Computational Statistics and Data Analysis 117 (2018) 32–44

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Network linear discriminant analysis☆

Wei Cai [a], Guoyu Guan [a,b], Rui Pan [c,*], Xuening Zhu [d], Hansheng Wang [d]

[a] Key Laboratory for Applied Statistics of the MOE, and School of Mathematics and Statistics, Northeast Normal University, Changchun, China
[b] School of Economics, Northeast Normal University, Changchun, China
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[d] Guanghua School of Management, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 15 September 2016
Received in revised form 18 July 2017
Accepted 22 July 2017
Available online 9 August 2017

Keywords:
Classification
Linear discriminant analysis
Misclassification rate
Network data

ABSTRACT

Linear discriminant analysis (LDA) is one of the most popularly used classification methods. With the rapid advance of information technology, network data are becoming increasingly available. A novel method called network linear discriminant analysis (NLDA) is proposed to deal with the classification problem for network data. The NLDA model takes both network information and predictive variables into consideration. Theoretically, the misclassification rate is studied and an upper bound is derived under mild conditions. Furthermore, it is observed that real networks are often sparse in structure. As a result, asymptotic performance of NLDA is also obtained under certain sparsity assumptions. In order to evaluate the finite sample performance of the newly proposed methodology, a number of simulation studies are conducted. Lastly, a real data analysis about Sina Weibo is also presented for illustration purpose.

1. Introduction

Linear discriminant analysis (LDA) is one of the most widely-used classification methods. Due to its simplicity and efficiency, LDA has attracted great attention in a number of fields, such as biomedical studies, face recognition, earth science and many others (Yu and Yang, 2001; Hand, 2006; Guo et al., 2007). To deal with different kinds of data, extensions of LDA widely exist in the past literatures. For instance, functional LDA is proposed where the predictive variables are curves or functions (James and Hastie, 2001). Various penalized LDA methods have been developed for multi-class classification (Witten and Tibshirani, 2011; Clemmensen et al., 2012). Recent studies of LDA mainly concentrate on high dimensional data, including feature screening (Fan and Fan, 2008; Pan et al., 2016) and sparse estimation (Shao et al., 2011).

All aforementioned classification methods are mainly developed under the assumption that all the individuals are mutually independent. With the rapid advance of information technology, relational information among individuals can be easily collected (e.g., friendship, kinship and common interest). As one typical relational information, network data are

☆ The research of Wei Cai and Guoyu Guan is supported in part by National Natural Science Foundation of China (NSFC, 11501093, 11690012), China Postdoctoral Science Foundation Funded Project (Grant No. 2015M581378), and the Fundamental Research Funds for the Central Universities (Grant Nos. 2412015KJ028, 2412017FZ030). The research of Rui Pan is supported in part by National Natural Science Foundation of China (NSFC, 11601539). The research of Xuening Zhu and Hansheng Wang is supported in part by National Natural Science Foundation of China (NSFC, 71532001, 11525101) and Center for Statistical Science at Peking University. The research of all the authors is supported by the Fundamental Research Funds for the Central Universities (Grant Nos. 130028613, 130028729), and National Natural Science Foundation of China (NSFC, 11631003).

* Corresponding author.
E-mail address: panrui_cufe@126.com (R. Pan).

http://dx.doi.org/10.1016/j.csda.2017.07.007
0167-9473/

becoming increasingly available. A network refers to a group of individuals and the corresponding relationships among them. In the past decades, there are abundant literatures on model-based statistical analysis of network data. These models include but are not limited to the ER model (Erdös and Rényi, 1959), the p_1 model (Holland, and Leinhardt, 1981; Wasserman and Pattison, 1996; Robins et al., 2007), the stochastic blockmodel (Holland et al., 1983; Nowicki and Snijders, 2001; Karrer and Newman, 2011), and the latent space model (Hoff et al., 2002; Sewell and Chen, 2015). As one can see, existing statistical classification methods are no longer appropriate for network data since individuals are correlated with each other.

For network data, classification methods firstly arise in the field of machine learning, where *collective classification* (CC) is popularly used (Neville and Jensen, 2000; Taskar et al., 2002; McDowell et al., 2007). The spirit of collective classification is to make use of the information collected from one's neighbors when predicting one particular individual's class label. As a result, network structure can be taken into consideration and the resulting prediction performance can be enhanced. However, collective classification has three main disadvantages. First of all, due to its complex model setup, the results of collective classification are lack of interpretation. Secondly, although there are a number of applications of collective classification, its theoretical properties are not clear. Thirdly, most collective classification methods rely on iterative algorithms which lead to high computational cost. This motivates us to develop a novel statistical classification model for network data.

In this paper, we propose a new methodology called network linear discriminant analysis (NLDA). It makes use of both the predictive variables and network structure. As a result, relational information can be incorporated, which leads to improved prediction performance compared with traditional LDA. Furthermore, certain sparsity assumptions are imposed for large-scale network data. This makes the computation of NLDA feasible when the network size is huge. At the same time, the newly proposed NLDA method possesses excellent theoretical properties in terms of misclassification rate. Under mild assumptions, the method of NLDA outperforms that of LDA for different kinds of network structures. Lastly, the classification ability with information only from network structure is also investigated.

The rest of this article is organized as follows. In Section 2, we introduce the NLDA model and establish its theoretical properties. In Section 3, sparse networks are considered and the corresponding asymptotic results are derived. A number of numerical studies are conducted in Section 4 to demonstrate the finite sample performance of our newly proposed methodology. A real data analysis is also presented for illustration purpose. Some concluding remarks are given in Section 5. All the technical proofs are left in Appendix.

2. Network linear discriminant analysis

2.1. Model and notations

To describe the network structure, define an adjacency matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{n\times n}$, where $a_{i_1 i_2} = 1$ if the i_1th node follows the i_2th node, and $a_{i_1 i_2} = 0$ otherwise. We follow the tradition and let $a_{ii} = 0$ for $1 \le i \le n$. In addition, let (Y_i, X_i) be the observation collected from the ith node, where $Y_i \in \{0, 1\}$ is the binary class label with $P(Y_i = k) = \pi_k$, and $\pi_0 + \pi_1 = 1$. Furthermore, $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ is the associated p-dimensional predictor. Given $Y_i = k \in \{0, 1\}$, X_i is assumed to follow a p-dimensional multivariate normal distribution with mean $\mu_k = (\mu_{k1}, \ldots, \mu_{kp})^\top \in \mathbb{R}^p$ and covariance $\Sigma = (\sigma_{j_1 j_2}) \in \mathbb{R}^{p\times p}$. For convenience, we write $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ and $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$. Theoretically, the adjacency matrix is assumed to be random and generated according to some probability distribution. Specifically, we assume that conditional on $\mathbb{Y}$ and $\mathbb{X}$, different edges (i.e., $a_{i_1 i_2}$ s) are mutually independent with

$$P(a_{i_1 i_2} = 1|\mathbb{Y}, \mathbb{X}) = P(a_{i_1 i_2} = 1|Y_{i_1} = k_1, Y_{i_2} = k_2) = \omega_{k_1 k_2}, \tag{1}$$

where $\omega_{k_1 k_2} \in (0, 1)$ is the link probability from class k_1 to class k_2. In addition, let $\omega = (\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top \in \mathbb{R}^4$.

Traditional LDA predicts the class label of node i by maximizing the posterior probability $P(Y_i = k|X_i)$. It can be easily proved that this probability is proportional to $\pi_k \exp(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i)$. However, given the network structure A, we are able to employ not only the nodal information from node i but also the information from its connected nodes. Then, the corresponding prediction problem becomes maximizing $P(Y_i|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$, where $\mathbb{Y}_{(-i)} = (Y_{i'} : i' \ne i)^\top \in \mathbb{R}^{n-1}$. By assuming (1), i.e., the network structure A and the nodal covariates $\mathbb{X}$ are conditionally independent given the class labels $\mathbb{Y}$, we have

$$\begin{aligned} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A) &= P(Y_i = k)P(X_i|Y_i = k)P(A|\mathbb{Y}_{(-i)}, Y_i = k) \propto \pi_k \exp\left(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i\right) \\ &\times \prod_{j\ne i}\prod_{l}\left\{(\omega_{lk})^{a_{ji}}(1-\omega_{lk})^{1-a_{ji}}(\omega_{kl})^{a_{ij}}(1-\omega_{kl})^{1-a_{ij}}\right\}^{I(Y_j=l)}, \end{aligned} \tag{2}$$

where some constants independent of k are ignored and $I(\cdot)$ is the indicator function. We denote the optimal prediction of Y_i as $Y_i^* = \arg\max_{k\in\{0,1\}} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$. Regarding (2), we have the following three remarks.

Remark 1. Note that the right hand side of (2) consists of two components. One is in proportion to the posterior probability $P(Y_i = k|X_i)$, i.e., $\pi_k \exp(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i)$. If the network information is ignored, this is just the result derived from the traditional LDA. The other component is due to the network information.

Remark 2. If the covariances of two classes (i.e., Σ_1 and Σ_2) are different, the method of quadratic discriminant analysis (QDA) should be applied. As an extension, we could incorporate the network information into QDA as well, which leads to the network quadratic discriminant analysis (NQDA). Similar to NLDA, the posterior probability of NQDA can be derived as $P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A) \propto \pi_k|\Sigma_k|^{-1/2}\exp\{-2^{-1}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\}\prod_{j\neq i}\prod_l\{(\omega_{lk})^{a_{ji}}(1-\omega_{lk})^{1-a_{ji}}(\omega_{kl})^{a_{ij}}(1-\omega_{kl})^{1-a_{ij}}\}^{I(Y_j=l)}$. As one can see, this formula consists of two components as well, i.e., the traditional QDA part and the network structure part, which is identical to the second part of (2).

Remark 3. Although we only discuss the binary classification problem in this study, the proposed method can be easily extended to the multi-class case. Specifically, one need to compare more than two posterior probabilities when predicting the class label. Next, the label with the maximum posterior probability is determined as the predicted class, i.e., $Y_i^* = \arg\max_{1\leq k\leq K} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$ with $K > 2$, which can be similarly derived as the binary case.

2.2. Network linear discriminant analysis

For convenience, denote $\mathbf{E}_i = \{a_{ji}, a_{ij} : j \neq i\}$ as the collection of adjacent relationships of node i. By taking logarithm of the right hand side of (2), we get the *network linear discriminant function* as

$$\delta_k(X_i, \mathbf{E}_i) = \delta_k^{LDA}(X_i) + \delta_k^N(\mathbf{E}_i), \quad k = 0, 1, \tag{3}$$

where $\delta_k^{LDA}(X_i) = \log\pi_k - 2^{-1}\mu_k^\top\Sigma^{-1}\mu_k + \mu_k^\top\Sigma^{-1}X_i$ is the linear discriminant function of traditional LDA and

$$\delta_k^N(\mathbf{E}_i) = \sum_{l=0}^{1}\left(d_{+i}^{(l)}\log\frac{\omega_{lk}}{1-\omega_{lk}} + d_{i+}^{(l)}\log\frac{\omega_{kl}}{1-\omega_{kl}}\right) + \sum_{l=0}^{1} n_i^{(l)}\log\Big\{(1-\omega_{lk})(1-\omega_{kl})\Big\}.$$

More specifically, $d_{+i}^{(l)} = \sum_{j\neq i} a_{ji}I(Y_j = l)$ is node i's in-degree associated with class l (i.e., the number of i's followers in class l), $d_{i+}^{(l)} = \sum_{j\neq i} a_{ij}I(Y_j = l)$ is node i's out-degree associated with class l (i.e., the number of nodes followed by i in class l) and $n_i^{(l)} = \sum_{j\neq i} I(Y_j = l)$ represents the number of nodes in class l except for node i. Note that $n_i^{(1)} + n_i^{(0)} = n-1$ for each $1 \leq i \leq n$. As one can see, $\delta_k^N(\mathbf{E}_i)$ can be regarded as the network-based linear discriminant function. It is influenced by node i's in-degree and out-degree associated with each class and the size of each class. As a result, we refer $\delta_k(X_i, \mathbf{E}_i)$ in (3) as network linear discriminant function. The optimal prediction of Y_i can be equivalently written as $G(X_i, \mathbf{E}_i) = I(\delta_1(X_i, \mathbf{E}_i) \geq \delta_0(X_i, \mathbf{E}_i)) =$

$$I\left(\sum_{j\neq i} Z_{ij} + \log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1+\mu_0)^\top\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1-\mu_0)^\top\Sigma^{-1}X_i \geq 0\right), \tag{4}$$

where $Z_{ij} =$

$$\sum_{k=0}^{1}\left\{a_{ji}\log\frac{\omega_{k1}(1-\omega_{k0})}{\omega_{k0}(1-\omega_{k1})} + a_{ij}\log\frac{\omega_{1k}(1-\omega_{0k})}{\omega_{0k}(1-\omega_{1k})} + \log\frac{(1-\omega_{k1})(1-\omega_{1k})}{(1-\omega_{k0})(1-\omega_{0k})}\right\}I(Y_j = k)$$

can be regarded as node j's impact on the discrimination of node i. More specifically, assume link probabilities of inter-class are larger than that of intra-class (i.e., $\min\{\omega_{11}, \omega_{00}\} > \max\{\omega_{10}, \omega_{01}\}$). If node j is from class 1, then the terms $\log\frac{\omega_{11}(1-\omega_{10})}{\omega_{10}(1-\omega_{11})}$ and $\log\frac{\omega_{11}(1-\omega_{01})}{\omega_{01}(1-\omega_{11})}$ in Z_{ij} are both positive. As a result, if node i follows node j and/or node j follows node i (i.e., $a_{ij} = 1$ and/or $a_{ji} = 1$), node i is more likely to be classified to class 1.

As long as we define the network linear discriminant function, we next are particularly interested in evaluating the prediction accuracy of NLDA. More specifically, we address this problem from the Bayesian perspective. To this end, denote all the unknown parameters by $\boldsymbol{\theta} = (\pi^\top, \mu^\top, \mathrm{vec}(\Sigma)^\top, \omega^\top)^\top \in \mathbb{R}^{p^2+2p+6}$, where $\pi = (\pi_1, \pi_0)^\top \in \mathbb{R}^2$, $\mu = (\mu_1^\top, \mu_0^\top)^\top \in \mathbb{R}^{2p}$ and $\mathrm{vec}(\cdot)$ represents the vectorization of a matrix. First denote $G^{LDA}(X_i) = I(\delta_1^{LDA}(X_i) \geq \delta_0^{LDA}(X_i))$ as the optimal prediction by traditional LDA. Then define $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) = \pi_1 P(G(X_i, \mathbf{E}_i) = 0|Y_i = 1) + \pi_0 P(G(X_i, \mathbf{E}_i) = 1|Y_i = 0)$ and $R(G^{LDA}(X_i), \boldsymbol{\theta}) = \pi_1 P(G^{LDA}(X_i) = 0|Y_i = 1) + \pi_0 P(G^{LDA}(X_i) = 1|Y_i = 0)$ as the theoretical misclassification rate of NLDA and LDA methods respectively (Bickel and Levina, 2004; Fan and Fan, 2008).

In order to study misclassification rate of NLDA, we assume that there exists a positive constant ϵ, such that

$$\min_{k,l\in\{0,1\},k\neq l}\Big\{|\omega_{kl} - \omega_{kk}|, |\omega_{kl} - \omega_{ll}|\Big\} \geq \epsilon. \tag{5}$$

This condition means that the absolute difference of link probabilities between inter-class (i.e., ω_{11} and ω_{00}) and intra-class (i.e., ω_{10} and ω_{01}) should stay away from 0 with a moderate margin. Intuitively, this indicates that the social relationship between any pair of nodes depends on whether they are in the same class or not. Furthermore, from (5) it can be seen that the network is dense, i.e., $\max_{k,l\in\{0,1\}}\omega_{kl} = O(1)$. The probability that two nodes are connected is a constant even if the network is expanding. We then have the following theorem.

Theorem 1. *Assume that all the link probabilities are finite constants satisfying (5). We then have the following results.*

(a) *If there is no network information, the theoretical misclassification rate of NLDA is identical to that of LDA, i.e.,* $R(G^{LDA}(X_i), \boldsymbol{\theta})$. *It can be derived as* $\pi_1\Phi(-\mu_1^*/\Delta) + \pi_0\Phi(\mu_0^*/\Delta)$, *which is a constant and does not change along with network size n.*

(b) *An upper bound of the theoretical misclassification rate of NLDA, i.e.,* $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta})$, *can be obtained as*

$$\pi_1\left\{\Phi\left(\frac{(1-n)\mu_{Z|1} - 2\mu_1^*}{2\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|1}^2}{\sigma_{Z|1}^2 + 6^{-1}\mu_{Z|1}(\xi + \mu_{Z|1})}\right)\right\} + \pi_0\left\{\Phi\left(\frac{(n-1)\mu_{Z|0} + 2\mu_0^*}{2\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|0}^2}{\sigma_{Z|0}^2 - 6^{-1}\mu_{Z|0}(\xi - \mu_{Z|0})}\right)\right\}, \tag{6}$$

where $\mu_{Z|k} = E(Z_{ij}|Y_i = k)$, $\sigma_{Z|k}^2 = var(Z_{ij}|Y_i = k)$, $\Delta = \{(\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 - \mu_0)\}^{1/2}$, $\mu_1^* = \log(\pi_0^{-1}\pi_1) + 2^{-1}\Delta^2$, $\mu_0^* = \log(\pi_0^{-1}\pi_1) - 2^{-1}\Delta^2$, ξ *is a function of link probabilities given in Appendix A.3 and* $\Phi(\cdot)$ *is the cumulative distribution function of standard normal distribution. Furthermore, it can be proved that* $\lim_{n\to\infty} R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) \to 0$.

The proof of Theorem 1 is given in Appendix A.3. Regarding Theorem 1, we have the following two remarks.

Remark 4. For sufficiently large n, it can be easily shown that $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) \leq R(G^{LDA}(X_i), \boldsymbol{\theta})$. As a result, incorporating the information of network structure into traditional LDA indeed improves the prediction accuracy. This can be further demonstrated by the numerical studies in Section 4.

Remark 5. It is not difficult to rewrite $\sum_{j\neq i} Z_{ij}$ in (4) as $d_{+i}^{(1)} \log\frac{\omega_{11}}{\omega_{10}} + (n_i^{(1)} - d_{+i}^{(1)}) \log\frac{1-\omega_{11}}{1-\omega_{10}} + d_{i+}^{(1)} \log\frac{\omega_{11}}{\omega_{01}} + (n_i^{(1)} - d_{i+}^{(1)}) \log\frac{1-\omega_{11}}{1-\omega_{01}} + d_{+i}^{(0)} \log\frac{\omega_{01}}{\omega_{00}} + (n_i^{(0)} - d_{+i}^{(0)}) \log\frac{1-\omega_{01}}{1-\omega_{00}} + d_{i+}^{(0)} \log\frac{\omega_{10}}{\omega_{00}} + (n_i^{(0)} - d_{i+}^{(0)}) \log\frac{1-\omega_{10}}{1-\omega_{00}}$, which is the weighted sum of the number of node i's friends (in-degree and out-degree) and non-friends from two classes. From the perspective of geometry, it can be regarded as the intercept of the personalized discriminant hyperplane of node i. To get a more intuitional insight of the effect of network structure on misclassification rate, we take $d_{+i}^{(1)} \log\frac{\omega_{11}}{\omega_{10}}$ for instance, where $d_{+i}^{(1)}$ represents node i's in-degree associated with class 1, and $\log(\omega_{10}^{-1}\omega_{11})$ represents the strength for a node in class 1 following a node in class 1 relative to class 0. Assume any two nodes in the same class are more likely to be connected, i.e., $\min\{\omega_{11}, \omega_{00}\} > \max\{\omega_{10}, \omega_{01}\}$, then $\log(\omega_{10}^{-1}\omega_{11}) > 0$. Thus, if node i has more followers in class 1 (i.e., larger $d_{+i}^{(1)}$), it is more likely to classify node i to class 1. As a result, the misclassification rate of NLDA becomes smaller. In addition, under other settings of ω, it can also be verified that incorporating network information (i.e., $\sum_{j\neq i} Z_{ij}$) can indeed reduce the misclassification rate.

3. Sparse NLDA

According to our practical experience, real social networks are typically large in scale and extremely sparse in structure. In order to better understand the impact of network structure on classification, we consider sparse NLDA in this section. Assume that there exist three positive constants c_1, c_2 and γ, such that

$$c_1 n^{-\gamma} \leq \omega_{kl} \leq c_2 n^{-\gamma} \quad \text{for any } k,\ l \in \{0, 1\}. \tag{7}$$

Under this assumption, the link probability $\omega_{kl} \to 0$ as $n \to \infty$, which guarantees the sparsity property for large-scale networks. We next discuss the reasonable choice for γ. Simple calculation shows that $E(a_{ij}) = \sum_{k=0}^{1}\sum_{l=0}^{1} \omega_{kl}\pi_k\pi_l = O(n^{-\gamma})$. So the expected out-degree of node i is $E(d_{i+}) = O(n^{1-\gamma})$, where $d_{i+} = \sum_{j\neq i} a_{ij}$. It is clear that if $\gamma > 1$, $E(d_{i+}) \to 0$ as $n \to \infty$. This is obviously impractical, because in real network platform, the number of nodes followed by one particular node should not diminish as the network size becomes larger. As a consequence, $0 < \gamma \leq 1$ is sensible. We next discuss two scenarios regarding $0 < \gamma \leq 1$.

Scenario 1 ($\gamma = 1$). When $\gamma = 1$, we have $E(d_{i+}) = O(1)$. This suggests that, as the network size increases to infinity, the nodal out-degree should be bounded by a finite constant. This is fairly reasonable, because some network platforms impose an upper bound for nodal out-degree. For example, Sina Weibo (www.weibo.com), the largest Twitter-type social media in China, allows each user to follow no more than 2000 other users (i.e., $d_{i+} \leq 2000$). Hence, $\gamma = 1$ is of great importance in both theory and application.

Scenario 2 ($0 < \gamma < 1$). In the second scenario, we consider $0 < \gamma < 1$. It can be easily verified that $E(d_{i+}) \to \infty$ as $n \to \infty$. This indicates that as the network grows, the expected value of nodal out-degree expands with the network. As a result, we should also pay attention to $0 < \gamma < 1$. Similar results can also be derived for nodal in-degree, i.e., $d_{+i} = \sum_{j\neq i} a_{ji}$.

Throughout the rest of the article, we assume that $0 < \gamma \leq 1$. Under sparsity assumption (7), it is desirable to study the misclassification rate of NLDA method. To this end, assume that there exists a positive constant c_3, such that

$$\min_{k,l\in\{0,1\},k\neq l}\left\{|\omega_{kl} - \omega_{kk}|, |\omega_{kl} - \omega_{ll}|\right\} \geq c_3 n^{-\gamma}. \tag{8}$$

This condition means that the link probabilities between inter-class and intra-class at least have a small distance in the order of $O(n^{-\gamma})$. Then we have the following theorem.

Theorem 2. *Under the assumptions (7) and (8), we have*

(a) *For $0 < \gamma < 1$, an upper bound of $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta})$ is obtained as*

$$\pi_1\left\{\Phi\left(\frac{(1-n)\mu_{Z|1} - 2\mu_1^*}{2\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|1}^2}{\sigma_{Z|1}^2 + 6^{-1}\mu_{Z|1}(\xi + \mu_{Z|1})}\right)\right\} + \pi_0\left\{\Phi\left(\frac{(n-1)\mu_{Z|0} + 2\mu_0^*}{2\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|0}^2}{\sigma_{Z|0}^2 - 6^{-1}\mu_{Z|0}(\xi - \mu_{Z|0})}\right)\right\}. \tag{9}$$

(b) *For $\gamma = 1$, if $\sigma_{Z|1}^{-1}\mu_{Z|1} > \Delta^{-1}\mu_1^*$ and $\sigma_{Z|0}^{-1}\mu_{Z|0} < \Delta^{-1}\mu_0^*$,*

$$\lim_{n\to\infty} R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) < R(G^{LDA}(X_i), \boldsymbol{\theta}).$$

The proof of Theorem 2 is left to Appendix A.4. Recall that μ_k^*, Δ, $\mu_{Z|k}$, $\sigma_{Z|k}^2$ and ξ are defined in Theorem 1 with $k \in \{0, 1\}$. Furthermore, it is proved in Lemma 2 in Appendix A.2 that μ_k^*, Δ and ξ are constants, while $\mu_{Z|k}$ and $\sigma_{Z|k}^2$ are $O(n^{-\gamma})$. There are two main results in Theorem 2. First of all, the upper bound (9) given in (a) approaches 0 as $n \to \infty$, which means that $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta})$ tends to 0 as $n \to \infty$. As a result, when $\gamma < 1$, $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) \le R(G^{LDA}(X_i), \boldsymbol{\theta})$ for large enough n. Secondly, when $\gamma = 1$, the order of $\mu_{Z|k}$ and $\sigma_{Z|k}^2$ are $O(n^{-1})$. The assumptions $\sigma_{Z|1}^{-1}\mu_{Z|1} > \Delta^{-1}\mu_1^*$ and $\sigma_{Z|0}^{-1}\mu_{Z|0} < \Delta^{-1}\mu_0^*$ in (b) can be regarded as signal-to-noise ratios which guarantees that the network can provide more information than the predictor. As a result, we obtain that $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) < R(G^{LDA}(X_i), \boldsymbol{\theta})$ as $n \to \infty$. To sum up, taking network information into account can improve prediction accuracy in terms of the theoretical misclassification rate in sparse networks.

In real practice, it occurs occasionally that only network structure can be obtained. This inspires us to develop a new discriminant method with information only from the network structure. According to (3), we define the pure network discriminant rule as

$$G^N(\mathbf{E}_i) = I\big(\delta_1^N(\mathbf{E}_i) \ge \delta_0^N(\mathbf{E}_i)\big), \tag{10}$$

which is referred to as the Pure Network Linear Discriminant Analysis (PNLDA). The following theorem shows that PNLDA could be an alternative to NLDA under certain conditions.

Theorem 3. *Assume $0 < \gamma \le 1$, under the assumptions (7) and (8), we have*

(a) *For $\gamma = 1$,*

$$P(G(X_i, \mathbf{E}_i) \ne G^N(\mathbf{E}_i)) \le \pi_1\{2\pi(n-1)\}^{-1/2}\sigma_{Z|1}^{-1}\{1 - 2\Phi(-\mu_1^*/\Delta)\} + \pi_0\{2\pi(n-1)\}^{-1/2}\sigma_{Z|0}^{-1}\{1 - 2\Phi(-\mu_0^*/\Delta)\} + O(n^{-1/2}). \tag{11}$$

(b) *For $0 < \gamma < 1$, $\lim_{n\to\infty} P(G(X_i, \mathbf{E}_i) \ne G^N(\mathbf{E}_i)) \to 0$.*

The proof of Theorem 3 is left to Appendix A.5. Recall that $\sigma_{Z|k}$ is defined in Theorem 1, and it is proved in Lemma 2 that $\sigma_{Z|k}$ is of $O(n^{-\gamma/2})$ for $k \in \{0, 1\}$. As a result, by (a) in Theorem 3, when $\gamma = 1$ the right hand side of (11) is a constant. While as shown in (b), the probability tends to 0 as $n \to \infty$.

4. Numerical studies

4.1. Simulation models

To evaluate the finite sample performance of the newly proposed method, we present a number of simulation examples. There are two main differences in these examples. The first lies in the network structure, which reflects on the generating mechanisms of the link probability vector ω and the resulting adjacency matrix A. The second is the design of the nodal predictor $\mathbb{X}$. In each model setup, both balanced and unbalanced case of simulating $\mathbb{Y}$ are considered. More specifically, in balanced case $\pi_0 = \pi_1 = 0.5$, while in unbalanced case $\pi_0 = 0.9$ and $\pi_1 = 0.1$.

Example 1 (*Homophily*). Recall $\omega_{k_1k_2}$ is the link probability from class k_1 to k_2 ($k_1, k_2 \in \{0, 1\}$). We first study the phenomenon of "homophily". In this scenario, the nodes have higher probability to connect if they belong to the same class, i.e., $\min\{\omega_{00}, \omega_{11}\} > \max\{\omega_{01}, \omega_{10}\}$. Accordingly, we set $\omega = (\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top = (5\rho n^{-\gamma}, 2\rho n^{-\gamma}, 2\rho n^{-\gamma}, 5\rho n^{-\gamma})^\top$, where the positive constants ρ and γ can control the network density (i.e., $n^{-1}(n-1)^{-1}\sum_{i,j} a_{ij}$). Furthermore, we follow Witten and Tibshirani (2011) to generate the predictors $\mathbb{X}$. More specifically, given the class label $Y_i = k$ ($k \in \{0, 1\}$), the nodal predictors $X_i = (X_{i1}, \ldots, X_{i5})^\top \in \mathbb{R}^5$ are independently simulated from a multivariate normal distribution with mean μ_k and covariance $\Sigma = (\sigma_{j_1j_2})$, where $\sigma_{j_1j_2} = 0.5^{|j_1 - j_2|}$, $\mu_0 = (0, \ldots, 0)^\top$ and $\mu_1 = (1, -1, 1, -1, 1)^\top$.

Example 2 (*Heterophily*)**.** Next, we study the phenomenon of "heterophily". In this case, the link probabilities between different classes are higher, i.e., $\min\{\omega_{01}, \omega_{10}\} > \max\{\omega_{00}, \omega_{11}\}$. Accordingly, we set $(\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top = (\rho n^{-\gamma}, 3\rho n^{-\gamma}, 3\rho n^{-\gamma}, \rho n^{-\gamma})^\top$. In this case, we follow Guo et al. (2007) and generate $\mathbb{X}$ with different covariance matrix for different classes. Given the class label $Y_i = k\,(k \in \{0, 1\})$, the nodal predictors $X_i = (X_{i1}, \ldots, X_{i5})^\top \in \mathbb{R}^5$ are independently simulated from a multivariate normal distribution with mean μ_k and covariance Σ_k, where $\mu_0 = (0, \ldots, 0)^\top$, $\mu_1 = (1, -1, 1, -1, 1)^\top$, $\Sigma_1 = 0.5^{|j_1 - j_2|}$, $\Sigma_2 = I$ and I is the five-dimension identity matrix.

Example 3 (*Core–periphery*)**.** Lastly, we present the phenomenon of "core–periphery". In this scenario, the link probability of nodes within "core" class is the largest and the link probability within "periphery" class is the smallest among all link probabilities. Without loss of generality, we assume class 1 to be "core" class and class 0 as "periphery" class. Then we set $(\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top = (5\rho n^{-\gamma}, 3\rho n^{-\gamma}, 3\rho n^{-\gamma}, 0.5\rho n^{-\gamma})^\top$. In the last case, we follow Fan and Fan (2008) to test the robustness of our methodology when nodal predictors are not from normal distribution. To be more specific, given the class label $Y_i = k\,(k \in \{0, 1\})$, the nodal predictors $X_i = (X_{i1}, \ldots, X_{i5})^\top \in \mathbb{R}^5$ are independently simulated from uniform distribution $U(-2, 1)$ and $U(-1, 2)$ respectively.

4.2. Performance measurements and simulation results

For each case, we conduct the simulation studies under a fixed network size $n = 3000$ and randomly partition the generated data into training set and testing set with different values of training proportion (i.e., $r = 0.3, 0.5, 0.7$). Let $n_0 = n * r$ represents the training sample size. The parameters are estimated on the training set and the test set is used for validation. We set different ρ values (i.e., $\rho = 0.1, 2$) to control network density. Besides, $\gamma = 0.5$ and $\gamma = 1$ are considered for the sake of comparison. The experiment is randomly repeated $T = 1000$ times for each model setup. For the tth replication ($1 \le t \le T$), let $\hat{\boldsymbol{\theta}}^{(t)} = (\hat{\pi}^{(t)\top}, \hat{\mu}^{(t)\top}, \mathrm{vec}(\hat{\Sigma}^{(t)})^\top, \hat{\omega}^{(t)\top})^\top$ be the MLE for the NLDA model (see details in Appendix A.1), and $A^{(t)} = (a_{ij}^{(t)}) \in \mathbb{R}^{n\times n}$ be the associated adjacency matrix. We then consider the following measurements to gauge the performance of different methods.

Firstly, to evaluate the estimation accuracy of MLE, we consider the root mean square errors (RMSE) for the estimated parameters as $\mathrm{RMSE}_\pi = \{T^{-1}\sum_{t=1}^T \|\hat{\pi}^{(t)} - \pi\|^2\}^{1/2}$, $\mathrm{RMSE}_\mu = \{T^{-1}\sum_{t=1}^T \|\hat{\mu}^{(t)} - \mu\|^2\}^{1/2}$, $\mathrm{RMSE}_\Sigma = \{T^{-1}\sum_{t=1}^T \|\hat{\Sigma}^{(t)} - \Sigma\|_F^2\}^{1/2}$ and $\mathrm{RMSE}_\omega = \{T^{-1}\sum_{t=1}^T \|\hat{\omega}^{(t)} - \omega\|^2\}^{1/2}$ on training set, where $\|\cdot\|_F^2$ is the Frobenius norm of a matrix. Secondly, the randomly generated testing sample is applied to evaluate the prediction performance of the proposed method. Take a testing node i' as an example. In the tth replication, we generate its class label $Y_{i'}^{(t)}$ and the nodal predictor $X_{i'}^{(t)}$ according to NLDA model as previously. Then, the links between the testing node and existing n_0 nodes are generated according to (1), and collected by $\mathbf{E}_{i'}^{(t)} = \{a_{ii'}^{(t)}, a_{i'i}^{(t)}\}$, where $a_{ii'}^{(t)}$ and $a_{i'i}^{(t)}$ are the network links between the ith node in training set and the testing node in the tth replication. Given $\mathbf{E}_{i'}^{(t)}$ and the nodal predictor $X_{i'}^{(t)}$, we then make the discrimination about $Y_{i'}^{(t)}$ by four competing models. They are the traditional LDA model, the NLDA model (predicted by (4)), the PNLDA model (predicted by (10)) and the NQDA model (given by Remark 2 in Section 2.1). The discriminant accuracy is measured by the average misclassification error (AME), i.e., $\{(1-r)nT\}^{-1}\sum_{j=1}^{(1-r)n}\sum_{t=1}^T I(Y_j^{(t)} \neq G^*)$, where $G^* = G^{LDA}(X_j^{(t)})$, $G(X_j^{(t)}, \mathbf{E}_j^{(t)})$, $G^N(\mathbf{E}_j^{(t)})$, and $G^{NQDA}(X_j^{(t)}, \mathbf{E}_j^{(t)})$ for LDA, NLDA, PNLDA and NQDA respectively. Note that $G^{NQDA}(X_j^{(t)}, \mathbf{E}_j^{(t)})$ can be easily obtained similar to $G(X_j^{(t)}, \mathbf{E}_j^{(t)})$. Lastly, the average network density is reported on training set, i.e., $\mathrm{ND} = T^{-1}\sum_{t=1}^T\{n_0(n_0-1)\}^{-1}\sum_{i,j} a_{ij}^{(t)}$.

The simulation results for both the balanced and unbalanced cases are summarized in Tables 1 and 2 respectively. As we can see, the results of the three examples in two tables are quite similar. First, we find that for each setting of γ and ρ, the RMSE values are decreasing as the training sample size n_0 gets larger. This implies the parameters can be consistently estimated. Besides, the network is increasingly sparse as the training sample size n_0 increases. Lastly, with regard to the prediction accuracy, it can be seen that the AME for NLDA improves greatly with the training sample size n_0, but the AME for LDA improves slightly, which is consistent with the theoretical results in Theorem 2(a), i.e., the asymptotic misclassification rate for LDA converges to a strictly positive number while that for NLDA converges to zero for sparse network with $\gamma < 1$. For example, for "homophily" network with balanced classes (i.e., Table 1) and $r = 0.7$, the AME value of NLDA is 0.11% (at $\gamma = 0.5$ and $\rho = 0.1$), which is much lower than the AME of LDA (i.e., 3.60%). For $\gamma = 1$, NLDA also outperforms LDA which is consistent with the theoretical result in Theorem 2(b). In particular, the AME values of PNLDA are lower than LDA, and almost close to NLDA at $\gamma = 0.5$ as the training sample size n_0 increases. This corroborates with the results in Theorem 3. When the covariances of two classes are unequal, we find that the NQDA model outperforms other methods, which is reasonable. Besides, according to the results for "core–periphery" network, NLDA is still feasible when the normality assumption is violated.

In addition, we compare NLDA with the collective classification (CC) method proposed by Neville and Jensen (2000). Because of the expensive computational cost of CC, we conduct an extra simulation study under the balanced case of "homophily" with 200 replications. The results are given in Table 3, the performance is measured by AME and the CPU time is reported as well. The results in Table 3 show that NLDA method has higher accuracy with less CPU time.

Table 1
Simulation results for Examples 1–3 with 1000 replications for balanced case. The RMSE values ($\times 10^2$) are reported for $\boldsymbol{\theta}$ estimates. The AMEs (in %) of four competing models are reported to evaluate prediction accuracy. The network density is also reported as ND.

γ	ρ	r	RMSE ($\times 10^2$)				AME (%)				Density
			π	μ	Σ	ω	LDA	NLDA	PNLDA	NQDA	ND (%)
Homophily network											
0.5	0.1	0.3	1.741	4.687	3.781	0.024	3.622	0.342	2.151	0.347	1.167
		0.5	1.215	3.625	2.937	0.013	3.581	0.170	1.110	0.172	0.904
		0.7	1.097	3.033	2.490	0.008	3.577	0.112	0.642	0.113	0.763
1	2	0.3	1.741	4.687	3.781	0.019	3.622	0.744	4.995	0.751	0.778
		0.5	1.215	3.625	2.937	0.009	3.581	0.738	5.079	0.745	0.467
		0.7	1.097	3.033	2.490	0.006	3.577	0.741	5.069	0.748	0.333
Heterophily network											
0.5	0.1	0.3	1.753	4.671	22.75	0.018	8.580	1.335	3.688	1.108	0.667
		0.5	1.319	3.604	22.49	0.010	8.581	0.778	2.081	0.640	0.516
		0.7	1.056	3.095	22.38	0.006	8.593	0.518	1.351	0.428	0.437
1	2	0.3	1.753	4.671	22.75	0.015	8.580	2.460	7.307	2.074	0.445
		0.5	1.319	3.604	22.49	0.007	8.581	2.460	7.351	2.051	0.267
		0.7	1.056	3.095	22.38	0.004	8.593	2.490	7.378	2.067	0.191
Core–periphery network											
0.5	0.1	0.3	1.674	4.066	3.499	0.022	10.27	0.859	2.289	0.866	0.958
		0.5	1.300	3.145	2.681	0.011	10.22	0.456	1.153	0.458	0.742
		0.7	1.097	2.672	2.263	0.008	10.23	0.268	0.675	0.269	0.627
1	2	0.3	1.674	4.066	3.499	0.018	10.27	1.897	5.169	1.909	0.639
		0.5	1.300	3.145	2.681	0.008	10.22	1.885	5.161	1.892	0.383
		0.7	1.097	2.672	2.263	0.005	10.23	1.902	5.163	1.905	0.274

Table 2
Simulation results for Examples 1–3 with 1000 replications for unbalanced case. The RMSE values ($\times 10^2$) are reported for $\boldsymbol{\theta}$ estimates. The AMEs (in %) of four competing models are reported to evaluate prediction accuracy. The network density is also reported as ND.

γ	ρ	r	RMSE ($\times 10^2$)				AME (%)				Density
			π	μ	Σ	ω	LDA	NLDA	PNLDA	NQDA	ND (%)
Homophily network											
0.5	0.1	0.3	1.000	7.840	3.872	0.076	1.922	0.210	2.275	0.219	1.487
		0.5	0.793	6.229	2.962	0.040	1.886	0.108	1.128	0.110	1.152
		0.7	0.660	5.220	2.474	0.026	1.887	0.062	0.642	0.062	0.973
1	2	0.3	1.000	7.840	3.872	0.062	1.922	0.444	5.273	0.462	0.991
		0.5	0.793	6.229	2.962	0.030	1.886	0.425	5.250	0.438	0.595
		0.7	0.660	5.220	2.474	0.017	1.887	0.418	5.303	0.424	0.424
Heterophily network											
0.5	0.1	0.3	1.042	7.975	28.79	0.042	5.805	0.769	3.433	0.668	0.453
		0.5	0.754	6.102	28.58	0.021	5.720	0.437	1.829	0.373	0.351
		0.7	0.645	5.024	28.53	0.014	5.699	0.288	1.177	0.241	0.297
1	2	0.3	1.042	7.975	28.79	0.035	5.805	1.437	6.930	1.264	0.302
		0.5	0.754	6.102	28.58	0.016	5.720	1.425	7.280	1.236	0.181
		0.7	0.645	5.024	28.53	0.009	5.699	1.431	7.358	1.232	0.130
Core–periphery network											
0.5	0.1	0.3	1.012	6.807	3.499	0.077	4.730	0.139	0.644	0.143	0.332
		0.5	0.792	5.305	2.681	0.040	4.693	0.054	0.238	0.055	0.257
		0.7	0.652	4.488	2.263	0.026	4.715	0.023	0.109	0.023	0.217
1	2	0.3	1.012	6.807	3.499	0.064	4.730	0.437	1.958	0.449	0.221
		0.5	0.792	5.305	2.681	0.029	4.693	0.436	1.949	0.444	0.133
		0.7	0.652	4.488	2.263	0.017	4.715	0.425	1.933	0.428	0.095

4.3. A Sina Weibo dataset

We now illustrate a real data example from Sina Weibo, which is the largest Twitter-type social media in China. Specifically, $n = 4077$ nodes are collected from followers of an official MBA program. Four user generated labels are considered as, "Tsinghua", "Peking", "Cheung Kong" and "China Europe", which are all higher education institutions in China. Accordingly, the binary response Y_i is defined to be 1 if the ith user carries at least one of the above labels, otherwise $Y_i = 0$. To sum up, the number of nodes in Class 1 is 419. In addition, a number of nodal predictors are taken into consideration, which includes (1) the number of characters in personal labels (self-created by the users to describe their lifestyles), (2)

Table 3
Simulation results for Example 1 with 200 replications for balanced case. The AMEs (in %) of NLDA and CC are reported to evaluate prediction accuracy. The average CPU computing time is reported as well.

γ	ρ	r	AME (%)		CPU (s)	
			NLDA	CC	NLDA	CC
0.5	0.1	0.3	0.352	0.713	0.804	247.44
		0.5	0.183	0.469	0.907	464.40
		0.7	0.101	0.216	0.935	668.94
1	2	0.3	0.758	2.813	0.708	321.24
		0.5	0.733	1.886	0.904	597.36
		0.7	0.706	0.735	0.942	721.26

Table 4
The MLE of NLDA for Sina Weibo dataset.

	π	μ ($\times 10^2$)		
		Personal labels	Weibo posts	Tenure
Class 1	0.103	37.461	1.518	−3.464
Class 0	0.897	−0.174	−0.179	0.397

their cumulated number of Weibo posts, (3) the tenure (the time length since Weibo registration) measured in months. In addition, the Box–Cox transformation is conducted on all the predictors. In order to insure that all predictors are in the same scale, we also conduct the standardized transformation on all predictors. Lastly, the resulting network density is around 0.29%, which indicates a sufficiently sparse network.

The parameter estimation is conducted using the whole network and the MLEs of π and μ are reported in Table 4. The estimated π_1 value is 10.3%, which suggests an unbalanced class. Moreover, by the estimation result of μ, one could see that the network users in Class 1 tend to possess more personal labels and Weibo posts, but with slightly lower tenure. The estimated ω value is $(0.0491, 0.0127, 0.0069, 0.0006)^\top$, which indicates that the resulting network belongs to the "Core–Periphery" type network. Subsequently, we apply the five methods (i.e., LDA, NLDA, PNLDA, NQDA and CC) to the dataset to compare the prediction ability. Specifically, 70% of the nodes are randomly selected for model training, and the rest are used for prediction. To obtain a stable result, the experiment is repeated for $T = 100$ times. The resulting AMEs of the NLDA, PNLDA and NQDA model are 7.72%, 7.86% and 7.71%, which are very close and much smaller than that of the LDA model (i.e., 10.24%). Besides, the AME of CC method is 8.61%. This indicates a competitive prediction performance of the proposed NLDA, PNLDA and NQDA methods.

5. Concluding remarks

In this paper, we come up with a novel classification model named NLDA, which makes use of the information obtained from the network. After deriving the discriminant rule, an upper bound of the theoretical misclassification rate of NLDA is given, from which we can find that NLDA outperforms LDA for dense network, i.e., the results of Theorem 1. Next, we consider the sparse property of large-scale network and show that the theoretical misclassification rate of NLDA is smaller than LDA under different sparsity assumptions i.e., the results of Theorem 2. Besides, the PNLDA method is investigated in the situation that only the network information is available. With regard to the discriminant ability, it can be shown that the PNLDA model could be an alternative to NLDA under certain conditions, i.e., the results of Theorem 3. At last, we present several simulations which illustrate that NLDA always outperforms LDA with respect to AME. Similarly, for real application, we show that NLDA has better prediction performance for unbalanced datasets equipped with network structures. Note that, NLDA theoretically handles with predictors from normal distributions. Despite Example 3 in Section 4.1 shows that NLDA is still feasible when the predictors are from uniform distributions, we still provide some proposals when the predictors are from non-normal distributions. Lots of methods can be applied to test the normality, such as QQ-plot, Shapiro–Wilk W-test (Johnson and Wichern, 2014) and so on. If the normality assumption is violated, it is suggested transformations (e.g., Box–Cox transformation) can be conducted to make the predictors be more likely from the normal distribution.

In future work, it is of great interest to extend NLDA to the degree-corrected case, i.e., degree-corrected network linear discriminant analysis (DC-NLDA), which is inspired by the degree-corrected blockmodel (Karrer and Newman, 2011; Zhao et al., 2012; Qin and Rohe, 2013) for undirected network. Because of the limit of users' energy and restriction of Sina Weibo, the number of the users followed by each node (out-degree) is almost in the same scale, however there is no upper bound for followers (in-degree). Thus we define an in-degree strength coefficient parameter β_i ($1 \le i \le n$) which is positive and controls the expected in-degree of node i. Inspired by Jin (2015), we constrain that $\max_{1\le i\le n}\beta_i \le 1$. Then we assume that the link probability of a_{ij} not only depends on class labels of node i and node j but also on β_i. Based on above, we have

$P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A) \propto \pi_k \exp(-2^{-1}\mu_k^\top \Sigma^{-1}\mu_k + \mu_k^\top \Sigma^{-1}X_i)\prod_{j\neq i}\prod_l\{(\beta_i\omega_{lk})^{a_{ji}}(1-\beta_i\omega_{lk})^{1-a_{ji}}(\beta_j\omega_{kl})^{a_{ij}}(1-\beta_j\omega_{kl})^{1-a_{ij}}\}^{I(Y_j=l)}$. Note that the likelihood function cannot be directly factorized into the function of β_i $(1 \le i \le n)$ and the function of ω_{kl} $(k, l \in \{0, 1\})$, hence the maximum likelihood estimators (MLE) can be only obtained by iterative algorithm. Similarly, the optimal prediction of Y_i is $Y_i^* = \arg\max_{k\in\{0,1\}} P(Y_i = k|\mathbb{X}, \mathbb{Y}_{(-i)}, A)$. Due to the expensive computation cost of the iterative estimation algorithm, we have not applied this method in simulation studies and real example. We will further optimize the estimation algorithm and improve its practicability in future work.

Besides, some other directions are also appealing for further study. Firstly, as an extension of NLDA, the NQDA method has some distinct theoretical properties with NLDA, which is an attractive direction to be investigated in future work, especially when the network size is large. Secondly, the sparsity of network can be defined in various ways by different researchers. Thus, how to take into account different kinds of sparse networks and investigate the corresponding theoretical properties is also worth studying. Thirdly, although the proposed NLDA method is easy to extend to the multi-class classification problem (i.e., see Remark 3 in Section 2.1), it is noteworthy that the number of unknown parameters is increasing as the number of classes increases. Then how to enhance the estimation and computation efficiency in such situation can be an interesting topic to discuss.

Appendix

A.1. Derivation of MLE for NLDA

The likelihood function of NLDA model can be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= P\big(\mathbb{Y}, \mathbb{X}, A\big) = P\big(\mathbb{Y}\big)P\big(A|\mathbb{Y}, \mathbb{X}\big)P\big(\mathbb{X}|\mathbb{Y}\big)\\ &= \prod_{i=1}^{n}\pi_1^{Y_i}\pi_0^{1-Y_i}\prod_{i_1\neq i_2}\prod_{k_1,k_2}\Big\{(\omega_{k_1k_2})^{a_{i_1i_2}}(1-\omega_{k_1k_2})^{1-a_{i_1i_2}}\Big\}^{I(Y_{i_1}=k_1,Y_{i_2}=k_2)}\\ &\quad\times\prod_{i=1}^{n}(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\Big\{-\frac{1}{2}\sum_{k=0}^{1}I(Y_i=k)(X_i-\mu_k)^\top\Sigma^{-1}(X_i-\mu_k)\Big\}.\end{aligned}$$

Then, the maximum likelihood estimators (MLE) could be obtained as $\hat{\boldsymbol{\theta}} = (\hat{\pi}^\top, \hat{\mu}^\top, \mathrm{vec}(\hat{\Sigma})^\top, \hat{\omega}^\top)^\top = \arg\max_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$. After some calculations, the MLE can be shown as follows. For $k, l \in \{0, 1\}$, we obtain that $\hat{\pi}_k = n^{-1}\sum_{i=1}^{n}I(Y_i = k)$, $\hat{\omega}_{kl} = \{\sum_{i_1\neq i_2}I(Y_{i_1} = k, Y_{i_2} = l)\}^{-1}\sum_{i_1\neq i_2}a_{i_1i_2}I(Y_{i_1} = k, Y_{i_2} = l)$, $\hat{\mu}_k = \{\sum_{i=1}^{n}I(Y_i = k)\}^{-1}\sum_{i=1}^{n}I(Y_i = k)X_i$, and $\hat{\Sigma} = n^{-1}\sum_{k=0}^{1}\sum_{i=1}^{n}I(Y_i = k)(X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top$.

A.2. Technical lemmas

In order to prove the theoretical results , we first address some technical lemmas.

Lemma 1. *Under the assumption (5), given $Y_i = k \in \{0, 1\}$, we have*

$$\sqrt{n-1}\frac{(\bar{Z}_i - \mu_{Z|k})}{\sigma_{Z|k}} \to_d N(0, 1),$$

where $\bar{Z}_i = (n-1)^{-1}\sum_{j\neq i}Z_{ij}$, $\mu_{Z|k} = E(Z_{ij}|Y_i = k) = \pi_1^k(-\pi_0)^{1-k}\{g(\omega_{kk}, \omega_{10}) + g(\omega_{kk}, \omega_{01})\} + (-\pi_1)^{1-k}\pi_0^k\{g(\omega_{10}, \omega_{1-k,1-k}) + g(\omega_{01}, \omega_{1-k,1-k})\}$, $\sigma^2_{Z|k} = var(Z_{ij}|Y_i = k) = \pi_1^k\pi_0^{1-k}\{h(\omega_{kk}, \omega_{10}) + h(\omega_{kk}, \omega_{01})\} + \pi_1^{1-k}\pi_0^k\{h(\omega_{10}, \omega_{1-k,1-k}) + h(\omega_{01}, \omega_{1-k,1-k})\} - \mu^2_{Z|k} + 2\pi_1^k g(\omega_{kk}, \omega_{10})g(\omega_{kk}, \omega_{01}) + 2\pi_0^{1-k}g(\omega_{10}, \omega_{1-k,1-k})g(\omega_{01}, \omega_{1-k,1-k})$, and the notation "$\to_d$" represents convergence in distribution. Note that $g(a, b) \doteq a\log(b^{-1}a) + (1-a)\log\{(1-b)^{-1}(1-a)\}$ and $h(a, b) \doteq a\{\log(b^{-1}a)\}^2 + (1-a)[\log\{(1-b)^{-1}(1-a)\}]^2$.

Proof. This proof consists of three steps. In step 1, we show the independence of Z_{ij} s, given node i's class label Y_i. Next, the conditional expectation and variance of Z_{ij} are calculated in step 2. Lastly, we take step 3 to derive the asymptotic normality of $\bar{Z}_i$.

Step 1. We only consider the case of $Y_i = 1$ because of the symmetry of $Y_i = 1$ and $Y_i = 0$. Let

$$\begin{aligned}
Z_{ij} &= a_{ji}Y_j \log\frac{\omega_{11}}{\omega_{10}} + (1-a_{ji})Y_j \log\frac{1-\omega_{11}}{1-\omega_{10}} + a_{ij}Y_j \log\frac{\omega_{11}}{\omega_{01}} + (1-a_{ij})Y_j \log\frac{1-\omega_{11}}{1-\omega_{01}} \\
&\quad + a_{ji}(1-Y_j)\log\frac{\omega_{01}}{\omega_{00}} + (1-a_{ji})(1-Y_j)\log\frac{1-\omega_{01}}{1-\omega_{00}} \\
&\quad + a_{ij}(1-Y_j)\log\frac{\omega_{10}}{\omega_{00}} + (1-a_{ij})(1-Y_j)\log\frac{1-\omega_{10}}{1-\omega_{00}} \\
&= \tilde{z}_{j1}\log\frac{\omega_{11}}{\omega_{10}} + \tilde{z}_{j2}\log\frac{1-\omega_{11}}{1-\omega_{10}} + \tilde{z}_{j3}\log\frac{\omega_{11}}{\omega_{01}} + \tilde{z}_{j4}\log\frac{1-\omega_{11}}{1-\omega_{01}} \\
&\quad + \tilde{z}_{j5}\log\frac{\omega_{01}}{\omega_{00}} + \tilde{z}_{j6}\log\frac{1-\omega_{01}}{1-\omega_{00}} + \tilde{z}_{j7}\log\frac{\omega_{10}}{\omega_{00}} + \tilde{z}_{j8}\log\frac{1-\omega_{10}}{1-\omega_{00}} \\
&= z_{j1} + z_{j2} + z_{j3} + z_{j4} + z_{j5} + z_{j6} + z_{j7} + z_{j8}.
\end{aligned}$$

Given $Y_i = 1$, for any $j \neq q \neq i$, we can obtain that

$$\begin{aligned}
&P(\tilde{z}_{j1} = 1, \tilde{z}_{q1} = 1|Y_i = 1) = P(a_{ji} = 1, a_{qi} = 1, Y_j = 1, Y_q = 1|Y_i = 1) \\
&\quad = P(a_{ji} = 1, a_{qi} = 1|Y_j = 1, Y_q = 1, Y_i = 1)P(Y_j = 1, Y_q = 1|Y_i = 1) \\
&\quad = P(a_{ji} = 1|Y_j = 1, Y_i = 1)P(a_{qi} = 1|Y_q = 1, Y_i = 1)P(Y_j = 1)P(Y_q = 1) \\
&\quad = P(\tilde{z}_{j1} = 1|Y_i = 1)P(\tilde{z}_{q1} = 1|Y_i = 1) = \pi_1^2\omega_{11}^2,
\end{aligned}$$

and, $P(\tilde{z}_{j1} = 0, \tilde{z}_{q1} = 1|Y_i = 1) = P(\tilde{z}_{j1} = 0|Y_i = 1)P(\tilde{z}_{q1} = 1|Y_i = 1) = \pi_1\omega_{11}(1-\pi_1\omega_{11})$, $P(\tilde{z}_{j1} = 1, \tilde{z}_{q1} = 0|Y_i = 1) = P(\tilde{z}_{j1} = 1|Y_i = 1)P(\tilde{z}_{q1} = 0|Y_i = 1) = \pi_1\omega_{11}(1-\pi_1\omega_{11})$ and $P(\tilde{z}_{j1} = 0, \tilde{z}_{q1} = 0|Y_i = 1) = P(\tilde{z}_{j1} = 0|Y_i = 1)P(\tilde{z}_{q1} = 0|Y_i = 1) = (1-\pi_1\omega_{11})^2$ can be derived in a similar way. As above, we know that $\tilde{z}_{j1}$ and $\tilde{z}_{q1}$ are independent and identically distributed given $Y_i = 1$. Because the log-type terms can be considered as constants, z_{j1} and z_{q1} are also independent and identically distributed given $Y_i = 1$, as well as z_{jp} and z_{qp} for any $1 \leq p \leq 8$ by virtue of the similarity in configuration. Moreover, since $P(z_{jp}z_{qs}|Y_i = 1) = P(z_{jp}|Y_i = 1)P(z_{qs}|Y_i = 1)$ for any $1 \leq p, s \leq 8$, z_{jp} and z_{qs} are independent as well.

As a result, we have proved that for any $j \neq q \neq i$, Z_{ij} and Z_{iq} are independent and identically distributed given $Y_i = 1$. The same results can be made for the case of $Y_i = 0$ similarly, i.e., given $Y_i = 0$, Z_{ij} and Z_{iq} are independent and identically distributed for any $j \neq q \neq i$.

Step 2. Now, we calculate the conditional expectation and variance of Z_{ij}. For any $j \neq i$, $\mu_{Z|1} = E(Z_{ij}|Y_i = 1) =$

$$\begin{aligned}
&\pi_1\left\{\omega_{11}\log\frac{\omega_{11}}{\omega_{10}} + \omega_{11}\log\frac{\omega_{11}}{\omega_{01}} + (1-\omega_{11})\log\frac{1-\omega_{11}}{1-\omega_{10}} + (1-\omega_{11})\log\frac{1-\omega_{11}}{1-\omega_{01}}\right\} \\
&\quad + \pi_0\left\{\omega_{01}\log\frac{\omega_{01}}{\omega_{00}} + \omega_{10}\log\frac{\omega_{10}}{\omega_{00}} + (1-\omega_{01})\log\frac{1-\omega_{01}}{1-\omega_{00}} + (1-\omega_{10})\log\frac{1-\omega_{10}}{1-\omega_{00}}\right\},
\end{aligned}$$

which can be written as $\pi_1\{g(\omega_{11}, \omega_{10}) + g(\omega_{11}, \omega_{01})\} + \pi_0\{g(\omega_{10}, \omega_{00}) + g(\omega_{01}, \omega_{00})\}$, where $g(a, b) = a\log(b^{-1}a) + (1-a)\log\{(1-b)^{-1}(1-a)\}$. In the same way, we have $\mu_{Z|0} = E(Z_{ij}|Y_i = 0) = -\pi_1\{g(\omega_{10}, \omega_{11}) + g(\omega_{01}, \omega_{11})\} - \pi_0\{g(\omega_{00}, \omega_{01}) + g(\omega_{00}, \omega_{10})\}$. Next we can obtain that $\log(\omega_{kk}/\omega_{kl}) = O(1)$ and $\log\{(1-\omega_{kk})/(1-\omega_{kl})\} = O(1)$. Hence, under the assumption (5) we have $\mu_{Z|k} = O(1)$ for $k \in \{0, 1\}$. In addition, by Jensen's inequality, we can easily obtain that $\mu_{Z|1} > 0$ and $\mu_{Z|0} < 0$.

As for the conditional variance, we have that $\mathrm{var}(Z_{ij}|Y_i = 1) = \sum_{p=1}^{8}\mathrm{var}(z_{jp}|Y_i = 1) + \sum_{p\neq q}\mathrm{cov}(z_{jp}z_{jq}|Y_i = 1)$. We only calculate some terms of the above equation for simplicity and the other terms have the similar formation, which are $\mathrm{var}(z_{j1}|Y_i = 1) = \{\log(\omega_{11}/\omega_{01})\}^2\pi_1\omega_{11}(1-\pi_1\omega_{11})$, $\mathrm{var}(z_{j5}|Y_i = 1) = [\log\{(1-\omega_{11})/(1-\omega_{01})\}]^2\pi_1(1-\omega_{11})\{1-\pi_1(1-\omega_{11})\}$ and $\mathrm{cov}(z_{j1}z_{j5}|Y_i = 1) = [\log\{(1-\omega_{11})/(1-\omega_{01})\}]\{\log(\omega_{11}/\omega_{01})\}\{-\pi_1^2\omega_{11}(1-\omega_{11})\}$. Then after some calculations, we have

$$\begin{aligned}
\mathrm{var}(Z_{ij}|Y_i = 1) &= \pi_1\{h(\omega_{11}, \omega_{10}) + h(\omega_{11}, \omega_{01})\} + \pi_0\{h(\omega_{10}, \omega_{00}) + h(\omega_{01}, \omega_{00})\} \\
&\quad + 2\pi_1 g(\omega_{11}, \omega_{10})g(\omega_{11}, \omega_{01}) + 2\pi_0 g(\omega_{10}, \omega_{00})g(\omega_{01}, \omega_{00}) - \mu_{Z|1}^2, \\
\mathrm{var}(Z_{ij}|Y_i = 0) &= \pi_1\{h(\omega_{10}, \omega_{11}) + h(\omega_{01}, \omega_{11})\} + \pi_0\{h(\omega_{00}, \omega_{10}) + h(\omega_{00}, \omega_{01})\} \\
&\quad + 2\pi_1 g(\omega_{10}, \omega_{11})g(\omega_{01}, \omega_{11}) + 2\pi_0 g(\omega_{00}, \omega_{10})g(\omega_{00}, \omega_{01}) - \mu_{Z|0}^2,
\end{aligned}$$

where $h(a, b) = a\{\log(b^{-1}a)\}^2 + (1-a)[\log\{(1-b)^{-1}(1-a)\}]^2$. Then for $k \in \{0, 1\}$, $\sigma_{Z|k}^2 = \mathrm{var}(Z_{ij}|Y_i = k) = O(1) < \infty$ can also be obtained similar to the computation of conditional expectation.

Step 3. Given $Y_i = k \in \{0, 1\}$, by the central limit theorem, we get $\sqrt{n-1}\sigma_{Z|k}^{-1}(\bar{Z}_i - \mu_{Z|k}) \to_d N(0, 1)$, as $n \to \infty$. The proof is completed.

Lemma 2. *Under the assumptions (7) and (8) with* $0 < \gamma \leq 1$, *given* $Y_i = k \in \{0, 1\}$, *we have* $\sqrt{n-1}\sigma_{Z|k}^{-1}(\bar{Z}_i - \mu_{Z|k}) \to_d N(0, 1)$.

Proof. The proof is similar to the proof of Lemma 1. Firstly, like Lemma 1, we can show that for any $j \neq q \neq i$, Z_{ij} and Z_{iq} are independent and identically distributed given $Y_i = k \in \{0, 1\}$.

Secondly, we could obtain the Z_{ij}'s conditional expectation and conditional variance as follows. For $k \in \{0, 1\}$,

$$\begin{aligned}
\mu_{Z|k} &= (-\pi_1)^{1-k}\pi_0^k\{g(\omega_{10}, \omega_{1-k,1-k}) + g(\omega_{01}, \omega_{1-k,1-k})\} + \pi_1^k(-\pi_0)^{1-k}\{g(\omega_{kk}, \omega_{10}) + g(\omega_{kk}, \omega_{01})\},\\
\sigma_{Z|k}^2 &= 2\pi_1^k g(\omega_{kk}, \omega_{10})g(\omega_{kk}, \omega_{01}) + \pi_1^{1-k}\pi_0^k\{h(\omega_{10}, \omega_{1-k,1-k}) + h(\omega_{01}, \omega_{1-k,1-k})\} - \mu_{Z|k}^2\\
&\quad + 2\pi_0^{1-k} g(\omega_{10}, \omega_{1-k,1-k})g(\omega_{01}, \omega_{1-k,1-k}) + \pi_1^k\pi_0^{1-k}\{h(\omega_{kk}, \omega_{10}) + h(\omega_{kk}, \omega_{01})\},
\end{aligned}$$

where $g(a, b) = a\log(b^{-1}a) + (1-a)\log\{(1-b)^{-1}(1-a)\}$ and $h(a, b) = a\{\log(b^{-1}a)\}^2 + (1-a)[\log\{(1-b)^{-1}(1-a)\}]^2$. Under assumptions (7) and (8), we could obtain that $\log(\omega_{kk}/\omega_{kl}) = O(1)$ and $\log\{(1-\omega_{kk})/(1-\omega_{kl})\} \approx |(\omega_{kl} - \omega_{kk})/(1-\omega_{kk})| = O(n^{-\gamma})$. Hence, $\mu_{Z|k} = E(Z_{ij}|Y_i = k) = O(n^{-\gamma})$ and $\sigma_{Z|k}^2 = \mathrm{var}(Z_{ij}|Y_i = k) = O(n^{-\gamma}) < \infty$ for $k \in \{0, 1\}$. In addition, by Jensen's inequality, we can easily obtain that $\mu_{Z|1} > 0$ and $\mu_{Z|0} < 0$.

Lastly, given $Y_i = k \in \{0, 1\}$, by the central limit theorem, we have $\sqrt{n-1}\sigma_{Z|k}^{-1}(\bar{Z}_i - \mu_{Z|k}) \to_d N(0, 1)$, as $n \to \infty$. The proof is completed.

A.3. Proof of Theorem 1

The results of Theorem 1(a) is easily to derive and we omit it for simplicity. See more details in Johnson and Wichern (2014). In addition, it is not difficult to find that $R(G^{LDA}(X_i, \boldsymbol{\theta}))$ is a positive number which does not change along with network size n.

The proof of Theorem 1(b) consists of two steps. Some Bernstein inequalities (Lin and Bai, 2011) about the network terms are given in step 1. We take the second step to derive an upper bound for $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta})$.

Step 1. Given $Y_i = 1$, let $\xi = |\log(\omega_{10}^{-1}\omega_{11})| + |\log(\omega_{01}^{-1}\omega_{11})| + |\log(\omega_{00}^{-1}\omega_{01})| + |\log(\omega_{00}^{-1}\omega_{10})| + |\log\{(1-\omega_{10})^{-1}(1-\omega_{11})\}| + |\log\{(1-\omega_{01})^{-1}(1-\omega_{11})\}| + |\log\{(1-\omega_{00})^{-1}(1-\omega_{01})\}| + |\log\{(1-\omega_{00})^{-1}(1-\omega_{10})\}|$, then we recall the formula of Z_{ij} and easily obtain that $Z_{ij} < \xi$. Because all the link probabilities are finite constants, the order of ξ is $O(1)$.

Because $|Z_{ij} - \mu_{Z|1}| \leq |Z_{ij}| + \mu_{Z|1} < \xi + \mu_{Z|1}$, then for any $t > 0$, by Bernstein inequality, we have

$$P\left(\sum_{j\neq i}(Z_{ij} - \mu_{Z|1}) > t\middle|Y_i = 1\right) \leq \exp\left(-\frac{2^{-1}t^2}{(n-1)\sigma_{Z|1}^2 + 3^{-1}t(\xi + \mu_{Z|1})}\right).$$

Apply the inequality to $-(Z_{ij} - \mu_{Z|1})$ and take $t = 2^{-1}(n-1)\mu_{Z|1}$, then

$$P\left(\sum_{j\neq i} Z_{ij} \leq 2^{-1}(n-1)\mu_{Z|1}\middle|Y_i = 1\right) < \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|1}^2}{\sigma_{Z|1}^2 + 6^{-1}\mu_{Z|1}(\xi + \mu_{Z|1})}\right).$$

Given $Y_i = 0$, we could obtain similar result as

$$P\left(\sum_{j\neq i} Z_{ij} > -2^{-1}(n-1)\mu_{Z|0}\middle|Y_i = 0\right) \leq \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|0}^2}{\sigma_{Z|0}^2 - 6^{-1}\mu_{Z|0}(\xi - \mu_{Z|0})}\right).$$

Step 2. Let $U = S_0 + (\mu_1 - \mu_0)^\top \Sigma^{-1}X_i$. Given $Y_i = 1$, U follows $N(\mu_1^*, \Delta^2)$, where $\mu_1^* = \log(\pi_0^{-1}\pi_1) + 2^{-1}\Delta^2$, and $\Delta = \{(\mu_1 - \mu_0)^\top \Sigma^{-1}(\mu_1 - \mu_0)\}^{1/2}$. Given $Y_i = 0$, U follows $N(\mu_0^*, \Delta^2)$, where $\mu_0^* = \log(\pi_0^{-1}\pi_1) - 2^{-1}\Delta^2$.

Since $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) = \pi_1 P(G(X_i, \mathbf{E}_i) = 0|Y_i = 1) + \pi_0 P(G(X_i, \mathbf{E}_i) = 1|Y_i = 0)$, we consider the part of $Y_i = 1$ and the part of $Y_i = 0$ respectively. Denote $Z_N = \sum_{j\neq i} Z_{ij}$, given $Y_i = 1$, we get

$$\begin{aligned}
P(G(X_i, \mathbf{E}_i) = 0|Y_i = 1) &= P(U + Z_N < 0|Y_i = 1) = \int_{-\infty}^{+\infty} f_U(x)P(Z_N \leq -x)dx\\
&= \int_{-\infty}^{-\frac{n-1}{2}\mu_{Z|1}} f_U(x)P(Z_N \leq -x)dx + \int_{-\frac{n-1}{2}\mu_{Z|1}}^{+\infty} f_U(x)P(Z_N \leq -x)dx\\
&< \int_{-\infty}^{-\frac{n-1}{2}\mu_{Z|1}} f_U(x)dx + P\left(Z_N \leq 2^{-1}(n-1)\mu_{Z|1}\right)\\
&< \Phi\left(-\frac{2^{-1}(n-1)\mu_{Z|1} + \mu_1^*}{\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|1}^2}{\sigma_{Z|1}^2 + 6^{-1}\mu_{Z|1}(\xi + \mu_{Z|1})}\right),
\end{aligned}$$

and the result for $Y_i = 0$ is

$$\begin{aligned} P(G(X_i, \mathbf{E}_i) = 1|Y_i = 0) &= P(U + Z_N \geq 0|Y_i = 0) = \int_{-\infty}^{+\infty} f_U(x)P(Z_N \geq -x)dx \\ &= \int_{-\infty}^{-\frac{n-1}{2}\mu_{Z|0}} f_U(x)P(Z_N \geq -x)dx + \int_{-\frac{n-1}{2}\mu_{Z|0}}^{+\infty} f_U(x)P(Z_N \geq -x)dx \\ &< P\big(Z_N > 2^{-1}(n-1)\mu_{Z|0}\big) + \int_{-\frac{n-1}{2}\mu_{Z|0}}^{+\infty} f_U(x)dx \\ &\leq \Phi\left(\frac{2^{-1}(n-1)\mu_{Z|0} + \mu_0^*}{\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|0}^2}{\sigma_{Z|0}^2 - 6^{-1}\mu_{Z|0}(\xi - \mu_{Z|0})}\right). \end{aligned}$$

To sum up, we have $R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta})$ is less than

$$\begin{aligned} &\pi_1\left\{\Phi\left(\frac{2^{-1}(1-n)\mu_{Z|1} - \mu_1^*}{\Delta}\right) + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|1}^2}{\sigma_{Z|1}^2 + 6^{-1}\mu_{Z|1}(\xi + \mu_{Z|1})}\right)\right\} + \pi_0\left\{\Phi\left(\frac{2^{-1}(n-1)\mu_{Z|0} + \mu_0^*}{\Delta}\right)\right. \\ &\left. + \exp\left(-\frac{8^{-1}(n-1)\mu_{Z|0}^2}{\sigma_{Z|0}^2 - 6^{-1}\mu_{Z|0}(\xi - \mu_{Z|0})}\right)\right\}. \end{aligned}$$

Besides, under assumption (5), $\mu_{Z|k}$, $\sigma_{Z|k}^2$, μ_k^* and Δ are all constants for $k \in \{0, 1\}$. By Jensen's inequality and Lemma 1 in Appendix A.2, we know that $\mu_{Z|1} > 0$ and $\mu_{Z|0} < 0$. Then we can see that as n tends to infinity, the upper bound in (6) approaches 0. As a result, the proof of Theorem 1(b) is completed.

A.4. Proof of Theorem 2

The proof of consequence (a) is similar to the proof of Theorem 1(b) which is omitted here. We only show the proof of consequence (b).

The proof of consequence (b) can be easily derived from the asymptotic normality of Z_N as

$$\begin{aligned} \lim_{n\to\infty} R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) &= \lim_{n\to\infty}\left\{\pi_1 \int_{-\infty}^{+\infty} f_{U|Y_i=1}(x)P(Z_N < -x|Y_i = 1)dx + \pi_0 \int_{-\infty}^{+\infty} f_{U|Y_i=0}(x)P(Z_N \geq -x|Y_i = 0)dx\right\} \\ &= \pi_1\Phi\left(\frac{-\tilde{\mu}_{Z|1} - \mu_1^*}{\sqrt{\Delta^2 + \tilde{\sigma}_{Z|1}^2}}\right) + \pi_0\Phi\left(\frac{\tilde{\mu}_{Z|0} + \mu_0^*}{\sqrt{\Delta^2 + \tilde{\sigma}_{Z|0}^2}}\right), \end{aligned}$$

where $\tilde{\mu}_{Z|k} = \lim_{n\to\infty}(n-1)\mu_{Z|k} = O(1)$ and $\tilde{\sigma}_{Z|k}^2 = \lim_{n\to\infty}(n-1)\sigma_{Z|k}^2 = O(1)$ for $k \in \{0, 1\}$. It is easy to verify that if $\sigma_{Z|1}^{-1}\mu_{Z|1} > \Delta^{-1}\mu_1^*$, $\sigma_{Z|0}^{-1}\mu_{Z|0} < \Delta^{-1}\mu_0^*$ and $\gamma = 1$, $\lim_{n\to\infty} R(G(X_i, \mathbf{E}_i), \boldsymbol{\theta}) < R(G^{LDA}(X_i), \boldsymbol{\theta})$. Then the proof is completed.

A.5. Proof of Theorem 3

As we know that the network-based discriminant rule of node i is $G^N(\mathbf{E}_i) = I(\delta_1^N(\mathbf{E}_i) \geq \delta_0^N(\mathbf{E}_i))$, we would like to find an upper bound of $P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i))$ as the network size tends to infinity. Denote $Z_N = \sum_{j\neq i} Z_{ij}$, after some calculations, we obtain that $P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i))$ is equal to

$$\begin{aligned} &\pi_1 P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)|Y_i = 1) + \pi_0 P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)|Y_i = 0) \\ &\quad = \pi_1\{P(U + Z_N \geq 0, Z_N < 0|Y_i = 1) + P(U + Z_N < 0, Z_N \geq 0|Y_i = 1)\} \\ &\qquad + \pi_0\{P(U + Z_N \geq 0, Z_N < 0|Y_i = 0) + P(U + Z_N < 0, Z_N \geq 0|Y_i = 0)\}. \end{aligned}$$

Given $Y_i = 1$, we have

$$\begin{aligned} P(U + Z_N \geq 0, Z_N < 0|Y_i = 1) &= \int_0^{+\infty}\int_{-y}^{0} f_{\tilde{Z}_i}(x)dx f_U(y)dy \\ &\leq \int_0^{+\infty}\left\{\Phi\left(\frac{-\sqrt{n-1}\mu_{Z|1}}{\sigma_{Z|1}}\right) - \Phi\left(\frac{-y - (n-1)\mu_{Z|1}}{\sqrt{n-1}\sigma_{Z|1}}\right) + \frac{C'}{\sqrt{n}}\frac{E|Z_1 - \mu_{Z|1}|^3}{\sigma_{Z|1}^3}\right\} f_U(y)dy \\ &\leq \frac{1}{\sqrt{2\pi(n-1)}\sigma_{Z|1}}\int_0^{+\infty} y f_U(y)dy + \frac{C}{\sqrt{n}}\frac{E|Z_1 - \mu_{Z|1}|^3}{\sigma_{Z|1}^3} \\ &= \frac{1}{\sqrt{2\pi(n-1)}\sigma_{Z|1}}\left\{1 - \Phi\left(-\frac{\mu_1^*}{\Delta}\right)\right\} + O(n^{-1/2}), \end{aligned}$$

where C and C' are two positive constants and the second line is from Berry–Esseen theorem (Lehmann, 1999). In a similar way, we could also obtain $P(U + Z_N < 0, Z_N \geq 0|Y_i = 1) \leq -\{2\pi(n-1)\}^{-1/2}\sigma_{Z|1}^{-1}\Phi(-\mu_1^*/\Delta) + O(n^{-1/2})$. As a result, $P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)|Y_i = 1) \leq \{2\pi(n-1)\}^{-1/2}\sigma_{Z|1}^{-1}\{1 - 2\Phi(-\mu_1^*/\Delta)\} + O(n^{-1/2})$. Given $Y_i = 0$, we obtain similar process results in the following inequality, $P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)|Y_i = 0) \leq \{2\pi(n-1)\}^{-1/2}\sigma_{Z|0}^{-1}\{1 - 2\Phi(-\mu_0^*/\Delta)\} + O(n^{-1/2})$. Based on the above, we have

$$\begin{aligned} P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)) &\leq \pi_1\{2\pi(n-1)\}^{-1/2}\sigma_{Z|1}^{-1}\{1 - 2\Phi(-\mu_1^*/\Delta)\} \\ &+ \pi_0\{2\pi(n-1)\}^{-1/2}\sigma_{Z|0}^{-1}\{1 - 2\Phi(-\mu_0^*/\Delta)\} + O(n^{-1/2}). \end{aligned} \tag{12}$$

By Lemma 2, when $\gamma = 1$, the order of $\sigma_{Z|k}$ is $O(n^{-1})$, while the order of $\sigma_{Z|k}$ is $O(n^{-\gamma/2})$ when $0 < \gamma < 1$. Then we have $P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i))$ is bounded by a positive constant when $\gamma = 1$. However when $0 < \gamma < 1$, the order of right hand side of (12) is $O(n^{(\gamma-1)/2})$ and then tends to 0 as $n \to \infty$. As a result, we have $\lim_{n\to\infty} P(G(X_i, \mathbf{E}_i) \neq G^N(\mathbf{E}_i)) \to 0$ when $0 < \gamma < 1$. Then the proof is completed.

References

Bickel, P.J., Levina, E., 2004. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10 (6), 989–1010. http://dx.doi.org/10.3150/bj/1106314847.

Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B., 2012. Sparse discriminant analysis. Technometrics 53 (4), 406–413. http://dx.doi.org/10.1198/TECH.2011.08118.

Erdös, P., Rényi, A., 1959. On random graphs, I. Publ. Math. 6, 290–297.

Fan, J., Fan, Y., 2008. High dimensional classification using features annealed independence rules. Ann. Statist. 36 (6), 2605–2637. http://dx.doi.org/10.1214/07-AOS504.

Guo, Y., Hastie, T., Tibshirani, R., 2007. Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8 (1), 86–100. http://dx.doi.org/10.1093/biostatistics/kxj035.

Hand, D.J., 2006. Classifier technology and the illusion of progress. Statist. Sci. 21 (1), 1–14. http://dx.doi.org/10.1214/088342306000000060.

Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. J. Amer. Statist. Assoc. 97 (460), 1090–1098. http://dx.doi.org/10.1198/016214502388618906.

Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: first steps. Social Networks 5 (2), 109–137. http://dx.doi.org/10.1016/0378-8733(83)90021-7.

Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. J. Amer. Statist. Assoc. 76 (373), 33–50. http://dx.doi.org/10.1080/01621459.1981.10477598.

James, G.M., Hastie, T.J., 2001. Functional linear discriminant analysis for irregularly sampled curves. J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3), 533–550. http://dx.doi.org/10.1111/1467-9868.00297.

Jin, J., 2015. Fast network community detection by score. Ann. Statist. 43 (1), 57–89.

Johnson, R.A., Wichern, D.W., 2014. Applied Multivariate Statistical Analysis. Pearson.

Karrer, B., Newman, M.E.J., 2011. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83 (1), 016107. http://dx.doi.org/10.1103/PhysRevE.83.016107.

Lehmann, E.L., 1999. Elements of Large-sample Theory. Springer.

Lin, Z., Bai, Z., 2011. Probability Inequalities. Beijing: Science Press.

McDowell, L.K., Gupta, K.M., Aha, D.W., 2007. Cautious inference in collective classification, In: The National Conference on Artificial Intelligence, pp. 596–601.

Neville, J., Jensen, D., 2000. Iterative classification in relational data. In: Proceedings of American Association for Artificial Intelligence Workshop on Learning Statistical Models from Relational Data, pp. 13–20.

Nowicki, K., Snijders, T.A.B., 2001. Estimation and prediction for stochastic blockstructures. J. Amer. Statist. Assoc. 96 (455), 1077–1087. http://dx.doi.org/10.1198/016214501753208735.

Pan, R., Wang, H., Li, R., 2016. Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening. J. Amer. Statist. Assoc. 111 (513), 169–179. http://dx.doi.org/10.2139/ssrn.2562126.

Qin, T., Rohe, K., 2013. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In: Advances in Neural Information Processing Systems. pp. 3120–3128.

Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P., 2007. Recent developments in exponential random graph (p*) models for social networks. Social Networks 29 (2), 192–215. http://dx.doi.org/10.1016/j.socnet.2006.08.003.

Sewell, D.K., Chen, Y., 2015. Latent space models for dynamic networks. J. Amer. Statist. Assoc. 110 (512), 1646–1657. http://dx.doi.org/10.1080/01621459.2014.988214.

Shao, J., Wang, Y., Deng, X., Wang, S., 2011. Sparse linear discriminant analysis by thresholding for high dimensional data. Ann. Statist. 39 (2), 1241–1265. http://dx.doi.org/10.1214/10-AOS870.

Taskar, B., Abbeel, P., Koller, D., 2002. Discriminative probabilistic models for relational data. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, pp. 485–492.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p*. Psychometrika 61 (3), 401–425. http://dx.doi.org/10.1007/BF02294547.

Witten, D.M., Tibshirani, R., 2011. Penalized classification using Fisher's linear discriminant. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (5), 753–772. http://dx.doi.org/10.1111/j.1467-9868.2011.00783.x.

Yu, H., Yang, J., 2001. A direct LDA algorithm for high-dimensional data-with application to face recognition. Pattern Recognit. 34 (10), 2067–2070. http://dx.doi.org/10.1016/S0031-3203(00)00162-X.

Zhao, Y., Levina, E., Zhu, J., 2012. Consistency of community detection in networks under degree-corrected stochastic block models. Ann. Statist. 40 (4), 2266–2292. http://dx.doi.org/10.1214/12-AOS1036.