BRIDGING ARCHETYPAL ANALYSIS AND STOCHASTIC BLOCK MODELS

A PREPRINT

Aleix Alcacer
Department of Mathematics
Jaume I University
Castelló de la Plana
aalcacer@uji.es

Morten Mørup
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Kongens Lyngby
mmor@dtu.dk

June 19, 2023

ABSTRACT

TODO

Keywords Archetypal Analysis • Stochastic Block Models • Community Detection • Network Analysis • Matrix Factorization

1 Introduction

Research Questions:

- BiAA as a bridge between AA and SBM.
- BiAA extensions:
  - Multiple likelihoods (Normal, Bernoulli, Poisson)
  - Hard assignment
  - Degree correction
- What are the merits of Bi-AA as opposed to SBM?
  - Soft clustering
  - Better interpretation

This research paper presents an extension of the Biarchetype Analysis (BiAA) method for bridging Archetypal Analysis and Stochastic Block Models analysis. The proposed extension includes the support of multiple likelihood functions (Normal, Bernoulli and Poisson), the use of a hard assignment mode, and the application of degree correction. The paper also compares the merits of BiAA to existing stochastic block model (SBM) algorithms, with a focus on the benefits of soft clustering and better interpretation of the results.

Biarchetype analysis is a statistical method that aims to uncover the underlying structure of data by grouping similar observations into clusters or archetypes. As it is said, the proposed extension of BiAA includes the incorporation of multiple likelihood functions, which allows for the use of different probability distributions to model the data. This enables the method to be applied to a wider range of data types, such as binary data, count data, and continuous data.

In addition, the proposed extension includes the use of a hard assignment mode, which assigns each observation to the archetype with the highest probability of membership. This allows for a more straightforward interpretation of the

results, compared to the soft assignment mode used in the original BiAA method, where each observation is assigned a probability of membership in each archetype.

Finally, a degree correction is incorporated to BiAA, which adjusts the weights assigned to the nodes in the network based on their degree (i.e., the number of connections they have to other nodes). This ensures that nodes with higher degrees are not given too much influence in the calculation of the archetypes, which can lead to more accurate and unbiased results.

The paper also compares the merits of BiAA to existing SBM algorithms, which are commonly used for detecting community structure in networks. One key advantage of BiAA is its ability to perform soft clustering, where each observation can belong to multiple archetypes with varying degrees of membership. This allows for a more nuanced representation of the data and the underlying structure of the network. In contrast, SBM algorithms typically perform hard clustering, where each observation is assigned to exactly one block or community.

Another advantage of BiAA is its ability to provide better interpretation of the results. The archetypes identified by BiAA are derived from the data itself, and can be interpreted as extreme points in the data space. This allows for a more intuitive understanding of the patterns and structures in the data, compared to the block assignments produced by SBM algorithms, which are based on the probabilities of connections between nodes in the network.

Overall, the proposed extension of BiAA offers a bridge between archetypal analysis and SBM analysis, providing a different framework for community structure detection analysis. Therefore, BiAA will be an alternative and valuable tool for uncovering the underlying structure of communities and gaining insights into the patterns and structures present in the data.

2 Methods

2.1 Likelihood functions

Likelihood functions are used in statistical modeling to evaluate the quality of a model's predictions. They are a measure of how well a model fits the observed data. By maximizing the likelihood of the model, it is possible to improve the model's accuracy and make better predictions.

In the article, log-likelihood functions will be used as error functions in the defined models. This means that the models will be trained to minimize the negative log-likelihood of their predictions, which will improve their accuracy. The Normal, Bernoulli, and Poisson log-likelihood functions will be used in this case, which are appropriate for models that make predictions using normal, binary, and count data, respectively.

2.1.1 Normal distribution

The normal log-likelihood function is used to calculate the log-likelihood of a normal distribution given a set of data. The formula for the normal log-likelihood function is as follows:

$$log(L) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}$$

where L is the likelihood function, n is the number of data points, x_i is the i-th data point, μ is the mean of the data, σ is the standard deviation of the data, and the summation is over all data points in the sample.

2.1.2 Bernoulli distribution

The Bernoulli log-likelihood function is used to calculate the log-likelihood of a Bernoulli distribution given a set of data. The formula for the Bernoulli log-likelihood function is as follows:

$$log(L) = \sum_{i=1}^{n}[y_i \cdot log(p) + (1 - y_i) \cdot log(1 - p)]$$

where L is the likelihood function, y_i is the outcome of the binary event (1 if the event happened, 0 if it did not), p is the probability of the event happening, and the summation is over all events in the data.

2.1.3 Poisson distribution

Finally, the Poisson log-likelihood function is used to calculate the log-likelihood of a Poisson distribution given a set of data. The formula for the Poisson log-likelihood function is as follows:

$$log(L) = \sum_{i=1}^{n} [n_i \cdot log(\lambda) - \lambda - log(n_i!)]$$

where L is the likelihood function, n_i is the number of events observed in the i-th group, λ is the rate parameter of the Poisson distribution, and the summation is over all groups in the data.

2.2 Stochastic Block Model

The Stochastic Block Model (SBM) is a popular model used to generate synthetic networks with community structure. In an SBM, the nodes in the network are divided into groups or "blocks", and the connections between nodes are determined by the blocks they belong to. The SBM is often used to test the performance of community detection algorithms, which are algorithms that aim to identify the groups or communities present in a network.

For bipartite networks, the SBM can be viewed as a matrix factorization problem where the adjacency matrix (which represents the connections between the nodes) is decomposed into three matrices, one representing the blocks and the others representing the connections between the blocks. More specifically, let X be the adjacency matrix of a network with M and N nodes respectively, where A_{ij} is the number of connections between nodes i and j. In the SBM, this adjacency matrix can be written as the product of three matrices, A, Z and C, where A is a $M \times K$ matrix, Z is a $K \times L$ matrix and D is a $L \times N$ matrix:

$$X \approx AZD$$

Here, A and D are the matrices that represents the blocks, where A_{ik} is the membership of node i in block k and D_{lj} is the membership of node j in block l. Z is the matrix that represents the connections between the blocks, where Z_{kl} is the probability of a connection between nodes in block k and block l.

The matrix factorization of the adjacency matrix in the SBM allows us to model the community structure of the network by defining the blocks and the connections between the blocks. The blocks can be thought of as the groups or communities in the network, and the connections between the blocks can be used to control the strength and structure of the communities. For example, if Z_{kl} is high, it indicates that there are many connections between nodes in block k and block l, and the corresponding communities are likely to be strongly connected. On the other hand, if Z_{kl} is low, it indicates that there are few connections between nodes in block k and block l, and the corresponding communities are likely to be weakly connected.

In summary, the Stochastic Block Model is a popular model for generating synthetic networks with community structure, and it can be viewed as a matrix factorization problem where the adjacency matrix of the network is decomposed into three matrices representing the blocks and the connections between the blocks.

References: [**?**, **?**].

```
# $$X_{MN} \sim A_{MK} Z_{KL} D_{LN}$$
```

2.3 Biarchetype Analysis

[Archetype and biarchetype analysis definition]

Therfore, this method can be seen as a particular case of the Stochastic Block Model where the matrix Z, that represents the relationships between groups, is defined by taking convex combinations of the data X. This means that the matrix Z, called the archetypal matrix, is constructed as $Z = BXC$ where X is the matrix representing the original data points and B and C are matrices that represent the coefficients of the convex combination. By defining the relationship matrix in this way, Biarchetype Analysis is able to interpret the matrix Z based on input data.

2.4 Degree correction

A degree corrected algorithm is a variant of an existing algorithm that has been modified to account for the fact that the nodes in a network may have different degrees (i.e., different numbers of connections to other nodes). In many networks, the nodes have varying degrees, and a standard algorithm may not perform well when applied to such a network because it may give more weight to nodes with higher degrees, leading to biased or inaccurate results. A degree corrected algorithm addresses this issue by taking into account the degrees of the nodes when calculating the algorithm's result.

For example, when applying a community detection algorithm to a network generated using the SBM, it is important to take into account the degree distribution of the nodes in the network. The degree distribution of a network is the

distribution of the number of connections (or "edges") that each node has to other nodes in the network. In a network generated using the SBM, the degree distribution of the nodes will be different depending on the blocks they belong to. For example, nodes in the same block will tend to have similar degrees, while nodes in different blocks will have different degrees.

If a standard community detection algorithm is applied to a network generated using the SBM, it may not perform well because it may give more weight to nodes with higher degrees, leading to biased or inaccurate results. A degree corrected version of the algorithm, on the other hand, would take into account the degree distribution of the nodes and give less weight to nodes with higher degrees, resulting in more accurate community detection.

2.5 Proposed Models

2.5.1 Soft SBM

$$\arg\max_{A,Z,D} l(AZD; X) \tag{1}$$

such that:

- $\sum_{k=1}^{K} A_{mk} = 1$ with $A_{mk} \in [0, 1]$ for each $m = 1, \ldots, M$.
- $\sum_{l=1}^{L} D_{ln} = 1$ with $D_{ln} \in [0, 1]$ for each $n = 1, \ldots, N$.
- If the negative Bernoulli log-likelihood is used as loss function, $Z \in [0, 1]$.

2.5.2 Hard SBM

$$\arg\max_{A,Z,D} l(AZD; X) \tag{2}$$

such that:

- $\sum_{k=1}^{K} A_{mk} = 1$ with $A_{mk} \in \{0, 1\}$ for each $m = 1, \ldots, M$.
- $\sum_{l=1}^{L} D_{ln} = 1$ with $D_{ln} \in \{0, 1\}$ for each $n = 1, \ldots, N$.
- If the negative Bernoulli log-likelihood is used as loss function, $Z \in [0, 1]$.

2.6 BiArchetype Analysis

References: [**?**, **?**, **?**]

BiAA is a special case of SBM ($Z_{KL} = B_{KM} X_{MN} C_{NL}$).

$$\arg\max_{A,B,C,D} l(ABXCD; X) \tag{3}$$

such that:

- $\sum_{k=1}^{K} A_{mk} = 1$ with $A_{mk} \in [0, 1]$ for each $m = 1, \ldots, M$.
- $\sum_{m=1}^{M} B_{km} = 1$ with $B_{km} \in [0, 1]$ for each $k = 1, \ldots, K$.
- $\sum_{n=1}^{N} C_{nl} = 1$ with $D_{nl} \in [0, 1]$ for each $l = 1, \ldots, L$.
- $\sum_{l=1}^{L} D_{ln} = 1$ with $D_{ln} \in [0, 1]$ for each $n = 1, \ldots, N$.

2.6.1 Hard assignment

$$\arg\max_{A,B,C,D} l(ABXCD; X) \tag{4}$$

such that:

1. $\sum_{k=1}^{K} A_{mk} = 1$ with $A_{mk} \in \{0, 1\}$ for each $m = 1, \ldots, M$.
2. $\sum_{m=1}^{M} B_{km} = 1$ with $B_{km} \in [0, 1]$ for each $k = 1, \ldots, K$.
3. $\sum_{n=1}^{N} C_{nl} = 1$ with $D_{nl} \in [0, 1]$ for each $l = 1, \ldots, L$.
4. $\sum_{l=1}^{L} D_{ln} = 1$ with $D_{ln} \in \{0, 1\}$ for each $n = 1, \ldots, N$.

NOTE: The prototypes are not in the boundary of the CH. References: [**?**]

Change restrictions 2, 3 such that

3 Results

Normalized Mutual Information: [**?**]

3 When Bernoulli likelihood is used for SBM synthetic graphs, the Z matrix of the SBM and the Z matrix ($Z = BXC$) of the BiAA are equal to the probabilities used to construct the graph.

3.1 SBM-based synthetic data

3.2 BiAA-based synthetic data

3.3 Real datasets

3.3.1 Restaurant recommender system

[**?**] Source: https://archive.ics.uci.edu/ml/datasets/Restaurant+%2526+consumer+data

3.3.2 Drug side-effect association

[**?**] Source: https://snap.stanford.edu/biodata/datasets/10018/10018-ChSe-Decagon.html

3.3.3 NIPS author collaboration

3.4 Discussion

3.5 Conclusions