Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE

Health Care Provider Clustering Using Fusion Penalty in Quasi-Likelihood

Lili Liu[1,2] | Kevin He[3] | Di Wang[3] | Shujie Ma[4] | Annie Qu[5] | Yihui Luan[2] | J. Philip Miller[1] | Yizhe Song[6] | Lei Liu[1]

[1]Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri, USA | [2]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China | [3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA | [4]Department of Statistics, University of California, Riverside, California, USA | [5]Department of Statistics, University of California, Irvine, California, USA | [6]Division of Biology and Biomedical Sciences, Washington University in St. Louis, St. Louis, Missouri, USA

Correspondence: Lei Liu (lei.liu@wustl.edu)

Received: 3 July 2023 | **Revised:** 25 March 2024 | **Accepted:** 29 March 2024

Funding: This research is partly supported by NIH grants R21 EY031884, R21 EY033518, R21 AG063370, R21 AG068955, R01 AG081244, UL1 TR002345, and R01 DK129539, Natural Science Foundation of Shandong Province ZR2021QA024, and China Postdoctoral Science Foundation 2021M691952.

Keywords: clustering | fixed effects | fusion penalty | latent class | quasi-likelihood | random effects

ABSTRACT

There has been growing research interest in developing methodology to evaluate the health care providers' performance with respect to a patient outcome. Random and fixed effects models are traditionally used for such a purpose. We propose a new method, using a fusion penalty to cluster health care providers based on quasi-likelihood. Without any priori knowledge of grouping information, our method provides a desirable data-driven approach for automatically clustering health care providers into different groups based on their performance. Further, the quasi-likelihood is more flexible and robust than the regular likelihood in that no distributional assumption is needed. An efficient alternating direction method of multipliers algorithm is developed to implement the proposed method. We show that the proposed method enjoys the oracle properties; namely, it performs as well as if the true group structure were known in advance. The consistency and asymptotic normality of the estimators are established. Simulation studies and analysis of the national kidney transplant registry data demonstrate the utility and validity of our method.

1 | Introduction

In recent years, much effort has been devoted to appropriately and efficiently assessing and comparing the performance of health care providers, including hospitals, nursing homes, transplant centers, and other medical practitioners. Evaluating the performance of the health care providers is important to assure safe and adequate delivery of health care to patients. It can guide providers towards enhancing their quality of care and help patients find the best providers for their health needs.

The classic research of comparing the performance of health care professionals often employs a test-based procedure, categorizing the providers into three groups: "worse than expected," "as expected," or "better than expected." The most prevalent methodologies are the random effects and fixed effects models. Normand and Shahian (2007) and K. He et al. (2013) explored the logistic random effects and fixed effects models for binary outcomes, respectively. Estes et al. (2020) proposed high-dimensional fixed effects Poisson and negative binomial regression models for count outcomes. However, both models have their limitations. The random effects model accounts for correlation within providers' observations through a random effect (e.g., random intercept), which leads to biases (shrinkage) in the estimates of provider effects (Ohlssen, Sharples, and Spiegelhalter 2007; Racz and Sedransk 2010). Additionally, a Gaussian distribution assumption

© 2024 Wiley-VCH GmbH.

may result in biased estimated effects for top or bottom providers if the true distribution is skewed (C. McCulloch and Neuhaus 2011). On the other hand, the fixed effects model can be less efficient in parameter estimates due to a larger number of fixed effects (usually one for each provider).

The Scientific Registry of Organ Recipients (SRTR) introduced a 5-tier evaluation of transplant centers based on the risk-adjusted 1-year graft survival of providers (Van Pilsum Rasmussen et al. 2018; Wey et al. 2018). Here, a rating of 5 signifies the best outcomes, while 1 indicates the least favorable. However, concerns have arisen about determining appropriate cutoffs for categorizing healthcare providers. Furthermore, the choice for the total number of tiers is arbitrary, as altering the number of tiers would inherently affect the list of evaluated providers. This highlights the pressing need for a more systematic classification approach for healthcare providers.

To better differentiate the performance of providers, a fused effects model (Liu et al. 2021) was proposed. This data-driven approach, which does not rely on prior knowledge of group classifications, is desirable for automatically clustering healthcare providers based on their performance. It operates on the presumption that provider effects belong to a limited set of groups, optimizing an objective function with a pairwise fusion penalty for clustering. Such a model is more flexible than the random effects model in that no distributional assumption is needed for the individual provider effect, and more efficient than the fixed effects model as only a small number of degrees (one for each group of providers rather than one for each provider) can capture the heterogeneity in providers' performance. However, its reliance on the sum of residual squares in the objective function makes it less applicable to, for example, binary or count outcomes.

In this paper, we consider a quasi-likelihood approach (Wedderburn 1974) which is more robust than the log-likelihood of an exponential family. It only needs to specify the mean-variance relationship, thus eliminating the need for the specific distributional assumption about the outcome. We thus maximize the penalized quasi-likelihood (PQL) (Breslow and Clayton 1993) with a pairwise fusion penalty for clustering. The PQL method can be easily implemented by iteratively fitting a penalized weighted least square model to a modified working response (P. McCullagh and Nelder 1989). The optimizing problem is computationally efficient with closed-form solutions. This data-driven procedure is able to automatically estimate the grouping structure with unknown grouping information. Note that we use the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) penalty function, which can correctly identify groups from a heterogeneous population (Ma and Huang 2017). The alternating direction method of multipliers (ADMM) algorithm is applied to implement the provider clustering (Boyd et al. 2011).

The rest of the paper is organized as follows. In Section 2, we describe the pairwise fusion penalization method to cluster providers. Section 3 introduces the theoretical properties of the proposed procedure. In Section 4, we assess the performance of our method via the Monte Carlo simulation studies. Section 5 illustrates the proposed method through the national kidney transplant registry data. We summarize our method and present some future directions in Section 6. The proofs are given in the Supporting Information Appendix.

2 | Model and Estimation

2.1 | Model

Let y_{ij} denote the outcome of the jth patient in provider i, and $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^\top$ denote the covariate vector, where $i = 1, \dots, m$ and $j = 1, \dots, n_i$. It is assumed that the observations y_{ij}'s have means $E(y_{ij}) = \mu_{ij}$ and variances $\text{var}(y_{ij}) = \phi V(\mu_{ij})$, where ϕ is an unknown dispersion parameter and $V(\cdot)$ is a known variance function. The mean μ_{ij} is related to the linear predictor $\eta_{ij} = a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta}$ through the known link function $\eta_{ij} = g(\mu_{ij})$. Here, a_i is the provider-specific effect, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is an unknown p-dimensional parameter vector.

We assume that a_i belongs to one of K groups $\mathcal{G}_1, \dots, \mathcal{G}_K$, which are mutually exclusive partitions of $\{1, \dots, m\}$. Moreover, the number of groups K and the group membership $\mathcal{G}_k$ are unknown in advance. We apply the penalty to shrink some of the pairs $|a_i - a_k|$ to zero. Consequently, we can automatically partition the provider effects into groups without the knowledge of a priori classification. Such a "fusion" penalty (Tibshirani et al. 2005) promotes both element similarity and element sparsity.

The log quasi-likelihood (QL) (Wedderburn 1974) is defined as

$$Q(\mu_{ij}, y_{ij}) = \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - t}{\phi V(t)} dt.$$

The QL works similarly to the maximum likelihood, but it does not require an exact distributional assumption. By merely defining the data's mean and the relationship between the mean and variance, the methodology remains robust even if there are slight inaccuracies in model details (C. E. McCulloch and Searle 2004). The PQL with the fusion penalty is

$$L_0(\boldsymbol{a}, \boldsymbol{\beta}, \lambda) = -\sum_{i=1}^{m} \sum_{j=1}^{n_i} \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - t}{\phi V(t)} dt + \sum_{1 \le i < k \le m} p_\gamma(|a_i - a_k|, \lambda), \tag{2.1}$$

where $\boldsymbol{a} = (a_1, \dots, a_m)^\top$ and $p_\gamma(t, \lambda)$ is a concave penalty function. Here we consider the SCAD penalty function, which is defined as

$$p_\gamma(t, \lambda) = \lambda \int_0^{|t|} \min\{1, (\gamma - x/\lambda)_+/(\gamma - 1)\} dx, \tag{2.2}$$

where $(x)_+ = x$ if $x > 0$ and $= 0$ otherwise, $\gamma > 0$ is a parameter that controls the concavity of the penalty function, and $\lambda \ge 0$ is the tuning parameter.

2.2 | Estimation Procedure

Let $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_m^\top)^\top$ with $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$, $\mathbf{X} = (\mathbf{X}_1^\top, \dots, \mathbf{X}_m^\top)^\top$ with $\mathbf{X}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^\top$, $N = \sum_{i=1}^m n_i$, $\mathbf{A} = \text{diag}(\mathbf{1}_{n_1}, \dots, \mathbf{1}_{n_m})$ with $\mathbf{1}_{n_i} = (1, \dots, 1)^\top$ being the vector with n_i ones. Denote $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{a} + \mathbf{X}\boldsymbol{\beta}$. The minimization problem (2.1) can be tackled by a quadratic approximation to the loss function based on a Taylor series

expansion around estimates $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\beta}^{(s)}$ at iteration s (see the Supporting Information Appendix) (P. McCullagh and Nelder 1989).

$$L_0(\boldsymbol{\alpha},\boldsymbol{\beta},\lambda) \approx \frac{1}{2}(\mathbf{z}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{z}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta}) + \sum_{1\le i<k\le m} p_\gamma(|a_i-a_k|,\lambda), \tag{2.3}$$

where the working vector $\mathbf{z}=(\mathbf{z}_1^\top,\ldots,\mathbf{z}_m^\top)^\top$ has components $\mathbf{z}_i = \{a_i + \mathbf{X}_i\boldsymbol{\beta} + \frac{\partial\eta_{ij}}{\partial\mu_{ij}}(y_{ij}-\mu_{ij})\}$. $\mathbf{W}$ is an $N\times N$ diagonal matrix of weights with diagonal elements $w_{ij} = \{\phi V(\mu_{ij})(\frac{\partial\eta_{ij}}{\partial\mu_{ij}})^2\}^{-1}$, where the dispersion parameter ϕ is estimated by

$$\widehat{\phi} = \frac{1}{N-p-m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(y_{ij}-\mu_{ij})^2}{V(\mu_{ij})} \tag{2.4}$$

(Liang and Zeger 1986; Wedderburn 1974). μ_{ij}, η_{ij}, ϕ, $\mathbf{W}$, and $\mathbf{z}$ are evaluated at $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\beta}^{(s)}$.

Note that the penalty function $p_\gamma(|a_i-a_k|,\lambda)$ cannot be written in the form of the addition of separate terms of $p_\gamma(|a_i|,\lambda)$ and $p_\gamma(|a_k|,\lambda)$ as in Lasso. Instead, we introduce a new set of parameters $\theta_{ik}=a_i-a_k$, which are equivalent to the pairwise differences of provider effects. As a result, the above minimization problem (2.3) can be transformed into the following constraint optimization problem:

$$\begin{aligned} L_0(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}) =& \frac{1}{2}(\mathbf{z}^{(s)}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{W}^{(s)}(\mathbf{z}^{(s)}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta}) \\ &+ \sum_{i<k} p_\gamma(|\theta_{ik}|,\lambda), \\ &\text{subject to} \quad a_i-a_k-\theta_{ik}=0, \end{aligned} \tag{2.5}$$

where $\boldsymbol{\theta}=\{\theta_{ik}, i<k\}^\top$. The ADMM algorithm is used to solve the constrained optimization (2.5). The estimates of the parameters are obtained by the augmented Lagrangian,

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\upsilon}) = L_0(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}) + \sum_{i<k}\upsilon_{ik}(\theta_{ik}-a_i+a_k) + \frac{\vartheta}{2}\sum_{i<k}(\theta_{ik}-a_i+a_k)^2, \tag{2.6}$$

where $\boldsymbol{\upsilon}=\{\upsilon_{ik}, i<k\}^\top$ are Lagrange multipliers, ϑ is the penalty parameter. We use the ADMM to iteratively compute the estimates of $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\upsilon})$. For given $\boldsymbol{\theta}^{(s)},\boldsymbol{\upsilon}^{(s)}$ at iteration s, we have:

$$(\boldsymbol{\alpha}^{(s+1)},\boldsymbol{\beta}^{(s+1)}) = \arg\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta}^{(s)},\boldsymbol{\upsilon}^{(s)}), \tag{2.7}$$

$$\boldsymbol{\theta}^{(s+1)} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\alpha}^{(s+1)},\boldsymbol{\beta}^{(s+1)},\boldsymbol{\theta},\boldsymbol{\upsilon}^{(s)}), \tag{2.8}$$

$$\upsilon_{ik}^{(s+1)} = \upsilon_{ik}^{(s)} + \vartheta(a_i^{(s+1)} - a_k^{(s+1)} - \theta_{ik}^{(s+1)}). \tag{2.9}$$

To update $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the minimization problem (2.7) is equivalent to minimizing

$$F(\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{z}^{(s)}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{W}^{(s)}(\mathbf{z}^{(s)}-\mathbf{A}\boldsymbol{\alpha}-\mathbf{X}\boldsymbol{\beta}) + \frac{\vartheta}{2}\|\boldsymbol{\Delta}\boldsymbol{\alpha}-\boldsymbol{\theta}^{(s)}+\vartheta^{-1}\boldsymbol{\upsilon}^{(s)}\|^2 + C, \tag{2.10}$$

where $\boldsymbol{\Delta} = \{(\mathbf{e}_i-\mathbf{e}_j), i<j\}^\top$ with $\mathbf{e}_i$ being the ith unit $m\times 1$ vector whose ith element is 1 and the remaining ones are 0, and C is a constant independent of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Thus, we set the derivatives $\partial F(\boldsymbol{\alpha},\boldsymbol{\beta})/\partial\boldsymbol{\alpha}=0$ and $\partial F(\boldsymbol{\alpha},\boldsymbol{\beta})/\partial\boldsymbol{\beta}=0$ to obtain the following updates $\boldsymbol{\alpha}^{(s+1)}$ and $\boldsymbol{\beta}^{(s+1)}$:

$$\boldsymbol{\alpha}^{(s+1)} = (\mathbf{A}^\top\mathbf{Q}_x\mathbf{A}+\vartheta\boldsymbol{\Delta}^\top\boldsymbol{\Delta})^{-1}[\mathbf{A}^\top\mathbf{Q}_x\mathbf{z}^{(s)} + \vartheta\boldsymbol{\Delta}^\top(\boldsymbol{\theta}^{(s)}-\vartheta^{-1}\boldsymbol{\upsilon}^{(s)})], \tag{2.11}$$

where $\mathbf{Q}_x = \mathbf{W}^{(s)} - \mathbf{W}^{(s)}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{(s)}$ and

$$\boldsymbol{\beta}^{(s+1)} = (\mathbf{X}^\top\mathbf{W}^{(s)}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}^{(s)}(\mathbf{z}^{(s)}-\mathbf{A}\boldsymbol{\alpha}^{(s+1)}). \tag{2.12}$$

To update $\boldsymbol{\theta}$, we need to minimize the function

$$\frac{\vartheta}{2}(\theta_{ik}-\pi_{ik}^{(s+1)})^2 + \sum_{i<k} p_\gamma(|\theta_{ik}|,\lambda),$$

where $\pi_{ik}^{(s+1)} = a_i^{(s+1)} - a_k^{(s+1)} + \vartheta^{-1}\upsilon_{ik}^{(s)}$. Due to the concave penalties, although the objective function $L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\upsilon})$ is no longer a convex function, it is convex with respect to each θ_{ik} when $\gamma > 1/\vartheta+1$ for the SCAD penalty. Moreover, for given $(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\upsilon})$, the minimizer of $L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\upsilon})$ with respect to θ_{ik} is unique with a closed-form expression. For the SCAD penalty with $\gamma > 1/\vartheta+1$, the explicit solution is

$$\theta_{ik}^{(s+1)} = \begin{cases} S(\pi_{ik}^{(s+1)},\lambda/\vartheta) & |\pi_{ik}^{(s+1)}| \le \lambda+\lambda/\vartheta, \\ \frac{S(\pi_{ik}^{(s+1)},\gamma\lambda/((\gamma-1)\vartheta))}{1-1/((\gamma-1)\vartheta)} & \lambda+\lambda/\vartheta < |\pi_{ik}^{(s+1)}| \le \lambda\gamma, \\ \pi_{ik}^{(s+1)} & |\pi_{ik}^{(s+1)}| > \lambda\gamma, \end{cases} \tag{2.13}$$

where $S(x,t) = x(1-t/|x|)_+$ is a groupwise soft thresholding operator. Finally, the Lagrange multiplier υ_{ik} is updated by (2.9). This process is repeated iteratively until the convergence over a grid of values for λ.

Due to the approximation in the computation, Equation (2.10) does not fully capture parameter information, and the iteration may also lead to a loss in efficiency (Liu et al. 2023). Let $\widehat{K}$ be the estimated number of groups, and let $\widehat{\mathcal{G}}_1,\ldots,\widehat{\mathcal{G}}_{\widehat{K}}$ be the estimated groups, and $\widehat{\delta}_k$ be the common value of a_i from group $\widehat{\mathcal{G}}_k$. Once the groups have been identified, we conduct a refitting step to estimate $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ by maximizing the following log quasi-likelihood with the grouped information:

$$\ell(\boldsymbol{\delta},\boldsymbol{\beta}) = \sum_{k=1}^{\widehat{K}}\sum_{i=\widehat{\mathcal{G}}_k}\sum_{j=1}^{n_i}\int_{y_{ij}}^{\mu_{ij}}\frac{y_{ij}-t}{\phi V(t)}dt, \tag{2.14}$$

where $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_{\widehat{K}})^\top$ and $g(\mu_{ij}) = \delta_k + \mathbf{x}_{ij}^\top\boldsymbol{\beta}$ for $i\in\widehat{\mathcal{G}}_k$. The above procedure can be conveniently implemented using the R function *glm* by specifying the family argument as a quasi-family, which also provides standard errors (SEs) and p-values of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$.

The initial values for the ADMM algorithm are taken as follows. The initial values $\boldsymbol{\alpha}^{(0)}$ are considered the fixed effects and obtained from the generalized linear model for the repeated measures data, then we set $\theta_{ik}^{(0)} = a_i^{(0)} - a_k^{(0)}$ and $\boldsymbol{\upsilon}^{(0)} = \mathbf{0}$. Following Boyd et al. (2011), the iterative algorithm terminates when primal residuals $\mathbf{r}^{(s+1)} = \boldsymbol{\Delta}\boldsymbol{\alpha}^{(s+1)} - \boldsymbol{\theta}^{(s+1)}$ and dual residuals $\mathbf{s}^{(s+1)} = \vartheta\boldsymbol{\Delta}^{\top}(\boldsymbol{\theta}^{(s+1)} - \boldsymbol{\theta}^{(s)})$ are close to zero, that is, $\|\mathbf{r}^{(s+1)}\| < \epsilon_{\text{primal}}^{(s+1)}$ and $\|\mathbf{s}^{(s+1)}\| < \epsilon_{\text{dual}}^{(s+1)}$, where $\epsilon_{\text{dual}}^{(s+1)}$ and $\epsilon_{\text{primal}}^{(s+1)}$ are specified as Boyd et al. (2011). If $\widehat{\theta}_{ik} = 0$, then a_i and a_k are classified into the same group.

Once we have calculated the solution path on a grid of λ values, we will select the optimal λ by minimizing the modified Bayesian information criterion (BIC) (Wang, Li, and Leng 2009) with a grid search. The BIC is given by

$$\text{BIC}(\lambda) = -2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{y_{ij}\widehat{\eta}_{ij}(\lambda) - b(\widehat{\eta}_{ij}(\lambda))\} + C_N(\widehat{K}(\lambda) + p)\log N,$$

where $\widehat{\eta}_{ij}(\lambda) = \widehat{a}_i(\lambda) + \mathbf{x}_{ij}\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{K}(\lambda)$ are the estimates at a given λ, p is the dimension of the parameter $\boldsymbol{\beta}$, and C_N is a positive number depending on the total number of observations $N = \sum_{i=1}^{m} n_i$. Following Ma and Huang (2017), $C_N = \log(\log(N + p))$.

We summarize the proposed method in Algorithm 1 below.

ALGORITHM 1 | ADMM algorithm for provider analysis.

Require: Initialize $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\upsilon}^{(0)}$, $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$.
  for $s = 0, 1, 2, \ldots$ **do**
    Compute the dispersion parameter ϕ using (2.4)
    Compute the weights $\mathbf{W}$ and the working response $\mathbf{z}$ based on $\boldsymbol{\alpha}^{(s)}$ and $\boldsymbol{\beta}^{(s)}$
    Compute $\boldsymbol{\alpha}^{(s+1)}$ using (2.11)
    Compute $\boldsymbol{\beta}^{(s+1)}$ using (2.12)
    Compute $\boldsymbol{\theta}^{(s+1)}$ using (2.13)
    Compute $\boldsymbol{\upsilon}^{(s+1)}$ using (2.9)
    if the convergence criterion is met, **then**
      Stop and denote the last iteration by $(\widehat{\boldsymbol{\beta}}(\lambda), \widehat{\boldsymbol{\alpha}}(\lambda))$,
    else
      $s = s + 1$.
    end if
  end for
Ensure: Output

3 | Theoretical Properties

In this section, we establish the theoretical properties of the proposed procedure. First, we briefly review some definitions. Let $|\mathcal{G}_{\min}| = \min_{1\le k\le K}|\mathcal{G}_k|$ and $|\mathcal{G}_{\max}| = \max_{1\le k\le K}|\mathcal{G}_k|$ be the minimum and maximum group sizes, respectively. For any vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_p)^T$, denote the supremum norm by $\|\boldsymbol{\zeta}\|_\infty = \max_{1\le j\le p}|\zeta_j|$, and for any matrix $\mathbf{A} = (A_{ij})_{i=1,j=1}^{s,t}$, denote $\|\mathbf{A}\|_\infty = \max_{1\le i\le s}\sum_{j=1}^{t}|A_{ij}|$. Let $\mathcal{M}_{\mathcal{G}}$ be the subspace of R^m, defined as

$$\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\alpha} \in R^m : a_i = a_j, \text{for any } i, j \in \mathcal{G}_k, 1 \le k \le K\}.$$

For each $\boldsymbol{\alpha} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\alpha} = \boldsymbol{\Psi}\boldsymbol{\delta}$, where $\boldsymbol{\Psi} = \{\psi_{ik}\}$ is the $m \times K$ matrix with $\psi_{ik} = 1$ for $i \in \mathcal{G}_k$ and $\psi_{ik} = 0$ otherwise; $\boldsymbol{\delta}$ is a $K \times 1$ vector of parameters with $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^{\top}$, and δ_k is the common value for the a_i's from the kth group. Denote $\mathbf{A}\boldsymbol{\alpha} = \mathbf{A}\boldsymbol{\Psi}\boldsymbol{\delta} = \mathbf{B}\boldsymbol{\delta}$.

We establish the properties of the oracle estimator, which is defined as the weighted least squares estimator assuming that the underlying group structure $(\mathcal{G}_1, \ldots, \mathcal{G}_K)$ is known. Specifically, the oracle estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$(\widehat{\boldsymbol{\alpha}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) = \underset{\boldsymbol{\alpha}\in\mathcal{M}_{\mathcal{G}}, \boldsymbol{\beta}\in R^p}{\arg\min}\frac{1}{2}(\mathbf{z} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{W}(\mathbf{z} - \mathbf{A}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}),$$

Correspondingly, the oracle estimators for the common effects $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are

$$\begin{aligned}(\widehat{\boldsymbol{\delta}}^{or}, \widehat{\boldsymbol{\beta}}^{or}) &= \underset{\boldsymbol{\delta}\in R^K, \boldsymbol{\beta}\in R^p}{\arg\min}\frac{1}{2}(\mathbf{z} - \mathbf{B}\boldsymbol{\delta} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{W}(\mathbf{z} - \mathbf{B}\boldsymbol{\delta} - \mathbf{X}\boldsymbol{\beta}) \\ &= [(\mathbf{B}, \mathbf{X})^{\top}\mathbf{W}(\mathbf{B}, \mathbf{X})]^{-1}(\mathbf{B}, \mathbf{X})^{\top}\mathbf{W}\mathbf{z},\end{aligned}$$

and the corresponding oracle estimator of $\boldsymbol{\alpha}$ is $\widehat{\boldsymbol{\alpha}}^{or} = \boldsymbol{\Psi}\widehat{\boldsymbol{\delta}}^{or}$. To avoid confusion, let $\boldsymbol{\beta}^0$ and $\boldsymbol{\alpha}^0$ be the true parameter values, and δ_k^0 be the true common provider effect for group k, $k = 1, \ldots, K$ and $\boldsymbol{\delta}^0 = (\delta_1^0, \ldots, \delta_K^0)^{\top}$.

The following conditions are introduced for analyzing the asymptotic behavior of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$.

(C1) The function $\rho_\gamma(t) = \lambda^{-1}p_\lambda(t, \lambda)$ is symmetric, nondecreasing and concave in t for $t \in [0, \infty)$. It is constant for $t \ge c\lambda$ with some constant $c > 0$, and $\rho(0) = 0$, where $\rho'(t)$ exists and is continuous except for a finite number values of t and $\rho(0+) = 1$.

(C2) The noise vector $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1n_1}, \ldots, \epsilon_{m1}, \ldots, \epsilon_{mn_m})^T$ has sub-Gaussian tails such that $P(|\boldsymbol{s}^T\boldsymbol{\epsilon}| > \|\boldsymbol{s}\|x) \le 2\exp(-c_1x^2)$ for any vector $\boldsymbol{s} \in R^N$ and $x > 0$, where $0 < c_1 < \infty$ and $N = \sum_{i=1}^{m} n_i$.

(C3) There exists a positive constant M_1 such that the diagonal elements w_{ij} in weight matrix $\mathbf{W}$ satisfying $\sup_{ij}|w_{ij}| \le M_1$, for $1 \le i \le m$ and $1 \le j \le n_i$. Assume $\|\mathbf{X}_h\| = \sqrt{N}$, for $1 \le h \le p$, $\lambda_{\min}[(\mathbf{B}, \mathbf{X})^{\top}\mathbf{W}(\mathbf{B}, \mathbf{X})] \ge C_1|\mathcal{G}_{\min}|$, $\|\mathbf{X}\|_\infty \le pC_2$ for some constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$.

The conditions above are analogous to those in Ma and Huang (2017). Conditions (C1) and (C2) are commonly used in penalized methods for high-dimensional settings. The concave penalty SCAD satisfies Condition (C1). Similar to the condition mentioned in Ma and Huang (2017), Condition (C3) also includes the bounded conditions for the weight matrix.

Theorem 3.1. *Assume the conditions C1–C3 hold, If $K = o(m)$, $p = o(m)$, and $|\mathcal{G}_{\min}| \gg \sqrt{(K+p)N\log m}$. With probability at least $1 - 2(K+p)m^{-1}$, the oracle estimators have the following*

asymptotic properties:

$$\left\| ((\widehat{\boldsymbol{\alpha}}^{or} - \boldsymbol{\alpha}^0)^\top, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^\top)^\top \right\|_\infty \leq \phi_m,$$

where $\phi_m = c_1^{-1/2} C_1^{-1} M_1 \sqrt{K+p} |\mathcal{G}_{\min}|^{-1} \sqrt{N \log m}$. *Moreover, for any vector* $t_m \in R^{K+p}$, *as* $m \to \infty$,

$$\sigma_m^{-1}(\mathbf{t}_m)\mathbf{t}_m^\top((\widehat{\boldsymbol{\delta}}^{or} - \boldsymbol{\delta}^0)^\top, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^\top)^\top \to N(0,1),$$

where $\sigma_m(\mathbf{t}_m) = \{\mathbf{t}_m^\top [\mathbf{U}^\top \mathbf{W} \mathbf{U}]^{-1} \mathbf{t}_m\}^{1/2}$ *and* $\mathbf{U} = (\mathbf{B}, \mathbf{X}) = (\mathbf{A}\boldsymbol{\Psi}, \mathbf{X})$.

The above result indicates that the bound $\phi_m \to 0$ as $m \to \infty$, implying that the oracle estimators $\widehat{\boldsymbol{\alpha}}^{or}$ and $\widehat{\boldsymbol{\beta}}^{or}$ are consistent with probability tending to one. It is worth noting that the oracle estimators $\widehat{\boldsymbol{\delta}}^{or}$ and $\widehat{\boldsymbol{\beta}}^{or}$ are normally distributed asymptotically. In addition, let $b_m = \min_{i \in \mathcal{G}_s, k \in \mathcal{G}_t, s \neq t} |a_i^0 - a_k^0| = \min_{s \neq t} |\delta_i^0 - \delta_k^0|$ be the minimal difference between any two groups. We have the following theorem.

Theorem 3.2. *Under the conditions C1–C3, if* $b_m > c\lambda$ *for some positive constant* $c > 0$, *and* $\lambda \gg \phi_m$ *with* ϕ_m *given in Theorem 3.1, then there exists a local minimizer* $(\widehat{\boldsymbol{\alpha}}(\lambda)^\top, \widehat{\boldsymbol{\beta}}(\lambda)^\top)^\top$ *of the objective function given in (2.3) satisfying as* $m \to \infty$,

$$P\left((\widehat{\boldsymbol{\alpha}}(\lambda)^\top, \widehat{\boldsymbol{\beta}}(\lambda)^\top)^\top = ((\widehat{\boldsymbol{\alpha}}^{or})^\top, (\widehat{\boldsymbol{\beta}}^{or})^\top)^\top \right) \to 1.$$

Theorem 3.2 indicates the oracle property; in other words, the oracle estimator $((\widehat{\boldsymbol{\alpha}}^{or})^\top, (\widehat{\boldsymbol{\beta}}^{or})^\top)^\top$ is a local minimizer $(\widehat{\boldsymbol{\alpha}}(\lambda)^\top, \widehat{\boldsymbol{\beta}}(\lambda)^\top)^\top$ of the objective function (2.3) with probability approaching one. Thus, the true groups can be recovered with a high probability. It also implies that $P(\widehat{\boldsymbol{\delta}}(\lambda) = \widehat{\boldsymbol{\delta}}^{or}) \to 1$, where $\widehat{\boldsymbol{\delta}}(\lambda)$ and $\widehat{\boldsymbol{\delta}}^{or}$ are the distinct values of $\widehat{\boldsymbol{\alpha}}(\lambda)$ and $\widehat{\boldsymbol{\alpha}}^{or}$, respectively. This property together with the asymptotic normality given in Theorem 3.1 leads to the following corollary regarding the asymptotic property of estimators $(\widehat{\boldsymbol{\delta}}(\lambda)^\top, \widehat{\boldsymbol{\beta}}(\lambda)^\top)^\top$.

Corollary 3.1. *Suppose the conditions in Theorem 3.2 hold, for any vector* $t_m \in R^{K+p}$, *as* $m \to \infty$,

$$\sigma_m^{-1}(\mathbf{t}_m)\mathbf{t}_m^\top((\widehat{\boldsymbol{\delta}}(\lambda) - \boldsymbol{\delta}^0)^\top, (\widehat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^0)^\top)^\top \to N(0,1),$$

where $\sigma_m(\mathbf{t}_m) = \{\mathbf{t}_m^\top [\mathbf{U}^\top \mathbf{W} \mathbf{U}]^{-1} \mathbf{t}_m\}^{1/2}$, *with* $\mathbf{U} = (\mathbf{B}, \mathbf{X}) = (\mathbf{A}\boldsymbol{\Psi}, \mathbf{X})$.

The asymptotic distribution of the estimators provides a theoretical justification for further conducting statistical inference on provider clustering.

4 | Simulation

In this section, we conduct simulation studies to assess the performance of clustering models. Traditionally, in Model (2.1), the intercept a_i is either taken as a random effect or a fixed effect. We compare the performance of our fused effects model with the SCAD penalty (SCAD) with the random effects model (RE), the fixed effects model (FE), and the latent class model (LC). The classification accuracy and the efficiency of estimation are evaluated from 100 replicates. Two different provider sizes $m = 50$ and 100 are considered.

Example 1. We first generate the outcomes from the standard Poisson model (dispersion parameter $\phi = 1$) with mean μ_{ij}. The log-likelihood is

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} \log \mu_{ij} - \mu_{ij} - \log y_{ij}!),$$

where the canonical parameter is $\eta_{ij} = a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta} = \log \mu_{ij}$ and the variance $V(y_{ij}) = V_{ij} = \mu_{ij}$. Then the weights are $\omega_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V_{ij}^{-1} = \mu_{ij}^2 (1/\mu_{ij}) = \mu_{ij}$. The observations with large means receive large weights. Among the m provider effects a_i's, 10% are set to the value -1 for overperformers, 20% are set to the value 1 for underperformers, and the remaining 70% are set to the value 0 for providers not different from the reference standard. The coefficient $\boldsymbol{\beta} = (0.2, 0.2)^\top$. The covariate $\mathbf{X}$ is generated from a multivariate normal distribution with mean 0, variance 1, and an exchangeable correlation $\rho = 0.2$. We consider the number of patients in each provider $n_i \sim$ Uniform (90, 100).

The Rand index (RI) (Rand 1971) is used to measure the degree of agreement between the estimated partitions and the true partitions. The RI takes a value between 0 and 1, and a larger RI value indicates more agreement. The top panel of Table 1 reports the mean, median, and SD of the estimated number of groups $\widehat{K}$, and the RI for measuring clustering accuracy. The median of $\widehat{K}$ is 3—the true number of groups. As expected, the RI value is close to 1.

Let $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$ be the common estimates of provider effects from the overperforming groups, the reference group, and the underperforming groups, respectively. The expectation values of $\widehat{\delta}_1, \widehat{\delta}_2$, and $\widehat{\delta}_3$ are -1, 0, and 1, respectively. The top panel of Table 2 presents the bias, the SD, the SE, and the coverage probability of the 95% confidence intervals (CP) of $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$ for the SCAD estimators and oracle estimators. The measures by SCAD are calculated based on replicates with the estimated number of groups equal to three. We observe that the SCAD method performs very close to the oracle, supporting the oracle property established in Theorem 3.2. Clearly, our estimators $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$ agree well with the corresponding true values on average for all cases. The results indicate that our method yields accurate estimates of provider effects. Moreover, there is good agreement between SD and SE values, and the coverage probabilities are close to the nominal level 0.95.

To compare the efficiency of estimates $\widehat{\boldsymbol{\alpha}}$ to the competing models, the mean squared error (MSE), defined as $\sum_{i=1}^{m} (\widehat{a}_i - a_i)^2/m$, is calculated for each replicated dataset. The top panel of Table 3 reports the MSE of $\widehat{\boldsymbol{\alpha}}$, the bias, the SD, the SE, and the coverage probability of 95% confidence intervals (CP) of the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$. For graphical illustration, we also show the boxplots of the MSE of $\widehat{\boldsymbol{\alpha}}$ in Figure 1 of the Supporting Information. The MSEs of $\widehat{\boldsymbol{\alpha}}$ by our method are smaller compared to those of RE and FE models and very close to those of oracle. For $\widehat{\beta}_1$ and $\widehat{\beta}_2$, all methods yield nearly identical and approximately unbiased results. The means of dispersion parameter estimates $\widehat{\phi}$ by our method are 1.006 for $m = 50$ and 1.014 for $m = 100$, which are very close to the true value. Thus, our method has the correct SEs and coverage probabilities.

TABLE 1 | The sample mean, median, and standard deviation (SD) of $\widehat{K}$, and the Rand index (RI) value for the SCAD and latent class (LC) methods with $m = 50, 100$.

	Method	m	Mean	Median	SD	RI
Poisson with three groups	SCAD	50	3.01	3.00	0.100	1.000
		100	3.02	3.00	0.141	1.000
	LC	50	3.00	3.00	0.000	1.000
		100	3.00	3.00	0.000	1.000
Logistic with three groups	SCAD	50	3.06	3.00	0.239	0.975
		100	3.14	3.00	0.403	0.975
	LC	50	3.00	3.00	0.000	0.977
		100	3.00	3.00	0.000	0.976
Overdispersed Poisson with three groups	SCAD	50	3.11	3.00	0.373	0.990
		100	3.15	3.00	0.386	0.988
	LC	50	4.71	5.00	0.640	0.986
		100	4.99	5.00	0.100	0.988
Poisson with five groups	SCAD	50	5.07	5.00	0.326	0.996
		100	5.13	5.00	0.338	0.998
	LC	50	4.09	4.00	0.288	0.981
		100	4.10	4.00	0.302	0.981

Example 2. We generate the outcomes from the logistic regression. The logistic model for a binary response variable $y \in \{0, 1\}$ is given by $P(y_{ij} = 1|\mathbf{x}_{ij}) = \mu_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}$ with $\eta_{ij} = a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta}$. The log-likelihood function is

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}(a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta}) - \log(1 + \exp(a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta})).$$

For the logistic model, the working response $\mathbf{z} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and $\mathbf{W}$ is a diagonal matrix of weights, with elements $\omega_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V_{ij}^{-1} = \mu_{ij}(1 - \mu_{ij})$. $\mathbf{X}$, $\boldsymbol{\beta}$, provider effects a_i's, and the number of observations in each provider n_i are generated in the same way as in the Poisson setting. P. McCullagh and Nelder (1989) pointed out that overdispersion is not possible if y_{ij} only takes values 0 and 1, and it must be distributed as Bernoulli. Therefore, we set the overdispersion parameter $\phi = 1$ and do not adjust for overdispersion.

The middle panel of Table 1 reports the mean, median, and SD of the estimated number of groups $\widehat{K}$, and the RI for measuring clustering accuracy. From Table 1, we can see the proposed method works well in terms of classification accuracy, with a high RI of classifying providers as top- or bottom-performers. The middle panel of Table 2 presents the bias, the SD, the SE, and the coverage probability of the 95% confidence intervals (CP) of the estimators $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$. Similar to Poisson models, the estimates of $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$ by our fused effects method perform as well as the oracle estimates. Moreover, our estimates have CPs close to the nominal level. The middle panel of Table 3 reports the MSE of the estimator $\widehat{\boldsymbol{\alpha}}$, the bias, the SD, the SE, and the coverage probability of 95% confidence intervals (CP) of the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$. We observe that the MSE values from our method are much smaller than those from the RE and FE methods. All methods performed similarly to each other in terms of estimating $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

Example 3. In this example, we show that our fused effects model with quasi-likelihood is robust in the presence of misspecification of variance structure. It is known that the relationship between the variance and mean in the Poisson distribution satisfies $\sigma^2 = \mu$. However, such a relation does not hold for the overdispersed Poisson distribution. To generate data from overdispersed Poisson, we consider the negative binomial distribution with number of failures r and success probability p $(0 < p \leq 1)$, which has a density function as below:

$$\frac{\Gamma(y_{ij} + r)}{\Gamma(r) y_{ij}!} p^r (1 - p)^{y_{ij}},$$

where $y_{ij} = 0, 1, 2, \ldots$. We have the mean $\mu_{ij} = \frac{r(1-p)}{p}$ and the variance $V_{ij} = \frac{r(1-p)}{p^2}$; therefore, $V_{ij} = \frac{\mu_{ij}^2}{r} + \mu_{ij}$. We set $p = 0.5$ and $r = e^{\eta_{ij}}$, so $\mu_{ij} = \mathrm{e}^{\eta_{ij}}$ and $V_{ij} = 2\mu_{ij}$. This is related to the overdispersed Poisson where the variance is proportional to the mean, that is, $V_{ij} = \phi\mu_{ij}$ with $\phi = 2$. The canonical parameter is $\eta_{ij} = a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta} = \log \mu_{ij}$, and the weight is $\omega_{ij} = \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}}\right)^2 V_{ij}^{-1} = \mu_{ij}^2 (\phi\mu_{ij})^{-1} = \mu_{ij}/\phi$. $\mathbf{X}$, $\boldsymbol{\beta}$, provider effects a_i's, and the number of observations in each provider n_i are generated in the same way as in the Poisson setting.

The third panel of Table 1 reports the mean, median, and SD of the estimated number of groups $\widehat{K}$, and the RI for measuring clustering accuracy. From Table 1, we can see the median of $\widehat{K}$ by our fused-effects model is 3—the true number of groups. Moreover, the RI value is very close to 1. The third panel of

TABLE 2 | The bias, standard deviation (SD), standard error (SE), and the coverage probability of the 95% confidence intervals (CP) of the estimators for the smoothly clipped absolute deviation (SCAD), latent class (LC), and oracle estimators (ORACLE) in Poisson, logistic, and overdispersed Poisson models. The results for the latent class model are not presented for Examples 3 and 4 due to its poor performance in identifying the correct number of classes (the median of $\widehat{K}$ is not the true number of groups in Table 1).

			$m = 50$				$m = 100$			
Model		**Method**	**Bias**	**SD**	**SE**	**CP**	**Bias**	**SD**	**SE**	**CP**
Poisson with three groups	$\widehat{\delta}_1$	SCAD	0.008	0.077	0.074	0.94	0.007	0.048	0.052	0.96
		ORACLE	0.008	0.080	0.074	0.93	0.007	0.048	0.052	0.96
		LC	0.009	0.080	0.074	0.93	0.007	0.047	0.052	0.96
	$\widehat{\delta}_2$	SCAD	0.001	0.016	0.017	0.96	0.002	0.012	0.012	0.95
		ORACLE	0.001	0.016	0.017	0.96	0.002	0.012	0.012	0.95
		LC	0.001	0.017	0.017	0.96	0.002	0.011	0.012	0.97
	$\widehat{\delta}_3$	SCAD	0.002	0.020	0.020	0.94	5e-4	0.016	0.014	0.91
		ORACLE	0.002	0.020	0.020	0.94	0.001	0.016	0.014	0.91
		LC	0.001	0.018	0.020	0.95	4e-4	0.016	0.014	0.94
Logistic with three groups	$\widehat{\delta}_1$	SCAD	0.013	0.116	0.105	0.90	0.008	0.070	0.074	0.93
		ORACLE	0.006	0.118	0.104	0.90	0.004	0.066	0.074	0.95
		LC	0.002	0.129	0.104	0.88	0.001	0.070	0.073	0.93
	$\widehat{\delta}_2$	SCAD	0.008	0.040	0.035	0.89	0.003	0.026	0.025	0.90
		ORACLE	0.005	0.035	0.035	0.94	0.002	0.025	0.025	0.95
		LC	0.006	0.039	0.035	0.90	0.003	0.025	0.025	0.94
	$\widehat{\delta}_3$	SCAD	0.013	0.077	0.075	0.96	0.013	0.058	0.053	0.91
		ORACLE	0.003	0.075	0.074	0.96	0.004	0.054	0.052	0.96
		LC	0.002	0.083	0.074	0.94	0.005	0.060	0.052	0.90
Overdispersed Poisson with three groups	$\widehat{\delta}_1$	SCAD	0.019	0.097	0.100	0.98	0.015	0.069	0.071	0.92
		ORACLE	0.025	0.094	0.100	0.96	0.012	0.075	0.071	0.90
	$\widehat{\delta}_2$	SCAD	0.001	0.029	0.030	0.97	0.003	0.026	0.021	0.90
		ORACLE	4e-4	0.027	0.030	0.99	0.002	0.025	0.021	0.90
	$\widehat{\delta}_3$	SCAD	0.003	0.049	0.050	0.96	0.005	0.039	0.036	0.94
		ORACLE	0.005	0.049	0.050	0.96	8e-5	0.037	0.035	0.95
Poisson with five groups	$\widehat{\delta}_1$	SCAD	0.048	0.198	0.134	0.94	0.004	0.089	0.087	0.94
		ORACLE	0.005	0.112	0.122	0.97	0.009	0.083	0.086	0.92
	$\widehat{\delta}_2$	SCAD	0.011	0.105	0.074	0.90	0.007	0.062	0.052	0.89
		ORACLE	0.008	0.082	0.074	0.93	0.001	0.052	0.052	0.96
	$\widehat{\delta}_3$	SCAD	0.001	0.017	0.019	0.99	0.002	0.013	0.013	0.95
		ORACLE	0.002	0.017	0.019	0.98	0.002	0.012	0.013	0.95
	$\widehat{\delta}_4$	SCAD	0.002	0.031	0.027	0.96	3e-4	0.019	0.019	0.94
		ORACLE	0.002	0.031	0.027	0.95	5e-4	0.020	0.019	0.95
	$\widehat{\delta}_5$	SCAD	1e-4	0.017	0.017	0.97	3e-4	0.012	0.012	0.95
		ORACLE	2e-4	0.017	0.017	0.96	3e-5	0.012	0.012	0.96

Table 2 presents the bias, the SD, the SE, and the coverage probability of the 95% confidence intervals (CP) of the estimators $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$. We observe that the estimates of $\widehat{\delta}_1$, $\widehat{\delta}_2$, and $\widehat{\delta}_3$ by our fused effects method perform as well as the oracle estimates. Moreover, our estimates have CPs close to the nominal level.

The third panel of Table 3 reports the MSE of the estimator $\widehat{\boldsymbol{a}}$, the bias, the SD, the SE, and the coverage probability of 95% confidence intervals (CP) of the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$. For the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$, all the methods have small biases. The means of dispersion parameter estimates $\widehat{\phi}$ by our method are 2.002 for $m = 100$ and 2.013 for $m = 50$, which are very close

TABLE 3 | The bias, standard deviation (SD), standard error (SE), and the coverage probability of 95% confidence intervals (CP) of the estimators for the smoothly clipped absolute deviation (SCAD), fixed effects (FE), random effects (RE), latent class (LC), and oracle estimators (ORACLE) in the Poisson, logistic, and overdispersed Poisson models.

			$\widehat{\alpha}$	$\widehat{\beta}_1$				$\widehat{\beta}_2$			
	m	Method	MSE	Bias	SD	SE	CP	Bias	SD	SE	CP
Poisson with three groups	50	SCAD	0.001	0.001	0.012	0.013	0.99	1e-5	0.014	0.013	0.95
		FE	0.011	0.001	0.012	0.013	0.98	1e-4	0.014	0.013	0.93
		RE	0.010	0.002	0.012	0.013	0.98	3e-4	0.014	0.013	0.92
		ORACLE	0.001	0.001	0.012	0.013	0.99	2e-5	0.014	0.013	0.95
		LC	0.001	0.001	0.012	0.013	0.99	6e-5	0.014	0.013	0.95
	100	SCAD	4e-4	5e-4	0.009	0.009	0.95	5e-5	0.009	0.009	0.95
		FE	0.011	4e-4	0.009	0.009	0.95	2e-5	0.009	0.009	0.96
		RE	0.010	0.001	0.009	0.009	0.95	5e-4	0.009	0.009	0.96
		ORACLE	4e-4	5e-4	0.009	0.009	0.95	5e-5	0.009	0.009	0.95
		LC	4e-4	0.001	0.009	0.009	0.95	1e-4	0.009	0.009	0.95
Logistic with three groups	50	SCAD	0.017	0.002	0.033	0.031	0.94	0.006	0.031	0.031	0.95
		FE	0.047	2e-4	0.033	0.032	0.94	0.008	0.032	0.032	0.95
		RE	0.040	0.002	0.033	0.031	0.94	0.005	0.032	0.031	0.95
		ORACLE	0.003	0.002	0.033	0.031	0.94	0.006	0.031	0.031	0.96
		LC	0.016	0.002	0.033	0.031	0.94	0.006	0.032	0.031	0.95
	100	SCAD	0.016	0.001	0.020	0.022	0.97	0.004	0.024	0.022	0.93
		FE	0.049	0.001	0.020	0.022	0.97	0.006	0.024	0.022	0.91
		RE	0.041	0.001	0.020	0.022	0.97	0.004	0.024	0.022	0.93
		ORACLE	0.001	0.001	0.020	0.022	0.97	0.004	0.023	0.022	0.93
		LC	0.015	0.001	0.020	0.022	0.97	0.004	0.023	0.022	0.92
Overdispersed Poisson with three groups	50	SCAD	0.009	0.002	0.026	0.026	0.94	0.003	0.027	0.026	0.96
		ORACLE	0.002	0.002	0.026	0.026	0.94	0.003	0.027	0.026	0.96
		FE	0.036	0.002	0.031	0.013	0.55	0.004	0.030	0.013	0.61
		RE	0.033	0.003	0.031	0.013	0.54	0.004	0.030	0.013	0.62
		FE.quasi	0.036	0.002	0.031	0.024	0.88	0.004	0.030	0.024	0.88
		RE.quasi	0.030	0.002	0.026	0.026	0.93	0.003	0.027	0.026	0.97
		LC	0.023	0.002	0.031	0.013	0.53	0.004	0.030	0.013	0.62
	100	SCAD	0.008	1e-4	0.018	0.018	0.96	0.001	0.017	0.018	0.98
		ORACLE	0.001	1e-4	0.018	0.018	0.96	0.001	0.017	0.018	0.97
		FE	0.035	0.001	0.020	0.009	0.63	2e-5	0.020	0.009	0.59
		RE	0.032	5e-4	0.020	0.009	0.63	4e-4	0.020	0.009	0.59
		FE.quasi	0.035	0.001	0.020	0.017	0.92	2e-5	0.020	0.017	0.89
		RE.quasi	0.029	2e-4	0.018	0.018	0.96	0.001	0.018	0.018	0.97
		LC	0.022	0.001	0.020	0.009	0.61	2e-5	0.020	0.009	0.60
Poisson with five groups	50	SCAD	0.011	4e-4	0.012	0.011	0.94	0.001	0.009	0.011	0.96
		FE	0.019	0.001	0.012	0.011	0.92	0.001	0.010	0.011	0.96
		RE	0.016	5e-4	0.012	0.011	0.92	0.001	0.010	0.011	0.96
		ORACLE	0.002	4e-4	0.012	0.011	0.93	0.001	0.009	0.011	0.96
		LC	0.050	4e-4	0.012	0.011	0.90	5e-4	0.009	0.011	0.96

(Continues)

TABLE 3 | (Continued)

			$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\beta}}_1$				$\widehat{\boldsymbol{\beta}}_2$			
	m	Method	MSE	Bias	SD	SE	CP	Bias	SD	SE	CP
	100	SCAD	0.006	2e-4	0.009	0.008	0.92	2e-4	0.008	0.008	0.94
		FE	0.017	2e-4	0.009	0.008	0.92	2e-4	0.008	0.008	0.94
		RE	0.016	2e-4	0.009	0.008	0.92	2e-4	0.008	0.008	0.94
		ORACLE	0.001	2e-4	0.009	0.008	0.92	2e-4	0.008	0.008	0.94
		LC	0.049	2e-4	0.008	0.008	0.94	2e-4	0.008	0.008	0.94

to its true value. As a result, our fused effects model with QL can estimate the correct mean-variance relationship, and our method has small biases in SE estimation, resulting in coverage probabilities close to the nominal level of 95%. In comparison, the RE and FE methods fitted by the regular Poisson regression yield the estimated SEs smaller than the sampling SD, resulting in undercoverage.

In addition to the Poisson likelihood, we also fit the data by the quasi-likelihood random effects (RE.quasi) and fixed effects (FE.quasi) methods by R function *glmer.nb* and *glm* with the quasi-Poisson family, respectively. The code is available on GitHub (https://github.com/liliwustl/quasi-likelihood). These models can yield SEs close to the SDs, leading to appropriate inference as shown for the CP. However, their MSEs for $\widehat{\boldsymbol{\alpha}}$ are close to those from the Poisson likelihood, which are much larger than those by our method. In conclusion, our method is more robust and has better performance than the RE and FE methods.

Example 4. In this example, we evaluate the performance of the proposed model with five groups. We generate the outcomes from the standard Poisson model. The covariates $\mathbf{X}$, $\boldsymbol{\beta}$, and n_i are generated from the same distributions as in Example 1. We assume that among the m provider effects a_i's, 10% are set to $\delta_1 = -2$, indicating providers performing "better than expected"; 10% are set to $\delta_2 = -1$, indicating providers performing "somewhat better than expected"; 10% are set to $\delta_4 = 1$, indicating providers performing "somewhat worse than expected"; 10% are set to $\delta_5 = 2$, indicating providers performing "worse than expected"; and the remaining 60% are set to 0, indicating providers performing "as expected."

The grouping results of $\widehat{a}_i$ using our method are presented at the bottom panel of Table 1, where the medians of $\widehat{K}$ over the 100 replicates are 5, which is the true number of subgroups, and the mean values are very close to 5. Furthermore, the RI values approach one. Therefore, our methods perform well in cases with a larger number of groups.

The bottom panel of Table 2 presents the bias, the SD, the SE, and the coverage probability of the 95% confidence intervals (CP) of the estimators $\widehat{\delta}_1$, $\widehat{\delta}_2$, $\widehat{\delta}_3$, $\widehat{\delta}_4$, and $\widehat{\delta}_5$. We observe that the estimates are generally accurate, with some minor biases in $\widehat{\delta}_1$ and $\widehat{\delta}_2$ along with their SEs.

The bottom panel of Table 3 reports the MSEs of the estimator $\widehat{\boldsymbol{\alpha}}$, the bias, the SD, the SE, and the coverage probability of 95% confidence intervals (CP) for the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$. From the bottom panel of Table 3, we observe that the MSE values of $\widehat{\boldsymbol{\alpha}}$ using SCAD are smaller than those of the random effects and fixed effects models. These results indicate that the proposed method performs well with more groups. For the estimators $\widehat{\boldsymbol{\beta}}$, all methods exhibit similar performance.

Compared to the Latent Class Model. In addition, our method can be considered as an alternative to the classical latent class model, where we use fusion penalty to "classify" providers into different latent groups. Per the suggestion of a reviewer, we compare our method to the latent class model (LC) (Aitkin 1996;1999), implemented in the R package *npmlreg*. The results are presented in the last panel of Tables 1 and 3. Of note, the results for the latent class model are not presented for Examples 3 and 4 in Table 2 due to its poor performance in identifying the correct number of classes (the median of $\widehat{K}$ is not the true number of group in Table 1). In summary, if the distributions of outcomes are correctly specified (as in Examples 1 and 2), both the latent class model and our method have good performances, though the latent class model is slightly better. However, when the distributional form is misspecified, as seen in Example 3, our quasi-likelihood method is more robust than the latent class model, yielding SEs close to the corresponding SDs and resulting in valid coverage probabilities in Table 2. Additionally, the latent class model might not perform well when dealing with an increased number of groups. In Example 4 with five groups, the latent class model struggles to identify the correct number of groups, as it can only identify the correct number of subgroups in less than 50% of the datasets (median of $\widehat{K}$ is not equal to 5, as shown in Table 1). In contrast, our method is consistently successful in selecting groups. This underscores the advantage of our model over the latent class model approach.

5 | Application

Kidney transplantation is the optimal solution for those suffering from the end stage renal disease (ESRD). According to SRTR, the first-year posttransplant graft survival, defined as the time to the first of death or graft failure within 1 year of transplant, has been widely used for regulatory monitoring of transplant centers (Kasiske et al. 2016; Salkowski et al. 2014). We would like to cluster the transplant centers by their performance on the first-year posttransplant graft survival.

Data from the U.S. Organ Procurement and Transplantation Network's national kidney transplant registry can be accessed at https://optn.transplant.hrsa.gov/data/. We restrict the study cohort to adults (age ≥ 18) of the kidney transplant recipients who underwent transplantation between January 1, 2009, and December 31, 2009. The analysis cohort included 5332 patients across 133 centers, with the number of patients per center ranging from 10 to 100. A significant portion of events occurred in the first year after kidney transplantation. Out of 5332 patients who received their kidney transplant in 2009, 485 experienced either death or graft failure within the subsequent year. The first-year posttransplant graft survival can be considered as a binary outcome (Jay and Schold 2017; Snyder et al. 2016). If a patient faces graft failure or death within a year of transplantation, the outcome is labeled as 1; and 0 otherwise. As in simulation setting II, we take the overdispersion parameter $\phi = 1$ for the binary data and do not adjust for overdispersion. We apply our method to investigate the risk factors of 1-year graft survival and assess the quality of care at transplant centers with respect to their graft outcomes.

The study cohort included 15 baseline characteristics of donors and recipients as follows: time on end-stage renal disease (ESRD; reference: <1 year), donor age (reference: 30–45), donor gender (male = 1, female = 0), donor body mass index (BMI, reference: normal), donor race (reference: white), donor history of hypertension (DON-HTN, yes = 1, no = 0), donor meeting expanded criteria (DON-EC, yes = 1, no = 0), recipient gender (male = 1, female = 0), recipient race (reference: white), recipient insulin-dependent diabetes (REC-DIAB Type I, yes = 1, no = 0), recipient noninsulin- dependent diabetes (REC-DIAB Type II, yes = 1, no = 0), recipient age at transplant (reference: 50–60), recipient BMI (reference: normal), recipient previous kidney transplant (REC-PREV-KI, yes = 1, no = 0), recipient total cold ischemia time (REC-COLD-ISCH, >20 h = 1, <20 h = 0).

We use the Akaike information criteria (AIC) to assess the performance of different methods, defined as AIC $= -2\ell(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\beta}}) + 2(k + p)$. A smaller value of AIC indicates better performance. Note here to make a valid comparison, $\ell(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\beta}})$ is the regular likelihood for the logistic model, with estimates of $\boldsymbol{a}$ and $\boldsymbol{\beta}$ obtained from the four methods. If the center effect a_i is treated as a random intercept, we obtain AIC = 3547.6. For the fixed effects model, the AIC value is 3283.0. If a_i is estimated by our fused effects model, the 133 centers are classified into three groups with AIC = 3064.2. If a_i is estimated by the latent class model, the 133 centers are classified into only one group with AIC = 3166.6. Our method leads to a notable improvement in the model fitting.

Table 4 reports the estimate (Est.), SE, and p-value of $\widehat{\boldsymbol{\beta}}$ for testing the significance of the coefficients by our fused effects model with the SCAD penalty (SCAD), the fixed effects model (FE), the random effects model (RE), and the latent class model (LC). Longer times on ESRD, donor history of hypertension, recipient noninsulin-dependent diabetes, recipient 60 or older at transplant, and total cold ischemia time show strongly significant effects by these methods with p-values less than 0.05. On the other hand, the obese donor BMI and donor meeting expanded criteria (DON-EC) are significant variables by our method, but nonsignificant in the FE and RE methods. The result of our method indicates that the obese donor BMI and DON-EC can increase the risk of 1-year posttransplant graft failure, which is consistent with the findings in the literature. The obesity status of a donor is known to influence organ quality and stands as an independent contributor to graft failure (Heinbokel et al. 2013; Kwan et al. 2016). The elevated risk of graft failure among recipients of kidneys from donors with a higher BMI might be attributed to preexisting structural damage, making these grafts more vulnerable to the ischemic, immune, and workload insults of kidney transplantation (Naik et al. 2020). In contrast, kidneys from expanded criteria donors (ECD) have been identified as a notable risk factor for graft survival (Ojo 2005). While opting for an ECD kidney can substantially reduce the waiting time for a transplant, comparisons of posttransplant outcomes between standard donor criteria and ECD kidneys revealed higher incidences of primary nonfunction, delayed graft function, and decreased graft survival with ECD kidneys (Ojo 2005).

We use the standardized event ratio (SER) to assess the performance of provider i relative to the population norm (e.g., national reference standard) (Estes et al. 2020). The SER for provider i is defined as the ratio of the observed number of events to the expected number of events:

$$\text{SER}_i = \frac{\sum_{j=1}^{n_i} \mu_{ij}}{\sum_{j=1}^{n_i} \mu_{ij,M}},$$

where $\mu_{ij} = g^{-1}(a_i + \mathbf{x}_{ij}^\top \boldsymbol{\beta})$ is the estimated number of events for patient j in provider i, and $\mu_{ij,M} = g^{-1}(a_M + \mathbf{x}_{ij}^\top \boldsymbol{\beta})$ is the corresponding expected number, with $a_M = \text{median}(a_1, \ldots, a_m)$ denoting the median provider effect, which serves as our population norm. This median effect is more robust than the average and has been widely applied in provider analysis (K. He et al. 2013). If the SER value is less than 1, it suggests the provider is outperforming the norm, whereas a value greater than 1 indicates underperformance in comparison to the norm. For our study, the estimated SER for overperforming, average, and underperforming centers were 0.592, 1.000, and 1.465, respectively.

Let $\widehat{\delta}_k$ be the common value of estimators $\widehat{a}_i$'s in group $\widehat{G}_k$ for $k = 1, 2, 3$. Table 5 reports the estimates the estimates $\widehat{\delta}_1$ and $\widehat{\delta}_3$ relative to a reference $\widehat{\delta}_2$. Additionally, the table provides the number of centers (num.) in each group, SE, p-value, and SER by our fused effects method. We identify 71% (95) of centers not different from the reference, 9% (12) of centers better than the reference, and 20% (26) of centers worse than the reference. The estimates $\widehat{\delta}_1$ and $\widehat{\delta}_2$ relative to the reference $\widehat{\delta}_2$ are −0.934 and 1.019, respectively.

6 | Discussion

In this work, we have presented a novel fused effects model to cluster the health care providers using the quasi-likelihood framework. Compared with the traditional random effects model and fixed effects model, our method not only correctly clusters providers but also obtains more accurate estimates of the true provider effects. Rather than relying on specific distributions, our approach is more flexible and robust. We have shown that the proposed method enjoys the oracle properties; namely, it

TABLE 4 | The estimates (Est.), standard error (SE), and *p*-value of $\widehat{\boldsymbol{\beta}}$ in Application.

	SCAD			FE			RE			LC		
Variable	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**
Time on ESRD												
<1 years	Ref.											
1–5 years	−0.059	0.071	0.405	−0.070	0.073	0.336	−0.067	0.067	0.318	−0.030	0.070	0.662
>5 years	0.231	0.067	0.001	0.229	0.072	0.001	0.211	0.067	0.002	0.223	0.066	0.001
Donor age												
<15	0.005	0.070	0.943	0.009	0.071	0.897	0.010	0.066	0.876	0.003	0.070	0.970
15–30	0.032	0.069	0.642	0.026	0.070	0.708	0.026	0.064	0.683	0.024	0.068	0.723
30–45	Ref.											
45–60	0.125	0.064	0.052	0.122	0.069	0.078	0.115	0.064	0.074	0.125	0.067	0.061
>60	0.037	0.069	0.597	0.036	0.072	0.621	0.030	0.068	0.663	0.050	0.068	0.463
Donor race												
White	Ref.											
Black	−0.002	0.049	0.969	−0.021	0.051	0.679	−0.018	0.048	0.709	−0.007	0.048	0.889
Asian	−0.004	0.053	0.944	−0.006	0.056	0.919	−0.007	0.052	0.889	−0.030	0.053	0.565
Donor gender	−0.060	0.049	0.221	−0.067	0.050	0.182	−0.064	0.047	0.176	−0.053	0.049	0.271
Donor BMI												
Normal	Ref.											
Under	0.082	0.062	0.183	0.083	0.062	0.181	0.076	0.058	0.188	0.060	0.062	0.328
Over	−0.016	0.059	0.790	−0.015	0.061	0.805	−0.016	0.056	0.778	−0.035	0.059	0.556
Obesity	0.114	0.056	0.040	0.106	0.057	0.064	0.098	0.053	0.066	0.116	0.055	0.035
DON-HTN	0.171	0.056	0.002	0.180	0.058	0.002	0.168	0.054	0.002	0.168	0.055	0.002
DON-EC	0.148	0.068	0.029	0.130	0.071	0.067	0.129	0.067	0.056	0.121	0.067	0.070
Recipient gender	0.029	0.050	0.569	0.024	0.052	0.644	0.021	0.048	0.658	0.018	0.050	0.710
Recipient race												
White	Ref.											
Black	0.005	0.051	0.927	−0.009	0.057	0.867	−0.004	0.053	0.943	0.020	0.050	0.696
Asian	−0.094	0.060	0.118	−0.105	0.064	0.102	−0.098	0.058	0.090	−0.105	0.059	0.076
REC-DIAB												
Type I	0.078	0.048	0.103	0.072	0.050	0.153	0.069	0.047	0.143	0.068	0.048	0.155
Type II	0.118	0.051	0.021	0.127	0.053	0.016	0.118	0.049	0.017	0.124	0.050	0.014
Recipient BMI												
Normal	Ref.											
Under	0.011	0.055	0.837	0.014	0.056	0.805	0.012	0.051	0.822	0.011	0.055	0.841
Over	0.052	0.063	0.410	0.064	0.066	0.329	0.058	0.061	0.336	0.059	0.063	0.341
Obesity	0.088	0.064	0.168	0.093	0.065	0.154	0.086	0.060	0.156	0.097	0.063	0.120
REC-PREV-KI	0.026	0.051	0.614	0.022	0.052	0.674	0.021	0.049	0.667	0.053	0.051	0.293
REC-COLD-ISCH	0.132	0.047	0.005	0.116	0.051	0.023	0.109	0.048	0.022	0.103	0.047	0.027

(Continues)

TABLE 4 | (Continued)

	SCAD			FE			RE			LC		
Variable	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**	**Est**	**SE**	***p*-value**
REC-AGE												
18–35	0.109	0.059	0.064	0.109	0.060	0.067	0.102	0.055	0.066	0.100	0.058	0.088
35–50	0.043	0.065	0.509	0.034	0.066	0.613	0.031	0.061	0.614	0.028	0.064	0.668
50–60	Ref.											
60–70	0.223	0.059	0.001	0.225	0.060	0.001	0.212	0.056	0.001	0.221	0.058	0.001
>70	0.175	0.048	0.001	0.171	0.051	0.001	0.160	0.048	0.001	0.156	0.048	0.001

TABLE 5 | Results of $\widehat{\delta}_i$ in the national kidney transplant study.

	$\widehat{\delta}_1$	$\widehat{\delta}_2$	$\widehat{\delta}_3$
Est	−0.934	Ref.	1.019
num	12	95	26
SE	0.243	0.064	0.095
p-value	<0.001	<0.001	<0.001
SER	0.592	1.000	1.465

performs as if the true group structure were known in advance, indicating that the clustering consistency property is achieved. The consistency and asymptotic normality of the estimators have been established. Finally, an efficient ADMM algorithm has been proposed to achieve fast convergence. The advantages of the proposed methods are demonstrated by simulations and an application to profile kidney transplant centers.

Our model bears a strong resemblance to latent class models but distinguishes itself through the use of a fusion penalty on pairwise provider effects. While conventional latent class models fit multiple models based on class numbers and employ AIC/BIC to determine the number of latent classes, we only fit a single model and use the fusion penalty to efficiently classify providers with similar characteristics into latent groups. Also, our strategy addresses potential challenges on identification and convergence in the classical latent class models, especially when dealing with a large number of latent classes (as in simulation Example 4).

There are some other fusion approaches for clustering. For example, B. He et al. (2021) used a sparse fusion penalization for high-dimensional finite mixture modeling, applying an additional sparsity penalty to select relevant features. Malsiner-Walli, Pauger, and Wagner (2018) proposed a Bayesian approach to automatically group levels of a categorical covariate that have the same effects. Austin, Pan, and Shen (2020) modeled heterogeneity in finite mixture models and utilized a group-truncated L_1 penalty to shrink the differences between estimated parameter vectors. While these methods primarily focus on estimating the homogeneity effects of covariates, our research seeks to identify subgroups of providers for clustered data. Our model also differs from Ma and Huang (2017), which was proposed for regression models with subject-specific intercepts.

There are several potential expansions for our method. First, it can be adapted to accommodate more sophisticated hierarchical data, such as a three-level structure comprising hospitals, physicians, and patients. Second, our method can be deployed for clustering regression coefficients in situations when parameters are partially heterogeneous across the subgroups. This would be particularly beneficial in tailoring precise medical treatment to heterogeneous subgroups of patients (Liu and Lin 2019; Liu et al. 2023; Ma and Huang 2020; Tang, Xue, and Qu 2021). Lastly, an advanced quasi-likelihood framework can be utilized to better elucidate the relationship between variance and mean, as discussed in Basu and Rathouz (2005), Chen et al. (2013), Nelder and Pregibon (1987), and Shi et al. (2023).

A modified BIC (Wang, Li, and Leng 2009) was employed in our methods. Bhattacharya and McNicholas (2014) introduced a LASSO-penalized BIC (LPBIC), designed specifically for optimal mixture model selection in high-dimensional multivariate data. It is important to note that the LPBIC is developed for clustering analysis, not for regression. In our approach, a fusion penalty is implemented to shrink the pairwise differences of individual provider effects in a regression model. In contrast, the LPBIC employed a lasso penalty to shrink the mean of multivariate Gaussian density. The framework in LPBIC thus cannot be applied to our methods in a straightforward way. Exploring alternative penalty functions for our model is an area of interest for future research.

Acknowledgments

This research is partly supported by NIH grants R21 EY031884, R21 EY033518, R21 AG063370, R21 AG068955, R01 AG081244, UL1 TR002345, R01 DK129539, Natural Science Foundation of Shandong Province ZR2021QA024, and China Postdoctoral Science Foundation 2021M691952.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Restrictions apply to the availability of real data, which was used under license for this study. The simulated data and code for implementing the proposed method are publicly available at https://github.com/liliwustl/quasi-likelihood.

Open Research Badges



This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to confidentiality issues.

References

Aitkin, M. 1996. "A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models." *Statistics and Computing* 6: 251–262.

Aitkin, M. 1999. "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models." *Biometrics* 55, no. 1: 117–128.

Austin, E., W. Pan, and X. Shen. 2020. "A New Semiparametric Approach to Finite Mixture of Regressions Using Penalized Regression via Fusion." *Statistica Sinica* 30, no. 2: 783.

Basu, A., and P. Rathouz. 2005. "Estimating Marginal and Incremental Effects on Health Outcomes Using Flexible Link and Variance Function Models." *Biostatistics* 6: 93–109.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers." *Foundations and Trends in Machine Learning* 3: 1–122.

Breslow, N. E., and D. G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88: 9–25.

Chen, J., L. Liu, D. Zhang, and Y. Shih. 2013. "A Flexible Model for the Mean and Variance Functions, With Application to Medical Cost Data." *Statistics in Medicine* 32: 4306–4318.

Estes, J. P., Y. Chen, D. Sentürk, et al. 2020. "Profiling Dialysis Facilities for Adverse Recurrent Events." *Statistics in Medicine* 39: 1374–1389.

Fan, J., and R. Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96: 1348–1360.

He, B., T. Zhong, J. Huang, Y. Liu, Q. Zhang, and S. Ma. 2021. "Histopathological Imaging-Based Cancer Heterogeneity Analysis via Penalized Fusion With Model Averaging." *Biometrics* 77, no. 4: 1397–1408.

He, K., J. D. Kalbfleisch, Y. Li, and Y. Li. 2013. "Evaluating Hospital Readmission Rates in Dialysis Facilities; Adjusting for Hospital Effects." *Lifetime Data Analysis* 19: 490–512.

Heinbokel, T., B. Floerchinger, A. Schmiderer, et al. 2013. "Obesity and Its Impact on Transplantation and Alloimmunity." *Transplantation* 96: 10–16.

Jay, C., and J. D. Schold. 2017. "Measuring Transplant Center Performance: The Goals Are Not Controversial but the Methods and Consequences Can Be." *Current Transplantation Reports* 4: 52–58.

Kasiske, B., N. Salkowski, A. Wey, A. Israni, and J. Snyder. 2016. "Potential Implications of Recent and Proposed Changes in the Regulatory Oversight of Solid Organ Transplantation in the United States." *American Journal of Transplantation* 16: 3371–3377.

Kwan, J. M., Z. Hajjiri, A. Metwally, P. W. Finn, and D. L. Perkins. 2016. "Effect of the Obesity Epidemic on Kidney Transplantation: Obesity Is Independent of Diabetes as a Risk Factor for Adverse Renal Transplant Outcomes." *PloS ONE* 11: e0165712.

Liang, K.-Y., and S. L. Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73, no. 1: 13–22.

Liu, L., M. Gordon, J. Miller, et al. 2021. "Capturing Heterogeneity in Repeated Measures Data by Fusion Penalty." *Statistics in Medicine* 40: 1901–1916.

Liu, L., K. He, D. Wang, et al. 2023. "Healthcare Center Clustering for Cox's Proportional Hazards Model by Fusion Penalty." *Statistics in Medicine* 42, no. 20: 3685–3698.

Liu, L., and L. Lin. 2019. "Subgroup Analysis for Heterogeneous Additive Partially Linear Models and Its Application to Car Sales Data." *Computational Statistics and Data Analysis* 138: 239–259.

Ma, S., and J. Huang. 2017. "A Concave Pairwise Fusion Approach to Subgroup Analysis." *Journal of the American Statistical Association* 112: 410–423.

Ma, S., J. Huang, Z. Zhang, and M. Liu. 2020. "Exploration of Heterogeneous Treatment Effects Via Concave Fusion." *The International Journal of Biostatistics* 16: 20180026.

Malsiner-Walli, G., D. Pauger, and H. Wagner. 2018. "Effect Fusion Using Model-Based Clustering." *Statistical Modelling* 18, no. 2: 175–196.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C., and J. Neuhaus. 2011. "Prediction of Random Effects in Linear and Generalized Linear Models Under Model Misspecification." *Biometrics* 67: 270–279.

McCulloch, C. E., and S. R. Searle. 2004. *Generalized, Linear, and Mixed Models*. Hoboken, NJ: John Wiley & Sons.

Naik, A. S., Y. Zhong, R. Parasuraman, et al. 2020. "The Temporal and Long-Term Impact of Donor Body Mass Index on Recipient Outcomes After Kidney Transplantation – A Retrospective Study." *Transplant International* 33: 59–67.

Nelder, J., and D. Pregibon. 1987. "An Extended Quasi-Likelihood Function." *Biometrika* 74: 221–232.

Normand, S. T., and D. M. Shahian. 2007. "Statistical and Clinical Aspects of Hospital Outcomes Profiling." *Statistical Science* 22: 206–226.

Ohlssen, D. I., L. D. Sharples, and D. J. Spiegelhalter. 2007. "A Hierarchical Modelling Framework for Identifying Unusual Performance in Health Care Providers." *Journal of the Royal Statistical Society A* 170: 865–890.

Ojo, A. 2005. "Expanded Criteria Donors: Process and Outcomes." *Seminars in Dialysis* 18: 463–468.

Racz, M., and J. Sedransk. 2010. "Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes." *Journal of the American Statistical Association* 105: 48–58.

Rand, W. M. 1971. "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association* 66: 846–850.

Salkowski, N., J. J. Snyder, D. A. Zaun, T. Leighton, A. K. Israni, and B. L. Kasiske. 2014. "A Scientific Registry of Transplant Recipients Bayesian Method for Identifying Underperforming Transplant Programs." *American Journal of Transplantation* 14: 1310–1317.

Shi, Y., H. Li, C. Wang, et al. 2023. "A Flexible Quasi-Likelihood Model for Microbiome Abundance Count Data." *Statistics in Medicine* 42, no. 25: 4632–4643.

Snyder, J., N. Salkowski, S. Kim, et al. 2016. "Developing Statistical Models to Assess Transplant Outcomes Using National Registries: The Process in the United States." *Transplantation* 100: 288–294.

Tang, X., F. Xue, and A. Qu. 2021. "Individualized Multidirectional Variable Selection." *Journal of the American Statistical Association* 116, no. 535: 1280–1296.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. "Sparsity and Smoothness via the Fused Lasso." *Journal of the Royal Statistical Society B* 67: 91–108.

Van Pilsum Rasmussen, S. E., A. G. Thomas, J. Garonzik-Wang, et al. 2018. "Reported Effects of the Scientific Registry of Transplant Recipients 5-Tier

Rating System on US Transplant Centers: Results of a National Survey." *Transplant International* 31, no. 10: 1135–1143.

Wang, H., B. Li, and C. Leng. 2009. "Shrinkage Tuning Parameter Selection With a Diverging Number of Parameters." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71, no. 3: 671–683.

Wedderburn, R. W. 1974. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method." *Biometrika* 61, no. 3: 439–447.

Wey, A., N. Salkowski, B. L. Kasiske, A. K. Israni, and J. J. Snyder. 2018. "A Five-Tier System for Improving the Categorization of Transplant Program Performance." *Health Services Research* 53, no. 3: 1979–1991.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.