Ph.D. Defense

Mixed Membership Models with Applications to Neuroimaging

Nicholas Marco

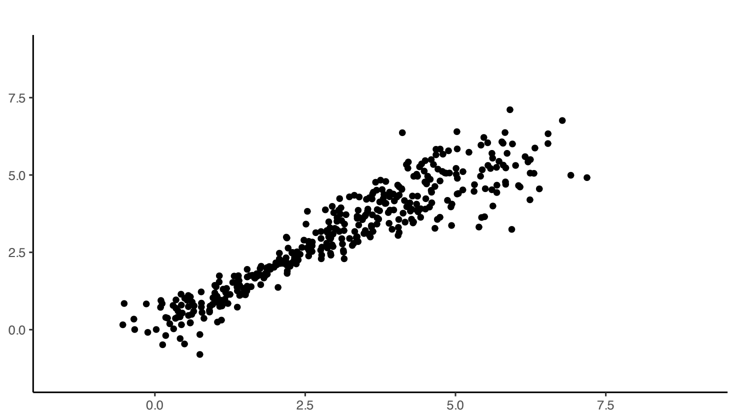
Advisor: Donatello Telesca
Additional Committee Members: Michele Guindani, Damla Şentürk, Joanne Weidhaas

Friday 26th May, 2023

Table of Contents

Overview of Clustering

- Clustering analysis is an exploratory task that aims to assign observations into homogeneous subgroups so that we can better understand the data (Hennig et al., 2015)

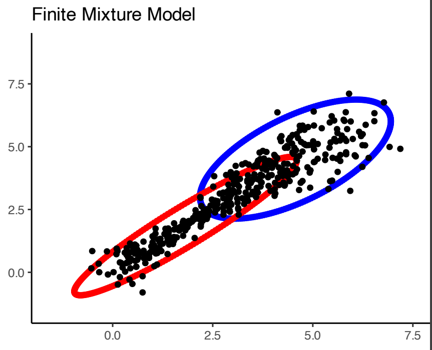


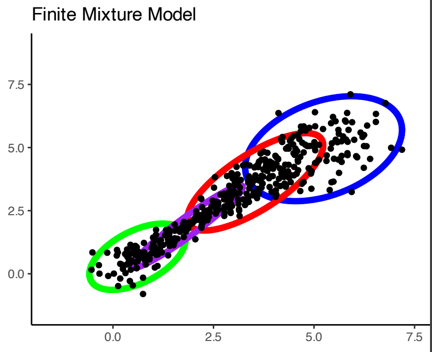
Overview of Clustering

- Clustering membership can generally be divided into two main categories:
  1. *Soft/Fuzzy clustering*: Each observation belong **partially** to each subgroup, akin to **Mixed Membership**
     - **Mixed Membership Models**, Fuzzy C-Means
  2. *Hard clustering*: Each observation comes from a **single** (but unknown) subgroup, akin to **Uncertain Membership**
     - Finite Mixture Models, K-Means

- Clustering models can generally be divided into two main categories:
  1. *Probabilistic/Model-Based clustering*: Construction of a fully probabilistic model of the data, with the clustering labels often thought of as latent variables
     - **Mixed Membership Models**, Finite Mixture Models
  2. *Cost-Based clustering*: Achieve clustering by minimizing a cost function to get the optimal clustering labels
     - Fuzzy C-Means, K-Means

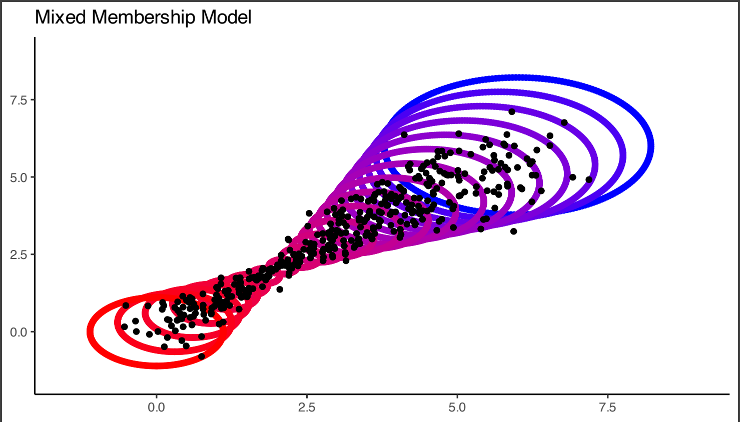
Overview of Finite Mixture Models





- Finite mixture models are probabilistic clustering models that assume each observation comes from one of the K clusters
  - The choice of the number of clusters (K) is user-specified

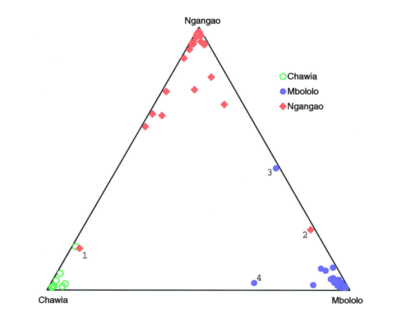
Overview of Mixed Membership Models



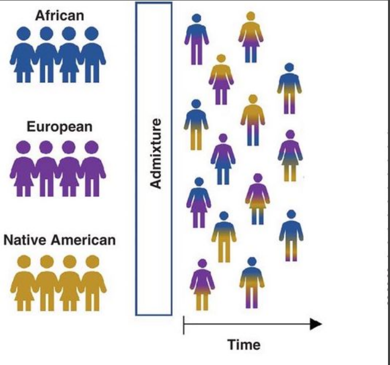
- Mixed membership models are a generalization of finite mixture models, where membership is considered to be on a spectrum

Mixed Membership Models in Genetics

- Mixed Membership Models often are referred to as *admixture models* in the genetics literature (Pritchard et al., 2000; Tang et al., 2005; Alexander et al., 2009)



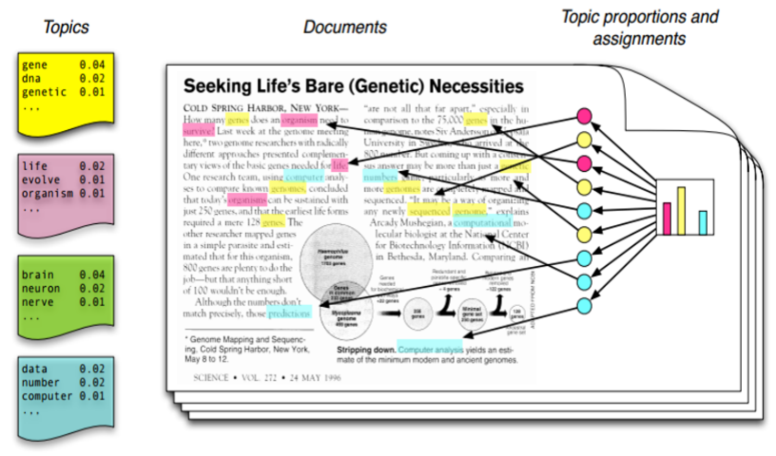
(Pritchard et al., 2000)



(Horimoto et al., 2022)

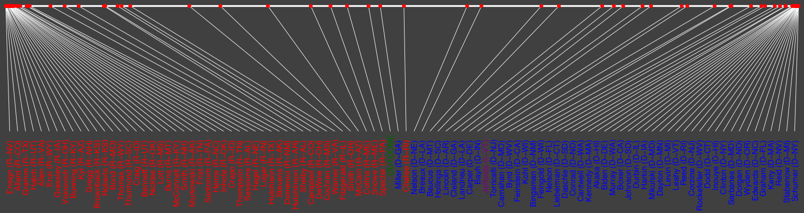
Latent Dirichlet Allocation

- Topic Models, such as Latent Dirichlet Allocation (Blei et al., 2003), aims to explain a collection of objects (referred to as *documents*) through a set of unobserved subgroups (referred to as *topics*)



Other Mixed Membership Models

- Erosheva et al. (2004) used a mixed membership model to classify scientific publications
- Heller et al. (2008) introduced a fully probabilistic mixed membership framework for data that is assumed to have come from the exponential family of distributions
  - Applied their framework to classifying senators based off of roll call data (binary voting records)



Example: Bivariate Normal (K=3 Features)

- The partial membership model framework proposed by Heller et al. (2008) leads to unwieldy implied sampling models, even in cases when we have more than 2 features in the mixed membership model

(Heller et al., 2006) (Marco et al., 2022)

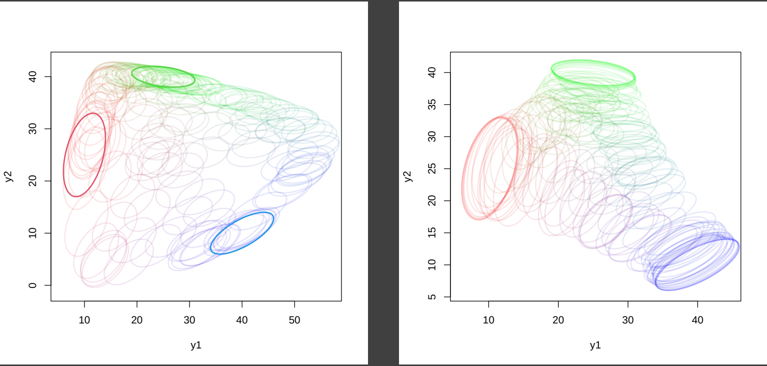
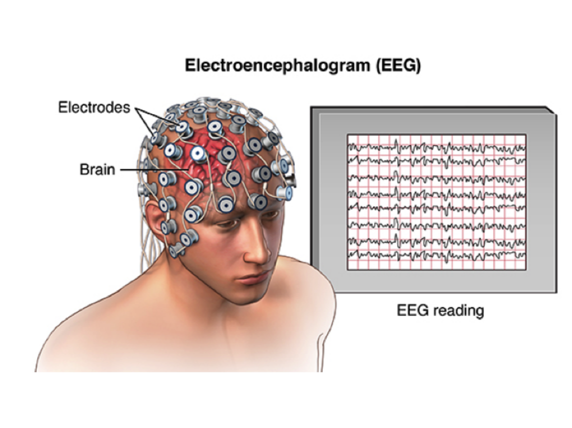


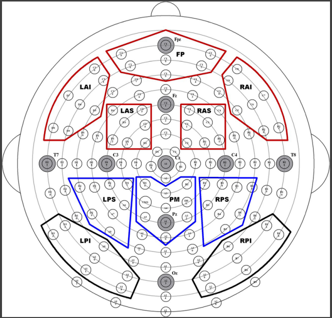
Table of Contents

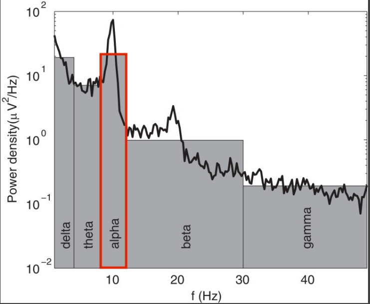
Motivation: EEG as a Functional Brain Imaging Modality



- EEG sensors measure distributed neuronal activity on cortical patches perpendicular to the sensors
- We study the response of a population of neurons – [Learning, memory formation, task execution, ...]

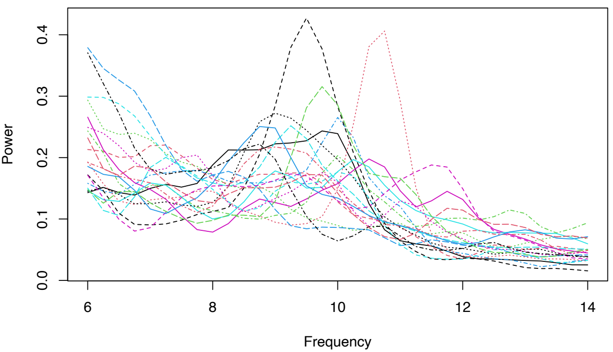
Resting State EEG and Spectral Features





- Power spectrum analysis associates spectral features in a specific frequency range, with bio-behavioral characterizations of brain activity
- We focus on the alpha frequency range, whose patterns at rest are thought to play a role in neural coordination and communication between distributed brain regions

EEG Spectral Power (ASD + TD)



- Can we use spectral power dynamics to identify latent neuro-developmental classes?
- Is the uncertain membership (clustering) framework appropriate for this application?

- Functional Data Analysis (FDA) focuses on methods used to analyze sample paths of an underlying continuous stochastic process Y
- Typically we consider:

  $$Y_i(t) = f_i(t) + \epsilon_i(t); \quad f_i(t) \sim GP\{\mu(t), C(\cdot,\cdot)\}; \quad \epsilon_i(t) \sim N(0, \sigma_\epsilon^2)$$

  Note: Often the literature on GP focuses on direct (parametrized) modeling of the covariance function $C(\cdot,\cdot)$

  Example: $C(s,t) = a^2 \exp\{-0.5||s-t||^2/\ell^2\}$

  FDA: Estimation of $C(s,t)$ from random samples $[Y_1(t), \ldots, Y_n(t)]$
- Established literature on flexible priors for $C(\cdot,\cdot)$ [Yang et al., 2017; Montagna et al., 2012; Shamshoian et al., 2022]

Functional Clustering (GP Mixtures)

- The FDA literature on clustering is very mature (James and Sugar, 2003; Chiu and Li, 2007)
- From a Bayesian perspective, assuming there exist K latent GPs

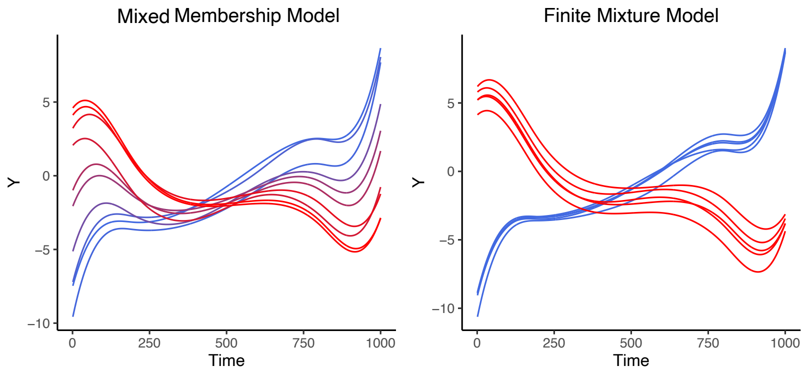
$$f^{(k)} \sim \mathcal{GP}\left(\mu^{(k)}, C^{(k)}\right), \quad k = 1, 2, \ldots, K$$

Each sample paths f_i, (i=1,2,…, N), follows a finite mixture of GPs:

$$p\left(f_i \mid \rho^{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \rho^{(k)} \, \mathcal{GP}\left(f_i \mid \mu^{(k)}, C^{(k)}\right);$$

where $\rho^{(k)} \in [0, 1]$ is the mixing proportion quantifying uncertain membership to the k^{th} GP

Functional Clustering vs. Functional Mixed Membership



Mixed Membership Functions

- Mixed membership process:

$$f_i \mid \mathbf{z}_i =_d \sum_{k=1}^{K} Z_{ik} f^{(k)}$$

- The proposed sampling model assumes

$$f_i \mid \boldsymbol{\Theta} \sim GP\left(\sum_k Z_{ik}\mu^{(k)},\ \sum_k Z_{ik}^2 C^{(k)} + \sum_k \sum_{k' \neq k} Z_{ik} Z_{ik'} C^{(k,k')}\right)$$

- Model K Gaussian Processes (GPs), $f^{(k)}$
  - K mean functions, $\mu^{(k)}(t)$
  - K covariance functions, $C^{(k,k)}(s,t)$
  - $\frac{K(K-1)}{2}$ cross-covariance functions, $C^{(k,j)}(t_k, t_j)$

Joint Representation of K Gaussian Processes

- We assume $f^{(k)}$ can be represented by a set of **uniformly continuous** basis functions.
- Let $\mathbf{B}(t)$ is a vector of the P basis functions evaluated at t
- The Multivariate Karhunen-Loève theorem (Happ and Greven, 2018) jointly decomposes K GPs:

$$f^{(k)}(t) = \boldsymbol{\nu}_k'\mathbf{B}(t) + \sum_{m=1}^{KP} \chi_m \boldsymbol{\phi}_{km}'\mathbf{B}(t), \tag{1}$$

  where $\boldsymbol{\nu}_k \in \mathbb{R}^P$, $\boldsymbol{\phi}_{km} \in \mathbb{R}^P$, and $\chi_m \sim \mathcal{N}(0,1)$
- Using this decomposition, we have:
  - $\mu^{(k)}(t) = \boldsymbol{\nu}_k'\mathbf{B}(t)$
  - $C^{(k,j)}(t_k, t_j) = \sum_{m=1}^{KP} \boldsymbol{\phi}_{km}'\mathbf{B}(t_k)\boldsymbol{\phi}_{jm}'\mathbf{B}(t_j)$

Multivariate Karhunen-Loève Theorem (cont.)

- The Karhunen-Loève theorem typically allows for a reduced dimensional representation with $M \leq KP$ components, s.t.

$$f^{(k)}(t) \approx \boldsymbol{\nu}_k' \mathbf{B}(t) + \sum_{m=1}^{M} \chi_m \boldsymbol{\phi}_{km}' \mathbf{B}(t), \qquad (2)$$

- Number of parameters needed to model the covariance structure:
  - Multivariate Karhunen-Loève: $\mathcal{O}(KPM)$
  - Naïve : $\mathcal{O}(K^2P^2)$

Finite Dimensional Margins

- $Z_{ik} \in (0,1) \longrightarrow$ mixed membership proportion of path i belonging to GP (k)
- Using the multivariate KL construction, we obtain:

$$y_i(t)|\boldsymbol{\Theta} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\left(\underbrace{\boldsymbol{\nu}_k'\mathbf{B}(t) + \sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}'\mathbf{B}(t)}_{f^{(k)}(t)}\right),\ \sigma^2\right) \tag{3}$$

- Integrating over χ_i yields

$$y_i(\mathbf{t}_i)|\boldsymbol{\Theta}_{-\chi} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\underbrace{\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\nu}_k}_{\boldsymbol{\mu}^{(k)}(\mathbf{t}_i)},\ \sum_{k=1}^{K}\sum_{j=1}^{K} Z_{ik}Z_{ij}\left(\underbrace{\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left(\boldsymbol{\phi}_{km}\boldsymbol{\phi}_{jm}'\right)\mathbf{S}(\mathbf{t}_i)}_{C^{(k,j)}(\mathbf{t}_i,\mathbf{t}_i)}\right) + \sigma^2\mathbf{I}_{n_i}\right) \tag{4}$$

Prior Distributions

- The ϕ parameters construct scaled eigenfunctions of the covariance operator
  - Mutually orthogonal
  - Magnitude of the scaled eigenfunctions should decrease
    - Multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011)

$$\phi_{kpm} | \gamma_{kpm}, \tilde{\tau}_{mk} \sim \mathcal{N}\left(0, \gamma_{kpm}^{-1} \tilde{\tau}_{mk}^{-1}\right),$$

$$\gamma_{kpm} \sim \Gamma\left(\nu_\gamma/2, \nu_\gamma/2\right), \quad \tilde{\tau}_{mk} = \prod_{n=1}^{m} \delta_{nk},$$

$$\delta_{1k} \sim \Gamma(a_{1k}, 1), \quad \delta_{jk} \sim \Gamma(a_{2k}, 1), \quad a_{1k} \sim \Gamma(\alpha_1, \beta_1), \quad a_{2k} \sim \Gamma(\alpha_2, \beta_2)$$

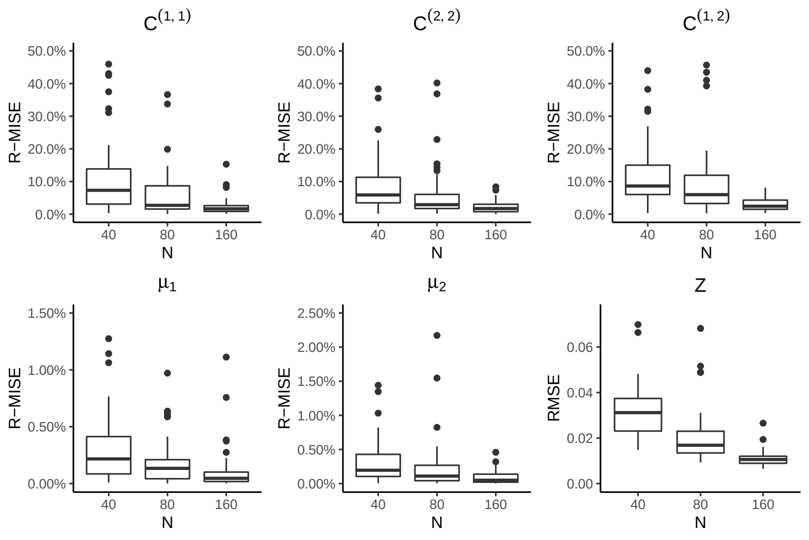
Posterior Distributions

- Let $\mathbf{\Sigma}_{jk} := \sum_{p=1}^{KP} \left(\boldsymbol{\phi}_{jp} \boldsymbol{\phi}'_{kp} \right)$ and

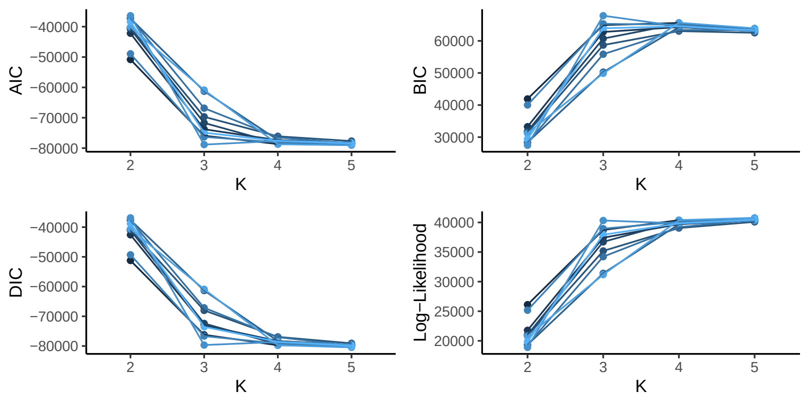
$$\boldsymbol{\omega} := \left\{ \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \mathbf{\Sigma}_{11}, \ldots, \mathbf{\Sigma}_{1K}, \ldots, \mathbf{\Sigma}_{KK}, \sigma^2 \right\}.$$

- The parameters in $\boldsymbol{\omega} \in \mathbf{\Omega}$ completely specify the mean and covariance structure of our model. We will denote the true set of parameters as $\boldsymbol{\omega}_0$
- Assumptions:
  1. $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are observed on a grid of R points ($R > KP$) in the domain, $\{t_1, \ldots, t_R\}$
  2. The variables Z_{ik} are fixed and known (not-random)
  3. $\sigma_0^2 > 0$
- Consider the fully saturated model (M = KP). Under these assumptions, the posterior distribution is weakly consistent at $\boldsymbol{\omega}_0 \in \mathbf{\Omega}$

Operating Characteristics on Engineered Data

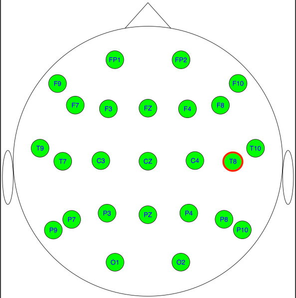


Selecting the Number of Features



Case Study: Peak Alpha Frequency (TD and ASD)

- Autism spectrum disorder (ASD) is a term used to describe individuals with a collection of social communication deficits and restricted or repetitive sensory-motor behaviors
- This case study contains electroencephalogram (EEG) data for 39 typically developing (TD) children and 58 children with ASD between the ages of 2 and 12 years old
- We fit a 2 functional feature mixed membership model on data from the T8 electrode



EEG Case Study Data

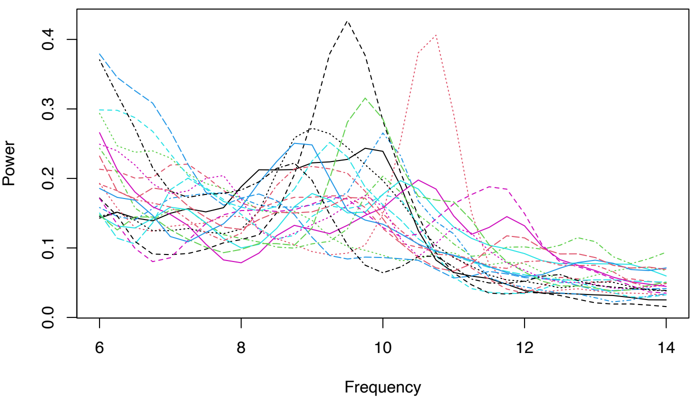


Figure: EEG data from the T8 electrode for 20 individuals (ASD and TD)

EEG Case Study Data (cont.)

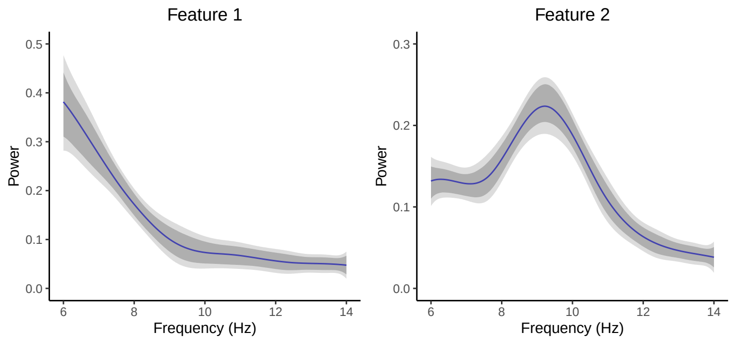
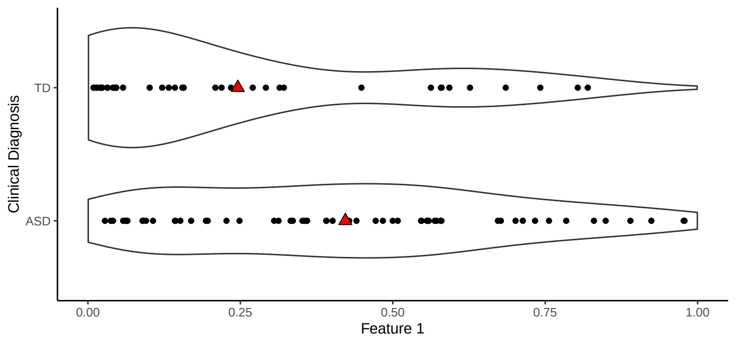


Figure: Posterior median and 95% credible (pointwise credible interval in dark gray and simultaneous credible interval in light gray) of the mean function for each latent functional feature.

EEG Case Study (cont.)



- Children with an TD clinical diagnosis are highly likely to load on the second functional feature, whereas children with ASD exhibit a higher level of heterogeneity

EEG Case Study Data (Functional Clustering)

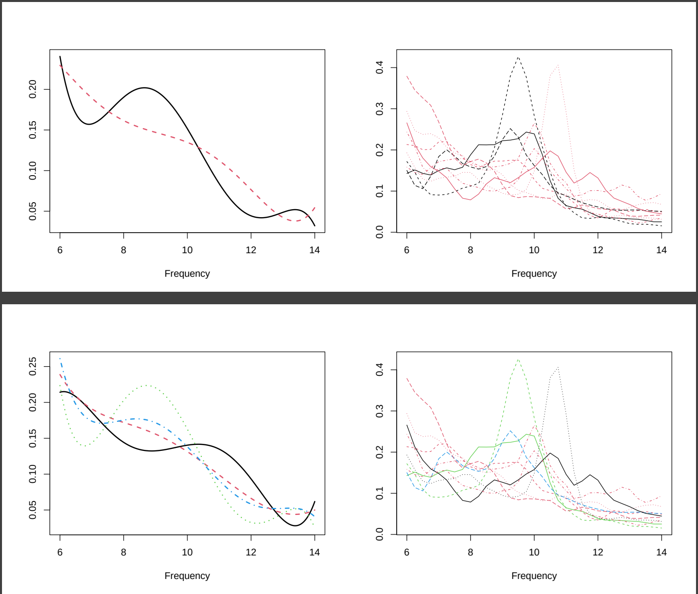
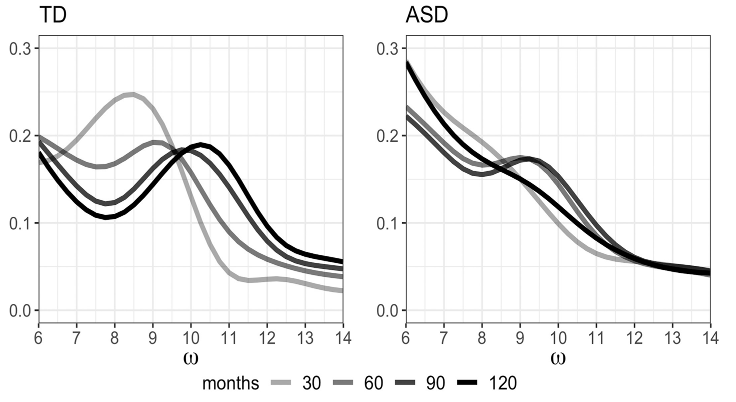


Table of Contents

Motivation: Peak Alpha Frequency Shift with Aging

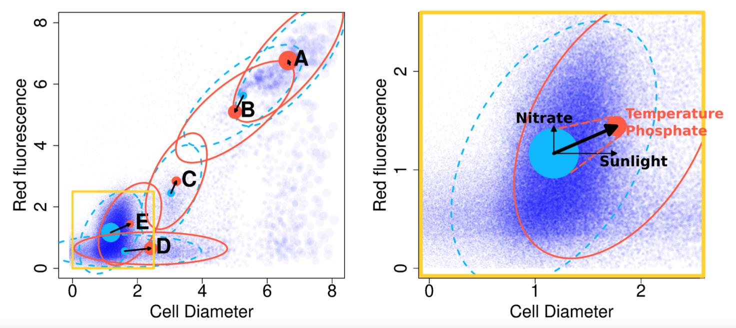
- As typically developing children grow, the alpha peak tends to becomes more prominent and the PAF shifts to a higher frequency (Rodríguez-Martínez et al., 2017; Scheffler et al., 2019)



(Scheffler et al., 2019)

Covariate Adjusted Clustering

- *Mixture of Experts models* and *Mixture of Regressions models* are two common covariate-dependent clustering models where the mean components of the mixtures are dependent on the covariates of interest
  - Mixture of Experts models also assume that cluster membership also depends on the covariates of interest



(Hyun et al., 2022)

- Gaussian finite mixture models (GFMM) can be expressed as

$$f_i \mid \boldsymbol{\pi}_i, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}, \sum_{k=1}^{K} \pi_{ik}C^{(k)}\right)$$

- Similarly, we can extend the framework of GFMMs to arrive at the Mixture of Regressions Model framework (5) and Mixture of Experts framework (6):

$$f_i \mid \mathbf{X}, \boldsymbol{\Theta} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}(\mathbf{x}_i), \sum_{k=1}^{K} \pi_{ik}C^{(k)}\right) \quad (5)$$

$$P\left(f_i \mid \mathbf{X}, \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k)\, \mathcal{GP}\left(f_i \mid \mu^{(k)}(\mathbf{x}_i), C^{(k)}\right) \quad (6)$$

- The mean function, $\mu^{(k)}(\mathbf{x}_i)$, is often modeled through a regression framework

Function-on-Scalar Regression

- Function-on-scalar regression is a common method in FDA which allows the mean structure of the continuous stochastic process to be covariate-dependent
  - The covariates of interest are scalar or vector-valued, while the response is functional
- The general form of function-on-scalar regression can be expressed as follows:

$$Y(t) = \mu(t) + \sum_{r=1}^{R} X_r \beta_r(t) + \epsilon(t); \quad t \in \mathcal{T}, \tag{7}$$

- The mean function ($\mu(t)$) and the functional coefficients ($\beta_r(t)$) are infinite dimensional parameters, making inference intractable
  - We typically assume that the data lie in the span of a finite set of basis functions $(b_1(t), \ldots, b_p(t))$
    - *A-priori* specified basis functions
    - Data-driven basis functions (F-PCA)

Function-on-Scalar Regression, Mixture of Regressions, and CAFMM Models

- Function-on-scalar regression can be considered a population level analysis, where the covariates have the same effects on each observation
- Gaussian mixture of regressions models can be considered a sub-population level analysis, where covariates the covariate effects on the mean structure depend on which cluster an observation belongs to

$$f_i \mid \mathbf{X}, \boldsymbol{\Theta} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\left(\mu_k + \sum_{r=1}^{R} X_{ir}\beta_{kr}\right), \sum_{k=1}^{K} \pi_{ik} C^{(k)}\right)$$

- Covariate adjusted functional mixed membership (CAFMM) models can be considered an individual level analysis, where each observation has a different allocation vector
  - Each underlying feature has a unique mean structure (covariate-dependent) and covariance structure

Extension to CAFMM Models

- The functional mixed membership model can be expressed as

$$\mathbf{x}_i | \mathbf{z}_{(1:N)} =_d \sum_{i=1}^{K} Z_{ik} \mathbf{f}_k,$$

where

$$f^{(k)} \sim \mathcal{GP}\left(\mu^{(k)}, C^{(k)}\right), \quad k = 1, 2, \ldots, K$$

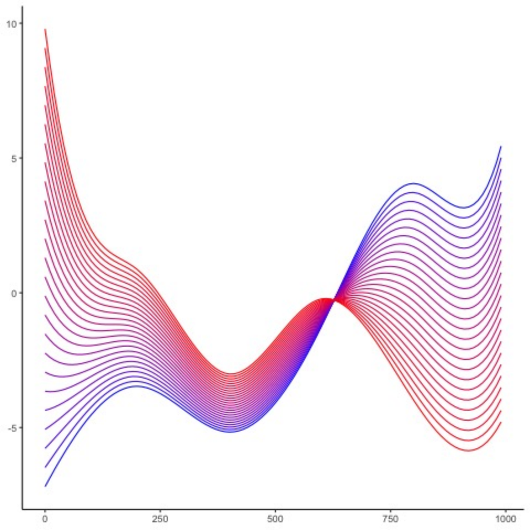
- This leads to the following likelihood:

$$f_i \mid \boldsymbol{\Theta} \sim GP\left(\sum_k Z_{ik}\mu^{(k)}, \sum_k Z_{ik}^2 C^{(k)} + \sum_k \sum_{k' \neq k} Z_{ik} Z_{ik'} C^{(k,k')}\right)$$

- Leveraging the function-on-scalar framework, we can arrive at the general form of the proposed CAFMM model

$$f_i \mid \boldsymbol{\Theta} \sim GP\left(\sum_k Z_{ik}\left(\mu^{(k)} + X_{ir}\beta_{rk}\right), \sum_k \sum_{k'} Z_{ik} Z_{ik'} C^{(k,k')}\right)$$

Example of a Covariate Adjusted Mean Structure



Finite Dimensional Marginal Distributions

- Let $\mathbf{x}_i \in \mathbb{R}^R$ be the vector of covariates for the i^{th} observation
- Using the multivariate KL construction and the assumption that the features lie in the user-defined basis, we obtain the functional model:

$$\mathbf{Y}_i(\mathbf{t}_i)|\boldsymbol{\Theta}, \mathbf{X} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\left(\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right) + \sum_{m=1}^{M} \chi_{im}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\phi}_{km}\right)\right), \; \sigma^2 \mathbf{I}_{n_i}\right\}$$

- Integrating our the χ_{im} parameters, we have

$$y_i(\mathbf{t}_i)|\boldsymbol{\Theta}_{-\chi} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right), \; \sum_{k=1}^{K}\sum_{j=1}^{K} Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left(\boldsymbol{\phi}_{km}\boldsymbol{\phi}_{jm}'\right)\mathbf{S}(\mathbf{t}_i)\right) + \sigma^2\mathbf{I}_{n_i}\right) \tag{8}$$

- $\boldsymbol{\eta}_k \in \mathbb{R}^{P\times R}$ represents the covariate adjustment to the mean structure of the k^{th} feature

Identifiability

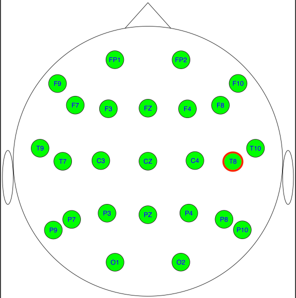
- Let $\boldsymbol{\omega}$ be a set of parameters
- The parameters $\boldsymbol{\omega}$ are unidentifiable if there exists at least one $\boldsymbol{\omega}^* \neq \boldsymbol{\omega}$ such that $\mathcal{L}(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i) = \mathcal{L}(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}^*, \mathbf{x}_i)$ for all sets of observations $\{\mathbf{Y}_i(\mathbf{t}_i)\}_{i=1}^N$
  - Otherwise, the parameters $\boldsymbol{\omega}$ are called identifiable
- The *label switching* problem is a common source of unidentifiability in finite mixture models.
- What conditions do we need on the parameters $\boldsymbol{\omega}$ and design matrix $\mathbf{X}$ to ensure identifiability?

Lemma: Consider a two feature ($K = 2$) covariate adjusted model as specified in Equation 39. The parameters $\boldsymbol{\nu}_k$, $\boldsymbol{\eta}_k$, Z_{ik}, $\sum_{m=1}^{M}(\boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm})$, and σ^2 are identifiable up to a permutation of the labels (i.e. label switching), for $k, k' = 1, 2$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$, given the following assumptions:

1. $\mathbf{X}$ is full column rank with $\mathbf{1}$ not in the column space of $\mathbf{X}$.
2. The separability condition holds on the allocation matrix (there exists $\tilde{i}_1, \tilde{i}_2$ such that $Z_{\tilde{i}_1 1} = 1$ and $Z_{\tilde{i}_2 2} = 1$). Moreover, there exists at least 2 observations with allocation parameters that lie in the interior of the unit simplex $\left(\text{i.e. } \mathbf{z}_i \in \left\{\mathbf{z} \in \mathbb{R}^2 \mid \sum_{k=1}^{2} Z_k = 1, 0 < Z_k < 1\right\}\right)$.
3. The sample paths $\mathbf{Y}_i(\mathbf{t}_i)$ are sampled such that $n_i \geq P$, and furthermore, there exists a sample path $\mathbf{Y}_i(\mathbf{t}_i)$ such that $n_i > 4M$.

Revisiting the ASD Case Study

- Autism spectrum disorder (ASD) is a term used to describe individuals with a collection of social communication deficits and restricted or repetitive sensory-motor behaviors
- This case study contains electroencephalogram (EEG) data for 39 typically developing (TD) children and 58 children with ASD between the ages of 2 and 12 years old
- We fit a 2 CAFMM model on data from the T8 electrode with Age as the covariate of interest



Function-on-Scalar Regression (Covariates: Age)

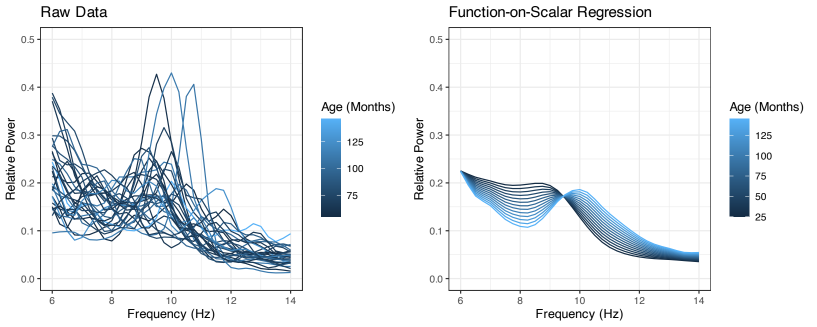


Figure: (Left) Data colored by age at the time of recording. (Right) Results from a function-on-scalar regression.

CAFMM Model (Covariates: Age)

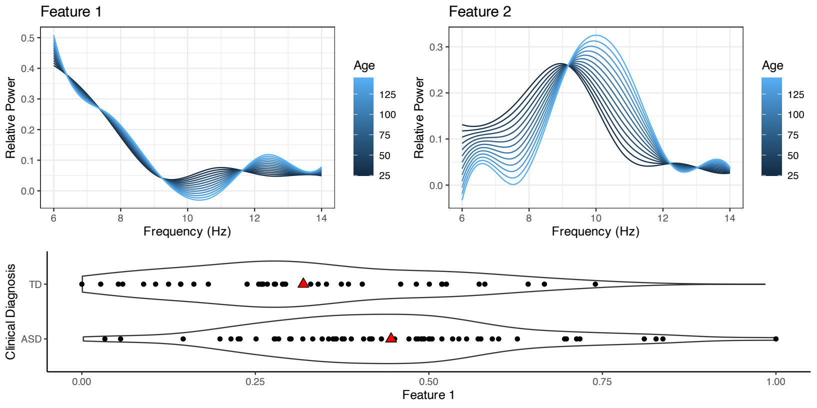


Figure: (Top Left) Mean of the first feature at various ages. (Top Right) Mean of the second feature at various ages. (Bottom) Estimated allocation features stratified by clinical diagnosis.

Function-on-Scalar Regression (Covariates: Age and Clinical Diagnosis)

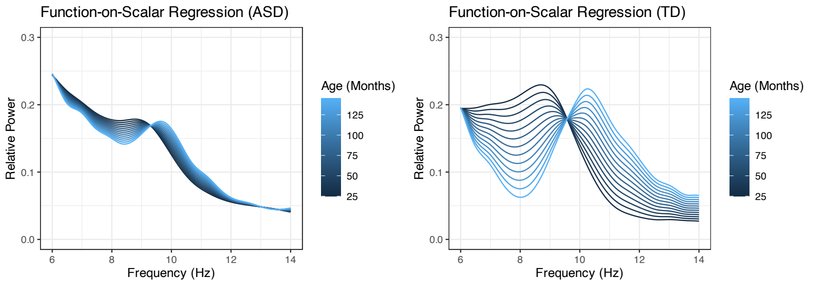
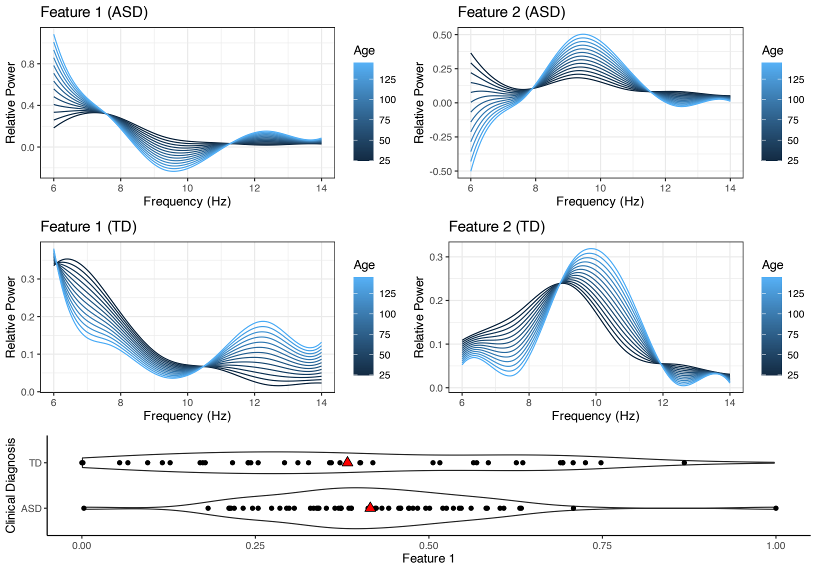


Figure: Results from a function-on-scalar regression with age and clinvial diagnosis as the covariates of interest.

CAFMM Model (Covariates: Age and Clinical Diagnosis)



CAFMM Model (Covariates: Age and Clinical Diagnosis)

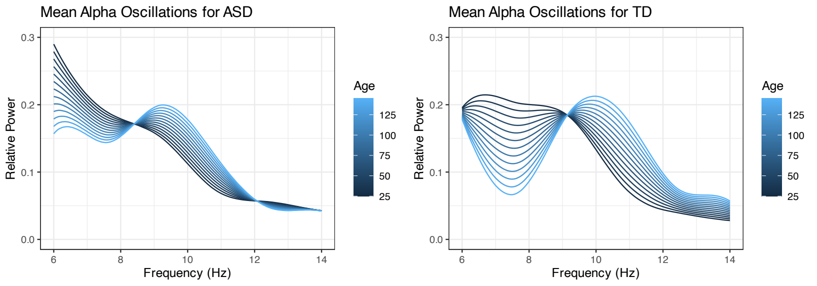


Figure: Estimated average developmental trajectory of alpha oscillations stratified by diagnostic group.

Summary

- Interpretable sampling models allow us to easily interpret the mean and covariance structure
- Multivariate KL constructions allow for efficient representation and dimension reduction of multivariate GPs
- In our applications, results are robust to increasing dimensionality (multi-channel analyses)
- Covariate adjusted functional mixture models can be thought of as a generalization of function-on-scalar regression

Thank You!

R Packages

BayesFMMM   Funct. Mixed Membership Models   https://github.com/ndmarco/BayesFMMM

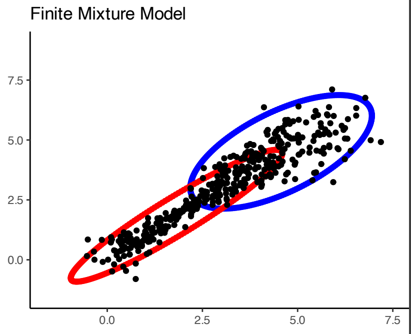
Manuscripts

- Marco N, Senturk D, Jeste S, Dickinson A and D. Telesca D (2022) *Functional Mixed Membership Models.* (arXiv:2206.12084).
- Marco N, Senturk D, Jeste S, Dickinson A and D. Telesca D (2022) *Flexible Regularized Estimation in High-Dimensional Mixed Membership Models* (arXiv:2212.06906)

Funding

R01 MH122428-01 (DS, DT) from the NIH/NIMH

Construction of a Finite Mixture Model



- Let $\pi_{ik} \in \{0, 1\}$ $(\sum_k \pi_{ik} = 1)$ denote whether or not the i^{th} observation belongs to the k^{th} cluster, by the law of total probability we have:

$$P(Y_i) = P(Y_i \mid \pi_{i1} = 1)\color{red}{P(\pi_{i1} = 1)} + \cdots + P(Y \mid \pi_{iK} = 1)\color{red}{P(\pi_{iK} = 1)}$$

$$= \sum_{k=1}^{K} \color{red}{\rho_k} P(Y_i \mid \pi_{ik} = 1)$$

Construction of a Finite Mixture Model

- Assuming that the distributions of the clusters are in the exponential family, we have

$$P(Y_i \mid \boldsymbol{\theta}_{1:K}) = \rho_k P(Y_i \mid \boldsymbol{\theta}_k)$$

- Using the latent variables $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{iK}]$ ($\pi_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} \pi_{ik} = 1$), we have

$$P\left(Y_i \mid \boldsymbol{\pi}_i, \boldsymbol{\theta}_{(1:K)}\right) = \sum_{\boldsymbol{\pi}_i} P(\boldsymbol{\pi}_i) \prod_{i=1}^{K} P\left(Y_i \mid \boldsymbol{\theta}_k\right)^{\pi_{ik}},$$

where $P(\pi_{ik} = 1) = \rho_k$

Extension to Heller's Characterization of Partial Membership Models

- Let $\mathbf{z}_i = [Z_{i1}, \ldots, Z_{iK}]$, where $Z_{ik} \in [0, 1]$ and $\sum_k Z_{ik} = 1$, represent the i^{th} observation's proportion of membership to the K^{th} feature
- Using these latent variables, we arrive at the general form proposed in Heller et al. (2008):

$$P\left(Y_i \mid \mathbf{z}_i, \boldsymbol{\theta}_{(1:K)}\right) \propto \int_{\mathbf{z}_i} P(\mathbf{z}_i) \prod_{i=1}^{K} P\left(Y_i \mid \boldsymbol{\theta}_k\right)^{Z_{ik}} \mathrm{d}\mathbf{z}_i$$

- Assuming the distributions of the features are in the exponential family (i.e. $Y_i \mid \boldsymbol{\theta}_k \sim \text{Expon}(\boldsymbol{\theta}_{\boldsymbol{k}})$), we have

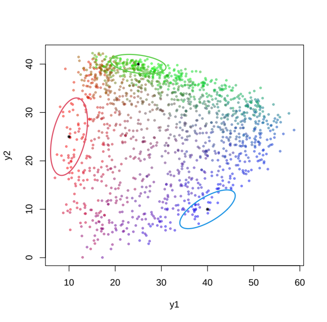
$$Y_i \mid \mathbf{z}_i, \boldsymbol{\theta}_{(1:K)} \sim \text{Expon}\left(\sum_k Z_{ik} \boldsymbol{\theta}_k\right)$$

Extension to Heller's Characterization of Partial Membership Models

- Assuming that the features follow a Gaussian distribution, where $\boldsymbol{\nu}_k$ and $\mathbf{C}_k$ denote the corresponding mean and covariance parameters of the k^{th} feature, we have that

$$Y_i \mid \mathbf{z}_i, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)} \sim \mathcal{N}\left(\mathbf{H}_i \mathbf{h}_i, \mathbf{H}_i\right),$$

where $\mathbf{h}_i = \sum_{k=1}^{K} \pi_{ik} \mathbf{C}_k^{-1} \boldsymbol{\nu}_k$ and $\mathbf{H}_i = \left(\sum_{k=1}^{K} \pi_{ik} \mathbf{C}_k^{-1}\right)^{-1}$



Extension to our Proposed Mixed Membership Model

- In a Gaussian finite mixture model, we have:

$$p\left(\mathbf{x}_i|\rho_{(1:K)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) = \sum_{\boldsymbol{\pi}_i} p(\boldsymbol{\pi}_i) \prod_{i=1}^{K} \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\nu}_k, \mathbf{C}_k\right)^{\pi_{ik}}$$

- If we condition on the membership parameters, we get:

$$\mathbf{x}_i|\boldsymbol{\pi}_{(1:N)} =_d \sum_{i=1}^{K} \pi_{ik}\mathbf{f}_k,$$

  where $\mathbf{f}_k \sim \mathcal{N}\left(\boldsymbol{\nu}_k, \mathbf{C}_k\right)$

- Thus we can rewrite the likelihood as:

$$\mathbf{x}_i|\boldsymbol{\pi}_{(1:N)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)} \sim \mathcal{N}\left(\sum_{k=1}^{K} \pi_{ik}\boldsymbol{\nu}_k, \sum_{k=1}^{K} \pi_{ik}\mathbf{C}_k\right)$$

Extension to our Proposed Mixed Membership Model

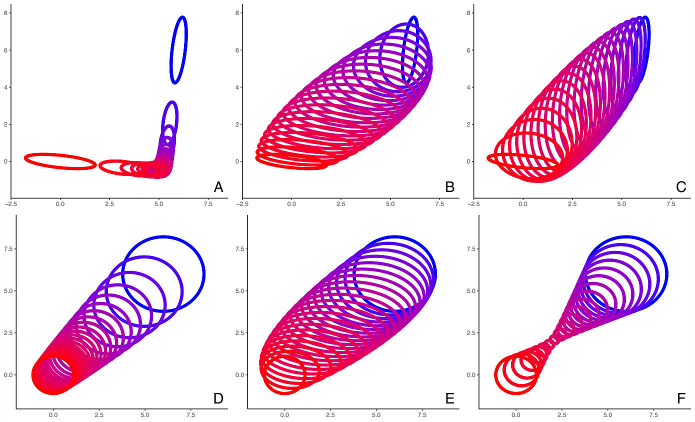
- We can extend this to our partial membership model by introducing variables $\mathbf{z}_i = [Z_{i1}, \ldots, Z_{iK}]$ ($Z_{ik} \in [0,1]$, $\sum_k Z_{ik} = 1$) such that:

$$\mathbf{x}_i | \mathbf{z}_{(1:N)} =_d \sum_{i=1}^{K} Z_{ik} \mathbf{f}_k$$

- We can't assume that the *features* ($\mathbf{f}_k$) are independent
- Let $\mathbf{C}^{(k,k')} = \mathrm{Cov}(\mathbf{f_k}, \mathbf{f_{k'}})$ denote the cross-covariance between the feature k and feature k'
- Letting $\boldsymbol{\mathcal{C}}$ denote the collection of covariance and cross-covariance matrices, we have

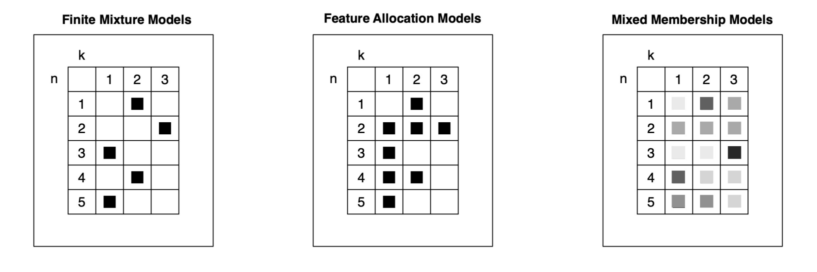
$$\mathbf{x}_i | \mathbf{z}_{(1:N)}, \boldsymbol{\nu}_{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik} \boldsymbol{\nu}_k, \sum_{k=1}^{K} Z_{ik}^2 \mathbf{C}_k + \sum_{k=1}^{K} \sum_{k \neq k'} Z_{ik} Z_{ik'} \mathbf{C}^{(k,k')} \right)$$

Relation to Other Mixed Membership Models



- The proposed representation is more flexible and interpretable compared to other MMMs (i.e. Heller et al., 2008).

Visualizations of Clustering Models



Joint Decomposition

- Letting $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K]$, we we have that

$$\text{Cov}(\text{vec}(\mathbf{F})) = \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{C}^{(1,1)} & \ldots & \mathbf{C}^{(1,K)} \\ \vdots & \ddots & \vdots \\ \mathbf{C}^{(K,1)} & \ldots & \mathbf{C}^{(K,K)} \end{bmatrix}$$

- Letting $\boldsymbol{\Phi}_m = [\boldsymbol{\phi}'_{1m} \ldots \boldsymbol{\phi}'_{Km}]'$ be scaled eigenvectors of $\boldsymbol{\Sigma}$, we have

$$\mathbf{C}^{(k,k')} = \sum_{m=1}^{PK} \phi_{km}\phi'_{k'm}$$

- Thus we have that $\text{vec}(\mathbf{F}) \approx \text{vec}(\boldsymbol{\mu}) + \sum_{m=1}^{M} \chi_m \boldsymbol{\Phi}_m$ or $\mathbf{f}_k \approx \boldsymbol{\nu}_k + \sum_{m=1}^{M} \chi_m \boldsymbol{\phi}_{km}$, where $\chi_m \sim \mathcal{N}(0,1)$

Model Specification

- Using the approximation, we obtain:

$$\mathbf{y}_i|\boldsymbol{\Theta} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik} \underbrace{\left(\boldsymbol{\nu}_k + \sum_{m=1}^{M} \chi_{im}\boldsymbol{\phi}_{km}\right)}_{f^{(k)}(t)}, \sigma^2\mathbf{I}_P\right)$$

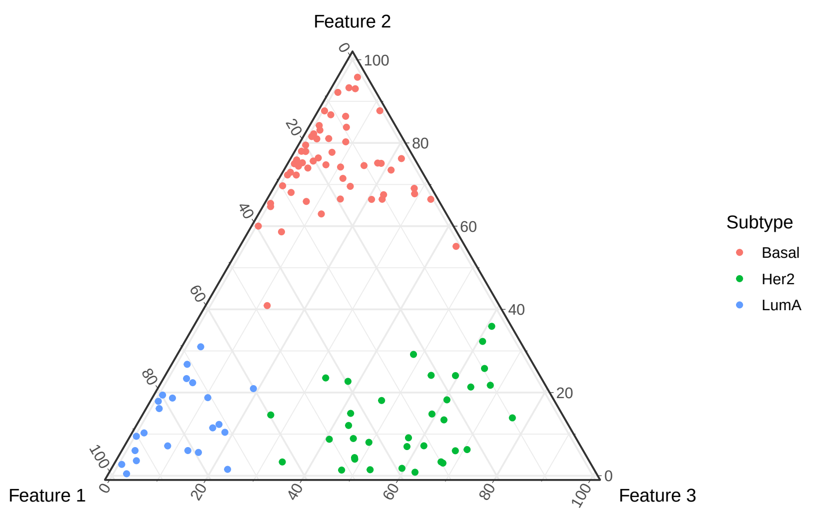
- If we integrate out the χ_{im} variables, we obtain:

$$\mathbf{y}_i|\boldsymbol{\Theta}_{-\chi} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\nu}_k, \left(\sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'} \underbrace{\left(\sum_{m=1}^{M} \boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm}\right)}_{C^{(k,k')}}\right) + \sigma^2\mathbf{I}_P\right)$$

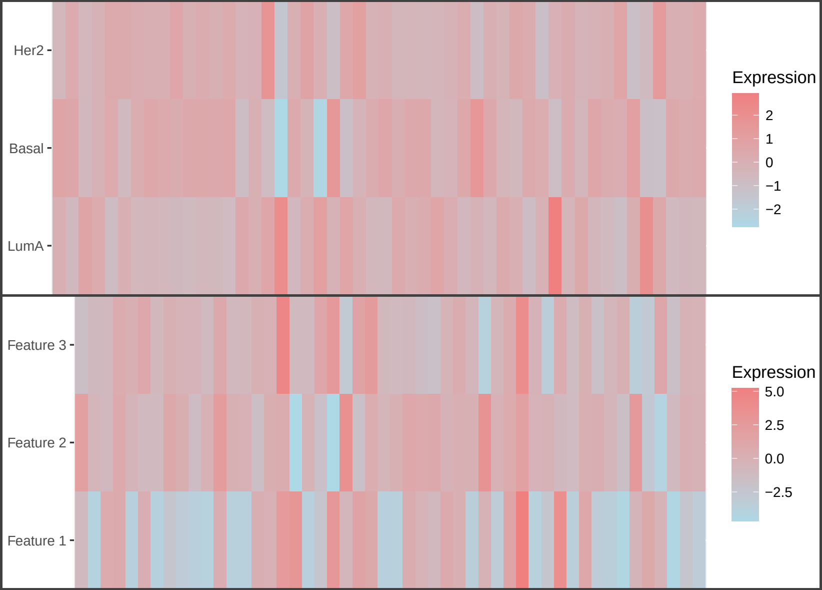
Case Study: Molecular Subtypes of Breast Cancer

- In 2014, there were an estimated 534,000 deaths due to breast cancer worldwide (Wang et al.,2016)
- In the past two decades, 5 molecular subtypes of breast cancer have been discovered; each with a different prognosis, risk factors, and treatment sensitivity (Prat et al., 2015)
- In 2009, Parker et al. discovered that the cancer subtype can be accurately classified by centroid-based prediction methods using gene expression data from 50 genes (PAM50)
- We fit a 3 feature mixed membership model on gene expression data from PAM50, using patients with LumA, Basal, and Her2 cancer subtypes

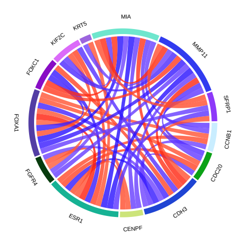
Case Study: Molecular Subtypes of Breast Cancer

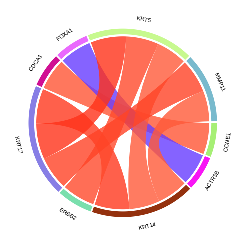


Case Study: Molecular Subtypes of Breast Cancer



Case Study: Molecular Subtypes of Breast Cancer





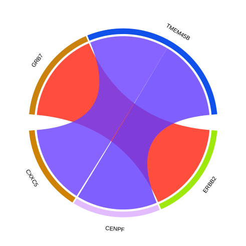
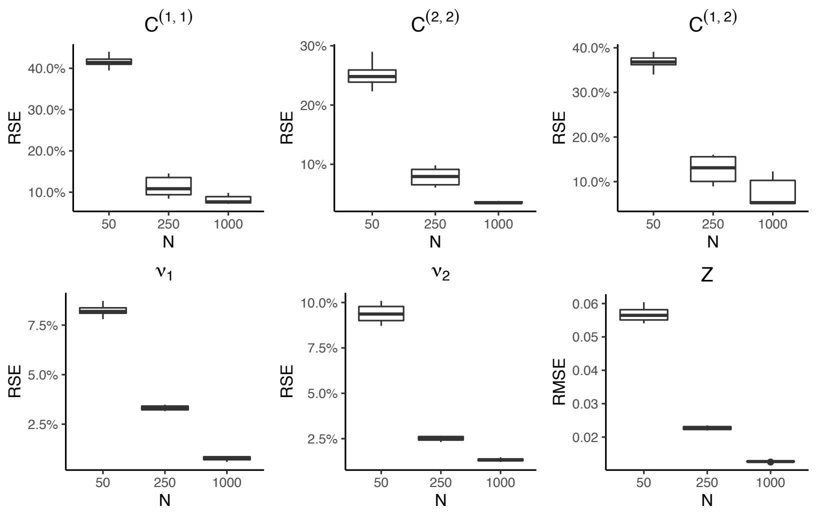
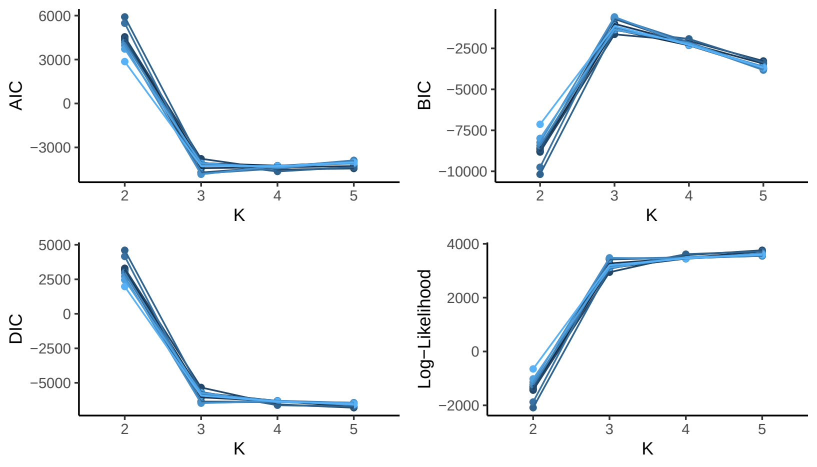


Figure: Visualization of the correlation structure of the each feature (Feature 1: Left, Feature 2: Middle, Feature 3: Right)

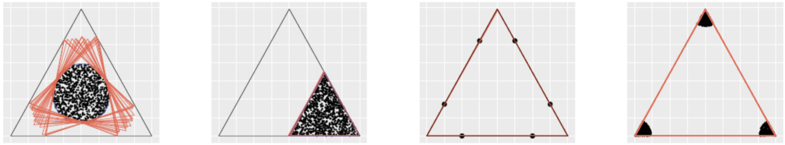
Simulation Study: Recovery of Parameters



Simulation Study: Information Criteria



Unidentifiability of Allocation Parameters



(a) non-identifiable (b) non-identifiable (c) identifiable (d) identifiable

Identifiability of Allocation Parameters

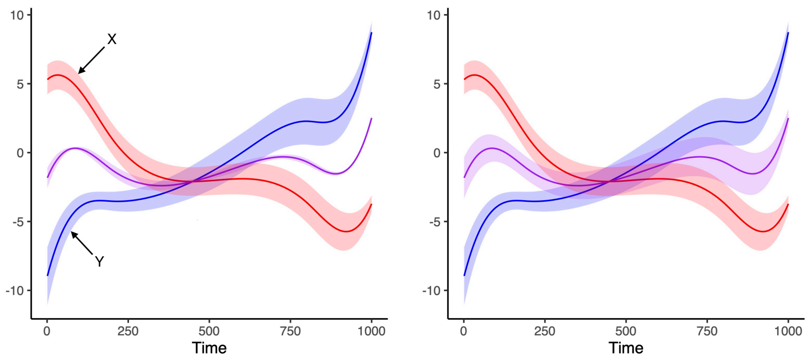
- *Seperability* condition: at least one observation belongs entirely in each feature
- *Sufficiently Scattered* condition: an allocation matrix $\mathbf{Z}$ is sufficiently scattered if:

  1. $\text{cone}(\mathbf{Z}')^* \subseteq \mathcal{K}$
  2. $\text{cone}(\mathbf{Z}')^* \cap bd\mathcal{K} \subseteq \{\lambda \mathbf{e}_f, f = 1, \ldots, k, \lambda \geq 0\}$

  where $\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 \leq \mathbf{x}'\mathbf{1}_K\right\}$,
  $bd\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 = \mathbf{x}'\mathbf{1}_K\right\}$,
  $\text{cone}(\mathbf{Z}')^* := \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x}\mathbf{Z}' \geq 0\right\}$, and $\mathbf{e}_f$ is a vector with the i^{th} element equal to 1 and zero elsewhere.

Effects of the Cross-Covariance Function

$$\text{Cov}^{(X,Y)}(s,t) = \text{Cov}\left(X(s), Y(t)\right)$$



EEG Case Study (cont.)

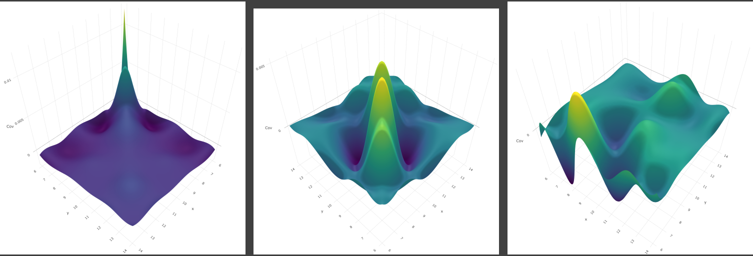
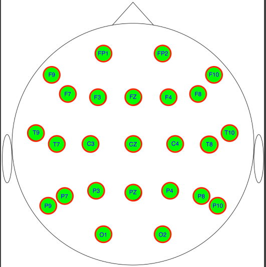


Figure: Posterior estimates of the covariance functions (From left to right: covariance of feature 1, covariance of feature 2, cross-covariance between features 1 and 2)

- In the previous case study, we only used the T8 electrode and discarded the information from the 24 other electrodes
- For this case study, we will model all electrodes using a functional model, assuming $\mathcal{T} \subset \mathbb{R}^3$
  - Two of the indices will contain the spatial location of the electrodes
  - The third index will contain the frequency domain



Analysis of Multi-Channel EEG Data (cont.)

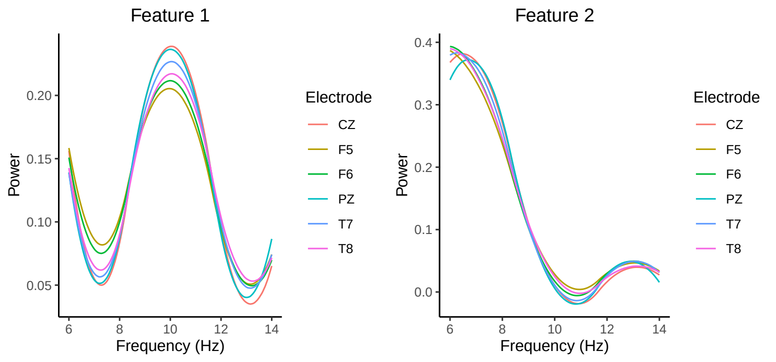


Figure: Posterior estimates of the means of the two functional features viewed at specific electrodes of interest

Analysis of Multi-Channel EEG Data (cont.)

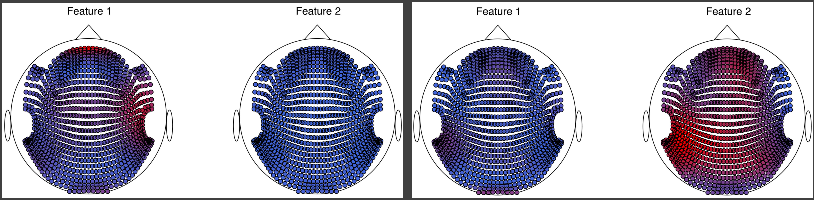
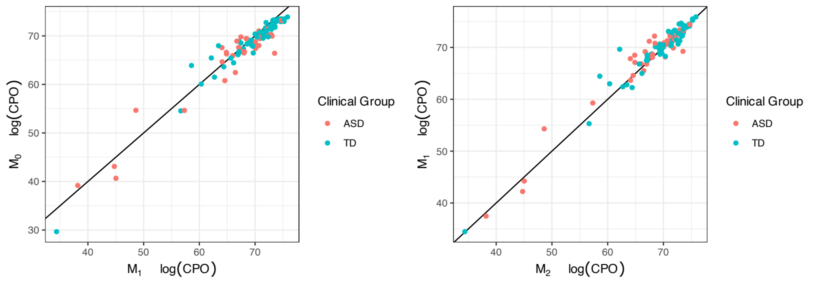


Figure: Variance of electrodes at 6 Hz (left) and 10 Hz (right)

- For the second functional feature, we can see that there is high heterogeneity around the T8 electrode at 6 Hz

Conditional Predictive Ordinate (CPO)



Sim Study (Covariate Adjusted)

Truth / Model (# Covariates)	Parameter	$N = 60$	$N = 120$	$N = 240$
2/2	μ_1	1.9% (0.3%, 24.7%)	1.1% (0.2%, 10.4%)	0.3%(0.1%, 8.8%)
	μ_2	1.5% (0.4%,14.5%)	1.0% (0.2%, 10.5%)	0.2% (0.1%, 10.9%)
	$C^{(1,1)}$	156.1% (2.1%, 112219.4%)	110.3% (0.1%, 1806067.0%)	6.1% (0.1%, 362938.9%)
	$C^{(2,2)}$	88.1% (1.8%, 60673.8%)	416.2% (1.9%, 1008651.0%)	4.9% (0.5%, 22725.8%)
	$C^{(1,2)}$	431.2% (3.5%, 35924.4%)	433.7% (2.2%, 246646.3%)	22.3% (0.6%, 29231.3%)
	Z	0.047 (0.020, 0.099)	0.030 (0.013, 0.074)	0.013 (0.008, 0.054)
		$N = 50$	$N = 100$	$N = 200$
1/1	μ_1	1.5% (0.2%, 7.6%)	0.8% (0.1%, 4.9%)	1.1%(0.2%,5.4%)
	μ_2	1.6% (0.3%,5.7%)	1.2% (0.2%, 7.6%)	1.2% (0.2%, 5.4%)
	$C^{(1,1)}$	218.5% (26.0%, 11299.6%)	30.8% (14.4%, 308.4%)	37.1% (9.5%, 421.2%)
	$C^{(2,2)}$	204.4% (22.5%, 2603.4%)	40.2% (8.3%, 597.6%)	25.5% (5.7%, 157.7%)
	$C^{(1,2)}$	219.8% (42.9%, 1912.9%)	89.1% (21.2%, 403.0%)	60.6% (13.0%, 350.2%)
	Z	0.067 (0.047, 0.085)	0.056 (0.042, 0.081)	0.051 (0.040, 0.065)
1/0	μ_1	382.2% (153.4%, 961.9%)	650.7% (91.1%, 1511.0%)	1076.7%(94.8%,2339.0%)
	μ_2	394.6% (117.5%,1292.3%)	751.4% (69.0%, 1721.0%)	885.1% (145.0%, 2313.0%)
	$C^{(1,1)}$	1581365.0% (81644.7%, 23059352.5%)	1328559.4% (64656.5%, 40230314.1%)	1348112.9% (98035.6%, 65828353.0%)
	$C^{(2,2)}$	730829.2% (133764.2%, 9829513.4%)	1015747.1% (86551.9%, 17361755.8%)	802590.5% (44704.4%, 21037857.8%)
	$C^{(1,2)}$	1271237.9% (90303.1%, 9356418.4%)	1917180.3% (91394.3%, 20373022.9%)	1392890.2% (81254.1%, 19419032.6%)
	Z	0.202 (0.180, 0.217)	0.172 (0.157, 0.184)	0.144 (0.121, 0.156)
		$N = 40$	$N = 80$	$N = 160$
0/1	μ_1	2.3% (0.3%, 36.7%)	2.5% (0.2%, 33.6%)	1.9%(0.2%,20.4%)
	μ_2	4.1% (0.3%,36.1%)	1.9% (0.3%, 21.6%)	3.8% (0.2%, 26.1%)
	$C^{(1,1)}$	27.1% (7.7%, 703.6%)	19.1% (3.3%, 95.5%)	20.3% (3.1%, 64.9%)
	$C^{(2,2)}$	28.9% (9.4%, 319.1%)	19.0% (3.7%, 206.9%)	13.5% (3.0%, 74.8%)
	$C^{(1,2)}$	31.4% (8.8%, 353.3%)	24.2% (7.7%, 61.2%)	26.9% (4.9%, 67.1%)
	Z	0.0957 (0.070, 0.148)	0.083 (0.061, 0.107)	0.068 (0.048, 0.088)
0/0	μ_1	0.23% (0.04%, 1.23%)	0.12% (0.01%, 0.35%)	0.04%(0.01%,0.31%)
	μ_2	0.27% (0.09%,0.88%)	0.12% (0.02%, 0.42%)	0.04% (0.01%, 0.31%)
	$C^{(1,1)}$	3.5% (0.9%, 16.0%)	1.9% (0.3%, 7.4%)	1.3% (0.3%, 4.4%)
	$C^{(2,2)}$	4.5% (0.6%, 18.0%)	1.6% (0.3%, 8.0%)	1.1% (0.2%, 4.5%)
	$C^{(1,2)}$	5.3% (1.1%, 19.9%)	2.0% (0.6%, 9.5%)	1.3% (0.6%, 5.4%)
	Z	0.032 (0.023, 0.049)	0.018 (0.013, 0.024)	0.011 (0.009, 0.015)