

GAUSSIAN MIXTURE MODELS FOR STOCHASTIC BLOCK MODELS WITH NON-VANISHING NOISE

Heather Mathews*

Vaishakhi Mayya†

Alexander Volfovsky*

Galen Reeves*†

* Department of Statistical Science, Duke University

† Department of Electrical and Computer Engineering, Duke University

ABSTRACT

Community detection tasks have received a lot of attention across statistics, machine learning, and information theory with work concentrating on providing theoretical guarantees for different methodological approaches to the stochastic block model. Recent work on community detection has focused on modeling the spectral embedding of a network using Gaussian mixture models (GMMs) in scaling regimes where the ability to detect community memberships improves with the size of the network. However, these regimes are not very realistic. This paper provides tractable methodology motivated by new theoretical results for networks with non-vanishing noise. We present a procedure for community detection using novel GMMs that incorporate truncation and shrinkage effects. We provide empirical validation of this new representation as well as experimental results using a large email dataset.

Index Terms— community detection, network analysis

1. INTRODUCTION

Network data are of paramount importance across many modern scientific fields [1–4]. One of the most common tasks in network analysis is the search for community structure among units in the network. Much of the statistical [5–7] and information theoretical [8–11] work on community detection is based on analyzing the performance of different algorithms in recovering communities from data generated from the stochastic block model (SBM) [12]. In this probabilistic network model, the probability of an edge between individuals i and j is governed exclusively by the community memberships, $X = (X_1, \dots, X_n)$, of the nodes. The complete network can then be represented by its adjacency matrix, A , where $A_{ij} = A_{ji} = 1$ if there is an edge between node i and node j and 0 otherwise. In this setting, the task of *community detection* is to recover the community labels X given the adjacency matrix A and possible side information.

A large body of work has considered spectral clustering methods for community detection with early work focusing on the behavior of clustering nodes via k -means [5] and more recent work focusing on Gaussian mixture models (GMMs) [6, 7]. Much of this work is derived under scaling regimes where

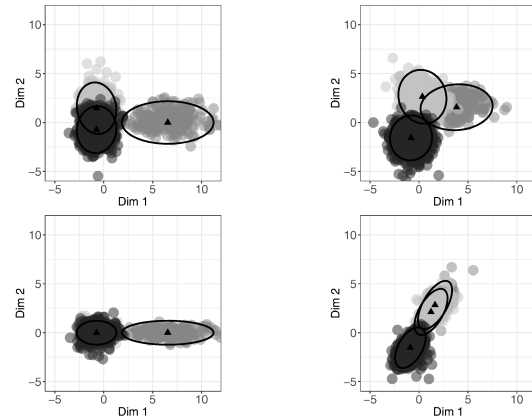


Fig. 1. Aligned eigenvectors from graphs generated from four different SBMs. The models in the top row have two eigenvalues of R that greater than one while the bottom row only has one eigenvalue greater than one. In the left column R is diagonal, while in the right column there is a rotation by an orthogonal matrix. Each $\blacktriangle$ and ellipse are based on the mean and covariance of GMM components from Section 3.

the ability to detect community memberships improves with the size of the graph. We call this the vanishing-noise regime.

In this paper we focus on the degree-balanced SBM in which each community has the same expected average degree d . In Section 2 we provide a formal problem formulation that describes how all the relevant information can be captured by in a $(K - 1)$ dimensional embedding of the eigenvectors of the adjacency matrix. We then develop a novel GMM representation for mid- to high-noise regimes that is able to appropriately quantify the uncertainty about the labels of the individual nodes. In Figure 1 we display our proposed eigenvector embedding and the mean and covariance matrices we derive in Section 3 for our GMM across models with both high- (lower row of plots) and low-noise (upper row of plots) levels. Section 4 provides an empirical validation of our method and a comparison to other state-of-the-art algorithms.

2. PROBLEM FORMULATION

Our approach is described in the context of a general latent space model for networks that includes the degree balanced SBM as a special case. Let $X = (X_1, \dots, X_n)$ be an $n \times s$ matrix of latent variables whose rows are drawn independently from a distribution P on $\mathbb{R}^s$ with mean zero, identity covariance, and bounded support. Conditional on X , the entries of the adjacency matrix $A = (A_{ij})$ are drawn independently according to

$$A_{ij} \sim \text{Bern}\left(\frac{d}{n} + \frac{\sigma}{n} X_i^T R X_j\right), \quad i < j, \quad (1)$$

where: d is a positive number that parameterizes the expected degree of each node in the network; R is a symmetric $s \times s$ matrix that describes how the probability of an edge depends on the latent variables; and $\sigma = \sqrt{d(n-d)/n}$ is a scaling factor that ensures that the signal-to-noise ratio is invariant to the choice of d . The tuple (n, P, d, R) is valid only if $d/n + (\sigma/n)x^T R \tilde{x}$ is between zero and one for all $x, \tilde{x}$ in the support of P .

The assumption that P has zero mean ensures that the model is degree-balanced in expectation. Specifically, $\mathbb{E}[A_{ij} | X_i] = d/n$, and thus the expected degree of node i is independent of X_i . The assumption that P has identity covariance is without loss of generality since any linear transformation of the latent variables X_i can be absorbed into the model parameter R .

When the support of P is finite this model reduces to the degree balanced SBM where each node is assigned to exactly one of K possible communities, independently with probability vector $p = (p_1, \dots, p_K)$. Specifically, we let X_i be supported on a set of K points $\{\mu_1, \dots, \mu_K\}$ in dimension $s = K - 1$ satisfying the moments constraints:

$$\sum_{k=1}^K p_k \mu_k = 0, \quad \sum_{k=1}^K p_k \mu_k \mu_k^T = I. \quad (2)$$

We leverage a unique specification of P given in [11, Remark 1] as a function of label probabilities p . We note that an alternative representation for the labels, used in previous work [5], is to associate the k -th label with the k -th standard basis vector in $\mathbb{R}^K$. An explicit mapping between these various representations is provided in [11].

3. THEORY AND METHODOLOGY

In this section, we present a general method for inference in degree-balanced networks. This method is based on a novel Gaussian approximation for the spectral embedding of the adjacency matrix. Recall that the leading eigenvector is correlated with the degree of the nodes and thus does not provide any information about the memberships. Therefore, we consider the spectral embedding of the normalized adjacency

matrix $A - (d/n)\mathbf{1}\mathbf{1}^T$. The projection of this matrix onto the space of rank- s matrices can be expressed as $V\Lambda V^T$ where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_s)$ contains the largest eigenvalues (in magnitude), in decreasing order, and $V = (V_1, \dots, V_n)^T$ is an $n \times s$ matrix with orthonormal columns corresponding to the eigenvectors. We note that the representation of the eigenvectors is not unique.

3.1. Gaussian Approximation of Eigenvectors

In the context of community detection, the basic principle underlying spectral clustering is that the rows of the leading eigenvectors are correlated with the latent variables. Applying standard clustering techniques, such as k -means, directly on the points $V_1, \dots, V_n \in \mathbb{R}^s$ provides a partition of the nodes in the network and can be used to estimate community memberships. *More generally, a principled approach to inference is to formulate a joint model for the eigenvectors and the latent parameters.* This applies in the general latent space model as well as in the specialized case of community detection where the parameter space is finite. This paper builds upon recent work [6, 7, 11], which provides both theoretical and empirical support for the use of GMMs.

To describe our approach, we introduce the scaled eigenvectors $Y_1, \dots, Y_n \in \mathbb{R}^s$ according to

$$Y_i = \sqrt{n} \text{diag}(r_1, \dots, r_s) V_i, \quad (3)$$

where $r_1 \geq r_2 \geq \dots r_s$ are the eigenvalues of R . The results in Athreya et al. [6, Theorem 4.8] can be used to characterize the asymptotic distribution of Y_i in the vanishing-noise regime where P is fixed while the model parameters (n, d, R) scale to infinity. Adapted to the setting of this paper, this result suggests the following approximation:

$$UY_i \sim \mathcal{N}\left(RX_i, \tilde{\Sigma}(X_i)\right), \quad (4)$$

where U is an orthogonal matrix that aligns the eigenvectors with the latent variables and the covariance is given by

$$\tilde{\Sigma}(x) = \mathbb{E}_{X_0 \sim P} [\nu(x, X_0) X_0 X_0^T] \quad (5)$$

$$\nu(x, \tilde{x}) = 1 + \left(\frac{n-2d}{n\sigma}\right) x^T R \tilde{x} - \frac{1}{n} (x^T R \tilde{x})^2. \quad (6)$$

The matrix U depends on the eigenspace of R as well as the particular choice of eigenvectors used in the eigendecomposition of A . In the proposed method described below, this matrix is estimated from the data using maximum likelihood.

It is important to emphasize that the approximation in (4) is adapted from the vanishing-noise regime where the eigenvalues of R scale with n . As a consequence, some important aspects of the moderate to high noise regimes are not captured. In particular, it is well known that an eigenvector is uninformative about latent structure unless its associated eigenvalue exceeds a threshold.

We propose a Gaussian approximation for the scaled eigenvectors Y_i that incorporates both truncation and shrinkage effects. Let $\bar{R}$ and $\underline{R}$ be the symmetric $s \times s$ matrices obtained by applying the mappings $r \mapsto \max(|r|, 1)$ and $r \mapsto \min(|r|, 1)$, respectively, to the eigenvalues of R . Our approximation is given by

$$UY_i \sim \mathcal{N}\left((\bar{R}^2 - I)^{1/2} X_i, \Sigma(X_i)\right), \quad (7)$$

where U is an orthogonal matrix and the covariance is

$$\Sigma(x) = (I - \bar{R}^{-2})^{-1/2} \tilde{\Sigma}(x) (I - \bar{R}^{-2})^{-1/2} + \bar{R}^{-1} \underline{R}^2 \bar{R}^{-1}. \quad (8)$$

The term $(\bar{R}^2 - I)^{1/2}$ provides the truncation and shrinkage to the eigenvalues of R . Note that any direction corresponding to an eigenvalue of magnitude less than one does not provide any information about X_i .

Our approximation follows from a leave-one-out argument combined with asymptotic properties of spiked Wigner matrices [13]. The full derivation is omitted due to space constraints.

Proposition 1. *The matrices $\Sigma(x)$ and $\tilde{\Sigma}(x)$ satisfy*

$$\Sigma(x) = \tilde{\Sigma}(x) + O(\lambda_{\min}^{-2}(R)) \quad (9)$$

$$\Sigma(x) = \underline{R}^2 + O(\sigma^{-1} + n^{-1}) \quad (10)$$

$$\tilde{\Sigma}(x) = \underline{R}^2 + O(\sigma^{-1} + n^{-1}). \quad (11)$$

Proof Sketch. These results follow straightforwardly from the assumption that P has bounded support and the fact that $\nu(x, \tilde{x}) = 1 + O(\sigma^{-1} + n^{-1})$. $\square$

Proposition 1 shows that as R increases, the approximation in (7) converges to the vanishing-noise approximation given in (4). Theorem 1 also shows that if either d or $n - d$ increase with n , then the covariance does not depend on X_i , and is given by $\underline{R}^2$. Interestingly, this result establishes a connection between the analysis of spectral methods and the information-theoretic analysis of dense networks given in [11], which also involves a GMM with common covariance across the mixtures.

3.2. Proposed Method

Let $\mathbb{P}_0(x, z)$ be the distribution of the pair (X_0, Z_0) where $X_0 \sim P$ and Z_0 is conditionally Gaussian:

$$Z_0 \sim \mathcal{N}\left((\bar{R}^2 - I)X_0, \Sigma(X_0)\right), \quad (12)$$

and $\Sigma(x)$ is defined by (8). Our method has three components:

1. (Spectral embedding) Let V be the eigenvectors of the rank- s projection of the normalized adjacency matrix $A - (d/n)\mathbf{1}\mathbf{1}^T$ and let $Y_1, \dots, Y_n$ be given by (3).

2. (Alignment via maximum likelihood) Let U^* be a solution to the optimization problem

$$\max_U \prod_{i=1}^n \mathbb{P}_0(UY_i), \quad (13)$$

where the maximum is over all orthogonal matrices U and $\mathbb{P}_0(z)$ is the marginal of $\mathbb{P}_0(x, z)$ with respect to x . For an SBM, this is the marginal over a GMM.

3. (Classification) For $i = 1, \dots, n$ output the the posterior $P_0(x | z)$ evaluated on the rotated data $z = (U^* Y_i)$. For an SBM, the posterior is represented by the probability vector $\hat{p}_i = (\hat{p}_{i1}, \dots, \hat{p}_{iK})$.

Besides the eigenvalue decomposition, the potentially computationally challenging step in our method is the optimization with respect to an orthogonal matrix. For convenience, this optimization can be carried out over a restricted set of orthogonal matrices belonging to the set $\{U : U \text{diag}(r_1, \dots, r_s)U^T = R\}$. In the simulations that follow, we obtain an approximate solution by searching over a set of representative orthogonal matrices.

4. EXPERIMENTAL RESULTS

In this section, we study the behavior of the degree-balanced SBM, parameterized by (n, p, d, R) from Section 2.

4.1. Numerical Simulations

We generate a network of $n = 5000$ nodes and $K = 3$ communities with probability vector $p = (0.1, 0.3, 0.6)$. The support of P is defined according to [11, Remark 1], which yields

$$\mu_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} -1/3 \\ 2\sqrt{5}/3 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} -1/3 \\ -\sqrt{5}/3 \end{pmatrix}.$$

The adjacency matrix is generated according to (1) with average degree $d = 15$ and

$$R = U \text{diag}(r_1, r_2)U^T, \quad U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We compare four different methods:

- (GMM) The method described in Section 3.2.
- (Low-Noise GMM) This is the version of the method described in Section 3.2 where $\mathbb{P}_0(x, z)$ corresponds to the low-noise approximation in (4).
- (Uninformed GMM) This method fits a GMM to the the rows of the eigenvectors associated with the $K - 1$ largest eigenvalues (in magnitude) of $A - (d/n)\mathbf{1}\mathbf{1}^T$,
- (K -Means) This method is applied to the same selected eigenvectors as the uninformed GMM.

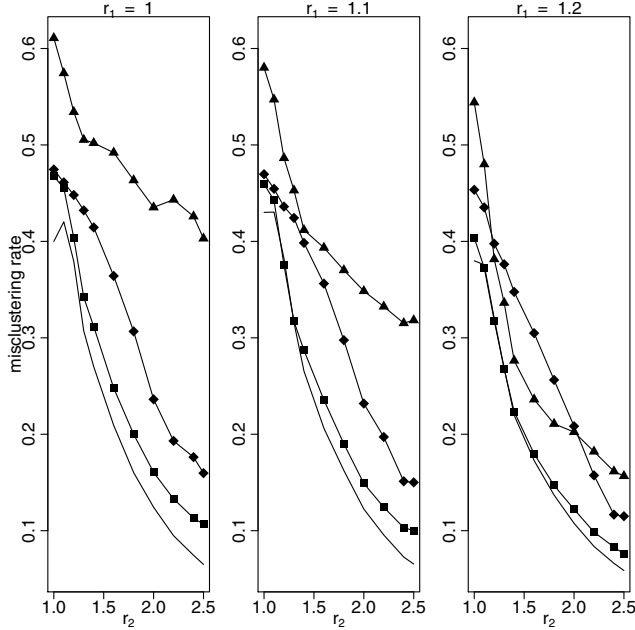


Fig. 2. Misclustering rate across varying regimes. Each point is the average misclustering rate over 100 independent networks. Methods are given by: GMM: ●, Low-Noise GMM: ■, Uninformed GMM: ◆, K -Means: ▲.

We study the regimes where $(\tilde{r}_1, \tilde{r}_2) \in \{1, 1.1, 1.2\} \times \{1, \dots, 2.6\}$ and $(r_1, r_2) = \text{sort}(\tilde{r}_1, \tilde{r}_2)$. Performance is assessed using the misclustering rate. For the GMM methods, we use the maximum a posteriori estimate of the of the community memberships. Following the usual convention in the literature, this metric optimized over permutations of the estimated labels, to mitigate the effects of label switching.

Figure 2 shows the misclustering rate across the different regimes. Each point in the Figure is the average misclustering rate over 100 networks generated for each set of parameters. As r_1 and r_2 increase, we see an improvement across all methods. When the eigenvalues of R are close to one, there is almost no correlation between the eigenvectors of the adjacency matrix and the community structure and so all methods perform poorly. Two interesting phenomena can be observed in the Figure: first, the relative advantage of our proposed approach to other approaches appears to increase as the larger eigenvalue of R grows (when $r_2 = 2.5$ we have a nearly 50% improvement over the next competing method). Second, as the eigenvalues of R grow, the performance of K -Means improves and surpasses that of the uninformed GMM.

4.2. Real-World Data Analysis

In this section, we apply our method to an email network from a large European research institution [14, 15]. In these data,

an undirected edge exists if one person sent another person at least one email. Only communication between individuals within the institution is considered and each person belongs to one of 42 known departments/communities. The number of communities is reduced to 3 such that 2 of the communities correspond to nodes in the two largest departments and the third community is made up of all other nodes. The average degree per community for this example is around 30 and thus approximately degree balanced.

We provide two types of analysis for this data: one based on oracle model parameters and one based on estimated ones. Consider that an oracle provides us with the true values of R and p (which we can compute based on our ground truth community information). Based on this, the misclustering rate of our approach is 0.2. In practice one rarely has access to true R and p values, we also estimate an $\hat{R}$ and $\hat{p}$ based on a 10% sample from the true communities. Using the estimated values of p and R we apply the method of Section 3 and achieve a misclustering rate of 0.3 where we have accounted for using 10% of the data to learn the model parameters. This compares favorably to the performance of K -means which achieve misclustering rates of 0.368.

5. CONCLUSION

In this paper we propose a Gaussian Mixture Model representation of a projection of the adjacency matrix that can be leveraged for community detection in the non-vanishing noise regime. In contrast to prior work, we use a model for the joint distribution between the eigenvectors and the latent community structure that includes truncation and shrinkage effects. We demonstrate empirically that this novel representation is able to improve on the performance of community detection in moderate to high noise regimes.

For future directions, it would be interesting to see if these empirical results can be proven rigorously and extended to regularized spectral methods based on the graph Laplacian [16] or other data-driven techniques [17]. This method can be extended to the settings of multiple observed networks on the same set of units [18].

6. REFERENCES

- [1] D. R. Farine and H. Whitehead, “Constructing, conducting and interpreting animal social network analysis,” *Journal of Animal Ecology*, vol. 84, no. 5, pp. 1144–1163, 2015.
- [2] R. I. Dunbar, V. Arnaboldi, M. Conti, and A. Passarella, “The structure of online social networks mirrors those in the offline world,” *Social networks*, vol. 43, pp. 39–47, 2015.
- [3] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, “Integration in emerging social networks explains aca-

- demic failure and success,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 3, pp. 792–797, 2019.
- [4] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on twitter during the 2016 us presidential election,” *Science*, vol. 363, no. 6425, pp. 374–378, 2019.
 - [5] K. Rohe, S. Chatterjee, and B. Yu, “Spectral clustering and the high -dimensional stochastic blockmodel,” *The Annals of Statistics*, vol. 39, no. 4, pp. 1878–1915, 2011.
 - [6] A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman, “A limit theorem for scaled eigenvectors of random dot product graphs,” *Sankhya A*, vol. 78, no. 1, pp. 1–18, 2016.
 - [7] S. Suwan, D. Lee, R. Tang, and et al, “Empirical bayes estimation for the stochastic blockmodel,” *Electronic Journal of Statistics*, vol. 10, pp. 761–782, 2016.
 - [8] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications,” *Physical Review E*, vol. 84, no. 6, p. 066106, 2011.
 - [9] E. Abbe, “Community detection and stochastic block models: Recent developments,” *Journal of Machine Learning Research*, vol. 18, no. 177, pp. 1–86, 2018.
 - [10] M. Lelarge and L. Miolane, “Fundamental limits of symmetric low-rank matrix estimation,” *Probability Theory and Related Fields*, 2018.
 - [11] G. Reeves, V. Mayya, and A. Volfovsky, “The geometry of community detection via the mmse matrix,” *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019.
 - [12] P. Holland, K. Laskey, and S. Leinhardt, “Stochastic blockmodels : First steps,” *Social Networks*, vol. 5, pp. 109–137, 1983.
 - [13] F. Benaych-Georges and R. Nadakuditi, “The singular values and vectors of low rank perturbations of large rectangular random matrices,” *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.
 - [14] H. Yin, A. Benson, J. Leskovec, and D. Gleich, ““local higher-order graph clustering,” in *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.
 - [15] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, 2007.
 - [16] Y. Zhang and K. Rohe, “Understanding regularized spectral clustering via graph conductance,” *arXiv preprint arXiv:1806.01468*, 2018.
 - [17] P. Zhang, “Robust spectral detection of global structures in the data by learning a regularization,” in *Advances in Neural Information Processing Systems*, 2016, pp. 541–549.
 - [18] V. Mayya and G. Reeves, “Mutual information in community detection with covariate information and correlated networks,” in *57th Annual Allerton Conference on Communication, Control, and Computing*, 2019.