Annals of the Institute of Statistical Mathematics (2021) 73:135–174
https://doi.org/10.1007/s10463-019-00745-z

Clustering of subsample means based on pairwise L1 regularized empirical likelihood

Quynh Van Nong[1] · **Chi Tim Ng**[2]

Received: 18 April 2018 / Revised: 30 October 2019 / Published online: 12 December 2019

Abstract
To classify a vast amount of strata or subsamples with unknown families of distributions according to their strata-means, a clustering approach is developed based on pairwise L_1 regularized empirical likelihood. Under such a clustering approach, all possible contradictory conclusions are ruled out automatically. On the contrary, the decision rules based on many existing pairwise comparison procedures can generate contradictory results. Moreover, under certain mild conditions, the proposed clustering method enjoys the consistency property that with probability going to one, all strata are classified correctly. An exterior point algorithm is presented for the clustering. The applications of the proposed methods are demonstrated using stock market data and microarray data of breast cancer patients.

Keywords Clustering · Empirical likelihood · Exterior point algorithm · Pairwise mean comparison · Pairwise L1 regularization

1 Introduction

The idea of L_1 regularization on the pairwise differences has been introduced in Pan et al. (2013) and Xie et al. (2008) for the clustering problems and in Zhu and Qu (2018) for the clustering of longitudinal curves. This paper generalizes such an idea from the clustering of observations from a sample to the clustering of strata-means from a population. To allow greater degree of robustness, the family of distribution

✉ Chi Tim Ng
easterlyng@gmail.com

Quynh Van Nong
quynhvandhsp@gmail.com

[1] Department of Mathematics, Thai Nguyen University of Education, Luong Ngoc Quyen Street, Thai Nguyen City, Vietnam

[2] Department of Statistics, Chonnam National University, 77, Yongbong-ro, Buk-gu, Gwangju 61186, South Korea

of the strata is not specified and the nonparametric empirical likelihood approach of Owen (1988) is adopted. In particular, consider the following two scenarios, (1) one-population case where the strata are classified according to the means and (2) two-population case where the strata are classified according to the population effects on the strata. In both medicine studies and social studies, it is interesting to classify age groups (strata) according to the sizes of the gender (population) effects.

Similar types of problems have been studied in the literature of multiple comparison and pairwise comparison and have found important applications in biology, social studies, and psychology, to name a few, Agresti et al. (2008), Gelman et al. (2012), Geman et al. (2004), and Lin et al. (2014). It is interesting to note that both clustering and pairwise comparison aim at determining if pairs of strata sharing the same mean though there are some philosophical differences. Classical multiple comparison and pairwise comparison methods, including Bonferroni's method in Bonferroni (1936), Tukey's method in Tukey (1949), and Duncan's multiple range procedure in Duncan (1955), are summarized in Dmitrienko et al. (2009), Miller (1981) and Hochberg and Tamhane (1987).

There is a clear difference between the clustering approach and the multiple comparison approach. Clustering approach can never generate contradictory conclusions. On the other hand, most existing multiple comparison (including pairwise comparison) methods are unable to rule out the possibility of concluding, e.g., $\mu_1 = \mu_2$, $\mu_2 = \mu_3$ but $\mu_1 \neq \mu_3$, where μ_1, μ_2, μ_3 are the strata-means. Though the concepts of coherence and consonance are introduced in the literature, e.g., Gabriel (1969), Sonnemann (2008), Zhao et al. (2010), Romano et al. (2011), possibility of drawing contradictory conclusions cannot be completely ruled out.

For the clustering problems, the pairwise L_1 penalty is applicable easily to any kinds of objective functions including the empirical likelihood discussed in this paper and the likelihood function as discussed in Pan et al. (2013), Xie et al. (2008), and Zhu and Qu (2018). On the contrary, there is a lack of literature discussing the multiple comparison methods under empirical likelihood approach. Though the empirical likelihood ratio test has been extensively studied in the literature, e.g., Qin and Lawless (1994), to the best of our knowledge, all these methods involve only one hypothesis. For example, Jing (1995) and Tsao and Wu (2006) study the empirical likelihood version two-sample mean test. Wu and Yan (2012) propose a weighted two-sample empirical likelihood method to test the difference of two-population means. Liu et al. (2008) constructs empirical likelihood method for the mean test of two d-dimensional samples that is done traditionally using Hotellings' T statistic. Cao and Van Keilegom (2006) develop an empirical likelihood-based test on the equality of the distributions of two populations.

To worst the situation, most existing multiple comparison methods rely heavily on model assumptions. For example, Tukey's method requires homogeneity of variance and normally distributed observations. The methods described in Geman et al. (2004) and Lin et al. (2014) are nonparametric. However, Geman et al. (2004) is designed specifically for the pairwise comparison of gene-expression-level data. It is not trivial to extend this method to the general pairwise comparison problems. The recent work Lin et al. (2014) focuses on the categorical data generated from multinomial distribution. To obtain a more general approach of nonparametric approach of pairwise difference

estimation, it is interesting to consider the empirical likelihood approach where each observed value constitutes one category.

The regularized likelihood (or penalized likelihood) has been widely studied in the context of regression analysis, e.g., Tibshirani (1996), Fan and Li (2001), Fan and Peng (2004), Fan and Lv (2010), Fu (1998), and Tibshirani and Taylor (2011). Moreover, the application of penalized empirical likelihood to regression analysis is also discussed in Tang and Leng (2010). In spite of these, its application to pairwise difference estimation is limited.

It is illustrated in this paper that the pairwise $L1$ regularized empirical likelihood approach allows us to control familywise error rate at a fixed level of say, 0.05. It is also interesting that when the sample size is large, it is possible to achieve consistency that all strata are classified correctly with high probability.

The rest of this article is organized as follows. In Sect. 2, the penalized estimation problem is described under empirical likelihood approach. In Sect. 3, the algorithm for solving penalized empirical likelihood maximization problem is presented. The consistency theory is established in Sect. 4. Simulation and real data examples are presented in Sects. 5 and 6, respectively, followed by the concluding remarks in Sect. 7, where possible extensions of the proposed method are discussed. The technical proofs are given in "Appendix."

2 Strata-mean clustering via regularized empirical likelihood

In this section, the ideas of pairwise $L1$ regularized empirical likelihood are demonstrated under two scenarios, (1) one-population case where the strata are classified according to the strata-means and (2) two-population case where the strata are classified according to sizes of the population effects.

2.1 L_1 regularized empirical likelihood estimation

First, consider the one-population case. Suppose that there are m independent strata of the same population. Denote the mean from a collection of independent random vectors $\{X_{ik}\}_{k=1}^{n_i}$ of the i-th strata by μ_i, where $i = 1, \ldots, m$. We are interested in sparse estimation of $\mu_i - \mu_j$. This means that zeros are preferred. Let $x_{i1}, x_{i2}, \ldots, x_{in_i}$ be the observations in the i-th strata. For simplicity, assume that there is no tie in the data. Let $p_i = (p_{i1}, \ldots, p_{in_i})$ and $p_{ik} = Pr(X_{ik} = x_{ik} | \text{ strata } i)$ fulfilling both $0 < p_{ik} < 1$ and $\sum_{k=1}^{n_i} p_{ik} = 1$. For the i-th strata, the empirical log-likelihood is

$$l_i(p_i) = \sum_{k=1}^{n_i} \log(p_{ik}).$$

Let

$$l(p_1, \ldots, p_m) = \sum_{i=1}^{m} \sum_{k=1}^{n_i} \log(p_{ik})$$

be the joint empirical log-likelihood function. Now, we propose to maximize the following regularized empirical likelihood function

$$Q(p_1, \ldots, p_m) = l(p_1, \ldots, p_m) - \lambda \sum_{1 \le i < j \le m} w_{ij} |\mu_i - \mu_j|, \tag{1}$$

subjected to the constraints

$$\sum_{k=1}^{n_i} p_{ik} = 1 \text{ and } \sum_{k=1}^{n_i} p_{ik} X_{ik} = \mu_i, \quad i = 1, \ldots, m,$$

where $\lambda > 0$ is a tuning parameter and $\mathbf{w} = (w_{ij})$ is a weight matrix that can be either fixed or computed from the data. If $w_{ij} = 1$ is chosen, the resulting pairwise L_1 penalty is a special case of the generalized Lasso in Tibshirani and Taylor (2011) and Zhang and Zhang (2012). In the context of regression analysis, it is well-documented (see, e.g., Zou 2006) that the Lasso shrinkage produces biased estimates when the coefficients are large. To reduce the bias of the Lasso and obtain a faster and more accurate algorithm, adaptive Lasso with weights $w_{ij} = \frac{1}{\left|\tilde{\mu}_i - \tilde{\mu}_j\right|^2}$ can be used, where $\tilde{\mu}_i, i = 1, 2, \ldots, m$ denote the initial estimates, for example, the sample means of the strata. The tuning parameter λ controls the errors in the test procedure. The choice of λ will be discussed later on in Sect. 2.2. Note that due to the non-differentiability of the L_1 penalty, exact zero is allowed in the solution.

The same idea can be extended to the two-population cases. Let $\mu_i^{(1)}$ and $\mu_i^{(2)}$ be the means of i-th strata from populations 1 and 2, respectively. We are interested in identifying strata-pairs (i, j) for which there are significant differences $(\mu_i^{(2)} - \mu_i^{(1)}) - (\mu_j^{(2)} - \mu_j^{(1)})$. Consider the reparameterization $a_i = \mu_i^{(2)} - \mu_i^{(1)}$. The penalized empirical log-likelihood function can be chosen as

$$\sum_{i=1}^{m} \sum_{k=1}^{n_i^{(1)}} \log\left(p_{ik}^{(1)}\right) + \sum_{i=1}^{m} \sum_{k=1}^{n_i^{(2)}} \log\left(p_{ik}^{(2)}\right) - \lambda \sum_{1 \le i < j \le m} w_{ij} |a_i - a_j|, \tag{2}$$

subject to the constraints

$$\sum_{k=1}^{n_i^{(1)}} p_{ik}^{(1)} = 1; \quad \sum_{k=1}^{n_i^{(2)}} p_{ik}^{(2)} = 1; \quad \sum_{k=1}^{n_i^{(2)}} x_{ik}^{(2)} p_{ik}^{(2)} - \sum_{k=1}^{n_i^{(1)}} x_{ik}^{(1)} p_{ik}^{(1)} = a_i,$$

for $i = 1, \ldots, m$.

2.2 Familywise error rate and Bayesian information criterion

Both familywise error rate (FWER) and Bayesian information criterion (BIC) can be employed to choose the tuning parameter λ.

References on FWER and multiple comparison can be found in Hochberg and Tamhane (1987) and Dmitrienko et al. (2009). It is common to control FWER at a pre-specified level of α, say, 0.05. However, due to the lack of explicit formula, we consider a bootstrapping approach of approximating FWER. See Efron and Tibshirani (1993) and Kleinman and Huang (2016) for the applications of bootstrapping methods to the multiple comparison. One can choose λ through simple grid-point search so that the bootstrap FWER is approximately 0.05. The bootstrap FWER can be obtained through the following steps:

1. Pool the data of all m-strata together. Re-sample from the pooled data without replacement.
2. Check if the number of detected cluster is greater than 1.
3. Repeat Step 1 and 2. The estimated FWER is the proportion of detecting more than one cluster.

Alternative to FWER, the concept of information criterion developed in model comparison and regression analysis is applicable too. Let $S(\lambda) = \{S_1, \ldots, S_c\}$ be the partition of $\{1, 2, \ldots, m\}$ obtained from the regularized estimation with λ. For the one-population problem, consider the following two definitions of BIC

$$\mathrm{BIC}_1(\lambda) = -2Q(\hat{p}_1, \ldots, \hat{p}_m) + \log(\log(m)) \cdot \sum_{s=1}^{c} \log\left(\sum_{i \in S_s} n_i\right),$$

$$\mathrm{BIC}_2(\lambda) = -2Q(\hat{p}_1, \ldots, \hat{p}_m) + \log(m) \cdot \sum_{s=1}^{c} \log\left(\sum_{i \in S_s} n_i\right),$$

where $\hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{in_i})$, $i = 1, \ldots, m$ are maximum regularized empirical likelihood estimator with tuning parameter λ. Choose λ so that BIC is minimized. Optimal λ can be obtained via grid-point search. For the two-population problems, one can replace n_i by $n_i^{(1)} + n_i^{(2)}$.

The term $\log(\log(m))$ in BIC_1 is adapted from the information criterion of Leng and Tang (2012) in the context of regression analysis. Similar definitions have also been discussed in the works, e.g., Fan and Tang (2013), Leng and Tang (2012), Wang et al. (2009), and Wang and Leng (2007). Moreover, Variyath et al. (2010) study the information criterion under empirical likelihood. Since the strata clustering problem can be reformulated as regularized likelihood estimation problem, it is reasonable to expect that the concepts of information criterion developed in regression analysis are also applicable. Our simulation experiences that will be discussed in Sect. 5 suggest that the greater penalty term $\log(m)$ in BIC_2 results in better performance of classification.

3 Algorithm

In this section, an iterative algorithm is presented to obtain the regularized estimation described in Sect. 2.

Pan et al. (2013) and Marchetti and Zhou (2014) propose algorithms for the clustering analysis. Comparing to the clustering analysis problem, the regularized empirical likelihood estimation involves additional constraints including

$$\sum_{k=1}^{n_i} \frac{x_{ik} - \mu_i}{n_i + \eta_i(x_{ik} - \mu_i)} = 0.$$

It is illustrated that some ideas of exterior point algorithm can be introduced to the so-called coordinate-wise algorithm of Pan et al. (2013), also Wu and Lange (2008), and Friedman et al. (2007).

3.1 One-population *m*-strata case

Let $\mu = (\mu_1, \ldots, \mu_m)'$ be the unknown parameters and $\eta = (\eta_1, \ldots, \eta_m)'$ be the Lagrange multipliers. The optimization problem of log-empirical likelihood (1) is equivalent to the optimization problem of

$$Q(\mu, \eta) = -\sum_{i=1}^{m}\sum_{k=1}^{n_i} \log\Big(n_i + \eta_i(x_{ik} - \mu_i)\Big) - \lambda \sum_{1\le i<j\le m} w_{ij}|\mu_i - \mu_j|, \qquad (3)$$

subjected to the constraints

$$\sum_{k=1}^{n_i} \frac{x_{ik} - \mu_i}{n_i + \eta_i(x_{ik} - \mu_i)} = 0, \quad i = 1, \ldots, m.$$

Consider the unconstrained minimization problem of

$$Q^*(\tau, \theta) = \sum_{i=1}^{m} f_i(\tau_i) + \frac{\beta}{2}\sum_{i<j}(\mu_i - \mu_j - \theta_{ij})^2 + \lambda \sum_{i<j} w_{ij}\,|\theta_{ij}|. \qquad (4)$$

where $\tau_i = (\mu_i, \eta_i)^T$ and

$$f_i(\tau_i) = \sum_{k=1}^{n_i} \log\left(n_i + \eta_i(x_{ik} - \mu_i)\right) + \frac{\beta}{2}\left(\sum_{k=1}^{n_i} \frac{x_{ik} - \mu_i}{n_i + \eta_i(x_{ik} - \mu_i)}\right)^2$$

for $i = 1, \ldots, m$. When $\beta \to \infty$, the solution to (4) approaches the solution to the original problem (3). The algorithm is as follows.

Step 1: Set $\beta = 1$.
Step 2: Fix $\tau_i, i = 1, \ldots, m$ and update θ_{ij}. Set

$$\theta_{ij}^{(\text{new})} = \begin{cases} \mu_i^{(\text{new})} - \mu_j^{(\text{new})} - \frac{\lambda w_{ij}}{\beta} & \text{if } \mu_i^{(\text{new})} - \mu_j^{(\text{new})} > \frac{\lambda w_{ij}}{\beta}, \\ \mu_i^{(\text{new})} - \mu_j^{(\text{new})} + \frac{\lambda w_{ij}}{\beta} & \text{if } \mu_i^{(\text{new})} - \mu_j^{(\text{new})} < -\frac{\lambda w_{ij}}{\beta}, \\ 0 & \text{if } \mu_i^{(\text{new})} - \mu_j^{(\text{new})} \in \left[-\frac{\lambda w_{ij}}{\beta}, \frac{\lambda w_{ij}}{\beta}\right]. \end{cases}$$

Step 3: Fix all θ_{ij} and update $\tau_1, \ldots, \tau_m$ at the same time. Here, Newton iteration can be performed on the function

$$Q^{***}(\tau, \theta) = \sum_{i=1}^{m} f_i(\tau_i) + \frac{\beta}{2} \sum_{i<j} (\mu_i - \mu_j - \theta_{ij})^2. \tag{5}$$

The formulas of $f_i(\tau_i)$ and their derivatives are given in "Appendix."
Step 4: Repeat steps 2 and 3 until converges.
Step 5: Increase β by doubling its current value.
Step 6: Repeat steps 2 and 5 until converges.

Note that Step 3 can be implemented with complexity $O(m)$. The inverse of Hessian matrix can be expanded as

$$H^{-1} = (D - \beta A A^T)^{-1} = D^{-1} + D^{-1} A \left(\frac{1}{\beta} I - A^T D^{-1} A\right)^{-1} A^T D^{-1}, \tag{6}$$

where D is $2m$-by-$2m$ diagonal matrix with all the entries are $\nabla^2_{\tau_i, \tau_i} f_i + B$ and

$$E_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ with } i = 1, \ldots, m; \quad A = \begin{bmatrix} E_1 \\ \vdots \\ E_m \end{bmatrix}_{2m \times 2}; \quad B = \begin{bmatrix} \beta \cdot m & 0 \\ 0 & 0 \end{bmatrix}.$$

Similar formula as (6) is used by some authors in the context of factor analysis, e.g., Lawley and Maxwell (1971) and Ng et al. (2015).

The tuning parameter λ is further obtained by a grid-point search with log-scale. $BIC(\lambda)$ is computed after minimizing $Q^*(\tau, \theta)$. Note that the computational burden of step 2 is $O(m^2)$. However, Newton step can be designed to be $O(mn)$. Therefore, the algorithm has iterative $O(m \max\{m, n\})$ complexity.

3.2 Two-population *m*-strata case

For the two-population cases, the constrained optimization problem is equivalent to minimizing

$$\sum_{i=1}^{m}\sum_{k=1}^{n_i^{(1)}}\log\left(n_i^{(1)}+\eta_i(x_{ik}^{(1)}-\mu_i^{(1)})\right)+\sum_{i=1}^{n^{(2)}}\sum_{k=1}^{n_i^{(2)}}\log\left(n_i^{(2)}+\eta_i(\mu_i^{(1)}-x_{ik}^{(2)}+a_i)\right)$$
$$+\lambda\sum_{i<j}w_{ij}|\theta_{ij}|, \tag{7}$$

subjected to the constraints

$$\sum_{k=1}^{n_i^{(1)}}\frac{x_{ik}^{(1)}-\mu_i^{(1)}}{n_i^{(1)}+\eta_i\left(x_{ik}^{(1)}-\mu_i^{(1)}\right)}=0 \quad \text{and} \quad \sum_{k=1}^{n_i^{(2)}}\frac{\mu_i^{(1)}-x_{ik}^{(2)}+a_i}{n_i^{(2)}+\eta_i\left(\mu_i^{(1)}-x_{ik}^{(2)}+a_i\right)}=0.$$

Similar to the one-population cases discussed in Sect. 3.1, the optimization problem for the two-population cases is equivalent to the unconstrained minimization problem of the objective function

$$Q^*(\tau,\theta)=\sum_{i=1}^{m}f_i(\tau_i)+\frac{\beta}{2}\sum_{i<j}(a_i-a_j-\theta_{ij})^2+\lambda\sum_{i<j}w_{ij}|\theta_{ij}|, \tag{8}$$

where $\tau_i=(\mu_i^{(1)},\eta_i,a_i)^T$ and

$$f_i(\tau_i)=\sum_{k=1}^{n_i^{(1)}}\log\left(n_i^{(1)}+\eta_i\left(x_{ik}^{(1)}-\mu_i^{(1)}\right)\right)+\sum_{k=1}^{n_i^{(2)}}\log\left(n_i^{(2)}+\eta_i\left(\mu_i^{(1)}-x_{ik}^{(2)}+a_i\right)\right)$$
$$+\frac{\beta}{2}\left(\sum_{k=1}^{n_i^{(1)}}\frac{x_{ik}^{(1)}-\mu_i^{(1)}}{n_i^{(1)}+\eta_i\left(x_{ik}^{(1)}-\mu_i^{(1)}\right)}\right)^2+\frac{\beta}{2}\left(\sum_{k=1}^{n_i^{(2)}}\frac{\mu_i^{(1)}-x_{ik}^{(2)}+a_i}{n_i^{(2)}+\eta_i\left(\mu_i^{(1)}-x_{ik}^{(2)}+a_i\right)}\right)^2,$$

for $i=1,\ldots,m$.

Procedure to solve (8) is similar that in Sect. 3.1, but with the following changes,

$$E_i=\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix}\text{ for }\ i=1,\ldots,m;\quad \eta=\begin{bmatrix}\beta\cdot m&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

and the updating formula for the elements of matrix θ,

$$\theta_{ij}^{(\text{new})} = \begin{cases} a_i^{(\text{new})} - a_j^{(\text{new})} - \frac{\lambda w_{ij}}{\beta} & \text{if } a_i^{(\text{new})} - a_j^{(\text{new})} > \frac{\lambda w_{ij}}{\beta}, \\ a_i^{(\text{new})} - a_j^{(\text{new})} + \frac{\lambda w_{ij}}{\beta} & \text{if } a_i^{(\text{new})} - a_j^{(\text{new})} < -\frac{\lambda w_{ij}}{\beta}, \\ 0 & \text{if } a_i^{(\text{new})} - a_j^{(\text{new})} \in \left[-\frac{\lambda w_{ij}}{\beta}, \frac{\lambda w_{ij}}{\beta}\right]. \end{cases}$$

4 Consistency theory

First, some notation is introduced. For simplicity, consider only the balanced cases, i.e., $n_1 = n_2 = \cdots = n_m = n$ for the one-population problem and $n_1^{(1)} = n_2^{(1)} = \cdots = n_m^{(1)} = n_1^{(2)} = n_2^{(2)} = \cdots = n_m^{(2)} = n$ for the two-population problem. Both m and n are allowed going to infinity. The results presented in this section can be generalized easily to the unbalanced cases if the ratio between maximum and minimum sample sizes is bounded above and below by nonzero constants.

For the one-population cases, without loss of generality suppose that the strata indexes $1, 2, \ldots, m$ have been rearranged so that the true means are

$$\begin{aligned} \mu_1^0 = \mu_2^0 = \cdots = \mu_{b_1}^0 < \mu_{b_1+1}^0 = \mu_{b_1+2}^0 = \cdots = \mu_{b_2}^0 < \cdots \\ < \mu_{b_{c-1}+1}^0 = \mu_{b_{c-1}+2}^0 = \cdots = \mu_{b_c}^0, \end{aligned}$$

where c is true number of clusters and $0 = b_0 < b_1 < b_2 < \cdots < b_c = m$. Set $m_1 = b_1, m_2 = b_2 - b_1, \ldots, m_c = b_c - b_{c-1}$. For the two-population cases, one can replace the notation μ by a.

Convention 1 *Let $\eta_i(\mu_i)$ be defined implicitly via*

$$\sum_{k=1}^{n} \frac{x_{ik} - \mu_i}{n + \eta_i(x_{ik} - \mu_i)} = 0.$$

Let $u_n^{(s)}, s = 1, \ldots, c$ be a sequence of constants so that

$$\max_{i=b_{s-1}+1,\ldots,b_s} n^{-1/2} \eta_i\left(\mu_{(s)}^0\right) \le O_p\left(u_n^{(s)}\right),$$

$$\max_{i=b_{s-1}+1,\ldots,b_s} n^{-1} \sup_{\mu \in \Xi} \left| \frac{d}{d\mu_i} \eta_i\left(\mu_{(s)}^0\right) \right| \le O_p\left(u_n^{(s)}\right),$$

and

$$\max_{i,j=b_{s-1}+1,\ldots,b_s} n^{-1/2} \sup_{\mu \in \Xi} \left| \frac{d}{d\mu_i} \eta_i(\mu) - \frac{d}{d\mu_i} \eta_j(\mu) \right| \le O_p\left(u_n^{(s)}\right).$$

The sequence $u_n^{(s)}$ can be bounded using extreme-value-distribution theory, see for example (Fisher and Tippett, 1928; Gnedenko, 1943). In the special cases where $Z_1, Z_2, \ldots, Z_m$ are independent $N(0, 1)$ random variables, McCormick (1980) shows

that $\max_i |Z_i| = O_p(\sqrt{2\log m})$. In the general cases where finite moment of order $\delta > 0$ exists, the inequality

$$\max_i |Z_i| \le \left(\sum_{i=1}^{m} |Z_i|^{\delta} \right)^{1/\delta}$$

can also be used to obtain the bound $u_n^{(s)}$.

Theorem 1 *Let $\Xi \subset R^m$ be a compact set containing μ^0 as an interior point. Suppose that c is finite. Consider the following conditions.*

(A1) *For $s = 1, 2, \ldots, c$, the weights w_{ij} satisfy*

$$\max_{i,j=b_{s-1}+1,\ldots,b_s} \left\{ w_{ij}^{-1} \sum_{t \ne s} \sum_{e=b_{t-1}+1}^{b_t} |w_{ie} - w_{je}| \right\} = o_p(m),$$

$$\max_{i,j=b_{s-1}+1,\ldots,b_s} \left\{ \sum_{t \ne s} \sum_{e=b_{t-1}+1}^{b_t} |w_{ie}| \right\} = O_p(m).$$

(A2) $\lambda m n^{-1} \to 0$ *and* $\lambda n^{-1/2} \min_{1 \le s \le c} \{m_s / u_n^{(s)}\} \to \infty$,
(A3) $\min_{i,j} \{|\mu_i^0 - \mu_j^0|, |\mu_i^0 - \mu_j^0| \ne 0\} > C$ *for some constant* $C > 0$.

Then, with probability going to one, for sufficiently large β, we have

(i) *for all* $s = 1, 2, \ldots, c$, $\max_{i,j=b_{s-1}+1,\ldots,b_s} |\hat{\mu}_i - \hat{\mu}_j| \le \lambda/\beta$,
(ii) $\hat{\theta}_{ij} = 0$ *if* $\mu_i^0 = \mu_j^0$, $1 \le i < j \le m$ *and* $\hat{\theta}_{ij} = \mu_i^0 - \mu_j^0 + o(1)$ *otherwise.*
(iii) *for all* $i = 1, 2, \ldots, m$, $|\hat{\mu}_i - \mu_i^0| = o(1)$.

Here, $o(1)$ in (ii) and (iii) can further be bounded by $\epsilon + \lambda/\beta$ and ϵ, respectively, where ϵ is any quantity dominating $\max\{(mn)^{-1/2}, \lambda m n^{-1}\}$.

In the limiting case $\beta = \infty$, result (i) becomes

(i') *for all* $s = 1, 2, \ldots, c$, $\max_{i,j=b_{s-1}+1,\ldots,b_s} |\hat{\mu}_i - \hat{\mu}_j| = 0$.

The proofs will be given in "Appendix." In the LASSO cases $w_{ij} = 1$, clearly,

$$\max_{i,j=b_{s-1}+1,\ldots,b_s} \left\{ w_{ij}^{-1} \sum_{t \ne s} \sum_{e=b_{t-1}+1}^{b_t} |w_{ie} - w_{je}| \right\} = 0,$$

$$\max_{i,j=b_{s-1}+1,\ldots,b_s} \left\{ \sum_{t \ne s} \sum_{e=b_{t-1}+1}^{b_t} |w_{ie}| \right\} = m.$$

Thus, Condition (A1) holds. Condition (A2) gives the range of λ so that consistency of the test holds. Condition (A3) requires that the pairwise difference is not too small to be detected.

We have similar results for the two-population cases.

Theorem 2 *Let $\Xi \in R^m$ be a compact set containing a^0 as an interior point. Consider the same conditions as Theorem 1. Then, with probability going to one, we have*

(i) *for all* $s = 1, 2, \ldots, c$, $\max_{i,j=b_{s-1}+1,\ldots,b_s} |\hat{a}_i - \hat{a}_j| \leq \lambda/\beta$

(ii) $\hat{\theta}_{ij} = 0$ *if* $a_i^0 = a_j^0$, $1 \leq i < j \leq m$ *and* $\hat{\theta}_{ij} = a_i^0 - a_j^0 + o(1)$ *otherwise.*

(iii) *for all* $i = 1, 2, \ldots, m$, $|\hat{a}_i - a_i^0| = o(1)$.

In the limiting case $\beta = \infty$*, result (i) becomes*

(i') *for all* $s = 1, 2, \ldots, c$, $\max_{i,j=b_{s-1}+1,\ldots,b_s} |\hat{a}_i - \hat{a}_j| = 0$.

The proofs of Theorem 2 are very similar to those of Theorem 1 and are omitted for brevity.

5 Simulation studies

Simulation studies are conduced in this section to evaluate the finite-sample performance of the proposed strata clustering method. One-population m-strata problem is investigated in Example 1, and two-population counterpart is discussed in Example 2.

All experiments in the examples are repeated for 100 times. The performances are measured in terms of two criteria. The first one is the mean misclassification, i.e., the proportion of correct conclusions among $m(m-1)/2$ hypotheses. The second one is C_s, $s = 1, 2, \ldots, c$, the cumulated number of strata of the first biggest s clusters.

Each simulated data are tested using the following three different choices of λ that based on BIC_1, BIC_2, and FWER, respectively. The FWER is controlled at the level of 0.05 using the method as described in Sect. 2.2.

Example 1 Consider one-population problems with two cases: balanced data and unbalanced data.

Balanced data cases: We assume that all the strata have the same sample size. Under this assumption, the influences of model, strata-variance, number of true clusters, and distances between cluster means are evaluated. Two models are compared, Chi-square distribution χ^2_μ, and Gamma distribution $\Gamma(\alpha = \mu^2/\upsilon, \beta = \mu/\upsilon)$ with variances υ fixed at 1, 2, 3, 4. Note that Chi-square distribution is a special case of Gamma distribution with mean-dependent variance. Consider two levels for the number of strata, $m = 40, 200$ and four levels of strata-sample size n, 20, 50, 100, 500. In the simulation, each cluster contains equal number of strata. The detailed results are reported in Tables 1, 2, 3, 4, 5, and 6.

Tables 1 and 2 summarize the results of misclassification for Chi-square distributed and Gamma distributed (with $\upsilon = 1$) data, respectively. Three different simulation settings are used, one cluster with mean 4, two clusters with means 4 and 8, and four clusters with means 3, 5, 7, 9. In general, all three criteria BIC_1, BIC_2, and FWER perform similarly well under the large sample size case $n = 500$. It can be observed that for small sample cases, BIC_2 and FWER tend to perform better than BIC_1, particularly in the one-cluster cases. In addition, the misclassification tends to be smaller

Table 1 Misclassification rate for Chi-square distribution

m	n	Cluster's means	BIC_1	BIC_2	FWER
40	20	4	0.4367436	0.05130769	0.0025
		(4, 8)	0.1870897	0.0665	0.02317949
		(3, 5, 7, 9)	0.1732179	0.1712436	0.1857949
	50	4	0.2549872	0.006397436	0.005
		(4, 8)	0.07524359	0.007602564	0.009871795
		(3, 5, 7, 9)	0.09183333	0.07003846	0.06597436
	100	4	0.1790128	0.000474359	0.002
		(4, 8)	0.02544872	0.0009102564	0.006948718
		(3, 5, 7, 9)	0.04015385	0.01952564	0.03453846
	500	4	0.06002564	0.006794872	0.0015
		(4, 8)	0.008448718	0.002282051	0.0110641
		(3, 5, 7, 9)	0.004358974	0.001602564	0.02916667
200	20	4	0.8168121	0.0006984925	4e−04
		(4, 8)	0.3328648	0.05642111	0.01423769
		(3, 5, 7, 9)	0.2055156	0.1834151	0.2677653
	50	4	0.5475382	0.000601005	2e−04
		(4, 8)	0.09393015	0.006221608	0.003516583
		(3, 5, 7, 9)	0.1195603	0.07835829	0.05212362
	100	4	0.3462849	0.001373869	5e−04
		(4, 8)	0.02475528	0.002126633	0.002432161
		(3, 5, 7, 9)	0.03880804	0.01506533	0.01520704
	500	4	0.06885678	0.0003040201	2e−04
		(4, 8)	0.003083417	0.0006251256	0.002234673
		(3, 5, 7, 9)	0.02048241	0.0004623116	0.006120101

under the fixed-variance Gamma cases. This suggests differences between varying-variance cases and fixed-variance cases. Table 3 compares Gamma distributions fixed at different levels of variance. It can be seen that misclassification rates for both BIC_1 and BIC_2 increase as the variance increases. However, BIC_2 always gives smaller misclassification than BIC_1. To summarize, the variance of the distribution plays an important role to the finite-sample performance of the classification. Therefore, a good control of the variance is essential to the good performance.

Table 4 shows the influence of the distances between cluster means. In the cases where the cluster means are closer to each other, the misclassification rates are higher when the sample size is small under both large and small variance cases.

Tables 5 and 6 show $C_s, s = 1, 2, \ldots, c$, the cumulative proportion of strata of the first s clusters. In the ideal cases without misclassification, $C_1 = 100\%$ for one-cluster cases, ($C_1 = 50\%, C_2 = 100\%$) for two-cluster cases, and ($C_1 = 25\%, C_2 = 50\%, C_3 = 75\%, C_4 = 100\%$) for the four-cluster cases. It can be seen that for both $m = 40$ and $m = 200$ cases, the misclassification decreases as the sample size

Table 2 Misclassification rate for Gamma distribution with $\upsilon = 1$

m	n	Cluster's means	BIC_1	BIC_2	FWER
40	20	4	0.4746923	0.02161538	0.034
		(4, 8)	0.1462051	0.004602564	0.03101282
		(3, 5, 7, 9)	0.05211538	0.007230769	0.1952179
	50	4	0.2456026	0.007166667	0.04567949
		(4, 8)	0.06691026	0.002961538	0.03623077
		(3, 5, 7, 9)	0.01202564	0.001653846	0.08307692
	100	4	0.1712949	0.002871795	5e−04
		(4, 8)	0.03139744	0.0007435897	0.005282051
		(3, 5, 7, 9)	0.005602564	2.564103e−05	0.03297436
	500	4	0.03264103	0.00825641	0
		(4, 8)	0.01879487	0.005679487	0.0007307692
		(3, 5, 7, 9)	0.002692308	0.0005641026	0.004269231
200	20	4	0.7277704	0.004079899	3e−04
		(4, 8)	0.3047025	0.002030151	0.01598291
		(3, 5, 7, 9)	0.07890251	0.004747236	0.2043121
	50	4	0.4124578	0.0005864322	0
		(4, 8)	0.140096	0.0005929648	0.003317588
		(3, 5, 7, 9)	0.007459296	0.0006080402	0.06458844
	100	4	0.1362683	0.0009899497	0.01403719
		(4, 8)	0.07299849	0.0002909548	0.01018543
		(3, 5, 7, 9)	0.004992462	0.0003050251	0.01538191
	500	4	0.009832663	0.0006944724	0
		(4, 8)	0.03094472	0.0004276382	0.003073367
		(3, 5, 7, 9)	0.003395477	0.0001979899	0.006879397

increases. In the one-cluster cases, the cumulative proportions C_s of FWER are always closer to the ideal proportion than those of BIC_1 and BIC_2. However, the performances of BIC_1 and BIC_2 are comparable to that of FWER.

Unbalanced data cases: To see the influence of the unbalanced strata-sample sizes on the performance of the proposed method, the strata-sample sizes are generated randomly from uniform distributions with ranges (20, 40), (90, 110), (190, 210), respectively. The data are generated from the Chi-square distribution, and two levels $m = 40, 100$ are considered for the number of strata. The settings of the cluster means are the same as those in Case 1. The results are summarized in Table 7. It is clear from the table that the proposed clustering method is also applicable to the unbalanced cases. Similar to the results of the balanced cases in Table 1, the two criteria BIC_2, FWER outperform BIC_1 in general. However, BIC_2 tends to have less misclassification in the small strata-sample-size cases.

Example 2 In the second example, consider two-population problems. In both populations, there are m-strata with sample sizes $n^{(1)}$ in Population One and $n^{(2)}$ in

Table 3 Misclassification rate for Gamma distribution with different υ

m	n	Cluster's means	$\nu = 2$		$\nu = 3$		$\nu = 4$	
			BIC_1	BIC_2	BIC_1	BIC_2	BIC_1	BIC_2
40	20	4	0.451359	0.02408974	0.4660256	0.02578205	0.4706026	0.02526923
		(4, 8)	0.1483462	0.006410256	0.1349359	0.007038462	0.1646538	0.01357692
		(3, 5, 7, 9)	0.05585897	0.01887179	0.06835897	0.03576923	0.09364103	0.06123077
	100	4	0.1592179	0.0005858974	0.1637179	0.000525641	0.1702821	0.004038462
		(4, 8)	0.03442308	0.00233333	0.02339744	0.001064103	0.02705128	0.002602564
		(3, 5, 7, 9)	0.009602564	0.000641025	0.009089744	0.001269231	0.01047436	0.002448718
200	20	4	0.8215543	0.004472864	0.8060568	0.001694472	0.8158618	0.009677387
		(4, 8)	0.2919844	0.00359397	0.2724719	0.005766332	0.2842055	0.007885427
		(3, 5, 7, 9)	0.0788	0.01375578	0.1017633	0.02948141	0.121192	0.0612402
	100	4	0.2567869	6e−04	0.298404	9.949749e−05	0.3319698	0.001092965
		(4, 8)	0.0355392	0.000496984	0.03892111	0.000918593	0.02693769	0.0001703518
		(3, 5, 7, 9)	0.003664824	0.0004979899	0.004663317	0.0009854271	0.005961809	0.001965829

Table 4 Misclassification rate for Gamma distribution with different sample means

m	n	Cluster's means	$\upsilon = 1$		$\upsilon = 3$	
			BIC_1	BIC_2	BIC_1	BIC_2
40	20	(1, 1.5, 2, 2.5)	0.2002692	0.2033462	0.2684487	0.3165641
		(3, 4.5, 6, 7.5)	0.05761538	0.01602564	0.1162821	0.09348718
	100	(1, 1.5, 2, 2.5)	0.04974359	0.03205128	0.1576795	0.1598846
		(3, 4.5, 6, 7.5)	0.007653846	0.0005641026	0.01528205	0.002846154
200	20	(1, 1.5, 2, 2.5)	0.2199281	0.2110307	0.2535035	0.2603392
		(3, 4.5, 6, 7.5)	0.0956201	0.01124623	0.1585779	0.09236482
	100	(1, 1.5, 2, 2.5)	0.0744593	0.03372312	0.1876482	0.1654226
		(3, 4.5, 6, 7.5)	0.004465829	0.0005065327	0.01231508	0.00278995

Population Two. The strata-means in Population One are sampled from $\{3, 5, 7, 9\}$ randomly. The differences a_i are then added to the strata-means of Population Two. In the simulation, each cluster contains equal number of strata. Within the same cluster, a_i are the same but μ_i are allowed different. Consider two levels of number of strata, $m = 40, 200$ and six settings of sample sizes, $(n^{(1)}, n^{(2)}) = (25, 25); (25, 50); (25, 100); (50, 50); (50, 100); (100, 100)$. The variance of Gamma distribution υ is set to 1.

Here, we only demonstrate the misclassification results of Gamma distribution with $\upsilon = 1$. The results are summarized in Table 8. The findings about the influences of model, strata-variance, number of true clusters, and distances between cluster means are similar to those in Example 1 and are not included for brevity. This example further confirms that the proposed penalized empirical likelihood method can be applied to the two-population m-strata classification problems.

6 Real data example

To demonstrate the applications of the regularized empirical likelihood method to the strata classification problems, we consider the following three datasets: chronic myelogenous leukemia survival data from Hehlmann et al. (1994), APPL (Apple) daily stock price data from Year 1981 to Year 2017 available at Yahoo finance, and the breast cancer data in Van't Veer et al. (2002). In the first two examples, both one-population m-strata problem and two-population m-strata problem are studied. The last example is a two-population m-strata problem. These three examples cover small $m = 3$ case, medium $m = 37$ case, and large m case. For the large m case, $m = 24{,}480$ and a splitting strategy is adopted so that each estimation involves only < 300 strata.

6.1 Chronic myelogenous leukemia survival data

This is a small m example. The original data contain 507 observations on the 7 variables (see Hehlmann et al. (1994) for details). Here, the variables "treatment," "gender," and

Table 5 Cumulative proportion of number groups in k-th cluster for $m = 40$

Cluster means	Number of groups	1	2	3	4	5	≥ 6
$n = 20$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.70150	0.85550	0.91950	0.95500	0.97525	1
	BIC_2	0.97	0.99350	0.99850	0.99975	1	1
	FWER	0.99875	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.41550	0.75025	0.85300	0.90975	0.96550	1
	BIC_2	0.48575	0.92675	0.97125	0.98825	0.99525	1
	FWER	0.502125	0.98406	0.99725	0.99950	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.2455	0.43050	0.57625	0.77175	0.83575	1
	BIC_2	0.30050	0.52325	0.68925	0.79575	0.86775	1
	FWER	0.41725	0.69	0.86875	0.95575	0.98425	1
$n = 50$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.84025	0.92325	0.96225	0.98100	0.99100	1
	BIC_2	0.99875	1	1	1	1	1
	FWER	0.9975	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.47652	0.90700	0.95500	0.97800	0.99050	1
	BIC_2	0.5	0.99425	0.99900	1	1	1
	FWER	0.49950	0.98975	0.99925	1	1	1

Table 5 continued

Cluster means	Number of groups	1	2	3	4	5	≥ 6
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.24050	0.45600	0.64	0.78475	0.91275	1
	BIC_2	0.26300	0.49875	0.70975	0.87875	0.93650	1
	FWER	0.26325	0.50250	0.71725	0.88875	0.94775	1
$n = 100$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.88075	0.95350	0.98250	0.99325	0.99750	1
	BIC_2	1	1	1	1	1	1
	FWER	0.99900	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.49575	0.97225	0.99150	0.99800	0.99950	1
	BIC_2	0.5	0.99925	1	1	1	1
	FWER	0.49975	0.99300	0.99950	1	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.24875	0.48875	0.70800	0.90150	0.95025	1
	BIC_2	0.25075	0.49950	0.74	0.95700	0.98700	1
	FWER	0.24975	0.49625	0.72225	0.91925	0.96625	1
$n = 500$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.96300	0.99350	0.99875	1	1	1
	BIC_2	1	1	1	1	1	1
	FWER	0.99925	1	1	1	1	1

Table 5 continued

Cluster means	Number of groups	1	2	3	4	5	≥ 6
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.49950	0.99175	0.99675	0.99875	0.99975	1
	BIC_2	0.5	1	1	1	1	1
	FWER	0.49900	0.98925	0.99875	1	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.25	0.5	0.74875	0.99150	0.99825	1
	BIC_2	0.25	0.5	0.75	0.99925	1	1
	FWER	0.24950	0.49400	0.72675	0.93000	0.96775	1

Table 6 Cumulative proportion of number groups in k-th cluster for $m = 200$

Cluster means	Number of groups	1	2	3	4	5	≥ 6
$n = 20$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.41200	0.57055	0.66985	0.73675	0.78665	1
	BIC_2	0.99970	0.99990	0.99995	1	1	1
	FWER	0.9998	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.31745	0.55835	0.65170	0.71540	0.76175	1
	BIC_2	0.48845	0.95670	0.97300	0.98085	0.98630	1
	FWER	0.50070	0.99200	0.99615	0.99820	0.99935	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.16395	0.28995	0.39495	0.48175	0.55490	1
	BIC_2	0.21905	0.39245	0.52855	0.63380	0.71140	1
	FWER	0.56575	0.80715	0.91100	0.95530	0.97395	1
$n = 50$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.67155	0.78235	0.84910	0.88880	0.91640	1
	BIC_2	0.99990	1	1	1	1	1
	FWER	0.99990	1	1	1	1	1

Table 6 continued

Cluster means	Number of groups	1	2	3	4	5	≥ 6
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.46915	0.90465	0.93540	0.95290	0.96515	1
	BIC_2	0.49915	0.99500	0.99795	0.99945	0.99990	1
	FWER	0.49965	0.99670	0.99930	0.99990	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.22535	0.42720	0.60195	0.74545	0.80180	1
	BIC_2	0.24450	0.47585	0.68750	0.86590	0.90930	1
	FWER	0.25680	0.49980	0.73100	0.93345	0.96070	1
$n = 100$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.81240	0.87525	0.91245	0.93850	0.95505	1
	BIC_2	0.99995	1	1	1	1	1
	FWER	0.99975	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.49450	0.97930	0.98750	0.99210	0.99455	1
	BIC_2	0.49985	0.99830	0.99965	0.99990	1	1
	FWER	0.49990	0.99765	0.99965	1	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.24660	0.48700	0.71805	0.93445	0.95205	1
	BIC_2	0.25030	0.49820	0.74210	0.97705	0.98980	1
	FWER	0.25	0.49790	0.74070	0.97460	0.98300	1

Table 6 continued

Cluster means	Number of groups	1	2	3	4	5	≥ 6
$n = 500$							
4	Theoretical	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.97815	0.98860	0.99380	0.99655	0.99770	1
	BIC_2	1	1	1	1	1	1
	FWER	0.99990	1	1	1	1	1
(4, 8)	Theoretical	0.5 (50%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.49980	0.99750	0.99935	0.99960	0.99975	1
	BIC_2	0.5	1	1	1	1	1
	FWER	0.5	0.99790	0.99975	1	1	1
(3, 5, 7, 9)	Theoretical	0.25 (25%)	0.5 (50%)	0.75 (75%)	1 (100%)	1 (100%)	1 (100%)
	BIC_1	0.24985	0.49965	0.74890	0.99655	0.99835	1
	BIC_2	0.25	0.5	0.75	0.99935	0.99995	1
	FWER	0.24995	0.49930	0.74715	0.98935	0.99450	1

Table 7 Misclassification rate for Chi-square distributed unbalanced data

m	Range of n	Cluster's means	BIC_1	BIC_2	FWER
40	(20, 40)	4	0.4135385	0.02601285	0.01569231
		(4, 8)	0.144	0.07302564	0.2457051
		(3, 5, 7, 9)	0.1374231	0.1355513	0.2061282
	(90, 110)	4	0.1269615	0.03821795	0.01070513
		(4, 8)	0.03244872	0.001692308	0.02278205
		(3, 5, 7, 9)	0.03779487	0.02491026	0.03094872
	(190, 210)	4	0.1423846	5e−04	0.004423077
		(4, 8)	0.02784615	0.004923077	0.01953846
		(3, 5, 7, 9)	0.01152564	0.004474359	0.009602564
100	(20, 40)	4	0.5886242	0.02130707	0.0163899
		(4, 8)	0.2065495	0.0858101	0.2341374
		(3, 5, 7, 9)	0.1513556	0.1437556	0.2038465
	(90, 110)	4	0.2954929	0.04849495	0.001775758
		(4, 8)	0.04690909	0.02019798	0.01463838
		(3, 5, 7, 9)	0.04308081	0.02912929	0.02160202
	(190, 210)	4	0.2896646	0.008967677	0.001981818
		(4, 8)	0.0223899	0.003919192	0.006410101
		(3, 5, 7, 9)	0.01474141	0.00420404	0.004072727

"time survival" are considered. There is a total of three treatment groups. In the original data, the sample sizes of these three treatment groups are inbalanced. For simplicity, the data are truncated randomly so that each of $m = 3$ treatment groups consists of $n = 120$ observations.

One-population m-strata problem: The objective is to compare the mean survival time of three treatments for chronic myelogenous leukemia. The penalized empirical likelihood method is applied. To chose the penalized parameter, we here take the grid-points for λ with log-scale as 0.001, 0.0021, 0.0046, 0.01, 0.021, 0.046, 0.1, 0.215, 0.464, 1, respectively. The detailed results are reported in Table 9.

Table 8 shows the detected cluster treatments after using our new clustering method. The second and third treatment share the same mean survival time of patients while that of the first treatment is different.

Two-population m-strata problem: The objective is to compare gender effect on the survival time under different treatments. Female is Population 1, and Male is Population 2. In this case, the three treatments are classified according to the additional gender effects on top of the treatment effects. To perform the estimation, λ is selected using the same grid-points as in the one-population m-strata case. It is interesting that the estimate suggests that there is only one cluster and thus, gender effects are similar under the three treatments.

Table 8 Misclassification rate for Example 2 Gamma with fixed variance

m	(n_1, n_2)	Cluster's means	BIC_1	BIC_2	FWER
40	(25, 25)	0	0.2191795	0.01175641	0
		(0, 4)	0.06938462	0.000474359	0.0004871795
		(0, 2, 4, 6)	0.03048718	0.005628205	0.01048718
	(25, 50)	0	0.1597179	0.001846154	0
		(0, 4)	0.04874359	0.0005128205	0.002192308
		(0, 2, 4, 6)	0.02653846	0.001794872	0.007192308
	(25, 100)	0	0.1352179	5e−04	0
		(0, 4)	0.02783333	0.0002564103	0.001461538
		(0, 2, 4, 6)	0.01602564	0.002833333	0.003538462
	(50, 50)	0	0.1163846	0.006974359	0
		(0, 4)	0.04230769	0	0.0002435897
		(0, 2, 4, 6)	0.01519231	0.001769231	0.004320513
	(50, 100)	0	0.1002692	0	0
		(0, 4)	0.02435897	0	0.0002435897
		(0, 2, 4, 6)	0.005269231	0.0004615385	0.006128205
	(100, 100)	0	0.05388462	0.001423077	0
		(0, 4)	0.02155128	0	0.0002435897
		(0, 2, 4, 6)	0.005217949	0.0002307692	0.003064103
200	(25, 25)	0	0.01700201	0.001559799	0
		(0, 4)	0.02635176	0.001421608	4.974874e−05
		(0, 2, 4, 6)	0.04450503	0.1834151	0.001914573
	(25, 50)	0	0.006042211	0.0002994975	0
		(0, 4)	0.01439296	0.0006909548	0
		(0, 2, 4, 6)	0.005829146	0.00218593	0.0008849246
	(25, 100)	0	0.006146734	0.0005959799	0
		(0, 4)	0.01159648	0.0002005025	0.0002487437
		(0, 2, 4, 6)	0.01647236	0.003116583	0.001758794
	(50, 50)	0	0.004151256	1.005025e−06	0
		(0, 4)	0.00869799	0	0
		(0, 2, 4, 6)	0.01774221	0.001542211	0.0006633166
	(50, 100)	0	0.002882915	1e−04	0
		(0, 4)	0.005376884	0.0001979899	9.949749e−05
		(0, 2, 4, 6)	0.004946734	0.0008115578	0.0008592965
	(100, 100)	0	0.001592462	0	0
		(0, 4)	0.002837688	9.899497e−05	4.974874e−05
		(0, 2, 4, 6)	0.002842211	0.0001708543	0.000638191

Table 9 Detected cluster treatments

Label of clusters	1	2
Number of treatments	2	1

Table 10 Detected cluster absolute returns of years

Label of clusters	1	2	3	4	5	6	7	8	9
Number of abs returns by year	1	2	5	11	8	2	4	3	1

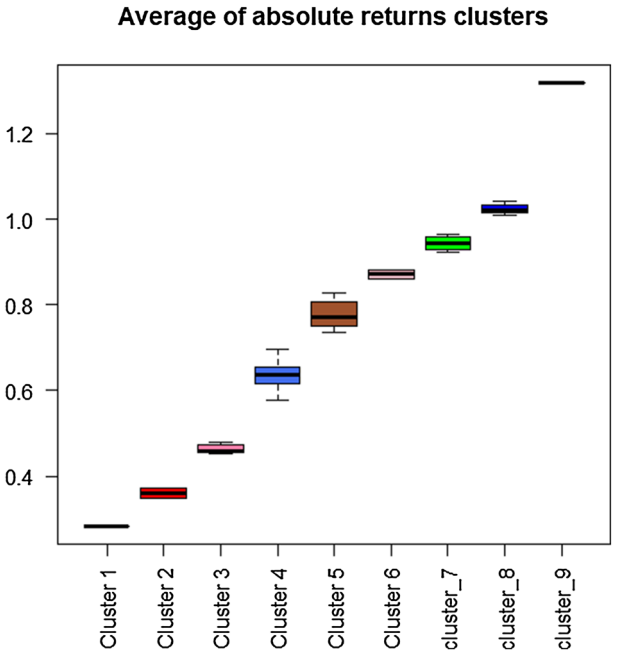

Fig. 1 Box plot for comparison average of absolute returns between clusters

6.2 Structural change and Monday effect in the stock market

Consider a medium m case with $m = 37$. The APPL (Apple) stock price data are used.

Case 1: Compare average daily absolute return of APPL (Apple) stock year by year from 1981 to 2017. Note that one year is considered as one group (strata) and the number of groups is $m = 37$. All groups share the same sample size. The absolute returns are computed as $|\log(S_t/S_{t-1})|$, where S_t refers to the stock price at time t. It is well-known among econometricians that financial data undergo regime switching and the stock returns are not identically distributed, see Andreou and Ghysels (2002). It is also a styled fact that the volatility exhibits certain time-varying pattern as described in the well-known autoregressive conditional heteroscedasticity (ARCH) model. In this example, the absolute returns are used to describe the volatility. Unlike the model-based methods that rely on the ARCH model and its variants, the proposed clustering method provides a non-model-based approach of studying heteroscedasticity in the financial market. The results are given in Table 9. Table 10 shows the classification of years using regularized empirical likelihood comparison method. Nine clusters are detected. Figures 1 and 2 show further details.

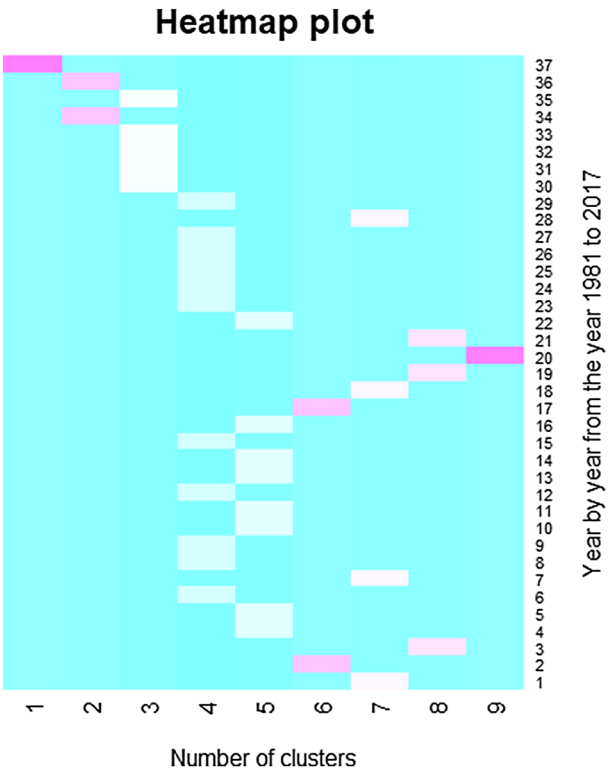

Fig. 2 Data arranged using Heatmap

Case 2: Consider a two-population m-strata problem. The Monday effect on APPL stock is studied year by year. Note that, one year is considered as one group (strata) and the absolute returns are the data. The purpose is to identify years with extraordinary Monday effects. Monday effect means that the Monday returns (close Friday to close Monday) are different from the returns on other days. There are many studies on the Monday effect in the financial markets. Some show that the Monday effect in the US stock market occurs strongly during the 1980's, see, e.g., French (1980), Rogalski (1984), etc. However, some recent works present evidence that Monday returns are not significantly different from returns during the rest of the week, see, e.g., Coutts and Hayes (1999), Steeley (2001), etc.

In order to illustrate the application of the regularized empirical likelihood approach and compare the conclusions of the above-mentioned works, we set Monday as Population 1 and other days of week (Tue,Wed,Thu,Fri) as Population 2. The finding is similar to those in Coutts and Hayes (1999). There is no Monday effect on APPL stock absolute returns year by year from the year 1981 to year 2017. 37 years share the same average absolute returns.

6.3 Microarray data of breast cancer patients

In this example, the breast cancer data in Van't Veer et al. (2002) are considered. The concepts of pairwise gene comparison are also used in Geman et al. (2004).

The data consist of gene expression profiles measured in 78 primary breast cancers cases: 34 from patients who developed distant metastases within 5 years (Population One) and 44 from patients who continued to be disease-free after a period of at least 5 years (Population Two). All patients were lymph node negative, and under 55 years of age at diagnosis. Profiles were obtained using Hu25K microarrays comprised of $G = 24{,}480$ human probe sequences.

In this real data example, we try to identify genes that can serve as indicators for distinguishing "good prognosis" from "poor prognosis" (long and short interval to distant metastases). By using the proposed regularized empirical likelihood method, the genes are classified according to the gene-expression-level-difference between two populations. Below, the notation μ refers to the gene-expression levels. The penalty takes the form

$$\lambda \sum_{i<j} \left|\mu_i^{(1)} - \mu_j^{(1)} - \mu_i^{(2)} + \mu_j^{(2)}\right|.$$

Consider reparameterization, $a_i = \mu_i^{(1)} - \mu_i^{(2)}$. Then, the penalty can be written as

$$\lambda \sum_{i<j} \left|a_i - a_j\right|.$$

Genes belonging to the cluster with greatest absolute a_i are identified.

To avoid the heavy computation, we use the proposed regularized empirical likelihood method in Sect. 2.1 together with the following data-splitting strategy:

- Step 1: We randomly split the breast cancer data into sub-data containing 200 genes each. For each sub-data, we do clustering and identify the genes not belonging to the biggest cluster. It is found that in all sub-data, there is one single cluster containing the majority of the genes.
- Step 2: For the out-genes identified in Step 1, do another clustering.

After the above procedure, we get 15 different clusters of genes, excluding the biggest clusters in Step 1. Among these 15 clusters, the biggest cluster contains 247 genes while the smallest cluster contains 1 gene. The detailed classification results are shown in Table 11. Figure 2 shows the box plot for mean-gene-expression-level differences in the 15 clusters, where the x-axis values are the sorted according to the ascending order of the means differences. In Fig. 2, the ranges of the 15 clusters do not overlap each other, suggesting significant differences between genes in different clusters. The gene from Cluster 15 (named SFRS2) shows greatest difference between the two populations (Fig. 3).

Table 11 Detected cluster genes

Number of clusters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of genes	1	1	1	13	2	2	247	23	4	1	1	1	2	1	1

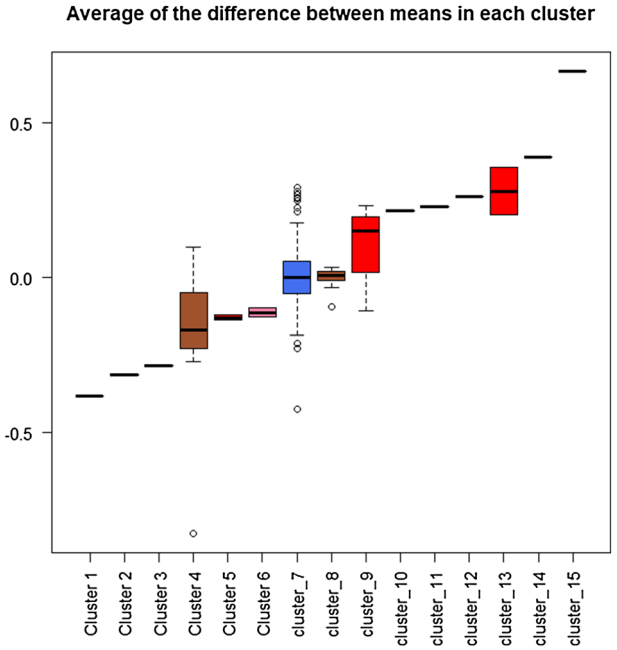

Fig. 3 Box plot for comparison average of means between clusters

7 Conclusion

It is illustrated that the strata clustering problem can be reformulated as pairwise $L1$ regularized estimation problem and the robustness can be achieved by using non-parametric empirical likelihood approach. One of the advantages is that contradictory conclusion can never be occurred under the proposed method. The proposed method allows one to control FWER and achieve consistency via BIC. Moreover, when the sample size is large, it is possible that all strata are classified correctly with probability going to one, see Theorems 1 and 2.

It is an interesting future research direction to study the influence of the dependence structure between the strata. This can further improve the applications in genetics and medicine where genes (strata) are known to be dependent on each other. Another interesting extension of the proposed approach is to explore the link between the new method and existing multiple comparisons methods under high-dimensional settings. It is also an interesting research direction is to establish formal asymptotic theory for the BIC for pairwise comparisons and strata clustering.

Acknowledgements Chi Tim, Ng's work is supported by National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No. NRF-2017R1C1B2011652) and 2018 Chonnam National University Research Program Grant (No. 2018-3428).

A Calculation of derivatives in Sect. 3.2

In this section, the formulas necessary to the computations in Step 3 of the algorithm in Sect. 3.2 are provided.

Step 3: Fix all θ_{ij} update $\tau_1, \ldots, \tau_m$ at the same time. Use Newton method to solve unconstrained minimization problem

$$Q^{***}(\tau, \theta) = \sum_{i=1}^{m} f_i(\tau_i) + \frac{\beta}{2} \sum_{i<j} (a_i - a_j - \theta_{ij})^2.$$

Gradient of Q^{***}:

$$\nabla_{\tau_i} Q^{***} = \frac{\partial f_i(\tau_i)}{\partial \tau_i} + \beta \sum_{i<j}^{m} (a_i - a_j - \theta_{ij}).$$

Let $f_i = k_{1i} + g_{1i} + k_{2i} + g_{2i}$, where

$$k_{1i} = \sum_{j=1}^{n^{(1)}} \log\left(n^{(1)} + \eta_i (x_{ij}^{(1)} - \mu_i^{(1)})\right),$$

$$k_{2i} = \sum_{j=1}^{n^{(2)}} \log\left(n^{(2)} + \eta_i (\mu_i^{(1)} - x_{ij}^{(2)} + a_i)\right),$$

$$g_{1i} = \frac{\beta}{2} h_{1i}^2; \quad h_{1i} = \sum_{j=1}^{n_i} \frac{n^{(1)}\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)}{n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)},$$

$$g_{2i} = \frac{\beta}{2} h_{2i}^2; \quad h_{2i} = \sum_{j=1}^{n_i} \frac{n^{(2)}\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)}{n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)}.$$

Then, $\nabla f_i = (v_1, v_2, v_3)$. The derivatives of f_i are as follows,

$$v_1 = \frac{\partial f_i}{\partial \mu_i^{(1)}} = \beta h_{1i} \frac{\partial h_{1i}}{\partial \mu_i^{(1)}} - \sum_{j=1}^{n^{(1)}} \frac{\eta_i}{n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)}$$

$$+\beta h_{2i} \frac{\partial h_{2i}}{\partial \mu_i^{(1)}} + \sum_{j=1}^{n^{(2)}} \frac{\eta_i}{n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)},$$

$$v_2 = \frac{\partial f_i}{\partial \eta_i} = h_{1i} + \beta h_{1i}\frac{\partial h_{1i}}{\partial \eta_i} + h_{2i} + \beta h_{2i}\frac{\partial h_{2i}}{\partial \eta_i},$$

$$v_3 = \frac{\partial f_i}{\partial a_i} = \frac{\partial k_{2i}}{\partial a_i} + \frac{\partial g_{2i}}{\partial a_i} = \sum_{j=1}^{n^{(2)}} \frac{\eta_i}{n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)} + \beta h_{2i}\frac{\partial h_{2i}}{\partial a_i},$$

and

$$\frac{\partial h_{1i}}{\partial \eta_i} = -n^{(1)} \sum_{j=1}^{n^{(1)}} \left(\frac{\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)}{n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)} \right)^2,$$

$$\frac{\partial h_{1i}}{\partial \mu_i^{(1)}} = -\sum_{j=1}^{n^{(1)}} \frac{\left(n^{(1)}\right)^2}{\left(n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)\right)^2},$$

$$\frac{\partial h_{2i}}{\partial a_i} = \frac{\partial h_{2i}}{\partial \mu_i^{(1)}} = \sum_{j=1}^{n^{(2)}} \frac{\left(n^{(2)}\right)^2}{\left(n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)\right)^2},$$

$$\frac{\partial h_{2i}}{\partial \eta_i} = -\sum_{j=1}^{n_i} n^{(2)} \left(\frac{\mu_i^{(1)} - x_{ij}^{(2)} + a_i}{n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)} \right)^2.$$

Hessian matrix of Q^{***}.

$$\nabla^2 Q^{***}_{\tau_i,\tau_i} = \frac{\partial^2 f_i(\tau_i)}{\partial \tau_i^2} + \beta(m-1), \quad \nabla^2 Q^{***}_{\tau_i,\tau_j} = -\beta, \quad \text{for } i \neq j,$$

$$\frac{\partial^2 f_i}{\partial a_i^2} = \frac{\partial^2 k_{2i}}{\partial a_i^2} + \frac{\partial^2 g_{2i}}{\partial a_i^2} = -\sum_{j=1}^{n^{(2)}} \left(\frac{\eta_i}{n^{(2)} + \eta_i(\mu_i^{(1)} - x_{ij}^{(2)} + a_i)} \right)^2 + \beta\left(\left(\frac{\partial h_{2i}}{\partial a_i}\right)^2 + h_2 \frac{\partial^2 h_{2i}}{\partial a_i^2} \right),$$

$$\frac{\partial^2 f_i}{\partial \eta_i^2} = \frac{\partial h_{1i}}{\partial \eta_i} + \beta\left(\left(\frac{\partial h_{1i}}{\partial \eta_i}\right)^2 + h_{1i}\frac{\partial^2 h_{1i}}{\partial \eta_i^2} \right) + \frac{\partial h_{2i}}{\partial \eta_i} + \beta\left(\left(\frac{\partial h_{2i}}{\partial \eta_i}\right)^2 + h_{2i}\frac{\partial^2 h_{2i}}{\partial \eta_i^2} \right),$$

$$\frac{\partial^2 f_i}{\partial (\mu_i^{(1)})^2} = -\sum_{j=1}^{n^{(1)}} \left(\frac{\eta_i}{n^{(1)} + \eta_i(x_{ij}^{(1)} - \mu_i^{(1)})} \right)^2 + \beta\left(\left(\frac{\partial h_{1i}}{\partial \mu_i^{(1)}}\right)^2 + h_{1i}\frac{\partial^2 h_{1i}}{\partial (\mu_i^{(1)})^2} \right) - \sum_{j=1}^{n^{(2)}} \left(\frac{\eta_i}{n^{(2)} + \eta_i(\mu_i^{(1)} - x_{ij}^{(2)} + a_i)} \right)^2 + \beta\left(\left(\frac{\partial h_{2i}}{\partial \mu_i^{(1)}}\right)^2 + h_{2i}\frac{\partial^2 h_{2i}}{\partial (\mu_i^{(1)})^2} \right),$$

$$\frac{\partial^2 f_i}{\partial a_i \partial \eta_i} = \frac{\partial h_{2i}}{\partial a_i} + \beta\left(\frac{\partial h_{2i}}{\partial \eta_i} \cdot \frac{\partial h_{2i}}{\partial a_i} + h_2 \frac{\partial^2 h_2}{\partial a_i \partial \eta_i}\right),$$

$$\frac{\partial^2 f_i}{\partial a_i \partial \mu_i^{(1)}} = -\sum_{j=1}^{n^{(2)}}\left(\frac{\eta_i}{n^{(2)} + \eta_i(\mu_i^{(1)} - x_{ij}^{(2)} + a_i)}\right)^2 + \beta\left(\frac{\partial h_{2i}}{\partial \mu_i^{(1)}} \cdot \frac{\partial h_{2i}}{\partial a_i} + h_2 \frac{\partial^2 h_{2i}}{\partial a_i \partial \mu_i^{(1)}}\right),$$

$$\begin{aligned}\frac{\partial^2 f_i}{\partial \eta_i \partial \mu_i^{(1)}} = {} & \frac{\partial h_{1i}}{\partial \mu_i^{(1)}} + \beta\left(\frac{\partial h_{1i}}{\partial \mu_i^{(1)}} \cdot \frac{\partial h_{1i}}{\partial \eta_i} + h_{1i} \frac{\partial^2 h_{1i}}{\partial \eta_i \partial \mu_i^{(1)}}\right) + \frac{\partial h_{2i}}{\partial \mu_i^{(1)}} \\ & + \beta\left(\frac{\partial h_{2i}}{\partial \mu_i^{(1)}} \cdot \frac{\partial h_{2i}}{\partial \eta_i} + h_{2i} \frac{\partial^2 h_{2i}}{\partial \eta_i \partial \mu_i^{(1)}}\right),\end{aligned}$$

where

$$\frac{\partial^2 h_{1i}}{\partial a_i^2} = \frac{\partial^2 h_{1i}}{\partial a_i \partial \mu_i^{(1)}} = 0,$$

$$\frac{\partial^2 h_{1i}}{\partial \eta_i^2} = 2n^{(1)} \sum_{j=1}^{n^{(1)}}\left(\frac{x_{ij}^{(1)} - \mu_i^{(1)}}{n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)}\right)^3,$$

$$\frac{\partial^2 h_{1i}}{\partial (\mu_i)^2} = -2(n^{(1)})^2 \sum_{j=1}^{n^{(1)}} \frac{\eta_i}{\left(n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)\right)^3},$$

$$\frac{\partial^2 h_{1i}}{\partial \eta_i \partial \mu_i^{(1)}} = 2(n^{(1)})^2 \sum_{j=1}^{n^{(1)}} \frac{(x_{ij}^{(1)} - \mu_i^{(1)})}{\left(n^{(1)} + \eta_i\left(x_{ij}^{(1)} - \mu_i^{(1)}\right)\right)^3},$$

$$\frac{\partial^2 h_{2i}}{\partial a_i^2} = \frac{\partial^2 h_{2i}}{\partial \left(\mu_i^{(1)}\right)^2} = -2(n^{(2)})^2 \sum_{j=1}^{n^{(2)}} \frac{\eta_i}{\left(n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)\right)^3},$$

$$\frac{\partial^2 h_{2i}}{\partial \eta_i^2} = 2n^{(2)} \sum_{j=1}^{n^{(2)}}\left(\frac{\mu_i^{(1)} - x_{ij}^{(2)} + a_i}{n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)}\right)^3,$$

$$\frac{\partial^2 h_{2i}}{\partial a_i \partial \eta_i} = \frac{\partial^2 h_{2i}}{\partial \eta_i \partial \mu_i^{(1)}} = -2(n^{(2)})^2 \sum_{j=1}^{n^{(2)}} \frac{(\mu_i^{(1)} - x_{ij}^{(2)} + a_i)}{\left(n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)\right)^3},$$

$$\frac{\partial^2 h_{2i}}{\partial a_i \partial \mu_i^{(1)}} = -2(n^{(2)})^2 \sum_{j=1}^{n^{(2)}} \frac{\eta_i}{\left(n^{(2)} + \eta_i\left(\mu_i^{(1)} - x_{ij}^{(2)} + a_i\right)\right)^3}.$$

B Proofs of main theorems

B.1 Notation

The following conventions are used throughout the proof. Let $\epsilon > 0$ and $\epsilon_2 > 0$ be some chosen infinitesimal quantities so that

(B1) $\epsilon \gg \epsilon_2$,
(B2) $\epsilon_2 \gg (mn)^{-1/2}$,
(B3) $\epsilon_2 \gg \lambda mn^{-1}$.
Choose β so that
(B4) $\beta \min_{i,\Xi} \left\{ \sum_{j=1}^{n} \left(\frac{x_{ij}-\mu_i}{n+\eta(\mu_i)(x_{ij}-\mu_i)} \right)^2 \right\} > 1$, $\beta \gg mn$, and $\lambda\beta \gg n^2 u_n^{(s)} \epsilon_2 (\min_{1\le s\le c} m_s)^{-2}$.

Some notation is introduced. Let $\mu^\dagger = (\mu_{(1)}, \ldots, \mu_{(c)})$ be c-dimensional vector and $\mu = (\mu_1, \ldots, \mu_m)$ be m-dimensional vector. Let $\bar{\mu}_{(s)} = m_s^{-1} \sum_{b_{s-1}+1}^{b_s} \mu_i$ and

$$\bar{\mu} = (\bar{\mu}_{(1)}, \ldots, \bar{\mu}_{(1)}, \ldots, \bar{\mu}_{(c)}, \ldots, \bar{\mu}_{(c)}).$$

Define

$$\begin{aligned} Q^{**}(\mu, \theta) &= \sum_{i=1}^{m} \ell_i(\mu_i) + \frac{\beta}{2} \sum_{i<j} (\mu_i - \mu_j - \theta_{ij}) + \lambda \sum_{i<j} w_{ij} |\theta_{ij}|, \\ L^\dagger(\mu^\dagger) &= \sum_{s=1}^{c} \ell_{(s)}(\mu_{(s)}), \\ Q^\dagger(\mu^\dagger) &= \sum_{s=1}^{c} \ell_{(s)}(\mu_{(s)}) + \lambda \sum_{s<t} \left(\sum_{i=b_{s-1}+1}^{b_s} \sum_{j=b_{t-1}+1}^{b_t} w_{ij} \right) |\mu_{(s)} - \mu_{(t)}|, \end{aligned} \quad (9)$$

where $\ell_i(\mu_i) = \sum_{k=1}^{n_i} \log(n + \eta_i(\mu_i)(x_{ik} - \mu_i))$ and $\ell_{(s)}(\mu_{(s)}) = \sum_{i=b_{s-1}+1}^{b_s} \ell_i(\mu_{(s)})$. Denote

$$\begin{aligned} \hat{\mu}^\dagger &= (\mu_{(1)}, \ldots, \mu_{(c)}) = \operatorname{argmin} Q^\dagger(\mu^\dagger) \quad \text{and} \\ \hat{\mu}^{\dagger\dagger} &= (\mu_{(1)}, \ldots, \mu_{(1)}, \ldots, \mu_{(c)}, \ldots, \mu_{(c)}). \end{aligned}$$

Lemma 3 guarantees the existence of $\hat{\mu}^\dagger$. Moreover, Lemma 1 suggests that $\ell_i(\mu_i) = \min_\eta f_i(\eta, \mu)$ if β is sufficiently large.

B.2 Proof of Theorem 1 and Theorem 2

The proof of Theorem 2 is similar to that of Theorem 1 so that the proof is omitted. In what follows, we establish Theorem 1. By Lemma 1, optimizing $Q^*(\eta, \mu, \theta)$ is

equivalent to optimizing $Q^{**}(\mu, \theta)$. It can be checked that fixed μ, $Q^{**}(\mu, \theta)$ is optimized at

$$\theta_{ij}(\mu) = \begin{cases} 0, & \text{when } |\mu_i - \mu_j| \leq \lambda w_{ij}/\beta, \\ \mu_i - \mu_j - \frac{\lambda w_{ij}}{\beta} \text{sgn}(\mu_i - \mu_j), & \text{otherwise.} \end{cases}$$

Define

$$\theta^*_{ij}(\mu) = \begin{cases} 0, & \text{when } \mu^0_i = \mu^0_j, \\ \mu_i - \mu_j - \frac{\lambda w_{ij}}{\beta} \text{sgn}(\mu_i - \mu_j), & \text{otherwise.} \end{cases}$$

In what follows, we show that $\min_{\mu \in \Xi} Q^{**}(\mu, \theta^*(\mu))$ exists in an infinitesimal neighborhood around $\hat{\mu}^{\dagger\dagger}$ and such a solution fulfills $|\hat{\mu}_i - \hat{\mu}_j| \leq \lambda(\min\{w_{ij} : \mu^0_i = \mu^0_j\})/\beta$ for all i, j belonging to the same "true cluster" with probability going to one. If these are so, the solution to $\min_{\mu \in \Xi} Q^{**}(\mu, \theta^*(\mu))$ solves $\min_{\mu \in \Xi} Q^{**}(\mu, \theta(\mu))$ too.

To establish the existence of $\min_{\mu \in \Xi} Q^{**}(\mu, \theta^*(\mu))$. Consider the neighborhood

$$\mathcal{N} = \left\{ \mu : |\mu_i - \bar{\mu}_i| \leq \epsilon \text{ and } |\bar{\mu}_i - \hat{\mu}^{\dagger\dagger}_i| \leq \epsilon, i = 1, 2, \ldots, m \right\}.$$

See "Appendix B.1" for the definition of ϵ. Since $\mathcal{N}$ is compact, $\min_{\mu \in \mathcal{N}} Q^{**}(\mu, \theta^*(\mu))$ must exists. It suffices to show that $Q^{**}(\mu, \theta^*(\mu)) > Q^{**}(\hat{\mu}^{\dagger\dagger}, \theta^*(\hat{\mu}^{\dagger\dagger}))$ for all boundary points of $\mathcal{N}$ with probability going to one. If this is so, the minimum cannot be attained on the boundary and therefore must be an interior point. Consequently, the local minimum appears inside $\mathcal{N}$. Under condition (A3), when $\mu \in \mathcal{N}$,

$$\begin{aligned} Q^{**}(\mu, \theta^*(\mu)) &= K + \sum_{i=1}^{m} \ell_i(\mu_i) + \frac{\beta}{2} \sum_{s=1}^{c} \sum_{b_{s-1}+1 \leq i < j \leq b_s} (\mu_i - \mu_j)^2 \\ &\quad + \lambda \sum_{s<t} \left(\sum_{i=b_{s-1}+1}^{b_s} \sum_{j=b_{t-1}+1}^{b_t} w_{ij}(\mu_j - \mu_i) \right) \\ &= K + Q_1(\mu) + Q_2(\mu) + Q_3(\mu), \end{aligned} \tag{10}$$

where

$$K = -\frac{\lambda^2}{2\beta} \sum_{s<t} \left(\sum_{i=b_{s-1}+1}^{b_s} \sum_{j=b_{t-1}+1}^{b_t} w^2_{ij} \right)$$

is a constant. Consider the approximation

$$\begin{aligned} Q_1(\mu) + Q_3(\mu) \approx\ & Q_1(\hat{\mu}^{\dagger\dagger}) + Q_3(\hat{\mu}^{\dagger\dagger}) + (Q_3(\mu) - Q_3(\hat{\mu}^{\dagger\dagger})) + \nabla Q_1(\hat{\mu}^{\dagger\dagger}) \\ & \cdot(\mu - \hat{\mu}^{\dagger\dagger}) + (\mu - \hat{\mu}^{\dagger\dagger})^T \nabla^2 Q_1(\hat{\mu}^{\dagger\dagger}) \cdot (\mu - \hat{\mu}^{\dagger\dagger}). \end{aligned}$$

Under conditions (A1), Lemma 2 suggests that

$$\nabla Q_1(\hat{\mu}^{\dagger\dagger}) \cdot (\mu - \hat{\mu}^{\dagger\dagger}) \leq O_p(mn\epsilon_2\epsilon)$$

and

$$|Q_3(\mu) - Q_3(\hat{\mu}^{\dagger\dagger})| \leq O_p(\lambda m^2 \epsilon).$$

There are two types of boundary points on $\partial\mathcal{N}$, (i) $|\mu_i - \bar{\mu}_i| = \epsilon$ for some i and $|\bar{\mu}_i - \hat{\mu}_i^{\dagger\dagger}| \leq \epsilon$ for all i and (ii) $|\mu_i - \bar{\mu}_i| \leq \epsilon$ for all i and $|\bar{\mu}_i - \hat{\mu}_i^{\dagger\dagger}| = \epsilon$ for some i. For type (i) boundary points, $|\mu_i - \mu_j| > \epsilon$ for some $i \neq j$. Without loss of generality, assume that $|\mu_1 - \mu_2| > \epsilon$. Then, $Q_2(\mu) \geq \beta(\mu_1 - \mu_2)^2/2 = O_p(\beta\epsilon^2)$. Conditions (B1) and (B3) guarantee that the positive definite term $Q_2(\mu)$ is dominating. For type (ii) boundary points, $|\bar{\mu}_i - \hat{\mu}_i^{\dagger\dagger}| = \epsilon$ for some i. Note that $\bar{\mu}_i - \hat{\mu}_i^{\dagger\dagger}$ share a common value within the same cluster. Then,

$$\|\mu - \hat{\mu}^{\dagger\dagger}\|_2 \geq \|\mu - \bar{\mu}\|_2 = O_p(m\epsilon^2)$$

and

$$(\mu - \hat{\mu}^{\dagger\dagger})^T \nabla^2 Q_1(\hat{\mu}^{\dagger\dagger}) \cdot (\mu - \hat{\mu}^{\dagger\dagger}) \geq O_p(mn\epsilon^2).$$

Conditions (B1) and (B3) guarantee that the positive definite term $(\mu - \hat{\mu}^{\dagger\dagger})^T \nabla^2 Q_1(\hat{\mu}^{\dagger\dagger})$ dominates all other terms excepting $Q_2(\mu)$. Moreover, the quantity $Q_2(\mu)$ is always positive. This completes the existence proof.

Next, we show with probability going to one that the solution to $Q^{**}(\mu, \theta^*(\mu))$ fulfills $|\hat{\mu}_i - \hat{\mu}_j| \leq \lambda/\beta$ for all $i, j = b_{s-1} + 1, \ldots, b_s$ within the same cluster s from $1, 2, \ldots, c$. Note that by the definition of $\mathcal{N}$, the function $Q^{**}(\mu, \theta^*(\mu))$ is differentiable if Condition (A3) holds and $\lambda/\beta \to 0$. Then,

$$\begin{aligned}\nabla_{\mu_i} Q^{**}(\mu_i, \theta^*(\mu_i)) &= \nabla_{\mu_i} Q_1(\mu_i) + \nabla_{\mu_i} Q_2(\mu_i) + \nabla_{\mu_i} Q_3(\mu_i) \\ &= \nabla_{\mu_i} \ell_i(\mu_i) + \beta \sum_{i<j=b_{s-1}+1}^{b_s} (\mu_i - \mu_j) - \lambda \sum_{s<t} \sum_{j=b_{t-1}+1}^{b_t} w_{ij},\end{aligned}$$

for $i = b_{s-1} + 1, \ldots, b_s$ and $s = 1, \ldots, c$. Consider $\hat{\mu}_i$, the minimizer of $Q^{**}(\mu_i, \theta^*(\mu_i))$. Suppose that i belongs to the s-th cluster. By Taylor expansion, it holds that

$$\begin{aligned}0 = \nabla_{\mu_i} Q^{**}(\mu_i, \theta^*(\mu_i))\Big|_{\mu_i=\hat{\mu}_i} &\approx \nabla_{\mu_i} \ell_i\left(\hat{\mu}_i^{\dagger\dagger}\right) + H_i\left(\hat{\mu}_i - \hat{\mu}_i^{\dagger\dagger}\right) \\ &+\beta \sum_{i<j=b_{s-1}+1}^{b_s} \left(\hat{\mu}_i - \hat{\mu}_i^{\dagger\dagger} - \hat{\mu}_j + \hat{\mu}_j^{\dagger\dagger}\right) - \lambda \sum_{t\neq s} \sum_{e=b_{t-1}+1}^{b_t} w_{ie},\end{aligned}$$

where $H_i = \nabla^2 Q^{**}(\hat{\mu}_i^{\dagger\dagger}, \theta^*(\hat{\mu}_i^{\dagger\dagger}))$. The bias $|\hat{\mu} - \hat{\mu}^{\dagger\dagger}|$ can therefore be established as follows,

$$\hat{\mu} - \hat{\mu}^{\dagger\dagger} = \begin{pmatrix} \mathcal{H}_1 & 0 & \cdots & 0 \\ 0 & \mathcal{H}_2 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \mathcal{H}_c \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{P}_1 - \mathcal{G}_1 \\ \mathcal{P}_2 - \mathcal{G}_2 \\ \vdots \\ \mathcal{P}_c - \mathcal{G}_c \end{pmatrix}$$

where $\mathcal{D}_s = \text{diag}(H_{b_{s-1}+1}, \ldots, H_{b_s}) + m_s\beta \mathbf{I}_{m_s}$, $\mathcal{H}_s = \mathcal{D}_s + m_s\beta \mathbf{I}_{m_s} - \beta \mathbf{1}_{m_s}\mathbf{1}_{m_s}^T$, $\mathcal{G}_s = \left(G_{b_{s-1}+1}, \ldots, G_{b_s}\right)$ for $s = 1, \ldots, c$, $G_i = \nabla \ell(\hat{\mu}_i^{\dagger\dagger})$ for $i = 1, 2, \ldots, m$, $\mathcal{P}_s = \left(P_{b_{s-1}+1}, \ldots, P_{b_s}\right)$ for $s = 1, \ldots, c$ and $\mathcal{P}_i = \lambda \sum_{t \neq s} \sum_{e=b_{t-1}+1}^{b_t} w_{ie}$ for $i = b_{s-1} + 1, \ldots, b_s$.

Consider the matrix inverse formula

$$\mathcal{H}_s^{-1} = (\mathcal{D}_s - \beta \mathbf{1}_{m_s}\mathbf{1}_{m_s}^T)^{-1} = \mathcal{D}_s^{-1} + \mathcal{D}_s^{-1}\mathbf{1}_{m_s}\left(\frac{1}{\beta} - \mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}\mathbf{1}_{m_s}\right)^{-1} \mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}.$$

Then,

$$\hat{\mu}_s - \hat{\mu}_s^{\dagger\dagger} = \mathcal{D}_s^{-1}(\mathcal{P}_s - \mathcal{G}_s) + \mathcal{D}_s^{-1}\mathbf{1}_{m_s}\left(\frac{1}{\beta} - \mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}\mathbf{1}_{m_s}\right)^{-1} \mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}(\mathcal{P}_s - \mathcal{G}_s).$$

It can be easily seen that

$$\mathcal{D}_s^{-1} = \begin{pmatrix} \frac{1}{H_{b_{s-1}+1} + m_s\beta} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{H_{b_s} + m_s\beta} \end{pmatrix}, \quad \mathcal{D}_s^{-1}\mathbf{1}_{m_s} = \begin{pmatrix} \frac{1}{H_{b_{s-1}+1} + m_s\beta} \\ \vdots \\ \frac{1}{H_{b_s} + m_s\beta} \end{pmatrix}$$

and

$$\mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}\mathbf{1}_{m_s} = \sum_{i=b_{s-1}+1}^{b_s} \frac{1}{H_i + m_s\beta}.$$

Note that $m_s\beta$ eventually dominates all H_{b_i}. By Lemma 2,

$$\begin{aligned} \frac{1}{\beta} - \mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}\mathbf{1}_{m_s} &= \frac{1}{\beta} - \sum_{i=b_{s-1}+1}^{b_s} \frac{1}{H_i + m_s\beta} = \frac{1}{\beta} - \frac{1}{m_s\beta} \sum_{i=b_{s-1}+1}^{b_s} \frac{1}{1 + \frac{H_i}{m_s\beta}} \\ &= \frac{1}{\beta} - \frac{1}{m_s\beta} \sum_{i=b_{s-1}+1}^{b_s} \left(1 - \frac{H_i}{m_s\beta} + \frac{H_i^2}{m_s^2\beta^2} - \cdots\right) \\ &\approx \frac{1}{m_s^2\beta^2} \sum_{i=b_{s-1}+1}^{b_s} H_i = O_p\left(\frac{n}{m_s\beta^2}\right). \end{aligned}$$

From the first-order condition of $Q^{\dagger}(\cdot)$, $\sum_{i=b_{s-1}+1}^{b_s}(P_i - G_i) = 0$. In addition,

$$\frac{P_i}{H_i + m_s\beta} = O_p\left(\frac{\lambda m\epsilon_2}{m_s\beta}\right) \quad \text{and} \quad \frac{G_i}{H_i + m_s\beta} = O_p\left(\frac{n\epsilon_2}{m_s\beta}\right).$$

Hence, under Condition (B3),

$$\begin{aligned}
\mathbf{1}_{m_s}^T \mathcal{D}_s^{-1}(\mathcal{P}_s - \mathcal{G}_s) &= \sum_{i=b_{s-1}+1}^{b_s} \frac{P_i - G_i}{H_i + m_s\beta} \\
&\approx \frac{1}{m_s^2\beta^2} \sum_{i=b_{s-1}+1}^{b_s} H_i(P_i - G_i) \\
&= O_p\left(\max\left\{\frac{\lambda mn}{m_s\beta^2}; \frac{n^2\epsilon_2}{m_s\beta^2}\right\}\right) \\
&= O_p\left(\frac{n^2\epsilon_2}{m_s\beta^2}\right).
\end{aligned}$$

Since both $A_1 = \left(\frac{1}{\beta} - \mathbf{1}_{m_s}^T D_s^{-1}\mathbf{1}_{m_s}\right)^{-1}$ and $A_2 = \mathbf{1}_{m_s}^T D_s^{-1}(P_s - G_s)$ are constants, for i, j belonging to the same cluster,

$$|\hat{\mu}_i - \hat{\mu}_j| \approx \frac{P_i - G_i}{H_i + m_s\beta} - \frac{P_j - G_j}{H_j + m_s\beta} + \left(\frac{1}{H_i + m_s\beta} - \frac{1}{H_j + m_s\beta}\right) \cdot A_1 \cdot A_2.$$

Under Condition A1, we see that $|P_i - P_j| = o_p(\lambda m w_{ij})$. Under assumption $m\beta \gg \max_i H_i$, Taylor expansion yields

$$\begin{aligned}
\frac{1}{H_i + m_s\beta} - \frac{1}{H_j + m_s\beta} &= \frac{1}{m_s\beta}\left(\frac{1}{1 + \frac{H_i}{m_s\beta}}\right) - \frac{1}{m_s\beta}\left(\frac{1}{1 + \frac{H_j}{m_s\beta}}\right) \\
&= \frac{1}{m_s\beta}\left(1 - \frac{H_i}{m_s\beta} + o(1)\right) - \frac{1}{m_s\beta}\left(1 - \frac{H_j}{m_s\beta} + o(1)\right) \\
&\approx \frac{1}{m_s^2\beta^2}(H_i - H_j) = O_p\left(\frac{nu_n^{(s)}}{m_s^2\beta^2}\right), \\
\frac{P_i}{H_i + m_s\beta} - \frac{P_j}{H_j + m_s\beta} &= \frac{P_i}{m_s\beta}\left(\frac{1}{1 + \frac{H_i}{m_s\beta}}\right) - \frac{P_j}{m_s\beta}\left(\frac{1}{1 + \frac{H_j}{m_s\beta}}\right) \\
&\approx \frac{P_i}{m_s\beta}\left(1 - \frac{H_i}{m_s\beta} + \frac{H_i^2}{m_s^2\beta^2}\right) \\
&\quad - \frac{P_j}{m_s\beta}\left(1 - \frac{H_j}{m_s\beta} + \frac{H_j^2}{m_s^2\beta^2}\right)
\end{aligned}$$

$$\approx \frac{1}{m_s\beta}(P_i - P_j) - \frac{1}{m_s^2\beta^2}(P_iH_i - P_jH_j)$$
$$\approx o_p\left(\frac{\lambda m w_{ij}}{m_s\beta}\right) + o_p\left(\frac{\lambda m n u_n^{(s)}}{m_s^2\beta^2}\right),$$
$$\frac{G_i}{H_i + m_s\beta} - \frac{G_j}{H_j + m_s\beta} = O_p\left(\frac{\sqrt{n}u_n^{(s)} + \sqrt{n}u_n^{(s)}\epsilon_2}{m_s\beta}\right).$$

This implies that when $\lambda\beta \gg n^2 u_n^{(s)}\epsilon_2 m_s^{-2}$ and $\beta \gg m n u_n^{(s)} m_s^{-2}$, it holds that $|\hat{\mu}_i - \hat{\mu}_j| = o_p\left(\frac{\lambda w_{ij}}{\beta}\right)$. This completes the proof of (i).

Using the same technique developed for the proof of (i) and from the definition of compact set $\mathcal{N}$, (ii) and (iii) hold. □

B.3 Technical Lemmas

Lemma 1 *Under assumption (B4), $\eta_i(\mu_i)$ minimizes $f_i(\eta_i, \mu_i)$ for all $\mu_i \in R$ and $f_i(\eta_i(\mu_i), \mu_i) = \ell(\mu_i)$. Let $(\hat{\mu}, \hat{\theta})$ be the local minimizer of the function $Q^{**}(\mu, \theta)$. Then, $(\eta(\hat{\mu}), \hat{\mu})$ is a local minimizer of $Q^*(\eta, \mu, \theta)$ too, or equivalently,*

$$(\hat{\mu}, \hat{\theta}) = \underset{\mu,\theta}{\operatorname{argmin}}\, Q^{**}(\mu, \theta) = \underset{\mu,\theta}{\operatorname{argmin}} \min_{\eta} Q^*(\eta, \mu, \theta).$$

Proof Clearly, when $\eta_i = \eta_i(\mu_i)$, we have $h_i = 0$. From "Appendix A,"

$$\frac{\partial f_i}{\partial \eta_i} = h_i + \beta h_i \frac{\partial h_i}{\partial \eta_i} = 0 \quad \text{and} \quad \frac{\partial^2 f_i}{\partial \eta_i^2} = \frac{\partial h_i}{\partial \eta_i} + \beta\left(\frac{\partial h_i}{\partial \eta_i}\right)^2.$$

Then, for any fixed μ, f_i (as a function of η_i) is convex when $\partial h_i/\partial\eta_i < -1/\beta$, or equivalently when (B4) holds. Consequently, $\eta_i(\mu_i) = \operatorname{argmin}_{\eta_i} f_i(\eta_i, \mu_i)$. It is obvious that $f_i(\eta_i(\mu_i), \mu_i) = \ell(\mu_i)$ because $h_i = 0$. The desired results follow immediately. □

Lemma 2 *Let $\hat{\mu}_i$ be a consistent estimator of μ_i^0. Suppose that an infinitesimal compact set $\mathcal{N} \subset R$ contains both μ_i^0 and $\hat{\mu}_i$ as interior points. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\mu_i} f_i(\eta_i(\hat{\mu}_i), \hat{\mu}_i) = O_p\left(\max\left\{n^{1/2}, n\left(\hat{\mu}_i - \mu_i^0\right)\right\}\right)$$

and

$$\inf_{\mu\in\mathcal{N}} \frac{\mathrm{d}^2}{\mathrm{d}\mu_i^2} f_i(\eta_i(\mu_i), \mu_i)$$

is bounded below by some positive $O_p(n)$ quantity.

Proof Since

$$\sum_{j=1}^{n} \frac{1}{n+\eta_i(\mu_i)(x_{ij}-\mu_i)} = 1 \quad \text{and} \quad \sum_{j=1}^{n} \frac{x_{ij}-\mu_i}{n+\eta_i(\mu_i)(x_{ij}-\mu_i)} = 0,$$

it can be verified after some algebraic manipulations that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\mu_i} f_i(\eta_i(\mu_i), \mu_i) &= -\sum_{j=1}^{n} \frac{\eta_i(\mu_i)}{n+\eta_i(\mu_i)(x_{ij}-\mu_i)} + \sum_{j=1}^{n} \frac{x_{ij}-\mu_i}{n+\eta_i(\mu_i)(x_{ij}-\mu_i)} \frac{\mathrm{d}\eta}{\mathrm{d}\mu} \\
&= -\eta_i(\mu_i), \\
\frac{\mathrm{d}^2}{\mathrm{d}\mu_i^2} f_i(\eta_i(\mu_i), \mu_i) &= -\frac{\mathrm{d}}{\mathrm{d}\mu_i}\eta_i(\mu_i) = n\left[\sum_{k=1}^{n}\left(\frac{x_{ik}-\mu_i}{n+\eta_i(\mu_i)(x_{ik}-\mu_i)}\right)^2\right]^{-1} \\
&\quad \left[\sum_{k=1}^{n}\left(\frac{1}{n+\eta_i(\mu_i)(x_{ik}-\mu_i)}\right)^2\right].
\end{aligned}$$

As shown in Owen (2001), $\frac{\mathrm{d}}{\mathrm{d}\mu_i} f_i(\eta_i(\mu_i^0), \mu_i^0) = O_p(n^{1/2})$ and $\eta_i(\mu_i^0) = O_p(n^{1/2})$, then using Taylor expansion, we have

$$\begin{aligned}
\eta_i(\hat{\mu}_i) &\approx \eta_i\left(\mu_i^0\right) - \left(\hat{\mu}_i - \mu_i^0\right) \left.\frac{\mathrm{d}\eta_i(\mu_i)}{\mathrm{d}\mu_i}\right|_{\mu_i=\mu_i^0} \\
&\approx \eta_i\left(\mu_i^0\right) - \left(\hat{\mu}_i - \mu_i^0\right) \sum_{j=1}^{n} \frac{n}{n+\eta_i\left(\mu_i^0\right)\left(x_{ij}-\mu_i^0\right)} \\
&= O_p(n^{1/2}) + n\left(\hat{\mu}_i - \mu_i^0\right) = O_p\left(\max\left\{n^{1/2}, n\left(\hat{\mu}_i - \mu_i^0\right)\right\}\right).
\end{aligned}$$

Next, we give a lower bound of $\frac{d^2}{d\mu_i^2} f_i(\eta(\mu_i), \mu_i)$. Note that $\sum_{k=1}^{n} p_i^2 \geq 1/n$ subjected to the constraint $\sum_{k=1}^{n} p_i$. Therefore,

$$\sum_{k=1}^{n}\left(\frac{1}{n+\eta_i(x_{ik}-\mu_i)}\right)^2 \geq \frac{1}{n}.$$

Suppose that μ_i belongs to an infinitesimal set containing μ_i^0 as interior point. Consider the approximation

$$\eta_i(\mu_i) \approx \frac{n\sum_{k=1}^{n}(x_{ik}-\mu_i)}{\sum_{k=1}^{n}(x_{ik}-\mu_i)^2}.$$

It is not difficult to see that

$$\sum_{k=1}^{n}\left(\frac{x_{ik}-\mu_i}{n+\eta_i(x_{ik}-\mu_i)}\right)^2 = \frac{1}{n^2}\sum_{k=1}^{n}(x_{ik}-\mu_i)^2\left\{1-\frac{2}{n}(x_{ik}-\mu_i)\eta_i(\mu_i)+\cdots\right\}$$
$$= O_p(n^{-1}).$$

Then, the second-order derivative is bounded below by some positive $O_p(n)$ quantity. □

Lemma 3 *The objective function* (9) *admits a local solution in the interior of the ball*

$$\mathcal{B} = \left\{(\mu_{(1)},\ldots,\mu_{(c)}) : \sum_{s=1}^{c}\left(\mu_{(s)}-\mu^0_{(s)}\right)^2 \le \epsilon_2^2\right\}$$

with probability going to one. See Convention A1 for the definition of ϵ_2.

Proof Since $\mathcal{B}$ is compact, there must be a minimum within $\mathcal{B}$. To complete the proof, we show with probability going to one that on the boundary $\partial\mathcal{B}$, we have $Q^\dagger(\mu_{(1)},\ldots,\mu_{(c)}) > Q^\dagger(\mu^0_{(1)},\ldots,\mu^0_{(c)})$. Then, such a minimum must not be attained on the boundary. There must be a local minimum in the interior of $\mathcal{B}$. Consider the approximation

$$\begin{aligned}
&Q^\dagger(\mu_{(1)},\ldots,\mu_{(c)}) - Q^\dagger\left(\mu^0_{(1)},\ldots,\mu^0_{(c)}\right)\\
&= \left(\mu_{(1)}-\mu^0_{(1)},\ldots,\mu_{(c)}-\mu^0_{(c)}\right)\cdot\nabla L^\dagger\left(\mu^0_{(1)},\ldots,\mu^0_{(c)}\right)\\
&\quad+\left(\mu_{(1)}-\mu^0_{(1)},\ldots,\mu_{(c)}-\mu^0_{(c)}\right)^T\nabla^2 L^\dagger\left(\mu^0_{(1)},\ldots,\mu^0_{(c)}\right)\\
&\quad\left(\mu_{(1)}-\mu^0_{(1)},\ldots,\mu_{(c)}-\mu^0_{(c)}\right)+\lambda\sum_{s<t}\left(\sum_{i=b_{s-1}+1}^{b_s}\sum_{j=b_{t-1}+1}^{b_t}w_{ij}\right)\\
&\quad\left(|\mu_{(s)}-\mu_{(t)}|-|\mu^0_{(s)}-\mu^0_{(t)}|\right)\\
&= R_1+R_2+R_3.
\end{aligned}$$

From Condition (A1) and Lemma 2, $R_1 \le O_p((mn)^{1/2}\epsilon_2)$, $R_2 \ge O_p(mn\epsilon_2^2)$, and $R_3 \le O_p(\lambda m^2\epsilon_2)$. Conditions (B2)–(B3) guarantee that the positive term R_2 dominates both R_1 and R_3 and thus $R_1+R_2+R_3>0$. □

References

Agresti, A., Bini, M., Bertaccini, B., Ryu, E. (2008). Simultaneous confidence interval for comparing binomial parameters. *Biometrics*, *64*, 1270–1275.

Andreou, E., Ghysels, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, *17*, 579–600.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita Pubblicazioni. *del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.
Cao, R., Van Keilegom, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *The Canadian Journal of Statistics*, *34*, 61–77.
Coutts, J. A., Hayes, P. A. (1999). The weekend effect, the stock exchange account and the financial times industrial ordinary shares index: 1987–1994. *Applied Financial Economics*, *9*, 67–71.
Dmitrienko, A., Tamhane, A., Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics*. Boca Raton: Chapman and Hall/CRC Press.
Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, *11*, 1–42.
Efron, B., Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.
Fan, J., Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, *96*, 1348–1360.
Fan, J., Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*, 101–148.
Fan, J., Peng, H. (2004). On nonconcave penalized likelihood with diverging number of parameters. *The Annals of Statistics*, *32*, 928–961.
Fan, Y., Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *The Annals of Statistics*, *38*, 3567–3604.
Fisher, R., Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of largest or smallest member of a sample. In *Proceedings of the Cambridge philosophical society* (Vol. 24, pp. 180–190).
French, K. (1980). Stock returns and the weekend effect. *Journal of Financial Economics*, *8*, 59–69.
Friedman, J., Hastie, T., Hofling, H., Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *2*, 302–332.
Fu, W. J. (1998). Penalized regression: The bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, *7*, 397–416.
Gabriel, K. (1969). Simultaneous test procedures—Some theory of multiple comparisons. *The Annals of Mathematical Statistics*, *40*, 224–250.
Gelman, A., Hill, J., Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211.
Geman, D., d'Avignon, C., Winslow, R. (2004). Classifying gene expression profiles from pairwise mRNA comparisons. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–19.
Gnedenko, B. V. (1943). Sur la distribution limite du terme d'une série aléatoire. *Annals of Mathematics*, *44*, 423–453.
Hehlmann, R., Heimpel, H., Hasford, J., Kolb, H. J., Pralle, H., Hossfeld, D. K., Queisser, W., Loeffler, H., Hochhaus, A., Heinze, B. (1994). Randomized comparison of interferon-alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. The German CML study group. *Blood*, *84*(12), 4064–4077.
Hochberg, Y., Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.
Jing, B. Y. (1995). Two-sample empirical likelihood method. *Statistics & Probability Letters*, *24*, 315–319.
Kleinman, K., Huang, S. S. (2016). Calculating power by bootstrap, with an application to cluster-randomized trials. *The Journal for Electronic Health Data and Methods*, *4*, 1202.
Lawley, D. N., Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York: American Elsevier Publisher Company.
Leng, C., Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, *99*, 703–716.
Lin, Y. Q., Cheung, S. H., Poon, W. Y., Lu, T. Y. (2014). Pairwise comparisons with ordered categorical data. *Statistics in Medicine*, *32*, 3192–3205.
Liu, Y., Zou, C., Zhang, R. (2008). Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters*, *78*, 548–556.
Marchetti, Y., Zhou, Q. (2014). Solution path clustering with adaptive concave penalty. *Electronic Journal of Statistics*, *8*, 1569–1603.
McCormick, W. P. (1980). Weak convergence for the maxima of stationary Gaussian processes using random normalization. *Annals of Probability*, *8*, 483–497.
Miller, R. (1981). *Simultaneous statistical inference*. New York: Springer.
Ng, C. T., Yau, C. Y., Chan, N. H. (2015). Likelihood inference for high-dimensional factor analysis of time series with applications in finance. *Journal of Computational and Graphical Statistics*, *24*, 866–884.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.
Owen, A. B. (2001). *Empirical likelihood*. New York: Chapman & Hall/CRC.
Pan, W., Shen, X., Liu, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research*, *14*, 1865–1889.
Qin, J., Lawless, J. F. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, *22*, 300–325.
Rogalski, R. J. (1984). New findings regarding day-of-the-week returns over trading and non-trading periods: A note. *The Journal of Finance*, *39*, 1603–1614.
Romano, J. P., Shaikh, A., Wolf, M. (2011). Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, *7*(1), 1–25.
Sonnemann, E. (2008). General solutions to multiple testing problems. *Biometrical Journal*, *50*, 641–656. (translation with minor corrections of the original article Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. *EDV in Medizin und Biologie* 13, 120–128 by Helmut Finner).
Steeley, J. M. (2001). A note on information seasonality and the disappearance of the weekend effect in the UK stock market. *Journal of Banking and Finance*, *25*, 1941–1956.
Tang, C. Y., Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika*, *97*, 905–920.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.
Tibshirani, R. J., Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics*, *39*, 1335–1371.
Tsao, M., Wu, C. (2006). Empirical likelihood inference for a common mean in the presence of heteroscedasticity. *The Canadian Journal of Statistics*, *34*, 45–59.
Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrics*, *5*(2), 99–114.
Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536.
Variyath, A. M., Chen, J., Abraham, B. (2010). Empirical likelihood based variable selection. *Journal of Statistical Planning and Inference*, *140*, 971–981.
Wang, H., Leng, C. (2007). Unified LASSO estimation via least squares approximation. *Journal of the American Statistical Association*, *101*, 1418–1429.
Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, B*, *71*, 671–683.
Wu, C., Yan, Y. (2012). Empirical likelihood inference for two-sample problems. *Statistics and Its Interface*, *5*, 345–354.
Wu, T. T., Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, *2*, 224–244.
Xie, B., Pan, W., Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, *2*, 168–212.
Zhang, C. H., Zhang, T. (2012). *A general framework of dual certificate analysis for structured sparse recovery problems*. ArXiv, e-prints. arXiv:1201.3302.
Zhao, H., Wang, B., Cui, X. (2010). General solutions to consistency problems in multiple hypothesis testing. *Biometrical Journal*, *52*, 735–746.
Zhu, X., Qu, A. (2018). Cluster analysis of longitudinal profiles with subgroups. *Electronic Journal of Statistics*, *12*, 171–193.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of American Statistical Association*, *101*, 1418–1429.