Journal of Machine Learning Research 17 (2016) 1-25 Submitted 11/15; Revised 7/16; Published 10/16

A New Algorithm and Theory for Penalized Regression-based Clustering

Chong Wu* WUXX0845@UMN.EDU
Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

Sunghoon Kwon* SHKWON0522@GMAIL.COM
Department of Applied Statistics, Konkuk University, Seoul, South Korea
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

Xiaotong Shen XSHEN@UMN.EDU
School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

Wei Pan† WEIP@BIOSTAT.UMN.EDU
Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

Editor: Inderjit Dhillon

Abstract

Clustering is unsupervised and exploratory in nature. Yet, it can be performed through penalized regression with grouping pursuit, as demonstrated in Pan et al. (2013). In this paper, we develop a more efficient algorithm for scalable computation and a new theory of clustering consistency for the method. This algorithm, called DC-ADMM, combines difference of convex (DC) programming with the alternating direction method of multipliers (ADMM). This algorithm is shown to be more computationally efficient than the quadratic penalty based algorithm of Pan et al. (2013) because of the former's closed-form updating formulas. Numerically, we compare the DC-ADMM algorithm with the quadratic penalty algorithm to demonstrate its utility and scalability. Theoretically, we establish a finite-sample mis-clustering error bound for penalized regression based clustering with the L_0 constrained regularization in a general setting. On this ground, we provide conditions for clustering consistency of the penalized clustering method. As an end product, we put R package *prclust* implementing PRclust with various loss and grouping penalty functions available on GitHub and CRAN.

Keywords: Alternating direction method of multipliers (ADMM), Difference of convex (DC) programming, Clustering consistency, Truncated L_1-penalty (TLP).

1. Introduction

Clustering analysis separates a set of unlabeled data points into disparate groups, or clusters, based on some common properties of these points. It is a fundamental tool in machine learning, pattern recognition, and statistics, and has been widely applied in many fields, ranging from image processing to genetics. Clustering analysis has a long history, and, naturally, a large number of clustering methods have been developed; see Jain (2010) for an excellent overview.

∗. These authors contributed equally.
†. WP is the corresponding author.

Clustering analysis is regarded as unsupervised learning in absence of a class label, as opposed to supervised learning. Over the last few years, a new framework of clustering analysis has been introduced by treating it as a penalized regression problem (Pelckmans et al., 2005; Lindsten et al., 2011; Hocking et al., 2011; Pan et al., 2013; Chi and Lange, 2015) based on over-parameterization. Specifically, we parameterize p-dimensional observations, say x_i, $1 \leq i \leq n$, with its own centroid, say μ_i. Two observations are said to belong to the same cluster if their corresponding μ_i's are equal. Then clustering analysis is formulated to identify a small subset of distinct values of these μ_i's via solving the following optimization problem

$$\min_{\mu} \quad \frac{1}{2}\sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \lambda \mathcal{J}(\mu),$$

where λ is a nonnegative tuning parameter controlling the trade-off between the model fit and the number of clusters, and $\mathcal{J}(\mu)$ is a penalty on $\mu = (\mu_1^{'}, \cdots, \mu_n^{'})^{'}$. Perhaps due to computational simplicity, a convex $\mathcal{J}(\mu)$ has been extensively studied. For example, sum-of-norms clustering (Lindsten et al., 2011) defines $\mathcal{J}(\mu) = \sum_{j=1}^{n}\sum_{i<j} ||\mu_i - \mu_j||_q$, where $||\cdot||_q$ is the L_q-norm. However, a convex $\mathcal{J}(\mu)$ usually yields biased parameter estimates, leading to difficulties in separating the clusters. To overcome this disadvantage, Pan et al. (2013) proposed penalized regression-based clustering (PRclust), which uses the non-convex grouped truncated lasso penalty (gTLP) $\mathcal{J}(\mu) = \sum_{i<j} \mathrm{TLP}\left(||\mu_i - \mu_j||_2; \tau\right)$. Specifically, TLP is defined as $\mathrm{TLP}(\alpha;\tau) = \min(|\alpha|, \tau)$ for a scalar α and a tuning parameter τ. It can be thought of as the L_1-penalty for a small $|\alpha| \leq \tau$, but no further penalization for a large $|\alpha| > \tau$. One benefit of PRclust is that it can treat some complex clustering situations, for example, in the presence of non-convex clusters, in which traditional methods such as K-means break down (Pan et al., 2013).

To deal with the nonseparable and non-convex grouping penalty in μ_i's, a quadratic penalty based algorithm (Pan et al., 2013) was developed by introducing some new parameters $\theta_{ij} = \mu_i - \mu_j$. This algorithm is relatively slow, and due to use of the quadratic penalty, the estimated centroids from the same cluster can never be exactly the same. To overcome these difficulties, we develop a novel and efficient computational algorithm called DC-ADMM, which combines the benefit of the alternating direction method of multipliers (ADMM) (Boyd et al., 2011) with that of the difference of convex (DC) method (Le Thi Hoai and Tao, 1997). As a result, DC-ADMM is much faster than the quadratic penalty based algorithm, in addition to that some estimated centroids can be exactly equal to each other when their corresponding observations come from the same cluster. As a by-product of this new method, we make R package *prclust* implementing both the quadratic penalty based algorithm and DC-ADMM available in CRAN (`https://cran.r-project.org`) and GitHub (`https://github.com/ChongWu-Biostat/prclust`).

Clustering consistency of PRclust remains unknown, though operating characteristics of PRclust have been studied via some simulations and real data analysis (Pan et al., 2013). In the penalized regression based clustering framework, clustering consistency of some related models has been studied (Radchenko and Mukherjee, 2014; Zhu et al., 2014). For example, Radchenko and Mukherjee (2014) studied clustering consistency of another method with univariate observations; Zhu et al. (2014) extended this result to multivariate observations by assuming only two clusters. In this paper,with some distributional assumptions, we

establish a general clustering consistency theory for a wide range of models, including PRclust as a special case. Our theory is applicable to multiple clusters and provide a finite-sample mis-clustering error bound in the absence of overlapping clusters. On this ground, we give sufficient conditions for PRclust to correctly identify clusters in terms of the expected Hellinger loss. As a result, PRclust not only reconstructs the true clusters, but also yields optimal parameter estimation through the L_0 grouping penalty.

The remaining of this paper is organized as follows. Section 2 introduces the new DC-ADMM algorithm and discusses a stability criterion to select the tuning parameters. A simulation study is then performed to demonstrate the numerical performance of the new algorithm as compared to other methods. This is followed by a theory for accuracy of clustering in Section 3. A discussion of the results is given in Section 4. The proofs of the main results are given in an Appendix.

2. New Algorithm

To treat non-convexity more efficiently, we introduce a DC algorithm based on the ADMM, called DC-ADMM. We prove DC-ADMM yields a Karush-Kuhn-Tucker (KKT) solution, and some extensions are discussed.

2.1 DC-ADMM

DC-ADMM contains three steps: first, it rewrites the original unconstrained cost function into a constrained one and introduces some new variables to simplify optimization with respect to the non-convex grouping penalty; second, DC programming is applied to convert the non-convex optimization problem into a sequence of convex relaxations; third, each relaxed convex problem is solved by a standard ADMM.

First, rewrite the PRclust cost function

$$\min_{\mu} \quad \frac{1}{2}\sum_{i=1}^{n}||x_i-\mu_i||_2^2+\lambda\sum_{i<j}\mathrm{TLP}\left(||\mu_i-\mu_j||_2;\tau\right) \tag{1}$$

as the equivalent constrained problem

$$\begin{aligned}\min_{\mu,\theta} \quad & S(\mu,\theta)=\frac{1}{2}\sum_{i=1}^{n}||x_i-\mu_i||_2^2+\lambda\sum_{i<j}\mathrm{TLP}\left(||\theta_{ij}||_2;\tau\right)\\ \text{subject to} \quad & \theta_{ij}=\mu_i-\mu_j, \quad 1\le i<j\le n,\end{aligned}$$

where $||\cdot||_2$ is the L_2-norm. Here, we introduce new variables $\theta_{ij}=\mu_i-\mu_j$ for the differences between the centroids and thus simplify optimization with respect to the grouping penalty.

To treat the non-convex gTLP on θ_{ij}'s, we apply DC programming (Le Thi Hoai and Tao, 1997). In particular, the cost function $S(\mu,\theta)$ is decomposed into a difference of two convex functions $S(\mu,\theta)=S_1(\mu,\theta)-S_2(\theta)$:

$$\begin{aligned}S_1(\mu,\theta)=&\frac{1}{2}\sum_{i=1}^{n}||x_i-\mu_i||_2^2+\lambda\sum_{i<j}||\theta_{ij}||_2,\\ S_2(\theta)=&\lambda\sum_{i<j}(||\theta_{ij}||_2-\tau)_+,\end{aligned}$$

where $(\alpha)_+$ denotes the positive part of α, which is α if $\alpha > 0$ and 0 otherwise.

Given the DC composition, we construct a sequence of upper approximations of $S(\mu, \theta)$ iteratively by replacing $S_2(\theta)$ at iteration $m+1$ with its piecewise affine minorization

$$S_2^{(m)}(\theta) = S_2\left(\hat{\theta}^{(m)}\right) + \lambda \sum_{i<j} \left(||\theta_{ij}||_2 - ||\hat{\theta}_{ij}^{(m)}||_2\right) I\left(||\hat{\theta}_{ij}^{(m)}||_2 \geq \tau\right)$$

at the current estimate $\hat{\theta}^{(m)}$ from iteration m, leading to an upper convex approximating function at iteration $m+1$:

$$\begin{aligned} S^{(m+1)}(\mu, \theta) = & \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 \\ & + \lambda \sum_{i<j} \left(||\theta_{ij}||_2\right) I\left(||\hat{\theta}_{ij}^{(m)}||_2 < \tau\right) + \lambda\tau \sum_{i<j} I\left(||\hat{\theta}_{ij}^{(m)}||_2 \geq \tau\right), \end{aligned} \tag{2}$$

where $I(\cdot)$ is the indicator function.

Then apply ADMM to solve the corresponding constrained convex problem at iteration $m+1$

$$\min_{\mu,\theta} \quad S^{(m+1)}(\mu, \theta), \qquad \text{subject to } \theta_{ij} = \mu_i - \mu_j, \quad 1 \leq i < j \leq n. \tag{3}$$

ADMM solves (3) by minimizing the corresponding scaled augmented Lagrangian

$$\begin{aligned} L_\rho(\mu, \theta) = & \frac{1}{2} \sum_{i=1}^{n} ||x_i - \mu_i||_2^2 + \lambda \sum_{i<j} \left(||\theta_{ij}||_2\right) I\left(||\hat{\theta}_{ij}^{(m)}||_2 < \tau\right) + \lambda\tau \sum_{i<j} I\left(||\hat{\theta}_{ij}^{(m)}||_2 \geq \tau\right) \\ & + y' \sum_{i<j} \left(\theta_{ij} - (\mu_i - \mu_j)\right) + (\rho/2) \sum_{i<j} ||\theta_{ij} - (\mu_i - \mu_j)||_2^2, \end{aligned} \tag{4}$$

where the dual variable y is a vector of Lagrange multipliers and ρ is a nonnegative penalty parameter. Using the scaled Lagrange multiplier $u = y/\rho$ (Boyd et al., 2011, §3.3.1), we can express ADMM as

$$\begin{aligned} \hat{\mu}_i^{k+1} = & \operatorname*{argmin}_{\mu_i} \frac{1}{2} ||x_i - \mu_i||_2^2 + \frac{\rho}{2} \sum_{j>i} ||\hat{\theta}_{ij}^k - (\mu_i - \hat{\mu}_j^k) + \hat{u}_{ij}^k||_2^2 \\ & + \frac{\rho}{2} \sum_{j<i} ||\hat{\theta}_{ij}^k - (\mu_i - \hat{\mu}_j^{k+1}) + \hat{u}_{ij}^k||_2^2, \\ \hat{\theta}_{ij}^{k+1} = & \operatorname*{argmin}_{\theta_{ij}} \begin{cases} \lambda\tau + \frac{\rho}{2} ||\theta_{ij} - (\hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1}) + \hat{u}_{ij}^k||_2^2, & \text{if} \quad ||\hat{\theta}_{ij}^{(m)}||_2 \geq \tau; \\ \lambda ||\theta_{ij}||_2 + \frac{\rho}{2} ||\theta_{ij} - (\hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1}) + \hat{u}_{ij}^k||_2^2, & \text{if} \quad ||\hat{\theta}_{ij}^{(m)}||_2 < \tau; \end{cases} \\ \hat{u}_{ij}^{k+1} = & \hat{u}_{ij}^k + \hat{\theta}_{ij}^{k+1} - (\hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1}), \quad 1 \leq i < j \leq n, \end{aligned} \tag{5}$$

where k stands for step k in the standard ADMM. Using some simple algebra, we obtain the updating formula for μ as follows

$$\hat{\mu}_i^{k+1} = \frac{x_i + \rho \sum_{j>i} \left(\hat{\mu}_j^k + \hat{\theta}_{ij}^k + \hat{u}_{ij}^k\right) + \rho \sum_{j<i} \left(\hat{\mu}_j^{k+1} - \hat{\theta}_{ji}^k - \hat{u}_{ij}^k\right)}{1 + \rho(n-1)}.$$

Applying a block soft thresholding operator for the group lasso penalty (Yuan and Lin, 2006), we have

$$\hat{\theta}_{ij}^{k+1} = \begin{cases} \hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1} - \hat{u}_{ij}^k & \text{if} \quad ||\hat{\theta}_{ij}^{(m)}||_2 \geq \tau; \\ \text{ST}\left(\hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1} - \hat{u}_{ij}^k; \lambda/\rho\right) & \text{if} \quad ||\hat{\theta}_{ij}^{(m)}||_2 < \tau; \end{cases} \tag{6}$$

where $\text{ST}(\theta;\gamma) = (||\theta||_2 - \gamma)_+\theta/||\theta||_2$. The convergence time of ADMM is highly related to the penalty parameter ρ. A poor selection of ρ can result in a slow convergence for the ADMM algorithm (Ghadimi et al., 2015) and thus DC-ADMM. In this paper, we fix $\rho = 0.4$ throughout for simplicity. For the subsequent relaxed convex problem (3), $\hat{\mu}^{(m+1)}$ and $\hat{\theta}^{(m+1)}$ are updated according to standard ADMM (5) until some stopping criteria, such as that both dual and primal residuals are small (Boyd et al., 2011), are met. We summarize the DC-ADMM algorithm in Algorithm 1.

Algorithm 1: DC-ADMM for penalized regression based clustering

Input : n observations $X = \{x_1, \ldots, x_n\}$; tuning parameters λ, τ and ρ.

Initialize: Set $m = 0$, $\hat{u}_{ij}^{(0)} = 0$, $\hat{\mu}_i^{(0)} = x_i$ and $\hat{\theta}_{ij}^{(0)} = x_i - x_j$ for $1 \leq i < j \leq n$.

while $m = 0$ *or* $S(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}) - S(\hat{\mu}^{(m-1)}, \hat{\theta}^{(m-1)}) < 0$ **do**

  $m \leftarrow m + 1$

  Update $\hat{\mu}^{(m)}$ and $\hat{\theta}^{(m)}$ based on (5) until convergence with a standard ADMM.

end

Output: Estimated centroids for the observations, $\hat{\mu}_1, \ldots, \hat{\mu}_n$, from which a cluster label for each observation is assigned.

In Algorithm 1, for each iteration m, $\hat{\mu}_i^0 = x_i$ and $\hat{\theta}^0 = x_i - x_j$ for $1 \leq i < j \leq n$ are used as the starting values for (5); $(\hat{\mu}^{(m+1)}, \hat{\theta}^{(m+1)})$ is the limit point of the ADMM iterations in (5), or equivalently, is a minimizer of (3). $(\hat{\mu}^{(m+1)}, \hat{\theta}^{(m+1)})$ is then used to update the objective function $S^{(m+1)}(\mu, \theta)$ in (2) as a new approximation to $S(\mu, \theta)$. The process is iterated until the stopping criteria are met.

Since the cost function (3) is a sum of a differentiable and convex function and a convex penalty in θ (while $\hat{\theta}^{(m)}$ is known), ADMM converges to its minimizer (Boyd et al., 2011). Then DC-ADMM's convergence in a finite number of steps follows by the facts that DC programming guarantees the decrease of the subsequent convex relaxations (2), and that $S^{(m+1)}(\mu, \theta)$ has only a finite set of possible forms across all m. Theorem 1 shows that the solution of the DC-ADMM converges to a KKT point.

Theorem 1 *In the DC-ADMM, $S\left(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}\right)$ converges in a finite number of steps; that is, there exists an $m^* < \infty$ with*

$$S\left(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}\right) = S\left(\hat{\mu}^{(m^*)}, \hat{\theta}^{(m^*)}\right) \qquad \textit{for} \quad m \geq m^*$$

Furthermore, $\left(\hat{\mu}^{(m^*)}, \hat{\theta}^{(m^*)}\right)$ is a KKT point.

DC-ADMM only guarantees a local instead of a global minimizer. As shown in simulations, DC-ADMM performed well in terms of clustering accuracy. This suggests that DC-ADMM typically yields a good local solution, though not necessarily global. A variant of DC algorithms called outer approximation method of Breiman and Cutler (1993) gives a global minimizer, but may converge slowly. For a large-scale problem, we prefer the present version for its faster convergence at an expense of possibly missing global solutions.

With different random starting values, DC-ADMM could yield different KKT points for the same data and parameters. However, our limited numerical experience suggests that DC-ADMM gives good solutions with our proposed starting values.

Let N_{admm}, N_{quad} be the numbers of iterations for running the standard ADMM and quadratic based algorithm, respectively. The computational complexity of updating θ and μ for one time is $O(pn^2)$. Note that the complexity of DC programming is $O(1)$ and N_{admm} typically scales as $O(1/\epsilon)$, where ϵ is the tolerance (He and Yuan, 2015). Then for the DC-ADMM algorithm, the computational complexity is $O(pn^2/\epsilon)$. In contrast, based on the empirical experience, N_{quad} relates to the number of observations n and quadratic based algorithm is much slower than DC-ADMM. In practice, especially in earlier iterations, one may not want to run the ADMM updates fully until convergence to save computing time. Another trick is that for the subsequent convex relaxations, we can initialize (warm start) $\hat{\mu}^0$, $\hat{\theta}^0$ and $\hat{u}^0$ at their optimal values from the previous relaxed convex problem, which significantly reduces the number of ADMM iterations.

In the DC-ADMM, the hard constraint guarantees that we can obtain exactly some $\hat{\mu}_i - \hat{\mu}_j - \hat{\theta}_{ij} = 0$; in contrast, in the quadratic penalty based algorithm (Pan et al., 2013), due to the use of soft constraint, we cannot obtain exactly $\hat{\mu}_i - \hat{\mu}_j - \hat{\theta}_{ij} = 0$ no matter how large the finite tuning parameter is chosen. Pan et al. (2013) provided an alternative algorithm (PRclust2) to force some $\hat{\mu}_i - \hat{\mu}_j - \hat{\theta}_{ij} = 0$ by running the quadratic based algorithm several times. Although PRclust2 leads to similar clustering results as DC-ADMM in our simulations, it is on average around 10 to 30 times slower than the quadratic based algorithm and is not feasible to large data sets.

2.2 Selection of the Number of Clusters

A generalized degrees of freedom (GDF) together with generalized cross validation (GCV) was proposed for selection of tuning parameters for clustering (Pan et al., 2013). This method, while yielding good performance, requires extensive computation and specification of a hyper-parameter, perturbation size. Here, we provide an alternative by modifying a stability-based criterion (Tibshirani and Walther, 2005; Liu et al., 2016) for determining the tuning parameters.

The main idea of the method is based on cross-validation. That is, (1) randomly partition the entire data set into a training set and a test set with an almost equal size; (2) cluster the training and test sets separately via PRclust with the same tuning parameters; (3) measure how well the training set clusters predict the test clusters. To be specific, first, randomly partition the entire data set into a training set X_{tr} and a test set X_{te} with a roughly equal size. Second, apply DC-ADMM (Algorithm 1) with the same tuning parameters to X_{tr} and X_{te}, leading to the corresponding clustering assignments l_{tr} and l_{te}, respectively. Third, assign X_{te} to clusters according to l_{tr}; that is, assign each observation

in X_{te} to the closest cluster of X_{tr} defined by l_{tr} in terms of the Euclidean distance, with $l_{\text{te|tr}}$ the corresponding clustering assignments. Note that the distance between an observation in X_{te} and a cluster of X_{tr} is the minimum distance between the observation and each observations in the cluster. To measure how well the training set clusters predict the test clusters, we compute the adjusted Rand index (Hubert and Arabie, 1985) between $l_{\text{te|tr}}$ and l_{te} as the prediction strength. Recall that the adjusted Rand index ranges between 0 and 1 with a higher value indicating a higher agreement. Repeat the above process T times and calculate the average prediction strength as the mean of T different prediction strengths. This process is repeated over various tuning parameter values, obtaining their corresponding average prediction strengths, then choose the set of the tuning parameters with the maximum average prediction strength. The intuition behind this idea is that if the tuning parameters lead to a stable clustering result, then the training set clusters will be similar to the test set clusters, and hence will predict them well, leading to a high average prediction strength.

2.3 Extensions

The K-means method uses squared L_2-norm distances to generate cluster centroids, which may be inaccurate if outliers are present (Xu et al., 2005). In contrast, K-medians uses the L_1-norm distance and is more robust to outliers. Corresponding to modifying the K-means to K-medians, we can extend PRclust by replacing the squared L_2-norm with the L_1-norm loss function and estimate the centroids μ through minimizing the following cost function

$$\min_{\mu} \quad S_{L_1}(\mu) = \frac{1}{2}\sum_{i=1}^{n} ||x_i - \mu_i||_1 + \lambda \sum_{i<j} \text{TLP}\left(||\mu_i - \mu_j||_2; \tau\right).$$

Due to the nature of the DC-ADMM algorithm, we just need to change the updating formula for $\hat{\mu}$ and leave the remaining updating formula (5), (6) unchanged. Note that

$$\begin{aligned}\hat{\mu}_i^{k+1} = \underset{\mu_i}{\text{argmin}} \frac{1}{2}||x_i - \mu_i||_1 + \frac{\rho}{2}\sum_{j>i} ||\hat{\theta}_{ij}^k - (\mu_i - \hat{\mu}_j^k) + \hat{u}_{ij}^k||_2^2 \\ + \frac{\rho}{2}\sum_{j<i} ||\hat{\theta}_{ij}^k - (\mu_i - \hat{\mu}_j^{k+1}) + \hat{u}_{ij}^k||_2^2.\end{aligned}$$

To solve the above problem, we define $\nu_i = x_i - \mu_i$ and simplify the cost function with the L_1-loss:

$$\begin{aligned}\hat{\mu}_i^{k+1} = \underset{\mu_i}{\text{argmin}} \frac{1}{2}||\nu_i||_1 + \frac{\rho}{2}\sum_{j>i} ||\hat{\theta}_{ij}^k - (x_i - \nu_i - \hat{\mu}_j^k) + \hat{u}_{ij}^k||_2^2 \\ + \frac{\rho}{2}\sum_{j<i} ||\hat{\theta}_{ij}^k - (x_i - \nu_i - \hat{\mu}_j^{k+1}) + \hat{u}_{ij}^k||_2^2.\end{aligned}$$

Using simple algebra and the soft thresholding operator for lasso (Tibshirani, 1996), we obtain an updating formula as:

$$\hat{\mu}_i^{k+1} = \text{STL}\left(\frac{\sum_{j>i}\left(\hat{\mu}_j^k + \hat{\theta}_{ij}^k + \hat{u}_{ij}^k - x_i\right) + \sum_{j<i}\left(\hat{\mu}_j^{k+1} - \hat{\theta}_{ji}^k - \hat{u}_{ij}^k - x_i\right)}{n-1}, \frac{1}{2\rho(n-1)} \right) + x_i,$$

where $\text{STL}(\alpha, \gamma) = \text{sign}(\alpha)(|\alpha| - \gamma)_+$. In this case, the scalar operation on a vector is element-wise.

In addition, we can also use other penalty functions. In an appendix, we provide details of the DC-ADMM algorithm for PRclust with lasso or TLP as grouping penalty.

2.4 Simulations

Consider two overlapped convex clusters with the same spherical shape in two dimensions. Specifically, a random sample of $n = 100$ observations was generated, with 50 from a bivariate Gaussian distribution $N((0,0)', 0.33\mathbb{I})$, while the other 50 from $N((1,1)', 0.33\mathbb{I})$, where $\mathbb{I}$ is the identity matrix.

For PRclust, we searched $\tau \in \{0.1, 0.2, \ldots, 1\}$ and $\lambda \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7, 1, 1.5, 2\}$. To evaluate the performance of selecting the tuning parameters, we used the Rand index (Rand, 1971) and adjusted Rand index (Hubert and Arabie, 1985), measuring the agreement between estimated cluster and the truth with a higher value indicating a higher agreement. PRclust with the stability based criterion selecting its tuning parameters performed well: the average number of clusters was 2.63, slightly larger than the truth $K_0 = 2$; the correspond clustering results had high degrees of agreement with the truth, as evidenced by the high indices. Table 1 shows the frequencies of the number of clusters selected by the stability criterion: for the overwhelming majority (93%), either the correct number of cluster $K_0 = 2$ was selected, or a slightly larger $K = 3$ or 4 was selected. As expected, applying the quadratic penalty based algorithm with the stability criterion yielded a similar result. GCV with GDF yielded the similar results for clustering accuracy. However, to use GCV with GDF, the user has to specify the perturbation size, a hyper-parameter. In contrast, the stability based criterion is insensitive to the repeat times T. For the simulation, the average numbers of clusters selected with $T = 10, 50$ and 100 were 2.63, 2.68 and 2.76, respectively.

Now we illustrate differences between the two algorithms. First, we demonstrate how two algorithms operated differently with respect to various values of the tuning parameter λ, while τ was fixed at 0.7 (Figure 1). Note that, due to the soft constraint of the quadratic penalty based algorithm, we cannot obtain exactly $\hat{\mu}_i - \hat{\mu}_j - \hat{\theta}_{ij} = 0$. Even for a sufficiently large λ, there were still quite some unequal $\hat{\mu}_{i,1}$'s, which were all remarkably close to their true values 0 or 1. In contrast, due to using the hard constraint on $\theta_{ij} = \mu_i - \mu_j$, DC-ADMM yielded some equal estimated centroids $\hat{\mu}_{i,1}$. In this simulation, the stability based criterion tended to select the most stable tuning parameters, confirming its selecting good tuning parameters and yielding good clustering results.

Figure 2 shows the run-time of two algorithms against the number of observations n and dimension p. As a matter of fact, the DC-ADMM is much faster than the quadratic penalty

	Stability Based Criterion				GCV with GDF			
Algorithm	Freq	$\hat{K}$	Rand	aRand	Freq	$\hat{K}$	Rand	aRand
DC-ADMM	All	2.63	0.950	0.901	All	3.29	0.956	0.912
	60	2.00	0.954	0.908	39	2.00	0.958	0.917
	26	3.00	0.949	0.898	22	3.00	0.965	0.930
	7	4.00	0.945	0.890	17	4.00	0.959	0.918
	5	5.00	0.924	0.847	8	5.00	0.940	0.881
	2	6.00	0.952	0.903	12	6.00	0.947	0.894
Quadratic	All	2.70	0.951	0.902	All	2.41	0.962	9.925

Table 1: Comparison of the tuning parameter selection criteria based on 100 simulated data sets each with 2 clusters.

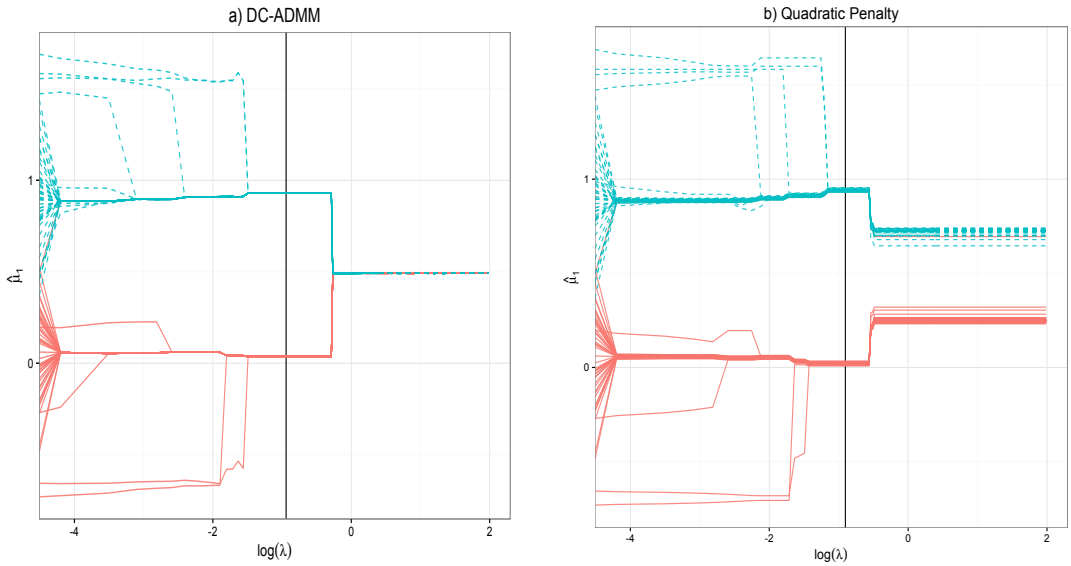

Figure 1: Solution paths of the first coordinate $\hat{\mu}_{i,1}$ for the first simulated data set. τ is fixed at 0.7. Vertical black line represents the tuning parameter selected by the stability based criterion.

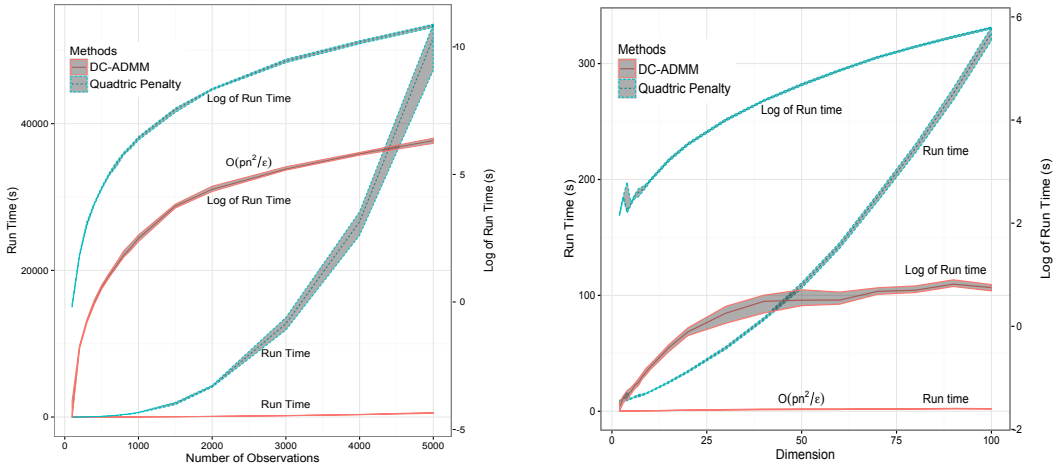

Figure 2: Comparison of run-times of DC-ADMM and quadratic penalty based algorithm based on the average of 100 simulations with different random seeds. Shaded regions represent the 25% and 75% quantiles of the run-times for corresponding algorithms. The complexity of DC-ADMM is $O(pn^2/\epsilon)$, whereas the quadratic penalty based algorithm is much slower.

based algorithm, particularly when either n or p is large. For DC-ADMM, the number of iteration was insensitive to the sample size and was around 100. In contrast, for the quadratic penalty based algorithm, it increased dramatically as the sample size increased; when the sample size was 200, the number of iteration was around $1,000$; however, the number of iteration increased to around $85,000$ when the sample size increased to $6,000$. The complexity of DC-ADMM is quadratic in the sample size n (the ratio of run-time to n^2 was around 10^{-5}) and linear in the dimension p (the ratio of run-time to p was around 0.05), confirming that the computational complexity is $O(pn^2/\epsilon)$.

Figure 3 shows the solution paths for other methods. PRclust2 provided very similar results as DC-ADMM (Figure 3a). However, PRclust2 is extremely slow (around 10 to 30 times slower than the quadratic penalty based algorithm) and not feasible to large data sets. Convex penalties, such as the lasso and the L_2-norm penalty, always shrink all the estimates towards zero and thus lead to severely biased parameter estimates. For example, if we used the L_2-norm (Figure 3b) or the lasso (Figure 3c) as the grouping penalty, the estimated centroids were shrunk towards each other, leading to their convergence to the same point at the end and thus much worse performance in clustering. The TLP (Shen et al., 2012) preformed much better than the lasso since it imposed no further penalty on large estimates (Figure 3d). Since the TLP does not borrow information from other variables, it performed slightly worse than its grouped version.

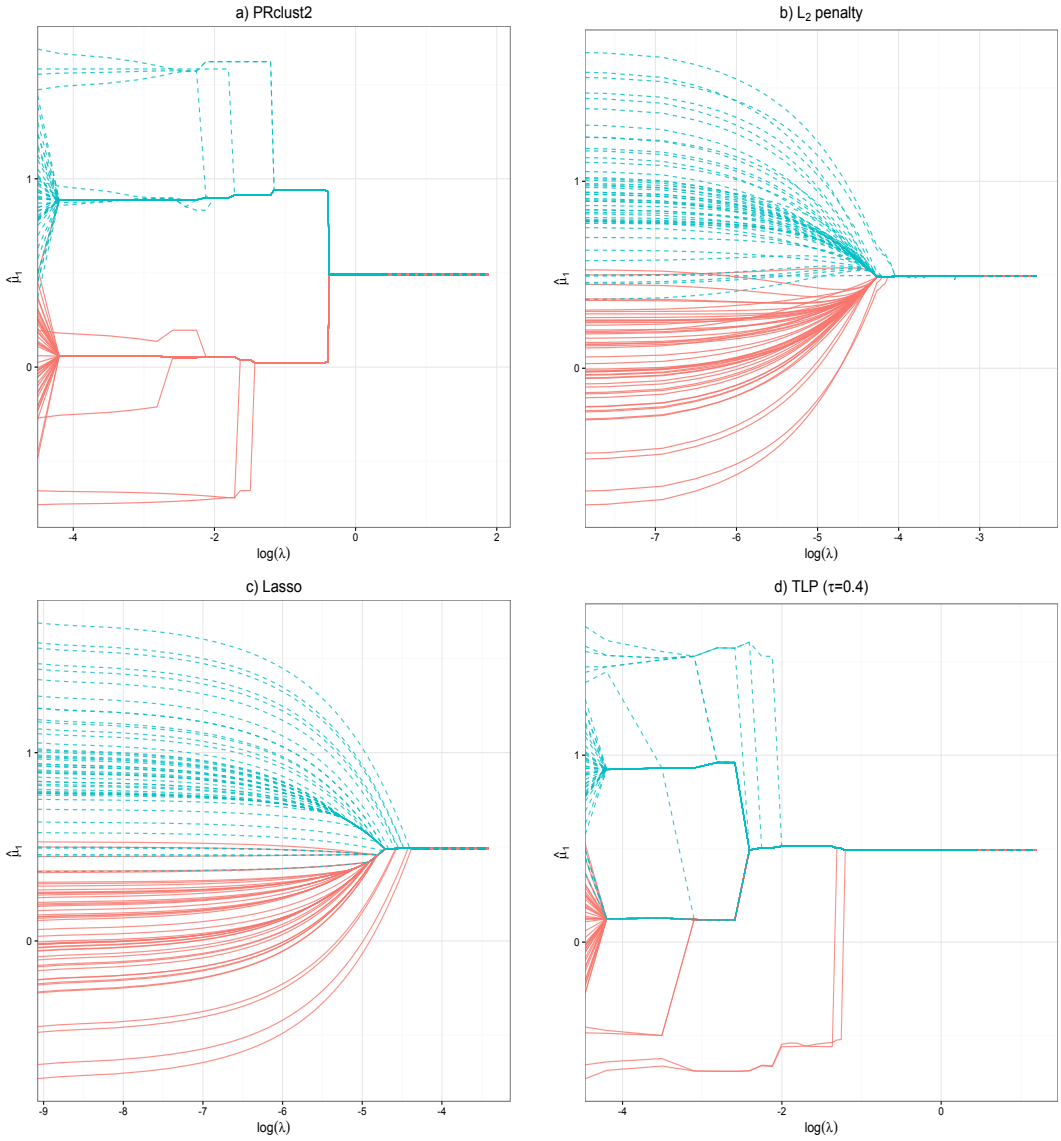

Figure 3: Solution paths of $\hat{\mu}_{i,1}$ for a) PRclust2, b) L_2 penalty, c) Lasso penalty and d) TLP for the first simulated data set.

3. Theory

Though operating characteristics of PRclust have been intensively studied, its clustering consistency properties remain unknown. In this section, based on the maximum likelihood estimation framework, we develop some theoretical properties for penalized regression based clustering method, which incorporates original PRclust (Pan et al., 2013) as a special case. Recall that PRclust does not put any distribution assumptions on the data; however, it can be treated as assuming a Gaussian distribution for the data implicitly as to be shown later. To avoid unaddressable complexity of over-parameterizing the underlying distribution, some mild technical assumptions are introduced. Then we develop a probability bound of clustering consistency which is slightly harder than clustering center consistency (Pollard, 1981).

3.1 PRclust in the Penalized Maximum Likelihood Framework

Assume $x_i \in \mathbb{R}^p \sim f_{\mu_i}(\cdot)$, $1 \leq i \leq n$ are n independent random samples, where f_{μ_i} is a probability density function of x_i with its centroid $\mu_i \in \mathbb{R}^p$. We obtain an estimate $\hat{\mu}^{L_0}$ of $\mu = (\mu_1^{'}, \ldots, \mu_n^{'})^{'} \in \mathbb{R}^{pn}$ via solving the following constrained L_0-problem:

$$\min_{\mu} \quad -\mathcal{L}(\mu) \quad \text{subject to} \quad \mathcal{J}(\mu) \leq J, \tag{7}$$

where J is a nonnegative tuning parameter controlling the trade-off between the model fit and the number of clusters, $\mathcal{L}(\mu) = \sum_{i=1}^{n} \log(f_{\mu_i}(x_i))$ is the log-likelihood that corresponds to the model fit, and $\mathcal{J}(\mu) = \sum_{i<j} I\{d(\mu_i, \mu_j) \neq 0\}$ is the grouping penalty that controls the number of clusters. $I(\cdot)$ is the indicator function and $d(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a distance, which can be defined $d(\mu_i, \mu_j) = ||\mu_i - \mu_j||_q = \{\sum_{m=1}^{p} |\mu_{im} - \mu_{jm}|^q\}^{1/q}$, $0 < q < \infty$. Then $\mathcal{J}(\mu)$ equals the number of distinct pairs of centroids $\mu_i \neq \mu_j$.

The regularization problem (7) is a constrained counterpart of the following penalized unconstrained L_0-problem:

$$\min_{\mu} \quad -\mathcal{L}(\mu) + \lambda \mathcal{J}(\mu), \tag{8}$$

where $\lambda \geq 0$ is a tuning parameter corresponding to J in (8). Note that (7) and (8) may not be equivalent in their global minimizers, which is unlike a convex problem.

In a high-dimensional situation, it is not computationally feasible to minimize a discontinuous cost function in (8) and (7). As a surrogate, we consider an estimator $\hat{\mu}^{tL_1}$ that minimizes the following truncated L_1-problem:

$$\min_{\mu} \quad -\mathcal{L}(\mu) + \lambda \mathcal{J}_{\tau}(\mu), \tag{9}$$

where $\mathcal{J}_{\tau}(\mu) = \sum_{i<j} \text{TLP}(d(\mu_i, \mu_j); \tau)$. Note that if assuming $x_i \sim MVN(\mu_i, \sigma^2 \mathbb{I})$, $1 \leq i \leq n$ and using L_2-distance, we get $-\mathcal{L}(\mu) = \sum_{i=1}^{n} ||x_i - \mu_i||_2^2$ after ignoring some constants and $\mathcal{J}_{\tau}(\mu) = \sum_{i<j} \text{TLP}\left(||\mu_i - \mu_j||_2; \tau\right)$, which indicate that (9) reduces to the original PRclust (1) under multivariate Gaussian distribution assumption. When τ is sufficiently small, the truncated L_1 constraint has a good approximation to the L_0 loss (Shen et al., 2012).

3.2 A Fundamental Assumption for Over-parameterization

To reduce the unaddressable complexity to an addressable level, we propose a fundamental assumption. Let $C_k, 1 \leq k \leq K$ be K clusters that satisfy $\cup_{k=1}^{K} C_k = \{x_1, \ldots, x_n\}$ and $C_i \cap C_j = \emptyset$, for $1 \leq i \neq j \leq K$. The number of partitions of n samples into K clusters is $(1/K!)\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n$, which in turn can be approximated by $K^n/K!$ (Steinley, 2006). Since PRclust is based on over-parameterization and assumes one parameter (centroid) for one corresponding sample, the complexity of PRclust is the same as all possible ways of constructing clusters based on all samples. Unfortunately, to the best of our knowledge, there is no possible probability bound that can cover this complexity that requires tail probability decreasing faster than $\exp(-n \log K)$. However, many of the clustering formulation lead to the overlapped clusters and there is no way to reconstruct the true clusters exactly. To recover non-overlapped true clusters, we put a mild technical restriction on the clustering formulation to reduce the complexity.

Assumption (A0): Partition samples $x_1, x_2, \ldots, x_n$ into K clusters. For any clusters $C_1, C_2, \ldots, C_K$, there exists m points $y_1^{(k)}, \ldots, y_m^{(k)} \in C_k$ such that $d(\overline{y}_m^{(k)}, x_k) \leq d(\overline{y}_m^{(k)}, x_c)$ for all $x_k \in C_k$ and $x_c \in C_k^c$, where $\overline{y}_m^{(k)} = \sum_{l=1}^{m} y_l/m$ and A^c denotes the complement of a set A. We define m as the *minimal disjoint centering number*.

Note that all the clusters are separated under (A0). Violating (A0) implies that there exist $x_k \in C_k$ and $x_c \in C_k^c$ such that $d(\overline{y}_m^{(k)}, x_k) > d(\overline{y}_m^{(k)}, x_c)$, indicating that there exists another cluster that overlaps with C_k. Interestingly, assumption of this kind seems necessary because clustering consistency is impossible when some clusters overlap, although it appears strong. Worth of note is that other papers, for instance Zhu et al. (2014), explicitly assume that different clusters are reasonably separable from each other for clustering consistency. Furthermore, (A0) excludes any irregular cluster structures, which are not constructed. Most importantly, Lemma 1 in the Appendix gives an upper bound of the number of ways of reconstructable clusters under (A0), reducing the overparameterization complexity from the super-exponential level in the sample size n, $\exp(-n \log K)$, to a polynomial level in n, $\exp(-mK \log n)$. Lastly, (A0) implies that all the clusters must include at least m samples, and guarantees cluster-center consistency asymptotically: for each $1 \leq k \leq K, \|\overline{y}_m^{(k)} - \mu_k\|_2 \to 0$ almost surely as $m \to \infty$, where μ_k is the centroid of the cluster C_k. Note that Pollard (1981) used a similar assumption for cluster-center consistency for the k-means method.

3.3 Clustering Consistency for L_0-constrained Problem

Define $\mathcal{C} = \{\mathcal{C}(\mu) : \mu \in \mathbb{R}^{pn}\}$, where $\mathcal{C}(\mu) = \{C_1, \ldots, C_K\}$ is a set of clusters based on μ such that for any cluster $\mathcal{C}_k$, $d(\mu_i, \mu_j) = 0$, $\forall i, j \in C_k$ and $d(\mu_i, \mu_j) \neq 0$, $\forall i \in C_k, j \in C_k^c$. Let $\mu^* = (\mu_1^{*'}, \ldots, \mu_n^{*'})' \in \mathbb{R}^{pn}$ with $\mu_i^* = (\mu_{i1}^*, \ldots, \mu_{ip}^*)' \in \mathbb{R}^p$ be the true centroid. We study asymptotic properties of $\hat{\mu}^{L_0}$ in (7) by giving a bound of the incorrect clustering probability: $P\left(\hat{\mu}^{L_0} \neq \hat{\mu}^o\right)$, where $\hat{\mu}^o = (\hat{\mu}_1^{o'}, \ldots, \hat{\mu}_n^{o'})' = \text{argmin}_{\mathcal{C}(\mu)=\mathcal{C}(\mu^*)}\mathcal{L}(\mu)$ is the oracle estimator that is usually unavailable unless the true clusters are known beforehand. Note that $\hat{\mu}^{L_0}$ is defined as a global minimizer of (7) and assume to be any global minimizer.

Before proceeding, we define a complexity measure for a given function space $\mathcal{F}$. For any $\epsilon > 0$, let $H(\epsilon, \mathcal{F})$ be the logarithm of the cardinality of the ϵ-bracketing of $\mathcal{F}$ of the

smallest size. To be specific, let $S(\epsilon, \mathcal{F}, r) = \{f_1^l, f_1^u, \ldots, f_r^l, f_r^u\}$ be the bracket covering of $\mathcal{F}$ that satisfies $\max_{1\le j\le r} ||f_j^u - f_j^l||_2 \le \epsilon$, where $||f||_2 = (\int f^2 dv)^{1/2}$ and there exists a j such that $f_j^l \le f \le f_j^u$ for any $f \in \mathcal{F}$, then $H(\epsilon, \mathcal{F}) = \log(\min\{r : S(\epsilon, \mathcal{F}, r)\})$. For more discussions about metric entropy of this type, see Kolmogorov and Tikhomirov (1959). To construct the clustering consistency properties, we need the following two assumptions.

Assumption (A1): There exists some constant $d_0 > 0$ such that, for any $\epsilon > 0$,

$$\sup_{C\in\mathcal{C}:|C|\le|\mathcal{C}(\mu^*)|} H(t, \mathcal{F}_C) \le d_0 m \log(n) \log(\epsilon^2/2^8 t), \quad \epsilon^2/2^8 < t < 2^{1/2}\epsilon \le 1,$$

where $\mathcal{F}_C = \{f_\mu : h_a^2(f_\mu, f_{\mu^*}) \le \epsilon^2, \mu \in \mathcal{B}_C\}$, f_μ is the density of $x = (x_1', \ldots, x_n')'$, $\mathcal{B}_C = \{\mu : \mathcal{C}(\mu) = C\}$, $|A|$ is the cardinality of a set A and m is the *minimal disjoint centering number* defined in (A0).

Note that (A1) puts some constraints on the size of parameter space, which is similar as Assumption A in Shen et al. (2012) and is a direct modification of the assumption in Wong and Shen (1995).

Define

$$\mathcal{C}_{\min}(\mu^*) \equiv \inf_{\mu\in\mathcal{B}} \frac{h_a^2(f_\mu, f_{\mu^*})}{|\mathcal{C}(\mu)|}$$

to be the degree-of-separation or the level of difficulty of clustering, where $\mathcal{B} = \{\mu : \mathcal{C}(\mu) \ne \mathcal{C}(\mu^*), \mathcal{J}(\mu) \le \mathcal{J}(\mu^*)\}$ is a parameter space of interest, $h_a(f_\mu, f_{\mu^*}) = \sum_{i=1}^n h(f_{\mu_i}, f_{\mu_i^*})/n^{1/2}$ is the averaged Hellinger metric with $h(f_{\mu_i}, f_{\mu_i^*}) = \left\{\frac{1}{2}\int (f_{\mu_i}^{1/2} - f_{\mu_i^*}^{1/2})^2 dv\right\}^{1/2}$.

Assumption (A2): There exists some constant $d_1 > 0$ such that

$$\mathcal{C}_{\min}(\mu^*) > d_1 m \log(n)/n,$$

where m is the *minimal disjoint centering number* defined in (A0).

Assumption (A2) describes the least favorable situation through $\mathcal{C}_{\min}(\mu^*)$ under which we can identify the true cluster partition. In fact, $\mathcal{C}_{\min}(\mu^*)$ depends on the number of true clusters and the minimum distance among true cluster centers induced by the Hellinger loss. Since (A2) puts some regularity conditions on log-likelihood function via Hellinger loss, we do not make any regularity conditions for the log-likelihood function explicitly. Similar assumptions as (A2) can be found in the literature of feature selection. For example, Shen et al. (2012) assumed

$$C_{\min}(\beta^*) \ge d_0 \log(p)/n, \tag{10}$$

where β^* is a true parameter vector of interest, d_0 is a positive constant, p is the dimension of β and n is the sample size. By assuming a probabilistic model such as the Gaussian distribution, (10) can be further specified as

$$\gamma_{\min}^2 = \min_{j:\beta_j^*\ne 0} |\beta_j^*| \ge d_0 \log(p)/n,$$

which implies the feature selection consistency can be constructed even when the minimal signal size is vanishing $\gamma_{\min} \to 0$ and the dimension of features is diverging $p \to \infty$. This assumption is much weaker than classical assumptions where $\gamma_{\min}$ and p are usually fixed constants.

(A2) serves similar roles for clustering consistency as (10) for feature selection consistency. As to be shown later in Proposition 1, by assuming the Gaussian distribution, (A2) can be explicitly specified. More importantly, it allows the minimum distance among different cluster centroids decreases toward zero, $\alpha_{\min} = \min_{\mu_i^* \neq \mu_j^*} \|\mu_i^* - \mu_j^*\|^2 \to 0$, and the number of cluster diverge to infinity, $K \to \infty$, indicating that the assumption used here is weaker than many other studies where $\alpha_{\min}$ and K are usually fixed constants (Pollard, 1981; Pelckmans et al., 2005; Radchenko and Mukherjee, 2014; Zhu et al., 2014). Then we establish the main theory for clustering consistency as follows.

Theorem 2 *Under Assumptions (A0) to (A2), if $J = \mathcal{J}(\mu^*)$, then, there exists some constant $c_2 > 0$, such that*

$$P\left(\hat{\mu}^{L_0} \neq \hat{\mu}^o\right) \leq \exp\left(-c_2 n \mathcal{C}_{\min}(\mu^*) + (m+1)\log(n) + 2\right),$$

provided that $d_1 > \max\{1/c_2, 2d_0(\log c_3)/c_4^2\}$. For example, we may use $c_2 = 4/(27 \times 1926)$, $c_3 = 10$ and $c_4 = (2/3)^{5/2}/512$. Further, $\hat{\mu}^{L_0}$ reconstructs the oracle estimator $\hat{\mu}^o$ with probability tending to one as $n \to \infty$. The following two asymptotic results hold as $n \to \infty$:
(A) (Clustering consistency) $P(\hat{\mu}^{L_0} \neq \hat{\mu}^o) \to 0$ and hence $P\left(\mathcal{C}(\hat{\mu}^{L_0}) \neq \mathcal{C}(\hat{\mu}^*)\right) \to 0$.
(B) (Optimal parameter estimation) $E\left[h_a^2(f_{\hat{\mu}^{L_0}}, f_{\mu^*})\right] = (1+o(1))E\left[h_a^2(f_{\hat{\mu}^o}, f_{\mu^*})\right]$, provided that $c_2 n \mathcal{C}_{\min}(\mu^*) + \log\left(E\left[h_a^2(f_{\hat{\mu}^o}, f_{\mu^*})\right]\right) \to \infty$.

Theorem 2 says that, under Assumptions (A0) to (A2), $\hat{\mu}^{L_0}$ consistently reconstructs the oracle estimator $\hat{\mu}^o$, and both an oracle clustering and an optimal parameter estimation with respect to expected Hellinger risk are asymptotically available by solving a constrained L_0-problem. As pointed out by a reviewer, the number of clusters K is an important but unknown tuning parameter. Theorem 2 shows that if the tuning parameter J is chosen to be $J = |\mathcal{J}(\mu^*)|$ then optimal clustering can be constructed asymptotically. We believe that theory established here can be a starting point in developing some new tuning parameter selection criteria, though we have not fully explored in this aspect here. Theorem 2 provides an insight into or gives theoretical justification on when or under which condition the proposed method is expected to give correct clustering. For instance, the theory suggests that the optimal tuning parameters may depend on the underlying true parameters, which needs to be estimated for real data. This, together with the tuning parameter selection criterion lead to the estimated data-dependent tuning parameter for this real data set.

Although theoretical properties of penalized clustering have been intensively studied (Radchenko and Mukherjee, 2014; Zhu et al., 2014), our result is new and different from the proceeding ones. For example, Radchenko and Mukherjee (2014) proved clustering consistency with univariate samples, which are not practical and, in fact, relatively easy to prove in our context without assumption (A0) since the complexity of over-parametrization falls down to an addressable level. Zhu et al. (2014) extended the clustering consistency to multivariate samples by assuming there are only two clusters, say C_1 and C_2 with centroids μ_1 and μ_2, respectively. To avoid some technical difficulties, Zhu et al. (2014) imposed an assumption that is not required in Theorem 2: two clusters C_1 and C_2 consist proportional number of samples in the sense that $|C_1|/|C_2| \to c$, where c is a positive constant. Theorem 2 established here extended clustering consistency to a more realistic situation: multivariate samples with many clusters.

3.4 Example: Truncated Multivariate Gaussian Distributions

In this example, we give a sufficient condition for (A2) to hold asymptotically, by constructing a lower bound of $C_{\min}(\mu^*)$ in terms of the minimum center distance $\alpha_{\min} = \min_{\mu_i^* \neq \mu_j^*} \|\mu_i^* - \mu_j^*\|_2$. Let $\phi_{\mu_i}, 1 \leq i \leq n$ be the multivariate Gaussian density function with mean $\mu_i \in \mathbb{R}^p$ and identity covariance matrix $I_{p\times p}$, that is, $\phi_{\mu_i}(z) = (2\pi)^{-p/2}\exp(-\|z - \mu_i\|_2^2/2), z \in \mathbb{R}^p, 1 \leq i \leq n$. For notation simplicity, we denote $\phi_{\mu_i} = \phi$ when $\mu_i = 0$. Note that it is not generally anticipated for clustering consistency under the usual Gaussian distribution assumption since the Gaussian distribution leads to overlapped clusters and violates the assumption (A0). Hence we modify the underlying distributions for the results in Theorem 2 by considering non-overlapping situations.

Consider a class of the truncated densities $\phi_{\mu_i,\alpha}, 1 \leq i \leq n$ with a truncation level $\alpha > 0$:

$$\phi_{\mu_i,\alpha}(z) = (1/c_\alpha)\phi_{\mu_i}(z) I\left(\|z - \mu_i\|_2^2 \leq \alpha/4\right), \tag{11}$$

where c_α is a normalizing constant. Note that$c_\alpha = \int_{A_{\mu_i,\alpha}} \phi_{\mu_i}(z)dz = \int_{A_\alpha} \phi(z)dz = \boldsymbol{\chi}_p(\alpha/4)$, where $A_{\mu_i,\alpha} = \{z : \|z - \mu_i\|_2^2 \leq \alpha/4\}$, $A_\alpha = \{z : \|z\|_2^2 \leq \alpha/4\}$ and $\boldsymbol{\chi}_p$ is the chi-square distribution function with p degrees of freedom. Given two mean vectors $\mu_i \neq \mu_j$, $\phi_{\mu_i,\alpha}$ does not overlap with $\phi_{\mu_j,\alpha}$ if $\|\mu_i - \mu_j\|_2^2 > \alpha$. Since the truncated densities $\phi_{\mu_i,\alpha}$ for $1 \leq i \leq n$ in (11) are not overlapped to each other if we take $\alpha = \alpha_{\min} = \min_{\mu_i \neq \mu_j} \|\mu_i - \mu_j\|_2^2$, we assume that the samples are independently distributed with true truncated densities $\phi_{\mu_i^*,\alpha_{\min}}; 1 \leq i \leq n$ with a truncation level $\alpha_{\min}$.

Now, ignoring constants, consider the problem in (7) for minimizing the minus log-likelihood $-\mathcal{L}(\mu) = \sum_{i=1}^n \|x_i - \mu_i\|_2^2/2$ under the constraint $\mathcal{J}(\mu) \leq J$. To derive a sufficient condition for (A2), we construct a lower bound of $\mathcal{C}_{\min}(\mu^*)$, the level of difficulty in recovering $\mathcal{C}(\mu^*)$. Asymptotic properties cannot be established when cluster $C_j \in \mathcal{C}(\mu)$ only shares a finite number of samples with true clusters, and thus we make the following assumption.

Assumption (A3): For any $\mu \in \mathbb{R}^{np}$, there exists m_1 such that $\inf_{C\in\mathcal{C}(\mu), C^*\in\mathcal{C}(\mu^*), C\cap C^*\neq\emptyset} |C \cap C^*| \geq m_1$.

Proposition 1 *Let* $r_{\alpha_{\min}} = \{\inf_{\mu\in\mathcal{B}} \inf_{\alpha_{\min} - \|\mu_i - \mu_i^*\|_2^2 \leq t \leq \alpha_{\min}} \chi_p(t/4)\}/\{4\boldsymbol{\chi}_p(\alpha_{\min}/4)\}$*, where* χ_p *and* $\boldsymbol{\chi}_p$ *are the chi-square density and distribution functions with* p *degrees of freedom, respectively. Under assumptions (A0), (A1) and (A3), if* $J = \mathcal{J}(\mu^*)$ *then the consistency results* (A) *and* (B) *in Theorem 2 hold, provided that*

$$r_{\alpha_{\min}}\alpha_{\min} \geq d_1 m K^* \log(n)/m_1, \tag{12}$$

for some constants $d_1 > \max\{1/c_2, 2d_0\log(c_3)/c_4^2\}$*, where* $K^* = |\mathcal{C}(\mu^*)|$.

Proposition 1 implies that (12) is a sufficient condition for (A2) for the truncated multivariate Gaussian distributions. In low dimensional situation, $\alpha_{\min}$, p and K^* may be fixed. $r_{\alpha_{\min}}$ is bounded below, which implies the clustering consistency follows when $m\log(n)/m_1 \to \infty$ as $n \to \infty$. Moreover, clustering consistency holds when $\alpha_{\min} \to 0$ and $p \to \infty$. From L'Hpital's rule, $\lim_{\alpha_{\min}\to 0} r_{\alpha_{\min}} \leq \lim_{\alpha_{\min}\to 0} \chi_p(\alpha_{\min}/4)/4\boldsymbol{\chi}_p(\alpha_{\min}/4) = \infty$ for any $p \geq 3$, which implies (12) is satisfied when $m_1\alpha_{\min}/mK^*\log(n) \to \infty$ as $n \to \infty$. For example, let $K^*\log(n) = n^k$, $m = n^h$ and $m_1 = n^{h_1}$ for some positive constants k, h and h_1, then the theorem holds provided that $\alpha_{\min} n^{h_1-(h+k)} \to \infty$ for any $k + h < h_1 < 1$,

implying that we can recover the true clusters even when $\alpha_{\min} \to 0$ and $K^* \to \infty$ as $n \to \infty$ for any $p \geq 3$.

At first sight, a truncated multivariate Gaussian distribution is an extreme example; however, after ignoring some constants the corresponding minus log-likelihood function $-\mathcal{L}(\mu) = \sum_{i=1}^{n} ||x_i - \mu_i||_2^2$, is used in the original PRclust. Moreover, truncated multivariate Gaussian distributions guarantee that different clusters are separated from each other; non-truncated Gaussian distributions lead to overlapping clusters, and consistency of distance-based clustering methods, including ours, is not expected as a result. For example, suppose we have n observations, $n/2$ form a Gaussian distribution $N(-0.5, 1)$, while the other $n/2$ from $N(0.5, 1)$. According to the K-means cluster center consistency theory (Pollard, 1981), the cluster centers determined by the K-means with $K = 2$ converge to $a_1 = -0.9$ and $a_2 = 0.9$, not the original clusters centers at $\mu_1 = -0.5$ and $\mu_2 = 0.5$. The reason is that all the negative observations from the second distribution/cluster are mis-clustered into the first cluster, while all positive observations from the first clusters are incorrectly assigned to the second cluster by the K-means, leading to an under-estimated center for the first cluster; similarly the over-estimation of the second cluster center can be explained. This simple example suggests that clustering consistency cannot be established when non-truncated Gaussian distributions are used in K-means. Furthermore, previous works focused on establishing clustering consistency with the distance between clusters growing at a sufficiently fast rate. For example, Zhu et al. (2014) showed that if the distance between two clusters and sample size n grow at the same rate as $n \to \infty$, then the corresponding method can separate the two clusters perfectly. In contrast, clustering consistency established here still holds when minimum distance between the cluster centroids $\alpha_{\min} \to 0$, implying that the assumptions used here are weaker than the previous ones.

4. Discussion

The proposed new algorithm DC-ADMM bears some similarity to the quadratic penalty based algorithm in terms of the cost function and using difference convex programming. However, they differ significantly in their specific formulations. Instead of using the quadratic penalty technique, we use a hard constraint and an augmented Lagrangian in DC-ADMM. Consequently, the DC-ADMM is much faster than the quadratic penalty based algorithm and can be relatively easy to be extended to other cost functions that may have some advantages for certain problems.

The theory that states some sufficient conditions for clustering consistency and optimal parameter estimation in the PRclust framework covers a much wide range of loss functions and grouping penalties, which helps us study theoretical results uniformly for some specific PRclust implementations in the future. For example, when graph information is available, by adding a constraint on the two connected nodes in the graph, we can estimate a cluster partition and grouping structure of variables simultaneously. The mis-clustering error bound and asymptotic properties of this graph-based PRclust can be obtained via a slight modification to the theory established here.

The methods can be extended in several directions. First, the convergence of the DC-ADMM algorithm is related to the penalty parameter ρ. A poor choice of ρ may result in a slow convergence for the ADMM algorithm (Ghadimi et al., 2015). One may use an

over-relaxed ADMM algorithm to speed up. Other options exist; for example, we may use different values of ρ in each iteration (Wang and Liao, 2001). Second, since the algorithm is relatively fast, it is now feasible to deal with high dimensional data, for which variable selection is necessary. In principle, we may add a new penalty into the cost function for variable selection (Pan and Shen, 2007). Third, we may modify PRclust for noisy big data. Others have developed an iterative sub-sampling approach to improve the computational efficiency of a solution path clustering and to handle noisy big data (Marchetti and Zhou, 2014). A modification of PRclust along this direction may be useful.

An R package *prclust* implementing the DC-ADMM algorithm and the quadratic penalty algorithm with various loss and penalty functions is available at GitHub (`https://github.com/ChongWu-Biostat/prclust`) and CRAN (`http://cran.r-project.org`).

Acknowledgments

The authors thank the reviewers for helpful comments. This research is partially supported by NIH grants R01GM113250, R01HL105397 and R01HL116720, and NSF grants DMS–1415500 and DMS–1207771. CW is supported by a University of Minnesota Doctoral Dissertation Fellowship.

Appendix A.

Proof of Theorem 1. The finite termination property of DC-ADMM follows from the following three facts. First, since (2) is closed, proper and convex and the augmented Lagrangian (4) has a saddle point, the standard ADMM converges (Boyd et al., 2011). Second, by construction of $S^{(m)}(\mu, \theta)$, for each $m \in N$,

$$\begin{aligned} 0 \leq S(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}) &= S^{(m+1)}(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}) \leq S^{(m)}(\hat{\mu}^{(m)}, \hat{\theta}^{(m)}) \\ &\leq S^{(m)}(\hat{\mu}^{(m-1)}, \hat{\theta}^{(m-1)}) = S(\hat{\mu}^{(m-1)}, \hat{\theta}^{(m-1)}), \end{aligned}$$

implying that $S(\hat{\mu}^{(m)}, \hat{\theta}^{(m)})$ decreases in m; otherwise the algorithm stops. Note that $(\hat{\mu}^{(m)}, \hat{\theta}^{(m)})$ is the limiting point of the ADMM iterations in (5). Third, since $S^{(m+1)}(\mu, \theta)$ depends on m only through that on the indicator functions $I(||\hat{\theta}_{ij}^{(m)}||_2 < \tau)$, which can be either 1 or 0, $S^{(m+1)}(\mu, \theta)$ has only a finite set of possible functional forms across all m, leading to a finite number of its possible and distinct minimal values. These facts imply that DC-ADMM terminates in a finite number of iterations.

To show that $(\hat{\mu}^{(m^*)}, \hat{\theta}^{(m^*)})$ is a KKT point of $S(\mu, \theta)$, we check if the solution satisfies a local optimality of $S(\mu, \theta)$. Since the subgradient of $S(\mu, \theta)$ and $S^m(\mu, \theta)$ are the same at the minimizer (Rockafellar, 2015), we verify the following requirement:

$$x_i + \rho \sum_{j>i} (\mu_j + \theta_{ij} + u_{ij}) + \rho \sum_{j<i} (\mu_j - \theta_{ji} - u_{ij}) - (1 + \rho(n-1))\mu_i = 0; \tag{13}$$

$$\lambda b_{ij} \theta_{ij} / ||\theta_{ij}||_2 + \rho(\theta_{ij} - (\mu_i - \mu_j) + u_{ij}) = 0; \tag{14}$$

$$\theta_{ij} - \mu_i - \mu_j = 0, \tag{15}$$

where b_{ij} is the regular subdifferential of $\min(||\theta_{ij}||_2, \tau)$ at $||\theta_{ij}||_2$. Easily, (15) is the hard constraint in the DC-ADMM and is met at convergence. Note that $(\hat{\mu}^{(m^*)}, \hat{\theta}^{(m^*)}, \hat{u}^{(m^*)}) = (\hat{\mu}^{(m^*-1)}, \hat{\theta}^{(m^*-1)}, \hat{u}^{(m^*-1)})$ at termination. Then (13) is satisfied with $(\mu, \theta, u) = (\hat{\mu}^{(m^*-1)}, \hat{\theta}^{(m^*-1)}, \hat{u}^{(m^*-1)})$. For (14), consider three cases.

- If $||\hat{\theta}_{ij}^{(m^*-1)}||_2 > \tau$, the $\hat{\theta}_{ij}^{(m^*-1)} = \hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - u_{ij}^{(m^*)}$, implying $\theta_{ij} = \hat{\theta}_{ij}^{(m^*)}$ since $b_{ij} = 0$.

- If $0 < ||\hat{\theta}_{ij}^{(m^*)}||_2 < \tau$ and $||\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}||_2 > \lambda/\rho$, then

$$\hat{\theta}_{ij}^{(m^*)} = \left(||\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}||_2 - \frac{\lambda}{\rho} \right) \frac{\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}}{||\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}||_2},$$

  hence that $||\hat{\theta}_{ij}^{(m^*)}||_2 = ||\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}||_2 - \frac{\lambda}{\rho}$. Then (14) is met when $\theta_{ij} = \hat{\theta}_{ij}^{(m^*)}$ since $b_{ij} = 1$.

- If $0 < ||\hat{\theta}_{ij}^{(m^*)}||_2 < \tau$ and $||\hat{\mu}_i^{(m^*)} - \hat{\mu}_j^{(m^*)} - \hat{u}_{ij}^{(m^*)}||_2 < \lambda/\rho$, then $||\hat{\theta}_{ij}^{(m^*)}||_2 = 0$, which contradicts to the fact that $0 < ||\hat{\theta}_{ij}^{(m^*)}||_2 < \tau$.

This completes the proof. ■

Lemma 1. *Given n observations $x_i \in \mathbb{R}^p$, $1 \le i \le n$, let the number of ways of constructing K clusters that satisfies disjoint condition (assumption A0) with the* minimal disjoint centering number *m be $c_{n,K,m}$. Then*

$$c_{n,K,m} \le (n - Km)^K \prod_{k=1}^{K} \binom{n-(k-1)m}{m}.$$

Proof of Lemma 1. Without loss of generality, we fix the first km points and form K disjoint subsets $S_k = \{x_{(k-1)m+1}, \ldots, x_{km}\} \subset \{x_1, \ldots, x_n\}$, $k \le K$. Let $r_i^{(k)} = d(\overline{x}_m^{(k)}, x_i)$, $km+1 \le i \le n$ with $\overline{x}_m^{(k)} = \sum_{j=(k-1)m+1}^{km} x_j/m$ and $\tilde{r}_i^{(k)}$ be an ordered sequence of $r_i^{(k)}$ that satisfies $\tilde{r}_{km+1}^{(k)} \le \cdots \le \tilde{r}_n^{(k)}$. Then a possible way of constructing a subset C_k based on S_k is including S_k and all the points within distance $\tilde{r}_i^{(k)}$. For a subset C_k, the number of constructing ways is $n - Km$ at most. Hence, the number of ways of constructing K subsets C_k, $k \le K$ based on S_k is $(n-Km)^K$ at most.

Note that the number of ways of fixing possible K disjoint subsets S_k, $k \le K$ is $\prod_{k=1}^{K} \binom{n-(k-1)m}{m}$. Hence the total number of ways of constructing K subsets along to the way described above is $(n-Km)^K \prod_{k=1}^{K} \binom{n-(k-1)m}{m}$ at most. Note that any cluster partition with K clusters that satisfies disjoint structure condition with the *minimal disjoint centering number* m can be constructed via the ways described above. Hence $c_{n,K,m} \le (n-Km)^K \prod_{k=1}^{K} \binom{n-(k-1)m}{m}$. This completes the proof. ■

Proof of Theorem 2. On the set $\tilde{\mathcal{B}} = \{\mu : \mathcal{C}(\mu) = \mathcal{C}(\mu^*)\} \subset \{\mu : \mathcal{J}(\mu) \le \mathcal{J}(\mu^*)\}$, we have $\hat{\mu}^{L_0} = \sup_{\mu \in \tilde{\mathcal{B}}} \mathcal{L}(\mu) = \hat{\mu}^o = \sup_{\mathcal{C}(\mu)=\mathcal{C}(\mu^*)} \mathcal{L}(\mu)$. Let the parameter space of interest be $\mathcal{B} = \{\mu : \mathcal{C}(\mu) \ne \mathcal{C}(\mu^*), \mathcal{J}(\mu) \le \mathcal{J}(\mu^*)\}$. Since $\mathcal{J}(\mu) \le \mathcal{J}(\mu^*)$ implies $|\mathcal{C}(\mu)| \le K^*$, we have $\mathcal{B} \subset \{\mu : \mathcal{C}(\mu) \ne \mathcal{C}(\mu^*), |\mathcal{C}(\mu)| \le K^*\} \subset \cup_{k=1}^{K^*} \cup_{C \in \mathcal{C}_k} \mathcal{B}_C$, where $\mathcal{B}_C = \{\mu : \mathcal{C}(\mu) = C\}$ and $\mathcal{C}_k = \{C \in \mathcal{C} : C \ne \mathcal{C}(\mu^*), |C| = k\}$. Hence, using $\mathcal{L}(\hat{\mu}^o) \ge \mathcal{L}(\mu^*)$, we have

$$\begin{aligned}
P\left(\hat{\mu}^{L_0} \ne \hat{\mu}^o\right) &\le P^*\left(\sup_{\mu \in \mathcal{B}} \{\mathcal{L}(\mu) - \mathcal{L}(\hat{\mu}^o)\} > 0\right) \\
&\le P^*\left(\sup_{\mu \in \mathcal{B}} \{\mathcal{L}(\mu) - \mathcal{L}(\mu^*)\} > 0\right) \\
&\le \sum_{k=1}^{K^*} \sum_{C \in \mathcal{C}_k} P^*\left(\sup_{\mu \in \mathcal{B}_C} \{\mathcal{L}(\mu) - \mathcal{L}(\mu^*)\} > 0\right),
\end{aligned}$$

where P^* is the outer probability. Now we apply Theorem 1 of Wong and Shen (1995) to bound each term. For any $\mu \in \mathcal{B}_C$ and $C \in \mathcal{C}_k$, $h_a^2(f_\mu, f_{\mu^*}) \ge k\mathcal{C}_{\min}(\mu^*)$, there exists a constant $c_2 > 0$ such that

$$P^*\left(\sup_{\mu \in \mathcal{B}_C} \{\mathcal{L}(\mu) - \mathcal{L}(\mu^*)\} > 0\right) \le 4\exp\left(-c_2 nk\mathcal{C}_{\min}(\mu^*)\right),$$

provided that the local entropy conditions are satisfied as follows: there exist constants $c_3 > 0$ and $c_4 > 0$ such that

$$\int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} H^{1/2}(t/c_3, \mathcal{F}_C)dt \leq c_4 n^{1/2}\epsilon^2 \tag{16}$$

for any $\epsilon^2 \geq k\mathcal{C}_{\min}(\mu^*)$. Let $\epsilon_n^2 = 2d_0 \log(c_3) m \log(n)/c_4^2 n$. Under (A1), ϵ_n solves the inequality

$$\max_{k \leq K^*} \sup_{C \in \mathcal{C}_k} \int_{\epsilon^2/2^8}^{2^{1/2}\epsilon} H^{1/2}(t/c_3, \mathcal{F}_C)dt \leq (d_0 m \log(n))^{1/2}(2^{1/2}\epsilon)(\log(c_3))^{1/2} \leq c_4 n^{1/2}\epsilon^2$$

with respect to ϵ provided that $\epsilon_n < \epsilon$. Hence, (16) follows if $\mathcal{C}_{\min}(\mu^*) \geq \epsilon_n^2$, and from (A2), this holds when $d_1 \geq 2d_0 \log(c_3) m/c_4^2$. From Lemma 1, $|\mathcal{C}_k| \leq (n-km)^k \prod_{j=1}^k \binom{n-(j-1)m}{m} \leq n^{k+mk}$. Hence,

$$\begin{aligned} P\left(\hat{\mu}^{L_0} \neq \hat{\mu}^o\right) &\leq \sum_{k=1}^{K^*} 4\exp\left(-c_2 nk\mathcal{C}_{\min}(\mu^*) + k(m+1)\log(n)\right) \\ &\leq 4R\left(\exp(-c_2 n\mathcal{C}_{\min}(\mu^*) + (m+1)\log(n)\right) \\ &\leq 5\exp\left(-c_2 n\mathcal{C}_{\min}(\mu^*) + (m+1)\log(n)\right) \\ &\leq \exp\left(-c_2 n\mathcal{C}_{\min}(\mu^*) + (m+1)\log(n) + 2\right), \end{aligned}$$

where $R(x) = x/(1-x)$ is exponentiated logistic function.

Now (A) follows from $P\left(\mathcal{C}(\hat{\mu}^{L_0} \neq \mathcal{C}(\mu^*)\right) \leq P\left(\hat{\mu}^{L_0} \neq \mu^*\right)$ and $d_1 > 1/c_2$. For the risk property, using $h_a^2(\hat{\mu}^{L_0}, \mu^*) \leq 1$,

$$\begin{aligned} E\left[h_a^2(\hat{\mu}^{L_0}, \mu^*)\right] &\leq E\left[h_a^2(\hat{\mu}^o, \mu^*)\right] + E\left[h_a^2(\hat{\mu}^{L_0}, \mu^*) I(\hat{\mu}^{L_0} \neq \hat{\mu}^o)\right] \\ &\leq E\left[h_a^2(\hat{\mu}^o, \mu^*)\right] + P(\hat{\mu}^{L_0} \neq \hat{\mu}^o) \\ &\leq (1 + o(1))E\left[h_a^2(\hat{\mu}^o, \mu^*)\right] \end{aligned}$$

provided that $\exp\left(-c_2 n\mathcal{C}_{\min}(\mu^*)\right)/Eh_a^2(\hat{\mu}^o, \mu^*) = o(1)$, and then (B) established. This completes the proof. ■

Proof of Proposition 1. It suffices to show that (12) is a sufficient condition for Assumption (A2). First, we give a lower bound of the Hellinger metric between $\phi_{\mu_i, \alpha_{\min}}$ and $\phi_{\mu_i^*, \alpha_{\min}}$ for a given $\mu \in \mathcal{B} = \{\mu : \mathcal{C}(\mu) \neq \mathcal{C}(\mu^*), \mathcal{J}(\mu) \leq \mathcal{J}(\mu^*)\}$. Let

$$A_{\alpha_{\min}} = \{z : \|z\|_2^2 \leq \alpha_{\min}/4\} \text{ and } A_{\mu_i, \alpha_{\min}} = \{z : \|z - \mu_i\|_2^2 < \alpha_{\min}/4\}, 1 \leq i \leq n.$$

Given $\mu_i \neq \mu_i^*$ with $\|\mu_i - \mu_i^*\|_2^2 \leq \alpha_{\min}$, let $\Delta_i^* = \mu_i - \mu_i^*$ then we have

$$\begin{aligned}
&h^2(\phi_{\mu_i,\alpha_{\min}}, \phi_{\mu_i^*,\alpha_{\min}})^2 \\
=&h^2(\phi_{\alpha_{\min}}, \phi_{\Delta^*,\alpha_{\min}})^2 \\
=&\frac{1}{2}\int \left(\phi_{\alpha_{\min}}^{1/2}(z) - \phi_{\Delta_i^*,\alpha_{\min}}^{1/2}(z)\right)^2 dz \\
=&\frac{1}{2\boldsymbol{\chi}_p^2(\alpha_{\min}/4)}\left(\int_{A_{\alpha_{\min}}} \phi(z)dz + \int_{A_{\Delta_i^*,\alpha_{\min}}} \phi_{\Delta_i^*}(z)dz - 2\int_{A_{\alpha_{\min}}\cap A_{\Delta_i^*,\alpha_{\min}}} \phi(z)^{1/2}\phi_{\Delta^*}(z)^{1/2}dz\right) \\
=&\frac{1}{\boldsymbol{\chi}_p^2(\alpha_{\min}/4)}\left(\int_{A_{\alpha_{\min}}} \phi(z)dz - \int_{A_{\alpha_{\min}}\cap A_{\Delta_i^*,\alpha_{\min}}} \phi(z)^{1/2}\phi_{\Delta^*}(z)^{1/2}dz\right).
\end{aligned}$$

Let $B_{\Delta_i^*,\alpha_{\min}} = \{z : \|z - \Delta_i^*\|_2^2 \leq \alpha/4 - \|\Delta_i^*\|_2^2/4\}$ then it is easy to see that

$$A_{\alpha_{\min}} \cap A_{\Delta_i^*,\alpha_{\min}} \subset B_{\Delta_i^*,\alpha_{\min}}.$$

By using the equality,

$$\phi(z)^{1/2}\phi_{\Delta_i^*}(z)^{1/2} = (2\pi)^{-p/2}\exp(-\|z - \Delta_i^*/2\|_2^2/2 - \|\Delta^*\|_2^2/8) = \exp(-\|\Delta_i^*\|_2^2/8)\phi_{\Delta_i^*/2}(z),$$

we have

$$\begin{aligned}
\int_{A_{\alpha_{\min}}\cap A_{\Delta_i^*,\alpha_{\min}}} \phi(z)^{1/2}\phi_{\Delta^*}(z)^{1/2}dz \leq& \exp(-\|\Delta_i^*\|_2^2/8)\int_{B_{\Delta_i^*,\alpha_{\min}}} \phi_{\Delta_i^*/2}(z)dz \\
=&\exp(-\|\Delta_i^*\|_2^2/8)\boldsymbol{\chi}_p(\alpha_{\min}/4 - \|\Delta_i^*\|_2^2/4) \\
\leq&\boldsymbol{\chi}_p(\alpha_{\min}/4 - \|\Delta_i^*\|_2^2/4).
\end{aligned}$$

According to the mean value theorem,

$$h^2(\phi_{\mu_i,\alpha_{\min}}, \phi_{\mu_i^*,\alpha_{\min}})^2 \geq \frac{\boldsymbol{\chi}_p(\alpha_{\min}/4) - \boldsymbol{\chi}_p(\alpha_{\min}/4 - \|\Delta_i^*\|_2^2/4)}{\boldsymbol{\chi}_p^2(\alpha_{\min}/4)} \geq r_{\alpha_{\min}}\|\mu_i - \mu_i^*\|_2^2, \quad (17)$$

where $r_{\alpha_{\min}} = \{\inf_{\mu\in\mathcal{B}}\inf_{\alpha_{\min}-\|\Delta_i^*\|_2^2\leq t\leq\alpha_{\min}} \chi_p(t/4)\}/4\boldsymbol{\chi}_p(\alpha_{\min}/4)$. Next, we find a lower bound of $C_{\min}(\mu^*)$. From (17), the inequality $h_a^2(f_\mu, f_{\mu^*}) \geq \sum_{i=1}^n h^2(f_{\mu_i}, f_{\mu_i^*})/n$ implies

$$nC_{\min}(\mu^*) = n\inf_{\mu\in\mathcal{B}} h_a^2(f_\mu, f_{\mu^*})/|\mathcal{C}(\mu)| \geq nr_{\alpha^*}\inf_{\mu\in\mathcal{B}}\|\mu - \mu^*\|_2^2/|\mathcal{C}(\mu)|. \quad (18)$$

It is easy to see that $\inf_{\mu\in\{\mu:|\mathcal{C}(\mu)|<K^*\}}\|\mu - \mu^*\|_2^2/|\mathcal{C}(\mu)| \geq \inf_{\mu\in\{\mu:|\mathcal{C}(\mu)|=K^*\}}\|\mu - \mu^*\|_2^2/K^*$, since the sum of within cluster variances, $\|\mu - \mu^*\|_2^2 = \sum_{i=1}^n \|\mu_i - \mu_i^*\|_2^2$, is minimized when $|\mathcal{C}(\mu)| = K^*$. Hence, we have

$$\inf_{\mu\in\mathcal{B}}\|\mu - \mu^*\|_2^2/|\mathcal{C}(\mu)| = \inf_{\mu\in\{\mu:|\mathcal{C}(\mu)|=K^*,\mathcal{C}(\mu)\neq\mathcal{C}(\mu^*)\}}\|\mu - \mu^*\|_2^2/K^*. \quad (19)$$

Let $\mathcal{C}(\mu) = \{C_1, \ldots, C_K\}$ and $\mathcal{C}(\mu^*) = \{C_1^*, \ldots, C_{K^*}^*\}$. Since $\mathcal{C}(\mu) \neq \mathcal{C}(\mu^*)$, without loss of generality, we may assume that $C_s \cap C_t^* \neq \emptyset$ for $s, t = 1, 2$. Then the right-hand side of (19) achieves its minimum when $\mu \in \mathcal{B}_{12} = \{\mu : C_s \cap C_t^* \neq \emptyset$ for $s, t = 1, 2$ and $\mu_i = \mu_i^*$ for $i \in$

$\cup_{3\le k\le K^*} C_k\}$. Let $\mu_i = \nu_s, i \in C_s$ for $s = 1, 2$ and similarly let $\mu_i^* = \nu_t^*, i \in C_t^*$ for $t = 1, 2$. Then it follows that

$$\begin{aligned}\inf_{\mu\in\mathcal{B}} \|\mu - \mu^*\|_2^2 &= \inf_{\mu\in\mathcal{B}_{12}} \sum_{t=1,2} \left(n_{1t}\|\nu_1 - \nu_t^*\|_2^2 + n_{2t}\|\nu_2 - \nu_t^*\|_2^2\right) \\ &= \inf_{n_{st},s,t=1,2} \sum_{t=1,2} \left(n_{1t}\|\bar{\nu}_1^* - \nu_t^*\|_2^2 + n_{2t}\|\bar{\nu}_2^* - \nu_t^*\|_2^2\right) \\ &= \inf_{n_{st},s,t=1,2} \left(\frac{n_{11}n_{12}}{n_{11}+n_{12}} + \frac{n_{21}n_{22}}{n_{21}+n_{22}}\right) \|\nu_1^* - \nu_2^*\|_2^2,\end{aligned}$$

where $n_{st} = |C_s \cap C_t^*|$ for $s, t = 1, 2$, and $\bar{\nu}_1^* = (n_{11}\nu_1^* + n_{12}\nu_2^*)/(n_{11}+n_{12})$ and $\bar{\nu}_2^* = (n_{21}\nu_1^* + n_{22}\nu_2^*)/(n_{21}+n_{22})$ are the weighted means of ν_1^*s and ν_2^*s in C_1 and C_2, respectively. From (A3), $n_{st} \ge m_1$ for $s, t = 1, 2$, which implies $n_{11}n_{12}/(n_{11}+n_{12}) = 1/(1/n_{11}+1/n_{12}) \ge m_1/2$ and similarly, $n_{21}n_{22}/(n_{21}+n_{22}) \ge m_1/2$. Hence the lower bound becomes

$$\inf_{\mu\in\mathcal{B}} \|\mu - \mu^*\|_2^2 \ge m_1\alpha_{\min}. \tag{20}$$

From (18), (19), (20) and definition of $C_{\min}(\mu^*)$, it is easy to see that (A2) is met if $C_{\min}(\mu^*) \ge r_{\alpha_{\min}} m_1\alpha_{\min}/nK^* \ge d_1 m \log(n)/n$ which is equivalent to

$$r_{\alpha_{\min}}\alpha_{\min} \ge d_1 m K^* \log(n)/m_1.$$

This completes the proof. ■

Appendix B.

The cost function of PRclust with lasso grouping penalty will be convex, and thus DC-ADMM is exactly same as ADMM and a global solution will be reached. Note that $\theta_{ij} = (\theta_{ij1}, \ldots, \theta_{ijp})'$, then the updating formulas can be summarized as follows:

$$\begin{aligned}\hat{\mu}_i^{(m+1)} &= \frac{x_i + \rho\sum_{j>i}\left(\hat{\mu}_j^{(m)} + \hat{\theta}_{ij}^{(m)} + \hat{u}_{ij}^{(m)}\right) + \rho\sum_{j<i}\left(\hat{\mu}_j^{(m+1)} - \hat{\theta}_{ji}^{(m)} - \hat{u}_{ij}^{(m)}\right)}{1+\rho(n-1)}; \\ \hat{\theta}_{ijl}^{(m+1)} &= \mathrm{ST}\left(\hat{\mu}_{il}^{(m+1)} - \hat{\mu}_{jl}^{(m+1)} - \hat{u}_{ijl}^{(m)}; \lambda/\rho\right) \\ \hat{u}_{ij}^{(m+1)} &= \hat{u}_{ij}^{(m)} + \hat{\theta}_{ij}^{(m+1)} - (\hat{\mu}_i^{(m+1)} - \hat{\mu}_j^{(m+1)}), \quad 1 \le i < j \le n; i, l = 1, 2, \ldots, p.\end{aligned}$$

The main difference between TLP and gTLP is that TLP is an element-wise penalty and the updating formulas (5) for PRclust with TLP can be summarized as follows, while the other part of DC-ADMM remains unchanged:

$$\begin{aligned}\hat{\mu}_i^{k+1} &= \frac{x_i + \rho\sum_{j>i}\left(\hat{\mu}_j^{k} + \hat{\theta}_{ij}^{k} + \hat{u}_{ij}^{k}\right) + \rho\sum_{j<i}\left(\hat{\mu}_j^{k+1} - \hat{\theta}_{ji}^{k} - \hat{u}_{ij}^{m}\right)}{1+\rho(n-1)}; \\ \hat{\theta}_{ijk}^{k+1} &= \begin{cases} \hat{\mu}_{il}^{k+1} - \hat{\mu}_{jl}^{k+1} - \hat{u}_{ijl}^{k} & \text{if} \quad |\hat{\theta}_{ijl}^{(m)}| \ge \tau; \\ \mathrm{STL}\left(\hat{\mu}_{il}^{k+1} - \hat{\mu}_{jl}^{k+1} - \hat{u}_{ijl}^{k}; \lambda/\rho\right) & \text{if} \quad |\hat{\theta}_{ijl}^{(m)}| < \tau; \end{cases} \\ \hat{u}_{ij}^{k+1} &= \hat{u}_{ij}^{k} + \hat{\theta}_{ij}^{k+1} - (\hat{\mu}_i^{k+1} - \hat{\mu}_j^{k+1}), \quad 1 \le i < j \le n; i, l = 1, 2, \ldots, p.\end{aligned}$$

References

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Leo Breiman and Adele Cutler. A deterministic algorithm for global optimization. *Mathematical Programming*, 58(1-3):179–199, 1993.

Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

Euhanna Ghadimi, André Teixeira, Iman Shames, and Mikael Johansson. Optimal parameter selection for the alternating direction method of multipliers (admm): quadratic problems. *IEEE Transactions on Automatic Control*, 60(3):644–658, 2015.

Bingsheng He and Xiaoming Yuan. On non-ergodic convergence rate of douglas–rachford alternating direction method of multipliers. *Numerische Mathematik*, 130(3):567–577, 2015.

Toby Dylan Hocking, Armand Joulin, Francis Bach, and Jean-Philippe Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th International Conference on Machine Learning (ICML)*, 2011.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, 1985.

Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31 (8):651–666, 2010.

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

An Le Thi Hoai and Pham Dinh Tao. Solving a class of linearly constrained indefinite quadratic problems by dc algorithms. *Journal of Global Optimization*, 11(3):253–285, 1997.

Fredrik Lindsten, Henrik Ohlsson, and Lennart Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *Statistical Signal Processing Workshop*, 2011.

Binghui Liu, Xiaotong Shen, and Wei Pan. Integrative and regularized principal component analysis of multiple sources of data. *Statistics in Medicine*, 35(13):2235–50, 2016.

Yuliya Marchetti and Qing Zhou. Iterative subsampling in solution path clustering of noisy big data. *arXiv preprint arXiv:1412.1559*, 2014.

Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research*, 8(1):1145–1164, 2007.

Wei Pan, Xiaotong Shen, and Binghui Liu. Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *The Journal of Machine Learning Research*, 14(1):1865–1889, 2013.

Kristiaan Pelckmans, Joseph De Brabanter, JAK Suykens, and B De Moor. Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*, 2005.

David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1): 135–140, 1981.

Peter Radchenko and Gourab Mukherjee. Consistent clustering using an l_1 fusion penalty. *arXiv preprint arXiv:1412.0753*, 2014.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton university press, 2015.

Xiaotong Shen, Wei Pan, and Yunzhang Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107(497):223–232, 2012.

Douglas Steinley. K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.

Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

S.L. Wang and L.Z. Liao. Decomposition method with a variable parameter for a class of monotone variational inequality problems. *Journal of Optimization Theory and Applications*, 109(2):415–429, 2001.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362, 1995.

Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

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

Changbo Zhu, Huan Xu, Chenlei Leng, and Shuicheng Yan. Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems*, 2014.