Clustering using Max-norm Constrained Optimization

Ali Jalali ALIJ@MAIL.UTEXAS.EDU
University of Texas at Austin, 1 University Station, Code: C0806, Austin, TX 78712 USA

Nathan Srebro NATI@UCHICAGO.EDU
Toyota Technological Institute at Chicago, 6045 S. Kenwood Ave., Chicago, IL 60637 USA

Abstract

We suggest using the max-norm as a convex surrogate constraint for clustering. We show how this yields a better exact cluster recovery guarantee than previously suggested nuclear-norm relaxation, and study the effectiveness of our method, and other related convex relaxations, compared to other approaches.

1. Introduction

Clustering as the problem of partitioning data into clusters with strong similarity inside the clusters and strong dissimilarity across different clusters is one of the main problems in machine learning. In this paper, we consider the problem of cut-based, or *correlation*, clustering (Bansal et al., 2002) that has received a lot of attention recently (Ailon et al., 2011; Mathieu & Schudy, 2010; Bagon & Galun, 2011): Given a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ on n nodes with normalized symmetric affinity matrix A (for all $u,v\in\mathcal{V}$: $0\leq A_{uv}\leq 1$ and $A_{uu}=1$), we want to partition $\mathcal{V}$ into clusters $\mathcal{C}=\{C_1,\ldots,C_k\}$ so as to minimize the total *disagreement*

$$D(\mathcal{C})=\sum_{i=1}^{k}\sum_{u,v\in C_i}(1-A_{uv})+\sum_{i\neq j=1}^{k}\sum_{u\in C_i,v\in C_j}A_{uv}.$$

The first term, captures the *internal* disagreement inside clusters, and the second term captures the *external* agreement between nodes in different clusters. In an ideal cluster, the affinities between all members of the same cluster are 1 and the affinities between members of two different clusters are zero and hence the objective is zero. This objective does not require the number of clusters to be known ahead of time—we may decide to use any number of clusters, and this is accounted for in the objective. Unfortunately, finding a clustering minimizing the disagreement $D(\mathcal{C})$ is NP-Hard (Bansal et al., 2002).

We formulate this problem as an optimization of a convex disagreement objective over a non-convex set of *valid clustering* matrices (Section 2) and then consider convex relaxations of this constraint. Recently, Jalali et al. (2011) suggested a trace-norm (aka nuclear-norm) relaxation, casting the problem as minimizing an ℓ_1 loss and a trace-norm penalty, and providing conditions under which the true underlying clustering is recovered. Instead of trace-norm, we propose using the max-norm (aka γ_2 norm) (Srebro et al., 2005), which is a tighter convex relaxation than the trace-norm. Accordingly, we establish an exact recovery guarantee for our max-norm based formulation that is strictly better than the trace-norm based guarantee. We show that if the affinity matrix is a corruption of an "ideal" clustering matrix, with a certain bound on the corruption, then the optimal solution of the max-norm bounded optimization problem is exactly the ideal clustering (Section 3.1). We also discuss even tighter convex relaxations related to the max-norm, and suggest augmenting the convex relaxation with a single-linkage post-processing step in case of non-exact recovery, showing the empirical advantages of these approaches (Section 5).

The approach we suggests relies on optimizing an ℓ_1 objective subject to a max-norm constraint. A similar optimization problem with a trace-norm constraint (or trace-norm regularization) has recently been the subject of some interest in the context of "robust PCA" (Candes et al., 2011; Xu et al., 2012) and recovering the structure of graphical models with latent variables (Chandrasekaran et al., 2010). As with the trace-norm regularized variant, the ℓ_1 + max-norm problem can be formulated as an SDP and solved using standard solvers, but this is only applicable to fairly small scale problems. In Section 4, we discuss various optimization approaches to this problems, including approaches which preserve the sparsity of the solution.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.
Copyright 2012 by the author(s)/owner(s).

1.1. Relationship to the Goemans Willimason SDP Relaxation

Our convex relaxation approach is related to the classic SDP relaxations of max-cut (Goemans & Williamson, 1995) and more generally the cut-norm (Alon & Noar, 2006). In fact, if we are interested in a partition to exactly two clusters, the correlation clustering problem is essentially a max-cut problem, though with both positive and negative weights (i.e. a symmetric cut-norm problem), and our relaxation is essentially the classic SDP relaxation of these problems. Our approach and results differ in several ways.

First, we deal with problems with multiple clusters, and even when the number of clusters *is not* predetermined. If the number of clusters k *is* predetermined, the correlation clustering problem can be written as an integer quadratic program, with a k variables per node, and can be relaxed to an SDP. But this SDP will be very different from ours, and will involve a matrix of size $nk \times nk$, unlike our relaxation where the matrix is of size $n \times n$ regardless of the number of clusters. Consequently, the rounding techniques based on (random) projections typically employed for classic SDP relaxations do not seem relevant here. Instead, we employ a single-linkage post-processing as a form of "rounding" imperfect solutions.

Second, the type of guarantees we provide are very different from those in the Theory of Computation literature. Most of the SDP relaxation work we are aware of (including the classical work cited above) focuses on worst case constant factor approximation guarantees. On one hand, this means the guarantee needs to hold even on "crazy" inputs where there is really no reasonable clustering anyway, and second, and on the other hand it is not clear how approximating the objective to within a constant factor translates to recovering an underlying clustering. Instead, we prove that when the affinity matrix is close enough to following some underlying "true" clustering, the true clustering will be recovered *exactly*. This type of guarantee is more in the spirit of compressed sensing, which where *exact* recovery of a support set is guaranteed subject to conditions on the input (Jalali et al., 2011).

1.2. Other Clustering Approaches

There are several classes of clustering algorithms with different objectives. In hierarchical clustering algorithms such as UPGMA (Sneath & Sokal, 1973), SLINK (Sibson, 1973) and CLINK (Defays, 1977), the goal is to generate a sequence of clusterings by merging/splitting two clusters at each step of the sequence according to a *local* disagreement objective as opposed to our global $D(\mathcal{C})$. Because of this locality, these methods are known to be very sensitive to outliers.

Cut-based clustering algorithms such as k-means/medians (Steinhaus, 1957; Jain & Dubes, 1981), ratio association (Shi & Malik, 2000), ratio cut (Chan et al., 1994) and normalized cut (Yu & Shi, 2003) try to optimize an objective function globally. The main issue is that they are typically NP-Hard and need to know the number of clusters in advance.

In contrast, spectral clustering algorithms(von Luxburg, 2007) try to find the first k principal component of the affinity matrix or a transformed version of that (Meilă & Shi, 2001). These methods require the number of clusters in advance and has been shown to be tractable (convex) relaxations to NP-Hard cut-based algorithms (Dhillon et al., 2005). These methods are again very sensitive to outliers.

2. Problem Setup

Our approach is based on representing a clustering $\mathcal{C}$ through its incidence matrix $K(\mathcal{C}) \in \mathbb{R}^{n\times n}$ where $K_{uv} = 1$ iff u and v belong to the same cluster in $\mathcal{C}$ (i.e. $u, v \in C_i$ for some i), and $K_{uv} = 0$ otherwise (i.e. if u and v belong to different clusters). The matrix $K(\mathcal{C})$ is thus a permuted block-diagonal matrix, and can also be thought of as the edge incidence matrix of a graph with cliques corresponding to clusters in $\mathcal{C}$. We will say that a matrix K is a **valid clustering matrix**, or sometimes simply **valid**, if it can be written as $K = K(\mathcal{C})$ for some clustering $\mathcal{C}$ (i.e. if it is a permuted block diagonal matrix, with 1s in the diagonal blocks).

The disagreement can then be written as either:

$$D(\mathcal{C}) = \|A - K(\mathcal{C})\|_1 = \sum_{u,v} |A_{uv} - K(\mathcal{C})_{uv}| \qquad (1)$$

or as:

$$D(\mathcal{C}) = \sum_{u,v} K(\mathcal{C})_{uv}(1 - 2A_{uv}) + \sum_{uv} A_{uv} \ , \qquad (2)$$

where the term $\sum_{uv} A_{uv}$ does not depend on the clustering $\mathcal{C}$ and can thus be dropped.

We now phrase the correlation clustering problem as matrix problem, where we would like to solve

$$\min_K D(K) \ \text{ s.t. } \ K \text{ is a valid clustering matrix.} \qquad (3)$$

The problem is that even though the objectives (1) and (2) are convex, the constraint that K is valid is certainly not convex. Our approach to correlation clustering will thus be to relax this non-convex constraint (the validity of K) to a convex constraint.

We note that although both the absolute error objective (1) and the linear objective (2) agree on valid clustering matrices (or more generally, on binary matrices K), they can differ when K is fractional, and especially when A is also fractional. The choice of

objective can thus be important when relaxing the validity constraint to a convex constraint. More specifically, as long as A is binary (i.e. $A_{uv} \in \{0,1\}$), and $0 \le K_{uv} \le 1$, even if K is fractional, the two objectives agree. Non-negativity of K_{uv} is ensured in some, but not all, of the convex relaxations we study. When non-negativity is not ensured, the absolute error objective (1) would tend to avoid negative values, but the linear objective might certainly prefer them. More importantly, once the affinities A_{uv} are also fractional, the two objectives differ even for $0 \le K_{uv} \le 1$. While the linear objective would tend to not care much about entries with affinities close to $1/2$, the absolute error objective would tend to encourage fractional values in thees cases.

The linear objective also has some optimization advantages over the absolute function as well. From a numerical optimization point of view, dealing with the linear objective function is easier since we do not need to compute the sub-gradients of the ℓ_1-norm.

3. Max-Norm Relaxation

As discussed in the previous Section, we are interested in optimizing over the non-convex set of valid clustering matrices. The approach we discuss here is to relaxing this set to the set of matrices with bounded *max-norm* (Srebro et al., 2005). The max-norm of a matrix K is defined as

$$\|K\|_{\max} = \min_{K=RL^T} \|R\|_{\infty,2}\|L\|_{\infty,2}$$

where, $\|\cdot\|_{\infty,2}$ is the maximum of the ℓ_2 norm of the rows, and the minimization is over factorization of any internal dimensionality. It is not hard to see that if K is a valid clustering matrix, with $K = K(\mathcal{C})$, then $\|K\|_{\max} = 1$. This is achieved, e.g., by a factorization with $R = L$, and where each row R_u of R is a (unit norm) indicator vector with $R_{ui} = 1$ for $u \in C_i$ and zero elsewhere.

Relaxing the validity constraint to a max-norm constraint, and using the absolute error objective, we obtain the following convex relaxation of the correlation clustering problem:

$$\widehat{K} \;=\; \arg\min_{K} \|A - K\|_1 \quad \text{s.t.} \quad \|K\|_{\max} \le 1. \tag{4}$$

Alternatively, we could have used the linear objective (2) instead. In any case, after finding $\widehat{K}$, it is easy to check whether it is valid, and if so recover the clustering from its block structure. If $\hat{K}$ is valid, we are assured the corresponding clustering is a globally optimal solution of the correlation clustering problem.

3.1. Theoretical Guarantee

Assuming there exists an underlying true clustering, we provide a worst-case (deterministic) guarantee for exact recovery of that clustering in the presence of noise when the affinity matrix A is a binary $0-1$ matrix using absolute objective. The flavor of our result is similar to (Jalali et al., 2011) for trace-norm, except that we show the max-norm constraint problem recovers the underlying clustering with larger noise comparing to trace-norm constraint. This matches our intuition that max-norm is a tighter relaxation than trace-norm for valid clustering matrices.

To present our theoretical result, we start by introducing an important quantity that our main result is based upon. Suppose $\mathcal{C}^* = \{C_1^*, \dots, C_k^*\}$ is the underlying true clustering. For a node u and a cluster C_i^*, let $d_{u,C_i^*} = \frac{\sum_{v\in C_i^*} A_{u,v}}{|C_i^*|}$ if $u \notin C_i^*$ and $d_{u,C_i^*} = 1 - \frac{\sum_{v\in C_i^*} A_{u,v}}{|C_i^*|}$ otherwise and

$$D_{\max}(A,K) \equiv D_{\max}(A, K(\mathcal{C}^*)) = \max_{u,i}\; d_{u,C_i^*}$$

be the maximum of the disagreement ratios on the adjacency matrix. This definition is inspired by (Jalali et al., 2011) but is slightly different. Notice that the larger $D_{\max}(A,K)$ is, the more noisy (comparing to ideal clusters) the graph is; and hence, the harder the clustering becomes. In particular for ideal clusters (fully connected inside and fully disconnected outside clusters), we have $D_{\max}(A,K) = 0$.

We would like to ensure that when $D_{\max}(A,K)$ is small enough, our method can recover K. The following lemma helps us understand the information theoretic limit of $D_{\max}(A,K)$, i.e. what value of D_{max} is certainly *not* enough to ensure recovery, even information theoretically:

Lemma 1. *For any clustering $\mathcal{C} = \{C_1, \dots, C_k\}$ and for all $\gamma > \frac{2}{5+r}$ with $r = \frac{n^2}{\sum_i |C_i|^2}$, there exists an affinity matrix A such that $D_{\max}(A, K(\mathcal{C})) = \gamma$ and the combinatorial program (3) does not output $\mathcal{C}$.*

Note that the minimum of $\frac{2}{5+r}$ is attained when all clusters have equal sizes. If we have k^* clusters of size $\frac{n}{k^*}$, then $r = k^*$ and the bound in Lemma 1 asserts that if $D_{\max}(A,K) > \frac{2}{k^*+5}$, then there are examples for which the original clustering cannot be recovered by the combinatorial program (3). This implies that $D_{\max}(A,K)$ cannot be scaled better than $\Theta(\frac{1}{k^*})$ in general even without convex relaxation.

Suppose there exist a true underlying clustering $\mathcal{C}^*$ with k^* clusters. Let C_{min} be the smallest size underlying true cluster and we are given an affinity matrix A with $D_{\max} = D_{\max}(A, K(\mathcal{C}^*))$. Introducing lagrange multiplier μ, we consider the optimization problem

$$\widehat{K}_\mu \;=\; \arg\min_{K} \frac{1-\mu}{n^2}\|A - K\|_1 \;+\; \mu\|K\|_{\max}. \tag{5}$$

The following theorem characterizes the noise regime under which the simple max-norm relaxation (5) recovers $\mathcal{C}^*$.

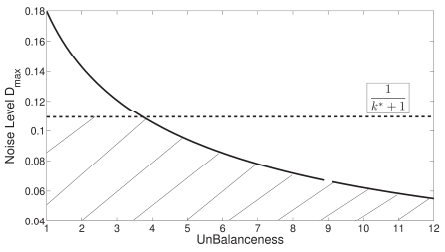


Figure 1. Theorem 1 guarantee region of the noise level $D_{\max}$ vs the unbalanceness parameter $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2$.

Theorem 1. *For $0-1$ matrix A, if $D_{\max} < \frac{1}{k^*+1}$ is small enough to satisfy $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2 \leq \frac{(1-3D_{\max})^2}{(1+D_{\max})D_{\max}}$ then, for any μ_0 satisfying $\frac{(1+D_{\max})}{(1-3D_{\max})|C_{\min}|^2} < \frac{(1-\mu_0)k^*}{\mu_0 n^2} < \frac{(1-3D_{\max})k^*}{D_{\max}\sum_i |C_i^*|^2}$, the matrix $\widehat{K}_{\mu_0}$ (the solution to (5)) is unique and equal to the matrix $K^* = K(\mathcal{C}^*)$ (the solution to (3)).*

Remark 1: Consider the parameter $\frac{1}{k^*}\sum_i \left(\frac{|C_i^*|}{|C_{\min}|}\right)^2$ in the theorem. Notice that for a balanced underlying clustering (k^* clusters of size n/k^*), this parameter is 1 and as the underlying clustering gets more and more unbalanced, this parameter increases. That motivates to call it *unbalanceness* of the clustering. It is clear that as unbalanceness parameter increases, the region of $D_{\max}$ for which our theorem guarantees the clustering recovery shrinks. We plot the admissible region of $D_{\max}$ due to unabalanceness in Fig 1.

Remark 2: According to the Lemma 1, the bound on $D_{\max}$ is order-wise tight and can be only improved by a constant in general.

3.2. Comparison to Trace-Norm Clustering

Since the max-norm constraint is strictly a tighter relaxation to the trace-norm constraint, we expect the max-norm algorithm to perform better. Our theorem shows improvement over the guarantees provided for trace-norm clustering. Comparing to the result of (Jalali et al., 2011) on trace-norm ($D_{\max} \leq \frac{|C_{\min}|}{4n}$), the max-norm tolerates more noise. To see this, consider a balanced clustering, then trace-norm requires $D_{\max} \leq \frac{1}{4k^*}$ and max-norm requires $D_{\max} \leq \min(\frac{1}{k^*+1}, 0.1789)$ which is larger than $\frac{1}{4k^*}$ for all k^*. The difference gets more clear for unbalanced clustering. Suppose we have one small cluster of constant size $|C_{\min}|$ and other clusters are approximately of size $\frac{n}{k^*}$. As (n, k^*) scales, trace-norm guarantee requires that $D_{\max} = o(\frac{1}{n})$ which is inverse proportional to the size of the smallest cluster, whereas, max-norm guarantee requires $D_{\max} = o(\frac{k^*}{n})$ which is inverse proportional to the size of the largest cluster. This is a huge theoretical advantage in our theorem.

Further, we compare our algorithm with trace-norm algorithm (Jalali et al., 2011) and SLINK on a probabilistic setup. Start from two different ideal clusters on 100 nodes: a) *Balanced* clusters: four ideal clusters of size 25, b) *Unbalanced* clusters: three ideal clusters of size 30 and one ideal cluster of size 10. Then, gradually increase $D_{\max}$ on both graphs and run all algorithms and report the probability of success in exact recovery of the underlying clusters. Although our theoretical guarantee is for binary affinity matrices, here, we run the same experiment for fractional affinity matrix. We run all experiments for both absolute and linear objectives. Fig. 3.1 shows that in all cases max-norm outperforms the trace-norm and the improvement is more significant for unbalanced clustering with fractional affinity matrix. Moreover, this experiments reveal that the absolute objective has slight advantage if the affinity matrix is binary and clusters are balanced; otherwise, the linear objective is better.

4. Max-norm + ℓ_1-norm Optimization

In this Section we consider optimization problems of the form (4). This problem recovers a sparse and low-rank matrix from their sum, considering max-norm as a proxy to rank. In Section 4.1, we discuss how (4) can be formulated as an SDP, allowing us to easily solve it using standard SDP solvers, as long as the problem size is relatively small. We then propose three other methods to numerically solve the optimization problem (4).

4.1. Semi-Definite Programming Method

Following Srebro et al. (2005), we introduce dummy variables $L, R \in \mathbb{R}^{n\times n}$ and reformulate (4) as the following SDP problem

$$\widehat{K} = \arg\min_{K,L,R} \|A-K\|_1 \quad \text{s.t.} \quad \begin{bmatrix} L & K \\ K^T & R \end{bmatrix} \succeq 0 \quad \text{and} \quad L_{ii}, R_{ii} \leq 1$$

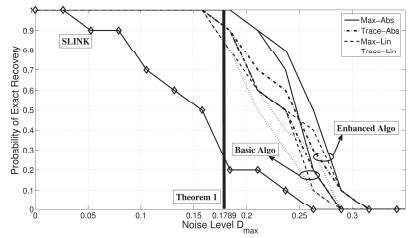
These constraints are equivalent to the condition $\|K\|_{\max} \leq 1$. This SDP can be solved using generic SDP solvers, though is very slow and is not scalable to large problems.

4.2. Factorization Method

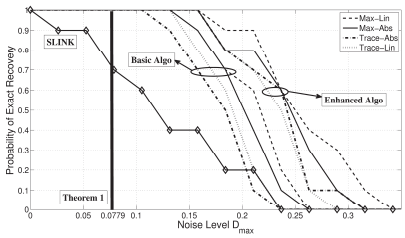
Motivated by Lee et al. (2010), we introduce dummy variables $L, R \in \mathbb{R}^{n\times n}$ and let $K = LR^T$. With this change of variable, we can reformulate (4) as

$$\widehat{K} = \widehat{L}\widehat{R}^T = \arg\min_{L,R} \|A - LR^T\|_1 \quad \text{s.t.} \quad \|L\|_{\infty,2}, \|R\|_{\infty,2} \leq 1.$$

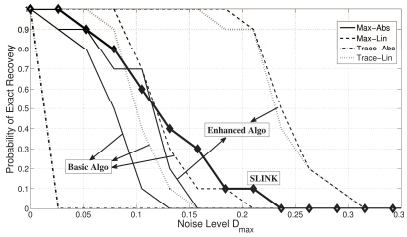
This problem is not convex, but it is guaranteed to have no local minima for large enough size of the problem (Burer & Choi, 2006). Furthermore, if we now the



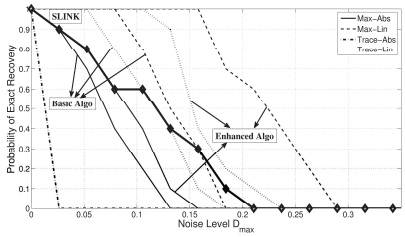
(a) Balanced; Binary



(b) UnBalanced; Binary



(c) Balanced; Fractional



(d) UnBalanced; Fractional

Figure 2. Probability of exact clustering recovery for max-norm and trace-norm constrained algorithms under absolute $\|A-K\|_1$ and linear $\sum_{i,j} K_{ij}(1-2A_{ij})$ objectives. There are 4 clusters of size 25 for the balanced case and three clusters of size 30 + one cluster of size 10 for the unbalanced case. We consider two cases for each graph; where the affinity matrix is binary and when it is not. We both show the results for simple max-norm relaxation (basic algorithm) and tighter relaxations presented in Section 5 (enhanced algorithm). The result shows that max-norm constrained optimization recovers the exact clustering matrix under higher noise regimes better than trace-norm and single-linkage algorithm. Also, the linear objective seems to be performing better than the absolute objective for the clustering problem in most cases.

optimal solution $\hat{K}$ has rank at most r, we can take L, R to be $\mathbb{R}^{n\times(r+1)}$. In practice, we truncate to some reasonably high rank r even without a known guarantee on the rank of the optimal solution. To solve this problem iteratively, Lee et al. (2010) suggest the following update

$$\begin{bmatrix} L \\ R \end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix} L \\ R \end{bmatrix}_k + \frac{\tau}{\sqrt{k}} \begin{bmatrix} \mathbf{Sign}(A-LR^T)\ R \\ \mathbf{Sign}(A-LR^T)^T\ L \end{bmatrix}_k\right).$$

The projection $\mathcal{P}_{\max}(\cdot)$ operates on rows of L and R; if ℓ_2-norm of a row is less than one, it remains unchanged, otherwise it will be rescaled so that the ℓ_2-norm becomes one.

A possible problem with the above formulation is the lack of "sparsity" in the following sense: The ℓ_1 objective is likely to yield and optimal solution K^* with many non-zeros in $A-K^*$, i.e. where K^* is *exactly* equal to A on some of the entries. However, gradient steps on the factorization are not likely to end up in exactly sparse solutions, and we are not likely to see any such sparsity in solutions of this method.

4.3. Loss Function Method

There are gradient methods such as truncated gradient (Langford et al., 2009) that produce sparse solution, however, these methods cannot be applied to this problem. We introduce a surrogate optimization problem to (4) by adding a loss function. For some large $\lambda \in \mathbb{R}$, solve

$$\widehat{K} = A - \widehat{Z} = \arg \min_{Z,L,R} \|Z\|_1 + \lambda\|A - Z - LR^T\|_2^2$$
$$\text{s.t.} \quad \|L\|_{\infty,2}, \|R\|_{\infty,2} \le 1.$$

Here, the matrix Z is sparse and includes the disagreements. For sufficiently large values of λ, the loss function ensures that the matrix $A-Z$ is close to the matrix LR^T that is a bounded max-norm matrix. To solve this problem iteratively, we use the update

$$Z_{k+1} = \mathcal{P}_{\ell_1}\left(Z_k + \frac{\tau\lambda}{\sqrt{k}}(A - Z - LR^T)_k\right)$$

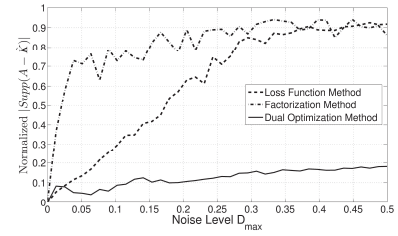
$$\begin{bmatrix} L \\ R \end{bmatrix}_{k+1} = \mathcal{P}_{\max}\left(\begin{bmatrix} L \\ R \end{bmatrix}_k + \frac{\tau\lambda}{\sqrt{k}} \begin{bmatrix} (A-Z-LR^T)\ R \\ (A-Z-LR^T)^T\ L \end{bmatrix}_k\right).$$

Here, $\mathcal{P}_{\ell_1}(\cdot)$ operates on entries; if an entry has the same sign before and after the update, it remains unchanged; otherwise, it will be set to zero. Solving directly for large values of λ might cause some problems due to the finite numerical precision. In practice, we start with some small value say $\lambda = 1$ and double the value of λ after some iterations. This way, we gradually put more and more emphasis on the loss function.

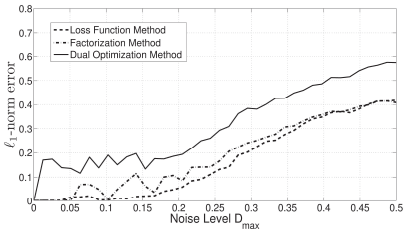
4.4. Dual Decomposition Method

Inspired by Rockafellar (1970), we first reformulate (4) by introducing a dummy variable $Z \in \mathbb{R}^{n\times n}$ as follows

$$\widehat{K} = \arg \min_{Z,K} \|A - K\|_1$$
$$\text{s.t.} \quad \|Z\|_{\max} \le 1 \qquad \text{and} \qquad Z = K.$$



(a) $|\mathrm{Supp}(A - \hat{K})|$



(b) $\|K^* - \hat{K}\|_1$

Figure 3. Comparison of the proposed numerical optimization methods in terms of the sparsity of the solution they provide and the ℓ_1 error of the estimation.

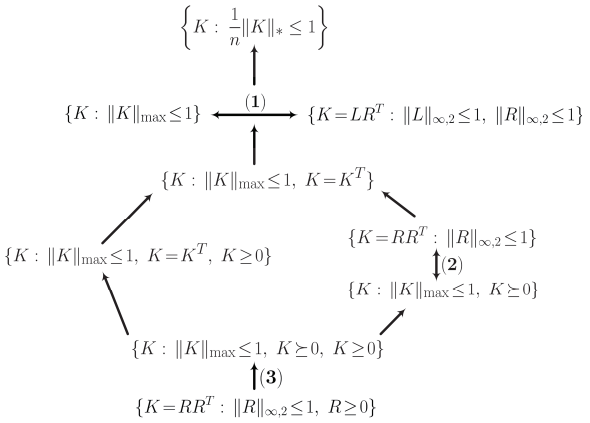


Figure 4. Summary of possible convex relaxations of the set of valid clustering matrices and their relations. Here, $\|\cdot\|_*$ represents the trace (nuclear) norm, $\|\cdot\|_{\infty,2}$ represents the maximum ℓ_2 norm of the rows, "$\ge$" is used for element-wise positiveness and "$\succeq$" is used for positive semi-definiteness. Each double-ended arrow represents the equivalence of two sets. Each single-ended arrow in this figure represents a *strict* sub-set relation between two sets.

Then, introducing a Lagrange multiplier $\Lambda \in \mathbb{R}^{n\times n}$, we propose the following equivalent problem:

$$\widehat{K} = \arg\max_{\Lambda} \min_{Z,K} \|A - K\|_1 + \langle\langle \Lambda,\ K - Z\rangle\rangle$$
$$\text{s.t.} \quad \|Z\|_{\max} \le 1.$$

Here, $\langle\langle \cdot,\ \cdot \rangle\rangle$ is the trace of the product. This problem is a saddle-point convex problem in (Z, K, Λ). To solve this, we iteratively fix Λ and optimize over (K, Z) and then, using those optimal values of (K, Z), update Λ.

For a fixed Λ, the problem can be separated into two optimization problems over K and Z as

$$\widehat{K}(\Lambda) = \arg\min_{K} \|A - K\|_1 + \langle\langle \Lambda,\ K\rangle\rangle$$

which can be solved using factorization method discussed above, and

$$\widehat{Z}(\lambda) = \arg\min_{Z} -\langle\langle \Lambda,\ Z\rangle\rangle \quad \text{s.t.} \quad \|Z\|_{\max} \le 1.$$

which is a soft thresholding; if $|\Lambda_{ij}| > 1$ then, $\widehat{K}(\Lambda)_{ij} = -\mathbf{Sign}(\Lambda_{ij})$; otherwise $\widehat{K}(\Lambda)_{ij} = A_{ij}$.

Using $\widehat{K}(\Lambda_k)$ and $\widehat{Z}(\Lambda_k)$, we update Λ as follows

$$\Lambda_{k+1} = \Lambda_k - \frac{\tau}{\sqrt{k}}(\widehat{K}(\Lambda_k) - \widehat{Z}(\Lambda_k))$$

until it converges. One criterion for the convergence of this method is to round both matrices $\widehat{K}, \widehat{Z}$ and check if they are equal. To use this criterion, we need to initialize the two matrices very differently to avoid the stopping due to the initialization.

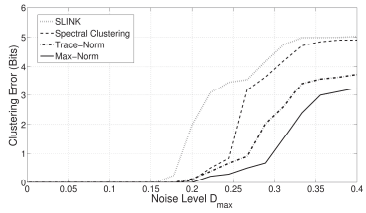
4.5. Numerical Comparison

We compare the performance of these methods. For three ideal clusters of size 20 with noise level $D_{\max}$, we run all three algorithms for 2000 iterations. We consider an initial step size $\tau = 1$ for all methods, and, for the loss function method, we doubel λ every 100 iterations. For the dual method, we update Λ for 20 times and run 100 iterations of the factorization method for the max-norm sub-problem at each update. We report the sparsity of the solution $A - \widehat{K}$ as well as the ℓ_1-norm of the error $\|\widehat{K} - K^*\|_1$ for each algorithm in Fig 3. This result shows that there is a trade-off between sparsity and the error – the dual optimization method provides consistently a sparse solution, where, factorization and loss function methods provide small error. The sparsity of loss function method gets worse as the noise increases.

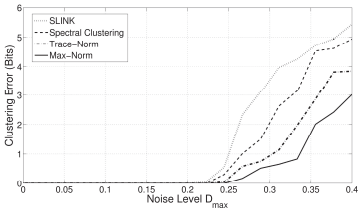
5. Tighter Relaxations

In this section, we improve our basic algorithm in two ways: first, we use a tighter relaxation for valid clustering constraint and second, we add a single-linkage step after we recovered the clustering matrix. Although max-norm is a tighter relaxation comparing to trace-norm, we would like to go further and introduce tighter relaxations. Figure 4 summarizes different possible relaxations based on max-norm. The tightest relaxation we suggest is $\{K = RR^T : \|R\|_{\infty,2} \le 1, R \ge 0\}$ based on the intuition that a clustering matrix is symmetric and has a trivial factorization $R \in \mathbb{R}^{n\times k}$, where, R_{ij} is non-zero if node i belongs to cluster j. Next lemma formalizes this result.

Lemma 2. *All relaxation sets shown in Fig. 4 are convex and the strict subset relations hold.*

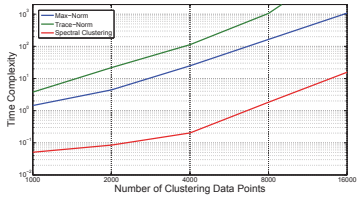


(a) Balanced; Fractional

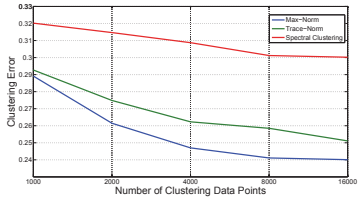


(b) UnBalanced; Fractional

Figure 5. Comparison of our *best* proposed method which is the linear objective over tight relaxation (followed by a single-linkage algorithm) with trace-norm counterpart, single-linkage algorithm and spectral clustering. Here, we plot the entropy-based distance of the recovered clustering with the underlying true clustering.



(a) Time Complexity



(b) Clustering Error

Figure 6. Comparison of our *best* proposed method which is the linear objective over tight relaxation (followed by k-means) with trace-norm and spectral clustering in terms of time complexity and clustering error on MNIST dataset.

This suggests using the tightest convex relaxation, that is constraining to K such that there exists $R >= 0, \|R\|_{\infty,2} <= 1$ with $K = RR^T$ (the set of matrices K with a factorization $K = RR^T, R >= 0$ is called the set of *completely positive matrices* and is convex (Berman & Shaked-Monderer, 2003)). We optimize over this relaxation by solving the following optimization problem over R:

$$\begin{aligned} \widehat{R} \;=\; \arg\min_{R} \; & \|A - RR^T\|_1 \\ \text{s.t.} \quad & \|R\|_{\infty,2} \le 1 \quad \& \quad R \ge 0. \end{aligned} \tag{6}$$

and setting $\widehat{K} = \widehat{R}\widehat{R}^T$. Although the constraint on $\widehat{K}$ *is* convex, the problem (6) is *not* convex in R.

5.1. Single-linkage Post Processing

The matrix $\widetilde{K}$ extracted from (6) might diverge from a valid clustering matrix in two ways: firstly, it might not have the structure of a valid clustering and secondly, even if it has the structure, the values might not be integer. We run SLINK on $\widetilde{K}$ as a "rounding scheme" to fix both of the above problems. SLINK gives a sequence of clusterings $\mathcal{C}_1, \ldots, \mathcal{C}_n$. To pick the best clustering, we choose

$$\widehat{K} = \arg\min_i \|A - K(\mathcal{C}_i)\|_1. \tag{7}$$

The matrix $\widetilde{K}$ can be viewed as a refined version of the affinity matrix A and hence the second step of the algorithm can be replaced by other hierarchical clustering algorithms.

5.2. Comparison with Other Algorithms

We compare our enhanced algorithm with the trace-norm algorithm (Jalali et al., 2011) followed by SLINK and SLINK itself. In all cases we pick a clustering from SLINK hierarchy using (7). The setup is identical to the experiment explained in Section 3.2. Fig 3.1 summarizes the results and shows that our algorithm outperforms all competitive methods significantly.

Besides the exact recovery, we would like to investigate that as noise level $D_{\max}$ increases, how bad the output of our algorithm gets. Using "variation of information" (Meilă, 2007) as a distance measure, we compare our algorithm with trace-norm, SLINK and spectral clustering (von Luxburg, 2007) for both balanced and unbalanced clusterings. For the spectral clustering method, we first find the largest $k = 4$ principal components of A and then, run SLINK on principal components. Fig 5 shows the result indicating that max-norm, even when the noise level is high, outputs a clustering that is not far from the true clustering.

5.3. MNIST Dataset

To demonstrate our method in a realistic and larger scale data set, we run our enhanced algorithm, trace-norm and spectral clustering on MNIST Dataset (LeCun et al., 1998). For each experiment, we pick a total of n data points from 10 different classes ($n/10$ from each class) and construct the affinities using Gaussian kernel as explained in (Bühler & Hein, 2009). We report the time complexities and clustering errors as previous experiment in Fig 5.1. For the spectral cluster-

ing, we take SVD using Matlab and pick the top 10 principal components followed by k-means.

References

Ailon, N., Avigdor-Elgrabli, N., and Liberty, E. An improved algorithm for bipartite correlation clustering. *European Symposium on Algorithms*, 2011.

Alon, N. and Noar, A. Approximating the cut-norm via grothendieck's inequality. *SIAM Journal on Computing*, 35, 2006.

Bagon, S. and Galun, M. Large scale correlation clustering optimization. *Arxiv preprint arXiv:1112.2903*, 2011.

Bansal, N., Blum, A., and Chawla, S. Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, 2002.

Berman, A. and Shaked-Monderer, N. *Completely Positive Matrices*. World Scientific Publication, 2003.

Bühler, T. and Hein, M. Spectral clustering based on the graph p-laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 81–88. ACM, 2009.

Burer, S. and Choi, C. Computational enhancements in low-rank semidefinite programming. *Optimization Methods and Software*, 21:493–512, 2006.

Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of ACM*, 58:1–37, 2011.

Chan, P., Schlag, M., and Zien, J. Spectral k-way ratio cut partitioning. *IEEE Trans. CAD-Integrated Circuits and Systems*, 13:1088–1096, 1994.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. arXiv:1008.1290, August 2010.

Defays, D. An efficient algorithm for a complete link method. *The Computer Journal (British Computer Society)*, 20:364–366, 1977.

Dhillon, I., Guan, Y., and Kulis, B. A unified view of kernel k-means, spectral clustering and graph cuts. *UTCS Technical Report TR-04-25, University of Texas at Austin*, 2005.

Goemans, M. and Williamson, D. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of ACM*, 42, 1995.

Gray, L.J. and Wilson, D.G. Nonnegative factorization of positive semidefinite nonnegative matrices. *Linear Algebra and its Applications*, 31:119–127, 1980.

Jain, A. K. and Dubes, R. C. *Algorithms for Clustering Data*. Prentice-Hall, 1981.

Jalali, A., Chen, Y., Sanghavi, S., and Xu, H. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

J.B., Hiriart-Urruty and C., Lembarechal. *Convex Analysis and Minimization Algorithms I*. Springer-Verlag, Netherland, 1991.

Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *The Journal of Machine Learning Research*, 10:777–801, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J., Recht, B., Salakhutdinov, R., Srebro, N., and Tropp, J.A. Practical large-scale optimization for max-norm regularization. In *NIPS*, 2010.

Lee, T., Shraibman, A., and Spalek, R. A direct product theorem for discrepancy. In *Proceedings of the IEEE 23rd Annual Conference on Computational Complexity*, 2008.

Mathieu, C. and Schudy, W. Correlation clustering with noisy input. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 712–728. Society for Industrial and Applied Mathematics, 2010.

Meilă, M. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98:873–895, 2007.

Meilă, M. and Shi, J. Learning segmentation by random walks. In *NIPS*, 2001.

Rockafellar, R. T. *Convex Analysis*. Prenticeton University Press; Princeton, NJ, 1970.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Sibson, R. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society)*, 16(1):30–34, 1973.

Sneath, P. M. A. and Sokal, R. R. *Numerical Taxonomy*. Freeman, San Francisco CA, 1973.

Srebro, N. *Learning with Matrix Factorizations*. PhD thesis, Massachusetts Institute of Technology, Boston, MA, 2004.

Srebro, N., Rennie, J., and Jaakkola, T. Maximum-margin matrix factorization. In *NIPS*, 2005.

Steinhaus, H. Sur la division des corps matériels en parties. *Bulletin de l'Acadbmie Polonaise des Sciences*, 4:801–804, 1957.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing, Springer*, 17, 2007.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 2012.

Yu, S. X. and Shi, J. Multiclass spectral clustering. In *International Conference on Computer Vision Proceedings*, 2003.