Statistica Sinica **33** (2023), 1533-1553
doi:https://doi.org/10.5705/ss.202021.0276

INTEGRATIVE ANALYSIS FOR HIGH-DIMENSIONAL STRATIFIED MODELS

Jian Huang[1], Yuling Jiao[2], Wei Wang[3], Xiaodong Yan[3] and Liping Zhu[4]

[1] *University of Iowa,* [2] *Wuhan University,*
[3] *Shandong University and* [4] *Renmin University of China*

Abstract: In modern economic studies, the population heterogeneity of multiple strata and high dimensionality of predictors pose major challenges. In this study, we introduce an integrative procedure that can be used to explore group and sparsity structures of high-dimensional and heterogeneous stratified models. Furthermore, we propose K-regression modelling as a hybrid of complex and simple models that exhibits arbitrary dependence on the stratum features, but linear dependence on the other variables. K-regression models exhibit the following features:(i) they are essentially nonparametric with respect to the stratified feature, and have parametric linear effects in the other variables with a 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 automatically integrates the strata pertaining to a common regression model, and simultaneously estimates the corresponding effects; (iii) the proposed method quickly recovers subpopulation and sparsity structure of the K-regression models within massive high-dimensional strata; and (iv) the resulting estimators exhibit two-layer oracle properties, that is, 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 sampling scheme improves the integration accuracy. The results of simulation show that the proposed method performs appropriately for finite samples, and we demonstrate the usefulness of the method using real data.

Key words and phrases: Group fixed effect, heterogeneity, high-dimensionality, integrative analysis, K-regressio, massive data, stratum-specific bootstrap.

1. Introduction

Researchers use strata to fit stratified models for the value of each categorical feature. For example, financial shares may be classified based on their prices or trading volumes, with firms classified into subgroups based on the pairwise interactions of their credit ratings and industry attributes. With the explosive

Corresponding author: Xiaodong Yan, Zhongtai Securities Institute for Financial Studies, Shandong University, Jinan 250100, China. E-mail: yanxiaodong128@163.com.

growth of raw data available in the social and econometric sciences (Varian (2014); Einav and Levin (2014)), come from a variety of sources. This provides multiple strata from, for example, different experimental methods, geographic locations (census, tract, county, state, etc.), external classifications, observable explanatory categories, nested (hierarchical) or non-nested data sets, pooled cross-sectional data sets and panel data sets.

Although the homogeneous assumption (Phillips and Sul (2007); Browning and Carro (2007); Su and Chen (2013)) facilitates the estimation and inference procedures for certain specified common parameters, the results may be misleading if multiple strata exhibit heterogeneous structures. The stratum phenomenon in big data arises because populations can be heterogeneous across strata because they are based on data sets from different sources (Zhao, Cheng and Liu (2016)). Several studies have investigated the importance of stratified models and controlling the latent heterogeneity of panel data models by regarding an individual as a stratum (Pesaran and Tosetti (2011); Su and Jin (2012); Song (2013); Chudik and Pesaran (2015); Yang, Yan and Huang (2019); Li, Cui and Lu (2020)). However, modeling in each stratum is inadvisable, because having too few observations in each stratum cause "incidental parameter" issues, such as in panel data models (Hsiao and Pesaran (2008); Lu, Cheng and Liu (2016)). Therefore, many works assume that multiple strata belong to several homogeneous groups within a broadly heterogeneous population (Lin and Ng (2012); Bester and Hansen (2016); Bonhomme and Manresa (2015); Yan, Yin and Zhao (2021); Yan et al. (2022)). That is 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, Shi and Phillips (2016); Su and Ju (2018); Sarafidis and Weber (2015)). We can also generate a time-specified stratum (Bai (2010); Kim (2011)) in panel data, where multiple strata are obtained at different times, and one stratum includes observations of some subjects at a particular time point (Qian and Su (2016); Li, Qian and Su (2017)).

Ma and Huang (2016, 2017) proposed a pairwise-fusion penalized approach to conduct a subgroup analysis for heterogeneous intercepts and coefficients.

The heterogeneity of big data is usually coupled with high dimensionality. However, existing methods cannot be used directly to analyze stratified models subject to different sparsity structures across latent groups to simultaneously achieve high dimensionality and heterogeneity. Here, we explore common features among multiple strata that exhibit high dimensionality by combining the strata that originally belonged to a common group into one group, and estimating the sparsity structure of group-specific parameters. Our devised penalty-based K-

regression demonstrates the following. First, the K stratified regression models serve as a hybrid of complex and simple models. They implies that it depends arbitrarily on the stratum feature, but is simply (typically linearly) dependent on the other covariates. That is it is essentially nonparametric with respect to the stratified features, and parametric with a simple form in the case of other variables. Second, the numerical algorithm exhibits the computational ease and speed with respect to the integration of the common structure and the recovery of the sparsity information in each stratum. Third, 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. The finite-sample performance of the proposed method is evaluated in Sections 4 and 5. Section 6 concludes the paper. Proofs are provided in the online Supplementary Material.

2. Models and Estimators

2.1. The K-regression method

Let us assume that we observe data items or records of the form (Z_i, X_i, Y_i), for $i = 1, \ldots, n$, with a triple population form (Z, X, Y). Here, Z_i is an ordered or unordered categorical variable with M classes $\mathcal{Z} = \{z_1, z_2, \ldots, z_M\}$, based on which we create strata, and the corresponding sample size with respect to stratum z_m is $n_m = \sum_{i=1}^n I(Z_i = z_m)$, where $I(\cdot)$ denotes the indicator function. Furthermore, $n_1 + \cdots + n_M = n$, X_i is another 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. This implies that

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

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

Big data in econometrics typically comprise multiple data sets obtained from various sources (Varian (2014); Einav and Levin (2014)). For instance, data may be corrected from several locations, during different periods, or using different data collection procedures, thus generating a set of strata $\mathcal{Z}$. In econometrics, financial data are often collected from more than one thousand daily transactions for tens of thousands of stocks, resulting in M=10,000 strata. Big data in real estate might include 10,000 daily accumulated observations obtained over 10 years from 344 communities. Here, we calculate and the number of strata (3,440) as the product of the numbers of communities and years. This implies that we need to generate the stratum set $\mathcal{Z}$ flexibly through the interaction of different categorical variables. Furthermore, $\mathcal{Z}$ can be formed based on the values of a continuous variable by adopting a slicing technique. For example, by slicing the confidence interval [10,50] of stock prices into four partitions [10,20), [20,30), [30,40), [40,50], we obtain four strata.

Stratified models are more flexible than single or average models in terms of representing the heterogeneous stratum-specified characteristics. However, similarity or generality may exist across strata, owing to homogeneous characteristics, implying that the distinct stratified models may belong to one common model. Furthermore, there may be very few observations in some strata. Even when $n_m = 1$ or when the stratum-specified number of covariates is considerably larger than the number of observations, that is, $p \gg n_m$, we should borrow strength from neighborhood models. In reality, we do not know 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 $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$ as a mutually exclusive partition of $\{z_1, \ldots, z_M\}$, for $z_m \in \mathcal{G}_k$, $\boldsymbol{\beta}_{z_m} = \boldsymbol{\alpha}_k$, where $\boldsymbol{\alpha}_k$ is the common value of $\boldsymbol{\beta}_{z_m}$. Furthermore, we propose K-regression models using

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

The K stratified regression models, simplified as the K-regression models in (2.2), inherit the advantages of the stratified models in (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. Therefore, they are essentially nonparametric with respect to the stratified feature, and parametric with a simple form in relation to the other variables. Compared with the classical mixture model, the $K-$regression model (2.2) inherits the semiparametric superiority and robustness. In contrast to the model of Li et al (2022), the $K-$ regression model characterizes the group-specific heterogeneity by introducing the latent group pattern parameter $\mathcal{G}_k$.

The K-regression model in (2.2) can flexibly accommodate heterogeneity for multiple strata in econometric analysis of pooled cross-sectional data sets and panel /longitudinal data sets. In the former case, strata are collected from M different time periods by observing different subjects during each temporal interval; the mth stratum contains n_m individuals. K-regression modelling is designed to detect K populations across the M strata, and specify which strata belong to common groups. In the case of panel/longitudinal data sets, where M strata are generated from M subjects (e.g., individuals, firms, countries, or regions) over n_m repeated measurements on the mth stratum. These data sets denote balanced panel/longitudinal data if $n_1 = \cdots = n_m$. The $K-$regression model specifies the K subgroup structures and integrates individuals belonging to a common subgroup.

Remark 1. We obtained an interesting discovery related to integrative analyses of balanced panel data sets by generating the M strata from M time points (not subjects), where the mth stratum covers n_m observations on n_m respective subjects. As such, the indices $\{z_1, \ldots, z_M\}$ are ordered categorical values, and do not exhibit a qualitative nature. Therefore, an integrative procedure searches for the locations of structural break jumps (Qian and Su (2016); Li, Qian and Su (2017)). Specifically, K subgroup divisions imply K-1 structural breaks, and the temporal intervals of structural breaks can be obtained as $\{(\max\{\mathcal{G}_{k-1}\}, \min\{\mathcal{G}_k\}) : k = 2, \ldots, K\}$, where $\max\{A\}$ and $\min\{A\}$ denote the maximum and minimum values, respectively, in set A.

Remark 2. Apart from being generated across multiple time periods, as in the case of pooled cross-sectional or panel/longitudinal data sets, M strata can also be collected from different geographic locations or by using experimental methods. We can also divide n observations into M strata based on s observable categorical variables, such as gender or race.

2.2. Estimator and computation

The K-regression models in (2.2) can be used to achieve our main objective of statistical estimation and inference with respect to the Kp coefficient vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_K^\top)^\top$ and the group parameter $\mathcal{G}$. Given the value K and the tuning parameter λ, the estimators of $\mathcal{G}$ and $\boldsymbol{\alpha}$ 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 the sparsity structure of α_k is considered, note that each of the potential K strata can also share information on α_k. For example, we can use a group penalty (Yuan and Lin (2006)) to detect the group structures and strata-specific coefficients, and can use a fused penalty (Tibshirani et al. (2005)) on the pairwise difference between α_{kj} and α_{kl} to check their order values.

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

$$\mathcal{G}_k(\boldsymbol{\alpha}) = \left\{ m : \left\{ \underset{k \in \{1,\ldots,K\}}{\operatorname{argmin}} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k\|^2 \right\} = k \right\}. \quad (2.4)$$

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

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

Subsequently, we can estimate $\mathcal{G}$ as $\widehat{\mathcal{G}} = (\widehat{\mathcal{G}}_1(\widehat{\boldsymbol{\alpha}}), \ldots, \widehat{\mathcal{G}}_K(\widehat{\boldsymbol{\alpha}}))^\top$.

In addition, we explore the close connection between (2.3) and the well-known k-means clustering algorithm to obtain a fast and efficient computing procedure. The simple and fast iterative algorithm, presented in Algorithm A.1 in the Supplemental Materials, generates a group estimator $\widehat{\mathcal{G}}$ and the coefficient estimator $\widehat{\boldsymbol{\alpha}}$ in (2.3) using the optimizations 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 the tuning parameter λ is fast, for two reasons. First, it quickly alternates between the integrative and update steps. The "integrative" step minimizes the objective function with respect to the membership assignment given fixed α_k and determines the integration of stratum m into the kth subpopulation with respect to the minimum quadratic loss, resulting in a rapid computation. In the "updated" step, we update the estimator $\boldsymbol{\alpha}_k^{(s+1)}$ separately for $k = 1, \ldots, K$ as

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

and the fast coordinate descent algorithm is used to calculate $\boldsymbol{\alpha}_k^{(s+1)}$. Second, the objective function becomes non-increasing during the iterative procedure, resulting in rapid numerical convergence. However, Algorithm A.1 in the Supplemental Materials is sensitive to the starting point $\boldsymbol{\alpha}^{(0)}$. Therefore, we use a k-means computational strategy to generate several initial values, and thus obtain stable estimators. We also consider a more efficient alternative in which we use the variable neighborhood search method as the heuristic to solve the minimum sum-of-squares partitioning problem (Bonhomme and Manresa (2015)), allowing for high-dimensional covariates. The procedure is presented as Algorithm A.2 in the Supplementary Material.

3. Theoretical Results

3.1. Notation

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

First, we introduce some notation and regularity conditions. Under the sparsity assumption of every subpopulation in high dimensionality, we obtain $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{k1}^\top, \boldsymbol{\alpha}_{k2}^\top)^\top$, where $\boldsymbol{\alpha}_{k1} \in \mathcal{R}^{q_k}$ and $\boldsymbol{\alpha}_{k2} \in \mathcal{R}^{p-q_k}$ are the nonzero and zero components, respectively, of $\boldsymbol{\alpha}_k$. Using this 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, by ranking the nonzero part of the parameters ahead of the zeros, $\boldsymbol{\alpha}$ can be rewritten as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\mathcal{K}1}^\top, \boldsymbol{\alpha}_{\mathcal{K}2}^\top)^\top$, where $\boldsymbol{\alpha}_{\mathcal{K}1} = (\boldsymbol{\alpha}_{11}^\top, \ldots, \boldsymbol{\alpha}_{K1}^\top)^\top$ and $\boldsymbol{\alpha}_{\mathcal{K}2} = (\boldsymbol{\alpha}_{12}^\top, \ldots, \boldsymbol{\alpha}_{K2}^\top)^\top$. Then the true coefficient vector $\boldsymbol{\alpha}_0 = (\boldsymbol{\alpha}_{0\mathcal{K}1}^\top, \mathbf{0}^\top)^\top$, $\mathrm{supp}(\boldsymbol{\alpha}_{0\mathcal{K}1}) = \sum_{k=1}^K q_k = q_{\mathcal{K}}$, and the estimator $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}_{\mathcal{K}1}^\top, \widehat{\boldsymbol{\alpha}}_{\mathcal{K}2}^\top)^\top$. Furthermore, $\mathcal{G}_0 = \{\mathcal{G}_{01}, \ldots, \mathcal{G}_{0K}\}$ and $\widehat{\mathcal{G}} = \{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_K\}$ 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 $\Pi = \widetilde{\Pi} \otimes I_p$, $\boldsymbol{Y} = \mathrm{diag}(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M)$ and $\mathbb{Y} = (\boldsymbol{Y}\widetilde{\Pi})^+$, where A^+ denotes a vector obtained from the row sums of matrix A. In addition, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{z_1}^\top, \ldots, \boldsymbol{\epsilon}_{z_M}^\top)^\top = (\epsilon_1, \ldots, \epsilon_n)^\top$, $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{X}_{z_1}, \ldots, \boldsymbol{X}_{z_M})$, $\mathbb{X} = (\boldsymbol{X}\Pi)_{n\times(Kp)}$, and $\mathbb{X}_1$ and $\mathbb{X}_2$ are $n \times q_{\mathcal{K}}$ and $n \times (Kp - q_{\mathcal{K}})$ submatrices, respectively, $\mathbb{X}$ corresponding to the decomposition of $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\mathcal{K}1}^\top, \boldsymbol{\alpha}_{\mathcal{K}2}^\top)^\top$.

Note that $\Pi^\top\Pi = \mathrm{diag}(|\mathcal{G}_{01}|, \ldots, |\mathcal{G}_{0K}|) \otimes I_p$. 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 the vector of the ith row of A. Let $\gamma_{\min}(A)$ and $\gamma_{\max}(A)$

be the smallest and largest eigenvalues, respectively, of A, and let

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

be the minimum difference between the coefficient vectors of all combinations of between every two populations.

Denote $d_n = (1/2)\min_{1\leq j\leq q_{\mathcal{K}}} |\alpha_{0\mathcal{K}1j}|$ as the half of minimum signal. Let $N_k = \sum_{z_m \in \mathcal{G}_{0k}} n_m$, and let $N_{\min} = \min_{1\leq k\leq K} N_k$ and $N_{\max} = \max_{1\leq k\leq K} N_k$ represent the true minimum and maximum sample sizes among all populations. Furthermore, let $n_{\min} = \min_{1\leq m\leq M} n_m$ and $n_{\max} = \max_{1\leq m\leq M} n_m$ denote the minimum and maximum sample sizes, respectively, among all strata, and let $p'_\lambda(a)$ and $p''_\lambda(a)$ denote the first and second derivations, respectively, of the 5 penalty $p_\lambda(a)$ about a. Let c and $c'_j s$ denote some positive constants.

3.2. Oracle property with a known group structure

If the underlying group parameter $\mathcal{G}_0$, that is, the matrix Π, is known, we can define an estimator as

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

Furthermore, $\widetilde{\boldsymbol{\alpha}}$ can be rewritten as $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1}^\top, \widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2}^\top)^\top$. However, the group relationship of the strata, that is, $\mathcal{G}_0$, is typically unknown in advance, and the oracle estimators are infeasible in practice. Nevertheless, this can provide some insight into 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 nonincreasing in $t \in (0, +\infty)$, and $p'_\lambda(t)$ is increasing in λ 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 $i = 1, \ldots, n$.

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

(C4) (i) $\gamma_{\min}(\mathbb{X}_1^\top \mathbb{X}_1) \geq c_2 N_{\min}$, $\gamma_{\max}(\mathbb{X}_1^\top \mathbb{X}_1) \leq c_3 n$. (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), including the 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 of the high-dimensional regression scenario and is 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 $\lambda = p'_\lambda(d_n) = O(K\sqrt{N_{\min}}/n)$ is incompatible with $\lambda = p'_\lambda(0_+) \gg K q_\mathcal{K} \sqrt{q_\mathcal{K}/N_{\min}}$. This contradiction implies that the LASSO-based K-regression estimator cannot, in general, attain the consistency rate obtained from Theorem 1 and the oracle property achieved in Theorem 2. Fan and Lv (2011) observed this issue under homogeneous high-dimensional data with $K = 1$, and $N_{\min} = n$, and the corresponding penalty conditions $p'_\lambda(d_n) = O(1/\sqrt{n})$, $d_n \gg \sqrt{s/n}$, 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

$$\mathbb{X}_1^\top \mathbb{X}_1 = \operatorname{diag}\left(\sum_{z_m \in \mathcal{G}_{0k}} \boldsymbol{X}_{z_m 1}^\top \boldsymbol{X}_{z_m 1}, k = 1, \ldots, K \right),$$

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 $\boldsymbol{X}_{z_m 1}$ denotes the submatrices of $\boldsymbol{X}_{z_m}$, formed by the columns in $\operatorname{supp}(\boldsymbol{\alpha}_{0k})$. Without loss of generality, the covariates can be scaled in every subpopulation, as assumed in Condition (C4) (ii), and then $\operatorname{tr}(\mathbb{X}_1^\top \mathbb{X}_1) = \sum_{k=1}^K q_k N_k$.

Theorem 1 (Consistency for estimator $\widetilde{\boldsymbol{\alpha}}_\mathcal{K}$ with group pattern known). *Under Conditions* (C1)–(C4) *and the additional condition* $\log(Kp) = o(n^2 q_\mathcal{K}/N_{\min}^2)$, *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 the objective function* (3.1) *such that* $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}2} = 0$ *with probability tending to one as* $N_{\min} \to \infty$ *and*

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

Theorem 1 establishes the consistency of the proposed penalized K-regression estimator $\widetilde{\boldsymbol{\alpha}}_{\mathcal{K}1}$; that is, there is a root-$(N_{\min}/q_\mathcal{K})$-consistent K-regression estimator of $\boldsymbol{\alpha}_{0\mathcal{K}1}$ under dimensionality p and population number K that satisfies $Kp = o\{\exp(n^2 q_\mathcal{K}/N_{\min}^2)\}$ 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 probability tending to one. Theorem 1 also addressed the strength of 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 the 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 the conditions of Theorem 1, *with Condition* (C3) *(i) replaced by* $p'_\lambda(d_n) = O(K/n)$, *and attaching the additional conditions* $q_{\mathcal{K}} = o(N_{\min})$, *and*

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

we conclude that

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

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* $\|a_n\| = 1$, *and* $\xrightarrow{D}$ *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 more slowly than $N_{\min}$ does. Combined with the conditions $N_{\min} \gg n^{5/6} q_{\mathcal{K}}^{1/2}$ and $N_{\min} \leq n/K$, we conclude that $K = o(n^{1/6})$, and thus Theorem 2 indicates that the number of subpopulations K is assumed to grow more slowly than $n^{1/6}$.

3.3. Theoretical property with group structure unknown

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

(C5) (i) $b_n \gg \sqrt{q_{\mathcal{K}} \log(n)/n_{\min}}$; (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 the 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 $n_{\min} \to \infty$.*

Theorem 3 implies that if the minimum difference of the common effects between any two subpopulations satisfies $b_n \gg \sqrt{q_{\mathcal{K}} \log(n)/n_{\min}}$, then 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.

Corollary 1. *Under the 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 $\|a_n\| = 1$, and $\xrightarrow{D}$ denotes convergence in distribution.

The asymptotic distribution of the K-regression estimators provides a theoretical justification for further statistical inference, such as testing for heterogeneity. Based on Corollary 1, we present a unified framework 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 $d = \text{rank}(\mathcal{B})$. This 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 α_k, and $H_0 : \alpha_j - \alpha_k = 0$, $j, k \in \{1, \ldots, K\}$, which can be used to test for the existence of effect heterogeneity among strata. We develop a χ^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 $\widehat{\sigma}^2 = (1/K)\sum_{k=1}^K (1/\sum_{z_m \in \widehat{\mathcal{G}}_k} n_m) \sum_{z_m \in \widehat{\mathcal{G}}_k} \|\boldsymbol{Y}_{z_m} - \boldsymbol{X}_{z_m}\widehat{\boldsymbol{\alpha}}_k\|^2$.

Theorem 4. *Under the null hypothesis and the conditions in Theorem 3, we have $\mathcal{T}_n(\mathcal{B}) \xrightarrow{D} \chi^2_d$ as $n \to \infty$, where χ^2_d denotes a 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 Wilk's theorem. The $100(1-\tau)\%$ approximated confidence region for $\mathcal{B}\boldsymbol{\alpha}$ is given by

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

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

4. Simulation Studies

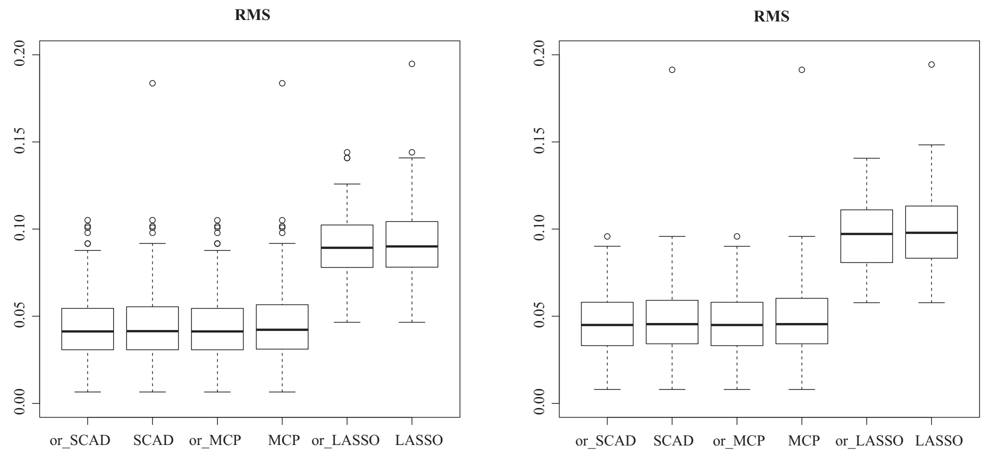
Next, we consider an example in which we evaluate the performance of our method. The preliminaries we adopt to measure the simulated results and additional examples are provided in the Supplemental Materials.

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

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

where $\boldsymbol{X}_{z_m}$ is assumed to be generated from a multivariate normal distribution with zero mean and covariance matrix $\Phi = (d_{jl})$, with $d_{jl} = 0.7^{|j-l|}$, $\boldsymbol{\epsilon}_{z_m}$ is assumed to follow the normal distribution $\mathcal{N}(0, 0.7^2)$. We randomly assigned the strata to two subpopulations, with equal probabilities; that is, $K = 2$ and $P(z_m \in \mathcal{G}_1) = P(z_m \in \mathcal{G}_2) = 1/2$, so that the coefficients are equal to $\boldsymbol{\alpha}_1$ for $z_m \in \mathcal{G}_1$, and are equal to $\boldsymbol{\alpha}_2$ for $z_m \in \mathcal{G}_2$, where $\boldsymbol{\alpha}_1 = (1, 0.8, \mathbf{0}_{p-2}^\top)^\top$, and $\boldsymbol{\alpha}_2 = (-1, -0.8, \mathbf{0}_{p-2}^\top)^\top$. We choose $n = 600$, and $p = 500$ or 1,000 with two numbers of strata, that is, $M = 100, 200$, and examine the performance of our proposed method under three penalized methods: the SCAD, MCP, and LASSO.

Table 1 and Figure 1 present the estimated results for Example 1. The results in parentheses denote the oracle estimates with known $\mathcal{G}_0$. We note several points. First, the simulated results in Table 1 in the considered measurements using the SCAD, MCP, and LASSO penalties are similar, and the estimates obtained using the three methods are close to their corresponding oracle estimates. Second, the K-regression method can accurately integrate the strata with a common population for estimated RI values that are approximately one. Third, based on sparsity-induced penalties, the proposed method behaves satisfactorily because the corresponding average numbers of accurately estimated zero components are similar to the true number $p-2$ of zero components in each population, whereas their corresponding average numbers of inaccurately estimated zero coefficients approach zero with respect to the PIZ



(a) (b)

Figure 1. Box plots of RMS under the three penalized methods with 100 replicates and $M = 100$, $p = 500$ (left), and $p = 1{,}000$ (right) in Example 1.

Table 1. Simulation results for different variable selection methods in Example 1 with $n = 600$; the results in parentheses denote the oracle estimates with known $\mathcal{G}_0$.

Selection method	Criterion	p=500 M=100	p=500 M=200	p=1,000 M=100	p=1,000 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)

and PCZ indices. Here, PCZ denotes the percentage of correct zeros with PCZ (%)$=(1/MT)\sum_{m=1}^{M}\sum_{t=1}^{T}\{(100\%/(p-|\mathcal{D}_{z_m}|))[\sum_{j=1}^{p} I(\widehat{\beta}_{mj(t)}=0)I(\beta_{0mj}=0)]\}$, and PIZ is the percentage of incorrect zeros with PIZ (%) $=(1/MT)\sum_{m=1}^{M}\sum_{t=1}^{T}$ $[(100\%/(|\mathcal{D}_{z_m}|))\{\sum_{j=1}^{p} I(\widehat{\beta}_{mj(t)}=0)I(\beta_{0mj}\neq 0)\}]$, where $\mathcal{D}_{z_m}=\{j:\beta_{0mj}\neq 0\}$ is the index of the true model in the mth stratum. Note that a larger PCZ and a smaller PIZ imply a good model-fitting procedure. Fourth, the K-regression can recover the subpopulation and sparsity structure in just a few seconds, even when the sample size n is large and the dimension p is high. Fifth, in Figure 1, the RMS values of the SCAD/MCP-based K-regression methods are smaller than those of the LASSO estimators and attain the oracle estimates with the known group structure $\mathcal{G}_0$. This 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 does not. Sixth, the proportion of the specified numbers of subpopulations $\widehat{K}$ equal to the true number is close to one using the K-regression method, based on the BIC. Seventh, increasing the number of strata (i.e., M) or the dimensionality (i.e., p) decreases the performance of the proposed method in relation to all the considered indices; Eighth, the level plots in Figure 2 adopts use to denote the component value of a coefficient matrix, and are generated based on the average estimates after 200 replicates, that is, the estimated $M\times p$ coefficient matrix $\widehat{\boldsymbol{\beta}}=(1/200)\sum_{t=1}^{200}\widehat{\boldsymbol{\beta}}_{(t)}$, under a fixed partition. The results imply that separate statistical modelling in each stratum (i.e., an M-penalty) dramatically reduces the quality of the estimates, whereas our proposed integrative analysis (i.e., the K-penalty) using K-regression methods ensures an the efficient estimation of the coefficient matrix by accurately recovering each subpopulation and its corresponding sparsity structure. Our simulation results show that the proposed K-regression method exhibits desirable behaviour in terms of integration, variable selection, parameter estimation, and computational speed, implying that the empirical results are consistent with those presented in Theorem 1.

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

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

Here, $\widehat{\mathcal{B}}=\widehat{\mathcal{D}}\otimes I_{q_{\mathcal{K}}}$, where $\widehat{\mathcal{D}}=\{(e_i-e_j),i<j\}^{\top}_{((K(K-1))/2)\times K}$, with e_i being the ith $K\times 1$ stratum vector, with the ith element equal to one and the remaining equal to zero. Furthermore, $I_{q_{\mathcal{K}}}$ is a $q_{\mathcal{K}}\times q_{\mathcal{K}}$ identity matrix and $\otimes$ is the Kronecker product. Theorem 4 indicates that rank$(\widehat{\mathcal{B}})=q_{\mathcal{K}}(K-1)$, and the mean of the p-values based on 200 replicates is given by $(1/200)\sum_{j=1}^{200}\chi^2_{q_{\mathcal{K}}(K-1)}(\mathcal{T}_n^{(j)}(\widehat{\mathcal{B}}))$, where

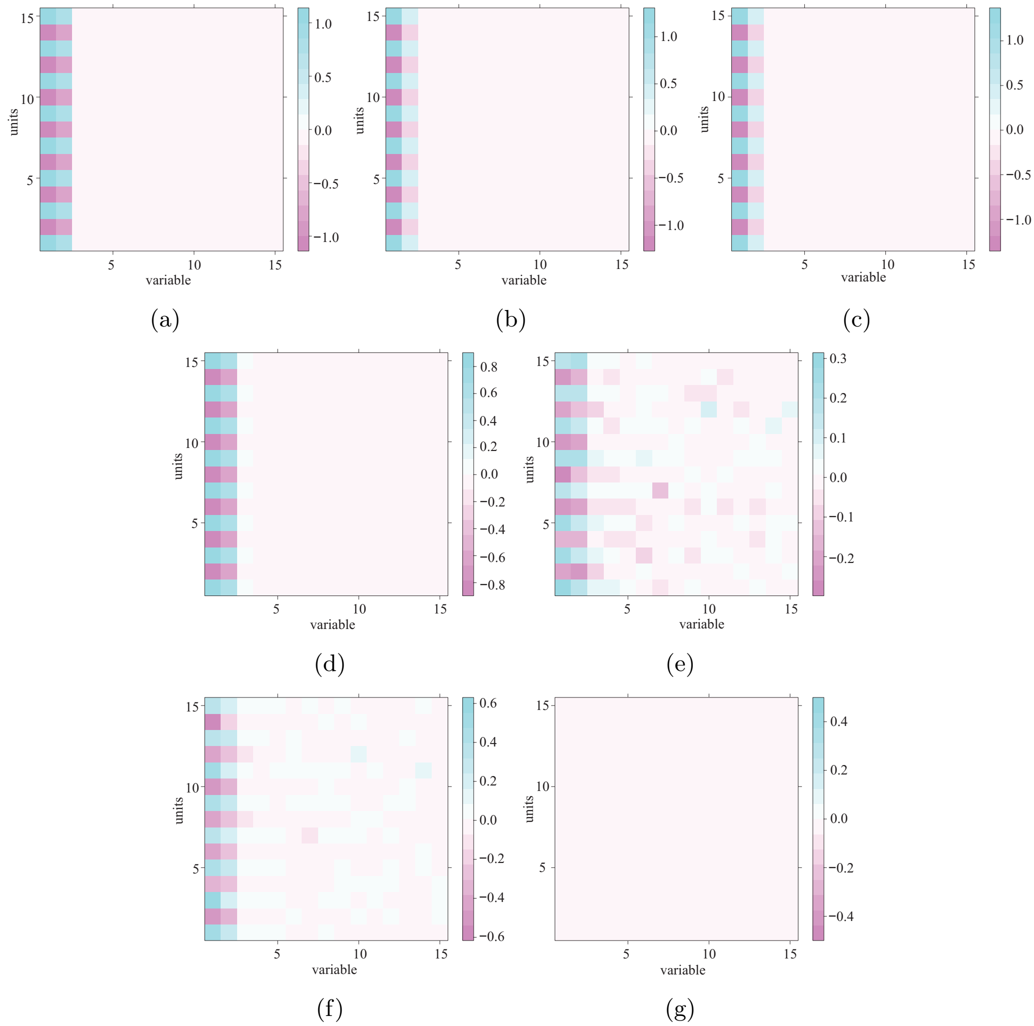
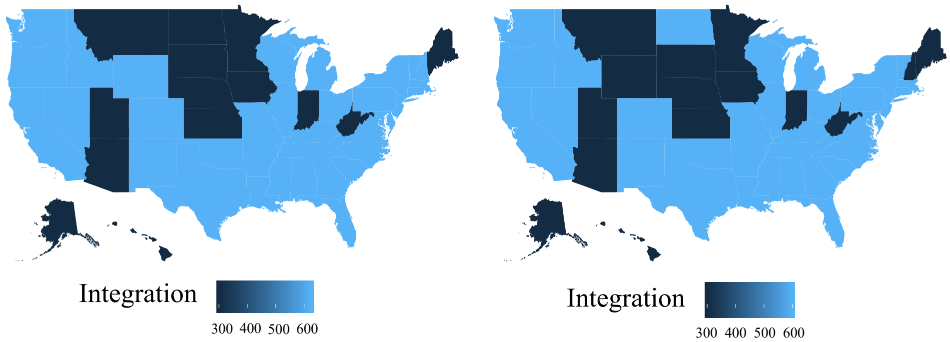


Figure 2. Level plots for the coefficient matrix estimation with $n = 60$, $M = 15$, and $p = 15$ in Example 1. 38TRUE shows a level plot of the true coefficient matrix; M-penalty represents methods that conduct statistical modelling based on each stratum separately; K-penalty denotes our penalty-based K-regression.

$\mathcal{T}_n^{(j)}(\widehat{\mathcal{B}})$ is the value of $\mathcal{T}_n(\widehat{\mathcal{B}})$ from the jth 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 χ^2 distribution with $q_{\mathcal{K}}(K-1)$ degrees of freedom. The mean p-values are all less than 0.001 in Example 1, which strongly supports the existence of effect heterogeneity in this example. The simulated results in Example 1 also suggest that the consistency of the estimation of the group-specific coefficient under the K-regression method is completely dependent on the integration accuracy, that is, $\widehat{\mathcal{G}}$.



(a) (b)

Figure 3. Subfigures (a) and (b) responds to K-SCAD and K-MCP, respectively. The proportion of crimes in the United States; the dark and light blues areas denote Populations 1 and 2, respectively.

5. Empirical Study

In this section, we apply our proposed K-regression method to communities and crime (CAC) data obtained from the UCI Machine Learning Repository. The data sets comprises information from different communities in the United States, 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 used to identify the community or state, explained by its corresponding abbreviated name, the data sets 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, $n = 2{,}215$ communities, and $p = 102$ covariates. Assuming that the samples of 48 states originate from K populations, the K-regression models can be defined as

$$\boldsymbol{Y}_{z_m} = \boldsymbol{X}_{z_m}\boldsymbol{\alpha}_k + \boldsymbol{\epsilon}_{z_m}, \quad k = 1, \ldots, K, z_m \in \mathcal{G}_k, \tag{5.1}$$

with $\sum_{k=1}^{K} |\mathcal{G}_k| = 48$. We also estimate the number of regression models K, group parameters $\mathcal{G}_k$, and coefficients $\boldsymbol{\alpha}_k$.

Based on the superior performance of the concave penalties in terms of estimation and variable selection, we applied SCAD- and MCP-based K-regression methods to recover the subpopulation and sparsity structure in the assumed model (5.1) on the CAC data set. We specify the optimal K using introduced

Table 2. Performance of the estimates of the CAC data using 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.

the BIC criterion in the Supplementary Materials. The eventual integration results are presented in Figure 3 and Table 2. First, as shown, the SCAD and MCP methods estimate the 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 the SCAD, whilst belonging to $\widehat{\mathcal{G}}_2$ by the MCP, and the states WY and NH are partitioned into $\widehat{\mathcal{G}}_2$ by the SCAD, regardless of being merged into $\widehat{\mathcal{G}}_1$ by the MCP. Second, the result of the coefficient estimation (Est.) and the corresponding p-values in Population 1 by the two concave penalties commonly and significantly specify the positive effects of the RPB and NIST on the murder ratio, whereas the NIST feature imposes zero effect on the response of interest in Population 2. Third, although the covariates HV, MPD, and PVB do not affect the murder ratio in Population 1, they exhibit considerable influence in Population 2. Fourth, the existence of heterogeneity is verified through $\mathcal{T}_n(\mathcal{B})$ in (3.2), where $\mathcal{B} = \mathcal{D} \otimes I_{q_{\mathcal{K}}}$, with $\mathcal{D} = \{(e_i - e_j), i < j\}^{\top}_{((K(K-1))/2)\times K}$ and $K = 2$, and $q_{\mathcal{K}} = 10$ and $q_{\mathcal{K}} = 9$ for the SCAD and MCP penalties, respectively. Theorem 4 indicates that $\mathrm{rank}(\mathcal{B}) = q_{\mathcal{K}}(K-1)$. Then, the calculated p-value is $\chi^2_{\mathrm{rank}(\mathcal{B})}(\mathcal{T}_n(\mathcal{B})) = 0.008$ and 0.010 by the 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, where a bigger population density corresponds to Population 2. In addition, the number of significant factors influencing the number of murders in Population 2 is apparently much larger, which may be attributed to a more complex environment along with a high population density.

6. Conclusion

In this study, we have developed a K-regression model to simultaneously integrate strata with a common regression structure and to estimate stratum-specific fixed effects, thus accommodating to accommodate the unobserved heterogeneity in the multiple strata. In simulations and real-data examples, the K-regression method exhibits superior performance with respect to fast integration and accurate variable selection. This is because massive data often comprise multiple high-dimensional strata, derived from a growing number of heterogeneous subpopulations with an unknown common structure and sparsity information, K-regression modelling is naturally scalable and can deal with heterogeneous issues related to massive data sets. We have also learned that the statistical inference in integrative analysis depends on the subpopulation and sparsity recovery, resulting in inference uncertainty. Thus, a “post-integration and selection” issue arises, requiring additional future research.

7. Supplementary Material

In the Supplementary Material, subsection 1 introduces the Stratum-specific bootstrap. Subsection 2 shows additional results of Monte Carlo simulations. section which contains a brief description of the online supplementary materials. Subsection 3 display further empirical analysis on CAC data sets. Subsection 4 provides the proofs of the theorems in this paper.

Acknowledgments

The authors are grateful to the editor, associate Editor, and two referees for their valuable suggestions and comments. Xiaodong Yan was supported by the National Natural Science Foundation of China (grant number 11901352), National Statistical Science Research Project (grant number 2022LY080); Jinan Science and Technology Bureau (grant number 2021GXRC056); Wei Wang was supported by the Nature Science Foundation of Shandong Province (grant number ZR202111150121); Jian Huang was supported in part by the U.S. National Science Foundation grant DMS-1916199. Liping Zhu was supported by National Natural Science Foundation of China (12171477, 11731011, 11931014) and Natural Science Foundation of Beijing Municipality (Z190002).

References

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

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.

Browning, M. and Carro, J. M. (2007). Heterogeneity and microeconometrics modelling. In *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*. Cambridge University Press, New York.

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. and Pesaran, M. H., (2008). Random coefficient panel data models. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*. 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., Cui, G. and Lu, L. (2020). Efficient estimation of heterogeneous coefficients in panel data models with common shocks. *Journal of Econometrics* **216**, 327–353.

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

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. (2016). Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv 1607.03717.*

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

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

Phillips, P. C. and Sul, D. (2007). Transition modeling and econometric convergence tests. *Econometrica* **75**, 1771–1855.

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.

Sarafidis, V. and Weber, N. (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 Chen, Q. (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. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics* **206**, 554-573.

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

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**, 91–108.

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

Yan, X., Wang, H., Zhou, Y., Yan, J., Wang, Y., Wang, W. et al. (2022). Heterogeneous logistic regression for estimation of subgroup effects on hypertension. *Journal of Biopharmaceutical Statistics* **16**, 1–17.

Yan, X., Yin, G. and Zhao, X. (2021). Subgroup analysis in censored linear regression. *Statistica Sinica* **31**, 1027–1054.

Yang, X., Yan, X. and Huang, J. (2019). High-dimensional integrative analysis with homogeneity and sparsity recovery. *Journal of Multivariate Analysis*, **174**, 104529.

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**, 49–67.

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

Zhao, T., Cheng, G. and Liu, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* **44**, 1400–1437.

Jian Huang

Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52246, USA.

E-mail: jian-huang@uiowa.edu

Yuling Jiao
Zhongnan University of Economics and Law, Wuhan 430073, China.
E-mail: yulingjiaomath@whu.edu.cn
Wei Wang
Zhongtai Securities Institute of Financial Studies, Shandong University, Jinan 250100, China.
E-mail: wangwei_0115@outlook.com
Xiaodong Yan
Zhongtai Securities Institute for Financial Studies, Shandong University; Shandong National Center for Applied Mathematics, Jinan 250100, China.
E-mail: yanxiaodong128@163.com
Liping Zhu
Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.
E-mail: zhu.liping@ruc.edu.cn

(Received August 2021; accepted February 2022)