Title

Regularized matrix data clustering and its application to image analysis.

Journal
Biometrics, 77(3)

Authors
Gao, Xu
Shen, Weining
Zhang, Liwen
et al.

Publication Date
2021-09-01

DOI
10.1111/biom.13354

Peer reviewed

Published in final edited form as:
Biometrics. 2021 September ; 77(3): 890–902. doi:10.1111/biom.13354.

Regularized matrix data clustering and its application to image analysis

Xu Gao[1], **Weining Shen**[1], **Liwen Zhang**[2], **Jianhua Hu**[3], **Norbert J. Fortin**[4], **Ron D. Frostig**[4,5], **Hernando Ombao**[6]

[1]Department of Statistics, University of California, Irvine, California

[2]Shanghai University of Finance and Economics, Shanghai, China

[3]Herbert Irving Comprehensive Cancer Center, Columbia University, New York

[4]Department of Neurobiology and Behavior, University of California, Irvine, California

[5]Department of Biomedical Engineering, University of California, Irvine, California

[6]Statistics Program, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Abstract

We propose a novel regularized mixture model for clustering matrix-valued data. The proposed method assumes a separable covariance structure for each cluster and imposes a sparsity structure (eg, low rankness, spatial sparsity) for the mean signal of each cluster. 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. In theory, we show that the proposed estimators are strongly consistent for various choices of penalty functions. Simulation and two applications on brain signal studies confirm the excellent performance of the proposed method including a better prediction accuracy than the competitors and the scientific interpretability of the solution.

Keywords

clustering; imaging; matrix normal distribution; mixture model; regularization; time-frequency analysis

Correspondence: Weining Shen, Department of Statistics, University of California, Irvine, CA. weinings@uci.edu.

DATA AVAILABILITY STATEMENT
The data that support the findings in this paper are available in the Supporting Information of this article.

SUPPORTING INFORMAION
Web Appendices, Tables, and Figures referenced in Sections 4–6 are available with this paper at the Biometrics website on Wiley Online Library. Computational code and data sets are available at the following website (see Supporting Information Section 4 for a detailed description): https://github.com/XuGaoUCI/Regularized_Matrix_Clustering.

1 | INTRODUCTION

The past decade has witnessed a dramatic development in technologies that generate high-volume data sets with complex statistical structure. Among them, matrix-valued data are commonly encountered in brain images and signals, where the sampling unit can be viewed as a two-dimensional array (ie, matrix), for example, electroencephalography (EEG) and local field potentials (LFPs). These signals are in general high-dimensional and possess complicated structure such as spatial/temporal correlation, low rankness, and sparsity (Gao *et al.*, 2019; Hu *et al.*, 2019; Wang *et al.*, 2019). 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 a fundamental problem in statistics and many scientific applications including studies that investigate brain function and its responses to stimuli (King, 2015). One key motivation for this paper is the nonspatial working memory experiment conducted by coauthor Fortin's lab to study the neuronal learning process on the sequential ordering of odors (Allen *et al.*, 2016). The 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 in-sequence (ABCDE) or out-sequence (eg, AABDE, ABCDD) by holding their nose in the port until the signal was confirmed or withdrawing before the signal, respectively. LFPs were recorded from 12 microelectrodes that were implanted into rats' hippocampus. The major scientific question of interest here is to explore and understand (latent) features in LFP signals that are associated with neural mechanism in developing sequential odor memory. To demonstrate the power of clustering analysis, we conducted an exploratory analysis. Figure 1 (this figure appears in color in the electronic version of this article, and any mention of color refers to that version) presents the smoothed LFPs from one rat across 12 microelectrodes for five odors (ABCDE) and their associated mean signals aggregated over odors. In each plot, x-axis represents time rescaled to interval [0, 1], while y-axis represents the different microelectrodes (channels). 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 tetrodes within each odor. Across all the six heatmaps, we observe that tetrodes 1 to 8 and tetrodes 9 to 12 form two different paradigms separately across tetrodes for odors A, B, and D. Moreover, those two "paradigms" also evolve as time changes, which suggests that 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. Specifically, an ideal clustering approach should be able to discover "latent features" in LFP in the following ways: (a) Row-wise (within tetrode) and column-wise (between tetrodes) correlation should be identified and hence provide insights on spatial and temporal dependencies; (b) it should be capable of uncovering the true mean difference by introducing regularization terms and thus improve the SNR of the LFP signals; (c) it should be able to shed light on the nature of sparsity inherent from the data (eg, detecting boundary of image signals); and (d) by applying the proposed method to both time and frequency domains, it should be able to identify different “structures” that are not easily discernible by mere visual inspection of these signals.

In this paper, we focus on using finite mixture models for the purpose of clustering because statistical inference can be carried out in a computationally efficient and conceptually simpler way and the results have a nice probabilistic interpretation. Existing approaches such as biclustering (Chi *et al.*, 2017), hierarchical clustering (Euan *et al.*, 2018, 2019), spectral clustering (Ng *et al.*, 2002), and mixture models (of random vectors) are not directly applicable for matrix data analysis since the set of input covariates are treated as a vector, where the matrix structure and its interpretability are not taken into account (Reiss and Ogden, 2007). Moreover, by vectorizing a matrix, the resulting dimension of the input space can be extremely large, ie, a $p \times q$ matrix will be converted to a pq-dimensional vector, which creates additional challenges in both computation and theory.

To solve the aforementioned issues, we propose a novel penalized mixture model for clustering matrix-valued data. The framework is inspired by the mixture model proposed by Viroli (2011) and Gao *et al.* (2019) where each mixture component 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 (Dawid, 1981; Dutilleul, 1999). The separable covariance structure provides both computational convenience, since it effectively reduces the number of covariance parameters, and useful 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 (ie, ℓ_1, ℓ_2 and the nuclear norm) to give the method flexibility and robustness in capturing the different features in the mean structures, and therefore, enhance the ability to correctly identify the clusters. For example, the use of nuclear norm provides a useful low-rank approximation of the true image (Zhou and Li, 2014), and the use of ℓ_1-norm is helpful for detecting image boundaries (Wang *et al.*, 2017). We introduce a new expectation maximization (EM)-type of algorithm that allows efficient computation for all three penalization norms. In theory, we show a strong consistency result for the proposed estimator using the technique by Fan and Li (2001) which we have to modify to accommodate the matrix-valued data.

Note that matrix normal distribution and its separable covariance structure serve as the building blocks for our proposed mixture model. Although not within the scope of our current paper, it is possible to generalize the matrix normal distribution and consider a more general class of covariance models such as the spiked covariance model (Donoho *et al.*, 2018). We expect our proposed idea of applying different regularization norms on matrix mean signals still useful in those scenarios.

The rest of the paper is organized as follows. In Section 2, we give a brief review of the matrix normal distribution and then introduce the proposed penalized matrix normal mixture model. We propose a new EM-type of algorithm for computation and discuss a related one-step-late (OSL) algorithm. In Section 3, we study the theoretical properties of the proposed method by showing a consistency result for the (regularized) estimators. We evaluate the finite-sample performance by simulation studies in Section 4. Finally, we apply the proposed method to analyze two LFP datasets obtained from imaging studies on odor memory and stroke experiment in Sections 5 and 6. Additional numerical results, technical proofs, and a description of computational code and data sets are provided in the Supporting Information.

2 | METHOD

We begin with a review of matrix normal distribution and its finite mixture in Sections 2.1 and 2.2. Mixture models of matrix normal distribution and their applications for classifying three-way data have been previously discussed by Viroli (2011). We will discuss and highlight our main contributions in Section 2.3.

2.1 | Matrix normal distribution

We give a brief review of matrix normal distribution. This distribution will be used to model the matrix-valued image data. We define an $r \times p$ random matrix Y to have 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

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

where $M \in \mathbb{R}^{r \times p}$, $U \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{p \times p}$, matrices U and V are treated as between- and within-covariance matrices, $|\cdot|$ and tr denote the determinant and the trace of a matrix, respectively. In Gupta and Nagar (1999), an equivalent definition of (1) is given by

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

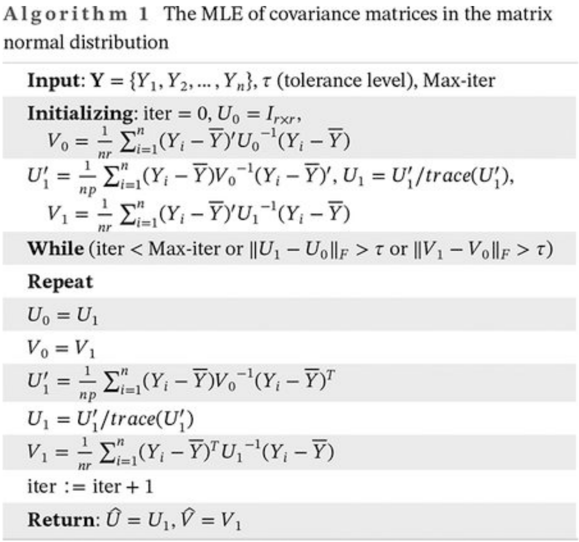
where vec is the column vectorization operation and ⊗ is the Kronecker product. The matrix normal distribution provides a natural extension of the usual multivariate normal distribution for modeling image and spatial-temporal data. For example, in spatial-temporal statistics, the row and column directions in the matrix correspond to the spatial and temporal attributes, respectively. The separability of the covariance structure is particularly useful in applications such as spatial statistics (Haas 1995; Cressie 2015) and electrophysiological data (Gao *et al.*, 2020). It is believed that such assumption drastically alleviates the number of parameters and thus provides a computational efficient and robust procedure for large scaled spatial-temporal data (Genton, 2007).

Statistical inference for the matrix normal distribution is usually conducted via the likelihood function. Given i.i.d. observations $Y_1, Y_2, \ldots, Y_n \sim MN_{r,p}(M, U, V)$, the log-likelihood function is

$$\begin{aligned}\ell(M, U, V) = & -\frac{npr}{2}\log 2\pi - \frac{nr}{2}\log|V| - \frac{np}{2}\log|U| \\ & -\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\left\{V^{-1}(Y_i - M)^T U^{-1}(Y_i - M)\right\}.\end{aligned} \tag{3}$$

The resulting maximum likelihood estimator (MLE) for M, U, and V are as follows:

$$\widehat{M} = \overline{Y}, \widehat{U} = \frac{1}{np}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\widehat{V}^{-1}\left(Y_i - \overline{Y}\right)^T,$$
$$\widehat{V} = \frac{1}{nr}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^T\widehat{U}^{-1}\left(Y_i - \overline{Y}\right). \tag{4}$$



Algorithm 1 The MLE of covariance matrices in the matrix normal distribution

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

Note that both U and V are identifiable only up to a constant multiple (Dutilleul, 1999), but their Kronecker product $U \otimes V$ remains invariant. There is no closed-form solution for $\widehat{U}$ and $\widehat{V}$. One can utilize iterative algorithm to obtain their values numerically as summarized in Algorithm 1.

2.2 | Matrix-normal mixture model

For the purpose of probabilistic clustering, we consider a matrix normal mixture model and its inference using the EM algorithm. Given i.i.d. $(r \times p)$-dimensional observations $Y_1, \ldots, Y_n$ 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

$$\sum_{k=1}^{K}\pi_k f(Y \mid \Theta_k) = \sum_{k=1}^{K}\pi_k MN_{r,p}(M_k, U_k, V_k), \tag{5}$$

where f is the matrix normal distribution with mean M_k and covariances U_k and V_k as defined by (1). We use $\Theta = (\Theta_1, \ldots, \Theta_K; \pi_1, \ldots, \pi_K)$ to denote the collection of parameters in (5). Then the log-likelihood function is

$$\ell_{obs}(\Theta) = \sum_{i=1}^{n}\log\left\{\sum_{j=1}^{K}\pi_j f\left(Y_i \mid \Theta_j\right)\right\}. \tag{6}$$

EM algorithm (Dempster *et al.*, 1977) has widely been used in all areas of statistics and has been demonstrated to work well for Gaussian mixture model. Thus, we also employ the EM algorithm to conduct inference, which is an iterative approach consisting of the expectation (E) and maximization (M) steps.

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

$$\alpha_{ij} = \frac{\pi_j f\left(Y_i \mid \Theta_j\right)}{\sum_{l=1}^{K} \pi_l f\left(Y_i \mid \Theta_l\right)}. \tag{7}$$

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

$$\widehat{\Theta}_j = \underset{\Theta_j}{\arg\max} \sum_{i=1}^{n} \sum_{j=1}^{K} \alpha_{ij} \log\left\{\pi_j f\left(Y_i \mid \Theta_j\right)\right\}.$$

By some algebra, we obtain the following relationship:

$$\begin{aligned} \widehat{\pi}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij}}{n}, \quad \widehat{M}_j = \frac{\sum_{i=1}^{n} \alpha_{ij} Y_i}{\sum_{i=1}^{n} \alpha_{ij}}, \\ \widehat{U}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij}\left(Y_i - \widehat{M}_j\right)\widehat{V}_j^{-1}\left(Y_i - \widehat{M}_j\right)'}{p\sum_{i=1}^{n} \alpha_{ij}}, \\ \widehat{V}_j &= \frac{\sum_{i=1}^{n} \alpha_{ij}\left(Y_i - \widehat{M}_j\right)'\widehat{U}_j^{-1}\left(Y_i - \widehat{M}_j\right)}{r\sum_{i=1}^{n} \alpha_{ij}}. \end{aligned} \tag{8}$$

Note that $\widehat{U}_j$ and $\widehat{V}_j$, $j = 1, \ldots, k$ do not have closed-form solutions, but can be obtained numerically following the similar steps in Algorithm 1.

2.3 | Penalized matrix normal mixture model

Mixture of matrix normal model has previously been discussed by Viroli (2011) 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 (eg, 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 (Zhou and Li, 2014) or total-variation-norm-based penalization (Wang *et al.*, 2017). 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 (Green, 1990), which gives results that are easily interpretable and also eases the computational burden. Specifically, we consider the penalized log-likelihood function

$$Q(\Theta; \lambda) = \sum_{i=1}^{n} \log\left\{ \sum_{j=1}^{K} \pi_j f\left(Y_i \mid \Theta_j\right) \right\} - \lambda \sum_{j=1}^{K} P\left(M_j\right), \tag{9}$$

where $P(\cdot)$ is some penalty function, such as ℓ_1, ℓ_2, and nuclear norms, and $\lambda \geq 0$ is the tuning parameter. For ℓ_1 and ℓ_2-penalty, the norms are defined on vectorized matrix means M_j, and for the nuclear norm penalty, it is defined as the sum of singular values of M_j. The proposed model in (9) is general and, in fact, setting $\lambda = 0$ is equivalent to the mixture model of matrix normal with no regularization as introduced in Viroli (2011).

Similar to Section 2.2, we propose a modified EM algorithm to estimate the parameters. The E-step proceeds in the same way as Equation (7). The M-step boils down to solving an optimization problem,

$$\widehat{\Theta} = \underset{\Theta}{\arg\max} \sum_{i=1}^{n} \sum_{j=1}^{K} \alpha_{ij} \log\left\{\pi_j f\left(Y_i \mid \Theta_j\right)\right\} - \lambda \sum_{j=1}^{K} P\left(M_j\right). \tag{10}$$

Note that the solution $\widehat{\Theta}$ may not have an explicit form. Lange (1995) proposed a gradient method related to the EM algorithm. It replaces the M-step by conducting one iteration of Newton's method. Alternative approaches, such as surrogate functions (Lange *et al.*, 2000) and overrelaxed EM algorithm (Yu, 2012), have also been introduced in the literature. Pan and Shen (2007) 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 subgradient approach. Green (1990) developed the "OSL" algorithm that can be applied to more general case. Inspired by the aforementioned methods, we develop a subgradient 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 Pan and Shen (2007), M_j can be updated by

$$\widehat{M_j} = \operatorname{sign}\left(\widetilde{M}_j\right)\left(\left|\widetilde{M}_j\right| - \frac{\lambda}{\sum_{i=1}^{n} \alpha_{i,j}} U_i \mathbf{1}_{r \times p} V_i\right)_+, \quad j = 1, \ldots, K, \tag{11}$$

where $\widetilde{M}_j = \frac{\sum_{i=1}^{n} \alpha_{i,j} Y_i}{\sum_{i=1}^{n} \alpha_{i,j}}$ is the update for M_i without the 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 partial derivative of the objective function $Q_{\ell_2}(\Theta)$ is

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

Therefore, M_j can be updated by

$$\widehat{\boldsymbol{M}}_j = \widetilde{\boldsymbol{M}}_j - \frac{2\lambda}{\sum_{i=1}^{n} \alpha_{ij}} U_j M_j V_j, \tag{12}$$

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

For the nuclear norm penalty $\| \cdot \|_*$, which is defined as the sum of singular values of a matrix, similar derivation yields

$$\widehat{\boldsymbol{M}}_j = \widetilde{\boldsymbol{M}}_j - \frac{\lambda}{\sum_{i=1}^{n} \alpha_{ij}} U_j \Phi_j \Omega_j^T V_j, \tag{13}$$

where M_j has the singular value decomposition $M_j = \Phi_j \Lambda_j \Omega_j^T$. The use of nuclear norm regularization is essentially equivalent with L_1-regularization on the singular values of M. More details of derivations of (11) to (13) are given in Section 3.3 of the Supporting Information.

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

I. (Initialization)—Pick a fixed clustering number K. We start with vectorizing the original matrix-valued observations $Y_1, \ldots, Y_n$ and apply K-means to achieve the initial cluster membership values, written as $S_1, \ldots, S_K$, where $S_j = \{i \mid Y_i \text{ in } j\text{th cluster}\}$. Note that alternative methods to K-means can also be used in this step, eg, randomly assign observations to different clusters. Then for each cluster, the initial value of Θ_j can be obtained following the approach in Section 2.1, and $\boldsymbol{\pi}_j$ can be directly estimated by $\hat{\pi}_j = |S_j|/n$, where $|S_j|$ is the cardinality of S_j.

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

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

III. (M-step)—Update the mean parameters M_j with respect to various penalties by Equations (11) to (13), respectively. Other parameters $\boldsymbol{\pi}_j$, U_j, V_j can be updated following Equation (8) and Algorithm 1.

IV. (Stopping criteria)—Repeat II and III until certain number of iterations have been reached or the change of the estimate of the mean parameter M_j (in terms of Frobenius norm) is below certain prespecified cutoff.

IV. (Choosing the number of clusters)—A key question in clustering is to determine the number of clusters. Inspired by Smyth (2000), we consider a predictive criteria by adopting the cross-validated penalized likelihood (CVPL) as the key measure. We split the dataset $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}$ into training and testing groups denoted by $\boldsymbol{Y}_{\text{train}}$, $\boldsymbol{Y}_{\text{test}}$, and then fit a k-mixture model (for $k = 1, \ldots$) on $\boldsymbol{Y}_{\text{train}}$ and then 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

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

where $\boldsymbol{Y}^{-l}$ is the data $\boldsymbol{Y}$ excluding $\boldsymbol{Y}^l$. Then we choose $\hat{k} = \operatorname{argmax}_k Q(L, k)$ for a preselected L. In numerical analysis, we choose $L = 3$, ie, three-fold cross validation, for computational convenience. Choosing a larger value for L is also possible, and it will be of future interest to investigate the stability of CVPL as a model selection criteria under different values for L. We will give more details for the calculation in the simulation study.

3 | THEORY

In this section, we study the theoretical properties for the proposed mixture of matrix normal model, assuming that the true number of clusters, denoted by K, is known. The theoretical understanding of mixture models (in particular Gaussian mixture model) has received attention (in particular, see Chen, 2017). Here, we first give a consistency result for the MLE of the matrix normal mixture model. It is well known that the likelihood for the mixture of univariate normal distributions is unbounded if there is no constraint on the parameter space. Therefore, we consider a constrained parameter space Ψ^{d_1, d_2} as

$$\Psi^{d_1, d_2} = \Big\{\Theta : (\pi_1, \ldots, \pi_k)^T \in \Delta_K, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_k \in \mathbb{R}^{r \times p} \times \boldsymbol{V}_1, \ldots, \boldsymbol{V}_K \in \mathbb{R}^{r \times r}, \boldsymbol{U}_1, \\ \ldots, \boldsymbol{U}_k \in \mathbb{R}^{p \times p} : \times \min_{1 \le h \ne j \le k} \rho\left(\boldsymbol{U}_h \boldsymbol{U}_j^{-1}\right) \ge d_1, \times \min_{1 \le h' \ne j' \le k} \rho\left(\boldsymbol{V}_{h'} \boldsymbol{V}_{j'}^{-1}\right) \ge d_2, \\ \rho(\boldsymbol{U}_l) > 0, \rho(\boldsymbol{V}_l) > 0 \text{ for } l = 1, \ldots, K\,.\Big\}, \qquad (14)$$

where d_1, d_2 are two fixed constants defined on (0, 1], Δ_K denotes a K-dimensional simplex, and $\rho(\cdot)$ denotes the minimum eigenvalue of a matrix. We first state a consistency result for the matrix normal mixture model. The proof is given in the Web Appendix C.

Theorem 1.

Let $Y_1, \ldots, Y_n$ be a random sample from a mixture of matrix normal distribution as defined in (5) indexed with parameter Θ that belongs to the space Ψ^{d_1, d_2} defined by (14). *Assume that the true parameter value $\Theta_0 \in \Psi^{d_1, d_2}$ and denote $\hat{\Theta}_n$ as the solution of* $\arg\max_{\Psi^{d_1, d_2}} \ell_{obs}(\Theta)$. *Then $\hat{\Theta}_n$ converges to Θ_0 almost surely.*

Here, $\Theta_0 = (\boldsymbol{\pi}_{10}, \ldots, \boldsymbol{\pi}_{K0}; M_{10}, \ldots, M_{K0}; V_{10} \otimes U_{10}, \ldots, V_{K0} \otimes U_{K0})$ is the collection of the true mixture model parameters, $\hat{\Theta}_n = \left(\hat{\pi}_1, \ldots, \hat{\pi}_K; \hat{M}_1, \ldots, \hat{M}_K; \hat{V}_1 \otimes \hat{U}_1, \ldots, \hat{V}_K \otimes \hat{U}_K\right)$, and the convergence of $\hat{\Theta}_n$ to Θ_0 means the convergence of each component in Θ. The condition of the parameter lying in (14) is not easy to check in practice. A sufficient condition is to bound all the eigenvalues within a given interval (a, b) for numerical stability (Ingrassia and Rocci, 2007).

Next, we show that under mild conditions, the penalized likelihood estimator of (9) is consistent. It provides a theoretical justification for the use of regularization within the matrix normal mixture model. We define a new parameter space $\bar{\Psi}^{d_1, d_2}$ as

$$\bar{\Psi}^{d_1, d_2} = \Big\{ (\pi_1, \dots, \pi_K, M_1, \dots, M_K, V_1 \otimes U_1, \dots, V_K \otimes U_K) \in \Psi^{d_1, d_2} : \frac{\sigma_i(U_h)}{\sigma_i(V_h)} = c_h \text{ for } i = 1, \dots, \min\{r, p\}, \text{ and } h = 1, \dots, K \Big\}, \tag{15}$$

where $\sigma_i(U_h)$ denotes the ith eigenvalue of matrix U_h and c_h is a positive constant for every $h = 1, \dots, K$.

Theorem 2.

Let $Y_1, \dots, Y_n$ be a random sample from a mixture matrix normal distribution (5). Let $\hat{\Theta}_\lambda$ be a maximizer of the penalized likelihood (9) where the penalty function takes in the form of ℓ_1, ℓ_2, or nuclear norms. Assume that the true parameter value $\Theta_0 \in \bar{\Psi}^{d_1, d_2}$, and $\lambda \to 0$ as $n \to \infty$, then $\|\hat{\Theta}_\lambda - \Theta_0\| = o_p(1)$.

Here, $\|\hat{\Theta}_\lambda - \Theta_0\| = \sum_{i=1}^{K} \left\{ |\hat{\pi}_i - \pi_{i0}| + \|\hat{M}_i - M_{i0}\|_F + \|\hat{V}_i \otimes \hat{U}_i - V_{i0} \otimes U_{i0}\|_F \right\}$. It is possible to establish the asymptotic normality and oracle property for the regularized estimator using the techniques in Fan and Li (2001) under stronger conditions. However, the focus here is the introduction of new methodology and its application in image clustering; therefore, we choose not to pursue this direction in the paper.

4 | SIMULATIONS

We first evaluate whether the proposed CVPL criteria is able to identify the correct number of clusters under different scenarios. All of the simulation results are based on 200 Monte-Carlo replications; and the reported CVPL and ARI values are averaged over those replications. In Scenarios I and II, we generate data from a mixture of two matrix normal distributions with equal proportions and mean structures of a cross and a rectangle shape, as shown in Web Figure 1. In both scenarios, the row-wise and column-wise covariance matrices follow an autoregressive setting where $\mathrm{cov}\{Y_{k_1, l_1}, Y_{k_2, l_2}\} = 0.9^{|k_1 - k_2| + |l_1 - l_2|}$, $1 \le k_i \le r$, $1 \le l_i \le p$. In Scenario I, we set the sample size $n = 500$ and image size $r = p = 60$. In Scenario II, we let $n = 100$, $r = p = 80$.

We apply the proposed method with ℓ_1, ℓ_2 and nuclear norm penalties, and summarize the results in Web Table 1. As noted, when $\lambda = 0$, our model is equivalent with the mixture model without regularization as proposed by Viroli (2011). It can be seen that the proposed method manages to choose the true number of clusters ($K = 2$) under most cases based on the maximum value of CVPL. Over the three regularization methods, by comparing the CVPL values, nuclear norm outperforms both the ℓ_1 and ℓ_2 norms and ℓ_2 norm performs slightly better than ℓ_1 norm. This is expected because the true mean structure has a low rank but is not entry-wise sparse.

To achieve a realistic signal noise ratio (SNR), in Scenario III, we utilize the estimated parameters from the real data example in Section 5.2 (Theta band time-frequency analysis under two clusters for Theta activity oscillations between 4 and 8 Hz). We set the sample size $n = 500$ and apply the proposed approach. Web Table 2 shows that the CVPL values are in favor of two clusters, which is consistent with our settings. In Scenario IV, we generate synthetic signals by adding white noise to the original images in Scenario II such that the covariances are no longer separable. As shown in Web Table 2, the proposed approach can identify the right number of clusters for most cases and ℓ_2 regularization seems a proper choice. Compared with simulation scenarios I and II, clustering is more challenging in Scenarios III and IV given the increased noise level. Therefore, the latent sparsity structure in images become more difficult to be detected, which is one reason why we do not observe a substantial difference in terms of CVPL values between $\lambda = 0$ (no regularization) and $\lambda \neq 0$. For Scenarios I to IV, we also present the percentage of correct detection of the number of clusters (over 200 Monte-Carlo replications) based on the highest CVPL values in Web Table 8. We find that in most cases, the correct K can be identified with more than 70% probability. For those cases with selection probability below 70%, one possible explanation is that there is another “competitive” K (usually next to the true K) whose CVPL value is very close to the CVPL value from the true K, eg, Scenario I with no penalty.

We also conduct an in-depth study to understand how the proposed three regularization methods perform over different sparsity levels. In Scenario V, we generate mean structures of a triangle and square with values 0 and 1. In Scenario VI, we follow the same setting as V except that we change all zero values of the triangle to small numbers close to 0 by adding a Gaussian noise. In Scenario VII, we consider a new mean structure of separated squares (see Web Figure 6). We then apply the proposed approach and calculate the mean square errors (MSEs) of the mean parameter estimates. From Web Table 7, it can be seen that using proper regularization penalties would result in more accurate mean parameter estimates. In Scenario V where values are 0 and 1, ℓ_1 with $\lambda = 0.5$ outperforms the other penalties and settings. The results are also consistent after comparing their CVPL values. Moreover, ℓ_2 with $\lambda = 0.2$ and nuclear norm with $\lambda = 1$ works the best for Scenarios VI and VII, respectively. The results confirm the intuition that ℓ_1 norm is in favor of zeroes, ℓ_2 norm shrinks small values, and nuclear norm pushes for a low-rank structure (eg, Scenario VII).

We then compare the performance of the proposed approach with other competing approaches: K-means, biclustering (Turner *et al.*, 2005), and spectral clustering (Ng *et al.*, 2002). We generate signals using the same settings as in Scenario I to VII. To evaluate the performance, we calculate the adjusted random index (ARI) (Milligan and Cooper, 1986) to compare clustering results with the underlying truth. ARI is a number that has a maximum value of 1; and it measures the agreement between two clustering solutions (even when the two clustering solutions have different number of clusters). In addition, we consider the prediction accuracy for all methods. Web Tables 3 to 6 summarize the results for Scenario I to IV. The proposed approach outperforms all three competing methods under all scenarios, possibly because our method takes account for the underlying sparse structure in the signals. For example, in Web Table 4, the proposed approach outperforms all three competing methods while biclustering performs slightly better than K-means and spectral clustering. In Scenario IV, our method still has a better predictive performance over the competitive methods even when the separable covariance assumption is violated in the data generation.

It is worth mentioning that our primary goal in the simulation is to demonstrate that CVPL is useful in guiding the selection of tuning parameters and the regularization effect on image clustering in an efficient manner. If one is more interested in selecting the "best" model based on the highest CVPL value. Then ideally, this can be done by conducting a fine grid search over all possible combinations of three tuning parameters (regularization norm, λ, and number of clusters). In our simulation studies, we first search over a wide range of values for both λ (eg, {0,.2,.5,.8,1,1.5,2,5,10}) and K (eg, 1 to 6) over a few Monte-Carlo replications, and then narrow down to a small range for computational convenience.

5 | ANALYSIS OF ODOR MEMORY DATA IN A RAT NEUROBIOLOGY EXPERIMENT

In this section, we use the proposed method to analyze an LFP dataset obtained from a memory coding experiment on nonspatial events (Allen *et al.*, 2016; Hu *et al.*, 2020). 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" yet $ABBDE$ is an "out-sequence." 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 Hz and $T = 2000$ time points. Web Figure 3 gives a snapshot of the LFP signals across 12 tetrodes. A clustering analysis is useful for this study because it may reveal latent pattern information in LFP signals and provide an in-depth understanding of their connections to different odors and in-/out-sequences.

5.1 | Time domain analysis on imaging clustering

As an initial step, we focus on the time domain to study the association between raw multimicroelectrode signals with "in-sequence" or "out-sequence" patterns. We implement the proposed method to the raw LFP signals across all the 247 trials. Table 1 summarizes the CVPL values among different number of clusters and penalties. Based on the highest CVPL value, our method chooses two clusters for all penalty norms, which correspond to the inand out-sequences. Moreover, our method has a significantly higher ARI value compared with that of K-means, which suggests that the proposed method has desired performance in detecting the latent structure that is related to "in-" or "out"-sequences.

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. Web Table 9 summarizes the CVPL and ARI values. Based on the highest CVPL value, nuclear norm regularization with $K = 2$ clusters is preferred. Based on ARI values, both ℓ_1 and nuclear norm regularization perform better than ℓ_2 norm, which may suggest a possible low-rank and sparse mean structure in the signals. For the number of clusters, $K = 2$ is preferred since its CVPL values are in general (6 out of 10 times) higher than those obtained by other choices of K under the use of different regularization norms. The ARI values are much higher from our method compared with K-means and other competing approaches.

5.2 | Time frequency clustering analysis

Next, we study the latent structure from a time-frequency perspective. Allen *et al.* (2016) suggest that two particular oscillatory bands (theta: 4 to 8 Hz and slow gamma: 20 to 40 Hz) yield strong power and play significant roles in detecting the in-/out-sequences. Web Figure 3 shows that 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 2 presents the results for the number of clusters and the ARI that compares with the true odor sequence for the theta band. Based on CVPL values, we find that a three-cluster model will be preferred if no penalty is used and a five-cluster model with nuclear norm regularization is the best among models that adopts a penalty. We then look into those two models in depth and it turns out that time-frequency signals related to odors A, B, and D are quite similar visually. That explains the different results between the two approaches in some way and demonstrates a clear advantage of our proposed model. By adding the nuclear norm penalty, the method is able to incorporate more information into the clustering procedure, which results in a larger number of clusters. The ARI value also falls slightly in favor of nuclear norm regularization. We then looked into results from other methods, similar to the previous observation, odors C and E result in the most consistent results, while odors A, B, and D lead to different groupings. Our approach provides some evidence indicating the association between the low-frequency band (theta) and the odor sequence. We also applied our method to the slow gamma band and obtained similar results (presented in the Supporting Information, Web Table 10).

6 | ANALYSIS OF RAT STROKE DATA

In this section, we apply the proposed approach to another LFPs dataset from a rat stroke experiment conducted by the Frostig laboratory at UC Irvine, where LFPs were recorded before and after the stroke. There are 32 microelectrodes implanted with four layers for each rat. The data we consider here are collected based on the signals of 5 minutes before and after the stroke. The sampling rate is 1000 Hz and each epoch is 1 second long. One of the scientific questions of interest from this experiment is to identify the latent patterns that change before and after the stroke. A preliminary time-frequency analysis in Figure 1 (bottom 4) shows the log power spectra of two microelectrodes. These results are obtained by averaging all epochs before and after stroke separately. The LFPs at most microelectrodes behave smoothly within each epoch and there appears to be a low level of dissimilarity before and after the stroke. However, for some microelectrode (in particular 10), there are some nonnegligible dynamics and obvious difference between prestroke and poststroke. These findings show that it may not be optimal to average over or vectorize all the channels when conducting a cluster analysis to identify the latent patterns that change before and after the stroke.

We also study the dynamics across all the 32 microelectrodes before and after the stroke. Figure 2 shows the time-frequency plot of beta and slow gamma frequency bands across the microelectrodes. The log power spectra are obtained by averaging over the epochs. By comparing the plots before and after the stroke, we observe strong dependence across microelectrodes for both bands. This highlights the importance of introducing regularization into the mixture model. Also, by comparing the plots before and after stroke, there is a clear sign of local discrepancy. Such difference will be easily ignored if one naively vectorizes the original signals when conducting cluster analysis without taking the matrix structure into account.

We apply the proposed approach to the time-frequency images across all the epochs before and after stroke. Table 3 shows the CVPL values across different number of clusters and penalty norms. With only one exception, all the scenarios suggest two clusters, which correspond to two statuses “normal” and “stroke.” We also compare the clustering results with the truth status index, and summarize the ARI of the proposed method and K-means in Table 4. There is a significant advantage in ARI for our method compared to the K-means. Our method, through regularization, produced ARI values that increase by 80% (comparing $\lambda = 0$ with $\lambda \neq 0$). In particular, slow gamma band performs perfectly with an ARI of 1.00 when a nuclear norm penalty is used with $\lambda = 2$. Similar results are also obtained for beta band. These findings are consistent with the findings in the preliminary analysis.

7 | CONCLUDING REMARKS

In this paper, we have proposed a regularized probabilistic clustering framework to analyze matrix data. Compared to the existing clustering approaches such as K-means, the advantages are as follows: (a) By working directly with matrix data, we are able to capture the row-wise and column-wise correlation simultaneously; (b) the proposed framework has the ability to uncover the nature sparsity that is inherent to the signals and images; (c) by using CVPL, the proposed method is able to find the optimal regularization method and the tuning parameters, which works for different types of signal structures and levels of sparsity; and (d) the proposed approach is grounded on theoretical foundations; provides straightforward interpretability; and has low computational cost (by parallel computing) and hence amenable to big datasets.

To incorporate different sparsity structures in images, we have proposed three regularization methods in this framework. In particular, ℓ_2 norm penalizes on large matrix element values, ℓ_1 norm targets on identifying nonzero ones, and nuclear norm encourages a low-rank approximation. In practice, one can choose the best regularization norm to use based on the CVPL criteria or prior knowledge about the image structure. Our simulation results confirm the excellent performance of the method in terms of image structure recovery and estimation accuracy improvement (eg, MSE) by using an appropriate regularization method under different scenarios.

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 relies on some prespecified measures (CVPL). As an extension of the framework introduced by Viroli (2011), one could introduce a Bayesian framework into the clustering analysis and conduct Bayesian inference on the number of clusters. Similarly with the use of elastic net in regression models, it will be of future interest to consider a combination of both ℓ_1 and ℓ_2 penalties in our model. Incorporating regularized covariance matrix estimation in the proposed mixture model will also be helpful, especially for analyzing images with large sizes. Another future working direction of interest is to develop inference procedures that can be used to quantify the similarity (and its uncertainty) between the cluster centroids and image signals for prediction purpose.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

The authors thank the editor, the associate editor, and two referees for their constructive and helpful comments on the earlier version of this paper. Shen’s research was partially supported by the Simons Foundation Award 512620 and the NSF Grant DMS 1509023. Hu’s effort was partially supported by the National Institute of Health Grants R01AI143886, R01CA219896, and CCSG P30 CA013696. Fortin’s research was supported by NIH grant R01-MH115697, NSF awards IOS-1150292 and BCS-1439267, and Whitehall Foundation award 2010-05-84. Frostig’s research was partially supported by the Leducq Foundation.

Funding information

Simons Foundation, Grant/Award Number: 512620; National Institute of Health, Grant/Award Numbers: R01AI143886, R01CA219896, CCSG P30 CA013696; NIH, Grant/Award Number: R01-MH115697; NSF, Grant/Award Numbers: IOS-1150292, BCS-1439267; Whitehall Foundation, Grant/Award Number: 2010-05-84

REFERENCES

Allen TA, Salz DM, McKenzie S and Fortin NJ (2016) Nonspatial sequence coding in ca1 neurons. Journal of Neuroscience, 36, 1547–1563. [PubMed: 26843637]

Chen J (2017) Consistency of the MLE under mixture models. Statistical Science, 32, 47–63.

Chi EC, Allen GI and Baraniuk RG (2017) Convex biclustering. Biometrics, 73, 10–19. [PubMed: 27163413]

Cressie N (2015) Statistics for Spatial Data. New York: John Wiley & Sons.

Dawid AP (1981) Some matrix-variate distribution theory: notational considerations and a Bayesian application. Biometrika, 68, 265–274.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39, 1–38.

Donoho DL, Gavish M and Johnstone IM (2018) Optimal shrinkage of eigenvalues in the spiked covariance model. Annals of Statistics, 46, 1742–1778. [PubMed: 30258255]

Dutilleul P (1999) The MLE algorithm for the matrix normal distribution. Journal of Statistical Computation and Simulation, 64, 105–123.

Euan C, Ombao H and Ortega J (2018) The hierarchical spectral merger algorithm: a new time series clustering. Journal of Classification, 35, 71–99.

Euan C, Sun Y and Ombao H (2019) Coherence-based time series clustering for brain connectivity visualization. Annals of Applied Statistics, 13, 990–115.

Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96, 1348–1360.

Gao X, Shen W, Hu J, Fortin N and Ombao H (2019) Modeling local field potentials with regularized matrix data clustering. In 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 597–602. IEEE.

Gao X, Shen W, Shahbaba B, Fortin N and Ombao H (2020) Evolutionary state-space model and its application to time-frequency analysis of local field potentials. Statistica Sinica, 30, 1561–1582. [PubMed: 32774073]

Gao X, Shen W, Ting C-M, Cramer SC, Srinivasan R and Ombao H (2019) Estimating brain connectivity using copula Gaussian graphical models. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 108–112. IEEE.

Genton MG (2007) Separable approximations of space-time covariance matrices. Environmetrics, 18, 681–695.

Green PJ (1990) On use of the EM for penalized likelihood estimation. Journal of the Royal Statistical Society. Series B (Methodological), 52, 443–452.

Gupta AK and Nagar DK (1999) Matrix Variate Distributions, Volume 104. Boca Raton, FL: CRC Press.

Haas TC (1995) Local prediction of a spatio-temporal process with an application to wet sulfate deposition. Journal of the American Statistical Association, 90, 1189–1199.

Hu L, Fortin NJ and Ombao H (2019) Modeling high-dimensional multichannel brain signals. Statistics in Biosciences, 11, 91–126.

Hu L, Guindani M, Fortin N and Ombao H (2020) A hierarchical Bayesian model for differential connectivity in multi-trial brain signals. Econometrics and Statistics, 15, 117–135. [PubMed: 33163735]

Ingrassia S and Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics &* Data Analysis, 51, 5339–5351.

King RS (2015) Cluster Analysis and Data Mining: An Introduction. Virginia: Stylus Publishing, LLC.

Lange K (1995) A gradient algorithm locally equivalent to the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 57, 425–437.

Lange K, Hunter DR and Yang I (2000) Optimization transfer using surrogate objective functions. Journal of Computational and Graphical Statistics, 9, 1–20.

Milligan GW and Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21, 441–458. [PubMed: 26828221]

Ng AY, Jordan MI and Weiss Y (2002) On spectral clustering: analysis and an algorithm. In Conference on Advances in Neural Information Processing Systems, pp. 849–856.

Pan W and Shen X (2007) Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research, 8, 1145–1164.

Reiss PT and Ogden RT (2007) Functional principal component regression and functional partial least squares. Journal of the American Statistical Association, 102, 984–996.

Smyth P (2000) Model selection for probabilistic clustering using cross-validated likelihood. Statistics and Computing, 10, 63–72.

Turner H, Bailey T and Krzanowski W (2005) Improved biclustering of microarray data demonstrated through systematic performance tests. Computational Statistics & Data Analysis, 48, 235–254.

Viroli C (2011) Finite mixtures of matrix normal distributions for classifying three-way data. Statistics and Computing, 21, 511–522.

Viroli C (2011) Model based clustering for three-way data structures. Bayesian Analysis, 6, 573–602.

Wang X, Zhu H and ADNI, (2017) Generalized scalar-on-image regression models via total variation. Journal of the American Statistical Association, 112, 1156–1168. [PubMed: 29151658]

Wang Y, Ting C-M, Gao X and Ombao H (2019) Exploratory analysis of brain signals through low dimensional embedding. In 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 997–1002. IEEE.

Yu Y (2012) Monotonically overrelaxed EM algorithms. Journal of Computational and Graphical Statistics, 21, 518–537.

Zhou H and Li L (2014) Regularized matrix regression. Journal of the Royal Statistical Society. Series B (Methodological), 76, 463–483.

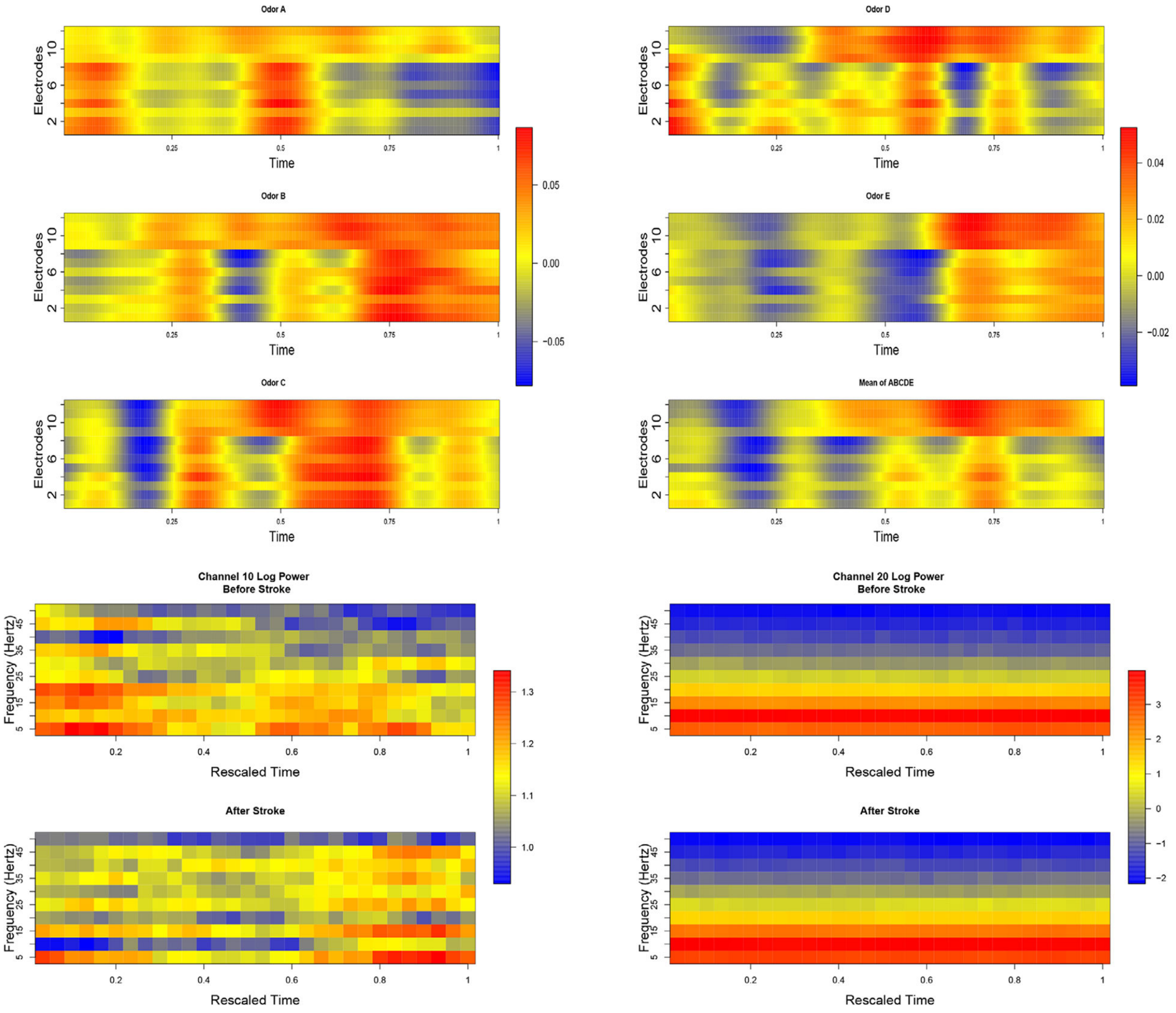


FIGURE 1.

Top 6 (memory coding experiment): smoothed LFPs for odors ABCDE and their means (aggregated over all the odors) where *x*-axis represents rescaled time and *y*-axis stands for different tetrodes (channels); Bottom 4(rat stroke study): the time-frequency plot of microelectrodes 10 and 20 among all the 600 epochs for before and after the stroke. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

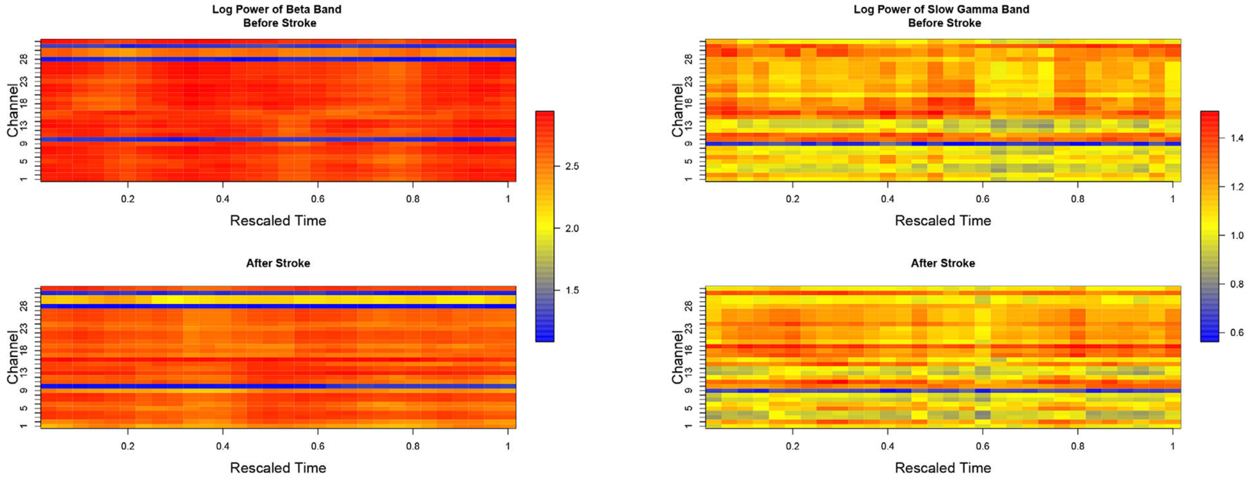


FIGURE 2.

Rat stroke study: the time-frequency plot of particular frequency bands among all the microelectrodes before and after stroke. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 1

Odor memory study (time domain): The cross-validated penalized likelihood (CVPL) and adjusted random index (ARI) values for different number of clusters and penalties

		CVPL				ARI			
Penalty	λ	$K = 1$	$K = 2$	$K = 3$	$K = 4$	**Our method**	***K*-means**	**Bicluster**	**Spectral**
	0	1.243	1.290[a]	1.285	1.281	0.768			
ℓ_1	0.5	1.222	1.253[a]	1.253[a]	1.246	0.786	0.499	0.583	0.512
	1	1.217	1.243[a]	1.206	1.204	0.768			
	1.5	1.208	1.249[a]	1.234	1.218	0.780			
ℓ_2	0.5	1.264	1.302[a]	1.107	1.240	0.768			
	1	1.267	1.301[a]	1.027	1.202	0.774	0.499	0.583	0.512
	1.5	1.263	1.298[a]	1.189	1.235	0.756			
Nuclear	0.5	1.253	1.309[a]	1.299	1.274	0.756			
	1	1.252	1.299[a]	1.287	1.277	0.733	0.499	0.583	0.512
	1.5	1.249	1.290[a]	1.286	1.214	0.711			

[a] *a*The highest CVPL value for each row (×10^5).

TABLE 2

Odor memory study (time-frequency domain): the CVPL and ARI values obtained from the “in-sequence” trials based on the theta band spectrum

		CVPL					ARI			
Penalty	λ	***K* = 1**	***K* = 2**	***K* = 3**	***K* = 4**	***K* = 5**	**Ours**	***K*-means**	**Bicluster**	**Spectral**
	0	9.432	11.001	11.300[a]	11.198	11.172	0.712			
ℓ_1	0.5	7.285	8.516	8.975	8.849	8.997[a]	0.692	0.674	0.681	0.678
	1	7.345	8.650	8.632	8.725	8.745[a]	0.703			
	1.5	7.324	8.571	8.705[a]	8.556	8.701	0.709			
ℓ_2	0.5	7.732	8.965[a]	8.881	8.671	7.277	0.693			
	1	7.634	8.719[a]	8.388	8.544	7.616	0.686	0.674	0.681	0.678
	1.5	7.534	8.650	8.632	8.825[a]	8.745	0.682			
Nuclear	0.5	7.430	9.034	9.196	9.183	9.259[a]	0.707			
	1	7.243	9.013	9.166	9.255	9.263[a]	0.714	0.674	0.681	0.678
	1.5	7.143	8.571	9.040[a]	8.995	8.969	0.712			

[aa] The highest value for each row (×10^3).

TABLE 3

Rat stroke study: the CVPL values for log power spectra from beta and slow gamma bands

		CVPL (slow gamma)					CVPL (beta)				
Penalty	λ	***K*** **= 1**	***K*** **= 2**	***K*** **= 3**	***K*** **= 4**	***K*** **= 5**	***K*** **= 1**	***K*** **= 2**	***K*** **= 3**	***K*** **= 4**	***K*** **= 5**
	0	2.781	2.941	2.964[a]	2.764	2.822	4.538	4.645[a]	4.627	4.526	4.598
ℓ_1	1	2.364	2.472[a]	2.031	1.513	0.421	3.832	3.980[a]	3.268	3.594	1.676
	2	2.043	2.106[a]	1.370	1.288	0.621	4.032	4.167[a]	3.373	3.227	3.277
ℓ_2	0.5	2.564	2.688[a]	2.474	2.306	2.184	4.129	4.245[a]	4.179	4.036	3.424
	1	2.354	2.484[a]	2.188	1.787	1.895	3.845	4.063[a]	3.557	3.429	3.329
	2	2.245	2.338[a]	2.163	1.539	1.733	3.743	4.024[a]	3.699	2.972	3.206
Nuclear	0.5	2.745	2.806[a]	2.627	2.502	2.303	4.356	4.464[a]	4.299	4.130	3.963
	1	2.434	2.556[a]	2.362	1.946	1.720	4.022	4.191[a]	3.977	3.618	3.371
	2	2.634	2.748[a]	1.689	1.257	0.684	3.543	3.687[a]	3.274	2.795	2.262

[aa] The highest CVPL value for each row over different frequency bands ($\times 10^4$).

TABLE 4

Rat stroke study: the ARI in relation to “Stroke” for spectrum from slow gamma and beta bands

		ARI (slow gamma)				ARI (beta)			
Penalty	λ	**Our method**	***K*-means**	**Bicluster**	**Spectral**	**Our method**	***K*-means**	**Bicluster**	**Spectral**
	0	0.507				0.887			
ℓ_1	0.5	0.981	0.751	0.783	0.764	0.942	0.715	0.741	0.724
	1	0.961				0.914			
	2	0.951				0.861			
ℓ_2	0.5	0.951				0.941			
	1	0.951	0.751	0.783	0.764	0.878	0.715	0.741	0.724
	2	0.961				0.787			
Nuclear	0.5	0.951				0.941			
	1	0.960	0.751	0.783	0.764	0.942	0.715	0.741	0.724
	2	1.000				0.951			