The Annals of Applied Statistics
2024, Vol. 18, No. 3, 2532–2550
https://doi.org/10.1214/24-AOAS1893

MIXTURE CONDITIONAL REGRESSION WITH ULTRAHIGH DIMENSIONAL TEXT DATA FOR ESTIMATING EXTRALEGAL FACTOR EFFECTS

BY JIAXIN SHI[1,a], FANG WANG[2,d], YUAN GAO[1,b], XIAOJUN SONG[3,e] AND HANSHENG WANG[1,c]

[1]*Guanghua School of Management, Peking University,* [a]*jxshi@stu.pku.edu.cn,* [b]*ygao_stat@outlook.com,* [c]*hansheng@gsm.pku.edu.cn*

[2]*Data Science Institute, Shandong University,* [d]*wangfang226@sdu.edu.cn*

[3]*Guanghua School of Management and Center for Statistical Science, Peking University,* [e]*sxj@gsm.pku.edu.cn*

Testing judicial impartiality is a problem of fundamental importance in empirical legal studies for which standard regression methods have been popularly used to estimate the extralegal factor effects. However, those methods cannot handle control variables with ultrahigh dimensionality, such as those found in judgment documents recorded in text format. To solve this problem, we develop a novel mixture conditional regression (MCR) approach, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector, which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a naïve Bayes type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, we are able to estimate the key parameters of interest as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offenses is analyzed for illustration purposes.

1. Introduction. Our research is empirically motivated by studies of equality, impartiality, and justice in jurisprudence (L'Heureux-Dube (2001), Meyerson (2006)). Fairness and justice have defined features of the judicial role, including aspects such as substantive decision-making by judges (Weiler (1968)), procedural justice (Krehbiel and Cropanzano (2000)), judicial independence (Meron (2005)), and the proper assessment of scientific evidence (Edmond (2002)). We emphasize criminal substantive justice in this paper. This means that judicial decisions must follow the principle of legality and should not be affected by prejudice regarding races, incomes, and other extralegal factors (Bright (2008), Lynch and Haney (2011)). Substantive justice represents the ultimate good of judicial impartiality. It is the core standard of good conduct and is essentially crucial for public confidence in the courts. In this regard, countries around the world have been promoting sustained reforms to unravel miscarriages of justice and safeguard judicial impartiality (Stith et al. (1998), Wadham (1993), Ye (2010)). By doing so, people wish the extralegal factor effects on judicial impartiality can be controlled and minimized.

Despite the fact that judicial impartiality has been universally promoted over the world for a long history, bias and prejudice, due to extralegal factors, do exist in practice. In fact, this is one of the most important research areas in empirical legal studies (Gross and Shaffer (2012),

Received September 2023; revised March 2024.
Key words and phrases. Expectation-maximization algorithm, judicial impartiality, mixture conditional regression, naïve Bayes model, ultrahigh dimensional data.

Nobles and Schiff (1995), Roberts (2003)). Researchers have made enormous efforts to document cases and analyze the reasons behind them. For example, Mishler and Sheehan (1993) studied about 4000 cases from the U.S. Supreme Court database in the period of 1956–1989 and found that the court has been highly responsive to public opinion. Its decisions have not only reflected the American public's overall policy preferences but also reinforced and legitimized emerging majoritarian concerns. Steffensmeier and Kramer (1998) investigated 139,000 criminal conviction cases from Pennsylvania in 1989–1992 and found blacks and males to be more likely to be incarcerated and to receive longer sentences, even after controlling for the type and severity of the offense and the offender's prior record. Bushway and Piehl (2001) analyzed 14,633 sentenced offenders from the state of Maryland and reported that African–Americans had 20% longer sentences, on average, than whites, holding constant age, gender, and recommended sentence length from the guide. Canes-Wrone, Clark and Kelly (2014) studied 2078 death penalty decisions issued by U.S. state courts under four judicial selection systems between 1980 and 2006 and found that judges were significantly more responsive to majority opinions on capital sentences in nonpartisan election systems than partisan systems. Glynn and Sen (2015) examined 2674 unique votes cast by 244 appeal judges from the U.S. Courts of Appeals on gender-related cases. They found that judges with daughters consistently tended to vote in a more feminist fashion on gender issues than judges with only sons. Bielen and Grajzl (2021) focused on 766 violent criminal cases that occured during the 12-week interval around the day of Theo van Gogh's assassination (November 2, 2004). They found that, immediately afterward, the prospects of prosecution for unrelated violent crimes with male suspects born in Muslim-majority countries increased by about 19%.

To summarize, there have been ample amount of empirical studies exploring a possible dependent relationship between judicial decisions and the interested extralegal factors, after controlling for a number of legal factors. Such problems can be nicely formulated as a regression problem with both the main covariates of interest and a set of control variables. Specifically, the dependent variable (Y) is a judicial decision, the key variables (X_1) are extralegal factors (e.g., race, gender, public opinion), and the control variables (X_2) are legal factors (e.g., severity of the current offense, the offense type, the previous criminal record). Then the interested problem becomes one of testing the statistical significance of the conditional regression relationship between Y and X_1, after controlling for the effect of X_2. Assuming that the law is perfectly just and self-contained and that no partiality or prejudice exists, we should expect Y and X_1 to be conditionally uncorrelated with each other, after controlling for the effect of X_2.

To fix the idea, consider for example the study of Pennsylvania criminal conviction cases in 1989–1992 (Steffensmeier and Kramer (1998)), where the response variable (Y) is the decision whether to incarcerate an offender; the extralegal factors of interest (the main covariates X_1) include race, gender, age, and their interaction, and the legal characteristic variables (the control variables X_2) include the type and severity of the offense and the offender's criminal record. To test the conditional regression relationship between Y and X_1 after controlling for X_2, a standard logistic regression model was employed. In a recent study of criminal cases happened before and after Theo van Gogh's assassination (Bielen and Grajzl (2021)), the dependent variable (Y) is the decision whether to prosecute a charge; the interested extralegal factors (the main covariates X_1) are the unrelated extraneous events (e.g., the Theo van Gogh's assassination), and the legal factors (the control variables X_2) include the criminal history and type of charge. In this case the difference-in-differences (DID) regression approach was used to study the conditional regression relationship between Y and X_1, after controlling for the effect of X_2. The literature suggests that testing conditional regression relationships should be of great importance for examining the effects of extralegal factors on judicial impartiality.

As one can see, to test the conditional regression relationship between the judicial decision (Y) and the extralegal factors (X_1), after controlling for the effects of legal factors (X_2), various standard regression models have been adopted for Y and (X_1, X_2). Those regression methods are easy to implement and have a nice interpretation. However, they also suffer from one serious limitation. That is, they can only handle control variables (X_2) with a relatively low dimension. For our empirical study of Chinese judicial decisions, the legal factors (X_2) contain features extracted from legal documents, which are lengthy text documents. Each element of the X_2 vector is then of a binary form, representing the existence or not of one particular keyword in the judgments. Since the original judgments are written in Chinese and contain a large number of legal issue-related keywords, the dimension of X_2 is very high, which precludes the immediate use of standard regression methods. Hence, how to test the conditional regression relationship between Y and X_1 with an ultrahigh dimensional X_2 becomes a problem of great importance.

Before we formally solve this problem, we consider splitting the original ultrahigh dimensional vector X_2 into two parts. The first part contains a subvector of X_2, which has a fixed dimension and is strongly correlated with Y. In our case this corresponds to those keywords in judgments that are not only high in frequency but also highly correlated with the response variable Y. This is merged with X_1 to form a new feature vector X, which has a fixed dimension and a strong correlation with Y. Therefore, a standard regression model can be used to describe the regression relationship for Y and X. The remaining part of X_2 is formed as another new vector, Z, which has a very high dimension and is weakly correlated with the response variable Y. Since it is related to the response (even weakly), it does carry useful information for predicting it. However, since it is of ultrahigh dimension and is weakly related to the response, it can hardly be directly incorporated into a usual regression model structure. Then how to effectively model the regression relationship between Y and (X, Z) with a fixed dimensional X and an ultrahigh dimensional Z becomes a key problem.

To solve this problem, we develop here a novel mixture conditional regression approach. Our method contains two important components. The first component is a mixture model. We assume that all the samples (i.e., legal cases) can be grouped into different classes. Within each group a standard linear regression model can be assumed for Y and X. We allow the intercepts of those regression models to be class-specific so that interclass heterogeneity can be modeled. We force the regression coefficients of the main covariates to be the same across classes so that the overall main covariate effect can be quantified. As one can see, the main regression model assumed between Y and X for every class is low-dimensional. This makes the subsequent parameter estimation and statistical inference very easy. However, the main challenge here is that the class membership for every sample is a latent variable that is not directly observed. For our cases the class membership is mainly determined by the judgments, which are represented by an ultrahigh dimensional binary vector. Therefore, it is theoretically appealing to assume for Z a naïve Bayes type mixture model so that the rich information contained in Z can be fully utilized to identify the latent membership for each sample.

To summarize, our model allows a feature vector to affect the response by two different mechanisms, as follows. The first mechanism is simply including a feature as an usual explanatory variable. The strength of this mechanism is that it can provide the best explanatory power from X to Y in a very direct way. The weakness of this approach is that the dimension of X cannot be too large. Otherwise, we would suffer from the curse of dimensionality. The second mechanism is to relate a feature Z with the response but indirectly through the latent class membership. The weakness of this approach is that the feature cannot affect the response directly. Therefore, the explanatory power is sacrificed to some extent. However, the strength of this approach is that it can easily accommodate as many features as possible. This leads to an interesting phenomenon, that is, the blessing of dimensionality. Simply speaking,

each mechanism has its own strength and weakness theoretically. Therefore, they need to be treated differently in practice.

For convenience, we refer to our model as a mixture conditional regression (MCR) model. To estimate it, we develop here a novel estimation method. It contains four steps. We show theoretically that the resulting estimators can be statistically as efficient as the oracle estimators, which are obtained by assuming that the latent class membership is known in advance. Extensive simulation studies are presented to demonstrate the finite sample performance of this method. A real data example of 6118 judgments is analyzed for illustration purposes. The rest of this article is organized as follows. Section 2 develops the MCR model, including the four-step estimators and their asymptotic statistical properties. Simulation studies are presented in Section 3 and a real data example of judgments is analyzed in Section 4. Finally, Section 5 concludes with a brief discussion. All technical details are relegated to the Supplementary Material (Shi et al. (2024)).

2. Methodology.

2.1. *A statistical model.* Let (Y_i, X_i) with $1 \le i \le n$ be the observation collected from the ith subject. Here $Y_i \in \mathbb{R}^1$ is the response of interest. In our case it is the log-transformed sentence length, and it is assumed to follow a continuous distribution. In the meanwhile, $X_i = (X_{i1}, \ldots, X_{iq})^\top \in \mathbb{R}^q$ is the associated main covariates. To model their regression relationship, we assume that

$$Y_i = \sum_{k=1}^{K} I(\mathcal{K}_i = k)\gamma_k + X_i^\top \theta + \varepsilon_i, \tag{2.1}$$

where $I(\cdot)$ is an indicator function, $\mathcal{K}_i \in \{1, 2, \ldots, K\}$ is a latent categorical variable identifying the latent class membership of the ith legal document, γ_k is an associated unknown coefficient, $\theta = (\theta_1, \ldots, \theta_q)^\top \in \mathbb{R}^q$ is the regression coefficient associated with confounding factors, and ε_i is random noise, which we assume to follow a normal distribution with mean 0 and variance σ^2. It is noteworthy that we allow different intercepts γ_ks according to classes so that the interclass heterogeneity can be modeled. In the meanwhile we assume the same coefficient θ for different classes so that the overall main covariate effect can be quantified.

We next consider how to model the dependence relationship between the latent class membership and the ultrahigh dimensional binary feature vector $Z_i = (Z_{i1}, \ldots, Z_{ip})^\top \in \mathbb{R}^p$ with $Z_{ij} \in \{0, 1\}$. Specifically, Z_{ij} is defined to be 1 if a prespecified keyword appears in the ith legal document and otherwise is 0. Note that Z_i is p-dimensional with a diverging p. In this work we might have $p > n$, and thus we allow more parameters than the observations. To model the dependence relationship between $\mathcal{K}_i$ and Z_i, a classical naïve Bayes model (Minnier et al. (2015), Rish et al. (2001), Spiegelhalter and Knill-Jones (1984), Yang (2018)) is assumed. Specifically, we assume that Z_i and (Y_i, X_i) are conditionally independent with $\mathcal{K}_i$ given. We also assume that Z_{ij}s for $1 \le j \le p$ are mutually conditionally independent, given $\mathcal{K}_i$. Mathematically, this jointly means that

$$P(Z_i|\mathcal{K}_i = k, Y_i, X_i) = P(Z_i|\mathcal{K}_i = k) = \prod_{j=1}^{p} P(Z_{ij}|\mathcal{K}_i = k) = \prod_{j=1}^{p} p_{kj}^{Z_{ij}}(1 - p_{kj})^{1-Z_{ij}},$$

where $p_{kj} = P(Z_{ij} = 1|\mathcal{K}_i = k)$. Finally, we assume that the class distribution probability $P(\mathcal{K}_i = k) = \pi_k$. Recall that ε_i follows a normal distribution with mean 0 and variance σ^2. Write $\phi(\gamma_k + X_i^\top\theta, \sigma^2) = (2\pi\sigma^2)^{-1}\exp\{-(Y_i - \gamma_k - X_i^\top\theta)^2/(2\sigma^2)\}$. Then a log-likelihood function can be written as

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k \phi\big(\gamma_k + X_i^\top\theta, \sigma^2\big) \prod_{j=1}^{p} p_{kj}^{Z_{ij}}(1 - p_{kj})^{1-Z_{ij}}\right\}, \tag{2.2}$$

where $\Theta = (\Omega^\top, \mathrm{vec}(P)^\top)^\top \in \mathbb{R}^{2K+q+1+pK}$, $\Omega = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2)^\top \in \mathbb{R}^{2K+q+1}$ and $P = (p_{kj}) \in \mathbb{R}^{K\times p}$. Here $\pi = (\pi_1, \ldots, \pi_K)^\top \in \mathbb{R}^K$ with $\sum_{k=1}^K \pi_k = 1$, $\gamma = (\gamma_1, \ldots, \gamma_K)^\top \in \mathbb{R}^K$, $\theta \in \mathbb{R}^q$, and $\sigma^2 \in \mathbb{R}^1$. For an arbitrary matrix A with dimension $M \times N$, $\mathrm{vec}(A)$ stands for an $MN \times 1$ column vector defined by stacking the columns of the matrix A on top of one another.

2.2. *Identifiability, interestingness, and relevance.* To estimate the model (2.2), it is crucial to ensure its identifiability. As one can see, if the response Y_i is integrated out, we are then left with binary feature Z_i and the covariate X_i only. In this case the identifiability becomes a serious issue. Nonidentifiable examples can be easily constructed. Therefore, for the model identification purpose, we cannot integrate Y_i out. Instead, we need to make full use of Y_i for model identification. To fix this idea, consider, for example, a highly simplified case with Y_i observed only (even without the binary feature vector Z_i); then the model (2.2) reduces to a standard mixture regression model. As pointed out by Grün and Leisch (2008), this model can be nicely identified under appropriate regularity conditions; see Section 3 in Grün and Leisch (2008). This discussion suggests that, even with the information Y_i only, we are able to identify the latent class membership in a probabilistic way. By supplying Y_i with additional information from Z_i, the identifiability can be further improved. Once the latent class membership is identified, the parameters $p_{kj}s$ associated Z_i can be estmated. This explains why when we develop our initial estimator for Ω in the next subsection, the response Y_i must be always involved.

As one can see, the statistical model (2.2) developed in Section 2.1 is a natural extension of the classical mixture linear regression model (De Veaux (1989), Wedel et al. (2000)). We modify this model slightly so that a large number of binary features can be included for a more accurate identification of the latent class membership. As one can see, the key assumption utilized here is the conditional independence structure imposed on $P(Z_i|\mathcal{K}_i = k)$ as $P(Z_i|\mathcal{K}_i = k) = \prod_{j=1}^p P(Z_{ij}|\mathcal{K}_i = k)$. Intuitively, this conditional independence assumption can be viewed as an implicit type of regularization, since it enforces a greatly simplified model structure for $P(Z_i|\mathcal{K}_i = k)$, which is a parsimonious approximation to reality. This natural extension is theoretically interesting due to the following reasons. First, this is an extension of the classical model from fixed-dimensional data to high-dimensional ones. Second, it allows us to extend the application of the classical mixture linear regression from structured data to unstructured text data, which are represented by a ultrahigh dimensional binary feature vector. Lastly, while the mainstream of the statistical literature complains about the curse of dimensionality, our model setup makes the high dimensionality a blessing. That is, higher feature dimension leads to more accurate identification of the latent class membership.

We then apply our methodology to the study of judicial impartiality. This is a problem of fundamental importance for empirical legal studies. In this regard a lot of statistical methods have been developed (Bushway and Piehl (2001), Glynn and Sen (2015), Peng and Cheng (2022), Steffensmeier and Kramer (1998)). The key feature of all those methods is to quantify the effect of the primary covariate of interest (e.g., ethnic, gender, age), after controlling for the confounding effects of legal factors. It is remarkable that most traditional statistical methods cannot handle ultrahigh dimensional data. Therefore, only a fixed number of legal factors can be included for controlling their confounding effects. That leaves ample amount of information contained in the legal documents in text format completely ignored. On the other side, this part of information is extremely useful for controlling the confounding effects of legal factors. Then how to solve this problem becomes practically important or even emergent. That inspires our methodology.

2.3. *An initial estimator.* We next consider how to practically estimate the model (2.1), which has a rather sophisticated structure and a large number of unknown parameters. It can hardly be optimized in a straightforward way by for example a standard Newton–Raphson algorithm. Thus, directly optimizing the joint log-likelihood function (2.2) might be practically extremely challenging or even infeasible. To solve this problem, we develop here an interesting estimation method, which starts with an initial estimator for the linear regression part and then progresses to a more sophisticated and also accurate final estimator. We are to show that this is a computationally more feasible solution with guaranteed statistical property. Specifically, if we focus on observations $\{(Y_i, X_i)\}_{i=1}^n$ only, we should have a log-likelihood function given by

$$\mathcal{L}(\Omega) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k \phi\big(\gamma_k + X_i^\top \theta, \sigma^2\big)\right\}, \tag{2.3}$$

where $\Omega = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2)^\top \in \mathbb{R}^{2K+q+1}$ is a finite-dimensional parameter and $\mathcal{L}(\cdot)$ is different from (2.2) with a slight abuse of notation. Therefore, an estimator for Ω can be defined as $\widehat{\Omega} = (\widehat{\pi}^\top, \widehat{\gamma}^\top, \widehat{\theta}^\top, \widehat{\sigma}^2)^\top = \operatorname{argmax}_\Omega \mathcal{L}(\Omega)$. To compute this estimator, we wish to obtain its first-order conditions. Note that π_k must to be optimized under the constraint $\sum_{k=1}^K \pi_k = 1$. We conduct the classical method of Lagrange multipliers (Breusch and Pagan (1980), Engle (1984)). The corresponding Lagrange function is given by $\mathcal{L}(\Omega, \lambda) = \mathcal{L}(\Omega) + \lambda(\sum_{k=1}^K \pi_k - 1)$, where λ is the Lagrange multiplier. Define $\dot{\mathcal{L}}(\Omega, \lambda) = (\dot{\mathcal{L}}_\Omega^\top(\Omega, \lambda), \dot{\mathcal{L}}_\lambda(\Omega, \lambda))^\top \in \mathbb{R}^{2K+q+2}$ be the first-order derivative of $\mathcal{L}(\Omega, \lambda)$ with respect to Ω and λ, where $\dot{\mathcal{L}}_\Omega(\Omega, \lambda) = \partial \mathcal{L}(\Omega, \lambda)/\partial \Omega \in \mathbb{R}^{2k+q+1}$ and $\dot{\mathcal{L}}_\lambda(\Omega, \lambda) = \partial \mathcal{L}(\Omega, \lambda)/\partial \lambda \in \mathbb{R}^1$. Subsequently, the first-order conditions are given by $\dot{\mathcal{L}}(\widehat{\Omega}, \widehat{\lambda}) = 0$. This leads to the estimation equations for Ω as

$$\begin{aligned}
\pi_k &= \mathcal{F}_{\pi_k}(\Omega) = \left(\sum_{i=1}^{n} \omega_{ik}\right)/n, \\
\gamma_k &= \mathcal{F}_{\gamma_k}(\Omega) = \left\{\sum_{i=1}^{n} (Y_i - X_i^\top \theta)\omega_{ik}\right\} \Big/ \left(\sum_{i=1}^{n} \omega_{ik}\right), \\
\theta &= \mathcal{F}_\theta(\Omega) = \big(X^\top X\big)^{-1} X^\top (Y - V), \\
\sigma^2 &= \mathcal{F}_{\sigma^2}(\Omega) = \sum_{i=1}^{n}\sum_{k=1}^{K} \big(Y_i - \gamma_k - X_i^\top \theta\big)^2 \omega_{ik}/n,
\end{aligned} \tag{2.4}$$

where $X = (X_1^\top, \ldots, X_n^\top)^\top \in \mathbb{R}^{n\times q}$, $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $V = (V_1, \ldots, V_n)^\top \in \mathbb{R}^n$, $V_i = \sum_{k=1}^K \omega_{ik}\gamma_k$, and $\omega_{ik} = P(\mathcal{K}_i = k | Y_i, X_i)$ is given by

$$\omega_{ik} = \mathcal{F}_{\omega_{ik}}(\Omega) = \frac{\pi_k \exp\{-\frac{1}{2\sigma^2}(Y_i - \gamma_k - X_i^\top \theta)^2\}}{\sum_{d=1}^K \pi_d \exp\{-\frac{1}{2\sigma^2}(Y_i - \gamma_d - X_i^\top \theta)^2\}}, \tag{2.5}$$

which is the mixture probability of the ith observation being the kth class.

Next, we turn this set of estimation equations (2.4) into a classical expectation-maximization (EM) type of algorithm (Bilmes et al. (1998), Dempster, Laird and Rubin (1977), Reynolds (2009)). Specifically, let $\widehat{\Omega}^{(0)} = (\widehat{\pi}^{(0)\top}, \widehat{\gamma}^{(0)\top}, \widehat{\theta}^{(0)\top}, \widehat{\sigma}^{2(0)})^\top$ be an arbitrarily specified initial estimator. For example, we can set $\widehat{\pi}_k^{(0)} = 1/K$, $\widehat{\gamma}^{(0)} = 0$, $\widehat{\theta}^{(0)} = 0$, and $\widehat{\sigma}^{2(0)} = 1$. Write $\widehat{\Omega}^{(t)} = (\widehat{\pi}^{(t)\top}, \widehat{\gamma}^{(t)\top}, \widehat{\theta}^{(t)\top}, \widehat{\sigma}^{2(t)})^\top$ as the estimator obtained in the tth step. Next, by (2.4) we can update $\widehat{\Omega}^{(t)}$ to be $\widehat{\Omega}^{(t+1)}$, as $\widehat{\pi}_k^{(t+1)} = \mathcal{F}_{\pi_k}(\widehat{\Omega}^{(t)})$, $\widehat{\gamma}_k^{(t+1)} = \mathcal{F}_{\gamma_k}(\widehat{\Omega}^{(t)})$, $\widehat{\theta}^{(t+1)} = \mathcal{F}_\theta(\widehat{\Omega}^{(t)})$, and $\widehat{\sigma}^{2(t+1)} = \mathcal{F}_{\sigma^2}(\widehat{\Omega}^{(t)})$. Here $V^{(t)} = (V_1^{(t)}, \ldots, V_n^{(t)})^\top$, $V_i^{(t)} = \sum_{k=1}^K \widehat{\omega}_{ik}^{(t)} \widehat{\gamma}_k^{(t)}$, and $\widehat{\omega}_{ik}^{(t+1)} = \mathcal{F}_{\omega_{ik}}(\widehat{\Omega}^{(t)})$.

2.4. *Estimating the response probability.* We next consider how to estimate the response probability p_{kj} for every k and j. To this end, we consider the joint log-likelihood function for (Y_i, X_i) and Z_{ij} as

$$\mathcal{L}^{(j)}(\Omega, p_j) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k \phi(\gamma_k + X_i^\top \theta, \sigma^2) p_{kj}^{Z_{ij}} (1 - p_{kj})^{1-Z_{ij}}\right\}, \tag{2.6}$$

where $p_j = (p_{1j}, \ldots, p_{Kj})^\top \in \mathbb{R}^K$ and the superscript "j" means that the log-likelihood is related to Z_{ij}. Ideally, we should optimize $\mathcal{L}^{(j)}(\Omega, p_j)$ with respect to all unknown parameters (i.e., Ω and p_j). It can be directly optimized by for example a Newton–Raphson type iterative algorithm. The associate computational cost should be practically very acceptable for a fixed j. However, if the feature dimension p is large, the total computational cost becomes much heavier. We are then inspired to search for computationally more efficient alternative. In fact, with the help of the initial estimator $\widehat{\Omega} = (\widehat{\pi}^\top, \widehat{\gamma}^\top, \widehat{\theta}^\top, \widehat{\sigma}^2)^\top$, given in the previous subsection, we can simply replace $\Omega = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2)^\top$ by its initial estimator $\widehat{\Omega}$. This leads to a simplified log-likelihood function as $\mathcal{L}^{(j)}(\widehat{\Omega}, p_j)$. Therefore, an estimator for p_j can be defined as $\widehat{p}_j = (\widehat{p}_{1j}, \ldots, \widehat{p}_{Kj})^\top = \operatorname{argmax}_p \mathcal{L}^{(j)}(\widehat{\Omega}, p)$. Similar to the previous subsection, we can obtain the estimation equation for p_{kj} as

$$p_{kj} = \mathcal{G}_{p_{kj}}(\widehat{\Omega}, p_j) = \left\{\sum_{i=1}^{n} \widehat{\pi}_{ik}^{(j)} Z_{ij}\right\} \Big/ \left\{\sum_{i=1}^{n} \widehat{\pi}_{ik}^{(j)}\right\} \tag{2.7}$$

with

$$\widehat{\pi}_{ik}^{(j)} = \mathcal{G}_{\pi_{ik}^{(j)}}(\widehat{\Omega}, p_j) = \frac{\widehat{\pi}_k \exp\{-\frac{1}{2\widehat{\sigma}^2}(Y_i - \widehat{\gamma}_k - X_i^\top \widehat{\theta})^2\} p_{kj}^{Z_{ij}} (1 - p_{kj})^{1-Z_{ij}}}{\sum_{d=1}^{K} \widehat{\pi}_d \exp\{-\frac{1}{2\widehat{\sigma}^2}(Y_i - \widehat{\gamma}_d - X_i^\top \widehat{\theta})^2\} p_{dj}^{Z_{ij}} (1 - p_{dj})^{1-Z_{ij}}}.$$

Then we can immediately obtain the EM algorithm for p_{kj} as follows. Specifically, let $\widehat{p}_{kj}^{(0)}$ be an arbitrarily specified initial estimator, such as $\widehat{p}_{kj}^{(0)} = 1/2$ for every k and j. Write $\widehat{p}_{kj}^{(t)}$ be the estimator obtained in the tth step. We can then compute $\widehat{p}_{kj}^{(t+1)}$ is $\widehat{p}_{kj}^{(t+1)} = \mathcal{G}_{p_{kj}}(\widehat{\Omega}, \widehat{p}_j^{(t)})$ with $\widehat{\pi}_{ik}^{(j,t+1)} = \mathcal{G}_{\pi_{ik}^{(j)}}(\widehat{\Omega}, \widehat{p}_j^{(t)})$ and $\widehat{p}_j^{(t)} = (\widehat{p}_{1j}^{(t)}, \ldots, \widehat{p}_{Kj}^{(t)})^\top \in \mathbb{R}^K$. By doing so, we can make full use of this analytical formula (2.7) so that the algorithm is no longer iterative beween Ω and p_j. Consequently, the computational cost can be significantly reduced.

Note that the initial estimator $\widehat{\Omega}$ is the standard M-estimator, which we know is $\sqrt{n}$-consistent under appropriate regularity conditions (Shao (2003), van der Vaart (1998)). Then we can further prove that $\widehat{p}_j = (\widehat{p}_{1j}, \ldots, \widehat{p}_{Kj})^\top \in \mathbb{R}^K$ is also $\sqrt{n}$-consistent for every j. To study the theoretical properties of $\widehat{p}_j$ over every j, we shall focus on the log-likelihood function $\mathcal{L}^{(j)}(\widehat{\Omega}, p_j)$. The technical details of the derivatives of $\mathcal{L}^{(j)}(\widehat{\Omega}, p_j)$ are given in Section 2.1 of the Supplementary Material (Shi et al. (2024)). Let $\|a\| = (\sum_{j=1}^{p} a_j^2)^{1/2}$ denote the ℓ_2 norm for an arbitrary vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$. Write $\pi_{\min} = \min_k \pi_k$ and $\pi_{\max} = \max_k \pi_k$ with $0 < \pi_{\min} \le \pi_{\max} < 1$. Write $\gamma_{\max}^2 = \max_k \gamma_k^2 > 0$. Then the uniform consistency for $\widehat{p}_j$ over every j can be established by Theorem 1, which is proved in Section 2.1 of the Supplementary Material (Shi et al. (2024)). By Theorem 1 we known that $\widehat{p}_j$ is uniformly consistent for p_j over $1 \le j \le p$. The uniform convergence rate is slightly slower than the standard rate of $\sqrt{n}$ by a slowly diverging factor C_n.

THEOREM 1. *Assume the technical conditions* (C1), (C2), *and* (C3), *as given in Section* 1 *of the Supplementary Material* (*Shi et al.* (2024)) *hold. Furthermore, assume that* $C_n > 0$ *is an arbitrary positive sequence such that*: (*I*) $C_n/\sqrt{n} \to 0$, *and* (*II*) $C_n^2/\log p \to \infty$ *as* $n \to \infty$. *We then have* $\max_j \|\widehat{p}_j - p_j\| = O_p(C_n/\sqrt{n})$.

2.5. *Class membership identification*. With the help of the response probability estimators, we are able to estimate the latent class membership with extraordinarily high accuracy. This is mainly because the feature dimension p is extremely high. That leads to an ample amount of information for class membership identification. Specifically, we are still interested in estimating the mixture probability for $\mathcal{K}_i = k$ but with (Y_i, X_i, Z_i) information given. Write $a_{ik} = I(\mathcal{K}_i = k)$ and $\pi_{ik} = P(\mathcal{K}_i = k|Y_i, X_i, Z_i)$. Then direct computation suggests that $\pi_{ik} = \mathcal{W}_{ik}(\Theta)$ with

$$\mathcal{W}_{ik}(\Theta) = \frac{\pi_k \exp\{-\frac{1}{2\sigma^2}(Y_i - \gamma_k - X_i^\top \theta)^2\} \prod_{j=1}^p p_{kj}^{Z_{ij}} (1 - p_{kj})^{1-Z_{ij}}}{\sum_{d=1}^K \pi_d \exp\{-\frac{1}{2\sigma^2}(Y_i - \gamma_d - X_i^\top \theta)^2\} \prod_{j=1}^p p_{dj}^{Z_{ij}} (1 - p_{dj})^{1-Z_{ij}}}. \tag{2.8}$$

Next, we can replace the unknown parameters in (2.8) with the estimators given in Sections 2.3 and 2.4. This leads to an another estimator for the mixture probability for each observation i as given by $\widehat{\pi}_{ik} = \mathcal{W}_{ik}(\widehat{\Theta})$.

Recall that the same mixture probability was also evaluated in Section 2.3 but with (Y_i, X_i) information only, which is denoted by ω_{ik}. Note that the feature involved in ω_{ik} has a fixed dimension, and thus there is very limited information. Therefore, the mixture probability estimator ω_{ik}, provided in (2.4), cannot be consistent for a_{ik}. In other words, the difference between ω_{ik} and a_{ik} will not converge to 0, even if $n \to \infty$. However, the story dramatically changes for the new mixture probability estimator $\widehat{\pi}_{ik}$ for which we assume that $p \to \infty$ as $n \to \infty$. Consequently, a sufficient amount of information can be accumulated for the latent class membership $\mathcal{K}_i$. As a direct consequence, it is more likely to be consistent for a_{ik}. In fact, this conjecture is formally verified by Theorem 2, whose proof is given in Section 2.2 of the Supplementary Material (Shi et al. (2024)). From this we know that the mixture probability estimator $\widehat{\pi}_{ik}$ is extremely close to the true membership indicator function $a_{ik} = I(\mathcal{K}_i = k)$, with a tiny error of order $O(\exp(-\nu p))$. This makes the subsequent estimator and inference for the main regression model of Y and X very easy.

THEOREM 2. *Assume the technical conditions* (C1)–(C6), *as given in Section* 1 *of the Supplementary Material* (*Shi et al.* (2024)) *hold. Then for an arbitrary constant* $0 < \nu < \Delta_{\min}/2$, *where* $\Delta_{\min}$ *is defined in Condition* (C5), *there exists a positive constant* $M > 0$ *such that* $P\{\max_{i,k} |\widehat{\pi}_{ik} - I(\mathcal{K}_i = k)| > \exp(-\nu p)\} = o(1)$ *as long as* $p > M$.

2.6. *The final main estimator.* By Theorem 2, we know that the true class membership can be consistently estimated by $\widehat{\pi}_{ik}$ with super excellent accuracy. As a consequence, we should be able to reestimate the interested regression parameter $\Omega = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2)^\top \in \mathbb{R}^{2K+q+1}$ with much-improved estimation accuracy. Define $X_i^a = (a_i^\top, X_i^\top)^\top \in \mathbb{R}^{K+q}$, where $a_i = (a_{i1}, \ldots, a_{iK})^\top \in \mathbb{R}^K$. Recall that $\sum_{k=1}^K a_{ik} = 1$. Specifically, if the latent class membership $\mathcal{K}_i$ is known in advance, we should estimate the parameters of the linear regression model by minimizing the following classical least squares (OLS) objective function as

$$\mathcal{L}(\Phi) = \sum_{i=1}^n \sum_{k=1}^K a_{ik}(Y_i - \gamma_k - X_i^\top \theta)^2 = \sum_{i=1}^n (Y_i - X_i^{a\top} \Phi)^2, \tag{2.9}$$

where $\Phi = (\gamma^\top, \theta^\top)^\top \in \mathbb{R}^{K+q}$ and $\mathcal{L}(\cdot)$ is different from (2.2) and (2.3) with a slight abuse of notation. By optimizing $\mathcal{L}(\Phi)$ with respect to Φ, we obtain $\widehat{\Phi}_{\text{oracle}} = \operatorname{argmin}_\Phi \mathcal{L}(\Phi) = (\widehat{\Sigma}_{\text{oracle}}^{\text{xx}})^{-1} \widehat{\Sigma}_{\text{oracle}}^{\text{xy}}$, where $\widehat{\Sigma}_{\text{oracle}}^{\text{xx}} = \sum_{i=1}^n X_i^a X_i^{a\top}/n$ and $\widehat{\Sigma}_{\text{oracle}}^{\text{xy}} = \sum_{i=1}^n X_i^a Y_i/n$.

As one can see, $\widehat{\Phi}_{\text{oracle}}$ is an oracle estimator, which cannot be practically computed, since the latent class membership $\mathcal{K}_i$ is not directly observed. However, by Theorem 2 we know that this binary indicator random variable $a_{ik} = I(\mathcal{K}_i = k)$ can be estimated consistently and accurately by $\widehat{\pi}_{ik}$. We are then motivated to approximate X_i^a by X_i^π, where $X_i^\pi =$

$(\widehat{\pi}_i^\top, X_i^\top)^\top \in \mathbb{R}^{K+q}$, and $\widehat{\pi}_i = (\widehat{\pi}_{i1}, \ldots, \widehat{\pi}_{iK})^\top \in \mathbb{R}^K$. Then a practically feasible estimator can be constructed as $\widehat{\Phi}_{\text{real}} = (\widehat{\Sigma}^{\text{xx}}_{\text{real}})^{-1}\widehat{\Sigma}^{\text{xy}}_{\text{real}}$, where $\widehat{\Sigma}^{\text{xx}}_{\text{real}} = \sum_{i=1}^n X_i^\pi X_i^{\pi\top}/n$ and $\widehat{\Sigma}^{\text{xy}}_{\text{real}} = \sum_{i=1}^n X_i^\pi Y_i/n$. Thereafter, an oracle estimator for σ^2 can be defined as $\widehat{\sigma}^2_{\text{oracle}} = \sum_{i=1}^n (Y_i - X_i^{a\top}\widehat{\Phi}_{\text{oracle}})^2/n$. Once $\widehat{\Phi}_{\text{real}}$ is obtained, $\widehat{\sigma}^2_{\text{oracle}}$ can be estimated as $\widehat{\sigma}^2_{\text{real}} = \sum_{i=1}^n (Y_i - X_i^{\pi\top}\widehat{\Phi}_{\text{real}})^2/n$. Lastly, we define the final estimator for π_k, as $\widehat{\pi}_{\text{real}} = (\widehat{\pi}_{\text{real},1}, \ldots, \widehat{\pi}_{\text{real},K})$, where $\widehat{\pi}_{\text{real},k} = (\sum_{i=1}^n \widehat{\pi}_{ik})/n$, and its ideal counterpart as $\widehat{\pi}_{\text{oracle}} = (\widehat{\pi}_{\text{oracle},1}, \ldots, \widehat{\pi}_{\text{oracle},K})$, where $\widehat{\pi}_{\text{oracle},k} = (\sum_{i=1}^n a_{ik})/n$. Recall that $\Omega = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2)^\top \in \mathbb{R}^{2K+q+1}$. Then we can immediately have the real estimator $\widehat{\Omega}_{\text{real}} = (\widehat{\pi}^\top_{\text{real}}, \widehat{\gamma}^\top_{\text{real}}, \widehat{\theta}^\top_{\text{real}}, \widehat{\sigma}^2_{\text{real}})^\top$ and the oracle estimator $\widehat{\Omega}_{\text{oracle}} = (\widehat{\pi}^\top_{\text{oracle}}, \widehat{\gamma}^\top_{\text{oracle}}, \widehat{\theta}^\top_{\text{oracle}}, \widehat{\sigma}^2_{\text{oracle}})^\top$. Theorem 3 characterizes the difference between the real and oracle estimator. From this we find that the resulting estimator enjoys the same convergence rate and asymptotic distribution as its ideal counterpart, which is defined by assuming that $\mathcal{K}_i$ is known in advance. The proof of Theorem 3 is given in Section 2.3 of the Supplementary Material (Shi et al. (2024)).

THEOREM 3. *Assume that the technical conditions* (C1)–(C6), *as given in Section* 1 *of the Supplementary Material* (*Shi et al.* (2024)) *hold. Then we have* $\|\widehat{\Omega}_{real} - \widehat{\Omega}_{oracle}\| = o_p(1/\sqrt{n})$.

3. Simulation studies.

3.1. *The simulation setup*. To demonstrate the finite sample performance of the proposed MCR method, we performed a number of simulation studies. Specifically, we would like to study the finite sample performance of: (a) the initial estimator $\widehat{\Omega}$, (b) the response probability estimators $\widehat{p}_j s$, (c) the mixture probability estimators $\widehat{\pi}_{ik} s$, and (d) the final main estimator $\widehat{\Omega}_{\text{real}}$. For the entire simulation study, we considered various sample sizes with $n = 1000$, 2000, or 5000. For each n the dimension of the binary feature vector was set to be $p = n/10$, $n/5$, $n/2$, n, or $2n$. Once n and p are given, following Tibshirani (1996), we generated $X_i \in \mathbb{R}^q$ with $q = 8$ from a multivariate normal distribution with mean 0 and $\text{cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ with $\rho = 0.5$ for $1 \le j_1, j_2 \le 8$. The number of classes was fixed at $K = 5$ (Vermunt and Magidson (2002)). The true value of $\Theta = (\pi^\top, \gamma^\top, \theta^\top, \sigma^2, \text{vec}(P)^\top)^\top \in \mathbb{R}^{2K+q+1+pK}$ was set at $\pi = (0.15, 0.2, 0.3, 0.25, 0.1)^\top$, $\gamma = (-4, -1, 2, 5, 8)^\top$, $\theta = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$, $\sigma^2 = 1$, and $P = (p_{kj}) \in \mathbb{R}^{K\times p}$. Here the matrix P was divided into K^2 block matrices, where the block diagonal elements were generated from a uniform distribution between 0.8 and 0.95, and other elements were generated from a uniform distribution between 0.01 and 0.3. Then $Z_i = (Z_{ij}) \in \mathbb{R}^p$ could be generated. The residual term ε_i was independently generated from a standard normal distribution. This leads to the final response variable Y_i according to the model (2.1).

3.2. *The initial estimator* $\widehat{\Omega}$. We start with the initial estimator $\widehat{\Omega} = (\widehat{\pi}^\top, \widehat{\gamma}^\top, \widehat{\theta}^\top, \widehat{\sigma}^2)^\top \in \mathbb{R}^{2K+q+1}$. For a given sample size n and binary feature dimension p, the experiment was randomly replicated for a total of $R = 500$ times. We use $\widehat{\tau}^{(r)}$ to represent one particular estimator (e.g., $\widehat{\theta}^{(r)}$) obtained in the rth replication ($1 \le r \le R$). The true parameter is denoted by τ. Then the estimator error (Err) can be evaluated as $\text{Err} = \|\widehat{\tau}^{(r)} - \tau\|$ for every $1 \le r \le R$. This leads to a total of R Err values, which are then log-transformed and boxplotted in Figure 1. By Figure 1 we find that, for essentially every estimator of interest (i.e., $\widehat{\pi}$, $\widehat{\gamma}$, $\widehat{\theta}$, and $\widehat{\sigma}^2$), large sample sizes always lead to smaller estimation errors. This numerical finding confirms that the initial estimator $\widehat{\Omega}$ is indeed consistent. Comparatively speaking, the estimation errors for $\widehat{\gamma}$ and $\widehat{\theta}$ are much larger than those of the others. Take $n = 5000$ as an example. The median log(Err) values of $\widehat{\gamma}$ and $\widehat{\theta}$ are given by -2.2034 and -2.3362, respectively. In contrast, those of $\widehat{\pi}$ and $\widehat{\sigma}^2$ are given by -4.2469 and -3.7232, respectively.

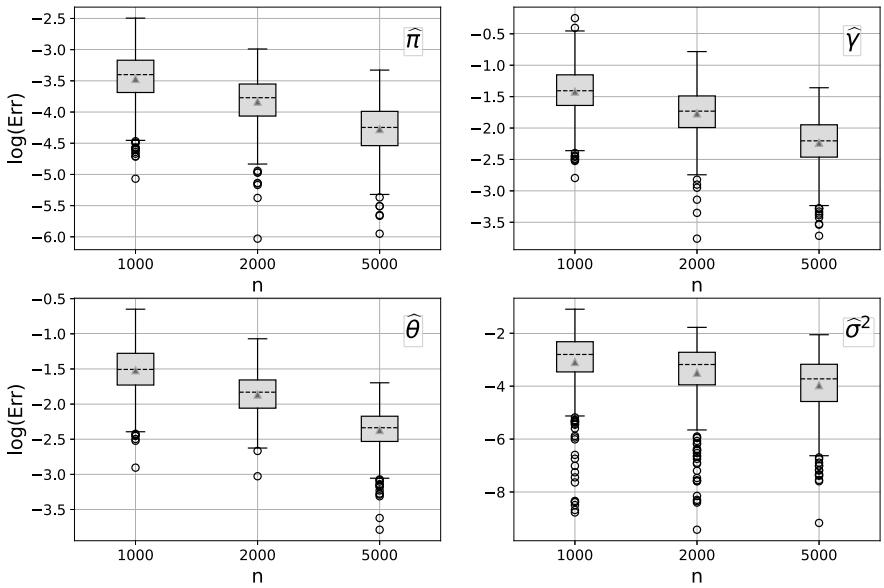

FIG. 1. *Log(Err) values for various initial estimators*: $\widehat{\pi}$ (*the upper-left panel*), $\widehat{\gamma}$ (*the upper-right panel*), $\widehat{\theta}$ (*the lower-left panel*), *and* $\widehat{\sigma}^2$ (*the lower-right panel*). *Three different sample sizes are considered. They are* $n = 1000$, 2000, *and* 5000, *respectively.*

The absolute difference are given by 2.0435, 1.9107, 1.5198, and 1.387, respectively. Considering that those are differencing in log-scale, we should conclude that the log(Err) values of $\widehat{\gamma}$ and $\widehat{\theta}$ are relatively large, compared to those of $\widehat{\pi}$ and $\widehat{\sigma}^2$.

3.3. *The response probability estimator* $\widehat{p}_j$. Next, we study $\widehat{p}_j \in \mathbb{R}^K$. Similarly, we can compute for each $\widehat{p}_j$ an Err value decoded by Err_j $(1 \le j \le p)$, and its maximum error (MaxErr) over j is given by $\text{MaxErr} = \max_j \text{Err}_j$. Recall that we have $R = 500$ random replications. This leads to a total of R MaxErr values, which are then log-transformed and box-plotted in Figure 2. By Figure 2 we obtain the following two interesting findings. First, for a fixed p, we find that the larger the sample size n, the smaller the maximum error (MaxErr). This confirms that $\widehat{p}_j$ is uniformly consistent for p_j over $1 \le j \le p$. Second, with a fixed sample size n, the maximum error (MaxErr) seems to be slightly larger as p increases. This interesting numerical finding suggests that the uniform convergence rate of $\widehat{p}_j$ is slightly

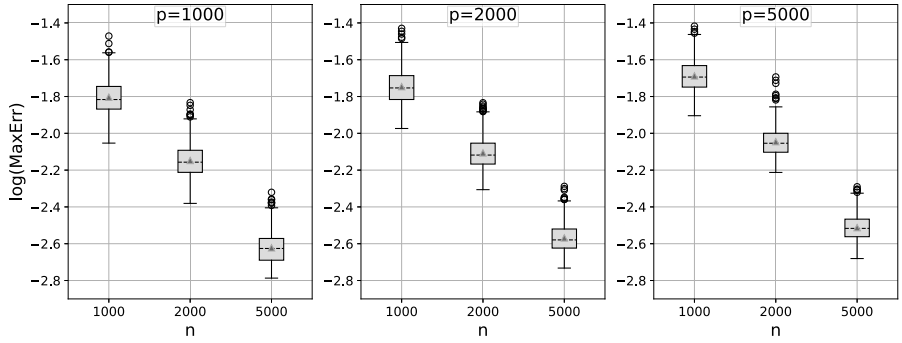

FIG. 2. *Log(MaxErr) values for the response probability estimator* $\widehat{p}_j$. *Different panels correspond to different feature dimensions*: $p = 1000$ (*the left panel*), 2000 (*the middle panel*), *and* 5000 (*the right panel*). *For a given panel*, *different boxplots correspond to different sample sizes with* $n = 1000$, 2000, *and* 5000, *respectively.*

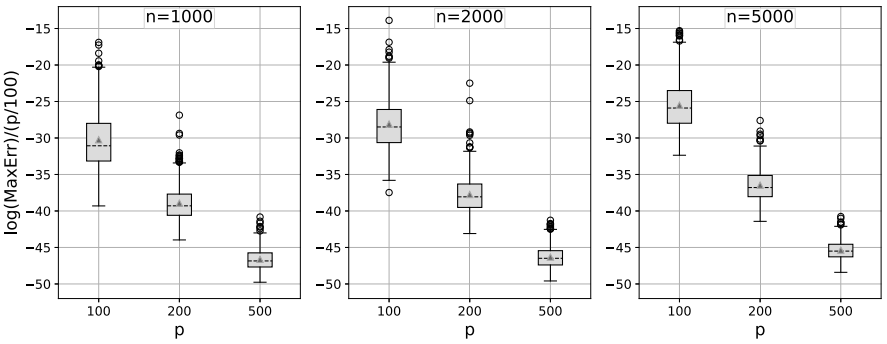

FIG. 3. *Log(MaxErr)/(p/100) values for the mixture probability estimator $\widehat{\pi}_{ik}$. Different panels correspond to different sample sizes: $n = 1000$ (the left panel), 2000 (the middle panel), and 5000 (the right panel). For a given panel, different boxplots correspond to different feature dimensions with $p = 100$, 200, and 500, respectively.*

slower than the standard rate of $\sqrt{n}$ if $p \to \infty$ as $n \to \infty$. All these results are in line with our theoretical findings in Theorem 1.

3.4. *The estimated mixture probability $\widehat{\pi}_{ik}$.* We then study $\widehat{\pi}_{ik}$. Recall that $\widehat{\pi}_{ik}$ estimates the latent class membership. We are extremely interested in evaluating the difference between $\widehat{\pi}_{ik}$ and the true membership indicator function $a_{ik} = I(\mathcal{K}_i = k)$. Thus, we can compute $\text{Err}_{i,k} = |\widehat{\pi}_{ik} - a_{ik}|$ for every i $(1 \le i \le n)$ and k $(1 \le k \le K)$, whose maximum error (MaxErr) over i and k is given by $\text{MaxErr} = \max_{i,k} \text{Err}_{i,k}$. Similarly, this leads to a total of R MaxErr values, which are then log-transformed and box-plotted in Figure 3. By Figure 3 we find that with a fixed sample size n, a larger p leads to smaller MaxErr values. The larger the p value is, the more feature information can be provided, and thus the more accurate the mixture probability could be. This results verify that the feature information helps us to estimate the latent class membership with extraordinarily high accuracy, which is in line with our theoretical findings in Theorem 2.

3.5. *The final main estimator $\widehat{\Omega}_{real}$.* Finally, recall that $\widehat{\Omega}_{\text{real}}$ is a practically feasible estimator to approximate the oracle estimator $\widehat{\Omega}_{\text{oracle}}$. Thus, we would like to study the difference between $\widehat{\Omega}_{\text{real}}$ and $\widehat{\Omega}_{\text{oracle}}$, as evaluated by $\text{Diff} = \|\widehat{\Omega}_{\text{real}} - \widehat{\Omega}_{\text{oracle}}\|$. Similarly, this leads to a total of R Diff values in log-scale, which are then box-plotted in Figure 4. By Figure 4 we find that, for a fixed small feature dimension p (e.g., $p = 100$), the larger the sample size n, the

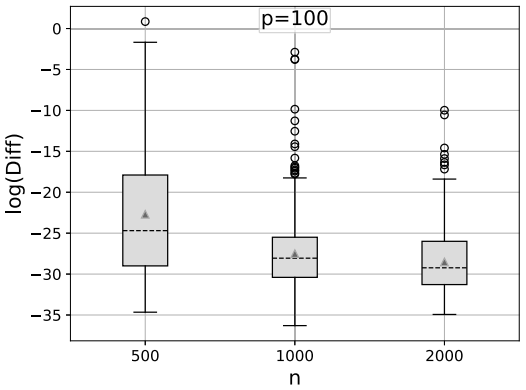

FIG. 4. *Log(Diff) values between $\widehat{\Omega}_{real}$ and $\widehat{\Omega}_{oracle}$ with the feature dimension $p = 100$. Different boxplots correspond to different sample sizes.*

smaller the mean error. Furthermore, the difference between $\widehat{\Omega}_{\text{real}}$ and $\widehat{\Omega}_{\text{oracle}}$ rapidly shrinks to an extremely tiny value as n increases. In fact, for a slightly large p (e.g., $p = 200$) and a reasonably large sample size (e.g., $n = 500$), the Diff values are too tiny to be distinguished from 0 due to the limited precision of a computer system. This indicates that $\widehat{\Omega}_{\text{real}}$ is almost identical to $\widehat{\Omega}_{\text{oracle}}$. All these results are in line with our theoretical findings in Theorem 3.

3.6. *A BIC method for K.* The simulation results presented in the previous subsections are based on the assumption that the true number of latent classes (i.e., K) is known in advance. Unfortunately, this is an unknown parameter that need be estimated. To this end, we follow the idea of Schwarz (1978) and develop here a BIC method. Specifically, let $K_{\max}$ be the maximum number of latent classes to be considered. For example, various $K_{\max}$ values (e.g., 10 and 20) have been considered. The resulting numerical performance is nearly identical. Therefore, we fix $K_{\max} = 10$ for the rest of the simulation study. Next, for any $1 \le K \le K_{\max}$, the interested model parameters can be estimated and denoted as $\widehat{\Theta}^{(K)} = (\widehat{\pi}^{(K)\top}, \widehat{\gamma}^{(K)\top}, \widehat{\theta}^{(K)\top}, \widehat{\sigma}^{2(K)\top}, \operatorname{vec}(\widehat{P}^{(K)}))^\top \in \mathbb{R}^{d^*}$ with $d^* = K + K + q + 1 + pK = 2K + q + 1 + pK$. Then a BIC selection criterion can be developed as

$$\begin{aligned}\text{BIC}(K) = -2\sum_{j=1}^{p}\Bigg[\sum_{i=1}^{n}\log\Bigg\{\sum_{k=1}^{K}\widehat{\pi}_k^{(K)}\phi\big(\widehat{\gamma}_k^{(K)} + X_i^\top\widehat{\theta}^{(K)}, \widehat{\sigma}^{2(K)}\big) \\ \times \big(\widehat{p}_{kj}^{(K)}\big)^{Z_{ij}}\big(1 - \widehat{p}_{kj}^{(K)}\big)^{1-Z_{ij}}\Bigg\}\Bigg] + \text{df} \times \log n,\end{aligned} \tag{3.1}$$

where $\text{df} = d^* - 1 = 2K + q + pK$ is the degree of freedom, due to the whole parameter $\Theta \in \mathbb{R}^{d^*}$ and $\sum_{k=1}^{K}\pi_k = 1$, and the penalization factor ($\text{df} \times \log n$) is due to the seminal work of Schwarz (1978). Therefore, the optimum K can be estimated as $\widehat{K} = \operatorname{argmin}_{1\le K\le K_{\max}}\text{BIC}(K)$. Following the simulation setting in the previous subsections, this experiment was randomly replicated $R = 500$ times. The percentage of the experiments with $\widehat{K} = K = 5$ is shown in Table 1. As one can see, for any fixed feature dimension p, the percentage of experiments with $\widehat{K} = K = 5$ converges to 100% rapidly as the sample size n increases. This suggests that $\widehat{K}$ should be a consistent estimator of K.

4. Real data analysis.

4.1. *The China judgments online data.* We present here a real case study. The dataset is obtained from *China Judgments Online* (CJO). The full dataset contains a total of 1,361,354 cases that happened in China from 2017 to 2018. For illustration purposes we study here the criminal cases only. This is mainly because the CJO dataset is a highly unbalanced dataset, with sample sizes varying considerably by crime. Obviously, we cannot work on crimes with extremely tiny sample sizes. In the meanwhile, past literature suggests that theft is one of the

TABLE 1
Percentage (%) of experiments with different n and p

	p					
n	10	50	100	200	500	1000
200	36.2%	28.2%	28.0%	25.0%	26.0%	25.4%
500	89.2%	85.0%	84.0%	83.8%	83.4%	84.4%
1000	98.0%	97.6%	98.0%	97.0%	97.2%	97.2%
2000	100%	99.6%	99.8%	99.8%	99.8%	99.8%

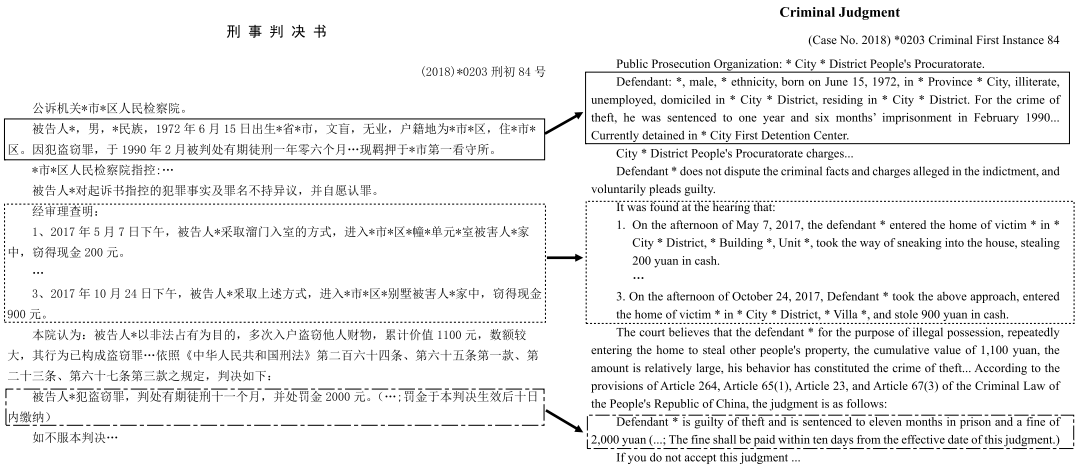

FIG. 5. *An arbitrarily selected judgment example* (*left*) *along with its English translation* (*right*). *The top solid box includes the defendant's demographic characteristics. The middle dotted box presents the court's findings of major facts. The bottom dot-dashed box shows the court's sentencing decisions.*

most common crimes worldwide (Felson and Boba (2010), Sheley and Ashkins (1981)). It happens that this is also the case for our CJO dataset, where theft accounts for about 24.44% of all cases (Simmons and Flood-Page (2002), Xu et al. (2022)). For illustration purposes we take burglaries in theft-related cases as an example. It accounts for about 13.59% of all theft cases. Moreover, to render our analysis in a more straightforward way, only those first trials and fixed-term imprisonment cases without any missing information are kept. This leads to a final sample size of $n = 6118$ cases.

For each case the CJO dataset collects a judgment document written in Chinese, including the defendant's demographic characteristics, the court's findings of major facts, and the court's sentencing decisions (Simmons and Flood-Page (2002)). Figure 5 shows an arbitrarily selected judgment example (left) along with its English translation (right). The defendant's demographic characteristics are typically included in the first paragraph of the judgement documents. It typically contains important information, such as age, gender, and ethnicity; see, for example, the top solid box in Figure 5. The court's findings of major facts are located in the judgment between "It was found at the hearing that" and "The court believes that;" see, for example, the middle dotted box in Figure 5. Lastly, the court's sentencing decisions are shown in the paragraph following 'The judgement is as follows;" see, for example, the bottom dot-dashed box in Figure 5.

4.2. *Variable description.* The primary variable of interest in our study is the length of the prison sentence, reported in months. We take the log-transformed length of the prison sentence as our response variable Y_i for every $1 \le i \le n$. We next consider a set of five extralegal factors (i.e., X_{1i}) available in our CJO dataset. These are mainly the demographic variables of the defendants, including age, gender, ethnicity, employment status, and education level. All variables are coded as dummies, except for age. Age of the defendant (X_{1i1}) is measured in years, with values ranging from 16 to 78 years. The mean age is about 33.7 years, with a standard deviation of 9.8 years. Gender (X_{1i2}) is coded as $X_{1i2} = 1$ for males and $X_{1i2} = 0$ for females. More than 90% (97.7%) of the defendants in our sample are males. For ethnicity (X_{1i3}), we represent Han Chinese by $X_{1i3} = 1$ and other minorities by $X_{1i3} = 0$. More than three-quarters (85.4%) of the defendants are Han Chinese. Employment status (X_{1i4}) is coded as $X_{1i4} = 1$ if the defendant is employed and $X_{1i4} = 0$ otherwise. About 38.4% of the defendants are employed. Lastly, the education level is coded as $X_{1i5} = 1$ if the defendant is in

elementary school (42.7%), $X_{1i6} = 1$ for junior middle school (40.6%), and $X_{1i7} = 1$ for high school or above (7.7%); illiterate defendants (9.0%) are coded as $X_{1i5} = X_{1i6} = X_{1i7} = 0$. This leads to the extralegal factor vector $X_{1i} \in \mathbb{R}^7$ for every $1 \le i \le n$.

Next, we extract keywords from criminal facts as legal factors, that is, control variable $X_{2i} = (X_{2ij})$ with $X_{2ij} \in \{0, 1\}$. To this end, we first cut the Chinese judgment documents into keywords and then compute their frequencies. For illustration purposes, only those keywords with a frequency of more than 10 times are kept. These account for about 4.8% of the total number of keywords but 95.99% of the total frequency. Among those keywords there are many keywords, which are very high in frequency but have little actual meaning. Those keywords are then excluded from our subsequent analysis. Those excluded keywords include, for example, "the defendant, the victim, the plaintiff," and "the Public Prosecution Service." This leads to a final set of 6578 comparatively more informative keywords, such as "knife, break-in," and "gold." We then code for each keyword as a dummy variable X_{2ij}, whose value is 1 if the jth keyword actually appears in the ith document and is 0 otherwise. Following the idea of the proposed MCR method, we then split X_{2i} into two parts. The first part contains a set of keywords that are not only high in frequency but also high in correlation with Y. In the meanwhile the size of the first part is determined by the BIC score in the standard linear regression, with a final size of 64. This subvector of X_{2i} is then merged with X_{1i} so that the main covariates $X_i \in \mathbb{R}^q$ with $q = 71$ can be formed. Then the rest of X_{2i} is formed as $Z_i \in \mathbb{R}^p$, with $p = 6514$ for every $1 \le i \le n$.

4.3. *The estimation results*. To apply the proposed MCR method, we first estimate the number of the latent classes (i.e., K), using the BIC selection criterion (3.1) proposed in Section 3.6. Specifically, we fix $K_{\max} = 20$. This leads to the final estimate $\widehat{K} = \operatorname{argmin}_{1 \le K \le K_{\max}} \mathrm{BIC}(K) = 7$. Subsequently, we fix $K = \widehat{K} = 7$ so that the main parameters of interest can be estimated. To gain some quick understanding about the practical meaning of those $K = 7$ classes, we compute the term frequency inverse document frequency (TF-IDF) of each keywords to determine the most representative keywords for each class (Qaiser and Ali (2018), Ramos et al. (2003), Wu et al. (2008)). We next summarize the top five keywords with the hightest TF-IDF scores for each class in Table 2. For each reported keyword, we also compute the percentage of the documents containing this keyword for each class. The results are reported in percentage right below the corresponding keyword. The sample sizes of each class are reported in the first column in Table 2. As one can see, class 1 mainly concerns about the cases of burglary with criminal records. Class 2 contains cases where the object of theft is tobacco or electronic devices, such as computers and iPhones. Class 3 comprises cases of burglary with a large amount. Class 4 primarily involves cases where the object of theft is gold, watches, or jewelry. Class 5 consists of cases where the

TABLE 2

The top five keywords with the highest TF-IDF scores for each class along with the class sample size and the keyword appearing percentage

Sample	Top 1	Top 2	Top 3	Top 4	Top 5
Class 1 (2809)	home (67%)	break in (24%)	record (19%)	window (10%)	climb (9%)
Class 2 (1265)	photo (79%)	cigarette (31%)	laptop (25%)	video (23%)	apple (17%)
Class 3 (807)	larger amount (63%)	possession (59%)	secrecy (31%)	indemnity (21%)	weapon (15%)
Class 4 (763)	gold (97%)	necklace (51%)	ring (46%)	bracelet (37%)	watch (32%)
Class 5 (288)	cash (92%)	phone (79%)	bedroom (76%)	gate (60%)	wardrobe (58%)
Class 6 (168)	liability (76%)	drive (49%)	force (46%)	motorcycle (35%)	joint (23%)
Class 7 (18)	knife (55%)	robbery (55%)	violence (44%)	escape (39%)	threat (22%)

TABLE 3
Estimation results of MCR model and OLR model

	MCR			OLR		
Variable	Estimate	SE	P-value	Estimate	SE	P-value
X_1 (Age)	0.0012	0.001	0.042	0.0016	0.001	0.032
X_2 (Male)	0.0662	0.039	0.088	0.1081	0.048	0.025
X_3 (Han)	−0.0633	0.017	0.000	−0.0761	0.021	0.000
X_4 (Employed)	−0.0118	0.012	0.324	−0.0002	0.015	0.988
X_5 (Elementary)	0.0361	0.021	0.092	0.0574	0.027	0.031
X_6 (Middle)	0.0214	0.022	0.326	0.0432	0.027	0.110
X_7 (High)	−0.0381	0.029	0.184	0.0040	0.036	0.911

object of theft is general goods and cash. Class 6 is about the joint crime. Class 7 includes cases transformed to robbery.

As we are most interested in estimating the effects of extralegal factors on judicial impartiality, our interpretation should focus on X_j $(1 \le j \le 7)$ only, since they are related to these factors. The detailed estimation results are summarized in the left panel of Table 3. Note that the reported standard errors (SE) in Table 3 are computed by simply treating $\widehat{\pi}_{ik}s$ as fixed. By Theorem 2 we know that the estimated mixture probability $\widehat{\pi}_{ik}$ should converge to the true membership indicator function $a_{ik} = I(\mathcal{K}_i = k)$ with super fast convergence rate. This fact has been numerically verified in Section 3.5; see Figure 4. Therefore, the difference between $\widehat{\pi}_{ik}$ and a_{ik} becomes asymptotically ignorable. Consequently, the "simple" standard errors estimator, as reported in Table 3, is indeed a statistically valid estimator for the asymptotic variance. By Table 3 and focusing on the 5% level of significance, we find that X_1 (Age) and X_3 (Ethnicity) seem to be statistically significant. Consider, for example, the age effect. The corresponding coefficient of X_1 is 0.0012, indicating that those at an older age tend to receive longer sentences, even after controlling for the effects of legal factors, as reflected in Z_i. Specifically, holding all other factors fixed, if the age of the defendant is increased by 10 years, the average sentence is expected to be about 1.21% longer.

For the sake of comparison, the results of an ordinary linear regression (OLR) model is also presented. For the OLR model, a linear regression model is directly fitted for Y_i and X_i, when the information contained in Z_i is completely ignored. The detailed results are summarized in the right panel of Table 3. The OLR results seem to suggest that the gender of the defendant (i.e., X_2) and the education level (i.e., X_6, junior middle school) are also statistically significant. Consider, for example, the gender effect. By OLR results we find that males tend to receive longer average sentences than females. However, after conditioning on the legal factors, as reflected in Z_i by MCR, this effect becomes no longer statistically significant. Therefore, it seems to us that the seemingly significant gender effect, as detected by the OLR method, is very questionable. It might be due to the fact that male defendants are often involved in more severe criminal acts. Once those legal factor effects are well-controlled by Z_i, this seemingly significant gender effect disappears.

To further support the MCR method, we next demonstrate that the MCR method also leads to more accurate prediction results than the typically used OLR method. To this end, we randomly split the whole CJO dataset into a training dataset (50%) and a testing dataset (50%). Here we use $\mathcal{T} = \{(X_i^*, Y_i^*, Z_i^*) : 1 \le i \le n^*\}$ to represent the testing dataset. Consider an arbitrary testing sample $(X_i^*, Y_i^*, Z_i^*) \in \mathcal{T}$. By model (2.1) we should have $E(Y_i^*|X_i^*, Z_i^*) = \sum_{k=1}^K I(\mathcal{K}_i = k)\gamma_k + X_i^{*\top}\theta$. Since the value of Y_i^* should not be known in advance, we need to define a new mixture probability estimator as $\widehat{\pi}_{ik}^* = \widehat{\pi}_k \prod_{j=1}^p \widehat{p}_{kj}^{Z_{ij}} (1 - \widehat{p}_{kj})^{1-Z_{ij}} / \sum_{k=1}^K \widehat{\pi}_k \prod_{j=1}^p \widehat{p}_{kj}^{Z_{ij}} (1 - \widehat{p}_{kj})^{1-Z_{ij}}$. Thereafter, a predictor for

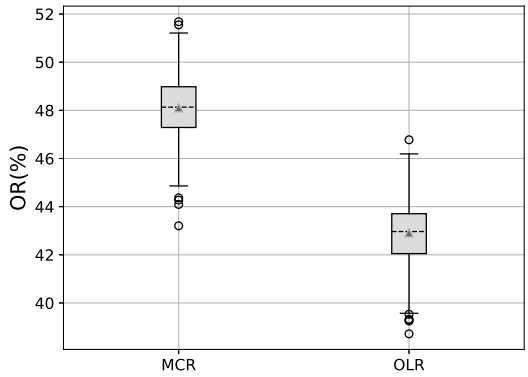

FIG. 6. *Out-of-sample R-squared (OR) values for the two competing models. The left and right boxplots represent the MCR and OLR methods, respectively.*

Y_i^* can be constructed as $\widehat{Y_i^*} = \sum_{k=1}^{K} \widehat{\pi}_{ik}^* \widehat{\gamma}_k + X_i^{*\top} \widehat{\theta}$, where the unknown parameters are estimated on the training dataset by the final estimator $\widehat{\Phi} = (\widehat{\gamma}^\top, \widehat{\theta}^\top)^\top \in \mathbb{R}^{K+q}$. Accordingly, the out-of-sample R-squared (OR) can be evaluated as

$$\mathrm{OR} = \left\{ 1 - \sum_{i=1}^{n^*} (Y_i^* - \widehat{Y_i^*})^2 / \sum_{i=1}^{n^*} (Y_i^* - \overline{Y_i^*})^2 \right\} \times 100\%,$$

where $\overline{Y_i^*} = \sum_{i=1}^{n^*} Y_i^* / n^*$. For a reliable evaluation, this experiment was randomly replicated for $M = 1000$ times. This leads to a total of M OR values, which are then box-plotted in Figure 6; see the left boxplot in Figure 6. Repeating the experiment for the OLR method with the interested response Y_i^* predicted by $\widehat{Y_i}^{*(\mathrm{OLR})} = \widehat{\gamma}^{(\mathrm{OLR})} + X_i^{*\top} \widehat{\theta}^{(\mathrm{OLR})}$, where $\widehat{\gamma}^{(\mathrm{OLR})}$ and $\widehat{\theta}^{(\mathrm{OLR})}$ are the ordinary least squared estimators obtained on the training dataset. That leads to the right boxplot in Figure 6. We find that the MCR method outperforms the OLR method clearly. The mean of OR in OLR method is about 42.91%, while that of MCR is about 48.10%, which is almost 5.2% better than OLR. The standard deviations of OR in OLR and MCR methods are about 1.20% and 1.24%, respectively. Then a standard two-sample t-test is conducted to compare 42.91% with 48.10%. This leads to an extremely significant p-value smaller than 10^{-8}. Therefore, it seems safe to conclude that, by utilizing information provided by Z_i appropriately, the MCR method should be a more accurate regression tool for testing judicial impartiality.

5. Concluding remarks. In this study we develop a novel MCR methodology for estimating extralegal factor effects with an application to Chinese burgary offenses. Both the theoretical and numerical properties of the MCR method are carefully studied. The main advantage of our MCR method is the ability to embrace more parameters than observations. The main limitation is the requirement that the ultrahigh dimensional control variables Z must be binary. Extensive simulation studies are presented to demonstrate the proposed MCR method. For the real data analysis, we find that the effect of some seemingly significant extralegal factors (such as gender) by an usual regression analysis disappears, once the latent class membership $\mathcal{K}$ can be correctly estimated and then controlled by Z. To summarize, we aim to provide here two important contributions to the existing literature. First, we provide the statistics literature a new regression tool for testing the interested conditional independence, when there exists an ultrahigh dimensional and binary control variable. Second, we provide the legal study literature, a new perspective for testing judicial impartiality, and demonstrate its usefulness on a large-scale Chinese burglary judgment dataset.

To conclude this article, we wish to discuss a few interesting new topics for future study. First, the MCR method assumes that the rich information contained in Z_i is represented by a binary feature vector. By doing so, the existence of a bag of keywords can be well-represented. Nevertheless, the associated frequency information is completely ignored. Then how to take the frequency information into consideration should be a good topic for further study (Kim et al. (2006), Kononenko (1991)). Second, we treat an EM algorithm as if it sufficiently converge if the difference between two consecutive estimates is sufficiently small. Our numerical experiments suggest that this simple method works fairly well. However, whether the final estimator obtained by our EM algorithm indeed converges numerically to the global optimizer is not theoretically investigated and thus not guaranteed in this work. A further research along this line seems quite involved and should be a good topic for future studies (Balakrishnan, Wainwright and Yu (2017), Bilmes et al. (1998), Xu and Jordan (1996)). Third, the number of latent classes K is estimated by a BIC selection criterion here, which our preliminary numerical experiments suggest that this BIC method works fairly well. However, its theoretical properties remain unknown. Then how to fill this important theoretical gap is another interesting direction for future exploration (Biernacki, Celeux and Govaert (2000), Zhao, Jin and Shi (2015)). Finally, the current MCR method can be viewed as a natural extension of the ordinary linear regression models. How to develop similar methods for many other popularly used generalized regression models (e.g., logistic regression) is also worth pursuing (Jansen (1993), Sedghi, Janzamin and Anandkumar (2016)).

Acknowledgments. Fang Wang is the correpsonding author. The authors would like to thank the Editor, the Associate Editor, and the referees for their constructive comments and advice that improved the quality of this paper.

Funding. Fang Wang's research is supported by National Natural Science Foundation of China (T2293773, 72371145) and Taishan Scholars Project (tsqn202211004).

Yuan Gao's research is partially supported by the Postdoctoral Fellowship Program of CPSF (GZC20230111).

Xiaojun Song's research is partially supported by National Natural Science Foundation of China (72373007, 72333001).

Hansheng Wang's research is partially supported by National Natural Science Foundation of China (12271012).

SUPPLEMENTARY MATERIAL

Theorem proofs (DOI: 10.1214/24-AOAS1893SUPPA; .pdf). The Supplementary Material contains technical conditions, proofs of main theoretical results, and verification details.

Data and code (DOI: 10.1214/24-AOAS1893SUPPB; .zip). The data and implementation Python code, along with instructions in README.md, are available as an online supplement (https://github.com/Shi12056/MCR.git).

REFERENCES

BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

BIELEN, S. and GRAJZL, P. (2021). Prosecution or persecution? Extraneous events and prosecutorial decisions. *J. Empir. Leg. Stud.* **18** 765–800.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BILMES, J. A. et al. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *TR—Int. Comput. Sci. Inst.* **4** 126.
BREUSCH, T. S. and PAGAN, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Rev. Econ. Stud.* **47** 239–253. MR0611118 https://doi.org/10.2307/2297111
BRIGHT, S. B. (2008). The failure to achieve fairness: Race and poverty continue to influence who dies. *Univ. Pa. J. Const. Law* **11** 23.
BUSHWAY, S. D. and PIEHL, A. M. (2001). Judging judicial discretion: Legal factors and racial discrimination in sentencing. *Law Soc. Rev.* 733–764.
CANES-WRONE, B., CLARK, T. S. and KELLY, J. P. (2014). Judicial selection and death penalty decisions. *Amer. Polit. Sci. Rev.* **108** 23–39.
DE VEAUX, R. D. (1989). Mixtures of linear regressions. *Comput. Statist. Data Anal.* **8** 227–245. MR1028403 https://doi.org/10.1016/0167-9473(89)90043-1
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. With discussion. MR0501537
EDMOND, G. (2002). Constructing miscarriages of justice: Misunderstanding scientific evidence in high profile criminal appeals. *Oxf. J. Leg. Stud.* **22** 53–89.
ENGLE, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handb. Econom.* **2** 775–826.
FELSON, M. and BOBA, R. L. (2010). Crime and everyday life. Sage.
GLYNN, A. N. and SEN, M. (2015). Identifying judicial empathy: Does having daughters cause judges to rule for women's issues? *Amer. J. Polit. Sci.* **59** 37–54.
GROSS, S. and SHAFFER, M. (2012). Exonerations in the United States.
GRÜN, B. and LEISCH, F. (2008). Finite mixtures of generalized linear regression models. In *Recent Advances in Linear Models and Related Areas* 205–230. Springer, Heidelberg. MR2523852 https://doi.org/10.1007/978-3-7908-2064-5_11
JANSEN, R. (1993). Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* 227–231.
KIM, S.-B., HAN, K.-S., RIM, H.-C. and MYAENG, S. H. (2006). Some effective techniques for naive Bayes text classification. *IEEE Trans. Knowl. Data Eng.* **18** 1457–1466.
KONONENKO, I. (1991). Semi-naive Bayesian classifier. In *Machine Learning—EWSL*-91 (*Porto*, 1991). *Lecture Notes in Computer Science* **482** 206–219. Springer, Berlin. MR1101397 https://doi.org/10.1007/BFb0017015
KREHBIEL, P. J. and CROPANZANO, R. (2000). Procedural justice, outcome favorability and emotion. *Soc. Justice Res.* **13**.
L'HEUREUX-DUBE, C. (2001). Beyond the myths: Qquality, impartiality, and justice. *J. Soc. Distress Homeless*.
LYNCH, M. and HANEY, C. (2011). Mapping the racial bias of the white male capital juror: Jury composition and the "Empathic divide". *Law Soc. Rev.* **45** 69–102.
MERON, T. (2005). Judicial independence and impartiality in international criminal tribunals. *Amer. J. Int. Law* **99** 359–369.
MEYERSON, D. (2006). *Understanding Jurisprudence*, 1st ed. Routledge, London.
MINNIER, J., YUAN, M., LIU, J. S. and CAI, T. (2015). Risk classification with an adaptive naive Bayes kernel machine model. *J. Amer. Statist. Assoc.* **110** 393–404. MR3338511 https://doi.org/10.1080/01621459.2014.908778
MISHLER, W. and SHEEHAN, R. S. (1993). The Supreme Court as a countermajoritarian institution? The impact of public opinion on Supreme Court decisions. *Amer. Polit. Sci. Rev.* **87** 87–101.
NOBLES, R. and SCHIFF, D. (1995). Miscarriages of justice: A systems approach. *Mod. Law Rev.* **58** 299.
PENG, Y. and CHENG, J. (2022). Ethnic disparity in Chinese theft sentencing. *China Rev.* **22** 47–71.
QAISER, S. and ALI, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *Int. J. Comput. Appl.* **181** 25–29.
RAMOS, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning* **242** 29–48, NJ, USA.
REYNOLDS, D. A. (2009). Gaussian mixture models. *Encycl. Biometrics* **741**.
RISH, I. et al. (2001). An empirical study of the naive Bayes classifier. In *IJCAI* 2001 *Workshop on Empirical Methods in Artificial Intelligence* **3** 41–46.
ROBERTS, S. (2003). 'Unsafe' convictions: Defining and compensating miscarriages of justice. *Mod. Law Rev.* **66** 441–451.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
SEDGHI, H., JANZAMIN, M. and ANANDKUMAR, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics* 1223–1231. PMLR.
SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2002723 https://doi.org/10.1007/b97553

SHELEY, J. F. and ASHKINS, C. D. (1981). Crime, crime news, and crime views. *Public Opin. Q.* **45** 492–506.
SHI, J., WANG, F., GAO, Y., SONG, X. and WANG, H. (2024). Supplement to "Mixture conditional regression with ultrahigh dimensional text data for estimating extralegal factor effects." https://doi.org/10.1214/24-AOAS1893SUPPA, https://doi.org/10.1214/24-AOAS1893SUPPB
SIMMONS, J. and FLOOD-PAGE, C. (2002). *Crime in England and Wales*. Home Office, London.
SPIEGELHALTER, D. J. and KNILL-JONES, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *J. R. Stat. Soc., A* **147** 35–58.
STEFFENSMEIER, D. and KRAMER, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, black, and male. *Criminology* **36** 763–798.
STITH, K., CABRANES, J. A. et al. (1998). *Fear of Judging*: *Sentencing Guidelines in the Federal Courts*. Univ. Chicago Press, Chicago.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
VERMUNT, J. K. and MAGIDSON, J. (2002). Latent class cluster analysis. In *Applied Latent Class Analysis* 89–106. Cambridge Univ. Press, Cambridge. MR1927666 https://doi.org/10.1017/CBO9780511499531.004
WADHAM, J. (1993). Unravelling miscarriages of justice. *New Law J.* **143** 1650–1650.
WEDEL, M., KAMAKURA, W. A., WEDEL, M. and KAMAKURA, W. A. (2000). *Mixture Regression Models*. Springer, Berlin.
WEILER, P. (1968). Two models of judicial decision-making. *Canadian Bar Review* **46** 406.
WU, H. C., LUK, R. W. P., WONG, K. F. and KWOK, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans. Inf. Syst.* **26** 1–37.
XU, K., LIU, H., WANG, F. and WANG, H. (2022). 'This crime is not that rrime'-classification and evaluation of four common crimes. *Law Probab. Risk* **20** 135–152.
XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** 129–151.
YANG, F.-J. (2018). An implementation of naive Bayes classifier. In 2018 *International Conference on Computational Science and Computational Intelligence* (*CSCI*) 301–306. IEEE Press, New York.
YE, X. (2010). The impact and direction of national standardized sentencing reform in China. *Columbia J. Asian Law* **24** 247.
ZHAO, J., JIN, L. and SHI, L. (2015). Mixture model selection via hierarchical BIC. *Comput. Statist. Data Anal.* **88** 139–153. MR3332023 https://doi.org/10.1016/j.csda.2015.01.019