FUNCTIONAL GRAPHICAL MODEL CLASSIFICATION

P. Li, T. Maiti
Michigan State University
East Lansing, USA
e-mail: `maiti@msu.edu`

Abstract

The functional magnetic resonance imaging (fMRI) records signals coming from human brains, which show activities and states of brains. This measurements result in a high-dimensional time series, and each dimension represents a region of brains. In this paper, we propose a functional Gaussian graphical model to describe the distribution and the correlation structure for this type of high-dimensional time series data, and we find a quadratic discriminant analysis can be effective on functional graphical model. There are two kernel estimators introduced in our work to estimate the node set and the edge set of the functional graphical model, and they are used in our discriminant functions. The simulation study showed that this classification method outperforms other existing methods, and it demonstrated the idea of choosing tuning parameters with different simulated data set. In addition, we present two real data applications. One is an alcoholic condition detection with Electroencephalography (EEG) data collected from electrodes placed on subject's scalps, and the other is a resting state detection using resting state fMRI data from the OpenfMRI database. In both applications, our proposed methodology performs better than other competitive methodologies.

Keywords: data science, classification, functional graphical model

1 Introduction

The advent of functional graphical model [1] has provided a new statistical model for high-dimensional time series data such as Electroencephalography (EEG) data and functional magnetic resonance imaging (fMRI) data. While these two types of data have different features, however, both of them are collected in a sequential form and have high dimensions. To analyze these data, we need to consider the data collected at a specific time point as a structured object and explore the correlation structure of these data. Graphical model turns out to be an appropriate tool to describe the distribution of the data, since the model can contain both vertex information as well as the conditional correlation structure. This can significantly help us, for example, to study the whole brain instead of separating it to different regions.

The classification problem then becomes significant once we can study the brain as a whole object. Since both EEG data and fMRI data record signals from human brain under different conditions, we may be interested about detecting human brain activities or states via these data. Currently, some algorithms, both regression based algorithm and optimization based algorithm, have been proposed to solve this problem. From the regression part, some logistic regression and sparse logistic regression model have been applied for the whole brain classification problem. In the statistical learning aspect, for example, [2] proposed a support vector machine for temporal classification in fMRI data, and [6] has further developed a sparse version of support vector machine for the whole brain classification problem. Moreover, [7] tried to design feature vectors of the brain image data using independent component analysis and apply support vector machine to this feature vector for classification. All these algorithms work well in their application. However, there are some issues remaining unclear in all these methods. The first one is how to vectorize and reduce the rank of the data. As we know, the brain image data, especially fMRI data, usually come in a form of high dimensional time sequence. At each scan, the data we can get is a 3 dimensional real-valued matrix, which cannot be applied to support vector machine model directly. Vectorization is the only way we can do to handle it, but it will lose the temporal information contained in the original data and will result in a super high dimensional vector. Simply reduce the rank of this vector with principle component analysis or independent component analysis will be difficult to interpret. Second, all these classification methods treat the high dimensional time sequence as repeated observation. This restricts the classification can only be made at certain discrete time point. However, the brain activity is a continuous procedure, so we should be able to do classification at any arbitrary time point.

In this paper, we proposed a new classification method, primly for brain image data but also fine for EEG data, to remedy the issues mentioned above. We tried to capture the time varying property of these data by estimating a functional graphical model with two Nadaraya-Waston estimators. These estimates are consistent in non-parametric statistics according to [5] and [4]. We then constructed a discriminant analysis classification with the estimators. The innovation of this work is that we preserved the spatial temporal information in model estimation. The dynamic of the model can help to improve the performance of classification. To compare our classification performance, we select two support vector machine based methods as benchmarks.

- **Sparse SVM**: Refer [6]
- **ICA SVM**: Refer [7]

The paper is formed in the following way. First, we introduce our functional graphical model, and present our estimators for this model. Then, we propose a classification basing on Bayes Classifier and Graphical Lasso. In simulation part, we use data generated from two different Brownian Motions to validate our method. Finally, we compare our method with two other methods in two different real data analysis, one is EEG data and the other is resting state fMRI data.

2 Methodology

2.1 Functional Gaussian Graphical Model

We shall start describing our methodology from the extension of Gaussian Graphical Model. Graphical model is a popular tool to present the distribution of some high-dimensional data, especially some data having complicate correlation structures. We refer [8] for the introduction of graphical models. A graphical model, G, consists of two components, the node set V and the edge set E. While the node set contains all the data information we observed, the edge set describes the underlying correlation between each nodes. Besides the data distribution, we are more eager to know the conditional correlation between each node. At this stage, Gaussian Graphical model is the most popular option, since the edge set becomes to be the precision matrix of the Gaussian distribution with the property that node i and j are conditionally independent if and only if (i, j) th element of the precision matrix is 0. What we are going to do here is extending the Gaussian graphical model into a functional setup.

Assume the graphical model $G(t) = (V(t), E(t))$ is changing smoothly with time t. At a fixed time point t, our observation $X \in \mathbb{R}^p$ is a p-dimensional random vector, representing the observations from each part of the model at this specific time point. Further, we assume this vector, X, is coming from a multivariate normal distribution $\mathcal{N}(\mu(t), \Sigma(t))$. Here, both $\mu(t)$ and $\Sigma(t)$ are continuous functions with respect to time. As a result, $X(t)$ will be our node set $V(t)$, and the inverse matrix of $\Sigma(t)$ decides the conditional correlation structure, $E(t)$, at this time point since the model is basing on the Gaussian assumption. We call $G(t)$ is our functional Gaussian graphical model defined on the continuous time interval. In the next part, we will present our methods for estimating this functional graphical model with data observed at several time points.

2.2 Kernel Estimator

Before the description of methodology, let's define the high-dimensional time series data with the following notation. Suppose $i = 1, ..., n$ and $j = 1, ..., n_i$, which means there are n total individuals and each individual has n_i different observations. We also assume the observations are taken at time point $t_1, t_2, ..., t_d \in [0, 1]$. At each time point $t_k, k = 1, ..., d$, we observe a p dimensional vector, $X_{ij}(t_k)$, which represents the j-th observation from the i-th individual. More specifically, the j-th observation from i-th individual at time point k can be denoted by $X_{ij}(t_k) = (X_{ij}^{(1)}(t_k), ..., X_{ij}^{(p)}(t_k))^T$, $i = 1, ..., n; j = 1, ..., n_i; k = 1, ..., d$. The table 1 shows the general structure of the data.

To functionalize our model, we introduce a bounded continuous kernel function defined on the interval $[0, 1]$. With the observations, we propose two kernel estimators for the mean and covariance of the random vector X at any arbitrary time point t. Here are some necessary notations:

$$\overline{X}(t_k) = \frac{1}{\sum_i n_i} \sum_{i,j} X_{ij}(t_k),\ \Sigma(t_k) = \frac{1}{\sum_i n_i} \sum_{i,j} (X_{ij}(t_k) - \overline{X}(t_k))(X_{ij}(t_k) - \overline{X}(t_k))^T,$$

Table 1: Data Structure

X_i	X_{i1}	$X_{i1}(t_1) = (X_{i1}^{(1)}(t_1), ..., X_{i1}^{(p)}(t_1))^T$
		...
		$X_{i1}(t_d) = (X_{i1}^{(1)}(t_d), ..., X_{i1}^{(p)}(t_d))^T$
	...	...
	...	...
	...	...
	X_{in_i}	$X_{in_i}(t_1) = (X_{in_i}^{(1)}(t_1), ..., X_{in_i}^{(p)}(t_1))^T$
		...
		$X_{in_i}(t_d) = (X_{in_i}^{(1)}(t_d), ..., X_{in_i}^{(p)}(t_d))^T$

where $K(\cdot)$ is the kernel function. Our kernel estimators for mean and covariance matrix are

$$\hat{\mu}(t) = \sum_{k=1}^{d} K(\frac{|t-t_k|}{h})\overline{X}(t_k) \Big/ \sum_{k=1}^{d} K(\frac{|t-t_k|}{h})$$
$$\hat{S}_n(t) = \sum_{k=1}^{d} K(\frac{|t-t_k|}{h})\Sigma(t_k) \Big/ \sum_{k=1}^{d} K(\frac{|t-t_k|}{h})$$

h is a tunning parameter here, which is approximately $O(n^{-\frac{1}{5}})$. These two estimators then can be used to estimate the precision matrices of the functional Gaussian graphical model and classification procedure. The kernel function here are Gaussian kernel, Epanechnikov and Tri-cube kernels. We try all of them in our data applications. Some other kernel functions used in data smoothing can also be considered.

To complete the model fitting procedure, we need to get the precision matrix of the distribution. However, simply computing the inverse matrix of our second estimator does not work well. There are two issues for that, one is the high cost of computation, the other is the singularity of the inverse matrix. Hence, we use graphical lasso instead. The precision matrix will be estimated with the following optimization procedure:

$$\hat{\Theta}(t) = arg\max \log det(\Theta(t)) - tr(\hat{S}_n(t)\Theta(t)) - \lambda|\Theta(t)|_{m\neq l} \tag{1}$$

$|\Theta(t)|_{m\neq l}$ is the L1 norm of the off-diagonal elements in the precision matrix. Among all positive definite matrices, we find the optimizer of the loss function, and it is going to be our estimate for the precision matrix $\Omega(t)$.

2.3 Binary Classification

Finally, we can describe our classification part. We start with a traditional binary classification set up. Let $X_{ij}(t_k), i = 1, ..., n; j = 1, ..., n_i; k = 1, ..., d$ and $Y_{ij}(t_k), i = 1, ..., m; j = 1, ..., m_i; k = 1, ..., d$ be two groups of observations. Suppose we get a new sequence of observations, which is denoted by $Z(t) = (z(t_1), ..., z(t_l))$, l is any arbitrary number, not necessarily equal to d. A quadratic discriminant analysis is applied to this new observation. Let $\hat{\mu}^1(t)$ and $\hat{\Omega}^1(t)$ be the estimated mean and covariance matrix for group X at time t, and $\hat{\mu}^2(t)$ and $\hat{\Omega}^2(t)$ be the estimated mean and covariance matrix for group Y. π_q are the corresponding prior probabilities The discriminant function for quadratic discriminant is

$$\delta(z(t), q) = \tfrac{1}{2}\log|\hat{\Omega}^q(t)| - \tfrac{1}{2}[z(t) - \hat{\mu}^q(t)]^T\hat{\Omega}^q(t)[z(t) - \hat{\mu}^q(t)] + \log \pi_q$$

Then we think $z(t)$ is from class 1 if $\delta(z(t), 1) \geq \delta(z(t), 2)$.

We then can present the posterior probability of the classification result in the following way. Since $\mathbb{P}(Z \in Class1|Z = z) = \mathbb{E}[I\{Z \in class1|Z = z\}]$, the posterior probability is $\hat{\mathbb{P}} = \frac{1}{l}\sum\limits_{k=1}^{l} I\{Z(t_k) \in class1|Z(t_k) = z(t_k)\}$. I is an indicator function here.

3 Simulation

In order to demonstrate the performance of our classification method, we apply it to a data set generated from two different Brownian Motions. As time varying Gaussian processes, the mean vectors and covariance matrices of these two Brownian Motions are functions of time. Suppose we use $\mu^1(t), \mu^2(t)$ and $\Sigma^1(t), \Sigma^2(t)$ to denote the corresponding mean vectors and covariance matrices of these two Brownian Motions, then

$$\begin{aligned} \mu^1(t) &= (\frac{t}{100}, ..., \frac{t}{100})^T & \mu^2(t) &= (\frac{5t}{100}, ..., \frac{5t}{100})^T \\ \Sigma^1(t) &= [\min(\frac{it}{500}, \frac{jt}{500})]_{100\times 100} & \Sigma^2(t) &= [\min(\frac{5it}{500}, \frac{5jt}{500})]_{100\times 100} \end{aligned}$$

For each Brownian Motion, we randomly generate 100 samples at each time point, and the time points are equally spaced between 1 and 2, i.e. $t = \frac{101}{100}, \frac{102}{100}, ..., \frac{200}{100}$. Thus, we have 200 individuals in total. Each individual is a 100×100 matrix, whose rows are observations of that individual at a certain time point. The training set consists of 80 individuals from each group, and the remaining individuals are test set. The result in table 2 shows that our method has a much better performance. However, there is no difference between different kernel function in classification performance.

4 Real Data Analysis

4.1 EEG Data

In this section, we are going to apply our method to the Electroencephalography (EEG) Data from the UCI Machine Learning Repository, `https://archive.ics.uci.edu/ml/machine-learning-databases/eeg-mld/eeg.data.html`. Our EEG data set arises from a large study to examine EEG correlates of genetic predisposition to alcoholism. It contains measurements from 64 electrodes placed on the scalp sampled at 256 Hz (3.9-msec epoch) for 1 second. There were two groups of subjects: alcoholic

Table 2: Classification Performance Comparison

	Kernel	Sensitivity	Specificity	Precision	Accuracy
	Gaussian	0.9	1	1	0.95
FGM	Epanechnikov	0.9	1	1	0.95
	Tri-cube	0.9	1	1	0.95
	Gaussian	0.75	0.8	0.7894	0.775
Sparse SVM	Epanechnikov	0.75	0.8	0.7894	0.775
	Tri-cube	0.75	0.8	0.7894	0.775
ICA SVM		0.523	0.52	0.55	0.53

Table 3: Classification Performance Comparison: EEG Data

	Kernel	Sensitivity	Specificity	Precision	Accuracy
	Gaussian	0.958	0.917	0.92	0.938
FGM	Epanechnikov	0.938	0.938	0.938	0.917
	Tri-cube	0.917	0.917	0.917	0.917
	Gaussian	0.567	0.833	0.773	0.7
Sparse SVM	Epanechnikov	0.833	0.6	0.676	0.716
	Tri-cube	0.833	0.6	0.676	0.716
ICA SVM		0.57	0.55	0.56	0.582

and control. Each subject was exposed to different conditions and their response in terms of EEG data was collected. There were 122 subjects and each subject completed 120 trials. The electrode positions were located at standard sites (Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990). [9] describes in detail the data collection process. We apply our classification method to this data set and record its performance. Further, we compared our results with some other methods to check the efficiency of our functional graphical model.

The original EEG data set contains huge amount of observations, and we randomly select 180 observations to evaluate our classification method. Among these 180 observations, half of them are from alcoholic group and the other half are from the control group. The training set has 60 observations from each class, and the rest 60 observations are used as testing set.

In our results, we compute the sensitivity, specificity, precision and accuracy of our classification research to demonstrate the efficiency of our method. Sensitivity is the true positive rate, specificity is the true negative rate, precision is the positive predictive rate and accuracy is the accuracy rate. The results are in 3. Our functional quadratic discriminant analysis method outperforms the other two methods again. The difference between classification accuracy is pretty similar to the simulation results. We can also find that the choice of kernel function does not make a great difference in classification result, although Gaussian kernel works slightly better in functional QDA.

Table 4: Classification Performance Comparison: fMRI Data

	Kernel	Sensitivity	Specificity	Precision	Accuracy
FGM	Epanechnikov	0.8	0.7	0.727	0.71
	Uniform	0.7	0.6	0.636	0.65
	Tri-cube	0.7	0.7	0.54	0.7
Sparse SVM	Epanechnikov	0.8	0.4	0.57	0.6
	Uniform	0.7	0.5	0.53	0.6
	Tri-cube	0.7	0.5	0.53	0.6
ICA SVM		0.57	0.55	0.56	0.582

4.2 fMRI Classification

In this section, we are heading to the most important topic and adopt all our tensor based methods to a fMRI classification task. The data was obtained from the OpenNeuro database. This project is a functional brain imaging study where 48 younger (20-30 years old) and 36 older (65-75 years old) healthy participants underwent magnetic resonance imaging after having adquate sleep or partial deprived sleep in a crossover design. There are three experiments investigating emotional mimicry, empathy for pain, and cognitive reappraisal, as well as resting state functional magnetic resonance imaging (fMRI). It also contains T1- and T2-weighted structural images and diffusion tensor images. On the night before imaging, participants were monitored with ambulatory polysomnography and were instructed to sleep either as usual or only for three hours. Participants came to the scanner the following evening. In this application, we only use the resting state fMRI data, and try to identify whether patients have enough sleeping or are under partial sleep deprivation with the data.

The fMRI data is pre-processed with the pipeline introduced by [7]. Temporal, spatial correction and coregistration have been applied to the original scans with Matlab tool box SPM 12.0. After data pre-processing, there are 70 scans from the group where patients had adequate sleep, and another 70 scans from the group where patients did not. We performaned a principle components analysis to reduce the dimension of data. The table 4 shows that our functional graphical model classification works better than others.

5 Conclusion

In this work, the functional Gaussian graphical model provides a semi-parametric way to estimate the distribution of high-dimensional functional neuroimage data, while the spatial temporal information contained in the data are well preserved. With some appropriate choices of the tuning parameters and the kernel functions, the model estimates are proved to be consistent. Compared with regularized support vector machine [2], our classification method, which is based on this functional graphical model, works much better with different kernel functions in the simulation as well as the real data applications.

In future, we are going to optimize our computation steps and speed up the whole classification procedure. Currently, the procedure is not fast enough when the sample size in the training set is too large. We hope to make it capable for big data applications in which both the sample size and the data dimension are huge.

References

[1] Qiao X., Guo S., James G.M. (2018). Functional graphical models. Journal of the American Statistical Association, 1-12.

[2] LaConte S., Strother S., Cherkassky V., Anderson J., Hu X. (2005). Support vector machines for temporal classification of block design fMRI data. NeuroImage, 26(2), 317-329.

[3] Kolar M., Xing E. (2011). On time varying undirected graphs. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 407-415).

[4] Ramsay J. (2005). Functional data analysis. Encyclopedia of Statistics in Behavioral Science.

[5] Fan J. (2018). Local polynomial modelling and its applications: monographs on statistics and applied probability 66. Routledge.

[6] Rosa M.J., Portugal L., Hahn T., Fallgatter A.J., Garrido M.I., Shawe-Taylor J., Mourao-Miranda J. (2015). Sparse network-based models for patient classification using fMRI. Neuroimage, 105, 493-506.

[7] Anderson A., Cohen M. (2019). Functional Network Connectivity and SVM Classification of fMRI data using R. Journal of Statistical Software.

[8] Koller D., Friedman N., Bach F. (2009). Probabilistic graphical models: principles and techniques. MIT press.

[9] Gevins A., Brickett P., Costales B., Le J., Reutter B. (1990). Beyond topographic mapping: towards functional-anatomical imaging with 124-channel EEGs and 3-D MRIs. Brain topography, 3(1), 53-64.