Robust Convex Clustering Analysis

Qi Wang[1], Pinghua Gong[2], Shiyu Chang[3], Thomas S. Huang[3], Jiayu Zhou[1]
[1]Computer Science and Engineering, Michigan State University, East Lansing, MI 48824.
[2]University of Michigan, Ann Arbor, MI 48109.
[3]Beckman Institute, University of Illinois at Urbana-Champaign, IL 61801.

***Abstract*—Clustering is an unsupervised learning approach that explores data and seeks groups of similar objects. Many classical clustering models such as k-means and DBSCAN are based on heuristics algorithms and suffer from local optimal solutions and numerical instability. Recently convex clustering has received increasing attentions, which leverages the sparsity inducing norms and enjoys many attractive theoretical properties. However, convex clustering is based on Euclidean distance and is thus not robust against outlier features. Since the outlier features are very common especially when dimensionality is high, the vulnerability has greatly limited the applicability of convex clustering to analyze many real-world datasets. In this paper, we address the challenge by proposing a novel robust convex clustering method that simultaneously performs convex clustering and identifies outlier features. Specifically, the proposed method learns to decompose the data matrix into a clustering structure component and a group sparse component that captures feature outliers. We develop a block coordinate descent algorithm which iteratively performs convex clustering after outliers features are identified and eliminated. We also propose an efficient algorithm for solving the convex clustering by exploiting the structures on its dual problem. Moreover, to further illustrate the statistical stability, we present the theoretical performance bound of the proposed clustering method. Empirical studies on synthetic data and real-world data demonstrate that the proposed robust convex clustering can detect feature outliers as well as improve cluster quality.**

I. Introduction

Clustering is one of the most common unsupervised learning approaches. For a given set of objects, clustering aims at finding groups of objects such that similar objects are grouped together. The unique advantage of unsupervised learning approaches is that they are not limited by the number of labeled data. In the big data era, we are submerged by an avalanche of data and information, and it is almost impossible to label all the obtained data. Moreover, the lack of labels is especially prominent in the domains where data labeling is expensive. Therefore clustering remains as a popular data exploration technique and is nowadays widely used in many fields such as medical imaging [6] and social science [2].

Many clustering algorithms are developed based on different grouping heuristics and clustering mechanisms. k-means clustering [14], as one of the most popular technique, uses Euclidean distance as a metric to partition data points into k clusters. K-means algorithm generates even sizes of clusters of spherical cluster shape, and density based clustering methods such as DBSCAN [7] can overcome the difficulties by clustering the data points based on data density. Deciding the number of clusters is typically hard without proper prior knowledge, and thus hierarchical clustering [11] is proposed to build a hierarchy on top of clusters. There are some common drawbacks in existing clustering algorithms that usually limit their practical performance. First, many algorithms are based on heuristics and are in fact solving a local solution to a non-convex optimization problems. For example, it can be shown that k-means clustering is equivalent to optimizing a non-convex objective using a descending algorithm [20]. Such algorithms are thus easily trapped by a bad local optima, especially in high-dimensional problems. In practice k-means would converge to different clusters when using different initializations. Another common challenge comes from the hyper-parameter deciding in terms of the number of clusters. Related ones include the cluster number k in k-means algorithm and neighborhood size or density parameter in the density based algorithms. It is worth mentioning that even hierarchical clustering can be used when cluster number is unknown, it is also greedy by nature; and it may produce spurious clusters [3].

The challenges in the traditional clustering algorithms have motivated the recent development of *convex clustering* techniques [5], [10], [12], [15], which leverages the group sparsity induced by a fused regularizer, preserving the convexity of the problem. In convex clustering, the objective is to jointly search cluster centers and assign them to each data point. Because of the convexity, the global optimum can be achieved by solving a traceable convex programming regardless of the initializations, the statistical performance of the objective can be comprehensively analyzed [17], and the geometrical properties associated to the objective can be studied and exploited [9]. Moreover, by varying the parameter of the regularization, convex clustering can generate a sequence of clustering results, which provide hierarchical clustering and is commonly referred as *clusterpath* [10]. Therefore convex clustering does not need cluster number *a priori* or careful initializations. However, because that convex clustering uses the Frobenius norm to evaluate clusters and measure the distance between data points and cluster centers, this method is vulnerable to outlier features. The clustering process will be distorted by outlier features, leading to suboptimal results.

In this paper, we propose a novel robust convex clustering (RCC) to address the challenges from outlier features, by explicitly modeling the outlier features in the clustering objective and identifying those features to improve the cluster quality. Specifically, the proposed RCC learns to decompose data into two components: the *clustering component* captures the clustering structure and assigns each data point to a cluster center, whereas the *robust component* identifies the feature outliers in the data. The learning of the clustering component uses fused sparsity norms as did in the convex clustering, and the learning of the robust component leverages grouped sparsity [19]. We provide an efficient algorithm for solving RCC, which casts light on its mechanism: RCC is achieved via an iterative process that identifies features incoherent with the current clustering structure and performs convex clustering on the "purified data" after removing the effects from the outlier features. Since the algorithm involves multiple runs of convex clustering, we develop an efficient algorithm for convex clustering by exploiting the structure of its dual formulation.

The remaining of this paper is organized as follows: Section II discusses related work to convex clustering. Section III presents the formulation and algorithm of robust convex clustering. Section IV shows the theoretical performance bound. Section V provides experimental results on synthetic and real datasets. Section VI concludes the paper.

II. Related Work

A. Convex clustering

Convex clustering method has been introduced and studied from different perspectives in recent years. In [15], Pelckmans and De Moor viewed the clustering problem as a convex optimization problem. They proposed a shrinkage term resulting sparseness amongst the differences between the centroids. They showed that varying the trade-off shrinkage term yields a hierarchical clustering tree. In [10], Hocking *et. al.* presented a new algorithm for solving the clusterpath of the convex clustering to facilitate hierarchical clustering to efficiently compute the continuous regularization path for convex clustering. To solve the optimization problem of convex clustering in [5], the authors introduced two splitting methods to solve the convex clustering and applied them to solve various real-world problems. The one is based on the alternating direction method of multipliers (ADMM) and the other is an instance of the alternating minimization algorithm (AMA). The authors also accelerated both method and showed that AMA could be significantly more efficient based on both the complexity analysis and numerical experiments.

B. Robustness and feature selection

Robustness of learning models are critical when there are outliers in the data. The sparsity-inducing norms have been demonstrated to be effectiveness in detecting outliers when properly incorporated in the learning formulation. [8] proposed the robust multi-task feature learning which simultaneously captures a common set of features among relevant tasks and identifies outlier tasks. They decomposed the weight matrix into two components. The first component is used for capturing the shared features among relevant tasks, and the second component is used for identify the outlier tasks. In [4] the authors proposed a robust multi-task learning that integrate low-rank and group-sparse structures. The proposed algorithm captures the relationship of multiple related tasks using a low-rank structure and meanwhile identifies the outlier tasks using a group-sparse structure. The robustness has not been previous studied in the context of convex clustering. Motivated by robust analysis studies above, in this paper we propose a novel robust convex clustering formulation using the sparsity-inducing norms.

III. Robust Convex Clustering

A. Robustness in convex clustering

Assume that we are given n data points and each data point is described by a d dimensional feature vector. We collectively represent the data by a data matrix $X \in \mathbb{R}^{d\times n}$. The convex clustering method clusters data points into groups via solving a convex optimization problem:

$$\min_{P} \tfrac{1}{2}\|X-P\|_F^2 + \alpha \sum\nolimits_{i<j} w_{i,j}\|P_i - P_j\|_p, \qquad (1)$$

where $\|\cdot\|_p$ is the ℓ_p-norm[1], $P \in \mathbb{R}^{d\times n}$ is the matrix consisting cluster centers and assignments, P_i is the i^{th} column of the matrix P and represents the centroid of cluster that the i^{th} data point is assigned, α is a non-negative regularization parameter that controls the number of clusters. $w_{i,j}$ is the weight between i^{th} and j^{th} data point specified by the user. Clusters are implicitly given by P matrix and $P_i = P_j$ means the two data points are in the same cluster. When α is zero, each data point forms a cluster since P equals to X when Eq. (1) reaches the optimum. As α increases, $\|P_i - P_j\|_p$ becomes smaller for all pairs so that the whole function can reach the minimal value. Hence more data points will be grouped into a cluster.

Both the loss function $\|X-P\|_F^2$ and the regularization terms $\sum_{i<j} w_{i,j}\|P_i - P_j\|_p$ treat features equally, and therefore when outlier features present, the convex clustering algorithm may be misled to distorted clusters. In this paper we propose to explicitly model the outliers and identify them during the clustering. We assume that the data matrix can be decomposed into two parts: $X = P + Q$, where the *clustering component* P captures the clustering structure as did in the traditional convex clustering, whereas the *robust component* $Q \in \mathbb{R}^{d\times n}$ is a row sparse matrix that identifies the feature outliers in the data. Therefore the cluster structure $P = X - Q$ is defined on a "purified" data $X - Q$, i.e., after effects from outlier features are removed. Hence when we perform convex clustering, the cluster will not be distorted by the outliers. The proposed robust convex clustering (RCC) can be formulated by the following convex optimization problem:

$$\begin{aligned}\hat{P},\hat{Q} = \min\nolimits_{P,Q} &\tfrac{1}{2}\|X-(P+Q)\|_F^2 \\ &+ \alpha \sum\nolimits_{i<j} w_{i,j}\|P_i - P_j\|_p + \beta\|Q\|_{2,1}, \quad (2)\end{aligned}$$

where β is the regularization parameter indiating the level of noise in the data. Here we use group lasso penalty to regularize Q and achieve the desired row-wise sparsity. Recall that group

[1] The convex clustering requires $p \geq 1$, and in this paper we focus on the convex clustering so we consider only $p \geq 1$. However, we note that the algorithm can be extended to the non-convex case where $0 < p < 1$.

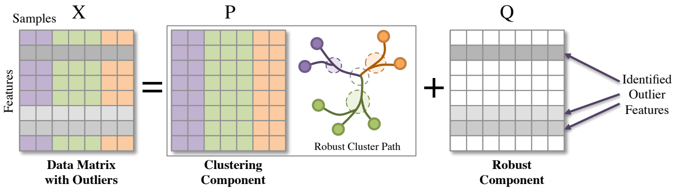


Fig. 1. The illustration of the proposed convex robust clustering analysis (RCC). The proposed method learns a *clustering component* P that captures the clustering structure and assigns each data point to a cluster center, and a *robust component* Q that identifies the feature outliers in the data.

lasso yield group sparsity within group. If one element in a row in Q is not zero, then all the elements in this row are likely to be non-zero. If one element in a row is 0, then the whole row elements are zero. Hence if a feature is useful, the corresponding row in Q will be zero. And if a feature is outlier, this row will be non-zero for all elements. The framework is illustrated in Figure 1.

B. Optimization

In this subsection, we show how to solve the robust convex clustering formulation in Eq. (2). We iteratively fix one of the P, Q blocks and optimize with respect to the other, until ceratin convergence criteria is achieved. The entire objective as well as the two subproblems are convex and thus this iterative algorithm will converge to one optimal solution [18]. When it converges, outlier features will be indicated by non-zero rows in the matrix Q. We use P^+ and Q^+ to denote the updated variables in the current iteration, P^- and Q^- to denote the variables in the previous iteration.

Solve P and fix Q. The optimization problem becomes:

$$P^+ = \underset{P}{\operatorname{argmin}} \tfrac{1}{2}\|(X - Q^-) - P\|_F^2 + \alpha \sum_{i<j} w_{i,j}\|P_i - P_j\|_2. \tag{3}$$

We see that in this step we perform a standard convex clustering on the modified data $X - Q^-$ by substracting the original data matrix with the outlier features and the effects Q^- identified in the last iteration. As discussed above, if a feature is useful feature, the corresponding row in Q is zero, whereas for outlier features, the corresponding rows in Q are non-zero. By subtracting Q^-, we remove the influence of outlier features on the current convex clustering.

Solve Q and fix P. The optimization problem becomes:

$$Q^+ = \operatorname{argmin}_Q \tfrac{1}{2}\|(X - P^+) - Q\|_F^2 + \beta\|Q\|_{2,1}, \tag{4}$$

where the optimization problem has the form of $\ell_{2,1}$ proximal projection and admits a closed form solution [13] $Q^+ = \mathcal{P}(X - P^+)$ whose i^{th} row is given by:

$$\max\left(0, 1 - \frac{\beta}{\sqrt{\sum_j (X-P)_{i,j}}}\right) \cdot \left(X - P^+\right)_i .$$

In Eq. (4) the algorithm seeks a sparse solution that captures the difference between the original data matrix X and the updated clustering structure P. When a feature is an outlier, then it will not conform the cluster structure, and thus the magnitude at the corresponding rows in $X - P^+$ is likely to be higher, and identified by the Q matrix via non-sparse rows.

C. An efficient solver for convex clustering

Since the update of the robust component Q has a closed form solution and the majority of computational cost of the proposed RCC comes from multiple calls of convex clustering to update the clustering component P. This motivates us to develop a more efficient algorithm to solve the convex clustering. There are attempts to solve the convex clustering in prior work, and however, directing solving the primal formulation is challenging because of the fused terms. The Lagrange dual [1] can shed light on some hidden geometric structures of many hard problems, which can be used to build efficient algorithms. In this paper we show that one of the dual formulations of the convex clustering leads to an efficient solver. In this section, we use U to denote $(X - Q)$ and the update of P in Eq. (3) is given by:

$$\min_P \tfrac{1}{2}\|P - U\|_F^2 + \alpha \sum\nolimits_{i<j} w_{i,j}\|P_i - P_j\|. \tag{5}$$

In order to obtain a dual formulation, we introduce a set of constraints and transform it into a constrained problem. Let $w_{i,j}(P_i - P_j) = \mathbf{z}_{ij}$. Then Eq. (5) is equivalent to:

$$\begin{aligned} &\min_{P,\{\mathbf{z}_{ij}\}} \tfrac{1}{2}\|P - U\|_F^2 + \alpha \sum\nolimits_{i<j} \|\mathbf{z}_{ij}\|, \\ &\text{s.t. } w_{i,j}(P_i - P_j) = \mathbf{z}_{ij}, \forall i < j \end{aligned} \tag{6}$$

which leads to the Lagrange function of Eq. (6) as follows:

$$\begin{aligned} \mathcal{L}(P, \{\mathbf{z}_{ij}\}, \{\boldsymbol{\theta}_{ij}\}) = &\tfrac{1}{2}\|P - U\|_F^2 + \alpha \sum\nolimits_{i<j} \|\mathbf{z}_{ij}, \| \\ &+ \sum\nolimits_{i<j} \boldsymbol{\theta}_{ij}^T (w_{i,j}(P_i - P_j) - \mathbf{z}_{ij}), \end{aligned}$$

where $\boldsymbol{\theta}$ is the Lagrange multiplier, $\boldsymbol{\theta}_{i,j}$ is a column of matrix $\boldsymbol{\theta}$, whose subindex i, j denote the column corresponding to the column vector $(w_{i,j}(P_i - P_j) - \mathbf{z}_{ij})$.

By minimizing $\mathcal{L}(P, \{\mathbf{z}_{ij}\}, \{\boldsymbol{\theta}_{ij}\})$ with respect to P and $\mathbf{z}_{ij}$, we obtain the objective function of the dual problem of Eq. (6) as follows:

$$\begin{aligned} \tilde{\mathcal{D}}(\boldsymbol{\theta}_{ij}) &= \min\nolimits_{P,\mathbf{z}_{ij}} \mathcal{L}(P, \mathbf{z}_{ij}, \boldsymbol{\theta}_{ij}) \\ &= \min_P \left\{ \tfrac{1}{2}\|P - U\|_F^2 + \sum\nolimits_{i<j} \boldsymbol{\theta}_{ij}^T w_{i,j}(P_i - P_j) \right\} \\ &\quad + \sum\nolimits_{i<j} \min\nolimits_{\mathbf{z}_{ij}} (\alpha\|\mathbf{z}_{ij}\| - \boldsymbol{\theta}_{ij}^T \mathbf{z}_{ij}), \end{aligned}$$

where we have rearranged all terms in the Lagrange function.

According to the property of the dual norm [1], we have the following:

$$\min_{\mathbf{z}_{ij}} \left\{ \alpha\|\mathbf{z}_{ij}\| - \boldsymbol{\theta}_{ij}^T \mathbf{z}_{ij} \right\} = \begin{cases} 0, & \text{if } \|\boldsymbol{\theta}_{ij}\| \le \alpha, \\ -\infty, & \text{otherwise.} \end{cases}$$

which implies that the feasible region is $\|\boldsymbol{\theta}_{ij}\| \le \alpha$. Hence we obtain:

$$\begin{aligned} \mathcal{D}(\boldsymbol{\theta}_{ij}) &= \min\nolimits_P \left\{ \tfrac{1}{2}\|P - U\|_F^2 + \sum\nolimits_{i<j} \boldsymbol{\theta}_{ij}^T w_{i,j}(P_i - P_j) \right\}. \\ &\text{s.t. } \|\boldsymbol{\theta}_{ij}\| \le \alpha \end{aligned} \tag{7}$$

Let $\mathbf{e}_i \in \mathbb{R}^n$ be the n-dimensional standard basis with only the i-th entry as 1. Then we have:

$$P_i = P\mathbf{e}_i.$$

Substituting the above equality into Eq. (7), we obtain:

$$\mathcal{D}(\boldsymbol{\theta}_{ij}) = \min\nolimits_P \left\{ \tfrac{1}{2}\|P - U\|_F^2 + \sum\nolimits_{i<j} \boldsymbol{\theta}_{ij}^T w_{i,j} P(\mathbf{e}_i - \mathbf{e}_j) \right\}.$$

The above problem admits a closed form solution as follows:

$$P = U - A, \tag{8}$$

where

$$A = \sum\nolimits_{i<j} \boldsymbol{\theta}_{ij} w_{i,j} (\mathbf{e}_i - \mathbf{e}_j)^T. \tag{9}$$

The dual problem of Eq. (6) is given by:

$$\max\nolimits_{\boldsymbol{\theta}_{ij}} \mathcal{D}(\boldsymbol{\theta}_{ij}), \text{ s.t. } \|\boldsymbol{\theta}_{ij}\| \leq \alpha, \tag{10}$$

We note that $\mathcal{D}(\boldsymbol{\theta}_{ij})$ is continuously differentiable. Moreover, the gradient of $\mathcal{D}(\boldsymbol{\theta}_{ij})$ is Lipchitz continuous and we have that:

$$\frac{\partial \mathcal{D}(\boldsymbol{\theta}_{ij})}{\partial \boldsymbol{\theta}_{ij}} = w_{i,j}(U - A)(\mathbf{e}_i - \mathbf{e}_j).$$

According to the property of dual objective function we solve Eq. (10) by computing the following proximal operator until it converges:

$$\begin{aligned} \boldsymbol{\theta}^{k+1} =& \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \tfrac{1}{2\eta_k} \|\boldsymbol{\theta} - (\boldsymbol{\theta}^k - \eta_k \nabla_{\mathcal{D}}(\boldsymbol{\theta}^k))\|^2 \right\} \\ & \text{s.t. } \|\boldsymbol{\theta}_{i,j}\| \leq \alpha \\ =& \mathcal{P}^T(\boldsymbol{\theta}^k - \eta_k \nabla_{\mathcal{D}}(\boldsymbol{\theta}^k), \alpha), \end{aligned}$$

where $\mathcal{P}^T(X, \alpha)$ is defined as: the j^{th} column of X is:

$$\max(0, 1 - \tfrac{\alpha}{\sqrt{\sum_i X_{i,j}}}) \cdot X_j.$$

When the optimization converges, we obtain $\boldsymbol{\theta}$. Substituting $\boldsymbol{\theta}$ into Eq. (8) and Eq. (9), we finally find the optimal solution P of Eq. (6) that solves Eq. (3). [2]

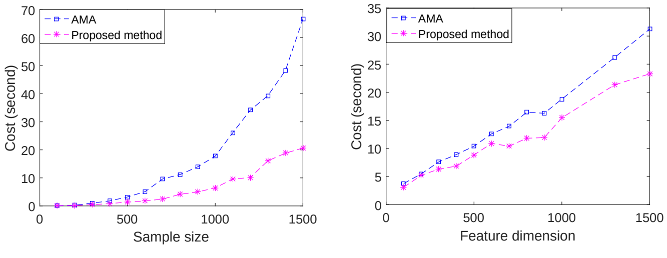
IV. Theoretical Analysis

In this section we present a theorem that gives the performance bound for the proposed robust convex clustering formulation in Eq. (2).

Theorem 1. *With probability of at least* $1 - \exp(-\frac{1}{2}(t - d\log(1 + \frac{t}{d})))$*, for a global minimizer* $\hat{P}, \hat{Q}$ *in Eq. (2), we have:*

$$\begin{aligned} &\frac{1}{2}\|(P^* + Q^*) - (\hat{P} + \hat{Q})\|_F^2 \\ &\quad \leq \alpha(gn+1)\frac{(\gamma_1+1)\sqrt{r}}{\kappa_1(r)}\left(\frac{2\alpha\sqrt{r}}{\kappa_1(r)} + \frac{2\beta\sqrt{c}}{\kappa_2(c)}\right) \\ &\quad\quad + 2\beta\frac{(\gamma_2+1)\sqrt{c}}{\kappa_2(c)}\left(\frac{2\alpha\sqrt{r}}{\kappa_1(r)} + \frac{2\beta\sqrt{c}}{\kappa_2(c)}\right). \end{aligned} \tag{11}$$

Here d is feature dimension and n is sample size. t, g, $\kappa_1(r)$, $\kappa_2(c)$, γ_1 and γ_2 are positive scalars. Detail definition of these numbers and proof of Theorem 1 can be found in Appendix.



(a) (b)

Fig. 2. (a) Computation cost (in second) with respect to sample size with 50 features. (b) Computation cost (in second) with respect to feature dimension with 500 samples.

V. Experiment

A. Efficiency: Convex Clustering Solvers

In this section, we evaluate the efficiency of the proposed convex clustering solver (based on dual formulation in Section III-C) and the state-of-the-art solver based on accelerated AMA [5]. We apply the two methods on synthetic data to compare how sample size and feature dimension affect the computation cost. All synthetic data contain 5 clusters. They are constructed as following: first, centroids for 5 clusters are generated: (1) select 10 entries and set the value as $\mathcal{N}(0, 16)$. These 10 entries are outlier features; (2) select $(d - 10)/5$ entries from entries that haven't been selected and generate non-zero values from $\mathcal{N}(0, 900)$ for these entries; (3) repeat (2) until the centroids for 5 clusters are all generated. Note that all the centroids are orthogonal to each other since we choose the appropriate locations of non-zero entries. Secondly, for each cluster we generate $n/5$ samples. The samples are generated by adding random variables $\mathcal{N}(0, 16)$ on their centroid. For the following experiment, we set all the regularization parameters to be 0.1 and all the weights are 1.

In the first experiment, we set the feature dimension to be 50 and the samples size vary in the set [100:100:1500] and repeat the experiment 50 times to eliminate noise. In the second experiment, we set the sample size to be 500 and vary feature dimension in the set [100:100:1500] and repeat the experiment 500 times. The results are shown in Figure 2. Since it is convex clustering, both methods converge to same solutions. In general, we see that the increases in sample size and dimension increase computation cost for both methods. But as the sample size increases when fixing feature dimension, the computation cost for the accelerated AMA increases exponentially while the computation cost for the proposed method only increases linearly. When we vary the feature dimension the computation costs for two methods increase almost linearly. But the slope for Dual is smaller than that of accelerated AMA. Hence the proposed method is faster than accelerated AMA, especially for large-scale cluster problem.

B. Efficiency of Robust Convex Clustering

We study the total cost of proposed algorithm under varying sample sizes and feature dimensions. We repeat all the experiment 20 times and report average time cost. The outlier feature dimension is fixed at 10. When varying sample size, we fix the feature dimension to be 50. When varying feature dimension, we fix the sample size to be 50. The results are

[2] The code is available at http://github.com/illidanlab/ConvexClustering

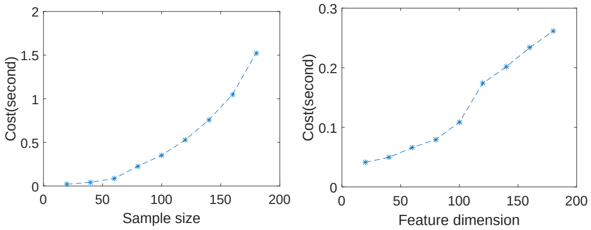


Fig. 3. These are the figures of total cost of proposed algorithm respect to sample size (left) and feature dimensions (right).

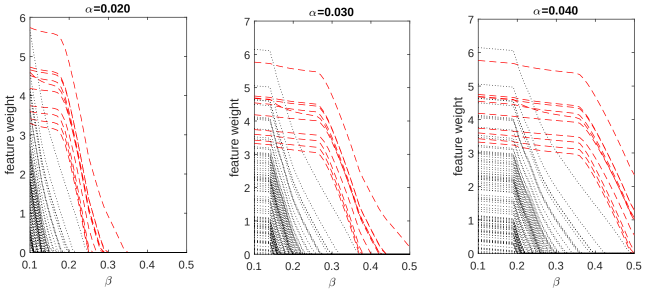


Fig. 4. The weight of relevant and outlier features in the robust component Q, where red lines are outlier features and black ones are relevant ones.

shown in Figure 3. We see that the cost increases linearly when feature dimension increases, and it increases super-linearly when sample size increase. This is consistent with the complexity of most clustering algorithms.

C. Effectiveness of Robust Component

Robust convex clustering can be used to select feature outliers. We report the weight of each feature in the robust component Q under different values of β in Figure 4. The weight of each feature is characterized by the norm of the corresponding row in Q. If the weight of a feature is high, then this feature is highly likely to be outlier. We set the total number of feature to be 100 and among the 100 features there are 10 outliers. The total sample number is 20. The cluster number is 2 and each cluster contains 10 data points. In Figure 4 the red lines are the 10 outlier features. The black lines are the 90 useful features. We see the weights of feature outliers are higher than the weights for almost all the other useful features. When α is larger, there is more overlap of outlier features and useful features. That is because when α is larger and larger, more points are grouped in the same cluster. The 5 designed clusters merged gradually when α is large enough. Hence useful features then are treated as outliers since the useful feature are designed to distinguish the 5 clusters. Thus the weights of useful features become larger and have more overlap with outlier features. Also we see that for a given α, when β decreases, weights of outlier features firstly become non-zero. Thus we conclude that robust convex clustering can effectively detect outlier features.

D. Cluster Quality: compare robust convex clustering and convex clustering

In this section we compare the cluster quality of the proposed robust convex clustering and convex clustering. We first apply the two methods on synthetic data with two clusters. The synthetic data are constructed similar to the data generated in efficiency part. The cluster quality is measured by *rand index* [16] between true clusters and clusters learned by each method. A high rand index indicates high cluster quality. The upper bound of rand index is 1. To simply our analysis, the synthetic data we create here only contain 2 clusters and each cluster has the same number of points. We use the distribution of $1.5 \times \mathcal{N}(0,1)$ to create outlier features and use $\mathcal{N}(0,1)$ to create centroids. The useful features are created by adding $\mathcal{N}(0,0.01)$ on centroids. The weights we use are $w_{i,j} = \exp(-\xi\|X_i - X_j\|_2^2)$, where ξ is a tunable parameter for optimal the clustering results.

TABLE I
CLUSTER QUALITY OF CONVEX CLUSTERING AND ROBUST CONVEX CLUSTERING RESPECT TO THE NUMBER OF DATA POINTS.

Sample size	16	18	20	22	24
CC	0.475	0.477	0.480	0.481	0.482
RCC	0.875	0.889	0.900	0.909	0.880

TABLE II
CLUSTER QUALITY OF CONVEX CLUSTERING AND ROBUST CONVEX CLUSTERING RESPECT TO NUMBER OF OUTLIER FEATURES.

Outlier features	2	4	6	8	10
CC	1.000	0.480	0.480	0.480	0.480
RCC	1.000	1.000	1.000	0.605	0.900

We first compare the cluster quality of two methods respect to the total number of data points, with 16 features and 4 outlier features. The results are shown in Table I. We see robust convex clustering outperforms traditional convex clustering in all the cases. When the sample size increases, we see the quality improves for both methods, because of the increased local density. Since Gaussian weights are used, high local density increases the cluster quality.

In the second experiment we fix the sample size to 20 and feature dimension to 30, and vary the number of outlier features. The results are shown in Table II. We see that when increasing the number of outlier features, the performance of convex clustering and robust convex clustering will become worse as expected. Convex clustering is very sensitive to outlier features: when outlier features is more than 2, the performance of convex clustering decreases rapidly and then it becomes stable. This is likely because the dominating effects of outlier features. On the other hand, robust convex clustering consistently performs much better than the convex clustering, thanks to its tolerance of feature outliers.

In the third experiment we vary the feature dimensionality, and fix the sample size at 16 and the number of outlier features to 4. We show the results in Table III. We see that when total feature dimension is high enough (the percentage of 'good' features are high), convex clustering can correct group the data. However when dimension is lower than 22 (more than 7% features are outliers), the performance is the same to all the cases, because that the clustering is purely based on the outliers. For robust convex clustering, with the increasing feature dimensionality, the performance increases and finally correctly group all points.

Besides the quantitative performance comparison of robust convex clustering and convex clustering. We also provide the cluster path of two method created by vary regularization

TABLE III
CLUSTER QUALITY OF CONVEX CLUSTERING AND ROBUST CONVEX CLUSTERING WITH VARYING FEATURE DIMENSION.

Rand index	14	16	18	20	22
CC	0.475	0.475	0.475	0.475	1.00
RCC	0.775	0.875	0.942	0.942	1.00

TABLE IV
CLUSTER QUALITY OF CONVEX CLUSTERING AND ROBUST CONVEX CLUSTERING ON REAL-WORLD DATA SETS.

Data set	Cardiotocography	Libras Movement	Seeds
CC	0.705	0.823	0.750
RCC	0.913	0.851	0.889

parameters. The clusters number is 2. The feature dimension is 16 and 4 features are outliers. Each cluster contains 8 points. To plot data in 2-dimension, we perform PCA and use the first two principal components. The results are shown in Figure 5. We see convex clustering first merges 15 points in one cluster and leaves one point as outlier point because of the distortion of feature outliers. But for robust convex clustering only one point is assigned to a wrong cluster. We use dash line to show its path. Hence robust convex clustering is much better than convex clustering when there exist outlier features.

Finally we compare robust convex clustering and convex clustering on three real-world data sets from UCI machine learning repository: Cardiotocography Data Set, Libras Movement Data Set, and Seeds Date Set. We report the rand index of the predicted clusters and real clusters in Table IV, and robust convex clustering is consistently better than convex clustering on these datasets.

VI. CONCLUSION

In this paper we proposed robust convex clustering which can identifies feature outliers and improves the cluster quality. Compared with convex clustering, we split the data matrix into two part: clustering component and robust component. By adding a group lasso regularization on robust component, we find feature outliers by the non-zero rows of robust component. We also show that robust convex cluster can improve the cluster quality compared with convex clustering when there exist feature outliers. Besides we present a new algorithm to solve convex clustering formulation which is a part of solving robust convex clustering formulation. Our method is faster as compared with the state-of-the-art solver based on accelerated AMA. In the paper we also further provided the theoretical bound of proposed robust convex clustering.

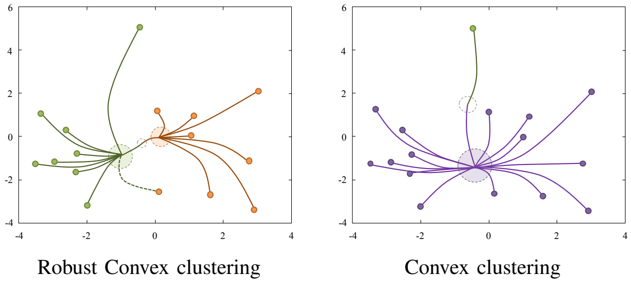
ACKNOWLEDGMENT

This research is supported in part by the NSF under grant number IIS-1565596 and IIS-1615597, and the ONR under grant number N00014-14-1-0631.

REFERENCE

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Ronald L Breiger, Scott A Boorman, and Phipps Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of mathematical psychology*, 12(3):328–383, 1975.



Robust Convex clustering

Convex clustering

Fig. 5. Cluster path of robust convex clustering and convex clustering.

[3] Gary K Chen, Eric C Chi, John Michael O Ranola, and Kenneth Lange. Convex clustering: An attractive alternative to hierarchical clustering. *PLoS Comput Biol*, 11(5):e1004228, 2015.

[4] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50. ACM, 2011.

[5] Eric C Chi and Kenneth Lange. Splitting methods for convex clustering. *arXiv preprint arXiv:1304.0499*, 2013.

[6] Keh-Shih Chuang, Hong-Long Tzeng, Sharon Chen, Jay Wu, and Tzong-Jer Chen. Fuzzy c-means clustering with spatial information for image segmentation. *computerized medical imaging and graphics*, 30(1):9–15, 2006.

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[8] Pinghua Gong, Jieping Ye, and Changshui Zhang. Robust multi-task feature learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 895–903. ACM, 2012.

[9] Lei Han and Yu Zhang. Reduction techniques for graph-based convex clustering. In *AAAI*, 2016.

[10] 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*, page 1, 2011.

[11] Stephen C Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[12] 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 (SSP), 2011 IEEE*, pages 201–204. IEEE, 2011.

[13] Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 339–348. AUAI Press, 2009.

[14] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[15] 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.

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

[17] Kean Ming Tan, Daniela Witten, et al. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324–2347, 2015.

[18] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

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

[20] Hongyuan Zha, Xiaofeng He, Chris Ding, Ming Gu, and Horst D Simon. Spectral relaxation for k-means clustering. In *Advances in neural information processing systems*, pages 1057–1064, 2001.

Robust Convex Clustering Analysis

Supplemental Materials

Qi Wang, Pinghua Gong, Shiyu Chang, Thomas S. Huang, Jiayu Zhou

APPENDIX

Here we show the proof of Theorem 1.

Assume a data matrix $X \in \mathbb{R}^{d\times n}$ with n samples and d features satisfies:

$$X = P^* + Q^* + \delta, \tag{12}$$

where $P^* \in \mathbb{R}^{d\times n}$ and $Q^* \in \mathbb{R}^{d\times n}$ are the underlying true decomposition of X. P^* is the component contianing cluster information of X. Q^* is the component containing outliers. $\delta \in \mathbb{R}^{d\times n}$ is the stochastic noise matrix. For the i^{th} row of δ, $\delta_{i,j} \sim \mathcal{N}(0, \sigma^2)$, where $i \in \mathbb{N}_d$ and $j \in \mathbb{N}_n$, i.e., for each feature there exists a normal distributed noise for all data points.

The optimization problem in Eq. (2) is:

$$(\hat{P}, \hat{Q}) = \underset{P,Q}{\operatorname{argmin}} \frac{1}{2}\|X - (P+Q)\|_F^2 + \alpha\|P\|_{FU} + \beta\|Q\|_{2,1}, \tag{13}$$

where $\hat{P} \in \mathbb{R}^{d\times n}$ and $\hat{Q} \in \mathbb{R}^{d\times n}$ are the optimal solution pair obtained by solving Eq. (13) and $\|P\|_{FU}$ is defined as:

$$\|P\|_{FU} = \sum_{i<j} w_{i,j}\|P_i - P_j\|_2.$$

We first present a theorem for the optimal solution pair which are important for our following theoretical analysis.

Theorem 2. *Consider the optimization problem in Eq. (2) for $n, d > 1$. Take the regularization parameters α and β as:*

$$\alpha > \lambda, \beta > \frac{\lambda}{\sqrt{n}}, \lambda = \sigma\sqrt{n(d+t)}, \tag{14}$$

where t is a positive scalar and σ is the standard deviation for each row of δ. Then with a probability of at lease $1 - \exp(-\frac{1}{2}(t - d\log(1+\frac{t}{d})))$, for a global minimizer $\hat{P}, \hat{Q}$ in Eq. (13) we have:

$$\frac{1}{2}\|(P^*+Q^*) - (\hat{P}+\hat{Q})\|_F^2 \leq \alpha(gn+1)\|(\hat{P}-P^*)^T\|_{2,1} + 2\beta\|\hat{Q}-Q^*\|_{2,1}, \tag{15}$$

where P^* and Q^* are the ground truth that generates the data matrix X and g is the maximum value of weight $w_{i,j}$.

Proof. Since $\hat{P}$ and $\hat{Q}$ are global minimizers, from Eq. (13) we have:

$$\frac{1}{2}\|X - (\hat{P}+\hat{Q})\|_F^2 + \alpha\|\hat{P}\|_{FU} + \beta\|\hat{Q}\|_{2,1} \leq \frac{1}{2}\|X - (P^*+Q^*)\|_F^2 + \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1},$$

By substituting the assumption we made about X in Eq. (12) we obtain:

$$\frac{1}{2}\|(P^*+Q^*) + \delta - (\hat{P}+\hat{Q})\|_F^2 + \alpha\|\hat{P}\|_{FU} + \beta\|\hat{Q}\|_{2,1} \leq \frac{1}{2}\|\delta\|_F^2 + \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1}. \tag{16}$$

Next we expand the first term in left hand side as:

$$\begin{aligned}&\frac{1}{2}\|(P^*+Q^*) + \delta - (\hat{P}+\hat{Q})\|_F^2\\ =&\frac{1}{2}\|(P^*+Q^*) - (\hat{P}+\hat{Q})\|_F^2 + \frac{1}{2}\|\delta\|_F^2\\ &+ \sum_{i,j}((P^*+Q^*) - (\hat{P}+\hat{Q}))\delta_{i,j}.\end{aligned} \tag{17}$$

By substituting Eq. (17) into Eq. (16) and rearranging all terms we obtain:

$$\begin{aligned}&\frac{1}{2}\|(P^*+Q^*) - (\hat{P}+\hat{Q})\|_F^2\\ &\leq \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1} - \alpha\|\hat{P}\|_{FU} - \beta\|\hat{Q}\|_{2,1}\\ &\quad + \sum_{i,j}(\hat{P}_{i,j} - P^*_{ij})\delta_{i,j} + \sum_{i,j}(\hat{Q}_{i,j} - Q^*_{ij})\delta_{i,j}\\ &\leq \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1} - \alpha\|\hat{P}\|_{FU} - \beta\|\hat{Q}\|_{2,1}\\ &\quad + \sum_j \|P^*_j - \hat{P}_j\|_2\|\delta_j\|_2 + \sum_i \|Q^*_i - \hat{Q}_i\|_2\|\delta_i\|_2,\end{aligned} \tag{18}$$

where the last line we used Cauchy-Schwartz inequality. P^*_j, $\hat{P}_j$,δ_j are the j^{th} column of P^*, $\hat{P}$ and δ respectively and Q^*_i, $\hat{Q}_i$,δ_i are the i^{th} row of Q^*, $\hat{Q}$ and δ respectively.

Next we compute the upper bounds for $\|\delta_i\|_2$ where i is the row index with the following Lemma about chi-squared random variable.

Lemma 1. *Let $\chi^2(d)$ be a chi-squared random variable with d degrees of freedom. Then the following holds [8]:*

$$Pr(\chi^2(d) \geq d+t) \leq \exp(-\frac{1}{2}(t - d\ \log(1+\frac{t}{d}))), t > 0. \tag{19}$$

Hence with the probability of at least $1 - \exp(-\frac{1}{2}(t - d\log(1+\frac{t}{d})))$, we have $\|\delta_i\|_2 \leq \sigma\sqrt{(d+t)} = \lambda/(\sqrt{n})$.

Similarly, we can compute the upper bound for $\|\delta_j\|_2$ where j is the column index:

$$\|\delta_j\|_2 \leq \sqrt{\|\delta\|_F^2} \leq \sigma\sqrt{n(d+t)} = \lambda. \tag{20}$$

Substitute the upper bound of $\|\delta_i\|_2$ and $\|\delta_j\|_2$ into Eq. (18) we obtain:

$$\begin{aligned}&\frac{1}{2}\|(P^*+Q^*) - (\hat{P}+\hat{Q})\|_F^2\\ &\leq \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1} - \alpha\|\hat{P}\|_{FU} - \beta\|\hat{Q}\|_{2,1}\\ &\quad + \lambda\sum_j \|P^*_j - \hat{P}_j\|_2 + \frac{\lambda}{\sqrt{n}}\sum_i \|Q^*_i - \hat{Q}_i\|_2.\end{aligned}$$

Now take the regularization parameter as Eq. (14), we obtain:

$$\begin{aligned}&\frac{1}{2}\|(P^*+Q^*) - (\hat{P}+\hat{Q})\|_F^2\\ &\leq \alpha\|P^*\|_{FU} + \beta\|Q^*\|_{2,1} - \alpha\|\hat{P}\|_{FU} - \beta\|\hat{Q}\|_{2,1}\\ &\quad + \alpha\|(\hat{P}-P^*)^T\|_{2,1} + \beta\|\hat{Q}-Q^*\|_{2,1}.\end{aligned} \tag{21}$$

We divide the right hand side of Eq. (21) into two parts, the terms that only contain P and the terms that only contain

Q, and bound both term accordingly. First we bound the part only contain P:

$$\begin{aligned}
&\alpha\|P^*\|_{FU} - \alpha\|\hat{P}\|_{FU} + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
=&\alpha\sum_{i<j} w_{i,j}(\|P_i^* - P_j^*\|_2 - \|\hat{P}_i - \hat{P}_j\|_2) + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
\leq&\alpha\sum_{i<j} w_{i,j}\|P_i^* - P_j^* - \hat{P}_i + \hat{P}_j\|_2 + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
\leq&\alpha\sum_{i<j} w_{i,j}(\|P_i^* - \hat{P}_i\|_2 + \|\hat{P}_j - P_j^*\|_2) + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
\leq&\alpha\sum_{i<j} g(\|P_i^* - \hat{P}_i\|_2 + \|\hat{P}_j - P_j^*\|_2) + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
\leq&\alpha gn\|(P^* - \hat{P})^T\|_{2,1} + \alpha\|(P^* - \hat{P})^T\|_{2,1} \\
\leq&\alpha(gn+1)\|(\hat{P} - P^*)^T\|_{2,1},
\end{aligned} \tag{22}$$

where indices i and j are both used to denote the column of P. Next the following terms that only contain Q can be bounded:

$$\beta\|Q^*\|_{2,1} - \beta\|\hat{Q}\|_{2,1} + \beta\|\hat{Q} - Q^*\|_{2,1} \leq 2\beta\|\hat{Q} - Q^*\|_{2,1}. \tag{23}$$

Combine Eq. (22), Eq. (23) and Eq. (21), we obtain:

$$\begin{aligned}
\frac{1}{2}\|(P^* + Q^*) - (\hat{P} + \hat{Q})\|_F^2 \leq& \alpha(gn+1)\|(\hat{P} - P^*)^T\|_{2,1} \\
&+ 2\beta\|\hat{Q} - Q^*\|_{2,1}.
\end{aligned} \tag{24}$$

This completes the proof of Theorem 2. □

Before we move on, we introduce two notations: $\mathcal{J}(\cdot)$ and $\mathcal{J}_\perp(\cdot)$ denote the index sets for nonzero rows and zero rows and the definations are:

$$\mathcal{J}(Q) = \{i | Q_i \neq 0\}, \mathcal{J}_\perp(Q) = \{i | Q_i = 0\}. \tag{25}$$

Next we make the following assumptions about the X, P^* and Q^* and give a theoretical bound between $P^* + Q^*$ and $\hat{P} + \hat{Q}$ based on these assumptions:

Assumption 1. *For a matrix pair $\Gamma_P \in \mathbb{R}^{d\times n}$ and $\Gamma_Q \in \mathbb{R}^{d\times n}$, let r and c ($1 \leq r \leq n$, $1 \leq c \leq d$) be the upper bounds of $|\mathcal{J}((P^*)^T)|$ and $|\mathcal{J}(Q^*)|$, respectively. Let γ_1 and γ_2 be positive scalars. We assume there exists positive scalars $\kappa_1(r)$ and $\kappa_2(c)$ such that:*

$$\kappa_1(r) = \min_{\Gamma_P,\Gamma_Q\in R(r,c)} = \frac{\|\Gamma_P + \Gamma_Q\|_F}{\|(\Gamma_P^T)^{\mathcal{J}(P^T)}\|_{2,1}} > 0, \tag{26}$$

$$\kappa_2(c) = \min_{\Gamma_P,\Gamma_Q\in R(r,c)} = \frac{\|\Gamma_P + \Gamma_Q\|_F}{\|(\Gamma_Q)^{\mathcal{J}(Q)}\|_{2,1}} > 0, \tag{27}$$

where $R(r,c)$ is defined as:

$$\begin{aligned}
R(r,c) =& \{\Gamma_P \in \mathbb{R}^{d\times n}, \Gamma_Q \in \mathbb{R}^{d\times n} | \Gamma_P \neq 0, \Gamma_Q \neq 0, | \\
&|\mathcal{J}(P^T)| \leq r, |\mathcal{J}(Q)| \leq c, \\
&\|(\Gamma_P^T)^{\mathcal{J}_\perp(P^T)}\|_{2,1} \leq \gamma_1\|(\Gamma_P^T)^{\mathcal{J}(P^T)}\|_{2,1}, \\
&\|(\Gamma_Q)^{\mathcal{J}_\perp(Q)}\|_{2,1} \leq \gamma_2\|(\Gamma_Q)^{\mathcal{J}(Q)}\|_{2,1}\},
\end{aligned}$$

where $\mathcal{J}(\cdot)$ and $\mathcal{J}_\perp(\cdot)$ are defined in Eq. (25) and $|\mathcal{J}|$ denotes the number of elements in the set $\mathcal{J}$.

Lemma 2. *Under Assumption 1 the following results hold with probability of at least $1 - \exp(-\frac{1}{2}(t - d\log(1 + \frac{t}{d})))$ $(t > 0)$:*

$$\|(P^* - \hat{P})^T\|_{2,1} \leq \frac{(\gamma_1 + 1)\sqrt{r}}{\kappa_1(r)}\left(\frac{2\alpha\sqrt{r}}{\kappa_1(r)} + \frac{2\beta\sqrt{c}}{\kappa_2(c)}\right), \tag{28}$$

$$\|Q^* - \hat{Q}\|_{2,1} \leq \frac{(\gamma_2 + 1)\sqrt{c}}{\kappa_2(c)}\left(\frac{2\alpha\sqrt{r}}{\kappa_1(r)} + \frac{2\beta\sqrt{c}}{\kappa_2(c)}\right). \tag{29}$$

The proof of a similar lemma can be found in [8].

Theorem 1 can be obtained by substituting Eq. (28) and Eq. (29) into Eq. (15).