2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, October 2-5, 2018

Topological and Algebraic Properties of Chernoff Information between Gaussian Graphs

Binglin Li[1], Shuangqing Wei[2], Yue Wang[1], Jian Yuan[1]

***Abstract*— In this paper, we want to find out the determining factors of Chernoff information in distinguishing a set of Gaussian graphs. We find that Chernoff information of two Gaussian graphs can be determined by the generalized eigenvalues of their covariance matrices. We find that the unit generalized eigenvalues do not affect Chernoff information and their corresponding dimensions do not provide information for classification purpose. In addition, we can provide a partial ordering using Chernoff information between a series of Gaussian trees connected by independent grafting operations. By exploiting relationship between generalized eigenvalues and Chernoff information, we can do optimal classification linear dimension reduction with least loss of information for classification.**
Key words: Gaussian graphs, Generalized eigenvalue, Chernoff information, Dimension reduction

I. Introduction

Gaussian graphical models are widely used in constructing the conditional independence of continuous random variables. It is used in many applications such as social networks [1], economics [2], biology and so on. Among Gaussian graphical models, we are particularly interested in Gaussian trees because of its sparse structure and the existence of computationally efficient algorithms in learning the underlying topologies.

In our study, we focus on classification i.e. hypothesis testing against a set of given Gaussian distributions with sparse graph structures. In this M-ary hypothesis testing problem, we infer which hypothesis a data sequence is generated from. The error probability decreases while data sequence size increases. So we use error exponent to measure how fast error probability decreases along with data. Error exponent is important when we want to estimate how much testing data we need to achieve a given error probability.

In particular, we aim at the error exponent associated with average error probabilities. The resulting error exponent characterizing the vanishing rate of average error probability approaching zero is thus determined by the minimum Chernoff information among all M-choose-2 pairs of hypotheses [3]. It should be noted that in literature (e.g. [4], [5]), because of the complexity in attaining closed form solutions to Chernoff information, KL distance was often adopted as a bound to Chernoff information.

In algebraic analysis of hypothesis testing problem, we also use generalized eigenvalues of covariance matrices as a metric of the difference between them [6], [7]. Clearly, Chernoff information and generalized eigenvalues of covariance matrices are respectively probabilistic and algebraic ways to describe the difference between two Gaussian graphs. There must be relation among topology, statistical distributions (Chernoff information), and algebra (generalized eigenvalues). This paper shows how Chernoff information can be determined by generalized eigenvalues. In addition, we show how topological differences affect generalized eigenvalues and thus Chernoff information. Our work, to the best of our knowledge, is the first one investigating such relationship from Chernoff information point of view.

More specifically, we find that two Gaussian graphs can be linearly and inversely transformed to two graphs whose covariance matrices are diagonal. Entries of the diagonal matrices are related to generalized eigenvalues. Thus we find that Chernoff information between two Gaussian graphs is an expression of generalized eigenvalues and a special parameter λ^*, which is also determined by generalized eigenvalues. In addition, we find that the unit generalized eigenvalues do not affect Chernoff information and the corresponding dimensions make no contribution to differentiating two Gaussian graphs for classification problem.

Our former paper [8] dealt with the classification problem related to Gaussian trees. We found that some special operations on Gaussian trees, namely adding operation and division operation, do not change Chernoff information between them. Now in this paper, we find that these two operations only add one extra unit generalized eigenvalue and do not affect other generalized eigenvalues. We can use generalized eigenvalues to prove the same proposition. Paper [8] also dealt with two Gaussian trees connected by one grafting operation and showed that Chernoff information between them is the same as that of two special 3-node trees whose weights are related to the underlying operation. In this paper, we extend this result to a Gaussian tree chain connected by independent grafting operations and provide a partial ordering of Chernoff information between these trees.

In practical scenarios, we may not have access to all the output of the model. Instead, we may have some constraint on observation costs, which prompts us to reduce the dimension of observation vectors in order to meet such constraints [9], [10]. A good choice here is doing linear dimension reduction in collection stage. We name this

This material is based upon work supported in part by the National Science Foundation (USA) under Grant No. 1320351, and the National Natural Science Foundation of China under Grant 61673237.

[1]B. Li, Y. Wang and J. Yuan are with Department of Electronic Engineering, Tsinghua University, Beijing, P. R. China, 100084. libl13@mails.tsinghua.edu.cn; {wangyue, jyuan}@mail.tsinghua.edu.cn

[2]S. Wei is with the school of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803, USA. swei@lsu.edu

dimension reduction as classification dimension reduction to distinguish it from traditional dimension reduction. We only deal with a 2-ary hypothesis testing in this part. We linearly transform an N dimensional Gaussian vector $\mathbf{x}$ to an $N_O < N$ dimensional vector $\mathbf{y} = \mathbf{A}\mathbf{x}$, through an $N_O \times N$ matrix $\mathbf{A}$. We want to find the optimal linear transformation $\mathbf{A}^*$ which can maximize Chernoff information of two low-dimensional distributions. Our former paper [8] only dealt with a simple, but non-trivial case with $N_O = 1$. In this work, we offer an optimal method to maximize the resulting Chernoff information after a linear transformation for an arbitrary $N_O \geq 1$.

We can divide the features of two distributions into two parts, namely shared features and discrepant features. The aim of classification dimension reduction is to keep discrepant features while discard shared features. Traditional dimension reduction methods, such as Principal Component Analysis (PCA) and other Representation Learning [11], aim to find the optimal features with maximum information. In traditional dimension reduction, we can also divide the features of high-dimensional distributions into two parts. But they are main features and minor features. The aim of traditional dimension reduction is to keep main features while discard minor features. In this way, traditional dimension reduction can recover most high-dimensional informtion from low-dimensional data. Our classification dimension reduction problem have different purpose compared to traditional dimension reduction methods. Some important features in traditional dimension reduction methods may be useless in our method because these features in two hypotheses are similar. In addition, our method needs to compare two distributions, while traditional dimension reduction methods, however, only consider one distribution.

Our major and novel results can be summarized as follows. We first provide the relationship between Chernoff information and generalized eigenvalues, which shows that generalized eigenvalues which are equal to 1 make no contribution to Chernoff information. We use this result to explain why adding and division operations of [8] do not affect Chernoff information between Gaussian trees. These results build a relationship between topology, statistical distribution and algebra. In addition, we deal with Gaussian trees connected by more than one grafting operation and show a partial ordering inside the chain. At last, we provide an optimal classification linear dimension reduction method.

This paper is organized as follows. In Section II, we propose the models of our analysis. The relationship between topology, Chernoff information and generalized eigenvalues is shown in Section III. The partial ordering of Gaussian trees in independent grafting chain is presented in Section IV. Section V shows the optimal classification linear dimension reduction method. In Section VI, we conclude the paper.

II. System Model

Gaussian tree models can represent the dependence of multiple Gaussian random variables by tree topologies. For simplification, we normalize the variance of all Gaussian variables to be 1 and the mean values to be 0. For an N-node tree $\mathbf{G} = (V, E, W)$ with vertex set $V = \{1, \ldots, N\}$, edge set $E = \left\{e_{ij} | (i,j) \subset V \times V\right\}$ and edge weights set $W = \{w_{ij} \in [-1,1] | e_{ij} \in E\}$, E satisfies $|E| = N - 1$ and contains no cycles. A distribution $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T \sim N(\mathbf{0}, \Sigma)$ is said to be a normalized Gaussian distribution on the tree $\mathbf{G} = (V, E, W)$ if

$$\sigma_{ij} = \begin{cases} 1 & i = j \\ w_{ij} & e_{ij} \in E \\ w_{im} w_{mn} \ldots w_{pj} & e_{ij} \notin E \end{cases} \quad (1)$$

where σ_{ij} is the (i,j) term of Σ and $e_{im} e_{mn} \ldots e_{pj}$ is the unique path from node i to node j.

A normalized covariance matrix of a Gaussian tree has a very simple inverse matrix and determinant, as shown in Proposition 1 which has been proved in our former paper [8].

Proposition 1: Assume Σ is a normalized covariance matrix of Gaussian tree $G = (E, V, W)$, so $|\Sigma| = \prod_{e_{ij} \in E} (1 - w_{ij}^2)$ and the elements $[u_{ij}]$ of Σ^{-1} follow the following expressions:

$$u_{ij} = \begin{cases} \frac{-w_{ij}}{1 - w_{ij}^2} & i \neq j \text{ and } e_{ij} \in E \\ 0 & i \neq j \text{ and } e_{ij} \notin E \\ 1 + \sum_{p: e_{ip} \in E} \frac{w_{ip}^2}{1 - w_{ip}^2} & i = j. \end{cases} \quad (2)$$

Consider a set of Gaussian trees, namely, $G_k(\mathbf{x}), k = 1, 2, \ldots, M$, with their prior probabilities given by $\pi_1, \pi_2, \ldots, \pi_M$. We want to do an M-ary hypothesis testing to find out from which Gaussian distribution the data sequence $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_t]$ ($\mathbf{x}_l = [x_{1,l}, \ldots, x_{N,l}]^T$) comes from. We define the average error probability of the hypothesis testing to be P_e, and let $E_e = \lim_{t \to \infty} \frac{-\ln P_e}{t}$ be the resulting error exponent, which depends on the smallest Chernoff information between the trees [5] , namely

$$E_e = \min_{1 \leq i \neq j \leq M} CI(\mathbf{\Sigma}_i || \mathbf{\Sigma}_j) \quad (3)$$

where $CI(\mathbf{\Sigma}_i || \mathbf{\Sigma}_j)$ is the Chernoff information between the i^{th} and j^{th} trees.

For two $\mathbf{0}$-mean N-dim Gaussian joint distributions, $\mathbf{x}_1 \sim N(0, \mathbf{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(0, \mathbf{\Sigma}_2)$, their Kullback-Leibler divergence is as follows

$$D(\mathbf{\Sigma}_1 || \mathbf{\Sigma}_2) = \frac{1}{2} \ln \frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|} + \frac{1}{2} tr(\mathbf{\Sigma}_2^{-1} \mathbf{\Sigma}_1) - \frac{N}{2} \quad (4)$$

where $tr(\mathbf{X}) = \sum_i x_{ii}$ is the trace of square matrix $\mathbf{X}$. We define a new distribution $N(0, \mathbf{\Sigma}_\lambda)$ in the exponential family of the $N(0, \mathbf{\Sigma}_1)$ and $N(0, \mathbf{\Sigma}_2)$, namely

$$\mathbf{\Sigma}_\lambda^{-1} = \mathbf{\Sigma}_1^{-1} \lambda + \mathbf{\Sigma}_2^{-1} (1 - \lambda) \quad (5)$$

so that Chernoff information can be given as

$$CI(\Sigma_1 || \Sigma_2) = D(\Sigma_{\lambda^*} || \Sigma_2) = D(\Sigma_{\lambda^*} || \Sigma_1) \quad (6)$$

where λ^* is the unique point in $[0, 1]$ at which the latter equation is satisfied [3].

We already know that the overall Chernoff information in an M-ary testing is bottlenecked by the minimum pairwise difference [3], thus we next focus on the calculation of Chernoff information of a pair of Gaussian trees.

We also consider classification dimension reduction problem in 2-ary hypothesis testing problem. If we can only observe a low N_O-dim vector, namely $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $N_O \times N$ matrix and $\mathbf{x} \in R^N, \mathbf{y} \in R^{N_O}$, the new low dimensional variables follow joint distributions $N(\mathbf{0}, \hat{\Sigma}_1)$ and $N(\mathbf{0}, \hat{\Sigma}_2)$, where $\hat{\mathbf{\Sigma}}_i = \mathbf{A}\mathbf{\Sigma}_i\mathbf{A}^T, i = 1, 2$. For fixed N_O, we want to find out the optimal $\mathbf{A}^*$ and its Chernoff information result $CI(\hat{\mathbf{\Sigma}}_1^* || \hat{\mathbf{\Sigma}}_2^*)$, s.t.

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} CI(\hat{\mathbf{\Sigma}}_1 || \hat{\mathbf{\Sigma}}_2) \tag{7}$$

III. Generalized eigenvalues, Chernoff information and topology

Chernoff information is the measurement of the difference between statistical distributions. It is hard to be calculated directly and we rarely study its insights about the relationship between Chernoff information and structure characters. Chernoff information and generalized eigenvalues are both important parameters to describe the difference of Gaussian distributions. We expose the relationship among generalized eigenvalues, Chernoff information and topology.

A. Linear transformation to diagonal covariance matrix related to generalized eigenvalues

For two N-node 0-mean Gaussian graphs G_1 and G_2 on random variables $\mathbf{x}$, whose covariance matrices are $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we can use an inverse linear transformation matrix $\mathbf{P}$ to transform them to $\mathbf{x}' = \mathbf{P}\mathbf{x}$ whose covariance matrices $\mathbf{\Sigma}_1'$ and $\mathbf{\Sigma}_2'$ are diagonal and related to the generalized eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$.

$\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are real symmetric positive definite matrices, so the eigenvalues of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$ are all positive, as shown in [12]. The eigenvalue decomposition of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$ is $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$, where $\mathbf{Q}$ is an $N \times N$ matrix and $\mathbf{\Lambda} = Diag(\{\lambda_i\})$ is a diagonal matrix of eigenvalues, in which we put repetitive eigenvalues adjacent. $\{\lambda_i\}$ are the eigenvalues of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$, namely the generalized eigenvalues of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Note that $\mathbf{Q}$ may be non-orthogonal when $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$ is not symmetric.

Proposition 2: For two N-node 0-mean Gaussian graphs G_1 and G_2 whose covariance matrices are $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ respectively, we can construct a linear transformation matrix $\mathbf{P} = \left(\mathbf{Q}^{-1}\mathbf{\Sigma}_2(\mathbf{Q}^{-1})^T\right)^{-\frac{1}{2}}\mathbf{Q}^{-1}$ and thus

$$\mathbf{\Sigma}_2' = \mathbf{P}\mathbf{\Sigma}_2\mathbf{P}^T = \mathbf{I}_N \tag{8}$$

$$\mathbf{\Sigma}_1' = \mathbf{P}\mathbf{\Sigma}_1\mathbf{P}^T = \mathbf{\Lambda} \tag{9}$$

where eigenvalue decomposition of $\mathbf{\Sigma}_1\mathbf{\Sigma}_2^{-1}$ is $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$.

The proof of Proposition 2 is shown in [12].

We can treat $\mathbf{\Sigma}_1'$ and $\mathbf{\Sigma}_2'$ as two covariance matrices of graphs G_1' and G_2' on $\mathbf{x}'$ obtained from G_1 and G_2 by inverse linear transformation $\mathbf{P}$. G_1' and G_2' are graphs with N independent variables.

The distances of G_1' and G_2' are as follows. The distances between G_1 and G_2 are the same with the distances of G_1' and G_2' because $\mathbf{P}$ is non-singular.

$$D(\mathbf{\Sigma}_1||\mathbf{\Sigma}_2) = D(\mathbf{\Sigma}_1'||\mathbf{\Sigma}_2') = \frac{1}{2}\sum_i (-\ln\lambda_i + \lambda_i - 1) \tag{10}$$

$$D(\mathbf{\Sigma}_2||\mathbf{\Sigma}_1) = D(\mathbf{\Sigma}_2'||\mathbf{\Sigma}_1') = \frac{1}{2}\sum_i \left(\ln\lambda_i + \frac{1}{\lambda_i} - 1\right) \tag{11}$$

$$\begin{aligned} D(\mathbf{\Sigma}_\lambda||\mathbf{\Sigma}_1) &= D(\mathbf{\Sigma}_\lambda'||\mathbf{\Sigma}_1') \\ &= \frac{1}{2}\sum_i \left(\ln(\lambda + (1-\lambda)\lambda_i) + \frac{1}{\lambda + (1-\lambda)\lambda_i} - 1\right) \end{aligned} \tag{12}$$

$$\begin{aligned} D(\mathbf{\Sigma}_\lambda||\mathbf{\Sigma}_2) &= D(\mathbf{\Sigma}_\lambda'||\mathbf{\Sigma}_2') \\ &= \frac{1}{2}\sum_i \left(\ln\frac{\lambda + (1-\lambda)\lambda_i}{\lambda_i} + \frac{\lambda_i}{\lambda + (1-\lambda)\lambda_i} - 1\right) \end{aligned} \tag{13}$$

$$CI(\mathbf{\Sigma}_1||\mathbf{\Sigma}_2) = CI(\mathbf{\Sigma}_1'||\mathbf{\Sigma}_2') = D(\mathbf{\Sigma}_{\lambda^*}'||\mathbf{\Sigma}_1') = D(\mathbf{\Sigma}_{\lambda^*}'||\mathbf{\Sigma}_2') \tag{14}$$

where $\mathbf{\Sigma}_\lambda^{-1} = \mathbf{\Sigma}_1^{-1}\lambda + \mathbf{\Sigma}_2^{-1}(1-\lambda)$ and $\mathbf{\Sigma}_\lambda'^{-1} = \mathbf{\Sigma}_1'^{-1}\lambda + \mathbf{\Sigma}_2'^{-1}(1-\lambda)$. Matrix $\mathbf{\Sigma}_\lambda'$ is also diagonal. Parameter λ^* in (14) is the unique root of $D(\mathbf{\Sigma}_{\lambda^*}'||\mathbf{\Sigma}_1') = D(\Sigma_{\lambda^*}'||\mathbf{\Sigma}_2')$, namely $\sum_i \left(\frac{1-\lambda_i}{\lambda^* + (1-\lambda^*)\lambda_i} + \ln\lambda_i\right) = 0$.

In this way, we can conclude that the KL and CI divergences between two Gaussian graphs can be determined by their generalized eigenvalues.

B. Relationship between generalized eigenvalues and Chernoff information

Here we show the relationship between Chernoff information of two Gaussian graphs and generalized eigenvalues of their covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. We define $\prod_{i=1}^N \lambda_i = |\mathbf{\Sigma}_1|/|\mathbf{\Sigma}_2| = \beta$ here.

Proposition 3: For two N-node Gaussian distributions whose covariance matrices are $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ and $|\mathbf{\Sigma}_1|/|\mathbf{\Sigma}_2| = \beta$, their Chernoff information satisfies

$$\begin{aligned} &CI(\mathbf{\Sigma}_1||\mathbf{\Sigma}_2) = \\ &\frac{1}{2}\sum_i \left\{\ln\{(1-\lambda^*)\sqrt{\lambda_i} + \frac{\lambda^*}{\sqrt{\lambda_i}}\}\right\} + \frac{1}{2}(\lambda^* - \frac{1}{2})\ln\beta \end{aligned} \tag{15}$$

where $\{\lambda_i\}$ are the generalized eigenvalues of $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$, and $\lambda^* \in [0, 1]$ is the unique result of

$$\sum_i \frac{1}{\lambda^* + (1-\lambda^*)\lambda_i} = N + (\lambda^* - 1)\ln\beta \tag{16}$$

$$\sum_i \frac{\lambda_i}{\lambda^* + (1-\lambda^*)\lambda_i} = N + \lambda^*\ln\beta \tag{17}$$

The results of equation (16) and (17) are the same. We can prove this proposition from equation (14), as shown in [12].

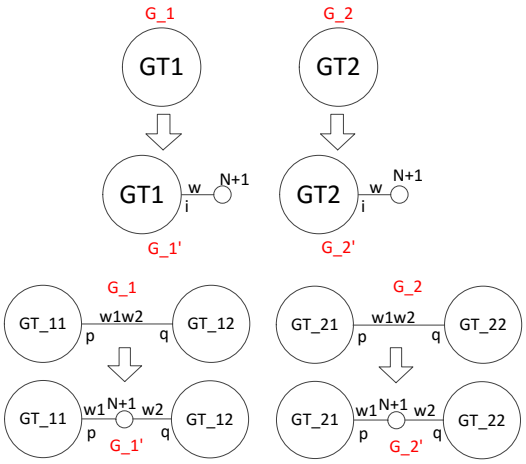

Fig. 1. Adding and Division operations of two trees

We find that generalized eigenvalues of covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are the key parameters of Chernoff information. We can get Chernoff information with these N generalized eigenvalues, so these N parameters contain all the information about the difference between two Gaussian trees. The generalized eigenvalues are so important that we need more properties about them.

C. Effect of Unit Generalized Eigenvalue on Chernoff Information

By carefully examining the relationship between generalized eigenvalues and Chernoff information, we have found that unit generalized eigenvalue has some interesting properties as detailed in this section.

Proposition 4: Assuming that the generalized eigenvalues of $(N+1)$-node G_1' and G_2' are the same with that of N-node G_1 and G_2 except a newly additional unit generalized eigenvalue, the optimal parameter λ^* of $(\mathbf{\Sigma}_1', \mathbf{\Sigma}_2')$ is the same with that of $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ and $CI(\mathbf{\Sigma}_1'||\mathbf{\Sigma}_2') = CI(\mathbf{\Sigma}_1||\mathbf{\Sigma}_2)$.

Under this assumption, $\frac{|\mathbf{\Sigma}_1'|}{|\mathbf{\Sigma}_2'|} = (\prod_{i=1}^{N}\lambda_i)\times 1 = \beta = \frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_2|}$. Proposition 4 can be proved from Proposition 3 as shown in [12].

Proposition 4 shows a possible way to do dimension reduction that we can reduce the dimension of Gaussian graphs from $N+1$ to N without decreasing their Chernoff information.

Paper [8] dealt with the classification on Gaussian trees. In that paper, we defined two special operations on Gaussian tree pairs, namely adding operation and division operation as shown in Fig. 1. The circles in the figure represent Gaussian trees. For adding operation, we add the same leaf node $N+1$, which has the same neighbor i with weight w, to both trees. Division operation only appears when two trees have the same edge e_{pq} with the same weight $w_1 w_2$, for which we split this edge into two edges and add a node $N+1$ in the path of $p-q$ which has edges $e_{(N+1)p}$ and $e_{(N+1)q}$ with weights w_1 and w_2, respectively. After these operations, we get a new pair of Gaussian trees with different dimensions compared to original Gaussian tree pairs. These two operations do not change Chernoff information between two Gaussian trees. Next we show how generalized eigenvalues change after adding or division operation. The covariance matrices of Gaussian trees are normalized.

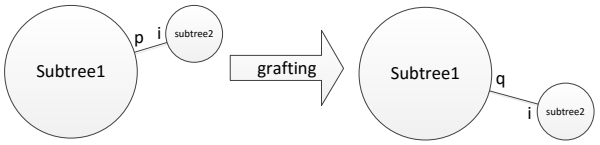

Fig. 2. T_2 is obtained from T_1 by grafting operation

Proposition 5: Assume that Gaussian trees G_1' and G_2' are obtained from G_1 and G_2 by adding operation or division operation. Their covariance matrices are $(\mathbf{\Sigma}_1', \mathbf{\Sigma}_2'), (\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ respectively. The generalized eigenvalues of $(\mathbf{\Sigma}_1', \mathbf{\Sigma}_2')$ are the same with that of $(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)$ except a newly added unit eigenvalue.

Proposition 5 is proved in [12]. From Proposition 4 and 5, we can conclude that adding and division operations do not affect Chernoff information between two Gaussian trees, which has been proved in our former paper [8]. We prove it using generalized eigenvalues now.

IV. Partial Ordering in Independent Grafting Chain

Grafting operation is a kind of topological operation by cutting down a subtree from another tree and pasting it to another location as shown in Fig. 2. In this figure, i, p, q are the nodes in both trees, representing random variables x_i, x_p, x_q. We separate subtree1 and subtree2 by cutting the edge e_{ip} with weight w and paste subtree2 to subtree1 by adding new edge e_{iq} with the same weight w. We do not actually cut any edges from Gaussian trees, though the name of the operation suggests otherwise. We use it to describe the topological difference between two Gaussian trees.

In our former paper [8], we have shown that two Gaussian trees connected by one grafting operation have the same Chernoff information with two special 3-node Gaussian trees whose weights are related to the underlying operation. Now we consider a more complex situation: two Gaussian trees connected by more than one grafting operation. According to Proposition 1, grafting operations do not change determinant of normalized Gaussian trees.

Gaussian trees connected by more grafting operations can not be simplified to a fixed couple of small trees because the interaction of these grafting operations varies. Our initial expectation was that bigger difference in topology between two Gaussian trees leads to larger Chernoff information. This may not be true for all situations.

Before we deal with a sequence of grafting operations, we need to constrain the interaction among them. We define the independence of grafting operations at first.

Definition 1: If all the grafting operations can be divided into different subtrees, as shown in Fig. 3, then these grafting operations are independent. After regrouping all the nodes, the whole tree has star-shaped topology. The subtree in

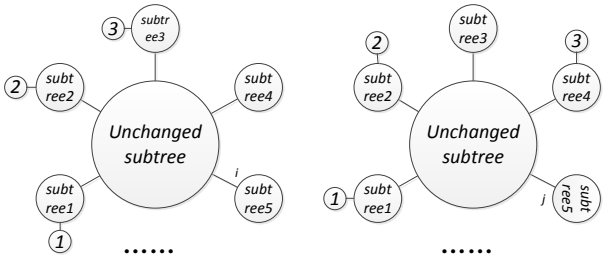

Fig. 3. Independent grafting operations

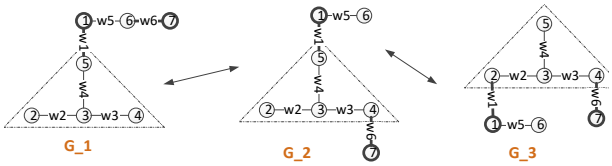

Fig. 4. Example for dependent grafting operations

the center is unchanged during grafting operations. Grafting operations are involved in disjoint super leaf nodes of the star.

In Fig. 3, we show 4 independent grafting operations around the unchanged subtree. There are three types of grafting operations in the star-shaped topology. From left to right, the 1-st, 2-nd grafting operations belong to the first type, the 4-th one belongs to the second type and the 3-rd operation belongs to the third type. For the first type, we can cut a subtree, represented by a small circle with a number in it, from the super leaf node and paste it to another part of this super leaf node. In the second type, we can cut the unchanged subtree outside the super leaf node and paste it to another location in this super leaf node. But for the third type, we cut a subtree, represented by a small circle with a number in it, from a super leaf node and paste it to another super leaf node. The third kind of grafting operations involve two super leaf nodes of the star while the first and second kinds of operations only involve one.

If all the grafting operations are independent, then we can make the following conclusion.

Proposition 6: For two Gaussian trees connected by several independent grafting operations, $\lambda^* = 1/2$ holds.

We can prove it as follows. The trees have the same number of nodes and the same entropy due to grafting operations. Parameter λ^* satisfies $tr(\mathbf{\Sigma}_{\lambda^*}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})) = 0$, which can be transformed from the definition formulas of λ^*. Expression $tr(\mathbf{\Sigma}_{\lambda^*}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}))$ is a summation formula with $4n$ term, where each 4 terms are related to one single grafting operation. We can deal with the terms respectively and prove $tr(\mathbf{\Sigma}_{0.5}(\mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1})) = 0$ eventually. More details can be found in [12].

Proposition 7: For the grafting chain $T_1 \leftrightarrow T_2 \leftrightarrow T_3 \leftrightarrow \cdots \leftrightarrow T_n$ where all the grafting operations in the chain are independent, we can conclude that $CI(T_i||T_j) \leq CI(T_p||T_q)$ if $p \leq i \leq j \leq q$.

Detail of the proof can be found in [12].

If we want to find out the minimum Chernoff information in this chain, we only need to try $n-1$ pairs of $T_i - T_{i+1}$, rather than all the $\binom{n}{2}$ pairs. The number of candidates is significantly reduced.

We can not compare $CI(T_1||T_2)$ and $CI(T_2||T_3)$ even in a simple chain $T_1 \leftrightarrow T_2 \leftrightarrow T_3$ without knowing the weights. We can only compare Chernoff information pairs $CI(T_i||T_j), CI(T_p||T_q)$ with $p \leq i \leq j \leq q$ ordering, so this result is a partial ordering relationship, rather than a full ordering relationship.

In Proposition 7, we constrain the grafting operations independent. We may wonder whether the result suits for all the possible grafting chains without independent assumption. Taking grafting chain $T_1 \leftrightarrow T_2 \leftrightarrow T_3$ in Fig. 4 as an example, the path of the first operation is $6-1-5-3-4$, which is destroyed in the second operation. The chain does not conform to definition 1 and the two grafting operations are not independent. Intuition tells us that $CI(T_1||T_3)$ is likely larger than $CI(T_1||T_2)$ and $CI(T_2||T_3)$, because the difference between T_1 and T_3 is the accumulation of $T_1 - T_2$'s difference and $T_2 - T_3$'s difference. Some numerical cases are shown in Table I. In this table, we can find that $CI(T_1||T_3) < CI(T_1||T_2)$ can hold. This is a counter-intuitive result because more topological differences can not lead to larger Chernoff information between Gaussian trees.

V. Dimension Reduction

The situations in the former section are all about full observation cases, where we can observe all Gaussian variables in the trees each time. In practice, we may have some constraints on observation costs, which prompts us to reduce the dimension of observation vectors in order to meet such constraints. We can only use linearly transformed low-dimensional samples to do the classification. The linear transformation matrix should make sure that the reduced data have the maximum information for classification.

In section III-A, we have shown Proposition 2. With this proposition, we can inversely and linearly transform two original Gaussian graphs into isolated node graphs of new linear space. That is to say, variables of two distributions in new space are independent from each other. In this way, we can decompose difference information into independent dimensions. After decomposition, we can choose the dimensions with most difference information as classification dimension reduction result. The choice of dimensions is shown as follows.

In these new space, x'_i, the i-th variable of $\mathbf{x}' = \mathbf{P}\mathbf{x}$, follows $N(0,1)$ in hypothesis 2 and $N(0,\lambda_i)$ in hypothesis 1. If λ_i is farther from 1, this dimension can provide more information for classification than other dimensions.

Assume that m of all the N eigenvalues $\{\lambda_i\}$ are greater than 1 and the other $N-m$ eigenvalues are no more than 1. If we want to reduce the observation dimension from N to N_O, we choose the dimensions of $\mathbf{x}'$ corresponding to the first k rank and last $N_O - k$ rank of $\{\lambda_i\}$, where $\max\{N_O + m - N, 0\} \leq k \leq \min\{m, N_O\}$. The $N_O \times N$

Cases	$\lambda^*_{T_1\|\|T_3}$	λ	$CI_{T_1\|\|T_3}$	$CI_{T_1\|\|T_2}$	$CI_{T_2\|\|T_3}$
1	0.5191	$19.5746, 0.0433, 1.5439, 0.7642, 1, 1, 1$	0.8983	0.9142	0.0251
2	0.5073	$9.2341, 0.1019, 1.2982, 0.8185, 1, 1, 1$	0.5402	0.5418	0.0113
3	0.5254	$9.4328, 1.653, 0.0844, 0.7603, 1, 1, 1$	0.5982	0.6103	0.0392
4	0.5082	$5.0195, 0.1863, 1.2201, 0.8766, 1, 1, 1$	0.3102	0.3132	0.0056

TABLE I

NUMERICAL CASES DISSATISFYING PROPOSITION 7 IN THE CHAIN OF FIG. 4

linear transformation matrix $\mathbf{A}_k$ is the corresponding N_O rows of $\mathbf{P}$ corresponding to the chosen eigenvalues.

Matrices $\mathbf{A}_k$ are candidate matrices of optimal classification linear dimension reduction matrix $\mathbf{A}^*$, as shown below.

Proposition 8: Matrix $\mathbf{A}^*$ is the optimal $N_O \times N$ linear transformation matrix to maximize the Chernoff information in transformed space, namely $\mathbf{A}^* = \arg\max_{\mathbf{A}_{N_O \times N}} CI(\hat{\mathbf{\Sigma}}_1 || \hat{\mathbf{\Sigma}}_2)$ where $\hat{\mathbf{\Sigma}}_i = \mathbf{A}\mathbf{\Sigma}_i\mathbf{A}^T$ for $i = 1, 2$.

$$\mathbf{A}^* \in \{\mathbf{A}_k | \max\{N_O + m - N, 0\} \leq k \leq \min\{m, N_O\}\} \tag{18}$$

Proof of proposition 8 can be seen in [12] and this proposition ensure the optimality of our method.

The observation is $\mathbf{y} = \mathbf{A}^*\mathbf{x}$ and the covariance matrices of $\mathbf{y}$ in two hypotheses are

$$\mathbf{\Sigma}_2'' = \mathbf{A}^*\mathbf{\Sigma}_2\mathbf{A}^{*T} = \mathbf{I}_{N_O} \tag{19}$$

$$\mathbf{\Sigma}_1'' = \mathbf{A}^*\mathbf{\Sigma}_1\mathbf{A}^{*T} = Diag(\{\mu_i\}) \tag{20}$$

where $\mathbf{\Sigma}_1''$ and $\mathbf{\Sigma}_2''$ are $N_O \times N_O$ diagonal matrices and $\{\mu_1, \mu_2, \ldots, \mu_{N_O}\}$ (including multiple eigenvalues) are N_O chosen eigenvalues.

The main idea of our method is similar as that of PCA. We first decompose the information into independent dimensions and then choose the dimensions with largest weights. But the methods by which we decompose information and choose dimensions are quite different.

VI. CONCLUSION

In this paper, we show the relationship between topology, statistical distribution and algebra. Chernoff information between two Gaussian graphs can be determined by the generalized eigenvalues of their covariance matrices. Unit generalized eigenvalues are very special and do not affect the Chernoff information. Adding and division operations on Gaussian trees only add a newly unit generalized eigenvalues and do not change other generalized eigenvalues. Thus these operations keep the Chernoff information. We also extend our former result about grafting operation to Gaussian trees connected by more than one independent grafting operation and provide a partial ordering among these trees. In addition, we provide an optimal classification linear dimension reduction method with the metric of Chernoff information.

REFERENCES

[1] F. Vega-Redondo, *Complex social networks*. Cambridge University Press, 2007, no. 44.

[2] A. Dobra, T. S. Eicher, and A. Lenkoski, "Modeling uncertainty in macroeconomic growth determinants using Gaussian graphical models," *Statistical Methodology*, vol. 7, no. 3, pp. 292–306, 2010.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[4] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

[5] M. B. Westover, "Asymptotic geometry of multiple hypothesis testing," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 3327–3329, 2008.

[6] G. Doretto and Y. Yao, "Region moments: Fast invariant descriptors for detecting small image structures," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3019–3026.

[7] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2496–2503.

[8] B. Li, S. Wei, Y. Wang, and J. Yuan, "Chernoff information of bottleneck Gaussian trees," in *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2016, pp. 970–974.

[9] E. Nowakowska, J. Koronacki, and S. Lipovetsky, "Dimensionality reduction for data of unknown cluster structure," *Information Sciences*, vol. 330, pp. 74 – 87, 2016.

[10] H. Guan, J. Zhou, B. Xiao, M. Guo, and T. Yang, "Fast dimension reduction for document classification based on imprecise spectrum analysis," *Information Sciences*, vol. 222, pp. 147 – 162, 2013.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[12] B. Li, S. Wei, Y. Wang, and J. Yuan, "Topological and algebraic properties of chernoff information between gaussian graphs," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1712.09741