9th International IEEE EMBS Conference on Neural Engineering
San Francisco, CA, USA, March 20 - 23 , 2019

Modeling Local Field Potentials with Regularized Matrix Data Clustering

Xu Gao[1], Weining Shen[1], Jianhua Hu[2], Norbert Fortin[3] , Hernando Ombao[4]

***Abstract*— In this paper, we propose a novel regularized mixture model for clustering matrix-valued image data. The new framework introduces a sparsity structure (e.g., low rank, spatial sparsity) and separable covariance structure motivated by scientific interpretability. We formulate the problem as a finite mixture model of matrix-normal distributions with regularization terms, and then develop an Expectation-Maximization-type of algorithm for efficient computation. Simulation results and analysis on brain signals show the excellent performance of the proposed method in terms of a better prediction accuracy than the competitors and the scientific interpretability of the solution.**

I. INTRODUCTION

High volume data sets with complex structures have been widely studied in many fields such as genetics, medicine and transportation (see [1], [2], [3], [4], [5], [6]). Among them, matrix-valued data is commonly encountered in brain images and signals, where the sampling unit can be viewed as a two-dimensional array (i.e., matrix), for example, electroencephalography (EEG) [7], [8], functional magnetic resonance imaging (fMRI) and local field potentials (LFPs) [9]. These signals are in general high-dimensional and possess complicated structure such as spatial/temporal correlation, low rankness and sparsity. The main goal of this paper is to provide a novel approach for clustering matrix-valued data while taking their complex structure into account.

Clustering is an important problem to understand brain function and responses to shocks and stimuli. One key motivation for this paper is the non-spatial working memory experiment conducted by co-author Fortin to study the neuronal learning process on the sequential ordering of odors [10]. Their groundbreaking discovery of temporal coding by the hippocampal neurons extends our basic understanding of the episodic memory neurobiology and thus provides cross-species foundations for clarifying the underlying neural mechanism in memory impairments. Throughout the experiment, series of five odors (denoted as ABCDE) were presented to rats from the same odor port. Each odor presentation was initiated by a nose poke. Rats were tested to correctly identify whether sequence of odor presentation was correct (ABCDE) or incorrect (e.g., AABDE, ABCDD, etc) by holding their nose in the port until the signal or withdrawing before the signal, respectively. LFPs were recorded from twelve microelectrodes that were implanted into rats' cortex. Due to the high dimensionality and complexity of the dataset, the biggest challenge lies in understanding features in LFP signals that are associated with neural mechanism in developing sequential odor memory.

Figure 1 presents the smoothed LFPs across 12 microelectrodes and the associated mean signals for five odors and the whole ABCDE sequence. It is clear that the mean patterns vary dramatically across different sequences, and there is a strong spatial dependence as we compare the signals among different electrodes within each odor. In particular, two "paradigm" can be found across electrodes for odors A, B and D by a rough visual inspection. Therefore a clustering analysis in this study will be helpful to reveal the latent patterns/structure in LFP and hence provide more insights on their connections to different odors. In this paper,

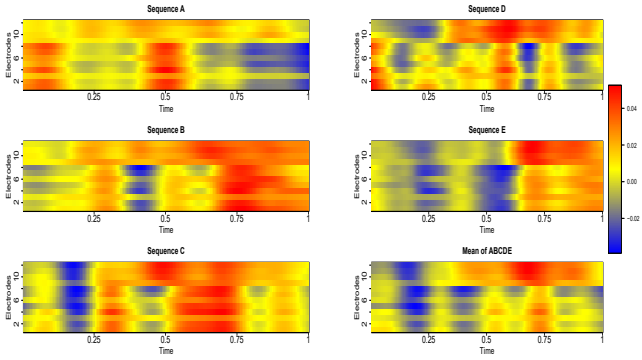

Fig. 1. The mean LFPs across different odors.

we focus on using finite mixture models for clustering purpose because statistical inference can be carried out in a convenient way and the results have a nice probabilistic interpretation. Existing approaches based on mixture models are not directly applicable for matrix data since the set of input covariates are treated as a vector, where the matrix structure and its interpretability are not taken into account. Moreover, by vectorizing a matrix, the resulting dimension of the input space can be extremely large, i.e., a $p \times q$ matrix will be converted to a pq-dimensional vector, which creates additional challenges in computation and theory.

To solve the aforementioned issues, we propose a novel penalized mixture model for clustering the matrix-valued data. Each mixture is represented by a matrix normal distribution, whose covariance matrix can be factorized into the Knocker product of two separate column and row covariance matrices [11], [12]. This representation provides

[1]Xu Gao and Weining Shen are with Department of Statistics, University of California, Irvine, California, 92697, USA, [2]Jianhua Hu is with Department of Biostatistics, Columbia University, New York, 10025, USA, [3]Norbert Fortin is with Department of Neurobiology and Behavior, University of California, Irvine, California, 92697, USA and [4]Hernando Ombao is with Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia. Email:xgao2@uci.edu

both computational convenience and practical interpretation as it separates the variations according to time and spatial domains. In addition, we consider a penalization approach equipped with three different norms (i.e., ℓ_1, ℓ_2 and the nuclear norm) as they may suit for different image structure regularization purpose. For example, the use of nuclear norm provides a useful low-rank approximation of the true image [13], and the use of ℓ_1-norm is helpful for detecting image boundaries [14]. Numerically, we introduce a new EM-type of algorithm that allows efficient computation for all three penalization norms and provide stable and interpretable results.

II. Method

A. Matrix normal distribution

We give a brief review of matrix normal distribution. A $r \times p$ random matrix Y follows a matrix normal distribution with mean M and covariance matrices U and V, denoted by $MN_{r,p}(M, U, V)$, if its density function is

$$f(Y|M,U,V) = \frac{\exp(-\frac{1}{2}\text{tr}(V^{-1}(Y-M)^T U^{-1}(Y-M))}{(2\pi)^{rp/2}|V|^{r/2}|U|^{p/2}}, \tag{1}$$

where $M \in \mathbb{R}^{r\times p}$, $U \in \mathbb{R}^{r\times r}$, $V \in \mathbb{R}^{p\times p}$ and matrices U and V are treated as between- and within-covariance matrices. In [15], a equivalent definition of (1) is given by

$$vec(Y) \sim N(vec(M), V \otimes U), \tag{2}$$

where vec is the column vectorization operation and $\otimes$ is the Kronecker product.

Statistical inference for the matrix normal distribution is usually conducted via the likelihood function. One can utilize iterative algorithm to obtain estimates of covariance matrices as summarized in Algorithm 1.

Algorithm 1 The MLE of covariance matrices

Input: $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_n\}$, τ (tolerance level), Max-iter
Initializing: iter $= 0$, $U_0 = I_{r\times r}$, $V_0 = \frac{1}{nr}\sum_{i=1}^n (Y_i - \bar{Y})' U_0{}^{-1}(Y_i - \bar{Y})$
$U_1 = \frac{1}{np}\sum_{i=1}^n (Y_i - \bar{Y}) V_0{}^{-1}(Y_i - \bar{Y})'$, $V_1 = \frac{1}{nr}\sum_{i=1}^n (Y_i - \bar{Y})' U_1{}^{-1}(Y_i - \bar{Y})$
While (iter $<$ Max-iter or $\|U_1 - U_0\|_F > \tau$ or $\|V_1 - V_0\|_F > \tau$)
Repeat
$U_0 := U_1$
$V_0 := V_1$
$U_1 = \frac{1}{np}\sum_{i=1}^n (Y_i - \bar{Y}) V_0{}^{-1}(Y_i - \bar{Y})^T$
$V_1 = \frac{1}{nr}\sum_{i=1}^n (Y_i - \bar{Y})^T U_1{}^{-1}(Y_i - \bar{Y})$
iter := iter $+ 1$
Return: $\hat{U} := U_1$, $\hat{V} := V_1$

B. Matrix-normal mixture model

We consider a matrix normal mixture model and its inference using EM algorithm. Given i.i.d. observations $Y_1, \cdots, Y_n$ from a mixture of K matrix normal distributions, each indexed by $\Theta_j = (M_j, U_j, V_j)$, and the weights $\pi_1, \ldots, \pi_K$ that belong to a K-dimensional simplex, denoted by Δ_K. Then the mixture density can be written as

$$\sum_{k=1}^K \pi_k f(Y|\Theta_k) = \sum_{k=1}^K \pi_k MN(M_k, U_k, V_k), \tag{3}$$

where f defined by (1). We use $\Theta = (\Theta_1, \ldots, \Theta_K; \pi_1, \ldots, \pi_K)$ to denote the collection of parameters in (3). Then the log-likelihood function is

$$\ell_{obs}(\Theta) = \sum_{i=1}^n \log\left\{\sum_{j=1}^K \pi_j f(Y_i|\Theta_j)\right\}. \tag{4}$$

We can employ Expectation Maximization (EM) algorithm to obtain the estimates of parameter.

In the E-step, the posterior probability of observation Y_i belongs to the j-th cluster is obtained by Bayes Theorem as follows,

$$\alpha_{ij} = \frac{\pi_j f(Y_i|\Theta_j)}{\sum_{l=1}^K \pi_l f(Y_i|\Theta_l)}. \tag{5}$$

In the M-step, the estimates of the parameter vector are obtained by solving the non-constraint optimization problem

$$\widehat{\Theta}_j = \arg\max_{\Theta_j} \sum_{i=1}^n \sum_{j=1}^K \alpha_{ij} \log\{\pi_j f(Y_i|\Theta_j)\}$$

By some algebra, we have

$$\begin{aligned}
\hat{\pi}_j &= \frac{\sum_{i=1}^n \alpha_{ij}}{n}, \quad \widehat{M}_j = \frac{\sum_{i=1}^n \alpha_{ij} Y_i}{\sum_{i=1}^n \alpha_{ij}}, \\
\widehat{U}_j &= \frac{\sum_{i=1}^n \alpha_{ij}(Y_i - \widehat{M}_j)\hat{V}_j^{-1}(Y_i - \widehat{M}_j)'}{p\sum_{i=1}^n \alpha_{ij}}, \\
\widehat{V}_j &= \frac{\sum_{i=1}^n \alpha_{ij}(Y_i - \widehat{M}_j)'\hat{U}_j^{-1}(Y_i - \widehat{M}_j)}{r\sum_{i=1}^n \alpha_{ij}}.
\end{aligned} \tag{6}$$

Note that $\widehat{U}_j$ and $\widehat{V}_j, j = 1, \ldots, k$ can be obtained numerically using the similar method to Algorithm 1.

C. Penalized matrix normal mixture model

Mixture of matrix normal model has been discussed by [16] for classifying three-way array data. However, for many imaging studies, there is a underlying spatial (matrix) structure that needs to be taken into account (see the motivating example in Section 1). Such structure can be effectively modeled by the use of penalty functions on the mean matrix signals, such as the low-rank approximation [13] or total-variation-norm-based penalization [14]. In this paper, we propose a penalized approach by including a penalization term on the means of each mixture component in the matrix normal mixture model. The penalty function takes the form of ℓ_1, ℓ_2, or nuclear norms of the mean matrices $M_1, \ldots, M_k$. The choice of the penalty function depends on the domain knowledge such as sparsity, smoothness and low rankness for the mean structure of each cluster [17], which

gives the results that are more easily interpretable and also eases the computational burden. Specifically, we consider the penalized log-likelihood function

$$Q(\Theta;\lambda)=\sum_{i=1}^{n}\log\left\{\sum_{j=1}^{K}\pi_j f(Y_i|\Theta_j)\right\}-\lambda\sum_{j=1}^{K}P(M_j), \quad (7)$$

where $P(\cdot)$ is some penalty function, such as ℓ_1, ℓ_2, and nuclear norms, and $\lambda \geqslant 0$ is the tuning parameter.

Similar to Section II-B, we propose a modified EM algorithm to estimate the parameters. The E-step proceeds in the same way as Equation (5). The M-step boils down to solving an optimization problem,

$$\widehat{\Theta}=\arg\max_{\Theta}\sum_{i=1}^{n}\sum_{j=1}^{K}\alpha_{ij}\log\{\pi_j f(Y_i|\Theta_j)\}-\lambda\sum_{j=1}^{K}P(M_j). \quad (8)$$

Note that the solution $\hat{\Theta}$ may not have an explicit form. [18] proposed a gradient method related to EM algorithm. It replaces the M-step by conducting one iteration of Newton's method. Alternative approaches, such as surrogate functions [19] and overrelaxed EM algorithm [20] have also been introduced in the literature.

In this article, we mainly focus on three types of penalties: ℓ_1, ℓ_2 and nuclear norms. For ℓ_1 and ℓ_2-penalty, the norms are defined on the vectorized matrix means M_j, and for the nuclear norm penalty, it is defined as the sum of singular values of M_j. [21] introduced ℓ_1-penalty to the mean parameters for mixture of univariate normal models. They obtained an explicit solution for the M-step using a sub-gradient approach. [17] developed the "one-step-late" (OSL) algorithm that can be applied to more general case. Inspired by the aforementioned results, we develop a sub-gradient approach when ℓ_1-norm is used and an OSL approach for ℓ_2 and nuclear norms.

For the ℓ_1-norm penalty, following a similar derivation by [21], M_j can be updated by

$$\hat{M}_j=\text{sign}(\tilde{M}_j)\left(|\tilde{M}_j|-\frac{\lambda}{\sum_{i=1}^{n}\alpha_{i,j}}U_i\mathbf{1}_{r\times p}V_i\right)_+, \quad (9)$$

where $j=1,\ldots,K$, $\tilde{M}_j=\frac{\sum_{i=1}^{n}\alpha_{i,j}Y_i}{\sum_{i=1}^{n}\alpha_{i,j}}$ is the update for M_i without penalty, $B_+=\max(B,0)$, $\mathbf{1}_{r\times p}$ is a matrix of all 1's, and sign() and $(.)_+$ are all component-wise operators.

For the ℓ_2 norm penalty, the objective function is derived to be

$$Q_{\ell_2}(\pi,\Theta)=\sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ij}\log\{\pi_j f(Y_i|\Theta_j)\}-\lambda\sum_{j=1}^{k}\|M_j\|_2.$$

After some calculation, we have

$$\frac{\partial Q_{\ell_2}(\pi,\Theta)}{\partial M_j}=U_j^{-1}\sum_{i=1}^{n}\alpha_{i,j}(Y_i-M_j)V_j^{-1}-2\lambda M_j.$$

Therefore M_j can be updated by

$$\hat{M}_j=\tilde{M}_j-\frac{2\lambda}{\sum_{i=1}^{n}\alpha_{ij}}U_jM_jV_j, \quad (10)$$

where U_j, M_j, V_j are the updates from the previous step.

For the nuclear norm penalty, similar derivation yields

$$\hat{M}_j=\tilde{M}_j-\frac{\lambda}{\sum_{i=1}^{n}\alpha_{ij}}U_j\Phi_j\Omega_j^TV_j, \quad (11)$$

where M_j has the singular value decomposition $M_j=\Phi_j\Lambda_j\Omega_j^T$.

In summary, the proposed estimation procedure involves algorithms of initialization and alternating from E-step and M-step. Here we provide more details.

I. (Initialization) We start with vectorizing the original matrix-valued observations $Y_1,\cdots,Y_n$ and apply K-means to achieve the initial cluster membership values, written as $S_1,\cdots,S_K$, where $S_j=\{i \mid Y_i \text{ in j-th cluster}\}$. Note that we consider alternative methods to K-means in this step, for example, by randomly assign observations to different clusters. Then for each cluster, the initial value of Θ_i can be obtained following the same manner as in Section II-A, and π_j can be directly estimated by $\hat{\pi}_j=|S_j|/n$.

II. (E-step) We update the posterior membership by

$$\alpha_{ij}=\frac{\pi_j f(Y_i|\Theta_j)}{\sum_{l=1}^{k}\pi_l f(Y_i|\Theta_l)}.$$

III. (M-step) The estimate of the mean parameter M_j with respect to various penalties is updated by Equations (9), (10) and (11) respectively. Updates of the estimates of π_j, U_j, V_j follow Equation (6) and Algorithm 1.

IV. (Stopping criteria) Repeat **II.** and **III.** until certain number of iterations have been reached or the Frobenius norm change of the estimate of the mean parameter M_j between consecutive iterates is below some pre-specified cutoff.

Choosing the number of clusters A key question in mixture models is to determine the number of clusters. Inspired by [22], we consider a predictive criteria by adopting the cross validated penalized likelihood (CVPL) as the key measure. We split the dataset $\boldsymbol{Y}=\{Y_1,\cdots,Y_n\}$ into training and testing groups denoted by $\boldsymbol{Y}_{\text{train}}, \boldsymbol{Y}_{\text{test}}$, and then fit a k-mixture model on $\boldsymbol{Y}_{\text{train}}$ and use the estimated parameters to obtain the penalized log-likelihood function on $\boldsymbol{Y}_{\text{test}}$, denoted by $Q(\boldsymbol{Y}_{\text{test}} \mid \boldsymbol{Y}_{\text{train}}, k)$. One nice property of the CVPL is that its expectation is the Kullback-Leibler (KL) divergence between the true penalized likelihood function and the k-mixture penalized likelihood plus some constant. Given this measure, we can define CVPL by first dividing $\boldsymbol{Y}=(\boldsymbol{Y}^1,\ldots,\boldsymbol{Y}^L)$ equally into L parts randomly, and then consider

$$Q(L,k)=L^{-1}\sum_{l=1}^{L}Q(\boldsymbol{Y}^l \mid \boldsymbol{Y}^{-l},k),$$

where $\boldsymbol{Y}^{-l}$ is the data $\boldsymbol{Y}$ excluding $\boldsymbol{Y}^l$. We choose the number of clusters k such that $Q(L,k)$ is maximized. We will give more details for the calculation in the Simulation study.

III. Simulations

We first evaluate whether the proposed CVPL approach is able to identify the correct number of clusters under different scenarios. We generate a mixture of two matrix normal distributions with equal proportions and mean structures of Crossing [23] and Rectangular. The row-wise and column-wise covariance matrices follow an autoregressive setting where $cov\{Y_{k_1,l_1}, Y_{k_2,l_2}\} = 0.9^{|k_1-k_2|+|l_1-l_2|}, 1 \leqslant k_i \leqslant r, 1 \leqslant l_i \leqslant p$. In Scenario I, we set the sample size $n = 100$ and $r = p = 60$. In Scenario II, we let $n = 50,\ r = p = 30$.

We applied the proposed method with ℓ_1, ℓ_2 and Nuclear penalties and summarized the results based on 200 simulations in Table I. It can be seen that the proposed method manages to choose the true number of cluster ($k = 2$) under all cases. Note that ℓ_2 and nuclear norm penalties seem to work better than that of ℓ_1 here. This is expected because the true mean structure has low rank but not entry-wise sparse mean structure.

TABLE I

The cross validated penalized likelihood (CVPL) values obtained from different number of clusters and penalties under two scenarios.

Penalty	λ	CVPL (Scenario I)			CVPL (Scenario II)		
		$k=2$	$k=3$	$k=4$	$k=2$	$k=3$	$k=4$
ℓ_1	0.5	2.345*	2.337	2.333	0.458*	0.453	0.451
	1	2.344*	2.336	2.330	0.457*	0.455	0.452
	1.5	2.341*	2.337	2.332	0.458*	0.457	0.455
ℓ_2	0.5	2.351*	2.349	2.344	0.462*	0.449	0.431
	1	2.352*	2.350	2.345	0.450*	0.434	0.419
	1.5	2.352*	2.349	2.344	0.446*	0.429	0.413
Nuclear	0.5	2.351*	2.348	2.343	0.461*	0.456	0.452
	1	2.351*	2.348	2.344	0.461*	0.457	0.452
	1.5	2.353*	2.349	2.345	0.460*	0.456	0.454

* The highest values across different scenarios ($\times 10^5$)

We then compare the performance of the proposed approach with that of K-means. We generate signals using the same mean and covariance as in the previous section except that we set $n = 50$ and $r = p = 20$ in Scenario III, and $n = 100$ and $p = r = 60$ in Scenario IV. For K-means, we set the number of clusters at the true value (2 in our case). To evaluate the performance, we calculate the adjusted random index (ARI) [24], which is a number that has a maximum value of 1 and measures the agreement between two clustering solutions and works well even when the partitions compared have different numbers of clusters. In addition, we consider the prediction accuracy for both methods and summarize the results based on 200 replications in Table II. In Scenario III, where the sample size is relatively small, the benefit of the proposed method is significant compared to K means in terms of both ARI and prediction accuracy for different choices of penalty norms and the tuning parameter λ. This is expected because K-means does not take the matrix structure into account. For Scenario IV, as the dimension of image increases, the prediction accuracy becomes worse. Still our method works better than K-means. The performance of our method in general becomes better as the tuning parameter λ increases, which suggests that regularization helps improve the clustering performance in this case.

TABLE II

The adjusted random index (ARI) and accuracy obtained from the proposed method and K means under Scenario III and IV.

Penalty	λ	ARI (Scenario III)		Accuracy		ARI (Scenario IV)		Accuracy	
		our method	kmeans	our method	kmeans	our method	kmeans	our method	kmeans
ℓ_1	0	0.867		0.882		0.644		0.696	
	0.5	0.924	0.513	0.938	0.626	0.691	0.517	0.744	0.607
	1	0.962		0.980		0.781		0.822	
	1.5	0.966		0.985		0.788		0.824	
ℓ_2	0.5	0.879		0.892		0.632		0.687	
	1	0.907	0.513	0.918	0.626	0.665	0.517	0.715	0.607
	1.5	0.868		0.881		0.788		0.824	
Nuclear	0.5	0.898		0.909		0.645		0.697	
	1	0.860	0.513	0.876	0.626	0.660	0.517	0.710	0.607
	1.5	0.884		0.897		0.636		0.687	

IV. Analysis of Odor Memory Data

In this section, we apply the developed methodology to analyze a LFP dataset obtained from a memory coding experiment on non-spatial events [10]. In that experiment, rats were trained to identify a series of five odors during the experiment. For most of the cases, those five odors were in the same sequence ("*in-sequence*" odors) while there were some violations ("*out-sequence*" odors). For example, odor sequence *ABCDE* is an "in-sequence" odor yet *ABBDE* is an "out-sequence" odor. Rats were required to poke and hold their nose in the port to correctly identify whether the odors were "in" or "out" sequence. Throughout the experiment, spike and LFP data were collected based on 12 microelectrodes exhibiting task-critical single-cell activity. The LFP dataset contains 247 trials with a sampling rate of 1000 Hertz and $T = 2000$ time points. Figure 2 gives a snapshot of the LFP signals across 12 electrodes.

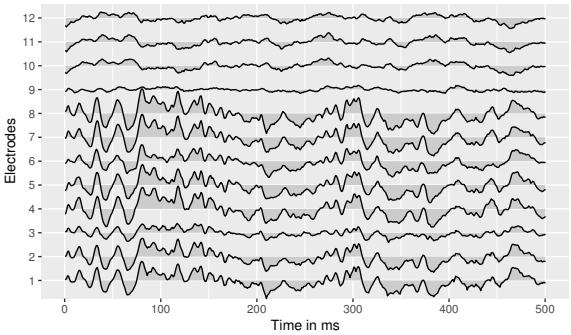

Fig. 2. Time series plot of LFP signals across 12 electrodes in trial 1. The plot only presents the first 500 time points.

A. Time Domain Analysis on Imaging Clustering

As an initial step, we focused on time domain to study the association between raw multi-microelectrode signals with "in-sequence" or "out-sequence" patterns. We implemented the proposed method to the raw LFP signals across all

the 247 trials. Table III summarizes the cross validated penalized likelihood (CVPL) values among different number of clusters and penalties. It is obvious that our method manages to choose the correct number of clusters (2) with the highest CVPL values for all penalty norms. Moreover, our method has a significantly higher adjusted random index (ARI) value compared with K-means, which suggests that the proposed method has desired performance in detecting the latent structure representing "in" or "out" sequences.

TABLE III

ODOR MEMORY STUDY (TIME DOMAIN): THE CROSS VALIDATED PENALIZED LIKELIHOOD (CVPL) AND ADJUSTED RANDOM INDEX (ARI) VALUES FOR DIFFERENT NUMBER OF CLUSTERS AND PENALTIES.

Penalty	λ	CVPL			ARI	
		$k=2$	$k=3$	$k=4$	our method	K means
ℓ_1	0	1.290*	1.285	1.281	0.768	0.499
	0.5	1.253*	1.253*	1.246	0.786	
	1	1.243*	1.206	1.204	0.768	
	1.5	1.249*	1.234	1.218	0.780	
ℓ_2	0.5	1.302*	1.107	1.240	0.768	0.499
	1	1.301*	1.027	1.202	0.774	
	1.5	1.298*	1.189	1.235	0.756	
Nuclear	0.5	1.309*	1.299	1.274	0.756	0.499
	1	1.299*	1.287	1.277	0.733	
	1.5	1.290*	1.286	1.214	0.711	

* The highest CVPL value ($\times 10^5$).

As a further step, researchers are also interested in understanding how LFP signals are related to rat's ability to correctly identify the odor sequence in this experiment. Due to the small sample size of the out-sequence trials, we only focus on those in-sequence trials. In other words, we focus on the "sensitivity" (true positive rate) of the experiment. Table IV summarizes the CVPL and ARI values. For different values of the penalty and λ, the proposed method manages to capture $K = 2$ clusters for most of the time. The ARI values are much higher from our method compared with K-means.

TABLE IV

ODOR MEMORY STUDY (TIME DOMAIN): THE CROSS VALIDATED PENALIZED LIKELIHOOD (CVPL) AND ADJUSTED RANDOM INDEX (ARI) VALUES OBTAINED BASED ON THE IN-SEQUENCE TRIALS.

Penalty	λ	CVPL				ARI	
		$k=2$	$k=3$	$k=4$	$k=5$	our method	K means
ℓ_1	0	1.135*	1.135*	1.126	1.131	0.762	0.506
	0.5	1.103*	1.076	1.084	1.094*	0.783	
	1	1.099*	1.070	1.077	1.136*	0.783	
	1.5	1.107*	1.1078	1.118*	1.068	0.609	
ℓ_2	0.5	1.142*	1.139	0.885	1.144*	0.769	0.506
	1	1.139*	1.016	1.101*	0.986	0.743	
	1.5	1.150*	0.865	1.016	1.061*	0.762	
Nuclear	0.5	1.159*	1.125	1.119	1.126*	0.769	0.506
	1	1.153*	1.116	1.136*	1.105	0.756	
	1.5	1.141*	1.142*	1.036	1.123	0.783	

* The top two CVPL values ($\times 10^5$).

B. Time Frequency Clustering Analysis

Next we study the latent structure from a time-frequency perspective. [10] suggests that two particular oscillatory bands (theta: 4 - 12 Hertz and slow gamma: 20 - 40 Hertz) yield strong power and play significant roles in detecting the in/out sequences. Figure 3 shows the time-frequency plot suggests that the low frequency theta band obtains much more power than the slow gamma band. We applied the proposed method to the spectrum of theta and slow gamma bands separately. Table V presents the results for the number of clusters and the ARI that compares with the true odor sequence for the theta band. The results seem to prefer a large number of clusters (e.g. 5), which may be explained by the fact that there are a total of 5 odors. Our approach provides some evidence indicating the association between the low frequency band (Theta) and the odor sequence. The method was applied to the slow gamma band and produced similar results.

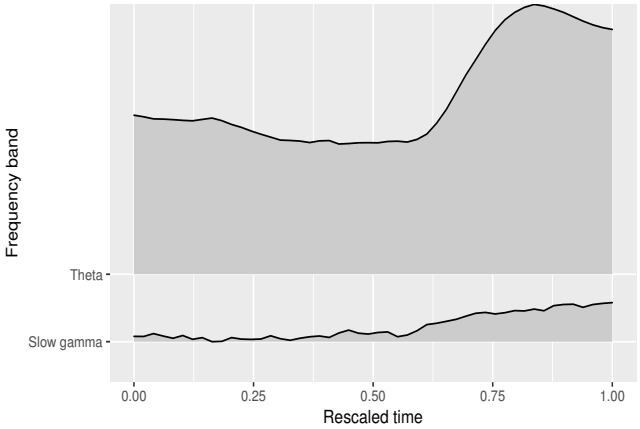

Fig. 3. The time frequency plot of Theta and Slow Gamma bands over the "in-sequence" trials.

TABLE V

ODOR MEMORY STUDY (TIME-FREQUENCY DOMAIN): THE CROSS VALIDATED PENALIZED LIKELIHOOD (CVPL) AND ADJUSTED RANDOM INDEX (ARI) VALUES OBTAINED FROM THE "IN-SEQUENCE" TRIALS BASED ON THE THETA BAND.

Penalty	λ	CVPL				ARI	
		$k=2$	$k=3$	$k=4$	$k=5$	our method	K means
ℓ_1	0	11.001	11.300*	11.198*	11.172	0.712	0.679
	0.5	8.516	8.975*	8.849	8.997*	0.692	
	1	8.650	8.632	8.725*	8.745*	0.703	
	1.5	8.571	8.705*	8.556	8.701*	0.709	
ℓ_2	0.5	8.965*	8.881*	8.671	7.277	0.693	0.679
	1	8.719*	8.388	8.544*	7.616	0.686	
	1.5	8.650	8.632	8.825*	8.745*	0.682	
Nuclear	0.5	9.034	9.196*	9.183	9.259*	0.707	0.679
	1	9.013	9.166	9.255*	9.263*	0.714	
	1.5	8.571	9.040*	8.995*	8.969	0.712	

* The top two highest values ($\times 10^3$).

V. Concluding remarks

In this paper, we proposed a regularized probabilistic clustering framework to analyze matrix data. Compared to the existing approaches such as K-means, the advantages are as follows: (1.) By working directly on matrices, we are able to capture the row-wise and column-wise correlation simultaneously; (2.) By introducing penalty terms into the likelihood function, the proposed framework has the ability to uncover the nature sparsity that are inherent to the signals and images; (3.) The proposed approach is grounded on theoretical foundations; provides straightforward interpretability; and has low computational cost and hence amenable to big datasets.

Although this paper provides some promising results, there remain many open problems that are encountered when analyzing matrix data. For instance, in the current work, choosing the number of clusters rely on some pre-specified measures (CVPL). As an extension, one could introduce a Bayesian framework into the clustering analysis and thus incorporate a more data-driven and interpretable optimal number of clusters.

Acknowledgement

Shen's research is partially supported by the Simons Foundation (Award 512620) and the National Science Foundation (NSF DMS 1509023).

References

[1] S. Gouveia, M. G. Scotto, C. H. Weiß, and P. J. S. Ferreira, "Binary auto-regressive geometric modelling in a dna context," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 66, no. 2, pp. 253–271, 2017.

[2] P. Chen, J. Jiao, M. Xu, X. Gao, and C. Bischak, "Promoting active student travel: a longitudinal study," *Journal of transport geography*, vol. 70, pp. 265–274, 2018.

[3] P. Chen, F. Sun, Z. Wang, X. Gao, J. Jiao, and Z. Tao, "Built environment effects on bike crash frequency and risk in beijing," *Journal of safety research*, vol. 64, pp. 135–143, 2018.

[4] Y. Guo, Y. Wang, T. Marin, E. Kirk, R. M. Patel, and C. D. Josephson, "Statistical methods for characterizing transfusion-related changes in regional oxygenation using near-infrared spectroscopy (nirs) in preterm infants," *arXiv preprint arXiv:1801.08153*, 2018.

[5] X. Gao, B. Shahbaba, and H. Ombao, "Modeling binary time series using gaussian processes with application to predicting sleep states," *Journal of Classification*, vol. 35, no. 3, pp. 549–579, 2018.

[6] X. Gao, D. Gillen, and H. Ombao, "Fisher information matrix of binary time series," *METRON*, vol. 76, no. 3, pp. 287–304, 2018.

[7] X. Gao, W. Shen, C.-M. Ting, S. C. Cramer, R. Srinivasan, and H. Ombao, "Modeling brain connectivity with graphical models on frequency domain," *arXiv preprint arXiv:1810.03279*, 2018.

[8] Y. Wang, C.-M. Ting, and H. Ombao, "Modeling Effective Connectivity in High-Dimensional Cortical Source Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 7, p. 1315, 2016.

[9] X. Gao, B. Shahbaba, N. Fortin, and H. Ombao, "Evolutionary state-space model and its application to time-frequency analysis of local field potentials," *arXiv preprint arXiv:1610.07271*, 2016.

[10] T. A. Allen, D. M. Salz, S. McKenzie, and N. J. Fortin, "Nonspatial sequence coding in ca1 neurons," *Journal of Neuroscience*, vol. 36, no. 5, pp. 1547–1563, 2016.

[11] A. P. Dawid, "Some matrix-variate distribution theory: notational considerations and a bayesian application," *Biometrika*, vol. 68, no. 1, pp. 265–274, 1981.

[12] P. Dutilleul, "The mle algorithm for the matrix normal distribution," *Journal of statistical computation and simulation*, vol. 64, no. 2, pp. 105–123, 1999.

[13] H. Zhou and L. Li, "Regularized matrix regression," *J R Stat Soc Series B Stat Methodol.*, vol. 76, no. 2, pp. 463–483, 2014.

[14] X. Wang, H. Zhu, and ADNI, "Generalized scalar-on-image regression models via total variation," *J Am Stat Assoc.*, vol. 112, pp. 1156–1168, 2017.

[15] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. CRC Press, 1999, vol. 104.

[16] C. Viroli, "Finite mixtures of matrix normal distributions for classifying three-way data," *Statistics and Computing*, vol. 21, pp. 511–522, 2011.

[17] P. J. Green, "On use of the em for penalized likelihood estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 443–452, 1990.

[18] K. Lange, "A gradient algorithm locally equivalent to the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 425–437, 1995.

[19] K. Lange, D. R. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[20] Y. Yu, "Monotonically overrelaxed em algorithms," *Journal of Computational and Graphical Statistics*, vol. 21, no. 2, pp. 518–537, 2012.

[21] W. Pan and X. Shen, "Penalized model-based clustering with application to variable selection," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1145–1164, 2007.

[22] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics and computing*, vol. 10, no. 1, pp. 63–72, 2000.

[23] H. Zhou and L. Li, "Regularized matrix regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 463–483, 2014.

[24] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis," *Multivariate Behavioral Research*, vol. 21, no. 4, pp. 441–458, 1986.