MATH 574M
Statistical Machine Learning
Final Project Report

Marium Yousuf

University of Arizona

May 19, 2020

Sparse inverse covariance estimation with G-Wishart distribution to classify cognitive impairment

Abstract: Alzheimer's Disease is a neurodegenerative disorder that weakens brain connectivity causing poor inter-regional communication. Functional Connectivity refers to the ongoing signal exchange in the brain that is drastically affected due to Alzheimer's. This change can be reflected using Gaussian precision matrix as it allows extraction of sparsity that represents the state of communication between brain regions. A Bayesian model that learns sparsity can be used to identify the cognitive state and further help identify the chances and predict any brain impairments.

I. Introduction

Neurons are involved in a signal exchange between brain regions creating a network-like structure within and between those regions. There are two kinds of configurations of such brain structure: anatomical and physiological [1]. Both of these configurations, also known as connections, allow the understanding of the brain as a network, which can be studied using the neuroimaging techniques [1], such as positron emission tomography (PET) scans and functional magnetic resonance imaging or functional MRI (fMRI).

In this report, we focus on the physiological connectivity, or functional connectivity, of the brain that describes a constant neuron communication between the regions. That is when one part of the brain receives signals for a task it shows activity and further sends signals to another regions to complete the task received. These mechanical observations provide insight into the patterns of statistical dependence among brain regions [1], which can further highlight the cognitive state of any given subject.

Subjects with weaker cognitive state are expected to have a dysfunctional communication, that is some signals sent from one region to another could be slower because of weak adaptivity or reception of the signals. Viewing the brain as a network, this dysfunctional structure can be understood as sparse in comparison to normally functioning brain. Exploiting the insight that functional connectivity provides, here is proposed a Bayesian inference model to learn sparsity structure in the brain network. To narrow down the focus of the problem, we only focus on the of age-related cognitive impairment, namely Alzheimer's disease. This model allows identification of patients with Alzheimer's disease (AD), Mild Cognitive Impairment (MCI), and Normal Control (NC) subjects.

Alzheimer's Disease (AD) is the most common form of dementia, which is a neurodegenerating illness that results in severe lifestyle changes. Development of AD gradually affects the cerebral cortex, which is a functional part of the brain operating on the sense perception, muscle movements, and assessment skills like problem-solving and critical thinking [2]. Other than this, the hippocampus - the brain region most affected by the AD - is disrupted affecting memory [3, 4].

AD is a progressive disease as it gets worse with time and is often only diagnosed when the patient is already showing extreme symptoms. Due to this, intensive research is dedicated to investigating ways to delay, cure, or prevent the progression of AD after, sometimes early,

diagnosis [5]. In the U.S. alone, the number of patients with AD is estimated at 5.5 million [6]. Typical life expectancy after an Alzheimer's diagnosis being four to eight years [6], on diagnosis, the annual cost incurred to patient care is estimated at over $100 billion per year [5].

There are two types of brain connectivity: structural connectivity and functional connectivity [1]. Structural connectivity concerns the white matter that forms due to long term neuron communication between brain regions, while functional connectivity refers to the active communication that takes place between brain regions through neurons [3]. Both types of brain connectivity eventually change [7].

Neuroimaging techniques reveal that AD is closely related to alteration in the functional connection of the brain and compared to NC, patients with AD show a decrease in the amount of functional connectivity, especially around the hippocampus area [4]. In contrast, there is an increase in the amount of connectivity within the frontal lobe that causes high sensitivity to sense of touch and a sense of visual information [4], revealing the strong connection within the individual regions but weaker communication transfer to another brain regions.

With a goal to learn sparsity and understand the strength of information exchange between brain regions, we consider p number of brain regions as variables and observe how they relate to each other. To build upon the choice of Bayesian analysis, we also discuss graphical LASSO, which is another widely covered tool in related literature to extract sparsity of a graphical structure as well as compare them to the Bayesian technique. Considering the complexity of our problem in terms high-dimensions of brain regions, we utilize probabilistic graphical modeling.

II. Background

(a) Undirected Gaussian Graphical Models and their Markov Property

A graph, $G = (V, E)$, consists of a set of vertices $V = v_1, \ldots, v_n$ and set of edges $E = e_1, \ldots, e_n$, such that each e_i connects any vertices $v_i, v_j \in V$, $i, j = 1, ..., n$. There are either directed graphs, also known as Bayesian Networks, or undirected graphs, also known as Markov Random Fields (MRFs). Vertices of the graph represent random variables, while edges represent the interaction between those random variables.

In this report, with an assumption that our random variable observations have a multivariate Gaussian distribution with mean μ and covariance matrix Σ, we consider an MRF with vertices representing the brain regions and edges representing the connection between them. For a NC subject, we assume a complete graph, which is a fully connected graph such that each pair of vertices is connected by an edge [8]. For MCI we introduce sparsity in a complete graph, and finally for for AD, we increase sparsity in the structure developed for MCI. Note that we assume a known structure for all three types of subjects.

Moreover, for our approach, we are concerned with two kinds of graphs: *decomposable* vs. *non-decomposable*. A graph is called decomposable if all the cycles ≥ 4 have a chord/edge between two vertices such that the edge is not a part of that cycle. Any graph that does not meet the said property of decomposable graph is called a non-decomposable graph.

Another important concept for this paper is of *prime components*. Prime components is related to the concept of graph decomposition (irrelevant to the decomposable/non-decomposable graph discussed above). Lauritzen [9] defines graph decomposition as a partitioning of V into a three subsets: A, B, S such that $AB|S$, and S is complete, called a separator set. A graph is called a prime component if it cannot be properly decomposed.

(b) LASSO and Graphical LASSO

LASSO is one the of shrinkage approaches used to trade-off bias in order to reduce large variance in least squares estimates to improve prediction accuracy and model interpretation [8]. LASSO estimate imposes a a penalty on the regression coefficients and is defined as:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j|,$$

where λ is the parameter used to control the amount of shrinkage - the larger λ, the greater the amount of shrinkage. Here $\lambda \geqslant 0$ and is also known as a complexity parameter. Now extending this concept of LASSO to graphs, we now discuss graphical LASSO (gLASSO). Similar to LASSO, gLASSO shrinks the regression coefficients in size and instead of the least squares estimates it penalizes the log-likelihood of the data defined as [8]:

$$l(\Theta) = \log\det(\Theta) - \mathrm{tr}(\Sigma \cdot \Theta),$$

with the L_1 LASSO penalty we have:

$$l(\Theta) = \log\det(\Theta) - \mathrm{tr}(\Sigma \cdot \Theta) - \lambda||\Theta||_1,$$

where $||\Theta||_1$ is the sum of of absolute values of the elements of the precision matrix.

(c) G-Wishart Distribution and Precision Matrices

MRF with missing edges assumes conditional independence between the vertices such that if an edge is absent between two random variables x_1 and x_2, then x_1 and x_2 are conditionally independent given all other variables [8] such that $x_1 \perp\!\!\!\perp x_2 | x_{i/\{1,2\}}$. This conditional independence is captured by the inverse covariance matrix, also known as precision matrix, $\Theta = \Sigma^{-1}$.

The role of Θ becomes explicit as we examine the conditional distribution of one variable against others [8]. Exploiting this property and using the fact that conditional distribution of Gaussian distribution is also Gaussian, we use G-Wishart distribution, which is a distribution over positive definite matrices and it represents the conditional independencies of a Gaussian graphical model. That is, given a graph structure G, G-Wishart distribution is a conjugate prior of the precision matrix of the multivariate normal random variables represented by the vertices of G.

Hence, sampling from G-Wishart using a known structure yields a positive definite matrix from a parameter set, $M^+(G)$, known as the cone of positive definite matrices [10, 11], that represents the conditional independencies of a Gaussian graphical model. Note that the complete graphs have no conditional independence interpretation [9]. Throughout this

report, we would use the notation $\mathbf{\Sigma}$ for covariance matrix and $\mathbf{\Theta} = \mathbf{\Sigma^{-1}}$ for precision matrix.

Another significant concept used in the sampling of G-Wishart related to precision matrix is of Cholesky decomposition, which is a decomposition of a positive definite matrix into the product of a lower (or upper) triangular matrix and its transpose [12]. For instance, given a matrix $D \in M^+(G)$, then D can be written as a Cholesky decomposition $D = T \cdot T^T$, where T is a lower (or upper) triangular matrix.

(d) Rejection Sampling

Rejection sampling simply involves approximating a target distribution, $p(x)$, for which direct sampling is intractable, using a proposal distribution, $q(x)$. There are two main properties of rejection sampling:

i. $q(x)$ is scaled by some factor, say α, such that $\alpha q(x) > p(x)$, and

ii. $\exists\, u$ that is uniformly sampled between $[0, 1]$.

If $u \leq \frac{p(x)}{\alpha q(x)}$, then the sample is accepted, otherwise rejected.

(e) Metropolis-Hastings Algorithm

Metropolis-Hastings (MH) is a Markov Chain Monte Carlo algorithm [13] that generates samples to approximate a distribution $p(x)$ which cannot be directly sampled. MH requires sampling from a proposal distribution $q(x^*|x)$ dependant on the current value of x such that the Markov chain moves towards x^* if and only if q satisfies the acceptance threshold otherwise it remains at x [13]. The following is a pseudocode adapted from [13]:

i. Initialize $x^{(0)}$

ii. For $i = 0$ to $N - 1$

- Sample $u \sim U_{[0,1]}$.
- Sample $x^* \sim q(x^*|x^{(i)})$.
- If $u < A(x^{(i)}, x^*) = \min\left\{1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^i)}\right\}$

$x^{(i+1)} = x^*$

else

$x^{(i+1)} = x$

III. Related Work

Huang et al. [4] introduce a model called the sparse inverse covariance model (SICE) to induce functional connectivity from PET scans. They use gLASSO to impose sparsity in a Gaussian graphical model. consistency of an inverse covariance matrix. Taking p number of brain regions, they assume their data to follow a Gaussian distribution and use SICE to find an estimate for the precision matrix of the p brain regions using the l_1 regularization with the following optimization.

$$\hat{\Theta} = \arg\max_{\Theta > 0} \log(\det(\Theta)) - \mathrm{tr}(\Sigma \cdot \Theta) - \lambda \|\Theta\|_1,$$

where Θ is the precision matrix, $\hat{\Theta}$ is the estimate of the precision matrix, Σ is the covariance matrix, and $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ are determinant and trace of a matrix, respectively. Moreover, $\|\cdot\|_1$ is the sum of all the absolute values of matrix entries, and lastly, λ is the regularization parameter. They apply SICE to identify AD, MCI, and NC subjects and claim that SICE effectively identifies the structure of a precision matrix such that it extracts the existence and non-existence of functional connections between brain regions. However, they disrecommend using SICE to estimate the magnitude of non-zero entries. That is, the SICE model can extract the sparsity correctly but not the extent of it. Huang et al. [4] technique utilizes a penalized likelihood and uses a classical approach in sense that their Θ is unknown but fixed. Over time, since the connection is understood to deteriorate, a classical approach did not seem ideal.

Hinne et al. [7] propose a Bayesian framework for the estimation of functional connectivity in the presence of uncertainty. They also use the precision matrices to understand the correlations between brain regions and depict the ongoing communication between them. They propose a Bayesian model using a Wishart prior for shrinkage that estimates a posterior density over precision matrices to analyse functional connectivity. Their approach is similar to ours except they do not consider cognitive impairments. They use diffusion imaging data and test their model using simulated data as well as resting-state fMRI data, eventually comparing their approach to the graphical lasso. They examined and depicted that the gLASSO approach is explicit in terms of describing the conditional independence. While their Bayesian approach yields a more nuanced correlation. Using inspiration from this appraoch, we also use G-Wishart prior to extract the sparseness of estimated precision matrices determined by a graph, G that corresponds to the brain's connectivity of the subject. Bayesian approach requires computation of a posterior density over sparse precision matrices [7]. This posterior is then used to compute marginal densities for precision matrices. However, computation of this posterior poses a challenge of estimating a normalising constant [10]. That is, the posterior distribution of sparse precision matrices is not defined in closed for and hence needs to be approximated.

Using the appraoch defined in [10], we resort to using a Monte Carlo method for computing the marginal likelihood in sparse Gaussian graphical models with rejection sampling and Metropolis-Hastings algorithm for efficient and independent sampling (more in next section). There are, however, hybrid approaches that use Bayesian framework with LASSO for covariance estimation [14, 15]. They use a LASSO prior with block Gibbs sampler to simulate covariance matrices. [14] also generalize the Bayesian graphical lasso to the Bayesian adaptive graphical lasso and claim that the Bayesian adaptive graphical LASSO performs better than other classical and Bayesian approaches. The only major differences between [14] and [15] is that [15] uses Metropolis-Hastings instead of block Gibbs sampler and that they don't explore the properties of graphical LASSO prior like [14].

There is plenty of other literature available that discusses the identification of AD from functional connectivity. Some papers mentioned here provide models on inferring functional connectivity, with a couple particularly focused on AD. Similar model to SICE is provided in Sun et al. [5] but with a focus on anatomical changes (structural connectivity) due to long-term impaired functional connectivity. They look at the connectivity patterns that provide image-based ways to distinguish between NC, MCI, and AD subjects. They propose an algorithm

based on *block coordinate descent* approach to estimate inverse covariance matrix that has a *user feedback* feature instilled in the estimation process. They apply this algorithm, and provide supportive experimental results, to PET images of 232 NC, MCI, and AD subjects that helps discover the connectivity patterns and differences between the three categories.

Deligianni et al. [16] also focuses on structural connectivity and observes it across subjects to predict functional connectivity. They present a probabilistic framework using cross-validation to learn the covariance structure of the brain. They consider the correlation of the parameters and introduce a model, based on conditional independence structure of structural connectivity, with a loss function independent to the parameters so to maintain the property of positive definite matrices that match the functional connectivity. Their results emphasize that using statistical learning, functional connectivity can be extracted from the anatomy of the brain.

Apart from the neuro-scientific aspect, sparsity matrix is also used to investigate other biological datasets. For instance, Knowles et al. [17] takes a Bayesian approach treating sparsity connectivity matrix as a random variable to model gene expression data of increasing complexity. In their paper, they show that inducing sparsity from datasets for E. Coli and breast cancer data improves predictive performance together with providing an easy interpretation.

IV. Model Approach and Experiment

Consider the mixture model from Figure 1. N represents the number of subjects and K represents the three categories: AD, MCI, and NC. For each subject, Θ_i is learned such that it represents the sparsity in the precision matrix. Z_i is the latent parameter that represents the functional connectivity state and defines if the subject has AD or not. The state of each subject is observed, and represented as X_i, which is dependent on Z_i, M_k, and Θ_k. The goal is to compute the posterior $p(Z_i|X_i)$ that reveals the cognitive state given the information associated with each subject.

For inference we plan to use a Gaussian mixture model with the following assumptions:

$$\begin{gathered} Z_i \sim \text{CAT}(\pi) \\ M_k \sim N(\mu, \sigma/\sqrt{p}) \\ X_i|Z_i, M_k, \Theta_k \sim N(M_k, \Theta_z^{-1}) \end{gathered}$$

Atay-Kayis et al [10] provides a G-Wishart sampler for sparse precision matrices, or non-decomposable graphs using the G-Wishart parameters δ and D refering to the degrees of freedom and inverse scale parameter, respectively. The degrees of freedom are defined as $\delta + |p| - 1$, where $\delta \geqslant 2$ and the inverse scale paramater is defined using a Cholesky decomposition such that $D = T^T T$, where T is an upper-triangular matrix. To compute the inverse scale parameter, we use an adjacency matrix of a known graph structure, forcing the diagonal elements to be non-zero and further inducing a diagonal dominance to ensure an invertible adjacency matrix, we inverted the adjacency matrix. Notice that Cholesky decomposition implies that D is a positive definite matrix.

Other than this, the sampling distribution is also based on the Cholesky decomposition of Θ such that $\Theta = \phi^T \phi$, such that $\exists \quad \psi = \phi T^{-1}$, where the diagonal elements of ψ follow a chi-squared distribution, while the non-diagonal elements that are non-zero, that is, the vertices

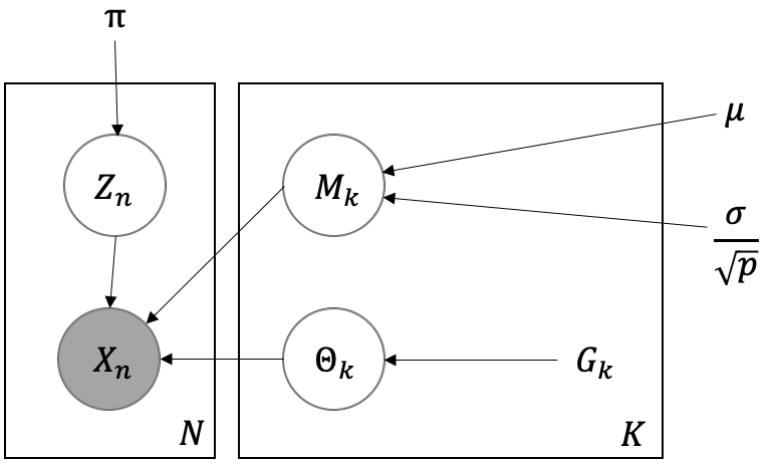

Figure 1: Mixture model representing the Bayesian network where $K = 3$ representing categories: AD, MCI, and NC and N number of subjects. G_k is G-Wishart prior on Gaussian precision matrices, Θ_k, M_k is the Gaussian mean with hyperparameters μ and $\sigma^2/\sqrt{p}$, where p is the number of regions. Z_n defines the cognitive state for each subject, while X_n is the observed data defining the sparsity for each subject.

that have an edge between (free elements) follow a independent univariate standard normal distribution. The non-free elements, defined as set of edges that are absent from G if G were complete and not sparse.

The sampler in [10] is introduced for the full-graph, which in high-dimension can be inefficient for inference. Carvalho et al. [18] introduces a junction tree decomposition method to compute the sampler in [10] at a prime component level. However, [19] refutes this method for assuming an independence between the free and mention that they are implicitly dependant on

$$\exp\left\{-\frac{1}{2}\sum_{(i,j)\in\bar{V}}\psi_{ij}^2\right\}.$$

As a solution, [19] suggest a rejection sampling method to induce independence and accept ψ samples only if they satisfy the following threshold:

$$u < \exp\left\{-\frac{1}{2}\sum_{(i,j)\in\bar{V}}\psi_{ij}^2\right\},$$

where u is uniformly sampled.

The algorithm below shows steps to sample ψ and is adapted from the Atay-Kayis et al. [10], Wang et al. [19], and Roverato et al. [12].

1. Given a graph G, for each prime component of G, G_{p_i}, compute a matrix, $D \in M^+(G)$, with an upper triangular Cholesky factor, T such that $D = T \cdot T^T$, where $T = (t_{ij})_{1\leqslant i\leqslant j\leqslant p}$.

2. For each $D = T \cdot T^T$, compute a relevant $t_{<ij]}$ such that $t_{<ij]} = \frac{t_{ij}}{t_{jj}}$.

3. Compute an upper triangular matrix, A, of size $p \times p$ from adjacency matrix of G_{p_i} plus an identity matrix of size $p \times p$, where any zero element, $a_{ij} = 0$, is off-diagonal and represents missing edges in G.

4. Compute $\bar{V} = W_{\sim V}$, where W is the arbitrary set of vertices if G_{p_i} were a complete graph.

Now, for a sample size, N:

5. For $(i, j) \in V$ and $i = 1, ..., p$, sample $\psi_{ii} = \sqrt{U_i}$ where $U_i \sim \chi^2$, and for $i = 1, ..., (p-1)$ and $j = (i+1), ..., p$, if $a_{ij} = 1$, sample $\psi_{ij} \sim V_{ij}$, where $V_{ij} \sim N(0, 1)$.

6. For $(i, j) \in \bar{V}$ and $i = 1, ..., (p-1)$ and $j = (i+1), ..., p$, if $a_{ij} = 1$, compute ψ_{ij}:

If $i = 1$ and $a_{ij} = 0$:

$$\psi_{ij} = -\sum_{k=i}^{j-1} \psi_{ik} t_{<kj]}$$

Otherwise, if $i > 1$ and $a_{ij} = 0$:

$$\psi_{ij} = -\sum_{k=i}^{j-1} \psi_{ik} t_{<kj]} - \sum_{r=1}^{i-1} \left(\frac{\psi_{ri} + \sum_{l=r}^{i-1} \psi_{rl} t_{<li]}}{\psi_{ii}} \right) \left(\psi_{rj} + \sum_{l=r}^{j-1} \psi_{rl} t_{<lj]} \right)$$

Built on top of the concept of graph decomposition, we initially computed a *junction tree* for efficient sampling, which is a tree representation of the prime components [18]. Each vertex of a junction tree represents a prime component and each edge represents a separator set, S, such that $v_i \in S$ are the common vertices from the two prime components that S connects. However, since the number of nodes in our chosen graphs is small, we moved our current focus to sampling the full graph. However, while the rejection sampling performed well for smaller components, the acceptance rate dropped quickly using the full graph. Therefore, for accepting psi samples, with using rejection sampling we also explored Metropolis-Hastings algorithm introduced in [20], which resulted in a worthwhile improvement in the acceptance rate. Therefore, the next step of the sampler follows:

7. The condition for rejection sampler is as follows:

$$u < \exp\left\{ -\frac{1}{2} \sum_{(i,j) \in \bar{V}} \psi_{ij}^2 \right\},$$

while the acceptance probability for Metropolis-Hastings algorithm from [20] is as follows:

$$\min\left\{ \frac{\omega(\psi_{prop})}{\omega(\psi_{curr})}, 1 \right\},$$

where $prop$ refers to proposed state, x^*, and $curr$ refers to the current state, x, while

$$\omega(\psi) = \exp\left\{ -\frac{1}{2} \sum_{(i,j)\in\bar{V}} \psi_{ij}^2 \right\}.$$

8. The final step is to compute Θ from the sampled ψ by first computing the matrix $\phi = \psi T$ and then finally, $\Theta = \phi^T \phi$.

Using this Θ, then, we simulated observations and then for the posterior distribution of Θ, following the section 4.3 of [10], we update the degree of freedom to $delta + n$, where n is the number of observations, x_i and inverse scale parameter to $D + U$, where U is a statistic computed from the observations such as:

$$U = \sum_{i=1}^{n} x_i^T x_i,$$

V. Empirical Results

Consider the graph in Figure 2 and Figure 3. Figure 2 shows the graph used for experimentation, while Figure 3 shows the acceptance rate against the index going from 100 to 2000 for sampling from the posterior distribution of Θ.

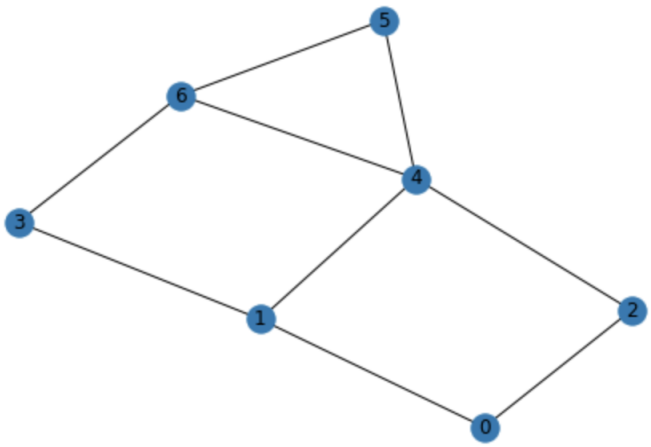

Figure 2: A toy graph used for experimentation. Notice that the vertices $\{0, 1, 2, 4\}$, $\{1, 3, 4, 6\}$, and $\{4, 5, 6\}$ form the prime components.

Applying the algorithm defined in the section above, we see the following acceptance rates:

- For rejection sampling: 314/500
- For MH algorithm: 381/500

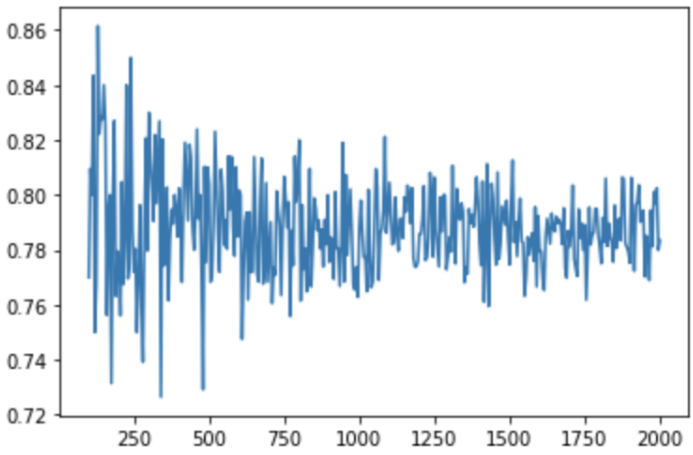

Figure 3: Results from Experiment with graph above, sampling from the prior distribution of Θ

Next steps include using three graphs with different levels of sparsity and use Gaussian Mixture modeling for inference. That is, we need to sample Z_i from categorical distribution and eventually compute the posterior $p(Z_i|X_i)$ and verify how the sparsity of each graph classifies.

References

[1] Olaf Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247, 2013.

[2] Medical Art Library. Cerebral cortex - functional areas., 2017. URL `https://medicalartlibrary.com/cerebral-cortex/`.

[3] Xavier Delbeuck. Alzheimer' disease as a disconnection syndrome? *Neuropsychology review*, 13(2):79–92, 2003.

[4] Shuai Huang. Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation. *NeuroImage*, 50(3):935–949, 2010.

[5] Liang Sun. Mining brain region connectivity for alzheimer's disease study via sparse inverse covariance estimation. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM:1335–1344, 2009.

[6] Alzheimers News Today Erum Naqvi. Alzheimer's disease statistics, 2019. URL `https://alzheimersnewstoday.com/alzheimers-disease-statistics/`.

[7] Max Hinne. Structurally-informed bayesian functional connectivity analysis. *NeuroImage*, 86: 294–305, 2014.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2017.

[9] Steffen L. Lauritzen. *Graphical models.* Clarendon Press, 1996.

[10] Aliye Atay-Kayis. A monte carlo method for computing the marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

[11] Hang Wang and Sophia Zhengzi Li. Efficient gaussian graphical model determination under g-wishart prior distributions. *Electronic Journal of Statistics*, 6:168–198, 2012.

[12] Alberto Roverato. Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

[13] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[14] Hao Wang. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886, 2012.

[15] Zakaria S Khondker, Hongtu Zhu, Haitao Chu, Weili Lin, and Joseph G. Ibrahim. The bayesian covariance lasso. *Stat Interface.*, 6(2):243–259, 2013.

[16] Fani Deligianni, Gael Varoquaux, Bertrand Thirion, Emma Robinson, David J. Sharp, A. David Edwards, and Daniel Rueckert. A probabilistic framework to infer brain functional connectivity from anatomical connections. *Biennial International Conference on Information Processing in Medical Imaging.*, Springer(Berlin, Heidelberg), 2011.

[17] David Knowles and Zoubin Ghahramani. Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

[18] Carlos M. Carvalho. Simulation of hyper-inverse wishart distributions in graphical models. *Biometrika*, 94(3):647–659, 2007.

[19] Hao Wang and Carvalho M. Carlos. Simulation of hyper-inverse wishart distributions for non-decomposable graphs. *Electronic Journal of Statistics*, 4:1470–1475, 2010.

[20] Hélene Massam Mitsakakis, Nicholas and Michael D. Escobar. A metropolis-hastings based method for sampling from the g-wishart distribution in gaussian graphical models. *Electronic Journal of Statistics*, 5:18–30, 2011.