Theory of Spectral Method for Union of Subspaces-Based Random Geometry Graph

Gen Li [1] Yuantao Gu [1]

Abstract

Spectral method is a commonly used scheme to cluster data points lying close to Union of Subspaces, a task known as Subspace Clustering. The typical usage is to construct a Random Geometry Graph first and then apply spectral method to the graph to obtain clustering result. The latter step has been coined the name Spectral Clustering. As far as we know, in spite of the significance of both steps in spectral-method-based Subspace Clustering, all existing theoretical results focus on the first step of constructing the graph, but ignore the final step to correct false connections through spectral clustering. This paper establishes a theory to show the power of this method for the first time, in which we demonstrate the mechanism of spectral clustering by analyzing a simplified algorithm under the widely used semi-random model. Based on this theory, we prove the efficiency of Subspace Clustering in fairly broad conditions. The insights and analysis techniques developed in this paper might also have implications for other random graph problems.

1. Introduction

1.1. Motivation

Union of Subspaces (UoS) model serves as an important model in statistical machine learning. Briefly speaking, UoS models those high-dimensional data, encountered in many real-world problems, which lie close to low-dimensional subspaces corresponding to several classes to which the data belong, such as hand-written digits (Hastie & Simard, 1998), face images (Basri & Jacobs, 2003), DNA microarray data (Parvaresh et al., 2008), and hyper-spectral images (Chen et al., 2011), to name just a few. A fundamental task in processing data points in UoS is to cluster these data points, which is known as Subspace Clustering (SC). Applications of SC has spanned all over science and engineering, including motion segmentation (Costeira & Kanade, 1998; Kanatani, 2001), face recognition (Wright et al., 2008), and classification of diseases (McWilliams & Montana, 2014) and so on. We refer the reader to the tutorial paper (Vidal, 2011) for a review of the development of SC.

Considering the wide applications of SC, numerous algorithms have been developed for subspace clustering (Tipping & Bishop, 1999; Tseng, 2000; Vidal et al., 2005; Yan & Pollefeys, 2006; Elhamifar & Vidal, 2009; Peng et al., 2018; Meng et al., 2018), together with computation-efficient scheme (Li & Gu, 2017; Li et al., 2020; Xu et al., 2020). Arguably, spectral method is the most popular and efficient method for solving SC, which obtains the clustering result by applying the spectral clustering (Ng et al., 2002; Von Luxburg, 2007) on the constructed random graph (or the adjacent matrix equivalently), named as Union of Subspaces-based Random Geometry Graph (UoS-RGG), depending on the relative position among data points, referring to Sparse Subspace Clustering (SSC) and its variants (Elhamifar & Vidal, 2009; Liu et al., 2012; Dyer et al., 2013; Heckel & Bölcskei, 2015; Chen et al., 2017).

In spite of the spectral method that practically works well for many applications, theoretical analysis is lacked for the performance of clustering results. The fundamental difficulty in the analysis of spectral method for SC may be the change of view required in treating UoS-RGG (or general Random Geometry Graph, RGG), which has non-independent edges, in contrast with the traditional approach to analyzing clustering algorithms via Stochastic Block Model (SBM) which assumes independent edges. Section 1.2 offers a detailed discussion of this difficulty, as well as a survey of the existing attempts in theoretical aspects. We therefore propose the critical question that this paper aims to explore:

- Why does spectral method work for RGG or UoS-RGG?

This paper focuses on the analysis on the spectral method for UoS-RGG. We consider a naive and prototypical SC algorithm (Algorithm 1) here, and prove this algorithm,

[1] The authors are with the Beijing National Research Center for Information Science and Technology (BNRist) and the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. Correspondence to: Yuantao Gu <gyt@tsinghua.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

though oversimplified, can still deliver an almost correct clustering result even when the subspaces are quite close to each other and when the number of samples is far less than the subspace dimension (see Theorem 1). To the best of our knowledge, this is the first ever theory established to demonstrate that spectral clustering method can efficiently correct false connections between clusters for SC. It not only constitutes the first theoretical guarantee for accuracy of spectral method for subspace clustering, but also provides the interesting insight that the widely-conjectured oversampling requirement for subspace clustering is redundant, and that subspace clustering is quite robust in existence of closely aligned subspaces. We also verify our results by numerical experiments in Section 4. Although our theoretical results is proved only for the simplified algorithm we choose, it should be quite convincing that more carefully-designed random graph for SC would give even better performance than what we guarantee here, and our proof could serve as a prototype to the analysis of these algorithms.

1.2. Related Works and Challenges

We now briefly review the literature on the adjacent matrix and spectral method and discuss their shortcomings. Since this paper mainly deals with theory, we shall focus on theoretical aspects of existing results.

1.2.1. Analysis of Random Graphs for UoS

To analyze the random graphs associated to UoS model in an abstract setting without referring to any specific algorithms, most researches focus on the Subspace Detection Property (Soltanolkotabi et al., 2012; Liu et al., 2012; Soltanolkotabi et al., 2014; Heckel & Bölcskei, 2015, SDP,), a property which indicates that there are no edge connections between the data points in different subspaces, but are many connections between the data points in the same subspace. Under some technical conditions on the parameters of SC, the random graphs constructed by a variety of SC algorithms have been proved to enjoy SDP. Readers may consult Section 3 in Soltanolkotabi et al. (2014) for details.

There are, however, two main deficiencies of SDP which render SDP hard to use in further analysis. The first one is that SDP does not imply a correct clustering result for spectral method. Actually, one can easily construct a counterexample where SDP holds but the clustering result is unsatisfying. The second one is that SDP requires too restrictive conditions on affinity between subspaces and sampling rate to hold. These conditions are provably unnecessary, as will be demonstrated in Section 3 of this paper.

1.2.2. Analysis of Spectral Method for Random Graphs

Compared with SDP, a more concrete approach to analyze SC algorithm is to investigate the performance of spectral method on random graphs associated to UoS model. To this end, analysis of spectral method for general random graphs (not necessarily associated to UoS model) is relevant. Note that the spectral method is explored deeply in the literature of community detection, which is an important problem in statistics, computer vision, and image processing (Abbe, 2017). Stochastic Block Model (SBM) is a widely used theoretical model in this field, which we briefly introduce as follows. For simplicity, we consider the two-block case, where the vertices of random graph are divided into two "blocks", i.e. sets of vertices that ought to be closely-related, each of size of $N/2$. Then each edge of random graph is independently generated from the following distribution: for $p > q > 0$, vertices $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are connected with probability p if $\boldsymbol{x}_i, \boldsymbol{x}_j$ belong to the same block, and with probability q if they belong to different blocks. Given an instance of this graph, we would like to identify the two blocks. Recently, a series of theoretical works are devoted to analyze the performance of spectral method on this problem in different settings (Coja-Oghlan, 2010; Vu, 2014; Chin et al., 2015; Abbe et al., 2017), and extensions (Sankararaman & Baccelli, 2018). We refer the reader to (Chen et al., 2020) for a summary of analysis of spectral method.

As far as we know, all existing results make essential use of the independence of different edges, which is unfortunately not the case in SC algorithms. In fact, it is a generic and natural phenomenon in RGG that when $\boldsymbol{x}_i, \boldsymbol{x}_j$ and $\boldsymbol{x}_i, \boldsymbol{x}_k$ are connected, the probability that $\boldsymbol{x}_j, \boldsymbol{x}_k$ are connected will be higher, hence the independence assumption does not hold for RGG.

With this fundamental gap in mind, it is crucial to develop a theory for RGG to provide a rigorous theoretical guarantee for SC algorithms.

2. Preliminaries and Problem Formulation

The generative model for data points in UoS we adapt in this paper is the semi-random model introduced in Soltanolkotabi et al. (2012), which assumes that the subspaces are fixed with points distributed uniformly at random on each subspace. This is arguably the simplest model providing a good starting point for a theoretical investigation. We begin with the two-cluster case to demonstrate the mechanism of spectral method. These assumptions are made only to simplify the exposition and are by no means crucial to the analysis. In fact, the above simplest case already captures the essential point of the problem. We shall show momentarily that by investigating this simple case we will develop all

Algorithm 1 Thresholding inner-product subspace clustering (TIP-SC)

```
Require: Normalized data set $\{\boldsymbol{x}_i\}_{1\le i\le N}$, threshold $\tau$.
Construct Adjacent Matrix $\boldsymbol{A}$:
    $A_{ij} = 1$ if $i \neq j, |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle| \ge \tau$, or $A_{ij} = 0$ otherwise.
Apply Spectral Method on $\boldsymbol{A}$:
    Calculate $\boldsymbol{W}$, the eigenspace corresponding to the top two eigenvalues of $\boldsymbol{A}$.
    Use $\mathrm{sgn}(\boldsymbol{w})$ as clustering result, where $\boldsymbol{w}$ is the vector in $\boldsymbol{W}$ perpendicular to the projection of all-ones vector in $\boldsymbol{W}$.
```

the ideas and techniques required for handling the general case. Without much effort, the results obtained here can be generalized to more subspaces with different dimensions and samples, which will be discussed briefly in Section 6.

Specifically, assume that the data consists of two clusters, corresponding to two fixed subspaces S_1, S_2 in $\mathbb{R}^n$, each with $N/2$ data points uniformly sampled from the unit spheres $\mathcal{S}_1^{d-1}$ and $\mathcal{S}_2^{d-1}$ respectively in S_1 and S_2 Here d is the subspace dimension and n is the ambient dimension. The goal of SC is to cluster the normalized data points $\{\boldsymbol{x}_i\}_{1\le i\le N}$.

Given the general description of SC, we turn our attention to a simple prototypical SC algorithm detailed in Algorithm 1, which we call Thresholding Inner-Product Subspace Clustering (TIP-SC). Considering that the angle between the data points in the same subspaces would be smaller statistically, we construct for some threshold $\tau \in (0, 1)$ the random graph by computing its adjacent matrix $\boldsymbol{A}$, where $A_{ij} = 1$ if $i \neq j, |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle| \ge \tau$, and $A_{ij} = 0$ otherwise. The TIP-SC algorithm concludes with applying the spectral clustering method on $\boldsymbol{A}$. The construction method of the adjacent matrix $\boldsymbol{A}$ is very close to TSC proposed in Heckel & Bölcskei (2015), and hence is ought to enjoy similar performance. We refer to Heckel & Bölcskei (2015) for a thorough evaluation of simulated and practical performances of this kind of construction method.

The main task of this paper is to prove this simple algorithm can achieve a high clustering accuracy under fairly general condition, which will be done in the next section.

Notations. Let $\boldsymbol{U}_1, \boldsymbol{U}_2$ denote the orthonormal bases for the subspaces S_1, S_2, respectively, and $\lambda_1 \ge \ldots \ge \lambda_d \ge 0$ denote the singular values of $\boldsymbol{U}_1^\top \boldsymbol{U}_2$. We also use S and S' to denote the subspaces to which $\boldsymbol{x}_i$ does and doesn't belong, respectively. Then $\boldsymbol{x}_i = \boldsymbol{U}\overline{\boldsymbol{a}}_i$ where $\boldsymbol{U}$ denotes the orthonormal bases for S, $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I}_d) \in \mathbb{R}^d$, and $\overline{\boldsymbol{a}}_i = \boldsymbol{a}_i/\|\boldsymbol{a}_i\|$ denotes its normalization. We use p, q to represent the probability that $A_{ij} = 1$ for $j \neq i, \boldsymbol{x}_j \in S$ and $\boldsymbol{x}_j \in S'$, respectively. Conditioned on $\boldsymbol{x}_i$, let p_i denote the probability of $A_{ij} = 1$ for $j \neq i, \boldsymbol{x}_j \in S$, and q_i denote the probability of $A_{ij} = 1$ for $j, \boldsymbol{x}_j \in S'$. Denote

$$\mathrm{aff} := \sqrt{\frac{\sum_i \lambda_i^2}{d}},$$

$$\kappa := 1 - \mathrm{aff}^2,$$

$$\rho := \frac{N}{2d}.$$

Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N$ with $u_i = \frac{1}{\sqrt{N}}$, and $v_i = \frac{1}{\sqrt{N}}$, if $\boldsymbol{x}_i \in S_1$, and $v_i = -\frac{1}{\sqrt{N}}$, if $\boldsymbol{x}_i \in S_2$, then $\boldsymbol{v}$ is the ground truth. $\boldsymbol{W}$ denotes the eigenspace corresponding to the top two eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{w}$ denotes the vector in $\boldsymbol{W}$, which is perpendicular to the projection of $\boldsymbol{u}$ in $\boldsymbol{W}$.

Further, $a_N = O(b_N)$ and $a_N \lesssim b_N$ mean that $|a_N/b_N| < c$ for some constant $c > 0$; $a_N = \Omega(b_N)$ and $a_N \gtrsim b_N$ mean that $|a_N/b_N| > c$ for some constant $c > 0$; $a_N \sim b_N$ means that $c_1 < |a_N/b_N| < c_2$ for some constant $c_1, c_2 > 0$.

3. Error Rate of TIP-SC Algorithm

This section presents our main theoretical results concerning the performance of TIP-SC. By the perturbation analysis of $\boldsymbol{A}$ from $\mathbb{E}\boldsymbol{A}$, the success of spectral method for SBM has been proved in various statistical assumptions. However, such analysis is insufficient to establish our result, since for UoS-RGG, the independence condition doesn't hold, which is the crux leading to the failure of the existing methods for analyzing spectral method on random graph. As a substitute, we discover the conditional independence property for $\boldsymbol{A}$, based on which we prove that the clustering result of TIP-SC is almost correct under some mild condition on affinity and sampling rate, which is explained in the following theorem through the clustering error rate, the proportion of the number of data points which are erroneously labelled.

Theorem 1. Choosing $\tau = O\left(\frac{1}{\sqrt{d}}\right)$ such that $p = O(1)$, there exists some numerical constant $c > 0$, such that whenever $\kappa > c\frac{1}{\sqrt[4]{d}}$, the clustering error rate of TIP-SC is less than $O\left(\frac{1}{\kappa^2}(\frac{\log N}{N} + \frac{\log d}{d})\right)$ with probability at least $1 - O(\frac{1}{N^{10}})$.

Parameter selection is often critical for the success of algorithms. The above result suggests that a dense graph ($p = O(1)$) is usually a good choice, which is quite different with SDP.

In this regime, the above result indicates that the algorithm works correctly in fairly broad conditions compared with existing analysis for SC. A fascinating insight revealed by the above theorem is that even when the number of samples

$N \ll d$, we can succeed to cluster the data set, which demonstrates the commonly accepted opinion that $\rho > 1$ is necessary for SC is partially inaccurate.

To clarify the condition on κ, namely on affinity, assume these two subspaces overlap in a smaller subspace of dimension s, but are orthogonal to each other in the remaining directions. In this case, the affinity between the two subspaces is equal to $\sqrt{s/d}$. Our assumption on κ indicates that subspaces can have intersections of almost all dimensions, i.e., $s = (1-o(1))d$. In contrast, previous works (Soltanolkotabi et al., 2012; 2014) imposes that the overlapping dimension should obey $s = o(1)d$, so that the subspaces are practically orthogonal to each other. In addition, under a slightly stronger condition, i.e., $\mathrm{aff} = O(1)$, we can prove the clustering error rate of TIP-SC can be exponentially small, i.e., $O(e^{-d})$, which is stated in Theorem 3.

In the noisy case, we assume each data point is of the form

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{x}$ denotes the clean data used in the above theorem, and $\boldsymbol{z} \sim \mathcal{N}(0, \frac{\sigma^2}{n}\boldsymbol{I})$ is an independent stochastic noise term. We have the following robustness guarantee for TIP-SC.

Theorem 2. Choosing $\tau = O\left(\frac{1}{\sqrt{d}}\right)$ such that $p = O(1)$, there exists some numerical constant $c, \sigma^* > 0$, such that whenever $\kappa > c\frac{1}{\sqrt[4]{d}}$ and $\sigma < \sigma^*$, the clustering error rate of TIP-SC is less than $O\left(\frac{(1+\sigma^2 d/n)^2}{\kappa^2}(\frac{\log N}{N} + \frac{\log d}{d})\right)$ with probability at least $1 - O(\frac{1}{N^{10}})$.

The proof is similar to that of Theorem 1, and both are deferred to Section 5.

4. Numerical Experiments

In this section, we perform numerical experiments validating our main results. We evaluate the algorithm and theoretical results based on the clustering accuracy. The impacts of κ, ρ, p, q on the clustering accuracy are demonstrated. Besides, we also show the efficiency of TIP-SC in the presence of noise.

According to the definition of semi-random model, to save computation and for simplicity, the data are generated by the following steps.

1) Given $d \ll n$ and $\mathrm{aff} = \sqrt{s/d}$, define $\boldsymbol{e}_i \in \mathbb{R}^n$, whose entries are zero but the i-th entry is one. Let $\boldsymbol{U}_1 = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d]$ be the orthonormal basis for subspace for S_1, and $\boldsymbol{U}_2 = [\boldsymbol{e}_{d-s+1}, \boldsymbol{e}_{d-s+2}, \ldots, \boldsymbol{e}_{2d-s}]$ be the orthonormal basis for subspace for S_2, such that the affinity between S_1 and S_2 is $\sqrt{s/d}$.
2) Given $N = \rho d$, generate N vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N \in \mathbb{R}^d$ independently from $\mathcal{N}(0, \frac{1}{d}\boldsymbol{I})$. Let $\boldsymbol{x}_i = \boldsymbol{U}_1\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}$ for $1 \le i \le N/2$ and $\boldsymbol{x}_i = \boldsymbol{U}_2\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}$ for $N/2+1 \le i \le N$.
3) In the presence of noise, given $\sigma > 0$, generate N random noise terms $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N \in \mathbb{R}^n$ independently from $\mathcal{N}(0, \frac{\sigma^2}{n}\boldsymbol{I})$. Let the normalized data of $\boldsymbol{x}_i + \boldsymbol{z}_i$ be the input of Algorithm 1.

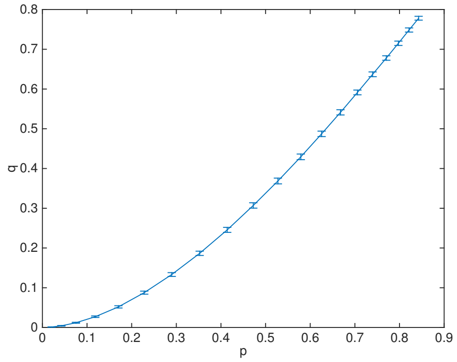

Figure 1. The relation between p and q, when $d = 100, n = 5000, \kappa = 1 - \sqrt{1/2}\,(s = d/2), \rho = 1$.

Since there are too many factors we need to consider, we always observe the relation between two concerned quantities, while keep others being some predefined typical values, i.e., $d^* = 100, n^* = 5000, \kappa^* = 1-\sqrt{1/2}\,(s^* = d/2), \rho^* = 1$, and τ is chosen to be τ^* such that the connection rate $\frac{p+q}{2} = 0.2$. We conduct the experiments in noiseless situations, except the last one which tests the robustness of Algorithm 1. Moreover, the curves are plotted by 100 trials in each experiment, while the mean and the standard deviation are represented by line and error bar, respectively. We can find that the randomness is eliminated in all experiments when the error rate is small.

It is obvious that p will decrease simultaneously if q decreases by increasing τ, which is also demonstrated in Figure 1. Combining the result of the second experiment (c.f. Figure 2), we can find that it is better to make p, q both large than to choose $q = 0$, although $q = 0$ is suggested by SDP, which is consistent with our result, while shows that SDP is somewhat inadequate for SC.

In the third and fourth experiments, we inspect the impacts of affinity and sampling rate on the performance of TIP-SC. From Figure 3 and Figure 4, the claim that SC works well in fairly broad conditions is verified. In addition, according to (1), we have

$$\mathrm{SNR} = 10 \log \frac{1}{\sigma^2},$$

then the last experiment (c.f. Figure 5) shows that the algorithm is robust even though SNR is low.

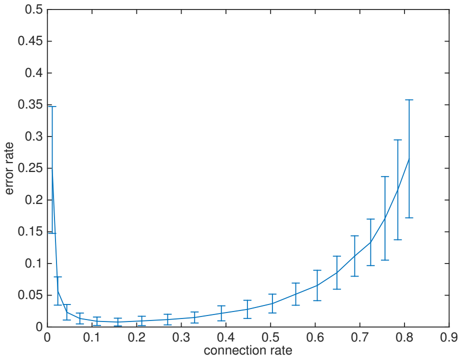

Figure 2. This figure demonstrates the clustering error rate versus the connection rate ($\frac{p+q}{2}$) in a general interval, when $d = 100, n = 5000, \kappa = 1 - \sqrt{1/2}(s = d/2), \rho = 1$.

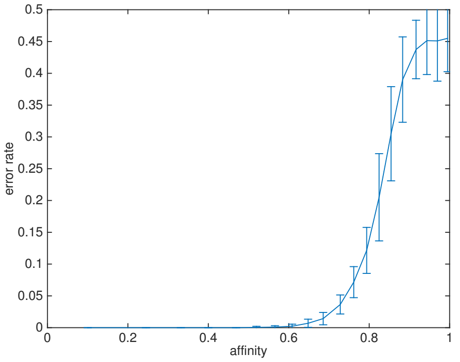

Figure 3. This figure demonstrates the clustering error rate versus the affinity in a general interval, when $d = 100, n = 5000, \rho = 1, \frac{p+q}{2} = 0.2$.

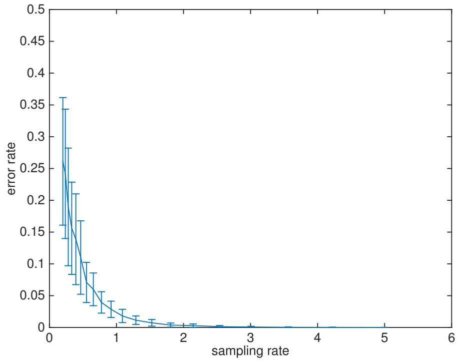

Figure 4. This figure demonstrates the clustering error rate versus the sampling rate ρ in a general interval, when $d = 100, n = 5000, \kappa = 1 - \sqrt{1/2}\ (s = d/2), \frac{p+q}{2} = 0.2$.

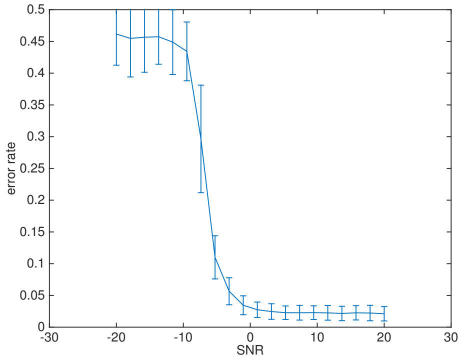

Figure 5. This figure demonstrates the clustering error rate versus the SNR in a general interval, when $d = 100, n = 5000, \kappa = 1 - \sqrt{1/2}\ (s = d/2), \rho = 1, \frac{p+q}{2} = 0.2$.

5. Proof of Main Results

5.1. Proof of Theorem 1

Recall the definition of $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{W}$ in Section 2, and notice that analyzing the error rate, denoted by γ, is equivalent to studying the difference between $\boldsymbol{w}$ and $\boldsymbol{v}$. Without loss of generality we may assume that $\langle \boldsymbol{w}, \boldsymbol{v} \rangle > 0$, thus the error rate is exactly

$$\gamma = \frac{1}{4} \left\| \frac{1}{\sqrt{N}} \operatorname{sgn}(\boldsymbol{w}) - \boldsymbol{v} \right\|_2^2 .$$

To estimate γ, it suffices to bound the distance between $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{W}$.

By simple geometric consideration, we have

$$\begin{aligned}
&\left\| \frac{1}{\sqrt{N}} \operatorname{sgn}(\boldsymbol{w}) - \boldsymbol{v} \right\|_2 \\
&\le 2 \| \boldsymbol{P}_{\boldsymbol{w}} \boldsymbol{v} - \boldsymbol{v} \|_2 \\
&\le 2 (\| \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{v} - \boldsymbol{v} \|_2 + \| \boldsymbol{P}_{\boldsymbol{w}} \boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{v} \|_2) \\
&= 2 (\| \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{v} - \boldsymbol{v} \|_2 + | \langle \overline{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}}, \boldsymbol{v} \rangle |) \\
&\le 2 (\| \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{v} - \boldsymbol{v} \|_2 + \| \overline{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}} - \boldsymbol{u} \|_2) \\
&\le 2 \| \boldsymbol{v} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{v} \|_2 + 4 \| \boldsymbol{u} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u} \|_2 ,
\end{aligned}$$

where $\overline{\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}}$ denote the normalization of $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{u}$. Moreover, for any $\lambda, \boldsymbol{x}$, we have

$$\| \boldsymbol{A} \boldsymbol{x} - \lambda \boldsymbol{x} \|_2 \ge (\lambda - \lambda_3(\boldsymbol{A})) \| \boldsymbol{x} - \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{x} \|_2 ,$$

where $\lambda_3(\boldsymbol{A})$ denotes the third largest eigenvalue of $\boldsymbol{A}$.

Summing up, for $\lambda_1, \lambda_2 > \lambda_3(\boldsymbol{A})$,

$$\gamma = \frac{1}{4}\left\|\frac{1}{\sqrt{N}}\mathrm{sgn}(\boldsymbol{w}) - \boldsymbol{v}\right\|_2^2 \lesssim \frac{\|\boldsymbol{A}\boldsymbol{u} - \lambda_1\boldsymbol{u}\|_2^2}{(\lambda_1 - \lambda_3(\boldsymbol{A}))^2} + \frac{\|\boldsymbol{A}\boldsymbol{v} - \lambda_2\boldsymbol{v}\|_2^2}{(\lambda_2 - \lambda_3(\boldsymbol{A}))^2},$$

Considering that $\mathbb{E}\langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{u}\rangle = p(N/2-1) + qN/2$, we expect $\lambda_1 = p(N/2-1)+qN/2$ is a good choice. Similarly, choose $\lambda_2 = p(N/2-1) - qN/2$.

From above discussion, to estimate γ we need to:

- Prove $\|\boldsymbol{A}\boldsymbol{u} - \lambda_1\boldsymbol{u}\|_2$ and $\|\boldsymbol{A}\boldsymbol{v} - \lambda_2\boldsymbol{v}\|_2$ are sufficiently small (see Lemma 1 and Lemma 2).
- Prove $\lambda_1 - \lambda_3(\boldsymbol{A})$ and $\lambda_2 - \lambda_3(\boldsymbol{A})$ are sufficiently large, which is equivalent to showing $p - q$ is large enough (see Lemma 1) and $\lambda_3(\boldsymbol{A})$ is small enough (see Lemma 3).

Before proceeding, we analyze the adjacent matrix $\boldsymbol{A}$ based on the conditional independence property, and provide probability estimations used in the proof of Theorem 1. Specifically, this refers to if conditioned on $\boldsymbol{x}_i, i \in \mathcal{S}$ for some subset $\mathcal{S}$ of $[N]$, A_{ij}, for $j \in \mathcal{S}^c$, are functions of $\boldsymbol{x}_j$, respectively, and then are independent from each other.

Moreover, recalling the definition of $\boldsymbol{x}_i, \boldsymbol{a}_i$, if conditioned on $\boldsymbol{x}_i, i \in \mathcal{S}$, A_{ij}, for $j \in \mathcal{S}^c$ are *nearly identically distributed*, and for some $j \in \mathcal{S}^c$, A_{ij}, for $i \in \mathcal{S}$ are *nearly independent* from each other, which will be explained and employed many times in the following analysis.

With above intuitions, we will provide some key lemmas for the analysis of spectral method.

Lemma 1. Choose $\tau_d = O(1)$, then $p = \Omega(1)$. Moreover, there exists some constant $c > 0$, such that if $\kappa = 1 - \mathrm{aff}^2 > c\sqrt{\frac{\log d}{d}}$,

$$p - q \gtrsim \kappa,$$

and

$$\frac{1}{N}\sum_i (q_i - q)^2 \lesssim \frac{\log d}{d}.$$

Proof. The proof can be found in Li & Gu (2019). □

Having finished the calculation about the probability of each entry, we now turn to the overall properties of $\boldsymbol{A}$.

Lemma 2. Conditioned on $\boldsymbol{x}_i$, for any $t > 0$

$$\mathbb{P}\left(\left|\frac{1}{N/2-1}\sum_{j:\boldsymbol{x}_j \in S} A_{ij} - p\right| > t\right) < \mathrm{e}^{-\frac{t^2(N/2-1)}{p+\frac{1}{3}t}},$$

and

$$\mathbb{P}\left(\left|\frac{1}{N/2}\sum_{j:\boldsymbol{x}_j \in S'} A_{ij} - q_i\right| > t\right) < \mathrm{e}^{-\frac{t^2 N/2}{q_i+\frac{1}{3}t}}.$$

Proof. The proof can be found in Li & Gu (2019). □

In the next lemma, we will analyze the eigenvalue of $\boldsymbol{A}$.

Lemma 3. With probability at least $1 - \frac{1}{N^{10}}$,

$$\lambda_3(\boldsymbol{A}) < c\sqrt{Np\log N + \frac{N^2p^2}{\sqrt{d}}},$$

where $\lambda_3(\boldsymbol{A})$ denotes the third largest eigenvalue of $\boldsymbol{A}$.

Proof. The proof can be found in Li & Gu (2019). □

Now, we have all the ingredients for the proof of Theorem 1.

Proof of Theorem 1. We begin with some inequalities for estimating the error. We have

$$\begin{aligned}
&\|\boldsymbol{A}\boldsymbol{u} - (\frac{(N-2)p}{2} + \frac{Nq}{2})\boldsymbol{u}\|_2^2 \\
&= \frac{1}{N}\sum_i (\sum_{j:\boldsymbol{x}_j \in S} A_{ij} - \frac{(N-2)p}{2} + \sum_{j:\boldsymbol{x}_j \in S'} A_{ij} - \frac{Nq}{2})^2 \\
&\le \frac{3}{N}\sum_i (\sum_{j:\boldsymbol{x}_j \in S} A_{ij} - \frac{(N-2)p}{2})^2 \\
&+ \frac{3}{N}\sum_i (\sum_{j:\boldsymbol{x}_j \in S'} A_{ij} - \frac{Nq_i}{2})^2 + \frac{3}{N}\sum_i (\frac{Nq_i}{2} - \frac{Nq}{2})^2.
\end{aligned}$$

According to Lemma 2, for all $1 \le i \le N$, we have, with probability at least $1 - \frac{1}{N^{10}}$,

$$\left(\sum_{j:\boldsymbol{x}_j \in S} A_{ij} - p(N/2-1)\right)^2 \lesssim N\log N,$$

and

$$\left(\sum_{j:\boldsymbol{x}_j \in S'} A_{ij} - q_i N/2\right)^2 \lesssim N\log N.$$

On the other hand, Lemma 1 gives,

$$\frac{3}{N}\sum_i (q_i N/2 - qN/2)^2 \lesssim \rho N\log d.$$

Summing up, we have, with probability at least $1 - \frac{1}{N^{10}}$,

$$\|\boldsymbol{A}\boldsymbol{u} - (p(N/2-1) + qN/2)\boldsymbol{u}\|_2^2 \lesssim N(\log N + \rho\log d).$$

Similarly, with probability at least $1 - \frac{1}{N^{10}}$,

$$\|\boldsymbol{A}\boldsymbol{v} - (p(N/2-1) - qN/2)\boldsymbol{v}\|_2^2 \lesssim N(\log N + \rho\log d).$$

According to Lemma 3, with probability at least $1 - \frac{1}{N^{10}}$, the third largest eigenvalue of $\boldsymbol{A}$ satisfies

$$\lambda_3(\boldsymbol{A}) \lesssim \sqrt{Np\log N + \frac{N^2p^2}{\sqrt{d}}} = O\left(\frac{N}{\sqrt[4]{d}}\right).$$

With these estimations at hand, recall

$$\gamma \lesssim \frac{\|\boldsymbol{A}\boldsymbol{u} - (p(N/2-1) + qN/2)\boldsymbol{u}\|_2^2}{|p(N/2-1) + qN/2 - \lambda_3(\boldsymbol{A})|^2} + \frac{\|\boldsymbol{A}\boldsymbol{v} - (p(N/2-1) - qN/2)\boldsymbol{v}\|_2^2}{|p(N/2-1) - qN/2 - \lambda_3(\boldsymbol{A})|^2}.$$

Lemma 1 gives $p \pm q \gtrsim 1 - \text{aff}^2$, then we have

$$\gamma \lesssim \left(\frac{\sqrt{N(\log N + \rho\log d)}}{N\left(1 - \text{aff}^2 - \frac{1}{\sqrt[4]{d}}\right)}\right)^2 \sim \frac{\log N + \rho \log d}{\kappa^2 N}.$$

We conclude the proof. □

5.2. Proof of Theorem 2

Robustness analysis can be completed by following the similar analysis method. We provide the differences in the analysis of noise, while omit the details.

Here, we only need to pay attention to the changes of Lemma 1, Lemma 2, and Lemma 3, when adding noise. Notice that the noise terms do not destroy the wonderful conditional independence property, then it's obvious that except the estimation for $p - q$, all other bounds still hold in a similar way. Through simple calculation, the contribution of noise has the form

$$p - q \gtrsim \frac{\kappa}{1 + \sigma^2 d/n}.$$

Taking this change into account, we can get the result easily.

6. Extension to Multi-cluster Case

Here, assume that the data consists of L clusters, corresponding to L fixed subspaces $S_1, S_2, \dots, S_L$ in $\mathbb{R}^n$, with dimension $d_1, d_2, \dots, d_L$ respectively. There are N_l data points uniformly sampled from the unit spheres $\mathcal{S}_1^{d_l-1}$ in S_l for $1 \le l \le L$. Then, we aim to cluster the normalized data points $\{\boldsymbol{x}_i\}_{1\le i\le N}$ for $N = N_1 + N_2 + \ldots + N_L$. In addition, TIP-SC, detailed in Algorithm 1, refers to the following procedure here. For some threshold $\tau \in (0, 1)$ the random graph by computing its adjacent matrix $\boldsymbol{A}$, where $A_{ij} = 1$ if $i \ne j, |\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle| \ge \tau$, and $A_{ij} = 0$ otherwise. Then we apply the spectral clustering method on $\boldsymbol{A}$, which is also done by two steps with slightly difference.

- Calculate $\boldsymbol{W}$, the eigenspace corresponding to the top L eigenvalues of $\boldsymbol{A}$.
- Use k-means to cluster $\{\boldsymbol{P_W}\boldsymbol{A}_i\}_{1\le i\le N}$, where $\boldsymbol{P_W}$ denote the projection matrix onto $\boldsymbol{W}$, and $\boldsymbol{A}_i$ denotes the i-th column of $\boldsymbol{A}$.

The main task of this section is to prove this algorithm is efficient for the multi-cluster case, which is completed by showing that $\|\boldsymbol{P_W}\boldsymbol{A}_i - \boldsymbol{P_W}\boldsymbol{A}_j\|_2$ is small if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are from the same subspace, and $\|\boldsymbol{P_W}\boldsymbol{A}_i - \boldsymbol{P_W}\boldsymbol{A}_j\|_2$ is large otherwise. This is stated in the following.

Before proceeding, let's introduce some notations used in this section. We use q_{ij} to represent the connection probability for two different points $\boldsymbol{x} \in S_i$ and $\boldsymbol{y} \in S_j$. Denote

$$\text{aff} := \max_{1\le i\ne j\le L} \sqrt{\frac{\sum_i \lambda_i^2}{\min\{d_i, d_j\}}},$$

For some data sets $\mathcal{X}$, let $R_{\text{inter}}(\mathcal{X})$ denote the maximal projection distance of $\|\boldsymbol{P_W}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2$ when $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ are from different subspaces, and $R_{\text{inner}}(\mathcal{X})$ denote the minimal projection distance of $\|\boldsymbol{P_W}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2$ when $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ are from the same subspace.

Theorem 3. Choosing $\tau = O\left(\frac{1}{\sqrt{\max_{1\le l\le L} d_l}}\right)$ such that $p_l = O(1)$ for all $1 \le l \le L$, there exists some numerical constant $c > 0$ depending only on $L, \max d_i/\min d_i, \max N_i/\min N_i$, such that whenever aff $< c$, with probability at least $1 - O(\frac{1}{N^{10}})$, we have $R_{\text{inter}}(\mathcal{X}) > 4R_{\text{inner}}(\mathcal{X})$ with $|\mathcal{X}| = (1 - O(e^{-d}))N$, and the TIP-SC will deliver a nearly correct clustering result. In addition, in the noisy case, the result still holds if $\sigma < \sigma^*$ for some numerical constant $\sigma^* > 0$.

In the multi-cluster case, we expose a more strict condition, i.e., aff $= O(1)$, ignoring some factor with respect to $L, \max d_i/\min d_i, \max N_i/\min N_i$. This assumption is still more generous than previous work, since there is no need to force data points in different subspaces to be disconnected. The detailed dependence on the above parameters and improvement condition for aff are beyond our scope, which we leave as future work. To avoid repetition, we will introduce the proof for the above result briefly, whose details are omitted due to the similarity with the two-cluster case.

Proof sketch. We consider the noiseless case, and the noisy case can be analyzed similarly. Let $\boldsymbol{A}^\star := \mathbb{E}\boldsymbol{A}$ whose entry is q_{ij} determined by its corresponding subspaces, and $\boldsymbol{W}^\star$ denote the eigenspace corresponding to the top L eigenvalues of $\boldsymbol{A}^\star$. The assumption on aff makes $\frac{\min_i q_{ii}}{\max_{i\ne j} q_{ij}} > C$ for some constant $C > 0$. Then the following facts hold for most data points $((1 - O(e^{-d}))N)$.

- If $\boldsymbol{x}_i, \boldsymbol{x}_j$ are from the same subspace, we have with probability at least $1 - \frac{1}{N^{10}}$,

$$\|\boldsymbol{P}_{\boldsymbol{W}^\star}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2^2 \ll N,$$

while if they are from different subspaces,

$$\|\boldsymbol{P}_{\boldsymbol{W}^\star}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2^2 \gtrsim N,$$

- Under the affinity condition, we have $\lambda_L(\boldsymbol{A}^\star) \asymp N$, and $\|\boldsymbol{A} - \boldsymbol{A}^\star\|_{\mathrm{op}} \lesssim \frac{N}{\sqrt[4]{d}}$ with probability at least $1 - \frac{1}{N^{10}}$ through the same analysis as Lemma 3. Then,

$$\begin{aligned} &\Big| \|\boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2 - \|\boldsymbol{P}_{\boldsymbol{W}^\star}(\boldsymbol{A}_i - \boldsymbol{A}_j)\|_2 \Big| \\ &\le \|\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{W}^\star}\|_{\mathrm{op}} \|\boldsymbol{A}_i - \boldsymbol{A}_j\|_2 \\ &\lesssim \frac{\|\boldsymbol{A} - \boldsymbol{A}^\star\|_{\mathrm{op}}}{\lambda_L(\boldsymbol{A}^\star) - \|\boldsymbol{A} - \boldsymbol{A}^\star\|_{\mathrm{op}}} \cdot \sqrt{N} \\ &\lesssim \sqrt{\frac{N}{\sqrt{d}}}. \end{aligned}$$

Putting everything together leads to $R_{\mathrm{inter}} > 4R_{\mathrm{inner}}$, which implies that k-means can succeed.

7. Conclusion

This paper establish a theory to analyze spectral method for Random Geometry Graph constructed by data points from Union of Subspaces. Based on this theory, we demonstrate the efficiency of Subspace Clustering in fairly broad conditions. To the best of our knowledge, the clustering accuracy of spectral method for SC has not been shown in the prior literature. The insights and analysis techniques developed in this paper might also have implications for other Random Geometry Graph.

Moving forward, one issue is to understand UoS-RGG constructed by more complex strategy, such as SSC. Additionally, ideally one would desire an exact recovery by spectral method, which needs entrywise analysis. We leave these to future investigation.

Acknowledgements

The authors are supported in part by the grant NSFC-61971266. We thank Xingyu Xu for numerous discussions about Subspace Clustering.

References

Abbe, E. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Basri, R. and Jacobs, D. W. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):218–233, 2003.

Chen, Y., Nasrabadi, N. M., and Tran, T. D. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10):3973–3985, 2011.

Chen, Y., Li, G., and Gu, Y. Active orthogonal matching pursuit for sparse subspace clustering. *IEEE Signal Processing Letters*, 25(2):164–168, 2017.

Chen, Y., Chi, Y., Fan, J., and Ma, C. Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*, 2020.

Chin, P., Rao, A., and Vu, V. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pp. 391–423, 2015.

Coja-Oghlan, A. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

Costeira, J. P. and Kanade, T. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

Dyer, E. L., Sankaranarayanan, A. C., and Baraniuk, R. G. Greedy feature selection for subspace clustering. *Journal of Machine Learning Research*, 14:2487–2517, 2013.

Elhamifar, E. and Vidal, R. Sparse subspace clustering. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2790–2797. IEEE, 2009.

Hastie, T. and Simard, P. Y. Metrics and models for handwritten character recognition. *Statistical Science*, pp. 54–65, 1998.

Heckel, R. and Bölcskei, H. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

Kanatani, K.-i. Motion segmentation by subspace separation and model selection. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 586–591. IEEE, 2001.

Li, G. and Gu, Y. Restricted isometry property of gaussian random projection for finite set of subspaces. *IEEE Transactions on Signal Processing*, 66(7):1705–1720, 2017.

Li, G. and Gu, Y. Theory of spectral method for union of subspaces-based random geometry graph. *arXiv preprint arXiv:1907.10906*, 2019.

Li, G., Liu, Q., and Gu, Y. Rigorous restricted isometry property of low-dimensional subspaces. *Applied and Computational Harmonic Analysis*, 49(2):608–635, 2020.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2012.

McWilliams, B. and Montana, G. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

Meng, L., Li, G., Yan, J., and Gu, Y. A general framework for understanding compressed subspace clustering algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1504–1519, 2018.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

Parvaresh, F., Vikalo, H., Misra, S., and Hassibi, B. Recovering sparse signals using sparse measurement matrices in compressed dna microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):275–285, 2008.

Peng, X., Feng, J., Xiao, S., Yau, W.-Y., Zhou, J. T., and Yang, S. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018.

Sankararaman, A. and Baccelli, F. Community detection on euclidean random graphs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2181–2200. SIAM, 2018.

Soltanolkotabi, M., Candes, E. J., et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Soltanolkotabi, M., Elhamifar, E., Candes, E. J., et al. Robust subspace clustering. *The Annals of Statistics*, 42(2): 669–699, 2014.

Tipping, M. E. and Bishop, C. M. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11 (2):443–482, 1999.

Tseng, P. Nearest q-flat to m points. *Journal of Optimization Theory and Applications*, 105(1):249–252, 2000.

Vidal, R. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

Vidal, R., Ma, Y., and Sastry, S. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Vu, V. A simple svd algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918*, 2014.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2008.

Xu, X., Li, G., and Gu, Y. Unraveling the veil of subspace rip through near-isometry on subspaces. *IEEE Transactions on Signal Processing*, 68:3117–3131, 2020.

Yan, J. and Pollefeys, M. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *European Conference on Computer Vision*, pp. 94–106. Springer, 2006.