Exact Recovery in the Latent Space Model

Chuyang Ke
Department of Computer Science
Purdue University
cke@purdue.edu

Jean Honorio
Department of Computer Science
Purdue University
jhonorio@purdue.edu

Abstract

We analyze the necessary and sufficient conditions for exact recovery of the symmetric Latent Space Model (LSM) with two communities. In a LSM, each node is associated with a latent vector following some probability distribution. We show that exact recovery can be achieved using a semidefinite programming approach.

1 Introduction

Community detection has been one of the most researched topics in the field of network models. With the emergence of social media in the past decade, researchers are now exposed to millions of records of interaction generated on the Internet everyday. One can note that the generic structure and organization of social media resemble certain network models, for instance, the Erdos-Renyi Model, the Stochastic Block Model (SBM) and the Latent Space Model (LSM) [Goldenberg et al., 2010, Newman et al., 2002]. The analogy comes from the fact that, in a social network each user can be modeled as a node, and the interpersonal interactions can be modeled as edges. One common assumption is that the users form several communities based on their own preferences, such as the user’s political view, music genre preferences, or whether the user is a cat or dog person. Another common assumption, often referred as homophily in prior literature, suggests that two nodes from the same community are more likely to be connected than those from different communities [Goldenberg et al., 2010, Hoff, 2008, Krivitsky et al., 2009]. The core task of community detection, also known as graph clustering, is to partition the nodes into communities based on the observed edge information [Abbe, 2018, Ke and Honorio, 2018, Fortunato, 2010].

In this paper, we are particularly interested in the Latent Space Model first proposed in Hoff et al. [2002]. In the Latent Space Model, there exists a small number of communities, and the label for each node (i.e., which community the node belongs to) is chosen randomly. After node labels are determined, each node is associated with a latent vector. It is natural to assume that for nodes from the same community, their associated latent vectors follow the same probability distribution [Ke and Honorio, 2018]. Finally, two nodes are randomly connected depending on the distance between their corresponding latent vectors. That is, nodes that are in the same community are more likely to be connected, than nodes in different communities. Note that we do not observe the node labels neither the latent vectors.

One of the most studied problems for network models is exact recovery [Abbe, 2018]. Prior works have analyzed the necessary conditions for exact recovery in various network models. For instance, Mossel et al. [2012], Abbe et al. [2016] provide information-theoretic lower bounds for exact recovery for the Stochastic Block Model. Ke and Honorio [2018] provides similar bounds for the Latent Space Model, the Preferential Attachment Model, the Small-World Model, among
others. Many previous works have also proposed efficient algorithms for exact recovery for the Stochastic Block Model; a partial list of works includes [Abbe et al. 2016, Bandeira 2018, Hajek et al. 2016, Chen and Xu 2014]. To the best of our knowledge, no previous work has provided efficient algorithms for exact recovery for the LSM. In general, problems involving latent variables are NP-hard and nonconvex, for instance, learning restricted Boltzmann machines or structural Support Vector Machines with latent variables [Long and Servedio 2010, Yu and Joachims 2009].

In this paper we address the problem of exact recovery for the Latent Space Model. More specifically, we analyze the recovery conditions of the Latent Space Model and propose a polynomial-time algorithm for the exact recovery of the true labels. Our work is motivated by the algorithms in [Abbe et al. 2016, Hajek et al. 2016, Bandeira 2018, Chen and Xu 2014]. We want to highlight that many techniques used in the analysis of the SBM from the works above, do not apply to the Latent Space Model. This is because in the LSM, edges are no longer statistically independent. As a result, our problem does not simply reduce to the SBM.

The paper is arranged as follows. In Section 2 we present the formal definition of the Latent Space Model and related notations. In Section 3 we analyze the information-theoretic lower bound for exact recovery of the symmetric Latent Space Model. Section 4 shows that under certain conditions, the true labels in the Latent Space Model can be exactly recovered with high probability using semidefinite programming (SDP) efficiently. Section 5 sums up our results from the perspective of synthetic experiments.

2 Preliminaries

First we provide the definition of the symmetric Latent Space Model with two communities.

**Definition 1** (Symmetric Latent Space Model). Let \( n \) be a positive even integer, \( \sigma > 0 \), \( d \in \mathbb{Z}^+ \) and \( \mu \in \mathbb{R}^d, \mu \neq 0 \). The pair \((y, X)\) is drawn under symmetric LSM\((n, d, \mu, \sigma)\) if \( y \) is a \( n \)-dimensional random vector uniformly drawn from \( Y = \{y : y \in \{\pm 1\}^n, \sum_i y_i = 0\} \), and \( X \in \mathbb{R}^{n \times d} \) is a random matrix such that for each \( i \in [n] \), \( x_i \) is generated from the \( d \)-dimensional Gaussian distribution \( N_d(y_i \mu, \sigma^2 I) \). A random graph \( G \) is generated from \((y, X)\) as follows. For each pair of vertices \( i, j \in [n] \), \((i, j)\) is an edge of \( G \) with probability \( \exp(-\|x_i - x_j\|^2) \).

The adjacency matrix \( A \) is an \( n \)-by-\( n \) binary matrix, such that \( A_{ij} = 1 \) if edge \((i, j)\) is in the graph \( G \), and 0 otherwise. We want to highlight that in the LSM, edges are not independent if not conditioning on the latent vectors. For example, suppose \( i, j \) and \( k \) are all from the same cluster. In the SBM edge \((i, j)\) and \((i, k)\) are independent, but this is not true in the LSM.

In later sections, we will frequently use the moment generating function of the distance between two or three latent vectors which, as we mention above, follow Gaussian distribution \( N_d(y_i \mu, \sigma^2 I) \). In particular, we are interested in the following expectation involving two variables

\[
\mathbb{E}_X \left[ \exp(-\|x_i - x_j\|^2) \right],
\]

and the following expectation involving three variables

\[
\mathbb{E}_X \left[ \exp(-\|x_i - x_k\|^2 - \|x_j - x_k\|^2) \right],
\]

when \( y_i, y_j, \) and \( y_k \) are taking different values. Note that the expectations above are related to the expected value of the Gaussian radial basis function (RBF) kernel.

We now define the following shorthand notations along with their analytic expressions, which will be used throughout the paper. Note that all of the following notations \((p, p', q, q', r, s_0, \) and \( s_1)\) are bounded in \((0, 1)\).
Lemma 1. In the case of two variables, we have

\[
p := \mathbb{E}_X \left[ \exp(-\|x_i - x_j\|^2) \mid y_i = y_j \right] = (4\sigma^2 + 1)^{-d/2},
\]

\[
p' := \mathbb{E}_X \left[ \exp(-2\|x_i - x_j\|^2) \mid y_i = y_j \right] = (8\sigma^2 + 1)^{-d/2},
\]

\[
q := \mathbb{E}_X \left[ \exp(-\|x_i - x_j\|^2) \mid y_i \neq y_j \right] / p = \exp\left(\frac{-4\|\mu\|^2}{4\sigma^2 + 1}\right),
\]

\[
q' := \mathbb{E}_X \left[ \exp(-2\|x_i - x_j\|^2) \mid y_i \neq y_j \right] / p' = \exp\left(\frac{-8\|\mu\|^2}{8\sigma^2 + 1}\right).
\]

In the case of three variables, we have

\[
r := \mathbb{E}_X \left[ \exp(-\|x_i - x_k\|^2 - \|x_j - x_k\|^2) \mid y_i = y_j = y_k \right] = ((2\sigma^2 + 1)(6\sigma^2 + 1))^{-d/2},
\]

\[
s_0 := \mathbb{E}_X \left[ \exp(-\|x_i - x_k\|^2 - \|x_j - x_k\|^2) \mid y_i = y_j \neq y_k \right] / r = \exp\left(\frac{-8\|\mu\|^2}{6\sigma^2 + 1}\right),
\]

\[
s_1 := \mathbb{E}_X \left[ \exp(-\|x_i - x_k\|^2 - \|x_j - x_k\|^2) \mid y_i = y_k \neq y_j \right] / r = \exp\left(\frac{-4(4\sigma^2 + 1)\|\mu\|^2}{12\sigma^4 + 8\sigma^2 + 1}\right).
\]

Proof sketch. The the value of \(p\) and \(q\) were given in Ke and Honorio [2018, Section A.4]. To the best of our knowledge, the results of \(p', q', r, s_0\), and \(s_1\), are novel.

The proof of \(p'\) and \(q'\) is similar to the one in Ke and Honorio [2018, Section A.4], except that we require \(t = -2\) in the moment generating function.

In the case of three variables, we introduce the \(3d \times 3d\) block matrix \(M\), where

\[
M = \begin{bmatrix}
I & 0 & -I \\
0 & I & -I \\
-I & -I & 2I
\end{bmatrix},
\]

and each entry above is a matrix of dimension \(d \times d\). Then, the expression

\[
-\|x_i - x_k\|^2 - \|x_j - x_k\|^2
\]

from (2) can be written as

\[
- \begin{bmatrix} x_i \\ x_j \\ x_k \end{bmatrix}^\top \begin{bmatrix} I & 0 & -I \\
0 & I & -I \\
-I & -I & 2I
\end{bmatrix} \begin{bmatrix} x_i \\ x_j \\ x_k \end{bmatrix},
\]

in which

\[
\begin{bmatrix} x_i \\ x_j \\ x_k \end{bmatrix} \sim N_{3d} \left( \begin{bmatrix} y_i \mu \\ y_j \mu \\ y_k \mu \end{bmatrix}, \sigma^2 I_{3d} \right).
\]

Then, by plugging the concatenate vector of \(x_i, x_j, x_k\) and the block matrix \(M\) into the Gaussian moment generating function theorem [Mathai and Provost [1992, Theorem 3.2a.1], we can solve for \(r, s_0\) and \(s_1\) for the different possibilities of the values of \((y_i, y_j, y_k)\).
3 Impossible Regime

In this section we analyze the necessary conditions for exact recovery of the symmetric Latent Space Model. Our analysis is based on Theorem 2 in [Ke and Honorio 2018], but we focus on the symmetric LSM.

Our goal is to characterize the information-theoretic limit of any algorithm in recovering the true labels $y^*$ in the symmetric LSM. More specifically, we would like to recover labels $\hat{y}$ given the observation of the adjacency matrix $A$. Also note that we do not observe the latent vectors $X$. Fano’s inequality [Cover and Thomas 2012, Yu 1997] allows us to obtain the following information-theoretic lower bound.

**Theorem 1.** In a symmetric Latent Space Model with parameters $(n,d,\mu,\sigma)$ where $n \geq 10$, if

$$-p \log q \leq \frac{3 \log 2}{10n},$$

then the probability of error $\mathbb{P}\{\hat{y} \neq y^*\}$, regardless of any algorithm that a learner could use for picking $\hat{y}$, is greater than or equal to $1/2$.

**Proof.** First we characterize the mutual information between the true labels $y^*$ and the observed adjacency matrix $A$. Following the analysis from [Ke and Honorio 2018], we have

$$I(y^*, A) \leq \max_{y,y' \neq y} D_{KL}(\mathbb{P}_{AX} \{A | y \} \| \mathbb{P}_{AX} \{A | y' \})$$

$$\leq \binom{n}{2} \max_{y,y' \neq y} D_{KL}(\mathbb{P}_{AX} \{A_{ij} | y_i, y_j \} \| \mathbb{P}_{AX} \{A_{ij} | y'_i, y'_j \})$$

$$\leq \binom{n}{2} \sum_{A_{ij}} \mathbb{P}_{AX} \{A_{ij} = 1 | y_i, y_j \} \log \frac{\mathbb{P}_{AX} \{A_{ij} = 1 | y_i, y_j \}}{\mathbb{P}_{AX} \{A_{ij} = 1 | y'_i, y'_j \}}$$

$$\leq \binom{n}{2} \mathbb{E}_{X} \left[ \exp(-\|x_i - x_j\|^2) \mid y_i = y_j \right] \cdot \log \frac{\mathbb{E}_{X} \left[ \exp(-\|x_i - x_j\|^2) \mid y_i = y_j \right]}{\mathbb{E}_{X} \left[ \exp(-\|x_i - x_j\|^2) \mid y_i \neq y_j \right]}$$

$$= - \binom{n}{2} p \log q$$

$$\leq - \frac{n^2}{2} p \log q,$$

where $D_{KL}(\cdot | \cdot)$ denotes the KL-divergence between two probability distributions. Then we can apply Fano’s inequality. For any predicted labels $\hat{y}$, we have

$$\mathbb{P}\{\hat{y} \neq y^*\} \geq 1 - \frac{I(y^*, A) + \log 2}{\log |\mathcal{Y}|}.$$ 

By definition of $\mathcal{Y}$ and counting, it follows that

$$|\mathcal{Y}| = \frac{n!}{2^\binom{n}{2}(\frac{n}{2})!}.$$
Note that $\sqrt{n}(n/e)^n \leq n! \leq e \sqrt{n}(n/e)^n$. It follows that
\[
|Y| \geq \frac{\sqrt{n}(n/e)^n}{e^2 n(n/2e)^n} = \frac{2^n}{e^2 \sqrt{n}},
\]
which indicates that
\[
\log |Y| \geq n \log 2 - \frac{\log n}{2} - 2 \geq \frac{1}{2} n \log 2,
\]
and the last inequality holds if $n \geq 10$.

Finally, by Fano’s inequality, for the probability of error to be at least $1/2$, it is sufficient to require the lower bound to be greater than $1/2$. Hence
\[
\mathbb{P}\{ \hat{y} \neq y^* \} \geq 1 - \frac{I(y^*, A) + \log 2}{\log |Y|}
\]
\[
\geq 1 - \frac{-n^2 p \log q + \log 2}{\frac{1}{2} n \log 2}
\]
\[
= 1 + np \frac{\log q}{\log 2} - \frac{2}{n}
\]
\[
\geq \frac{4}{5} + np \frac{\log q}{\log 2}
\]
\[
\geq \frac{1}{2},
\]
and the last inequality holds provided that $-p \log q \leq \frac{3 \log 2}{10n^2}$.

4 Polynomial-Time Regime with Semidefinite Programming

In this section we analyze the sufficient conditions for exact recovery of the symmetric Latent Space Model. Inspired by Bandeira [2018], Abbe et al. [2016], Hajek et al. [2016] we use a semidefinite programming (SDP) approach to reconstruct the two clusters.

Given the adjacency matrix $A$, we define the degree matrix $D$ to be the diagonal matrix $\text{diag}(d_1, \ldots, d_n)$, where each $d_i := \sum_{j \neq i} A_{ij} y_i y_j = \sum_{j \neq i, y_j = y_i} A_{ij} = \sum_{k \neq i, y_k \neq y_i} A_{ik}$.

The maximum likelihood estimator (MLE) for $y$ has the following form
\[
\max_y \quad y^T (2A - 11^T + I)y
\]
\[
s.t. \quad y \in \{+1, -1\}^n. \tag{4}
\]

We define $Y = yy^T$. Since $y_i$ is either $+1$ or $-1$, we have $Y_{ii} = 1$. Also note that $Y$ is a rank-1 positive semidefinite matrix. Thus (4) is equivalent to the following problem in matrix form
\[
\max_Y \quad \text{Tr} ((2A - 11^T + I)Y)
\]
\[
s.t. \quad Y_{ii} = 1
\]
\[
Y \succeq 0
\]
\[
\text{rank}(Y) = 1. \tag{5}
\]
Dropping the non-convex rank constraint from \(5\), we consider the following SDP relaxation

\[
\max_Y \quad \text{Tr} ((2A - 11^\top + I)Y) \\
\text{s.t.} \quad Y_{ii} = 1 \\
Y \succeq 0,
\]

which could be solved efficiently in polynomial-time \cite{Vandenberghe and Boyd 1996}. The dual problem \cite{Boyd and Vandenberghe 2004} of (6) is

\[
\min_v \quad 1^\top v \\
\text{s.t.} \quad \operatorname{diag}(v) \succeq 2A - 11^\top + I,
\]

which is equivalent to the following problem in the matrix form

\[
\min_M \quad \text{Tr} (M) \\
\text{s.t.} \quad M \text{ is diagonal} \\
M \succeq 2A - 11^\top + I.
\]

Note that, if (6) has a unique optimal solution \(Y = yy^\top\), then \(y\) is the optimal solution to (4). We want to show that \(yy^\top\) is the unique optimal solution to (6) for some \(y\).

To satisfy the optimality condition, it is sufficient to find a dual certificate \(M\) satisfying the constraints in (8). Optimality also requires the duality gap to be zero. In other words,

\[
\text{Tr} ((2A - 11^\top + I)yy^\top) = \text{Tr} (M).
\]

Observe that \(\text{Tr} ((2A - 11^\top + I)yy^\top) = \text{Tr} (2D + I)\) \cite{Abbe et al. 2016}. Thus, by setting \(M = 2D + I\), the duality gap is closed.

We want to highlight that by construction, \(M\) needs to satisfy the two constraints in (8). It is trivial to verify that \(M = 2D + I\) is diagonal. Feasibility of the solution requires that \(M = 2D + I \succeq 2A - 11^\top + I\), which is equivalent to

\[
2D - 2A + 11^\top \succeq 0.
\]

Furthermore, to satisfy the uniqueness condition, it is sufficient to ensure that

\[
\lambda_2(2D - 2A + 11^\top) > 0.
\]

This gives us the following result.

**Lemma 2.** If (9), (10), and (11) hold, then \(Y = yy^\top\) is the unique optimal solution to (6), and semidefinite programming achieves exact recovery.

Note that \(y\) is an eigenvector of the random matrix \(2D - 2A + 11^\top\), i.e.,

\[
\left(2D - 2A + 11^\top\right)y = 0.
\]

This enables us to drop the constraint of (10), because if one can prove that

\[
u^\top \left(2D - 2A + 11^\top\right)u > 0
\]
for any unit vector $u$ orthogonal to $y$, then the minimum eigenvalue of $(2D - 2A + 11^\top)$ is zero. Thus the matrix is automatically positive semidefinite.

The next goal is to prove that (11) holds with high probability. We want to highlight that the Latent Space Model does not reduce to the Stochastic Block Model. For example, in the SBM, $A_{ij}$ and $A_{ik}$ are independent if the underlying labels of node $i,j$ and $k$ are the same. This is not true for the LSM in general; without conditioning on the latent vectors $X$, $A_{ij}$ and $A_{ik}$ are not independent in the LSM.

Our proof relies on the use of matrix concentration inequalities. For the expectation we prove the following lemma.

**Lemma 3.** Assuming $p(1 + q) \leq 1$, we have

$$\lambda_2(2\mathbb{E}_{AX}[D] - 2\mathbb{E}_{AX}[A] + 11^\top) = np(1 - q).$$  \hspace{1cm} (12)

**Proof.** By Lemma 1, note that $\mathbb{E}_{AX}[A_{ij}]$ is equal to $p$ if $i \neq j, y_i = y_j$, equal to $pq$ if $y_i \neq y_j$, and equal to zero for diagonal entries. Thus we can rewrite the expectation of $A$ in the matrix form

$$\mathbb{E}_{AX}[A] = \frac{1}{2}p(1 + q)11^\top + \frac{1}{2}p(1 - q)yy^\top - pI.$$

By definition of $D$, it follows that $\mathbb{E}_{AX}[D_{ii}] = \frac{1}{2}np(1 - q) - p$ for every diagonal entry, and $\mathbb{E}_{AX}[D_{ij}] = 0$ for every $i \neq j$. Hence the expectation of $D$ can be written as

$$\mathbb{E}_{AX}[D] = \left(\frac{1}{2}np(1 - q) - p\right)I.$$

Then by the Courant-Fischer min-max theorem, we have

$$\lambda_2(2\mathbb{E}_{AX}[D] - 2\mathbb{E}_{AX}[A] + 11^\top) = \min_{u^\top y = 0, \|u\| = 1} u^\top(np(1 - q)I + (1 - p - pq)11^\top - p(1 - q)yy^\top)u = np(1 - q),$$

where the last equality holds by the assumption of $1 - p - pq \geq 0$, and the fact that $u$ and $y$ are orthogonal. \hfill $\Box$

Given the above result for the expectation, we can decompose (11) into the sum of several concentration parts. Once we prove that for every part the second smallest eigenvalue is greater than zero, the whole expression will be greater than zero.

We proceed in two steps. We first bound $A$ and $D$ from their expectations conditioning on some particular $X$, and then bound $\mathbb{E}_{X}[A]$ and $\mathbb{E}_{X}[D]$ from the total expectation $\mathbb{E}_{AX}[A]$ and $\mathbb{E}_{AX}[D]$. We present the following lemma.

**Lemma 4.** If each of the following

$$\frac{np(1 - q)}{8} + \min_i (d_i - \mathbb{E}_A[d_i]) > 0,$$  \hspace{1cm} (13)

$$\frac{np(1 - q)}{8} - \lambda_{\max}(A - \mathbb{E}_A[A]) > 0,$$  \hspace{1cm} (14)

$$\frac{np(1 - q)}{8} - \|\mathbb{E}_A[D] - \mathbb{E}_{AX}[D]\| > 0,$$  \hspace{1cm} (15)

$$\frac{np(1 - q)}{8} - \|\mathbb{E}_A[A] - \mathbb{E}_{AX}[A]\| > 0,$$  \hspace{1cm} (16)

holds, then semidefinite programming achieves exact recovery.
Proof. Our goal is to prove that the second smallest eigenvalue of $2D - 2A + 11^T$ is greater than zero. By Courant-Fischer min-max theorem, this is equivalent to proving
\[
\min_{u \perp y, \|u\|=1} u^T (2D - 2A + 11^T) u > 0
\]
holds with high probability. Here we are only interested in the space orthogonal to $y$, because $y$ is an eigenvector of the random matrix. Note that
\[
\|u\| = 1
\]
holds with high probability. Here we are only interested in the space orthogonal to $y$, because $y$ is an eigenvector of the random matrix. Note that
\[
\text{Proof.}
\]
\[
\min_{u \perp y, \|u\|=1} u^T (2D - 2A + 11^T) u = \min_{u \perp y, \|u\|=1} 2u^T (D - \mathbb{E}_A [D]) u
\]
\[
+2u^T (-A + \mathbb{E}_A [A]) u
\]
\[
+u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u.
\]

For the diagonal matrix part (17), we have
\[
\min_{u \perp y, \|u\|=1} 2u^T (D - \mathbb{E}_A [D]) u \geq 2\lambda_{\min} (D - \mathbb{E}_A [D])
\]
\[
= 2 \min_i (d_i - \mathbb{E}_A [d_i]).
\]

For the adjacency matrix part (18), we have
\[
\min_{u \perp y, \|u\|=1} 2u^T (-A + \mathbb{E}_A [A]) u \geq 2\lambda_{\min} (-A + \mathbb{E}_A [A])
\]
\[
= -2 \lambda_{\max} (A - \mathbb{E}_A [A]).
\]

We furthermore break down the remaining part (19) into three parts, including two concentration parts and one expectation part. The proof techniques used for the SBM in the prior literature [Hajek et al. 2016, Bandeira 2018, Abbe et al. 2016] do not work here because of the dependencies in the Latent Space Model.

\[
\min_{u \perp y, \|u\|=1} u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u = \min_{u \perp y, \|u\|=1} u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u
\]
\[
- u^T (2\mathbb{E}_A [A] - 2\mathbb{E}_A [A]) u
\]
\[
+ u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u.
\]

Next we transform the new terms above into spectral norms. For bounding (22), we obtain
\[
\min_{u \perp y, \|u\|=1} u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u \geq -2 \|\mathbb{E}_A [D] - \mathbb{E}_A [A]\|.
\]

For bounding (23), we obtain
\[
\min_{u \perp y, \|u\|=1} - u^T (2\mathbb{E}_A [A] - 2\mathbb{E}_A [A]) u \geq -2 \|\mathbb{E}_A [A] - \mathbb{E}_A [A]\|.
\]

Finally for bounding (24), assuming $p(1 + q) \leq 1$, we obtain
\[
\min_{u \perp y, \|u\|=1} u^T (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T) u = \lambda_2 (2\mathbb{E}_A [D] - 2\mathbb{E}_A [A] + 11^T)
\]
\[
= np (1 - q).
\]

The last equality holds by Lemma [3]

To complete our proof, we divide $np (1 - q)$ into four equal parts, and combine each part with the results from (20), (21), (25), and (26). □
We now state the conditions for exact recovery for the symmetric Latent Space Model using semidefinite programming.

**Theorem 2.** In a symmetric Latent Space Model with parameters \((n, d, \mu, \sigma)\), given that \(p(1+q) \leq 1\), if

\[
4n \left( 4n \cdot \frac{r(1 - s_0 - 2s_1)}{p^2 (1 - q)^2} - 1 \right) \\
+ 32n \left( \frac{2}{1 - q} - \frac{2p'}{p^2 (1 - q)^2} + \frac{r(3 - s_0 - 2s_1)}{p^2 (1 - q)^2} \right) \\
+ 64 \left( \frac{2r}{p^2 (1 - q)^2} - \frac{p'}{p^2 (1 - q)^2} - \frac{1}{(1 - q)^2} \right) \leq c_0, \tag{27}
\]

holds for some constant \(c_0 > 0\), and

\[
\frac{32n}{(1 - q)^2} \left( \frac{p'(1 + q)}{p^2} - (1 + q^2) \right) \leq c_1 \tag{28}
\]

holds for some constant \(c_1 > 0\), then as \(n\) gets large, semidefinite programming achieves exact recovery with high probability.

**Proof.** In order to prove that (11) holds with high probability, we first apply concentration inequalities to (13), (14), (15), and (16). Our goal is to prove that if condition (27) and (28) hold, each of the four inequality holds with probability at least \(1 - \frac{c}{n^1}\), where \(c > 0\) is some constant.

In the following proofs we denote \(f_{ij} := \mathbb{E}_A [A_{ij}] = \exp(-\|x_i - x_j\|^2)\), and \(\Delta_{ij}\) to be the matrix with 1 in entries \((i, j)\) and \((j, i)\), and 0 everywhere else.

To show that (13) holds with high probability, we use Hoeffding’s inequality in our proof. For any fixed latent vectors \(X\) and any \(i \in [n]\), we have

\[
P_A \{ d_i - \mathbb{E}_A [d_i] \leq -t \mid X \} \leq \exp \left( -\frac{2t^2}{n} \right),
\]

for every \(t > 0\).

By the union bound, it follows that

\[
P_A \left\{ \min_i (d_i - \mathbb{E}_A [d_i]) \leq -t \mid X \right\} \leq n \exp \left( -\frac{2t^2}{n} \right).
\]

Integrating over all possible \(X\), we get

\[
P_A X \left\{ \min_i (d_i - \mathbb{E}_A [d_i]) \leq -t \right\} \leq n \exp \left( -\frac{2t^2}{n} \right).
\]

Setting \(t = \frac{1}{8} np(1 - q)\), it follows that

\[
P_A X \left\{ \min_i (d_i - \mathbb{E}_A [d_i]) \leq -\frac{1}{8} np(1 - q) \right\} \leq n \exp \left( -\frac{np^2 (1 - q)^2}{32} \right),
\]

which is bounded above by \(n^{-1}\).

To show that (14) holds with high probability, we use the matrix Bernstein inequality [Tropp, 2012, Theorem 1.4] in our proof. Here we define \(M_{ij} := (A_{ij} - f_{ij}) \cdot \Delta_{ij}\). Note that for any latent vectors \(X\), \(\mathbb{E}_A [M_{ij} \mid X] = 0\).
By our construction, \( \sum_{i<j} M_{ij} = A - \mathbb{E}_A [A] \), and \( \Delta_{ij}^2 \) is the matrix with 1 in entries \((i,i), (j,j)\), and 0 everywhere else. As a result, the maximum eigenvalue of each \( M_{ij} \) is bounded above by

\[
\lambda_{\text{max}}(M_{ij}) = |A_{ij} - f_{ij}| \leq 1,
\]

and

\[
\left\| \sum_{i<j} \mathbb{E}_A [M_{ij}^2] \right\| = \left\| \sum_{i<j} \mathbb{E}_A [(A_{ij} - f_{ij})^2] \cdot \Delta_{ij}^2 \right\|
\]
\[
= \left\| \sum_{i<j} \text{Var}(A_{ij}) \cdot \Delta_{ij}^2 \right\|
\leq n - 1.
\]

Then applying the matrix Bernstein inequality, we have

\[
P_A \{ \lambda_{\text{max}}(A - \mathbb{E}_A [A]) \geq t \mid X \} \leq n \exp \left( \frac{-t^2/2}{(n-1) + t/3} \right)
\leq n \exp \left( \frac{-t^2/2}{n + t/3} \right).
\]

Integrating over all possible \( X \), we get

\[
P_{AX} \{ \lambda_{\text{max}}(A - \mathbb{E}_A [A]) \geq t \} \leq n \exp \left( \frac{-t^2/2}{n + t/3} \right).
\]

Setting \( t = \frac{1}{8} np(1 - q) \), it follows that

\[
P_{AX} \left\{ \lambda_{\text{max}}(A - \mathbb{E}_A [A]) \geq \frac{1}{8} np(1 - q) \right\} \leq n \exp \left( -\frac{np^2(1 - q)^2}{256} \right),
\]

which is bounded above by \( n^{-1} \).

To show that (15) holds with high probability, we use the matrix Chebyshev inequality \cite{Mackey14} in our proof. Note that

\[
P_X \{ \| \mathbb{E}_{AX} [D] - \mathbb{E}_A [D] \| \geq t \} \leq t^{-2} \cdot \mathbb{E}_X \left[ \| \mathbb{E}_{AX} [D] - \mathbb{E}_A [D] \|_F^2 \right]
\]
\[
= \frac{n}{t^2} \left( \frac{n^2}{4} \left( r(1 - s_0 - 2s_1) - p^2(1 - q)^2 \right)
+ n \left( p^2(1 - q) - p' + \frac{1}{2} r(3 - s_0 - 2s_1) \right) + 2r - p' - p^2 \right).
\]

The last equality follows from Lemma 1. Setting \( t = \frac{1}{8} np(1 - q) \), it follows that

\[
P_X \left\{ \| \mathbb{E}_{AX} [D] - \mathbb{E}_A [D] \| \geq \frac{1}{8} np(1 - q) \right\} \leq 4 \left( 4n \cdot \frac{r(1 - s_0 - 2s_1)}{p^2(1 - q)^2} - 1 \right)
+ 32 \left( \frac{2}{1 - q} - \frac{2p'}{p^2(1 - q)^2} + \frac{r(3 - s_0 - 2s_1)}{p^2(1 - q)^2} \right)
+ 64 \left( \frac{2r}{np^2(1 - q)^2} - \frac{p'}{np^2(1 - q)^2} - \frac{1}{n(1 - q)^2} \right),
\]

10
which is bounded above by $c_0 n^{-1}$ by condition (27).

To show that (16) holds with high probability, we use the matrix Chebyshev inequality [Mackey et al., 2014, Proposition 6.2] in our proof. Note that

$$P_X \{\|EAX - E[A]\| \geq t\} \leq t^{-2} \cdot E_X \left[\|EAX - E[A]\|^2_F \right] \leq \frac{n^2}{2t^2} \left(p'(1 + q') - p^2(1 + q^2)\right).$$

The last inequality follows from Lemma 1. Setting $t = \frac{1}{8} np(1 - q)$, it follows that

$$P_X \left\{\|EAX - E[A]\| \geq \frac{1}{8} np(1 - q)\right\} \leq \frac{32p'(1 + q')}{p^2} - \frac{32(1 + q^2)}{(1 - q)^2},$$

which is bounded above by $c_1 n^{-1}$ by condition (28).

Combining the results above and applying Lemma 4, the probability of the second smallest eigenvalue being greater than zero as in (11), is at least $1 - 4cn^{-1}$. This completes our proof.

5 Discussion

Here we provide the comparison between the impossible regime and the polynomial-time regime by synthetic experiments. The procedure is as follows. First we fix the number of nodes $n$ to be 300, and the number of trials to be 100. The variables in our simulations are $\|\mu\|$ and $\sigma$. We generate $y^*$ by randomly assigning $n/2$ nodes to the positive group ($y^*_i = 1$), and $n/2$ nodes to the negative group ($y^*_i = -1$). Then for every node $i$, we generate the corresponding latent vector $x_i$ according to the $d$-dimensional Gaussian distribution centered at $y_i \mu$ with covariance of $\sigma^2 I$, where $\mu = (\|\mu\|^2/d)^{1/2} 1_d$, and $1_d$ is the all-one vector in $\mathbb{R}^d$. Next we randomly sample the adjacency matrix $A$ based on the distances between the latent vectors. For each pair of $\|\mu\|$ and $\sigma$, we count how many times (out of 100) the matrix $2D - 2A + 11^\top$ is positive semidefinite and the second smallest eigenvalue is greater than zero, which allows us to compute an empirical probability of success. Lemma 2 guarantees that if the matrix is PSD and the second smallest eigenvalue is positive, then exact recovery can be performed efficiently by semidefinite programming.

Figure 1 shows our results. Area I in the figure (area to the left of the red curve) is the theoretical impossible regime given by Theorem 1. Area II in the figure (area to the right of the green curve, and below the blue curve) is the theoretical polynomial-time regime given by Theorem 2.

In the future, our research could be extended in several directions. Note that in our simulations, there exists a gap between the experiment results and the threshold given by Fano’s inequality. It might be possible to achieve better information-theoretic lower bounds using other approaches. Furthermore, from our simulations we notice there might exists a phase transition phenomenon similar to the one in the Stochastic Block Model [Abbe, 2018]. Finally, some more complicated latent space assumptions could also be introduced, for instance, using a non-diagonal covariance matrix in the Gaussian distribution, changing the Euclidean distance between latent vectors to other metrics, among others.

6 Appendix

Here, we reproduce the statements of the technical lemmas that we used in our proofs.
Figure 1: The empirical probability of the second smallest eigenvalue of $2D - 2A + 11^\top$ being positive. We fix the dimension of the latent space $d$ to be 2. The red curve matches the threshold from the impossible regime. The green curve matches the threshold from the polynomial-time regime with $c_0 = 500$, and the blue curve matches the threshold from the polynomial-time regime with $c_1 = 500$. Area I corresponds to the theoretical impossible regime, while Area II corresponds to the theoretical polynomial-time regime.

**Lemma 5** (Gaussian Moment Generating Function (Theorem 3.2a.1, Mathai and Provost [1992])). Let $x \sim N_p(\mu, \Sigma)$, $\Sigma$ is positive definite, $Q = x^\top Ax$, $A = A^\top$. Then the moment generating function of $Q$ is given by

$$M_Q(t) = \mathbb{E}_x \left[ \exp(t x^\top Ax) \right] = \int_x \frac{\exp(t x^\top Ax - \frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu))}{(2\pi)^{p/2} |\Sigma|^{1/2}} dx.$$  

Furthermore, if $(\Sigma^{-1} - 2tA)$ is symmetric positive definite, we have

$$M_Q(t) = |1 - 2t \Sigma^{1/2} A \Sigma^{1/2}|^{-1/2} \cdot \exp \left( t \mu^\top \Sigma^{-1/2} (\Sigma^{1/2} A \Sigma^{1/2}) \cdot (I - 2t \Sigma^{1/2} A \Sigma^{1/2})^{-1} \Sigma^{-1/2} \mu \right).$$

Next we introduce the following matrix concentration inequalities.

**Lemma 6** (Matrix Bernstein Inequality (Theorem 1.4, Tropp [2012])). Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices with dimension $d$. Assume that each random matrix satisfies

$$\mathbb{E} \left[ X_k \right] = 0$$

and

$$\lambda_{\text{max}}(X_k) \leq R$$

almost surely.

Then, for all $t \geq 0$,

$$\mathbb{P} \left\{ \lambda_{\text{max}} \left( \sum_k X_k \right) \geq t \right\} \leq d \exp \left( \frac{-t^2/2}{\sigma^2 + Rt/3} \right),$$

where

$$\sigma^2 := \left\| \sum_k \mathbb{E} \left[ X_k^2 \right] \right\|.$$
Lemma 7 (Matrix Chebyshev Inequality (Proposition 6.2, *Mackey et al.* 2014)). Let $X$ be a random matrix. Then, for all $t > 0$,

$$
P\{\|X\| \geq t\} \leq \inf_{p \geq 1} t^{-p} \cdot E\left[\|X\|_p^p\right],
$$

where $\|\cdot\|_p$ is the Schatten $p$-norm defined as

$$
\|X\|_p := (\text{Tr} |X|^p)^{1/p}.
$$

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

Afonso S Bandeira. Random laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, 18(2):345–379, 2018.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yudong Chen and Jiaming Xu. Statistical-computational phase transitions in planted models: The high-dimensional setting. In *International Conference on Machine Learning*, pages 244–252, 2014.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10):5918–5937, 2016.

Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in neural information processing systems*, pages 657–664, 2008.

Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Chuyang Ke and Jean Honorio. Information-theoretic limits for community detection in network models. *Advances in Neural Information Processing Systems*, 2018.

Pavel N Krivitsky, Mark S Handcock, Adrian E Raftery, and Peter D Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3):204–213, 2009.

Philip M Long and Rocco A Servedio. Restricted boltzmann machines are hard to approximately evaluate or simulate. 2010.
Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, Joel A Tropp, et al. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3): 906–945, 2014.

Arakaparampil M Mathai and Serge B Provost. *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM review*, 38(1):49–95, 1996.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, 423:435, 1997.

Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th annual international conference on machine learning*, pages 1169–1176. ACM, 2009.