Graph Coding for Model Selection and Anomaly Detection in Gaussian Graphical Models

Mojtaba Abolfazli, Anders Høst-Madsen, June Zhang
Department of Electrical Engineering
University of Hawaii, Honolulu, HI, USA
Email: {mojtaba, ahm, zjz}@hawaii.edu

Andras Bratincsak
Department of Pediatrics, John A. Burns School of Medicine
University of Hawaii, Honolulu, HI, USA
Email: andrasb@hphmg.org

Abstract—A classic application of description length is for model selection with the minimum description length (MDL) principle. The focus of this paper is to extend description length for data analysis beyond simple model selection and sequences of scalars. More specifically, we extend the description length for data analysis in Gaussian graphical models. These are powerful tools to model interactions among variables in a sequence of i.i.d Gaussian data in the form of a graph. Our method uses universal graph coding methods to accurately account for model complexity, and therefore provide more rigorous approach for graph model selection. The developed method is tested with synthetic and electrocardiogram (ECG) data to find the graph model and anomaly in Gaussian graphical models. The experiments show that our method gives better performance compared to commonly used methods.

I. INTRODUCTION

Multivariate Gaussian distributions are widely used in modeling real-world problems where the relationship between approximately normally distributed variables is of interest. Examples are the distribution of stock returns [1], [2], protein-protein interactions [3], and brain functional activities [4].

Let $X = (X_1, \ldots, X_p)$ be a $p-$dimensional random vector with multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The inverse of covariance matrix, $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega}$, is known as the precision matrix. If the (i, j) entry of the precision matrix is 0, then X_i and X_j are conditionally independent given all the other variables. Therefore, the conditional independence relationships of a multivariate Gaussian can be visualized by an unweighted, undirected graph, $G = (V, E)$ whose adjacency matrix, $\boldsymbol{A}$ is such that

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if} \quad \boldsymbol{\Omega}_{ij} \neq 0, i \neq j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The graph G is known as the conditional independence graph. The sparsity pattern of $\boldsymbol{\Omega}$ determines the topology of G. Figure 1 shows an example precision matrix $\boldsymbol{\Omega}$ and its associated conditional independence graph.

Often, we are interested in the problem of estimating G from i.i.d observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The problem is especially difficult when the Gaussian distribution has high dimensionality but with few observations ($N < p$). In most approaches, the estimator of $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Omega}$) contains a regularization
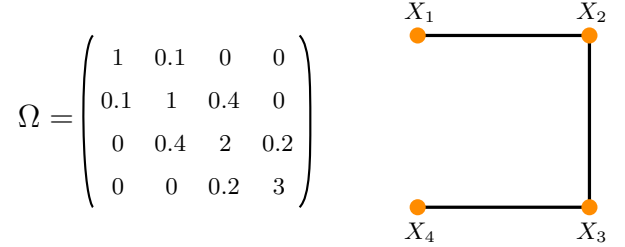
Fig. 1. Illustration of conditional independence graph G associated with a precision matrix.

term to 1) prevent overfitting, 2) control the sparsity of the estimated solution. In the case of estimating the precision matrix, the weight of the regularization term (as determined by the regularization parameter λ) has a high impact on the structure of the conditional independence graph, G.

Previously, we developed methods for lossless coding of arbitrary unweighted, undirected graph structures [5]. In this paper, we use the description length, the minimum number of bits required to describe the data losslessly, associated with these coding methods in conjunction with Rissanens's minimum description length (MDL) [6] to learn the best graph model of data. More specifically, we use the description length for model selection and anomaly detection in Gaussian graphical models.

Rissanen's MDL principle codifies model selection, which balances between how well a model describes the given data and the complexity of the model [6]. In practice, however, it is often difficult to compute an exact description length for some statistical model. For example, previous work in model selection of Gaussian graphical models does not account for the exact topology of the conditional independence graph.

In this paper, we leverage existing work in universal graph coding [5], [7], based on graph statistics such as edge probability, degree distribution and triangle distribution, to compute the exact description length of a Gaussian graphical model.

An overview of previous studies for model selection and anomaly detection in Gaussian graphical models is given in section II. In section III, we use our method in model selection of graphical lasso [8], a popular method for estimating sparse precision matrix. We show that our method outperforms competing model selection methods especially in the case where the number of observations is less than the number of

variables. In section IV, we use description length for anomaly detection using the principle of atypicality [9]. We show that atypicality can capture anomalous data by describing it with fewer bits in itself rather than the optimum coder for typical data. The experiment results on a real-world electrocardiogram (ECG) dataset show that our approach can find anomaly in multivariate normally distributed data with high accuracy.

II. PREVIOUS WORK

Different estimators have been introduced to fit a sparse graphical model to multivariate Gaussian data. Reference [10] proposed a method that works based on neighborhood selection, where conditional independence for each node is estimated by regressing on other variables with lasso. Another popular estimator is graphical lasso [8] that work based on l_1-penalized log-likelihood. This method is easy to implement and fast to compute. Other penalized methods have been introduced to solve the problem efficiently [11]–[15]. All of these estimators have a regularization parameter λ that controls the sparsity level of conditional independence graph ranging from fully disconnected nodes to a fully connected graph. Therefore, their performance is highly influenced by the value of regularization parameter λ.

To pick the best value of λ, different model selection techniques have been used in the literature. We can divide them into two main classes: information-based methods and resampling methods. Information-based methods such as Bayesian information criteria (BIC) [12], Akaike information criteria (AIC) [16], and extended Bayesian information criteria (EBIC) [17] work based on the performance on training dataset and the complexity of the model. This class offers a good statistical interpretation by adding a penalty term to compensate for the overfitting problem of more complex models. EBIC uses an extra penalization parameter that controls preference to simpler models and is set manually based on the cautious and application.

Resampling methods work based on measuring the performance on out-of-sample data. They split the data into a subset of samples to fit a model and the remaining samples to estimate the efficacy of the model. This process is repeated multiple times, and the results are aggregated to find the best model. The most common method in this class is cross-validation (CV) [18]. Other methods are Generalized Approximate Cross-validation (GACV) [19], Rotation Information Criterion (RIC) [20], and Stability Approach to Regularization Selection (StARS) [21]. The major shortcoming of resampling methods is their high computational cost since they require to solve the problem across all sets of subsamples.

Besides these two approaches, there are several studies that propose other methods. The authors in [22] suggested a method based on KL loss that improves the performance of BIC when the sample size is small. Also, several methods based on graph structure were proposed in [23] considering there is a high-level knowledge about graph structure. The paper [24] provides information-theoretic bounds for correct recovery of conditional independence graph based on sample size, number of variables, and the maximum node degree.

All of the aforementioned methods are missing a key aspect of graphical models. They do not consider the graph structure as an informative part of model selection. For example, EBIC only considers the number of edges in estimated graph structure while one may argue that we can have different graph structures with the same number of edges but various degrees of complexity. In this paper, we want to consider the whole graph structure for data analysis in Gaussian graphical models.

In addition to using description length for model selection, we will show another application for anomaly detection in Gaussian graphical models. Previous studies for anomaly detection in Gaussian graphical models are limited to change-points detection in graph structures over time. A time-varying graphical lasso method was developed in [25] by imposing a temporal penalty term that limits the evolution of network structure over time. Reference [26] introduced a multiscale method for changepoint detection that attains a trade-off between sensitivity and accuracy. In [27], structural changes are determined based on the difference between learned background precision matrix and sliding foreground precision matrix. Our approach is different as it uses description length for anomaly detection in data; that is, we encode data using a lossless source coder, and use the resulting codelength as the decision criteria.

III. MODEL SELECTION

In this section, we apply graph-coding techniques introduced in [5], [7] for model selection in Gaussian graphical models. In other words, our goal is to identify zeros in the precision matrix. For that, we use MDL principle and select λ that minimizes the sum of description length of the graph structure and data under that graph structure

$$\arg\min_{\lambda} L(G_\lambda) + L(D|G_\lambda) \tag{2}$$

where $L(G_\lambda)$ is the description length of conditional independence graph structure G_λ that depends on λ, and $L(D|G_\lambda)$ denote the description length of data D when encoded with G_λ. Our results show that G_λ^* associated with λ^* minimizing (2) gives a better model of data comparing to other methods such as CV, BIC, and EBIC.

In the following, we provide the details on how to compute terms in (2).

A. Encoding Graph Structure

To compute $L(G_\lambda)$, we first need to find the underlying conditional independence graph G_λ. By applying an estimator to the sequence of N i.i.d observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from $p-$variate Gaussian distribution, we obtain an estimated precision matrix $\hat{\mathbf{\Omega}}$ for each realization of λ. We dropped λ from $\hat{\mathbf{\Omega}}$ and other $\lambda-$related parameters for the sake of simplicity of notation.

In this paper, we use graphical lasso [8] as one of the most commonly used estimators in Gaussian graphical models. Our approach can be easily applied to any other estimator.

The estimated precision matrix $\hat{\Omega}$ obtained from the graphical lasso estimator is used to form the adjacency matrix $\mathbf{A}$ of the unweighted, undirected graph G_λ as described in (1). Next, we can use any of the following graph coders to compute $L(G_\lambda)$:

- **Degree distribution:** This coder uses the degree distribution of an underlying graph for encoding [5].
- **Triangle:** This coder looks for triangles to encode a graph structure [5].
- **IID structure:** This coder utilizes the edge probability to encode a graph structure. It is designed for Erdös-Rényi graphs and uses a two-stage compression algorithm [7]. First, the structure of graph is converted into two binary strings. Then, an arithmetic encoder is used to compress these strings. The decoder can reconstruct the graph that is isomorphic to the original graph.

For more details about these coders, an interested reader can see [5], [7].

B. Encoding Data with Graph Structure

Computing $L(D|G_\lambda)$ is not as straightforward as computing $L(G_\lambda)$. There are two challenges in this regard. First, we need to deal with real-valued data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where lossless source coding is not generalized directly. Second, we have to encode the data based on the underlying conditional independence graph G_λ. To encode real-valued data, we assume a fixed-point representation with a (large) finite number, r, bits after the period, and an unlimited number of bits before the period [28]. Therefore, the number of bits required to represent data x according to the pdf distribution $f(x)$ is given by

$$L(x) = -\log \int_x^{x+2^{-r}} f(t)dt \approx -\log\left(f(x)2^{-r}\right)$$
$$= -\log f(x) + r \qquad (3)$$

Since we are only interested in relative codelengths between different models, the dependency on r cancels out.

The second challenge is how to encode the data with respect to the graph structure. As mentioned, the data is generated by multivariate Gaussian distribution, which can be characterized by covariance matrix $\boldsymbol{\Sigma}$. Since we do not have access to true covariance matrix $\boldsymbol{\Sigma}$, we can find an estimation $\check{\boldsymbol{\Sigma}}$ based on sample covariance matrix $\boldsymbol{S}$. It should be noted that the structure of $\check{\boldsymbol{\Sigma}}^{-1}$ should match with the structure of G_λ obtained from the previous step:

$$\begin{cases} \check{\boldsymbol{\Sigma}}_{ij} = \boldsymbol{S}_{ij} & \text{if} \quad i = j \text{ or } \boldsymbol{A}_{ij} \neq 0 \\ \check{\boldsymbol{\Sigma}}_{ij}^{-1} = 0 & \text{otherwise.} \end{cases} \qquad (4)$$

This problem is known as *matrix completion* problem and is widely studied in the literature [29]. Dempster in [30], proved the existence and uniqueness of maximum likelihood solution for $\check{\boldsymbol{\Sigma}}$ when the sample covariance matrix $\boldsymbol{S}$ is positive definite. He presented a coordinate-descent algorithm to find the solution iteratively. It should be noted that $\hat{\Omega}$ is the precision matrix obtained from the graphical lasso solution and is not necessarily the inverse of estimated covariance matrix $\check{\boldsymbol{\Sigma}}$.

Once we estimate $\check{\boldsymbol{\Sigma}}$, we use predictive minimum description length (MDL) [6] to compute the codelength of data under the graph structure:

$$L(D|G_\lambda) = -\sum_{i=1}^{N-1} \log f\left(\mathbf{x}_{i+1}|\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_i)\right) \qquad (5)$$

where $f(\cdot|\cdot)$ is the conditional probability and $\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$ denote the maximum likelihood estimate of the parameter, which in this case is the estimated covariance $\check{\boldsymbol{\Sigma}}$ obtained under G_λ. The codelength in (5) is a valid codelength since it is sequentially decodable. It does not work for the first few samples, as there is no estimate of the covariance matrix. Instead, we encode the first few samples with a default distribution that is the same among different realizations of λ.

Finally, the best conditional independence graph structure G_λ is obtained by minimizing the summation of codelength of G_λ and data encoded under G_λ. In the following, we present an algorithm to find the best graph model of data G_{λ^*} associated with λ^*

Algorithm 1 Find the best graph model of data G_{λ^*} via graph-coding methods

---

Input: Samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and set of candidate regularization parameters $\{\lambda_1, \ldots, \lambda_K\}$.

Output: Best graph model of data G_{λ^*}.

1: **for** Each realization of $\lambda \in \{\lambda_1, \ldots, \lambda_K\}$. **do**
2:     Apply graphical model estimator to find $\hat{\Omega}$ and build G_λ based on (1).
3:     Use any graph coder in Section III-A to compute $L(G_\lambda)$.
4:     Compute $L(D|G_\lambda)$ by using (5) where $\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \ldots, \mathbf{x}_i)$ denote $\check{\boldsymbol{\Sigma}}$ obtained from (4).
5:     Add up codelengths resulted from step (3) and step (4).
6: **end for**
7: **return** G_{λ^*} with shortest total codelength in step (5).

C. Experiments

We now provide simulation results on synthetic data generated from zero-mean multivariate Gaussian distribution with known precision matrix, $\boldsymbol{\Omega}$. We will compare the performance of our approach against other methods in the recovery of conditional independence G. We use F1-score as a widely used metric. F1-score is the harmonic mean of precision and recall where the precision is the ratio of the number of correctly estimated edges to the total number of edges in the estimated graph, and the recall is the ratio of the number of correctly estimated edges to the total number of edges in the true graph [21]. We consider two cases: 1) $N > p$ with $N/p = 2$ and 2) $N < p$ with $N/p = 0.5$. The experiments were repeated for $p = 100, 200$. We applied the graphical lasso method described in [8] for 50 realizations of λ in the range $[\bar{\lambda}/10, \bar{\lambda}]$ where $\bar{\lambda}$ is the minimum value of λ that gives totally isolated nodes. The multivariate Gaussian data was generated with different

precision matrix structures have been frequently considered as test cases in the literature [31], [32]:

- Cycle structure with $\mathbf{\Omega}_{ii} = 1$, $\mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.5$, $\mathbf{\Omega}_{1p} = \mathbf{\Omega}_{p1} = 0.4$.
- Autoregressive process of order one AR(1) with $\mathbf{\Omega}_{ii} = 1$, $\mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.5$.
- Autoregressive process of order two AR(2) with $\mathbf{\Omega}_{ii} = 1$, $\mathbf{\Omega}_{i,i-1} = \mathbf{\Omega}_{i-1,i} = 0.5$, $\mathbf{\Omega}_{i,i-2} = \mathbf{\Omega}_{i-2,i} = 0.25$.
- Erdös-Rényi (ER) structure with $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + \delta \mathbf{I}_p$ where $\mathbf{\Omega}_1$ is a matrix with off-diagonal elements taking values randomly chosen from uniform distribution $\mathcal{U}(0.4, 0.8)$ with the probability of $2/p$ and diagonal values set to zero. To keep $\mathbf{\Omega}_2$ positive definite, we choose $\delta = \rho + 0.05$ where ρ is the absolute value of the minimum eigenvalue of $\mathbf{\Omega}_1$. Here $\mathbf{I}_p$ is the identity matrix of size p.
- Hub structure with two hubs. First, we create the adjacency matrix $\mathbf{A}$ by setting off-diagonal elements to one with the probability of 0.01 and zero otherwise. Next, we randomly select two hub nodes and set the elements of the corresponding rows and columns to 1 with the probability of 0.7 and zero otherwise. After that for each nonzero element $\mathbf{A}_{ij}$, we set $\mathbf{A}_{ij}$ with a value chosen randomly from uniform distribution $\mathcal{U}(-0.75, -0.25) \cup (0.25, 0.75)$. Then, we set $\mathbf{\Omega}_1 = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T)$. The final precision matrix $\mathbf{\Omega}_2$ is obtained by $\mathbf{\Omega}_2 = \mathbf{\Omega}_1 + \delta \mathbf{I}_p$ as set-up for Erdös-Rényi graph.

Table I and Table II show the results of applying different methods to recover the conditional independence graph by applying graphical lasso as the estimator for $N > p$ and $N < p$ cases, respectively. The results for CV are given for 5-fold CV, and for EBIC method are given for recommended value of $\gamma = 0.5$ [17]. The results for graph-coding methods are obtained by following the steps outlined in Algorithm 1. The values are the average of 50 Monte Carlo trials and the best value (values) at each row is (are) boldfaced. It can be seen that graph-coding techniques (columns named with Degree, IID, and Triangle) outperform other methods particularly when $N < p$. We observed that coding with degree distribution and triangle give better performance than coding with IID structure in most of the cases. This finding confirms the necessity of having graph coders with the capability to reflect more information about the graph in their codelength.

IV. ANOMALY DETECTION

Another application of our graph-coding approach is anomaly detection in multivariate Gaussian data. For detecting anomalous data, we utilize the atypicality concept developed in [33]. A sequence is atypical if it can be coded in fewer bits using a coder that is not the optimum coder for any typical sequence. Theoretical properties and experimental practicality of atypicality were described in [5], [33], [34]. In the following, we describe the approach for using atypicality to find anomalous data in graph-based data.

Type	p	N	Benchmark methods			Graph-coding methods		
			CV	BIC	EBIC	Degree	IID	Triangle
Cycle	100	200	0.26	0.65	0.75	**1**	**1**	**1**
Cycle	200	400	0.24	0.62	0.69	**0.99**	**0.99**	**0.99**
AR(1)	100	200	0.25	0.64	0.76	**0.99**	0.88	0.98
AR(1)	200	400	0.24	0.62	0.70	**1**	0.99	0.99
AR(2)	100	200	0.28	**0.64**	na	**0.64**	0.62	0.63
AR(2)	200	400	0.24	0.71	na	0.71	0.68	**0.72**
ER	100	200	0.28	0.63	0.51	**0.67**	0.60	**0.67**
ER	200	400	0.27	0.68	**0.76**	**0.76**	**0.76**	**0.76**
Hub	100	200	0.23	0.47	0.01	**0.49**	0.47	0.48
Hub	200	400	0.24	**0.44**	0.20	**0.44**	0.42	**0.44**

TABLE I
F1-SCORE OF CONDITIONAL INDEPENDENCE GRAPH RECOVERY USING DIFFERENT MODEL SELECTION METHODS. RESULTS ARE THE AVERAGE OVER 50 REPLICATIONS WHEN $N > p$.

Type	p	N	Benchmark methods			Graph-coding methods		
			CV	BIC	EBIC	Degree	IID	Triangle
Cycle	100	50	0.27	0.26	0.75	**0.89**	0.79	**0.89**
Cycle	200	100	0.23	0.29	0.68	**0.96**	0.80	**0.96**
AR(1)	100	50	0.28	0.26	0.73	**0.89**	0.80	0.88
AR(1)	200	100	0.23	0.29	0.70	**0.95**	0.78	0.94
AR(2)	100	50	0.32	0.30	na	0.36	**0.37**	0.36
AR(2)	200	100	0.27	0.33	na	0.52	**0.53**	0.52
ER	100	50	0.28	0.24	0.39	0.36	0.44	**0.46**
ER	200	100	0.29	0.35	0.62	0.60	0.63	**0.63**
Hub	100	50	0.18	0.16	0.02	0.25	**0.28**	0.18
Hub	200	100	0.15	0.19	0.01	**0.24**	**0.24**	0.10

TABLE II
F1-SCORE OF CONDITIONAL INDEPENDENCE GRAPH RECOVERY USING DIFFERENT MODEL SELECTION METHODS. RESULTS ARE THE AVERAGE OVER 50 REPLICATIONS WHEN $N < p$.

A. Anomalous Graph Detection Algorithm

In anomalous graph detection, we are given a set of N i.i.d observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from a $p-$variate Gaussian distribution as training set. The question is to determine whether a test set of M i.i.d observations $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ generated from a different distribution, where $N \gg M$, is anomalous or not.

The general procedure is as follows:

1) On the set of training data $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we use Algorithm 1 to find the best conditional independence graph G_T. It means that we learn the parameters of graph structure (e.g. degree distribution) that describes the data best. The typical coder is the estimated conditional independence graph structure G_T. Both encoder and decoder know the values of the parameters associated with G_T (i.e. degree distribution) and thus, they do not need to be encoded.

2) On the test data $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$, we first compute the codelength of typical coder $L(G_T) + L(D|G_T)$ by computing the codelength of data D and graph structure G_T

obtained from the previous step. We then use Algorithm 1 to find the best conditional independence graph G_A describing the test data; here for each realization of λ. We also need to add the overhead of encoding the parameters associated with that specific conditional independence graph. The atypical codelength, $L(G_A) + L(D|G_A)$, is now the minimum of these codelengths.

3) The atypicality measure is the difference between the atypical codelength and the typical codelength. If $L(G_A) + L(D|G_A) - L(G_T) - L(D|G_T) < 0$, or when it is smaller than some threshold, then the test data is considered as anomalous.

B. Experiments

We tested our method for anomaly detection on a real-world dataset. This dataset contains 12-lead ECG signals of a group of youth 16-21 years of age. The dataset was initially processed to select features with empirical distribution close to normal distribution. After screening all features and selecting those with approximately normal distribution, we ended up with 43 ECG variables for analysis.

The idea is that different age groups may show different patterns in their feature space, which may be too subtle to recognize by medical professionals or conventional ECG reading algorithms. Therefore, if we learn the underlying conditional independence graph for a specific age group, it might not perform well when it is applied to another age group. In the experiments, females of 16 years of age are considered as the typical data, and females of any other age group are considered as the atypical data. It is expected that the universal source coder gives a better model for the atypical data, shorter codelength, when it is compared to the typical coder.

To measure the performance, the receiver operating characteristic (ROC) curve and area under the curve (AUC) were used as the metrics. We estimated the typical coder based on 80% of data on the typical data. Test data comprises of two different sets 1) 20% of hold-out data from the typical dataset, 2) randomly chosen samples of the same size from each age group in the atypical data (one set for each age group). To compute the ROC curve, we repeated the experiment 50 times. Figure 2 depicts the ROC curve of the females of each age group versus 16-year-old females when we used degree distribution for coding the graph structure. Our methodology could distinguish a single age group (e.g. 17, etc. years) as atypical data compared to the training age-group (16 years) with a high level of accuracy. Despite the small age gap between the typical and atypical data and the fact that both datasets were from healthy individuals, there was a distinctive pattern of ECG variables that made the groups distinguishable by our novel application.

One may ask how other graph coders perform on this task and we may not need to have more efficient graph coders. Our experiments show when we use IID structure coder, the performance drops compared to the case when we use degree distribution or triangle coder as given in Table III. It means
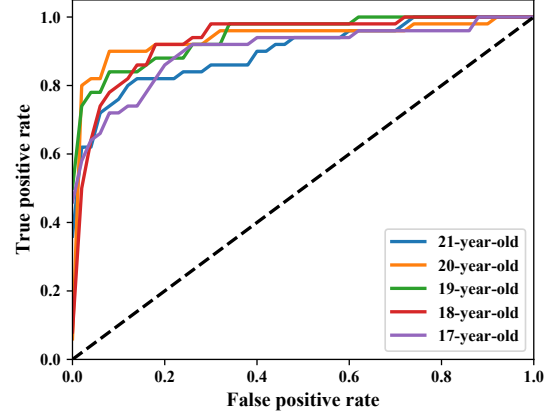

Fig. 2. ROC curve of atypicality measure when degree distribution is used. Each curve represents the measure when the corresponding age group is evaluated versus the 16-year-old group.

Age Group	Degree	IID	Triangle
17-year-old	**0.90**	0.85	0.89
18-year-old	**0.92**	0.90	0.90
19-year-old	0.94	0.93	**0.95**
20-year-old	**0.93**	0.92	0.92
21-year-old	**0.89**	0.85	**0.89**

TABLE III
AUC OF ATYPICALITY MEASURE FOR DIFFERENT GRAPH CODERS.

that the coders based on degree distribution and triangle can extract more information from the graph and use it toward shorter codelength, which leads to better performance.

V. CONCLUSION

In this paper, we extended the description length analysis to graph-based data by encoding graph structure and data according to the graph structure. We introduced two applications: 1) model selection, 2) anomaly detection in Gaussian graphical models. First, we demonstrated that our approach for model selection gives a more accurate graph model of data compared to commonly used methods that overlook the graph structure as an informative part in decision-making. Second, we utilized description length for anomaly detection by using codelength as the decision criteria. In the future, we will extend this approach to other types of graph-based data and applications.

REFERENCES

[1] S. J. Kon, "Models of stock returns—a comparison," *The Journal of Finance*, vol. 39, no. 1, pp. 147–165, 1984.
[2] V. Golosnoy, B. Gribisch, and R. Liesenfeld, "The conditional autoregressive wishart model for multivariate stock market volatility," *Journal of Econometrics*, vol. 167, no. 1, pp. 211–223, 2012.
[3] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, "Fast and accurate multivariate gaussian modeling of protein families: predicting residue contacts and protein-interaction partners," *PloS one*, vol. 9, no. 3, p. e92721, 2014.
[4] G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion, "Brain covariance selection: better individual functional connectivity models using population prior," *Advances in neural information processing systems*, vol. 23, pp. 2334–2342, 2010.

[5] A. Host-Madsen and J. Zhang, "Coding of graphs with application to graph anomaly detection," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1829–1833.

[6] J. Rissanen, "Stochastic complexity and modeling," *The Annals of Statistics*, no. 3, pp. 1080–1100, Sep. 1986.

[7] Y. Choi and W. Szpankowski, "Compression of graphical structures: Fundamental limits, algorithms, and experiments," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 620–638, 2012.

[8] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[9] A. Høst-Madsen, E. Sabeti, and C. Walton, "Data discovery and anomaly detection using atypicality: Theory," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5302–5322, 2019.

[10] N. Meinshausen, P. Bühlmann *et al.*, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[11] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine learning research*, vol. 9, no. Mar, pp. 485–516, 2008.

[12] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[13] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Annals of statistics*, vol. 37, no. 6B, p. 4254, 2009.

[14] D. M. Witten, J. H. Friedman, and N. Simon, "New insights and faster computations for the graphical lasso," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 892–900, 2011.

[15] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "Quic: quadratic approximation for sparse inverse covariance estimation," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2911–2947, 2014.

[16] P. Menéndez, Y. A. Kourmpetis, C. J. ter Braak, and F. A. van Eeuwijk, "Gene regulatory networks from multifactorial perturbations using graphical lasso: application to the dream4 challenge," *PloS one*, vol. 5, no. 12, p. e14147, 2010.

[17] R. Foygel and M. Drton, "Extended bayesian information criteria for gaussian graphical models," in *Advances in neural information processing systems*, 2010, pp. 604–612.

[18] A. J. Rothman, P. J. Bickel, E. Levina, J. Zhu *et al.*, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[19] H. Lian, "Shrinkage tuning parameter selection in precision matrices estimation," *Journal of Statistical Planning and Inference*, vol. 141, no. 8, pp. 2839–2848, 2011.

[20] S. Lysen, "Permuted inclusion criterion: a variable selection technique," *Publicly accessible Penn Dissertations*, p. 28, 2009.

[21] H. Liu, K. Roeder, and L. Wasserman, "Stability approach to regularization selection (stars) for high dimensional graphical models," in *Advances in neural information processing systems*, 2010, pp. 1432–1440.

[22] I. Vujačić, A. Abbruzzo, and E. Wit, "A computationally fast alternative to cross-validation in penalized gaussian graphical models," *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3628–3640, 2015.

[23] A. C. Mestres, N. Bochkina, and C. Mayer, "Selection of the regularization parameter in graphical models using network characteristics," *Journal of Computational and Graphical Statistics*, pp. 1–11, 2018.

[24] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for gaussian markov random fields," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1373–1377.

[25] D. Hallac, Y. Park, S. Boyd, and J. Leskovec, "Network inference via the time-varying graphical lasso," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 205–213.

[26] V. Avanesov, N. Buzun *et al.*, "Change-point detection in high-dimensional covariance structure," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 3254–3294, 2018.

[27] A. Maurya and M. Cheung, "Contrastive structured anomaly detection for gaussian graphical models," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 736–739.

[28] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, no. 2, pp. 416–431, 1983.

[29] R. Grone, C. R. Johnson, E. M. Sá, and H. Wolkowicz, "Positive definite completions of partial hermitian matrices," *Linear algebra and its applications*, vol. 58, pp. 109–124, 1984.

[30] A. P. Dempster, "Covariance selection," *Biometrics*, pp. 157–175, 1972.

[31] L. Li and K.-C. Toh, "An inexact interior point method for l 1-regularized sparse covariance selection," *Mathematical Programming Computation*, vol. 2, no. 3-4, pp. 291–315, 2010.

[32] K. M. Tan, P. London, K. Mohan, S.-I. Lee, M. Fazel, and D. Witten, "Learning graphical models with hubs," *arXiv preprint arXiv:1402.7349*, 2014.

[33] A. Høst-Madsen, E. Sabeti, and C. Walton, "Data discovery and anomaly detection using atypicality: Theory," *IEEE Transactions on Information Theory*, vol. 65, no. 9, pp. 5302–5322, 2019.

[34] E. Sabeti, C. Walton, S. J. Lim *et al.*, "Universal data discovery using atypicality," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 3474–3483.