AUGMENTED LATENT DIRICHLET ALLOCATION (LDA) TOPIC MODEL WITH GAUSSIAN MIXTURE TOPICS

Kedar S. Prabhudesai, Boyla O. Mainsah, Leslie M. Collins, and Chandra S. Throckmorton

Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.

ABSTRACT

Latent Dirichlet allocation (LDA) is a statistical model that is often used to discover topics or themes in a large collection of documents. In the LDA model, topics are modeled as discrete distributions over a finite vocabulary of words. The LDA is also a popular choice to model other datasets spanning a discrete domain, such as population genetics and social networks. However, in order to model data spanning a continuous domain with the LDA, discrete approximations of the data need to be made. These discrete approximations to continuous data can lead to loss of information and may not represent the true structure of the underlying data. We present an augmented version of the LDA topic model, where topics are represented using Gaussian mixture models (GMMs), which are multi-modal distributions spanning a continuous domain. This augmentation of the LDA topic model with Gaussian mixture topics is denoted by the GMM-LDA model. We use Gibbs sampling to infer model parameters. We demonstrate the utility of the GMM-LDA model by applying it to the problem of clustering sleep states in electroencephalography (EEG) data. Results are presented demonstrating superior clustering performance with our GMM-LDA algorithm compared to the standard LDA and other clustering algorithms.

***Index Terms*—** Topic models, latent Dirichlet allocation, Gaussian mixture models, clustering, sleep stage analysis.

1. INTRODUCTION

The latent Dirichlet allocation (LDA) topic model [1, 2] is an unsupervised algorithm commonly used for discovering topics or themes from a corpus of documents. The LDA model assumes that each document consists of multiple topics, where each topic is a distribution over words. Each word is associated with multiple topics with varying proportions, which makes the LDA an admixture or a mixed-membership model. The admixture nature of the LDA model makes it a powerful tool to model a collection of documents, as well as similar data types with discrete structure, such as population genetics [3] and social networks [4].

Corresponding author e-mail: leslie.collins@duke.edu

The LDA topic model can potentially be extended to applications with continuous data, such as inferring brain states from neurological data. However, to apply the LDA model to data in a continuous domain, discrete approximations to the observations have to be made in order to quantize the data space [5]. Making these discrete approximations can be detrimental when training the model, as information is lost when discretizing a continuous space. In this work, we propose an extension to the LDA topic model to handle continuous data by modeling topics as a mixture of Gaussians or a Gaussian mixture model (GMM). We denote this new model as the GMM-LDA model. We apply our model to the problem of clustering sleep stages from electroencephalography (EEG) signals to evaluate our proposed GMM-LDA model.

We first describe our extension to the LDA topic model in Section 2, and provide a method to infer model parameters in Section 3. We present an application of our GMM-LDA topic model for clustering sleep stages in Section 4 and Section 5 and conclusions in Section 6.

2. TOPIC MODEL SPECIFICATION

Consider a collection of D documents. Each document d, consists of words, $w_{d,n}, n = 1, \ldots, N_d$, where N_d is the number of words in the d^{th} document. Each document can be viewed as a mixture of various topics, and each word can be attributed to one of the topics [1]. The statistical distribution over the documents can be represented by a topic model. Probabilistic graphical representations of LDA and GMM-LDA topic models are shown in Figure 1 (a) and (b), respectively. A detailed description of the notations used in both models is provided in Table 1.

2.1. Latent Dirichlet allocation (LDA)

According to the LDA model, illustrated in Figure 1(a), each word in a collection of documents is assumed to be generated using a two step process [2]. First a topic assignment, $z_{d,n} \in [1, \ldots, K]$, is sampled from a distribution over topics, $z_{d,n} \sim \mathsf{Multinomial}(\boldsymbol{\theta}_d)$, where $\boldsymbol{\theta}_d$[1] is the document-specific proportion over topics. Next, a word $w_{d,n}$ is sampled

[1] $\boldsymbol{\theta}_d \sim \mathsf{Dirichlet}(\boldsymbol{\alpha})$

Table 1: Description of the topic model notations shown in Figure 1.

K : number of topics
M_k : number of Gaussian components in the k^{th} topic
D : number of documents
N_d : number of words in the d^{th} document
$\boldsymbol{\theta}_d$: proportion of topics in the d^{th} document
$\boldsymbol{\pi}_{d,k}$: proportion of Gaussian components in the k^{th} topic and the d^{th} document
$\boldsymbol{\psi}_k$: Discrete distributions over words in corpus for the LDA model
$\boldsymbol{\psi}_{k,m} \equiv \{\boldsymbol{\mu}_{k.m}, \boldsymbol{\Lambda}_{k,m}\}$: Mean and precision matrices of the m^{th} Gaussian component in the k^{th} topic, for the GMM-LDA model
$z_{d,n}$: topic indicator for the n^{th} word and the d^{th} document
$\xi_{d,n}$: component indicator for the n^{th} word and the d^{th} document
$w_{d,n}$: n^{th} word in d^{th} document
$\boldsymbol{\alpha}, \boldsymbol{\beta}$: Dirichlet distribution hyper-parameters for LDA model
$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$: Dirichlet distribution hyper-parameters for the GMM-LDA model
$\boldsymbol{\beta} \equiv \{\boldsymbol{\mu}_0, \lambda_0, \boldsymbol{W}_0, \nu_0\}$: Gaussian-Wishart distribution hyper-parameters for the GMM-LDA model

from the corresponding topic distribution, $w_{d,n}|z_{d,n} \sim \text{Multinomial}(\boldsymbol{\psi}_{z_{d,n}})$. Each topic distribution $\boldsymbol{\psi}_k$,[2] is a distribution over a finite set of words. The nature of the topic distribution limits the applicability of the LDA model to observations confined to a finite dictionary, e.g. a set of words in a document corpus.

2.2. Latent Dirichlet allocation with Gaussian mixture topics (GMM-LDA)

We modify the LDA model such that the topic distributions are distributions with support over a continuous space. In the new model, illustrated in Figure 1(b), each document is characterized by a mixture of topics, where each topic is distributed as a GMM. The data generation process for the GMM-LDA model can be described as follows. First, a topic assignment is sampled from a distribution of topics, $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$[3]. Since each topic is a GMM, we also need to sample a Gaussian component assignment, $\xi_{d,n} \in [1, \ldots, M_k]$, where M_k is the number of Gaussian components in the k^{th} topic. The document- and topic-specific distributions over Gaussian components are given by $\boldsymbol{\pi}_{d,k}$[4]. Given a topic assignment, $z_{d,n}$, a component assignment can be sampled $\{\xi_{d,n}|z_{d,n} \sim \text{Multinomial}(\boldsymbol{\pi}_{d,z_{d,n}})\}$. Finally, a word can be sampled from the $\xi_{d,n}^{th}$ Gaussian component in the $z_{d,n}^{th}$ topic, $w_{d,n}|z_{d,n}, \xi_{d,n} \sim \text{Gaussian}(\boldsymbol{\psi}_{z_{d,n},\xi_{d,n}})$. Thus each topic distribution $\boldsymbol{\psi}_{k,m}$[5] is a Gaussian distribution over the continuous space.

3. INFERENCE OF MODEL PARAMETERS

In the GMM-LDA model, the goal of inference is to estimate the joint posterior over the latent variables, $\boldsymbol{z}, \boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\psi}$:

[2] $\boldsymbol{\psi}_k \sim \text{Dirichlet}(\boldsymbol{\beta})$
[3] $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha}_1)$
[4] $\boldsymbol{\pi}_{d,k} \sim \text{Dirichlet}(\boldsymbol{\alpha}_2)$
[5] $\boldsymbol{\psi}_{k,m} \sim \text{Gaussian-Wishart}(\boldsymbol{\beta})$

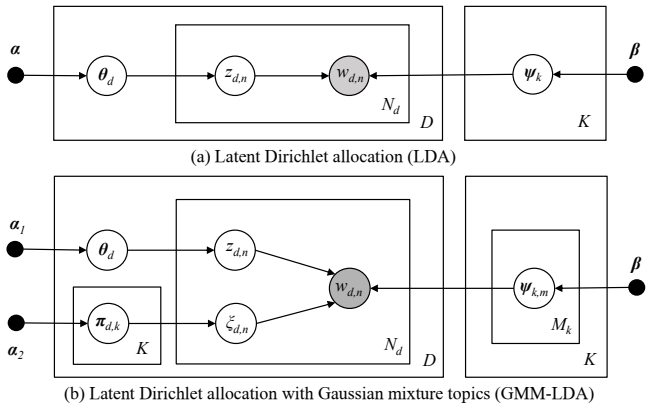

Fig. 1: Probabilistic graphical representation of (a) the LDA and (b) the augmented GMM-LDA. Unshaded nodes represent latent variables, shaded nodes represent observed variables and solid nodes represent hyper-parameters. Plate notation indicates replication of nodes, whereas directed arrows indicate dependencies between variables. A detailed description of the notations is provided in Table 1.

$$P(\boldsymbol{z}, \boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\psi} \mid \boldsymbol{w}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}) = \frac{P(\boldsymbol{z}, \boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{w} | \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta})}{P(\boldsymbol{w} | \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta})} .$$

Evaluating this posterior requires computing the marginal distribution of observed words, $P(\boldsymbol{w}|\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta})$, by integrating over the continuous, $\{\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\psi}\}$ and discrete $\{\boldsymbol{z}, \boldsymbol{\xi}\}$ latent variables. The discrete integral over $\boldsymbol{z}$ and $\boldsymbol{\xi}$ involves $\left[K\right]^{\sum_d N_d} + \left[\prod_k M_k\right]^{\sum_d N_d}$ operations, which makes the direct computation of the posterior intractable. Alternatively, approximate inference of the posterior can be performed using Gibbs sampling [6], which is a Markov chain Monte-Carlo sampling scheme. By choosing conjugate priors, we can analytically integrate out $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, and $\boldsymbol{\psi}$ and perform inference by sampling topic ($z_{d,n}$) and component assignments ($\xi_{d,n}$), based on their joint probability (details provided in Table 2):

$$\begin{aligned} P(z_{d,n} = k, \xi_{d,n} = & m \mid \boldsymbol{w}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}) \propto \\ & P(z_{d,n} = k \mid \boldsymbol{z}_-, \boldsymbol{\alpha}_1) \\ & \times P(\xi_{d,n} = m \mid z_{d,n}, \boldsymbol{z}_-, \boldsymbol{\xi}_-, \boldsymbol{\alpha}_2) \\ & \times P(w_{d,n} \mid z_{d,n}, \xi_{d,n}, \boldsymbol{w}_{k,m,-}, \boldsymbol{\beta}) \quad (1) \end{aligned}$$

4. APPLICATION: CLUSTERING SLEEP EEG DATA

We demonstrate the utility of the GMM-LDA model by applying it to the problem of clustering sleep stages using EEG signals. Monitoring EEG data during sleep allows for the diagnosis of sleep disorders and provides an understanding of sleep physiology [7, 8]. Further, class labels corresponding

Table 2: Equations to compute the joint probability in the GMM-LDA model in (1) using Gibbs sampling.

$$P(z_{d,n} = k \mid \boldsymbol{z}_-, \boldsymbol{\alpha}_1) = \frac{N_{k,-} + \alpha_{1,k}}{\sum_k N_{k,-} + \sum_k \alpha_{1,k} - 1}$$

$$P(\xi_{d,n} = m \mid z_{d,n} = k, \boldsymbol{z}_-, \boldsymbol{\xi}_-, \boldsymbol{\alpha}_2) = \frac{N_{k,m,-} + \alpha_{2,m}}{\sum_m N_{k,m,-} + \sum_m \alpha_{2,m} - 1}$$

$$P(w_{d,n} \mid \boldsymbol{w}_{-,k,m}, \boldsymbol{\beta}) \sim \mathcal{T}_{\nu_N - d + 1}\left\{\boldsymbol{\mu}_N, \frac{\boldsymbol{W}_N(\lambda_N + 1)}{\lambda_N(\nu_N - d + 1)}\right\}$$

$$\boldsymbol{\mu}_N = \frac{\lambda_0 \boldsymbol{\mu}_0 + N_{k,m,-}\overline{\boldsymbol{w}}}{\lambda_0 + N_{k,m,-}}$$

$$\boldsymbol{W}_N = \boldsymbol{W}_0 + \boldsymbol{S} + \frac{\lambda_0 N_{k,m,-}}{\lambda_0 + N_{k,m,-}}(\boldsymbol{\mu}_0 - \overline{\boldsymbol{w}})(\boldsymbol{\mu}_0 - \overline{\boldsymbol{w}})^T$$

$$\nu_N = \nu_0 + N_{k,m,-}$$

$$\lambda_N = \lambda_0 + N_{k,m,-}$$

$$\overline{\boldsymbol{w}} = \frac{1}{N_{k,m,-}} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \boldsymbol{w}_{-,k,m}$$

$$\boldsymbol{S} = \frac{1}{N_{k,m,-}} \sum_{d=1}^{D} \sum_{n=1}^{N_d} (w_{d,n} - \overline{\boldsymbol{w}})(w_{d,n} - \overline{\boldsymbol{w}})^T$$

$\boldsymbol{z}_-$: Set of topic indicators excluding $z_{d,n}$

$\boldsymbol{\xi}_-$: Set of Gaussian component indicators excluding $\xi_{d,n}$

$\boldsymbol{w}_{-,k,m}$: Set of observations allocated to the k^{th} topic and the m^{th} Gaussian component excluding $w_{d,n}$

$\mathcal{T}_\nu\{\boldsymbol{\mu}, \boldsymbol{W}\}$: Multivariate t-distribution ($d = 4$) with parameters $\boldsymbol{\mu}$, $\boldsymbol{W}$ and ν degrees of freedom

$N_{k,m,-}$: Count of observations allocated to the k^{th} topic and the m^{th} Gaussian component excluding $w_{d,n}$

to sleep stages are available, which allows us to compare the unsupervised clustering performance of our GMM-LDA model against an upper-bound on performance using a supervised algorithm.

4.1. Data

In this study, the Sleep-EDF database [9] was used for analysis. This database comprises EEG signals recorded from thirty healthy subjects during sleep. For each subject, 30s time segments of the signals are classified into one-of-eight classes. The six classes corresponding to sleep stages [10, 11] include: N1, N2, N3, N4, Wake (W) & REM (R). The remaining two classes represent movement artifacts and unscored segments, which were not used in this analysis.

4.2. Feature extraction and cross-validation details

Power spectral densities were estimated for each 30s time segment, by averaging over 2s Hamming windowed segments with 50% overlap [12]. Power estimates in Delta ($f < 4$Hz), Theta (4Hz $\leq f <$ 8Hz), Alpha (8Hz $\leq f <$ 12Hz) and Beta (12Hz $\leq f \leq$ 20Hz) bands were used to generate four frequency features, to train the model. Within the context of topic modeling, the 4-dimensional feature vector over a 30s time segment for a given subject is analogous to a word in a document, $w_{d,n}$, for $d = 1, \ldots, D$ and $n = 1, \ldots, N_d$, where D is the number of subjects and N_d is the number of 30s time segments for the d^{th} subject. In order to train the model in a reasonable amount of time with cross-validation, recordings from thirty subjects were split into groups of five. In a given group, data from four subjects were used for training and the fifth subject was used for testing. Hence, the model was trained five times for each group, with each subject used for testing.

4.3. Selection of GMM-LDA model hyper-parameters

To train the GMM-LDA model, we need to select the number of GMMs (K), the number of Gaussian components in the k^{th} GMM (M_k), and the prior hyper-parameters ($\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, $\boldsymbol{\beta}$). We used $K = 6$, which corresponds to the number of sleep stages. We selected M_k by minimizing the Bayesian information criterion (BIC), which is a function of the log-likelihood of the data with a penalty to restrict the number of model parameters [13], and the model with the lowest BIC is typically selected [14]. For each subject group, we fit data corresponding to each sleep stage using a GMM with a varied number of components and computed the BIC for each fit. Based on this analysis, we selected $\boldsymbol{M} = [5, 10, 5, 10, 10, 5]$, for sleep stages N1, N2, N3, N4, W and R, respectively. We chose non-informative priors for other model parameters. For the Gaussian-Wishart priors, we chose zero mean and identity covariance hyper-parameters for $\boldsymbol{\beta}$. For the Dirichlet priors, we chose uniform and symmetric hyper-parameters by setting $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ as vectors of ones.

4.4. Comparison to other models

We compared the unsupervised clustering performance of the GMM-LDA model with five other unsupervised methods: standard K-means, standard GMM, the LDA topic model and the Gaussian-LDA topic model. K-means and GMMs are standard single-membership clustering algorithms. For the LDA model, EEG signals for each subject were represented using the bag-of-patterns model [15, 16]. To obtain the bag-of-patterns representation, features from all subjects were first clustered using K-means, resulting in cluster centroids or codewords. Next, each subject's recording was represented as a vector of occurrence counts of codewords. The Gaussian-LDA is a special case of the GMM-LDA model, where each topic consists of just one Gaussian, i.e. $M_k = 1$, for $k = 1 \ldots K$. In addition to these unsupervised methods, we also tested performance using a supervised method viz. the support vector machine (SVM) classifier, to

compare the unsupervised methods against an upper bound on performance.

4.5. Performance evaluation

For performance evaluation, topic assignments for each subject were compared to the truth labels using Fowlkes-Mallows score (FMS) [17]. The FMS computes the geometric mean of precision and recall and is formally defined as follows:

$$\text{FMS}(\boldsymbol{z}_{d,-}, \boldsymbol{l}_{d,-}) = \frac{TP}{\sqrt{(TP+FP)\times(TP+FN)}}$$

where, $\boldsymbol{z}_{d,-}$ are topic allocations for all time segments for the d^{th} subject and $\boldsymbol{l}_{d,-}$ represent true class labels. True positives (TP) are the number of sample pairs belonging to the same clusters in both $\boldsymbol{z}_{d,-}$ and $\boldsymbol{l}_{d,-}$, false positives (FP) are the number of sample pairs belonging to the same clusters in $\boldsymbol{l}_{d,-}$, but not in $\boldsymbol{z}_{d,-}$ and false negatives (FN) are the number of sample pairs belonging to the same clusters in $\boldsymbol{z}_{d,-}$, but not in $\boldsymbol{l}_{d,-}$. FMS values range between 0 and 1, with higher values indicating greater similarity between topic allocations and class labels.

5. RESULTS AND DISCUSSION

Performance results of all algorithms are summarized in Figure 2(a). Repeated measures analysis of variance (ANOVA) was used to test statistical significance between the performances of various algorithms. Results of multiple comparison tests are shown in Figure 2(b). We observed that the GMM-LDA model provided the best performance amongst the clustering algorithms. This suggests that the GMM-LDA model provided the best fit to the sleep EEG data, by representing sleep stages as multi-modal GMMs over a continuous space. Clustering using K-means and GMMs assumes that each sleep stage is represented by a single unimodal Gaussian, and does not allow for different proportion of sleep stages per subject. The LDA allows for different proportion of sleep stages to be represented using a topic model; however, the data structure is lost when discretizing the continuous data space using the bag-of-patterns representation. While the Gaussian-LDA allows topics over a continuous domain, it limits them to unimodal distributions.

6. CONCLUSIONS AND FUTURE WORK

We have presented an extension to the LDA topic model that can handle data over a continuous domain. In this model, topics are represented as a mixture of Gaussians spanning a continuous domain instead of discrete distributions spanning a finite dictionary. We demonstrated the utility of the GMM-LDA model by clustering sleep EEG data and obtained the best performance amongst the unsupervised clustering algorithms, and most comparable to that of a supervised classification algorithm.

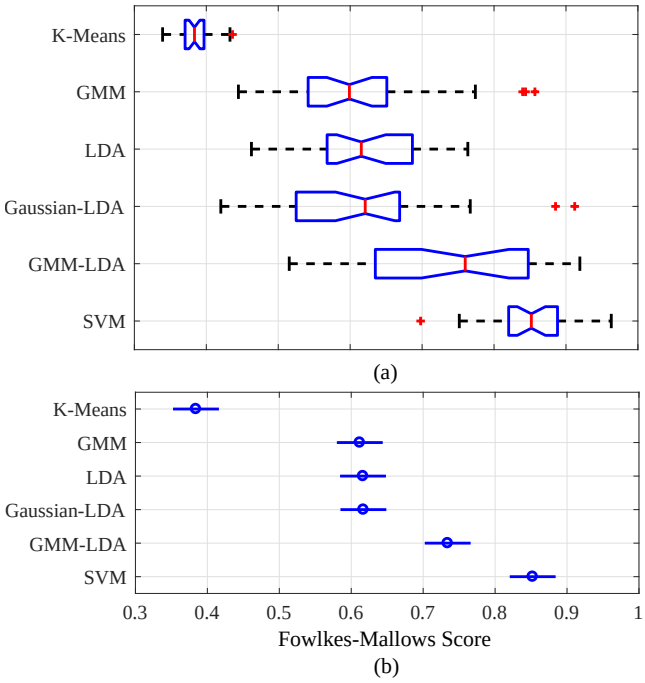

Fig. 2: (a) Box plots of Fowlkes-Mallows score (FMS) summarizing algorithm performance across thirty subjects. In each box, the red line indicates the median (q_2), and the lower and upper limits indicate $25^{th}(q_1)$ and $75^{th}(q_3)$ quantiles, respectively. The notch extremes correspond to the range $q_2 \pm \frac{1.58(q_3-q_1)}{\sqrt{N}}$; the lower and upper fences indicate extreme values not considered outliers, and outliers are indicated using '+'. (b) Plot of mean and confidence intervals of the FMS estimated using repeated measures ANOVA with Bonferroni correction ($\alpha = 0.05$). Two algorithms have significantly different means if their confidence intervals do not overlap.

Further improvements are needed for our GMM-LDA model. In this work, we incorporated knowledge about the sleep stage class labels to determine the number of topics, K, and the number of Gaussian components within each topic, M_k. However, there are certain cases where class labels may not be available or are not well defined. Hence, there is a need to automate the process of model selection, which could be achieved by using Bayesian non-parametric priors [18, 19], like Dirichlet processes, on topic proportions ($\boldsymbol{\theta}$), and GMM component proportions ($\boldsymbol{\pi}$). Also, the current model assumes that observations are exchangeable, i.e. the specific ordering of $w_{d,n}$ is neglected [1, 2]. This may be a limiting assumption for time series data with a specific sequence of events, such as sleep stages. An extension to the current model can incorporate Markovian dependence between observations. We will investigate these extensions to the GMM-LDA model in future work.

7. REFERENCES

[1] David Blei, Andrew Ng, and Michael Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, no. 1, pp. 993–1022, 2003.

[2] David Blei, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55–65, 2012.

[3] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly, "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, vol. 155, pp. 945–959, 2000.

[4] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang, "Topic and role discovery in social networks," *Computer Science Department Faculty Publication Series*, p. 3, 2005.

[5] Lawrence Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] Stuart Geman and Donald Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, , no. 6, pp. 721–741, 1984.

[7] Alexander Van Esbroeck and M. Brandon Westover, "Data-driven modeling of sleep states from EEG.," *International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2012, pp. 5090–3, 2012.

[8] J C Gillin, W Duncan, K D Pettigrew, B L Frankel, and F Snyder, "Successful separation of depressed, normal, and insomniac subjects by EEG sleep data.," *Archives of general psychiatry*, vol. 36, no. 1, pp. 85–90, 1979.

[9] AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, PCh Ivanov, RG Mark, JE Mietus, GB Moody, C-K Peng, and HE Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. 215–220, 2000.

[10] Michael H. Silber, Sonia Ancoli-Israel, Michael H. Bonnet, Sudhansu Chokroverty, Madeleine M. Grigg-Damberger, Max Hirshkowitz, Sheldon Kapen, Sharon a. Keenan, Meir H. Kryger, Thomas Penzel, Mark R. Pressman, and Conrad Iber, "The visual scoring of sleep in adults," *Journal of Clinical Sleep Medicine*, vol. 3, no. 2, pp. 121–131, 2007.

[11] A Rechtschaffen and A Kales, "A manual of standardized terminology, techniques and scoring for sleep stages of human subjects (nih publ. no. 204)," *US Washington: Government Printing Office*, 1968.

[12] Peter D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on TIme Averaging Over Short, Modified Periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.

[13] Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[14] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[15] Li Fei-Fei and P Perona, "A Bayesian hierarchical model for learning natural scene categories," *Computer Vision and Pattern Recognition, 2005.*, vol. 2, pp. 524–531, 2005.

[16] Jin Wang, Xiangping Sun, Mary F H She, Abbas Kouzani, and Saeid Nahavandi, "Unsupervised mining of long time series based on latent topic model," *Neurocomputing*, vol. 103, pp. 93–103, 2013.

[17] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[18] Peter Orbanz and Yee Whye Teh, "Bayesian Nonparametric Models," *Encyclopedia of Machine Learning*, vol. 25, no. 1, pp. 1–14, 2005.

[19] M. D. Escobar, "Nonparametric Bayesian methods in hierarchical models," *Journal of statistical planning and inference*, vol. 43, pp. 97–106, 1995.