $c'_j s$ denote some positive constants.

### 3.2 Oracle property with group structure known

If the underlying group parameter $\mathcal{G}_0$, i.e., matrix $\Pi$, is known, we can define an estimator as

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}} &= \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathcal{R}^{Kp}}\Big\{\frac{1}{2}\sum_{k=1}^{K}\sum_{z_m\in\mathcal{G}_{0k}}\|\boldsymbol{Y}_{z_m}-\boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k\|^2+\frac{n}{K}\sum_{k=1}^{K}\sum_{j=1}^{p}p_\lambda(|\alpha_{kj}|)\Big\},\\ &= \operatorname{argmin}_{\boldsymbol{\alpha}\in\mathcal{R}^{Kp}}\Big\{\frac{1}{2}\|\mathbb{Y}-\mathbb{X}\boldsymbol{\alpha}\|^2+\frac{n}{K}\sum_{k=1}^{K}\sum_{j=1}^{p}p_\lambda(|\alpha_{kj}|)\Big\}. \qquad (3.1)\end{aligned}$$

And $\widetilde{\boldsymbol{\alpha}}$ can be rewritten as $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1}^\top, \widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2}^\top)^\top$. Since the group relationship of the strata, i.e., $\mathcal{G}_0$, is typically unknown in advance, and the oracle estimators are infeasible in practice. However, this can shed light on the theoretical properties of the proposed estimators. Hereafter, we consider the following conditions:

(C1) $p_\lambda(t)$ is symmetric, increasing and concave in $t\in[0,+\infty)$, and $p_\lambda(0) = 0$. The derivative $p'_\lambda(t)$ is continuous and non-increasing in $t\in(0,+\infty)$, In addition, $p'_\lambda(t)$ is increasing in $\lambda$ and $\lambda^{-1}p'_\lambda(0+)\equiv\lambda^{-1}p'(0+) = c > 0$.

(C2) The noise vector $\boldsymbol{\epsilon}$ has sub-Gaussian tails such that $P(|\boldsymbol{a}^\top\boldsymbol{\epsilon}| < \|\boldsymbol{a}\|x) \geq 1 - 2\exp(-c_1x^2)$ for any vector $\boldsymbol{a} \in \mathcal{R}^n$ and $x > 0$, and $E(\epsilon_i^4) < \infty$ for $i = 1, \cdots, n$.

(C3) (i) $p'_\lambda(d_n) = O(\frac{K\sqrt{N_{\min}}}{n})$ and $d_n \gg \sqrt{q_{\mathcal{K}}/N_{\min}}$; (ii) $p'_\lambda(0_+) \gg Kq_{\mathcal{K}}\sqrt{q_{\mathcal{K}}/N_{\min}}$; (iii) For $\boldsymbol{b} \in \mathcal{N}_0$, where $\mathcal{N}_0 = \{\boldsymbol{b} \in \mathcal{R}^{q_{\mathcal{K}}} : \|\boldsymbol{b} - \boldsymbol{\alpha}_{0\mathcal{K}1}\| \leq d_n\}$, $\max_j p''_\lambda(|b_j|) = o(\frac{KN_{\min}}{n})$.

(C4) (i) $\gamma_{\min}(\mathbb{X}_1^\top\mathbb{X}_1) \geq c_2N_{\min}$, $\gamma_{\max}(\mathbb{X}_1^\top\mathbb{X}_1) \leq c_3n$. (ii) $\sum_{z_m\in\mathcal{G}_{0k}} \|\boldsymbol{X}_{mj}\|^2 = N_k$; (iii)$\|\mathbb{X}_2^\top\mathbb{X}_1\|_{2,\infty} = O(q_{\mathcal{K}}n)$; (iv) $\sup_i \|\mathbb{X}_{1i}\| \leq c_4\sqrt{q_{\mathcal{K}}}$; (v) $N_{\min} = O(N_{\max})$.

Lv and Fan (2009) considered the family of concave penalty functions in Condition (C1), which contained several examples, including the concave penalties SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The requirement with respect to the sub-Gaussian tails of the noise vector in Condition (C2) is basic in case of the high-dimensional regression scenario and invariant across multiple strata. The penalty assumption in Conditions (C3) (i) and (ii) implies a penalized level with respect to non-zero and zero components, respectively. With respect to LASSO penalty, Conditions (C3) (i) and (ii) cannot be simultaneously satisfied to ensure that $\lambda = p'_\lambda(d_n) = O(\frac{K\sqrt{N_{\min}}}{n})$ is incompatible with $\lambda = p'_\lambda(0_+) \gg Kq_{\mathcal{K}}\sqrt{q_{\mathcal{K}}/N_{\min}}$.

This contradiction implies that the LASSO-based $K$-regression estimator cannot generally attain the consistency rate obtained from Theorem 1 and the oracle property achieved in Theorem 2. This issue has be also observed by Fan and Lv (2011) under homogeneous high-dimensional data with $K = 1$, $N_{\min} = n$, and the corresponding penalty conditions are $p'_\lambda(d_n) = O(\frac{1}{\sqrt{n}})$, $d_n \gg \sqrt{s/n}$, and $p'_\lambda(0_+) \gg \sqrt{s/n}$, where $s$ denotes the true number of nonzero coefficients in a homogeneous setting. Bounding the smallest eigenvalue of the transposition of the active covariate matrix multiplied by the active covariate matrix under a heterogeneous structure is unavailable by $cn$, because

$$\mathbb{X}_1^\top \mathbb{X}_1 = \text{diag}\Big( \sum_{z_m \in \mathcal{G}_{0k}} \boldsymbol{X}_{z_m 1}^\top \boldsymbol{X}_{z_m 1}, k = 1, \cdots, K \Big),$$

and $\gamma_{\min}(\mathbb{X}_1^\top \mathbb{X}_1) \geq \gamma_{\min}\{\sum_{z_m \in \mathcal{G}_{0k}} \boldsymbol{X}_{z_m 1}^\top \boldsymbol{X}_{z_m 1}\} \geq c_2 N_{\min}$ for some constant $c_2$, where $\boldsymbol{X}_{z_m 1}$ denote the submatrices of $\boldsymbol{X}_{z_m}$ and formed by columns in $\text{supp}(\boldsymbol{\alpha}_{0k})$. Without loss of generality, the covariates have be scaled in every subpopulation, assumed in Condition (C4) (ii) and then $\text{tr}(\mathbb{X}_1^\top \mathbb{X}_1) = \sum_{k=1}^K q_k N_k$.

**Theorem 1.** *(Consistency for estimator $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}}$ with group pattern known)*

*Under Conditions (C1)-(C4) and additional condition* $\log(Kp) = o(\frac{n^2 g_{\mathcal{K}}}{N_{\min}^2})$, *there is a local minimizer* $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}} = (\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1}^\top, \widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2}^\top)^\top$ *of objective function (3.1) such*

*that* $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2} = 0$ *with probability tending to one as* $N_{\min} \to \infty$ *and*

$$\|\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1} - \boldsymbol{\alpha}_{0\mathcal{K}1}\| = O_p(\sqrt{q_{\mathcal{K}}/N_{\min}}).$$

Theorem 1 establishes the consistency of the proposed penalized $K$-regression estimator $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1}$, i.e., there is a root-$(\frac{N_{\min}}{q_{\mathcal{K}}})$-consistent $K$-regression estimator of $\boldsymbol{\alpha}_{0\mathcal{K}1}$ under dimensionality $p$ and population number $K$ that satisfies $Kp = o\{\exp(\frac{n^2 q_{\mathcal{K}}}{N_{\min}^2})\}$ at an exponential rate. The sparsity property of the proposed $K$-regression estimator $\widetilde{\boldsymbol{\alpha}}_{0\mathcal{K}2}$ is still valid, that is, zero components in $\boldsymbol{\alpha}_{0\mathcal{K}}$ are estimated as zero with a probability tending to one. Theorem 1 also answers three issues with respect to the the strength of the the minimum signal, its dimensionality, and the minimum sample size of the population that can be handled by the $K$-regression methods.

**Theorem 2.** *(Oracle property of estimators with group pattern known)*

*(i) (Sparsity) Under Conditions of Theorem 1, with probability tending to one as* $N_{\min} \to \infty$,

$$\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2} = 0.$$

*(ii) (Asymptotic normality) Under Conditions of Theorem 1 with Condition (C3) (i) replaced by* $p'_\lambda(d_n) = O(\frac{K}{n})$ *and attaching additional conditions* $q_{\mathcal{K}} = o(N_{\min})$,

$$N_{\min} \gg n^{5/6} q_{\mathcal{K}}^{1/2},$$

*then we conclude*

$$s_n(a_n)^{-1}a_n(\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1} - \boldsymbol{\alpha}_{0\mathcal{K}1}) \xrightarrow{D} \mathcal{N}(0,1),$$

*where*

$$s_n(a_n) = \sigma\{a_n(\mathbb{X}_1^\top \mathbb{X}_1)^{-1}a_n^\top\}^{1/2}$$

*and $a_n$ is a $1 \times q_{\mathcal{K}}$ row vector such that $\|a_n\| = 1$, $\xrightarrow{D}$ denotes convergence in distribution.*

Theorem 2 shows that the sparsity and asymptotic normality of the proposed $K$-regression estimator still hold when the nonsparsity size $q_{\mathcal{K}}$ diverges slower than $N_{\min}$, Combined with conditions $N_{\min} \gg n^{5/6}q_{\mathcal{K}}^{1/2}$ and $N_{\min} \leq n/K$ we conclude $K = o(n^{1/6})$, and thus Theorem 2 indicates that the number of subpopulation $K$ is assumed to grow slower than $n^{1/6}$.

### 3.3 Theoretical property with group structure unknown

Practically, the group structure is unknown. In this section, we provide sufficient conditions under which the induced local minimizer of the objective function (2.3) equal to the oracle least squares estimator $\widetilde{\boldsymbol{\alpha}}$ under a priori knowledge of group structure with high probability. We also derive the lower bound of the minimum difference of coefficients between subpopulation in order to be able to estimate the $K$ effects. Then we impose the following additional conditions.

(C5) (i) $b_n \gg \sqrt{\frac{q_{\mathcal{K}} \log(n)}{n_{\min}}}$; (ii) $\gamma_{\min}(\boldsymbol{X}_{z_m}\boldsymbol{X}_{z_m}) \geq c_5 n_m$, $\gamma_{\max}(\boldsymbol{X}_{z_m}\boldsymbol{X}_{z_m}) \leq c_6 n_m$ for some constants $c_5$, $c_6$.

**Theorem 3.** *If Conditions of Theorem 2 and (C5) hold, any local minimizer of the objective function can achieve the oracle estimator* $\widetilde{\boldsymbol{\alpha}}$ *with probability tending to one when* $n_{\min} \to \infty$.

The result of Theorem 3 implies that if the minimal difference of the common effects between any two subpopulations satisfies $b_n \gg \sqrt{\frac{q_{\mathcal{K}} \log(n)}{n_{\min}}}$, our method can actually recover the true group structure, which means any local solution produced by the proposed $K$-regression algorithm can achieve the oracle performance. Next, we conclude the following corollary.

**Corollary 1.** *Under Conditions of Theorem 2 and (C5), we have*

$$s_n(a_n)^{-1} a_n(\widehat{\boldsymbol{\alpha}}_{\mathcal{K}1} - \boldsymbol{\alpha}_{0\mathcal{K}1}) \xrightarrow{D} \mathcal{N}(0,1),$$

*where*

$$s_n(a_n) = \sigma\{a_n(\mathbb{X}_1^\top \mathbb{X}_1)^{-1} a_n^\top\}^{1/2}$$

*and* $a_n$ *is a* $1 \times q_{\mathcal{K}}$ *row vector such that* $\|a_n\| = 1$, $\xrightarrow{D}$ *denotes convergence in distribution.*

The asymptotic distribution of $K$-regression estimators provides a theoretical justification for further statistical inference, such as testing for het-

erogeneity. Based on the results in Corollary 1, a unified framework is presented for conducting hypothesis tests and constructing confidence regions for $\boldsymbol{\alpha}$. Specifically, we consider $H_0 : \mathcal{B}\boldsymbol{\alpha} = 0$ versus $H_1 : \mathcal{B}\boldsymbol{\alpha} \neq 0$, where $\mathcal{B}$ is a $d \times Kp$ matrix and $d = \text{rank}(\mathcal{B})$. Such hypothesis includes many special cases; for example, $H_{0k}$: $\alpha_k = 0$, $k \in \{1, \ldots, K\}$, which can be used to construct a confidence region for $\alpha_k$; and $H_0 : \alpha_j - \alpha_k = 0$, $j, k \in \{1, \ldots, K\}$, which can be used to test the existence of effect heterogeneity among strata. We develop a $\chi^2$-test statistic for testing $H_0$: $\mathcal{B}\boldsymbol{\alpha} = 0$,

$$\mathcal{T}_n(\mathcal{B}) = (\mathcal{B}\widehat{\boldsymbol{\alpha}})^\top (\mathcal{B}\widehat{\mathcal{V}}_n\mathcal{B}^\top)^{-1}(\mathcal{B}\widehat{\boldsymbol{\alpha}}), \tag{3.2}$$

where $\widehat{\mathcal{V}}_n = \widehat{\sigma}^2(\mathbb{X}_1^\top\mathbb{X}_1)^{-1}$ and $\widehat{\sigma}^2 = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{\sum_{z_m \in \widehat{\mathcal{G}}_k} n_m} \sum_{z_m \in \widehat{\mathcal{G}}_k} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\widehat{\boldsymbol{\alpha}}_k\|^2$.

**Theorem 4.** *Under the null hypothesis and conditions in Theorem 3, we have* $\mathcal{T}_n(\mathcal{B}) \xrightarrow{D} \chi^2_d$ *as* $n \to \infty$, *where* $\chi^2_d$ *denotes the chi-squared distribution with* $d$ *degrees of freedom.*

Theorem 4 provides the asymptotic distribution of the test statistic $\mathcal{T}_n(\mathcal{B})$ under the null hypothesis $H_0$: $\mathcal{B}\boldsymbol{\alpha} = 0$, indicating the validity of the renowned Wilk's phenomenon. The $100(1-\tau)\%$ approximated confidence region for $\mathcal{B}\boldsymbol{\alpha}$ is given by

$$R_\tau = \left\{ \iota : (\mathcal{B}\widehat{\boldsymbol{\alpha}} - \iota)^\top (\mathcal{B}\widehat{\mathcal{V}}_n\mathcal{B}^\top)^{-1}(\mathcal{B}\widehat{\boldsymbol{\alpha}} - \iota) \leq \chi^2_d(1-\tau) \right\},$$

where $\chi^2_d(1-\tau)$ is the $(1-\tau)$-quantile of the $\chi^2$ distribution with $d$ degrees of freedom.

## 4. Simulation studies

Next, we consider one example for checking the performance of our method and the preliminaries we adopt to measure the simulated results and additional examples can been seen in supplemental materials.

**Example 1.** In this example, we generated data from 2-regression models,

$$\boldsymbol{Y}_{z_m} = \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_{z_m}, z_m \in \mathcal{G}_k, k = 1, 2,$$

where $\boldsymbol{X}_{z_m}$ were assumed to be generated from a multivariate normal distribution with zero mean and covariance matrix $\Phi = (d_{jl})$ with $d_{jl} = 0.7^{|j-l|}$, $\boldsymbol{\epsilon}_{z_m}$ was assumed to follow the normal distribution $\mathcal{N}(0, 0.7^2)$, We randomly assigned the strata to two subpopulation with equal probabilities, i.e., $K = 2$ and $P(z_m \in \mathcal{G}_1) = P(z_m \in \mathcal{G}_2) = 1/2$, so that the coefficients equal to $\boldsymbol{\alpha}_1$ for $z_m \in \mathcal{G}_1$, and are $\boldsymbol{\alpha}_2$ for $z_m \in \mathcal{G}_2$, where $\boldsymbol{\alpha}_1 = (1, 0.8, \mathbf{0}^\top_{p-2})^\top$, $\boldsymbol{\alpha}_2 = (-1, -0.8, \mathbf{0}^\top_{p-2})^\top$. We choose $n = 600$, $p = 500$ or $1000$ with two numbers of strata, i.e., $M = 100, 200$, and exam performance of our proposed method under three penalized methods including SCAD, MCP and LASSO.

Tables 1 and Figures 1 present the estimated result of Examples 1. The results in parentheses denote the oracle estimates with known $\mathcal{G}_0$. We note that (i) the simulated results in Tables 1 in the considered measurements using the SCAD, MCP, and LASSO penalties are similar, the estimates obtained using the three methods are close to their corresponding Oracle ones; (ii) the $K$-regression method can accurately integrate the strata with a common population for the estimated RI values that are approximately one; (iii) based on sparsity-induced penalties, the proposed method behaves satisfactorily because the corresponding average numbers of accurately estimated zero components are considerably similar to the true number $p-2$ of zero components in each population, whereas their corresponding average numbers of the inaccurately estimated zero coefficients approach 0 with respect to the PIZ and PCZ indexes, where PCZ denotes percentage of correct zeros with PCZ $(\%)=\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\big\{\frac{100\%}{p-|\mathcal{D}_{z_m}|}[\sum_{j=1}^{p}I(\widehat{\beta}_{mj(t)}=0)I(\beta_{0mj}=0)]\big\}$ PIZ is the percentage of incorrect zeros with PIZ $(\%)=\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}[\frac{100\%}{|\mathcal{D}_{z_m}|}$ $\{\sum_{j=1}^{p}I(\widehat{\beta}_{mj(t)}=0)I(\beta_{0mj}\neq 0)\}]$, where $\mathcal{D}_{z_m}=\{j:\beta_{0mj}\neq 0\}$ as the index of true model in the $m$th stratum, note that larger PCZ and smaller PIZ imply good model-fitting procedure; (iv) $K$-regression can recover the subpopulation and sparsity structure in just a few seconds even though the sample size $n$ was large and the dimension $p$ was high; (v) in Figures 1,

the RMS values of the SCAD/MCP-based $K$-regression method are smaller than those of the LASSO estimators and attain Oracle estimates with the known group structure $\mathcal{G}_0$, which verifies that the concave penalty-based $K$-regression method achieves the Oracle property and $\sqrt{N_{\min}/q_{\mathcal{K}}}$ consistency, whereas the LASSO penalty is unavailable to attain such a result; (vi) the percentage of the specified numbers of subpopulations $\widehat{K}$ equal to the true number is close to one using the $K$-regression method through the considered BIC; (vii) increasing the number of strata (i.e., $M$) or dimensionality (i.e., $p$) decreases the performance of the proposed method in relation to all the considered indexes; (viii) the levelplots in Figure 2 adopts colors to denote the component value of coefficient matrix and are generated based on the averaging estimates after 200 replicates, i.e., the estimated $M \times p$ coefficient matrix $\widehat{\boldsymbol{\beta}} = \frac{1}{200}\sum_{t=1}^{200}\widehat{\boldsymbol{\beta}}_{(t)}$, under a fixed partition; the results imply that separate statistical modeling in each stratum (i.e., M-penalty) dramatically destroys the estimates, whereas our proposed integrative analysis (i.e., K-penalty) with $K$-regression methods can ensure the efficient estimation of the coefficient matrix by accurately recovering each subpopulation and its corresponding sparsity structure. The simulated results elucidate that the proposed $K$-regression method exhibits a desirable behavior in terms of the integration, variable selection, parameter estima-

tion, and computational speed, which implies that the empirical results are consistent with those presented in Theorems 1.

To verify the existence of treatment heterogeneity in Examples 1 for a case in which $M = 200$ and $p = 1000$, we apply the test statistic

$$\mathcal{T}_n(\widehat{\mathcal{B}}) = (\widehat{\mathcal{B}}\widehat{\boldsymbol{\alpha}})^\top(\widehat{\mathcal{B}}\widehat{\mathcal{V}}_n\widehat{\mathcal{B}}^\top)^{-1}(\widehat{\mathcal{B}}\widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\mathcal{B}} = \widehat{\mathcal{D}} \otimes I_{q_\mathcal{K}}$, $\widehat{\mathcal{D}} = \{(e_i - e_j), i < j\}^\top_{\frac{K(K-1)}{2} \times K}$ with $e_i$ being the $i$th $K \times 1$ stratum vector whose $i$th element is 1 and the remaining 0's, $I_{q_\mathcal{K}}$ is a $q_\mathcal{K} \times q_\mathcal{K}$ identity matrix and $\otimes$ is the Kronecker product. Theorem 4 indicates that $\text{rank}(\widehat{\mathcal{B}}) = q_\mathcal{K}(K-1)$, and the mean of $p$-values based on 200 replicates is given by $\frac{1}{200}\sum_{j=1}^{200} \chi^2_{q_\mathcal{K}(K-1)}(\mathcal{T}_n^{(j)}(\widehat{\mathcal{B}}))$, where $\mathcal{T}_n^{(j)}(\widehat{\mathcal{B}})$ is the value of $\mathcal{T}_n(\widehat{\mathcal{B}})$ from the $j$th replicate, $\chi^2_{q_\mathcal{K}(K-1)}(t) = P(\mathcal{Z}_{q_\mathcal{K}(K-1)} > t)$, and $\mathcal{Z}_{q_\mathcal{K}(K-1)}$ follows a $\chi^2$ distribution with $q_\mathcal{K}(K-1)$ degrees of freedom. The mean p-values are all less than 0.001 in Examples 1, which strongly supports the existence of effect heterogeneity this example. The simulated results in Examples 1 also suggests that the consistency of the estimation of group-specific coefficient under the $K$-regression method is completely dependent on the integration accuracy, i.e., $\widehat{\mathcal{G}}$.

## 5. Empirical study

In this section, we apply our proposed $K$-regression method to the "communities and crime (CAC) data" obtained from the UCI Machine Learning Repository. It comprises information from different communities of the U.S., socio-economic data from the 1990 U.S. Census and the 1990 U.S. Law Enforcement Management and Administrative Statistics Survey, and crime data from the 1995 U.S. FBI Uniform Crime Report. Apart from specific information to identify the community or state, explained by its corresponding abbreviated name, the datasets includes 125 variables and 18 crime indices. We selected the number of murders per 100K population in 1995 as a response of interest. After eliminating the covariates suffering from missingness, we obtained a dataset containing $M = 48$ states, $n = 2215$ communities, and $p = 102$ covariates. Assuming that the samples of 48 states originate from the $K$ populations, the $K$-regression models can be defined as

$$\boldsymbol{Y}_{z_m} = \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_{z_m}, k = 1, \cdots, K, z_m \in \mathcal{G}_k, \tag{6.1}$$

with $\sum_{k=1}^{K} |\mathcal{G}_k| = 48$. Further, we focus on estimating the number $K$ of regression models, group parameters $\mathcal{G}_k$'s and the coefficients $\boldsymbol{\alpha}_k$'s.

Enlightened by the superior performance of concave penalties in es-

timation and variable selection, we applied SCAD- and MCP-based $K$-regression methods to recover the subpopulation and sparsity structure in the assumed model (6.1) on the CAC dataset and specify the optimal $K$ through introduced the BIC criterion in the supplementary materials. The eventual integration results are presented in Figures 3 and Table 2, which visually elucidate that (i) the SCAD and MCP methods estimate common population number $K = 2$ and similar group structures $\widehat{\mathcal{G}}_1$ and $\widehat{\mathcal{G}}_2$, where the state ND is integrated into $\widehat{\mathcal{G}}_1$ by SCAD, whilst belonging to $\widehat{\mathcal{G}}_2$ by MCP, and the states WY and NH are partitioned into $\widehat{\mathcal{G}}_2$ by SCAD, regardless of being merged into $\widehat{\mathcal{G}}_1$ by MCP; (ii) the result of coefficient estimation (Est.) and corresponding P-values in Population 1 by the two concave penalties commonly and significantly specifies the positive effects of RPB and NIST on the murder ratio, whereas the NIST feature imposes zero effect on the response of interest in Population 2; (iii) although the covariates HV, MPD, and PVB do not impact on murder ratio in Population 1, they exhibit considerable influence in Population 2; and (iv) Furthermore, the existence of heterogeneity is verified through $\mathcal{T}_n(\mathcal{B})$ in (3.2), where $\mathcal{B} = \mathcal{D} \otimes I_{q_\mathcal{K}}$, $\mathcal{D} = \{(e_i - e_j), i < j\}^\top_{\frac{K(K-1)}{2} \times K}$, and $K = 2$, $q_\mathcal{K} = 10$ and $q_\mathcal{K} = 9$ for SCAD and MCP penalties, respectively. Theorem 4 indicates that $\text{rank}(\mathcal{B}) = q_\mathcal{K}(K-1)$ and then the calculated p-value is

$\chi^2_{\text{rank}(\mathcal{B})}(\mathcal{T}_n(\mathcal{B})) = 0.008$ and 0.010 by SCAD and MCP penalties, respectively, which confirms the existence of heterogeneity. Another interesting phenomenon is the tight connection of the model populations and the population density, the bigger population density corresponds to the model population 2. And the number of significant factors influencing the number of murders in population 2 apparently is much larger, which may be attributed to more complex environment along with high population density.

## 6. Concluding remarks

In this study, we develop the $K$-regression modeling to simultaneously integrate the strata with common regression structure and estimate stratum-specific fixed effects to accommodate the unobserved heterogeneity among multiple strata. The application of the $K$-regression method in simulations and real data examples exhibits superior performance with respect to fast integration and accurate variable selection. This is because massive data often comprises multiple high-dimensional strata derived from a growing number of heterogeneous subpopulations with an unknown common structure and sparsity information, $K$-regression modeling is naturally scalable and applicable to deal with heterogeneous issues for massive dataset. We also have learned that the statistical inference in integrative analysis is

dependent on the results of subpopulation and sparsity recovery, resulting in inference uncertainty. Thus, the "post-integration and selection" issue arises ; hence, further studies are required.

## Acknowledgments

The authors are grateful to the Editor, an Associate Editor, and two referees for their valuable suggestions and comments, which greatly improved the manuscript. Xiaodong Yan was supported by the National Natural Science Foundation of China (grant number 11901352), the Natural Science Foundation of Shandong Province (grant number ZR2019BA017), the Social Science Foundation of Shandong Province (grant number 19DTJJ03), and the Young Scholars Program of Shandong University (YSPSDU: 11020088964008); Jian Huang was supported in part by the U.S. National Science Foundation grant DMS-1916199.

## References

Bai, J. (2010). Common breaks in means and variances for panel data. *Journal of Econometrics* **157**, 78–92.

Breiman, L. (1996). Bagging predictors. *Machine learning* **24**, 123–140.

Browning, M. and J. M. Carro (2007). Heterogeneity and Microeconometrics Modelling. in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. 3, ed. by R. Blundell, W. K. Newey, and T. Persson. New York: Cambridge University Press, 45–74.

Bester, C. A. and Hansen, C. B. (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics* **190**, 197-208.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* **83**, 1147–1184.

Chudik, A. and Pesaran M. H. (2015). Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors. *Journal of Econometrics* **188**, 393-420.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, **346**, 1243089.

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

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-

dimensionality. *IEEE Transactions on Information Theory* **57**, 5467-5484.

Hsiao, C., Pesaran, M. H., 2008. Random coefficient panel data models. In L. Matyas and P. Sevestre (Eds.), The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice, pp. 187-216, 3rd Edition. Springer-Verlag, Berlin.

Kim, D., (2011). Estimating a common deterministic time trend break in large panels with cross sectional dependence. *Journal of Econometrics* **164**, 310–330.

Li, D., Qian, J. and Su, L. (2017). Panel data models with interactive fixed effects and multiple structural breaks. *Journal of the American Statistical Association* **111**, 1804–1819.

Li, K. and Lu, L. (2014). Efficient estimation of heterogeneous coefficients in panel data models with common shock. Working paper, Capital University of Economics and Business.

Li, Y., Yu, C., Zhao, Y., Yao, W., Aseltine, R.H. and Chen, K. (2022). Pursuing sources of heterogeneity in modeling clustered population. *Biometrics*.

Lin, C-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods* **1**, 42-55.

Lu, J., Cheng, G. and Liu, H. (2016). Nonparametric Heterogeneity Testing For Massive Data. arXiv preprint arXiv:1601.06212.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37**, 3498–3528.

Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. *Journal of American Statistical Association* **112**, 410–423.

Ma, S. and Huang, J. (2016). Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*.

Pesaran, M. H. and Tosetti, E. (2011). Large panels with common factors and spatial correlation. *Journal of Econometrics* **161**, 182-202.

Qian, J. and Su, L. (2016). Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso. *Journal of Econometrics* **191**, 86-109.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

Sarafidis, V. and N. Weber (2015). A Partially heterogenous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics* **77**, 274–296.

Song, M. (2013). Asymptotic theory for dynamic heterogeneous panels with cross-sectional dependence and its applications. Working paper, Columbia University.

Su, L. and Q. Chen, (2013). Testing Homogeneity in Panel Data Models With Interactive Fixed Effects. *Econometric Theory* **29**, 1079–1135.

Su, L. and Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. *Journal of Econometrics* **169**, 34–47.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica* **84**, 2215–2264.

Su, L. and Ju, G. (2018). Identifying Latent Grouped Patterns in Panel Data Models with Interactive Fixed Effects. *Journal of Econometrics*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(1), 91–108.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, **28**, 3–28.

Wang, W. (2019). Heterogeneous structural breaks in panel data models. *Journal of Econometrics*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), **68**(1), 49–67.

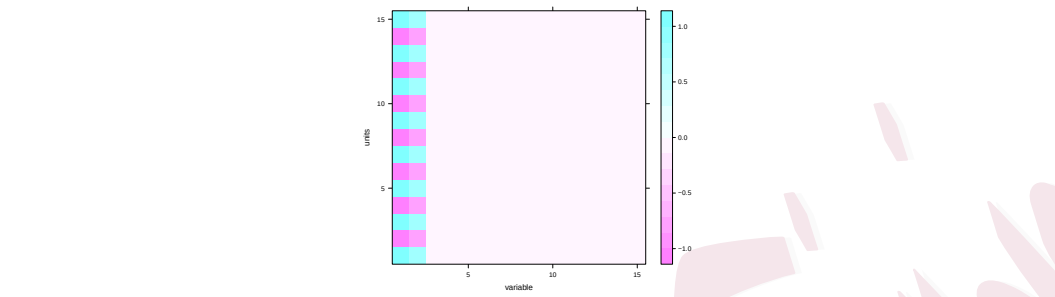
Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.


Figure 1: Boxplots of RMS under the three penalized methods after 100 replicates with $M$ =100, $p$ =500 (left) and $p$ =1000(right) in Example 1.

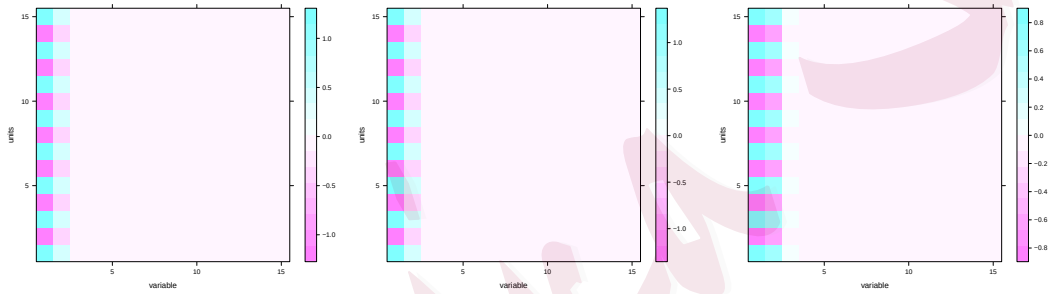
Table 2. Performance of estimates of CAC data through SCAD-based and MCP-based $K$-regression methods.

| | SCAD | | | | | | MCP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Population 1 | | | Population 2 | | | Population 1 | | | Population 2 | | |
| | AK | AZ | DC | AL | AR | CA | AK | AZ | DC | AL | AR | CA |
| | IA | IN | KS | CO | CT | DE | IA | IN | KS | CO | CT | DE |
| | MN | ND | SD | FL | GA | ID | MN | NH | SD | FL | GA | ID |
| | UT | WV | ME | IL | KY | LA | UT | WV | ME | IL | KY | LA |
| | | | | MA | MD | MI | WY | | | MA | MD | MI |
| | | | | MO | MS | NC | | | | MO | MS | NC |
| | | | | NH | NJ | NM | | | | NH | NJ | NM |
| | | | | NV | NY | NH | | | | NV | NY | ND |
| | | | | OK | OR | PA | | | | OK | OR | PA |
| | | | | RI | SC | TN | | | | RI | SC | TN |
| | | | | TX | VA | VT | | | | TX | VA | VT |
| | | | | WA | WI | WY | | | | WA | WI | |
| Variable | Est. | | P-value | Est. | | P-value | Est. | | P-value | Est. | | P-value |
| RPB | 13.17 | | 0.00 | 3.99 | | 0.00 | 12.94 | | 0.00 | 1.62 | | 0.01 |
| RPW | 0.00 | | – | 0.00 | | – | 0.00 | | – | -3.04 | | 0.00 |
| MPD | 0.00 | | – | 0.92 | | 0.05 | 0.00 | | – | 1.45 | | 0.03 |
| PWMYK | 0.00 | | – | -0.03 | | 0.45 | 0.00 | | – | -0.39 | | 0.33 |
| PWM | 0.00 | | – | -0.12 | | 0.39 | 0.00 | | – | 0.00 | | – |
| PPDH | 0.00 | | – | 1.93 | | 0.02 | 0.00 | | – | 0.00 | | – |
| HV | 0.00 | | – | 0.90 | | 0.04 | 0.00 | | – | 0.99 | | 0.04 |
| PVB | 0.00 | | – | 1.83 | | 0.01 | 0.00 | | – | 1.78 | | 0.01 |
| MRPHI | 0.00 | | – | 0.00 | | – | 0.00 | | – | 0.02 | | 0.79 |
| NIST | 2.47 | | 0.00 | 0.00 | | – | 2.72 | | 0.00 | 0.00 | | – |
| LPODU | 0.00 | | – | 0.03 | | 0.82 | 0.00 | | – | 0.00 | | – |

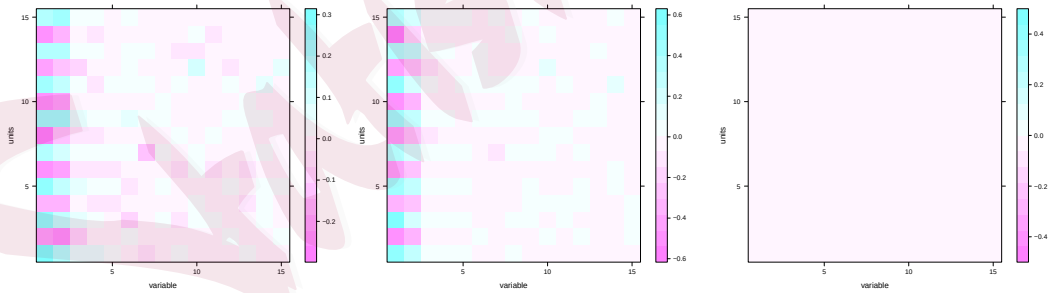
*Note*: RPB: racepctblack; RPW: racepctwhite; MPD: MalePctDivorce; PWMYK: PctWorkMomYoungKids; PWM: PctWorkMom; PPDH:PctPersDenseHous;HV: HousVacant;PVB:PctVacantBoarded; MRPHI:MedRentPctHousInc; NIST: NumInShelters; LPODU: LemasPctOfficDrugUn.




(a) TRUE




(b) K-SCAD (c) K-MCP (d) K-LASSO




(e) M-SCAD (f) M-MCP (g) M-LASSO

Figure 2: The levelplots of coefficient matrix estimation with $n = 60$, $M = 15$, $p = 15$ in Example 1. TRUE denotes the levelplot of true coefficient matrix; M-penalty represents the methods conducting statistical modeling based on each stratum separately; K-penalty denotes using our penalty-based $K$-regression.

(a) K-SCAD (b) K-MCP

Figure 3: The group picture of USA about crimes proportion and dark and light blues denote Populations 1 and 2, respectively.

Table 1: Simulation results by different variable selection methods in Example 1 with $n$=600 and the results in parenthesis denote the oracle estimates with known $\mathcal{G}_0$.

| | | p=500 | | p=1000 | |
|---|---|---|---|---|---|
| Selection method | Criterion | M=100 | M=200 | M=100 | M=200 |
| SCAD | CP(%) | 100.00(100.00) | 95.00(100.00) | 99.00(100.00) | 93.00(100.00) |
| | PCZ(%) | 100.00(100.00) | 100.00(100.00) | 100.00(100.00) | 100.00(100.00) |
| | PIZ(%) | 0.00(0.00) | 5.00(0.00) | 1.00(0.00) | 7.00(0.00) |
| | PER(%) | 100.00(100.00) | 97.00(100.00) | 99.00(100.00) | 93.00(100.00) |
| | RI(%) | 100.00(100.00) | 96.04(100.00) | 99.48(100.00) | 95.05(100.00) |
| | AMS(%) | 2.00(2.00) | 1.90(2.00) | 1.98(2.00) | 1.86(2.00) |
| | TIME | 2.08(0.77) | 3.57(0.89) | 2.48(1.23) | 3.81(1.41) |
| MCP | CP(%) | 99.00(100.00) | 99.00(100.00) | 98.00(100.00) | 96.00(100.00) |
| | PCZ(%) | 100.00(100.00) | 100.00(100.00) | 100.00(100.00) | 100.00(100.00) |
| | PIZ(%) | 1.00(0.00) | 1.00(0.00) | 2.00(0.00) | 4.00(0.00) |
| | PER(%) | 98.00(100.00) | 96.00(100.00) | 98.00(100.00) | 95.00(100.00) |
| | RI(%) | 99.49(100.00) | 97.87(100.00) | 98.98(100.00) | 96.43(100.00) |
| | AMS(%) | 1.98(2.00) | 1.98(2.00) | 1.96(2.00) | 1.92(2.00) |
| | TIME | 2.51(0.77) | 3.87(0.99) | 2.51(1.29) | 4.21(1.45) |
| LASSO | CP(%) | 99.00(100.00) | 91.00(100.00) | 97.00(100.00) | 82.00(100.00) |
| | PCZ(%) | 99.98(99.98) | 99.98(99.98) | 100.00(100.00) | 100.00(100.00) |
| | PIZ(%) | 1.00(0.00) | 9.00(0.00) | 3.00(3.00) | 18.00(0.00) |
| | PER(%) | 99.00(100.00) | 90.00(100.00) | 97.00(100.00) | 82.00(100.00) |
| | RI(%) | 99.47(100.00) | 94.01(100.00) | 98.46(100.00) | 89.66(100.00) |
| | AMS(%) | 2.06(2.06) | 1.92(2.12) | 1.99(2.06) | 1.71(2.09) |
| | TIME | 1.31(0.66) | 2.07(1.11) | 1.98(0.96) | 2.67(1.60) |