



THEORY AND METHODS

Modeling Missing at Random Neuropsychological Test Scores Using a Mixture of Binomial Product Experts

Daniel Suen and Yen-Chi Chen

Statistics, University of Washington, USA

Corresponding author: Daniel Suen; Email: dsuen@uw.edu

(Received 7 January 2025; revised 17 September 2025; accepted 18 September 2025)

Abstract

Multivariate bounded discrete data arises in many fields. In the setting of dementia studies, such data are collected when individuals complete neuropsychological tests. We outline a modeling and inference procedure that can model the joint distribution conditional on baseline covariates, leveraging previous work on mixtures of experts and latent class models. Furthermore, we illustrate how the work can be extended when the outcome data are missing at random using a nested EM algorithm. The proposed model can incorporate covariate information and perform imputation and clustering. We apply our model to simulated data and an Alzheimer's disease data set.

Keywords: binomial product mixture; latent variable models; missing data; mixture models; multivariate discrete data

1. Introduction

Modeling multivariate discrete data is a common problem across many fields, such as social sciences, psychology, and ecology. For instance, in education research, discrete data arises from students' test scores (Maydeu-Olivares & Liu, 2015), and in ecology, discrete data arises when the counts of a given animal population are measured over various areas and time periods (Anderson et al., 2019; Maslen et al., 2023). As the size of data sets continues to expand, missing data becomes increasingly prevalent (Kang, 2013), and it is common to have access to a rich set of discrete variables that is subject to missingness. Both the multivariate discreteness and the missingness alone can create modeling difficulties, but together, they present a unique challenge for statistical inference.

1.1. The National Alzheimer's Coordinating Center database

Although the general problem statement is widely important, our work is primarily motivated by the analysis of neuropsychological test scores in the database of the National Alzheimer's Coordinating Center (NACC).¹ The NACC is an NIH/NIA-funded center that maintains the largest longitudinal database for Alzheimer's disease in the United States. The NACC coordinates 33 Alzheimer's Disease and Research Centers (ADRC) in the United States.

¹<https://naccdata.org/>

In the NACC data, we have 41,181 individuals with data collected from 2005 to 2019. In this study, we are particularly interested in information regarding the neuropsychological tests. The neuropsychological tests are a set of examinations measuring an individual's cognitive ability in four different domains: language, attention, executive function, and memory. The scores from these tests are important features for dementia research because they are not based on a clinician's judgment. Similar to a conventional exam, the outcome of a neuropsychological test is a discrete number taking integer values and has a known range. Note that different tests have a different ranges. In our study, we consider eight neuropsychological test scores per individual.

A goal of this article is to introduce a simple and interpretable statistical model for modeling the neuropsychological tests. The discrete and correlated nature of the neuropsychological tests presents a challenge for modeling them. As far as we are aware, there is no statistical model for handling these test scores directly, not to mention the additional challenges from the missingness of the test scores. To simplify the problem, we focus on a static model that only considers the test scores from each individual's initial visit. While there are many covariates that could be incorporated in the analysis, we choose to include only four demographic variables (age, education, sex, and race) to reduce the complexity of the model. These variables are key demographic variables that are often included in any dementia research.

1.2. Research questions

Because the ranges of exams are generally large (30–300), it is infeasible to nonparametrically model the joint distribution of the eight test scores. Moreover, we have demographic variables, such as education, age, and sex. It is a nontrivial problem to study the relation between these demographic variables and the test scores. Furthermore, the existence of missing test scores among certain individuals further compounds the complexity of the overall analysis. In this article, our primary focus centers on addressing the following research questions:

1. *Finding a feasible and interpretable model for multiple discrete outcomes with missing values.* As mentioned before, our data contain multiple dependent discrete variables with missing values. We need to design a proper model to model the dependency among discrete variables and deal with the missing data problem. We also need an estimation and inference procedure that quantifies the uncertainty in the model while properly accounting for the missing data in a statistically principled way.
2. *Discovering latent groups of individuals using neuropsychological tests.* Clinicians have developed a set of rules to categorize individuals into different clinical groups (normal cognition, MCI, and dementia). However, this rule is based on clinical judgments and does not involve any neuropsychological information. Clustering has been of interest to the dementia research community (Alashwal et al., 2019). Thus, an ideal model should be able to cluster individuals into groups using neuropsychological test scores that represent multiple cognitive domains of interest.
3. *Investigating the association between neuropsychological tests and other variables.* It is often of a lot of interest to study how neuropsychological tests are associated with demographic variables or clinical judgments. For instance, there is a hypothesis that people with a higher education are more resilient to cognitive decline as a form of cognitive reserve (Meng & D'Arcy, 2012; Thow et al., 2018). We also want to use the NACC data to test this hypothesis.

1.3. Literature review

We provide some background on the current research in Alzheimer's disease and statistics methodology literature.

1.3.1. Dementia-related research

Multiple outcome variables are common in dementia-related research, but there is no clear widespread solution to the modeling problem. In some previous methods in the Alzheimer's disease literature, multiple test scores are standardized and averaged into a single holistic score for an individual (Boyle et al., 2018). This turns the multiple outcome problem into a single outcome problem and ignores the dependence structure. Mjørud et al. (2014) examined variables associated with multiple quality-of-life-related discrete outcomes and performed univariate regression on each outcome. Dimension reduction is also commonly used to reduce the number of outcome variables (Qiu et al., 2019; Yesavage et al., 2016), but this can lead to loss of interpretability as any downstream analysis is not using the original variables. Missing data is often ignored completely or when accounted for, single imputation methods or an off-the-shelf imputation approach are typically used (Brenowitz et al., 2017; Qiu et al., 2019).

Clustering has also become increasingly valuable in the dementia research community (Alashwal et al., 2019). For example, Escudero et al. (2011) used the *K*-means algorithm to divide subjects into pathologic and non-pathologic groups to study the early detection of Alzheimer's disease. Tosto et al. (2016) used *K*-means on a subset of NACC data to identify subgroups within Alzheimer's disease patients to understand the heterogeneity of the disease. Several papers also described model-based clustering approaches. For example, De Meyer et al. (2010) clustered biomarker data using a simple two-component Gaussian mixture model for early detection of Alzheimer's disease. Qiu et al. (2019) used neuropsychological test scores from a smaller data set than ours, imputing missing data using single imputation methods and the *mice* R package (van Buuren & Groothuis-Oudshoorn, 2011). Then, they applied principal components analysis before finally, using a Gaussian mixture model for clustering. In our work, we choose to model the raw test scores to avoid loss of information and preserve more interpretability. It is crucial to be cautious when employing single imputation methods or off-the-shelf packages, as they can potentially lead to an underestimation of the uncertainty associated with missing data, and the underlying assumptions regarding missing data may remain unclear.

Additionally, the cognitive reserve hypothesis emerged several decades ago when it was observed that after autopsies, some individuals with no dementia symptoms had brains that exhibited advanced Alzheimer's pathology (Katzman et al., 1989). Researchers hypothesized that some activities in life may provide individuals with a given resilience to cognitive decline, which is known as the cognitive reserve hypothesis. As such, there is deep interest in discovering the association between various covariates and dementia (Meng & D'Arcy, 2012; Stern, 2012; Thow et al., 2018; Zhang et al., 1990).

1.3.2. Methodology research

This article sits at the intersection of multiple subfields of statistics, including latent variable modeling, missing data, and clustering. The latent class model (LCM) was first proposed by Goodman (1974) for the purpose of modeling multivariate discrete data. The basic form of the LCM is of a mixture of multivariate multinomials. When the number of levels of the discrete random variable is large, this can lead to a large number of parameters (Bouveyron et al., 2019). There has been much work studying many aspects of the LCM, including the identifiability (Allman et al., 2009; Carreira-Perpinan & Renals, 2000; Gu & Xu, 2020; Gyllenberg et al., 1994) and incorporation of covariates (Huang & Bandeen-Roche, 2004; Ouyang & Xu, 2022; Vermunt, 2010). The traditional LCM has been extended in various forms, such as a mixture of Rasch models (Rost, 1990) and a mixture of item response models (McParland & Gormley, 2013).

The mixture of expert models is closely related to the mixture model framework and the LCM (Jacobs et al., 1991). It generalizes the standard mixture model by allowing the parameters to potentially depend on covariates (Gormley & Frühwirth-Schnatter, 2018). The increased flexibility of the models is associated with a large increase in the number of parameters. Many estimation procedures have been explored, including those based on an EM approach (Jacobs et al., 1991; Jordan & Jacobs, 1993), an EMM approach (Gormley & Murphy, 2008), or a Bayesian framework (Bishop & Svenskn, 2002; Frühwirth-Schnatter & Kaufmann, 2008).

Clustering has previously been explored in combination with missing data. Lee & Harel (2022) proposed a two-stage clustering framework by first clustering multiply imputed data sets individually and then, obtaining a final clustering by clustering the set of cluster centers obtained over all imputed data sets. In previous research on model-based clustering with missing data, Serafini et al. (2020) employed the EM algorithm and Monte Carlo methods to estimate Gaussian mixture models in the presence of missing at random (MAR) data. Sportisse et al. (2024) addressed model-based clustering under missing-not-at-random (MNAR) assumptions, employing a likelihood-based approach and EM algorithm. We include further comments on these methods in [Appendix D](#) of the Supplementary Material. These methods addressed the clustering problem but did not use covariates or discuss quantifying the uncertainty of the parameters. Since our method is motivated by the statistical analysis of Alzheimer's disease, handling both of these is critical to our work.

1.4. Outline

Our article is organized as follows. Section 2 presents a detailed description of our *mixture of binomial product experts* model. The handling of missing data is discussed in Section 3, where we describe the fitting of the model under a nonmonotone MAR assumption. In Section 4, we delve into the process of inference, specifically in constructing confidence intervals for the parameters of interest. We outline how to perform clustering in Section 5. In Section 6, we provide simulation results to examine the consistency and the coverage of our proposed estimator. In Section 7, we apply our method to an Alzheimer's disease data set, which motivates our model formulation. Finally, Section 8 concludes the article with a discussion of the results. In the Supplementary Material, we include further details on derivations and proofs (Appendices A and H of the Supplementary Material), identification theory (Appendix B of the Supplementary Material), code implementation ([Appendix C](#) of the Supplementary Material), remarks on related EM algorithms (Appendix D of the Supplementary Material), simulations ([Appendix E](#) of the Supplementary Material), comments on model selection ([Appendix F](#) of the Supplementary Material), comments on clustering (Appendix G of the Supplementary Material), and more comments on the real data analysis ([Appendix I](#) of the Supplementary Material). We provide a code implementation of our method in an R package.

2. Mixture of binomial product experts

2.1. A latent class model for neuropsychological test scores

Suppose we have n IID observations, indexed by $i = 1, \dots, n$. Let $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^\top \in \mathcal{X} \subset \mathbb{R}^p$ be the covariates and $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,d})^\top \in \mathcal{Y} \subset \mathbb{Z}^d$ be the outcome variables of interest, where $\mathcal{Y}$ is a bounded discrete set. In the present study, $\mathbf{Y}_i$ represents the neuropsychological test scores of individual i for a total of d total tests, but this framework can be applied to other problems as well. In the article, for notational convenience, we may elect to drop the index i to simplify notation and refer to the generic random variable that is not associated with a specific individual. The j th test score/outcome variable Y_j belongs to the set $\{0, 1, 2, \dots, N_j\}$. We assume that the covariates are always observed and that only the outcomes may be subject to missingness.

We start with a simpler scenario in which the data are completely observed (no absence of any outcome variables). The more complicated setting with missing outcome variables is addressed in Section 3. When we have multiple discrete outcome variables, the nonparametric approach is not feasible due to the curse of dimensionality. To see this, suppose we have four outcome variables, and each outcome variable has ten possible values. The joint distribution has a dimension of 10^4 , surpassing the order of magnitude of moderately sized samples. This motivates to use a parametric model.

To model the dependency among discrete variables, we consider a mixture model, which is based on the idea of the LCM. Originally presented in its modern form by Goodman (1974), the LCM takes the

form of the following mixture of multivariate multinomials distribution:

$$p(y; \theta) = \sum_{k=1}^K w_k p_k(y; \theta_k) = \sum_{k=1}^K w_k \prod_{j=1}^d p_{k,j}(y_j; \theta_{k,j}), \quad (2.1)$$

where $p_{k,j}$ denotes a multinomial with $\theta_{k,j}$ being its the parameter and the w_k 's are probability weights for each component such that $\sum_k w_k = 1$ and $w_k > 0$ for all k . As $p_{k,j}$ is a multinomial distribution, $\theta_{k,j}$ refers to the vector of probabilities (one for each level of Y_j) that sum to 1. There is a popular implementation for fitting these models through the *poLCA* package (Linzer & Lewis, 2011).

Let Z denote the component assignment such that $Z = \ell$ if and only if the associated observation is generated from mixture component ℓ . Namely, $P(Y = y|Z = k; \theta) = \prod_{j=1}^d p_{k,j}(y_j; \theta_{k,j})$. By construction, the LCM implicitly assumes that for $j \neq j'$, the variables Y_j and $Y_{j'}$ are conditionally independent given the latent class Z . This is known as the local independence assumption and is commonly used in the literature to model dependence between discrete random variables. While they are conditionally independent, the variables Y_j and $Y_{j'}$ are allowed to depend marginally for $j \neq j'$. The local independence assumption makes it convenient to calculate the conditional distributions, which are used in Section 3 when we discuss imputation. Some effort has been made to relax the local independence assumption, such as mixtures of log-linear models (Bock, 1994) and hierarchical LCMs (Chen et al., 2012; Zhang, 2004), but these have not been as widely adopted due to the increase in the number of parameters.

Throughout this article, we will refer to the data set containing tuples of the form $(\mathbf{X}, \mathbf{Y})$ as the *latent incomplete* (LI) data and the data set containing tuples of the form $(\mathbf{X}, \mathbf{Y}, Z)$ as the *latent complete* (LC) data. In the setting of no missing data, the LI data are what we typically have access to, but this is often insufficient for direct model fitting due to the unobserved latent variable Z . We will show that one can estimate the model parameters using an EM algorithm with the LI data. We use this nomenclature to avoid confusion in Section 3 when we encounter the traditional type of missingness with the outcome variables.

There has been previous work to incorporate covariates in the LCM (Huang & Bandeen-Roche, 2004; Vermunt, 2010). This is related to the *mixture of experts* models that arose from the machine learning literature (Jacobs et al., 1991; Jordan & Jacobs, 1993; Yuksel et al., 2012). These models generalize the classical mixture model formulation by allowing the model parameters w_k and θ_k to depend on covariates. The *simple mixture of experts* model (Gormley & Frühwirth-Schnatter, 2018) takes the following form:

$$p(y|x; \theta) = \sum_{k=1}^K w_k(x; \beta_k) p_k(y; \theta_k),$$

where β_k is introduced to allow the weights to depend on covariates. As written, the covariates only adjust the weights placed on each mixture component but do not affect the parameters of the individual component distribution.

To apply these ideas to the neuropsychological data, we combine the mixture of experts and LCM, leading to the following mixture of binomial product experts:

$$p(y|x; \beta, \theta) = \sum_{k=1}^K w_k(x; \beta_k) \underbrace{\prod_{j=1}^d \binom{N_j}{y_j} (\theta_{k,j})^{y_j} (1 - \theta_{k,j})^{N_j - y_j}}_{p_k(y; \theta_k)}, \quad (2.2)$$

where $w_k(x; \beta_k) = \exp(\beta_k^T \cdot (1, x)) / (\sum_{k' \in [K]} \exp(\beta_{k'}^T \cdot (1, x)))$ is from the multiple-class logistic regression model. Like in the classical LCM, we utilize the local independence assumption as each mixture decomposes as a product of binomials. However, there is a critical difference between our model and the

classical LCM. We use the binomial distribution instead of the multinomial distribution, and this leads to a large reduction in the number of parameters. Since test scores are ordered variables, we expect the reduction to a binomial distribution to be reasonable as scores further away from the mean score can be less likely. Similar to the simple mixture of experts model, each mixture weight depends on covariates.

Here, we will treat $k = 1$ as the reference group, so β_1 is the vector of all 0 by assumption and does not need to be estimated. For $k \in \{2, 3, \dots, K\}$, $\beta_k \in \mathbb{R}^{p+1}$, and for $k \in [K]$, $\theta_k \in \mathbb{R}^d$, we assume that the w_k utilize the logistic function, but it is possible to use other link functions, such as the probit and other monotonic functions (Ouyang & Xu, 2022). In our work, the outcome variables are neuropsychological test scores, and we treat the latent groups as subgroups of the population of varying cognitive ability. The five levels of the CDR score offer a natural choice for the number of mixture components.

In our model construction, we view each mixture component as representing a subgroup of the population, where the subgroups have varying cognitive ability that is summarized by θ . Therefore, we interpret the θ 's as attributes of population-level groups and do not allow them to depend on covariates. Since the weights $w_k(x; \beta_k)$ depend on covariates, every individual has their own weights associated with each population-level group. These covariates can be interpreted as a way to construct weights on an individual level for each of these population-level subgroups. A side product of this assumption is that this reduces model complexity. If the θ 's depend on covariates, then the number of parameters increases from $O(K(d+p))$ to $O(Kdp)$, which is a dramatic change even for moderately sized d and p and thus another reason we do not allow this as part of our model. Because the outcome variables are test scores, we can also interpret the quantity $N_j \cdot \theta_{k,j}$ as the mean of the j th test score of the k th latent group, and this is useful as a summary statistic for a given latent class.

Binomial product distributions have previously arisen in the literature. One such application was for modeling test scores for spatial tasks in child development in Thomas et al. (1993), but they considered a fairly limited setting with no covariates, two components, and two outcome variables. Binomial product and Poisson product distributions have also been used in the statistical ecology literature to model species abundance (Brintz et al., 2018; Haines, 2016; Kéry et al., 2005). These settings differ from ours because they often treat the total number counts N_j as a random quantity. In item response theory (IRT), the Rasch model posits that the probability of answering each question correctly depends on its difficulty and the individual's ability, assuming a Bernoulli distribution for each question (Rasch, 1960). However, in our data, only total test scores are available. Thus, assuming uniform difficulty across test questions, we model the scores using Binomial distributions. To our knowledge, our article is the first time binomial products has been used to analyze neuropsychological test score data while incorporating covariates.

Note that for the k th component, $\beta_k \in \mathbb{R}^{p+1}$ and $\theta_k \in \mathbb{R}^d$, there are $(K-1)(p+1)$ parameters associated with the covariates β and Kd parameters associated with θ and a total of $K(p+d+1) - p - 1$ parameters for the whole model. Thus, for fixed covariate and outcome dimensions, the number of parameters grows linearly in K .

2.2. Model fitting

We now describe our procedure for fitting the model. An intuitive approach is to estimate the parameters via the maximum likelihood approach. However, computing the maximum likelihood estimator (MLE) is a nontrivial task because the LI log-likelihood is not concave (Lemma 1 in the Supplementary Material). As such, solving the first-order conditions is no longer sufficient for determining the global maximizer.

Due to Lemma 1 (stated formally in Appendix H.1 of the Supplementary Material), maximizing the LI log-likelihood function directly is not straightforward. Here is an interesting insight from the mixture model literature: if we had access to the latent class label, then the MLE would be easy to construct. This insight motivates us to develop an EM algorithm (Dempster et al., 1977) by augmenting the observed data with the true component label Z to calculate the MLE for (β, θ) . For each $i = 1, \dots, n$, let $Z_i = \ell$ if the i th observation comes from mixture component ℓ . Such data augmentation allows one to bypass the

problem of taking the logarithm of a sum. To see this, let $\mathbf{X}_{1:n} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ and $\mathbf{Y}_{1:n} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ be the covariates and outcome variables of all n individuals, respectively. The LC log-likelihood function is written as

$$\begin{aligned}\ell_{\text{LC},n}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}, \mathbf{Z}_{1:n}) &= \log \left(\prod_{i=1}^n \prod_{k=1}^K (w_k(\mathbf{X}_i; \boldsymbol{\beta}_k) p_k(\mathbf{Y}_i; \boldsymbol{\theta}_k))^{1(Z_i=k)} \right) \\ &= \sum_{i=1}^n \sum_{k=1}^K 1(Z_i = k) \log w_k(\mathbf{X}_i; \boldsymbol{\beta}_k) + \sum_{i=1}^n \sum_{k=1}^K 1(Z_i = k) \log p_k(\mathbf{Y}_i; \boldsymbol{\theta}_k).\end{aligned}\quad (2.3)$$

In the EM algorithm, we start from an initial guess $(\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\boldsymbol{\theta}}^{(0)})$ and iterate between an expectation step (E-step) and a maximization step (M-step) to update our guess until convergence. Let t denote the t th iteration of the EM algorithm. In the E-step, we compute the expected value of the complete log-likelihood $\ell_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z})$ conditional on the observed data $\mathbf{X}$ and $\mathbf{Y}$. The expected value of the LC log-likelihood given the observed data forms the Q function in the EM algorithm

$$Q_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) := \mathbb{E}[\ell_{\text{LC}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, \mathbf{Y}, \mathbf{Z}) | \mathbf{X}, \mathbf{Y}; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}].$$

In practice, we do not have access to the true expectation, so we work with the sample analog $Q_{\text{LC},n}$. The sample analog can be expressed as

$$\begin{aligned}Q_{\text{LC},n}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) \\ := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log w_k(\mathbf{X}_i; \boldsymbol{\beta}_k) + \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) \log p_k(\mathbf{Y}_i; \boldsymbol{\theta}_k),\end{aligned}\quad (2.4)$$

where

$$\widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_i; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) = \frac{w_k(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}_k^{(t)}) p_k(\mathbf{Y}_i; \widehat{\boldsymbol{\theta}}_k^{(t)})}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}_{k'}^{(t)}) p_{k'}(\mathbf{Y}_i; \widehat{\boldsymbol{\theta}}_{k'}^{(t)})} \quad (2.5)$$

is the weight of observation $(\mathbf{X}_i, \mathbf{Y}_i)$ for the k th mixture component.

Since the LC log-likelihood function (2.3) decomposes as the sum of a function of $\boldsymbol{\beta}$ and a function of $\boldsymbol{\theta}$, the maximization step of the EM algorithm is separable. Note that since w_k and p_k are logistic regression and binomial product models, respectively, one can apply standard off-the-shelf model fitting procedures after reweighting each observation i by $\widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$. In the maximization step, the new estimate for $\boldsymbol{\theta}$ is updated as follows for each $k \in [K]$ and $j \in [d]$,

$$\widehat{\boldsymbol{\theta}}_{k,j}^{(t+1)} := \frac{\sum_{i=1}^n \frac{Y_{i,j}}{N_j} \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)}{\sum_{i=1}^n \widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)}, \quad (2.6)$$

which is simply the standard MLE formed by the sample proportion, reweighted by $\widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$. We update our estimate of $\boldsymbol{\beta}_k$ by fitting a multi-class logistic regression by reweighting each corresponding observation by $\widehat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$. This is equivalent to regressing the variable $W^{(t)}$ on the covariates X , where $W_i^{(t)} = (\widehat{\pi}_1^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \dots, \widehat{\pi}_K^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)) \in S^{K-1}$ for all i , where S^{K-1} is the $(K-1)$ th probability simplex. Note that this is not the standard logistic regression because the outcome variable belongs to a probability simplex. This logistic regression can still be fit using gradient descent in standard R packages.

Algorithm 1 describes the process of fitting the mixture of binomial product experts model in the presence of completely observed covariate and outcome data. For a single initialization point, the EM algorithm is not guaranteed to converge to the global maximizer. In practice, we run Algorithm 1 many times with multiple random initial guesses to explore the parameter space sufficiently and choose the parameter estimate from all of them that maximizes the log-likelihood.

Algorithm 1 Fitting the mixture model via EM.**Require:** $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n, (\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)})$

- 1: Start with an initial guess $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)})$.
- 2: Initialize the iteration step $t = 0$.
- 3: **repeat**
- 4: Compute the weights $\hat{\pi}_k^{(t)}(\mathbf{X}_i, \mathbf{Y}_i)$ for all observations i and mixture components k using (2.5).
- 5: Form the probability vector $W_i^{(t)} = (\hat{\pi}_1^{(t)}(\mathbf{X}_i, \mathbf{Y}_i), \dots, \hat{\pi}_K^{(t)}(\mathbf{X}_i, \mathbf{Y}_i))$ for each i .
- 6: Update $\hat{\boldsymbol{\theta}}^{(t+1)}$ using (2.6).
- 7: Update $\hat{\boldsymbol{\beta}}^{(t+1)}$ by fitting a logistic regression model of $W^{(t)}$ on $\mathbf{X}$.
- 8: Update $t \leftarrow t + 1$.
- 9: **until** Convergence
- 10: Let $\hat{\boldsymbol{\beta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\theta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\theta}}^{(t)}$.
- 11: **return** $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

2.3. Identifiability

We now make some brief comments on the identifiability of our model. Identifiability means that each parameter value corresponds to a unique probability distribution. That is, the mapping from the parameter space to the space of probability distributions is one-to-one. If identifiability does not hold, estimation becomes problematic because a unique MLE may not exist. However, this notion of identifiability may be too strong for many practical purposes. For example, even the common problem of label swapping violates this identifiability definition, so researchers often consider the idea of identifiability up to permutation of the parameters. We now consider the notion of *generic identifiability*, which relaxes the definition of identifiability even further. Generic identifiability means identifiability holds almost everywhere in the parameter space (Allman et al., 2009). More formally, this means that the mapping from the parameter space to the space of probability distributions may fail to be one-to-one only on a set of Lebesgue measure zero. Generic identifiability (holding almost everywhere) can be viewed as an intermediate assumption between two common assumptions: global identifiability (holding everywhere) and local identifiability (holding in a neighborhood of the true parameter). In a practical sense, this means that any such model fit on a given data set is unlikely to be unidentifiable, and we consider this notion to be sufficient for our applied data setting.

Proposition 1 (Sufficient conditions for generic identifiability). *Suppose the following conditions hold:*

- A1 *Each mixture is distinct such that $\theta_k \neq \theta_{k'}$ when $k \neq k'$.*
- A2 *The number of outcome variables d and the number of mixtures K satisfy the bound $d \geq 2\lceil \log_{1+\min_j N_j} K \rceil + 1$.*
- A3 *The design matrix is full-rank and $n > p$.*

Then, the mixture of binomial product experts is generically identifiable up to permutation of the parameters.

We note that the sufficient conditions outlined in Proposition 1 are fairly mild. Intuitively, one would not want the number of parameters to be too large when the data dimension d is small. The bound in Assumption A2 ensures that the number of mixture components K remains appropriately controlled relative to d and N_j , as a large K increases model complexity. When there are at least $d = 3$ outcome variables of moderate range, K can still be fairly large. Assumption A3 is to ensure identifiability of the

logistic regression parameters. We provide further comments on the assumptions and the proof of this proposition in [Appendix B](#) of the Supplementary Material. These assumptions are fairly mild, and we show that they can be met using relatively straightforward generating processes through our simulations in Section 6.

3. Missingness in the outcome variables

Another challenge in the NACC's neuropsychological data is the missingness in the outcomes. To properly describe the missing data problem, we first introduce some additional notation. Let $\mathbf{R}_i = (R_{i,1}, R_{i,2}, \dots, R_{i,d})^\top \in \mathcal{R} \subseteq \{0,1\}^d$ be the random binary vector that denotes the missing pattern associated with individual i . Element $R_{i,j}$ of the binary vector $\mathbf{R}_i$ is 1 if and only if $Y_{i,j}$ is observed. For a given pattern $r \in \mathcal{R}$, let $\mathbf{Y}_{i,r} = (Y_{i,j} : r_{i,j} = 1)$ denote the observed variables and $\mathbf{Y}_{i,\bar{r}} = (Y_{i,j} : r_{i,j} = 0)$ denote the missing variables. For example, when $d = 4$ and $r = 1,001$, then $\mathbf{Y}_{i,r} = (Y_{i,1}, Y_{i,4})$ and $\mathbf{Y}_{i,\bar{r}} = (Y_{i,2}, Y_{i,3})$. Similarly, for observation i with random missing pattern $\mathbf{R}_i$, denote $\mathbf{Y}_{i,\mathbf{R}_i}$ and $\mathbf{Y}_{i,\bar{\mathbf{R}}_i}$ as the observed and missing outcome variables for the i th observation, respectively. We place no restrictions on the set of possible patterns $\mathcal{R}$, allowing the pattern to be nonmonotone. Since $|\mathcal{R}| \leq 2^d$, missingness can easily become an exponentially complex problem in the dimension of the outcome variables.

In contrast to the previous section, we will use the term *fully complete* (FC) to refer to a data set containing IID tuples of the form $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{R})$. We call a data set containing IID observations of the form $(\mathbf{X}, \mathbf{Y}_R, \mathbf{R})$ as the *observed* data. Note that the observed data now have two kinds of incompleteness/missingness: incompleteness in the latent class variable (due to mixture models) and missingness in the outcomes. As the case with the LC data, we also do not have access to the FC data. We collect all of the observed outcome variables in one tuple $\mathbf{Y}_{1:n, \mathbf{R}_{1:n}} = (\mathbf{Y}_{1, \mathbf{R}_1}, \mathbf{Y}_{2, \mathbf{R}_2}, \dots, \mathbf{Y}_{n, \mathbf{R}_n})$.

3.1. Missing at random and an imputation strategy

Rubin (1976) outlined three types of missingness mechanisms: missing completely at random (MCAR), MAR, and MNAR. MCAR assumes that the missingness of the variable is independent of all variables in the data. MAR assumes that the missingness of a variable can only depend on observed variables. MNAR assumes that the missingness of a variable can depend on the value of the variable subject to missingness.

In the MCAR data, model fitting is straightforward because one can run Algorithm 1 on the complete cases, and the estimates of the parameters will be consistent. However, MCAR assumes that the missingness is irrelevant to observed outcomes, which is clearly false in the NACC data as a possible reason to miss some test scores is that the individual was too sick to finish the test. Therefore, we consider the MAR assumption. Formally, the definition of MAR is as follows.

Definition 1 (Missing at random). The outcome variables $\mathbf{Y}$ are MAR if the probability of missingness is dependent only on the variables that are observed for the given pattern. This assumption is written as

$$P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y}) \stackrel{\text{MAR}}{=} P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y}_r) \quad (3.1)$$

for all $r \in \mathcal{R}$.

Notice that the left-hand side of (3.1) represents the selection probability $P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y})$. This quantity is strictly unidentifiable because it depends on unobserved data, specifically $\mathbf{Y}_{\bar{r}}$. Thus, it cannot be estimated even given infinite observed data. The MAR assumption makes $P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y})$ identifiable because it equates it to a probability $P(\mathbf{R} = r | \mathbf{X}, \mathbf{Y}_r)$ that can be estimated from the observed data; the variables under the conditioning are all of the variables strictly observed under pattern r . The definition of MAR implies that the probability of a given missing pattern does not depend on variables that are unobserved under that pattern. A key advantage of this assumption is that we avoid the challenge of

modeling the joint distribution between $\mathbf{Y}$ and $\mathbf{R}$ since the missingness mechanism $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y})$ does not need to be modeled directly; this is known as the *ignorability property* (see the discussion later). Thus, we do not have to deal with making potentially unreasonable modeling assumptions on either the selection model $P(\mathbf{R} = r|x, y)$ or the extrapolation distribution $p(y_r|y_r, x, \mathbf{R} = r)$.

The MAR assumption is untestable because its validity relies strictly on data that is unobserved, and so it cannot be rejected by the observed data. Our primary reason for selecting this assumption is for modeling. This may lead to easier interpretability for scientists and practitioners since we fit a global model $p(y|x)$ across all missing patterns rather than a local model $p(y|x, R = r)$ for every pattern r . In the Alzheimer's disease literature, having more statistically sound work can be meaningful because of limitations in existing work that we have described previously in Section 1.3.1. Since we believe that it may be plausible because these tests are correlated due to underlying cognitive ability, we can use it as a starting point. This can be viewed as a baseline before proceeding with more complex MNAR assumptions. We recognize that MNAR may be more reasonable since missing test scores may be attributed to sickness. However, this is in itself a very open research question because MNAR is a broad class of assumptions, and performing mixture modeling with MNAR is not straightforward; we leave this for future work.

An additional feature of the MAR assumption is that this assumption offers a simple approach to impute the data, which makes the computation of the MLE a lot easier. Before describing the procedure of updating model parameters, we first introduce a multiple imputation procedure in Algorithm 2 that can be used to fill in the missing data. In the algorithm, for notational convenience, when we write a binary vector in the summation or product, this means we iterate over all indices whose elements are nonzero. For instance, when we write "For j in $(1, 0, 0, 1)$," this is equivalent to "For $j = 1, 4$." Therefore, if $\mathbf{R}_i = (1, 0, 0, 1)$, then "For j in $\mathbf{R}_i$ " corresponds to "For $j = 1, 4$ " as well.

As stated, Algorithm 2 describes how to impute the outcome variables, assuming that a good estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ is available. We discuss how to actually obtain such an estimate in Section 3.2. The multiple imputation algorithm exploits the fact that the conditional distribution $p(y_r|y_r, x)$ for any $r \in \mathcal{R}$ remains a mixture model. There are two steps to performing multiple imputation: 1) for every observation i with missing observations, we compute the weights of each component of the mixture distribution $p(y_{\mathbf{R}_i}|\mathbf{Y}_{i,\mathbf{R}}, \mathbf{X}_i)$ and 2) sample M times from the distribution $p(y_{\mathbf{R}_i}|\mathbf{Y}_{i,\mathbf{R}}, \mathbf{X}_i)$ for each observation i to form M completed data sets. The derivation of this procedure is provided in Appendix A.2 of the Supplementary Material.

3.2. Model fitting under a missing at random assumption

For each r , we assume that the selection probability $P(\mathbf{R} = r|\mathbf{X}, \mathbf{Y}_r; \gamma_r)$ belongs to a parametric family, indexed by γ_r . We collect these parameters into $\boldsymbol{\gamma} = (\gamma_r)_{r \in \mathcal{R}}$. For simplicity, we write the log-likelihood in terms of the probability model without the n samples. Under the MAR assumption, we can write the observed log-likelihood as

$$\begin{aligned} \ell_{\text{obs}}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\theta}; x, y_r, r) &= \log P(\mathbf{R} = r|x, y_r; \boldsymbol{\gamma}) + \log p(y_r|x; \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &= \ell_{\text{obs}}^{(1)}(\boldsymbol{\gamma}; x, y_r, r) + \ell_{\text{obs}}^{(2)}(\boldsymbol{\beta}, \boldsymbol{\theta}; y_r, x). \end{aligned}$$

The missingness mechanism is said to be ignorable because estimation of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is separated from the estimation of $\boldsymbol{\gamma}$. Model fitting under an MAR assumption typically occurs using an EM algorithm approach. We can proceed with estimating $(\boldsymbol{\beta}, \boldsymbol{\theta})$ by conditioning on the observed data $(\mathbf{X}, \mathbf{Y}_{\mathbf{R}})$ and using another EM algorithm approach. The population $Q_{\text{FC}, r}$ function writes as follows:

$$\begin{aligned} Q_{\text{FC}, r}(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}_r; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) &:= \mathbb{E}[\ell_{\text{LI}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})|\mathbf{X}, \mathbf{Y}_r; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}] \\ &= \sum_{y_r} \ell_{\text{LI}}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}, y_r, \mathbf{Y}_r) p(y_r|\mathbf{Y}_r, \mathbf{X}; \hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}) \end{aligned} \quad (3.2)$$

Algorithm 2 Multiple imputation.

Require: $\{(\mathbf{X}_i, \mathbf{Y}_{i,\mathbf{R}_i}, \mathbf{R}_i)\}_{i=1}^n, (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}), M$ (the number of imputed data sets)

```

1: for  $i = 1, \dots, n$  do
2:   if  $\mathbf{R}_i \neq \mathbf{1}_d$  then
3:     for  $k = 1, \dots, K$  do
4:       Form the weights  $\omega_{i,k} = w_k(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}_k) p_{k, \mathbf{R}_i}(\mathbf{Y}_{i, \mathbf{R}_i}; \widehat{\boldsymbol{\theta}}_k) / \sum_{k'} w_{k'}(\mathbf{X}_i; \widehat{\boldsymbol{\beta}}_{k'}) p_{k', \mathbf{R}_i}(\mathbf{Y}_{i, \mathbf{R}_i}; \widehat{\boldsymbol{\theta}}_{k'})$ 
5:   for  $m = 1, \dots, M$  do
6:     for  $i = 1, \dots, n$  do
7:       if  $\mathbf{R}_i \neq \mathbf{1}_d$  then
8:         Sample  $Z_i^{(m)} \sim \text{Categorical}(K, p = (\omega_{i,1}, \dots, \omega_{i,K}))$ 
9:         for  $j \in \bar{\mathbf{R}}_i$  do
10:          Sample  $\widetilde{Y}_{i,j}^{(m)}$  from  $\text{Binomial}(N_j, \widehat{\theta}_{Z_i^{(m)}, j})$ .
11:          Assign  $\widetilde{\mathbf{Y}}_i^{(m)} := (\mathbf{Y}_{i, \mathbf{R}_i}, \widetilde{\mathbf{Y}}_{i, \bar{\mathbf{R}}_i}^{(m)})$ .
12:   Let  $\widetilde{D}_m = \{(\mathbf{X}_i, \widetilde{\mathbf{Y}}_i^{(m)}, \mathbf{R}_i)\}_{i=1}^n$ .
13: return  $\{\widetilde{D}_m\}_{m=1}^M$ 

```

for every $r \in \mathcal{R}$. One major observation is that the conditional expectation relies on being able to fit estimate $p(y_{\bar{r}}|y_r, x; \boldsymbol{\beta}, \boldsymbol{\theta})$. We can leverage the consequences of the local independence assumption to obtain an easily computable form of the conditional distribution. Unfortunately, since the LI log-likelihood function ℓ_{LI} is not linear in the outcome variables, we are unable to reduce (3.2) to a more simple form. However, we can approximate this expectation stochastically by imputing the missing data $\mathbf{Y}_{\bar{r}}$ for every missing pattern r using the distribution $p(y_{\bar{r}}|y_r, x; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)})$.

To overcome estimating the conditional expectation of factorial terms, we approximate (3.2) stochastically with a Monte Carlo procedure. For large enough M , we expect that

$$Q_{\text{FC}, r}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}_r; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}) \approx Q_{\text{FC}, r, n}^{(M)}(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{X}_{1:n}, \mathbf{Y}_{1:n, \mathbf{R}_{1:n}}; \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)})$$

$$:= \frac{1}{Mn} \sum_{m=1}^M \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k(\mathbf{X}_i; \boldsymbol{\beta}_k) p_k(\widetilde{\mathbf{Y}}_i^{(m;t)}; \boldsymbol{\theta}_k) \right) \cdot \mathbf{1}(\mathbf{R}_i = r),$$

where $\widetilde{\mathbf{Y}}_i^{(m;t)} = (\mathbf{Y}_{i, \mathbf{R}_i}, \widetilde{\mathbf{Y}}_{i, \bar{\mathbf{R}}_i}^{(m;t)})$ denotes the m th imputed data for the i th observation on the iteration step t given the observed variables. The choice of the number of imputations M is important, and in practice, we use $M = 20$ to balance both computational time and good estimation performance. We discuss this in more detail with our simulations in Section 6 and Appendix E.2 of the Supplementary Material.

In the maximization step, we compute the MLE of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ using the completed data after the multiple imputation. Since we have completed data, the MLE can be found using the original EM algorithm, outlined in Algorithm 1. We summarize the procedure in the following algorithm. Throughout this article, we will use the notation $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ (without superscripts relating to t) to denote the point estimate obtained by Algorithm 3 after convergence is achieved. In practice, we assume that convergence is achieved if the L_2 norm of the difference between the current and old parameter estimates falls within a prespecified tolerance level (we choose $\varepsilon = 10^{-4}$).

Algorithm 3 is a nested EM procedure because we have two types of missingness: missingness in the outcome variables and missingness in the latent class labels. We embed an EM algorithm for the LCM fitting (this serves as the M-step in the outer EM algorithm) within an outer EM algorithm for handling the MAR data. Because we do not obtain a closed-form expression for the conditional expectations in the E-step but rather perform a stochastic approximation, Algorithm 3 is a Monte Carlo EM algorithm (Levine & Casella, 2001). We use the notation $\widetilde{\mathbf{Y}}_i^{(m;t)}$ to denote the m th imputed outcome variables using

Algorithm 3 Fitting the mixture model with MAR outcomes via a nested EM procedure.

Require: $\{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{R}_i)\}_{i=1}^n, (\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)}), M$

- 1: Start with an initial guess $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\theta}}^{(0)})$.
 - 2: Initialize the iteration step $t = 0$.
 - 3: **repeat**
 - 4: Let $\tilde{D}_t = \left\{(\mathbf{X}_i, \tilde{\mathbf{Y}}_i^{(m;t)})\right\}_{i=1, \dots, n; m=1, \dots, M}$ be the stacked imputed data sets obtained from Algorithm 2.
 - 5: Obtain $(\hat{\boldsymbol{\beta}}^{(t+1)}, \hat{\boldsymbol{\theta}}^{(t+1)})$ using Algorithm 1 on the imputed data set $\tilde{D}_t$ and starting point $(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$.
 - 6: Update $t \leftarrow t + 1$.
 - 7: **until** Convergence
 - 8: Let $\hat{\boldsymbol{\beta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\beta}}^{(t)}$ and $\hat{\boldsymbol{\theta}} = \lim_{t \rightarrow \infty} \hat{\boldsymbol{\theta}}^{(t)}$.
 - 9: **return** $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$
-

the model parameterized by $(\hat{\boldsymbol{\beta}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)})$. When the parameter is understood, we omit the t index. Note that when there is no missingness in the outcome variables, Algorithm 3 reduces to Algorithm 1 because the multiple imputation step is bypassed. Again, in practice, we run Algorithm 3 with multiple random initial points to ensure we explore the parameter space and converge to the global maximizer. We pick the point estimate that maximizes the observed log-likelihood, defined as

$$\ell_{\text{obs}, n} := \ell_{\text{obs}, n}(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{X}_{1:n}, \mathbf{Y}_{1:n}, \mathbf{R}_{1:n}) := \sum_{i=1}^n \log \left(\sum_{k=1}^K w_k(\mathbf{X}_i) p_{k, \mathbf{R}_i}(\mathbf{Y}_i, \mathbf{R}_i) \right), \quad (3.3)$$

where $p_{k,r}(y_r) = \prod_{j \in r} p_{k,j}(y_j)$ for any $r \in \mathcal{R}$. Note that when there is no missing data, this reduces to the LI log-likelihood function $\ell_{\text{LI}, n}$. Since we are working under a parametric model and estimating parameters using maximum likelihood, the estimators converge at $\sqrt{n}$ -rate and are asymptotically efficient. We provide theoretical justification for these methods and discussion of the asymptotic behavior in Appendix A of the Supplementary Material.

4. Inference

In this section, we discuss how to perform inference on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. There has been some previous work in obtaining asymptotic variance estimators for multiple imputation estimators (Robins & Wang, 2000; Wang & Robins, 1998), but this can be analytically challenging in our setting. The bootstrap is a widely adopted procedure for estimating the sampling distribution of an estimator and obtaining asymptotically valid confidence intervals (Efron, 1979) and has been previously used with the EM algorithm (Celeux & Diebolt, 1987). O'Hagan et al. (2019) have also explored the use of the bootstrap with a Gaussian mixture model. We provide theoretical justification for the bootstrap in our setting in Appendix H.3 of the Supplementary Material.

Bootstrapping works by resampling from the original data set (which is equivalent to sampling from empirical distribution), and with a large enough sample size and under sufficient regularity conditions, this mimics drawing samples from the true generating distribution. We bootstrap the data B times (where B is sufficiently large), including the missing data, and the estimation procedure is run on each of the B bootstrapped data sets.

Since our EM algorithm requires many initialization points in practice to properly explore the parameter space, there is a question of how to initialize the bootstrap procedure. We follow the recommendation outlined in Chen (2022b) to initialize the bootstrap at the same initial point on every

iteration, using the point estimator $(\hat{\beta}, \hat{\theta})$ returned from Algorithm 3. The primary goal of the bootstrap is to measure the stochastic variation of an estimator around the parameter of interest. If we perform the bootstrap with random initialization, we will capture additional uncertainty that arises from estimating different local modes of the log-likelihood function. Initializing at the same point also avoids the label switching identifiability problem, which occurs when the probability distribution remains identical after some parameters are permuted. Additionally, this saves on computational time because we are also not performing many random initializations. Our bootstrap procedure is summarized in Algorithm 4.

Algorithm 4 Bootstrap procedure for obtaining confidence intervals.

Require: $\{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{R}_i, \mathbf{R}_i)\}_{i=1}^n, (\hat{\beta}, \hat{\theta}), B$ (a large number, say 1,000)

- 1: **for** $b \in 1, \dots, B$ **do**
 - 2: Sample n draws uniformly with replacement from $\{1, 2, \dots, n\}$. Put these into index set I_b .
 - 3: Set the b th bootstrapped data set $D_b^* = \{\mathbf{X}_i, \mathbf{Y}_i, \mathbf{R}_i, \mathbf{R}_i\}_{i \in I_b}$.
 - 4: Obtain $(\beta^{*(b)}, \theta^{*(b)})$ using Algorithm 3 on data set D_b^* initialized with $(\hat{\beta}, \hat{\theta})$.
 - 5: **return** $\{\beta^{*(b)}\}_{b=1}^B, \{\theta^{*(b)}\}_{b=1}^B$
-

We now describe how to construct confidence intervals for a given parameter, using $\theta_{1,1}$ as an illustrating example. We first obtain a point estimate $(\hat{\beta}, \hat{\theta})$ using Algorithm 3. Then, we run Algorithm 4 using $(\hat{\beta}, \hat{\theta})$ as the initialization point for a large number of iterations B . We estimate the variance of $\theta_{1,1}$ using the bootstrap samples via the sample variance of $\{\theta_{1,1}^{*(b)}\}_{b=1}^B$

$$\widehat{V} = \frac{1}{B-1} \sum_{b=1}^B \left(\theta_{1,1}^{*(b)} - \frac{1}{B} \sum_{b'=1}^B \theta_{1,1}^{*(b')} \right)^2.$$

We form a 95% confidence interval using a normal approximation

$$(\hat{\theta}_{1,1} - 1.96 \cdot \sqrt{\widehat{V}}, \hat{\theta}_{1,1} + 1.96 \cdot \sqrt{\widehat{V}}).$$

Other approaches, such as the percentile method described in Chen (2022b) will also be equivalent asymptotically since the bootstrap distribution will be asymptotically normal. The key advantage of this method is that it avoids the difficulty of obtaining a closed-form expression for the Fisher information matrix, which can be difficult for complex mixture models. The process of the bootstrap is similar to that of attaining a point estimate because we run Algorithm 3 repeatedly many times. However, unlike obtaining a point estimate, we use the same initialization point on many different data sets (bootstrap samples) rather than many different initialization points on the same data set. White & Murphy (2014) have also discussed how bootstrap methods can struggle when parameters are close to the boundary, so we recommend that post-hoc visualizations are performed to ensure that label-switching has not occurred. In our data analysis and simulations, the mixture components are well-separated, and in practice, we order the groups based on the parameter estimates to avoid the label switching problem.

5. Cluster analysis

A feature of the mixture model that we propose is that we are able to cluster individuals into different groups according to the predictive probability. This allows us to discover latent groups inside our data. We assume that the mixture of binomial product experts model has already been fit successfully using Algorithm 3, and we have an estimator $(\hat{\beta}, \hat{\theta})$.

Each mixture component can be thought of as a cluster, so the method we propose can be used to perform model-based clustering (Bouveyron et al., 2019). To simplify the problem, we first consider the case where we do not have any missing outcome variables. That is, the data we have are only latent and

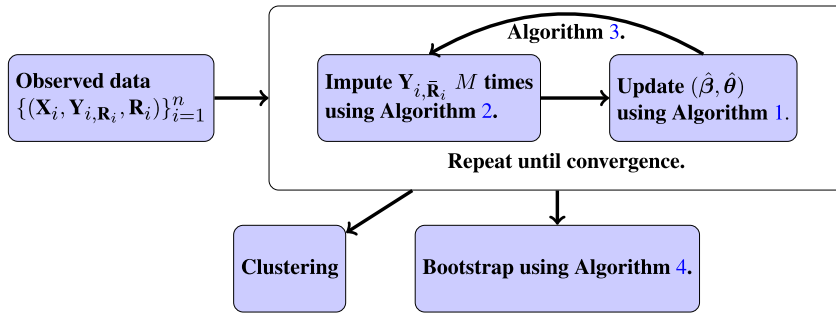


Figure 1. This flowchart describes overall inference procedure.

incomplete. For each observation i , define the probabilities as follows:

$$\hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \frac{w_k(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_k) p_k(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_k)}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_{k'}) p_{k'}(\mathbf{Y}_i; \hat{\boldsymbol{\theta}}_{k'})}. \quad (5.1)$$

The quantity $\hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i)$ is the predictive probability that individual i is from component k . We compute the probability of belonging to each class given the data $\mathbf{X}$ and $\mathbf{Y}$, leading to a probability vector $\hat{\boldsymbol{\pi}}(\mathbf{X}_i, \mathbf{Y}_i) = (\hat{\pi}_1(\mathbf{X}_i, \mathbf{Y}_i), \dots, \hat{\pi}_K(\mathbf{X}_i, \mathbf{Y}_i)) \in \mathbb{R}^K$. It is straightforward to create a cluster assignment of each individual by assigning each observation to the component with the maximum probability. Namely, we assign individual i to the cluster with

$$\hat{C}_i = \underset{k}{\operatorname{argmax}} \hat{\pi}_k(\mathbf{X}_i, \mathbf{Y}_i).$$

In the context of the neuropsychological test scores, we can interpret the latent class as a measure of latent cognitive ability over multiple cognitive domains. An individual's cognitive ability is a complex summary of many different attributes that we hope to measure using the neuropsychological test scores. The latent class can be predicted using the test scores and the baseline covariates.

When the data are missing, we perform clustering using the observed data. For any $r \in \mathcal{R}$, note that there is a closed-form expression for $P(Z = k | \mathbf{X}, \mathbf{Y}_r)$. Define the following quantity:

$$\hat{\pi}_{k,r}(\mathbf{X}_i, \mathbf{Y}_{i,r}) := P(Z = k | \mathbf{X}_i, \mathbf{Y}_{i,r}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \frac{w_k(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_k) p_{k,r}(\mathbf{Y}_{i,r}; \hat{\boldsymbol{\theta}}_k)}{\sum_{k'=1}^K w_{k'}(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_{k'}) p_{k',r}(\mathbf{Y}_{i,r}; \hat{\boldsymbol{\theta}}_{k'})}.$$

For observation i and every $k \in [K]$, we compute $\hat{\pi}_{k, \mathbf{R}_i}(\mathbf{X}_i, \mathbf{Y}_{i, \mathbf{R}_i})$. This quantity can be interpreted as the estimated probability of arising from component k given the data $\mathbf{X}_i$ and $\mathbf{Y}_{i, \mathbf{R}_i}$. When there are missing outcome variables, we compute these probabilities explicitly and use them to perform clustering. We summarize our overall proposed methodology in Figure 1. From the observed data, we run Algorithm 3 (which comprises Algorithms 1 and 2) to obtain a point estimate. We use the point estimate and the observed data in Algorithm 4 to construct confidence intervals using the bootstrap. Cluster assignments can also be obtained using the point estimates.

6. Simulations

We now examine the performance of our method in finite samples using simulated data. We have $p = 2$ covariates, $d = 4$ outcome variables, and $K = 3$ mixtures. Note that $\eta = 2$, $\eta = 2.5$, and $\eta = \infty$ correspond to approximately 75%, 85%, and 100% complete cases. We generate the data through a mixture of binomial product experts model $p(y|x)$, and the parameter η affects the strength of the missingness via the selection model $P(R = r|x, y)$. Further details on the data-generating process and the role of η are described in Appendix E.1 of the Supplementary Material. Each of the simulation studies is evaluated

Table 1. These results are the estimated MSEs and estimated coverage after imputing $M = 20$ times from our model and running Algorithms 3 and 4 for $U = 1,000$ replicates

$MSE_{\theta} (\times 100)$				Estimated coverage of θ s			
	Sample size				Sample size		
	$n = 500$	$n = 1,000$	$n = 2,000$		$n = 500$	$n = 1,000$	$n = 2,000$
$\eta = \infty$	0.203	0.099	0.05	$\eta = \infty$	0.946	0.948	0.950
$\eta = 2.5$	0.192	0.093	0.047	$\eta = 2.5$	0.943	0.943	0.945
$\eta = 2$	0.203	0.104	0.054	$\eta = 2$	0.939	0.939	0.932

$MSE_{\beta} (\times 100)$				Estimated coverage of β s			
	Sample size				Sample size		
	$n = 500$	$n = 1,000$	$n = 2,000$		$n = 500$	$n = 1,000$	$n = 2,000$
$\eta = \infty$	12.7	6.05	3.08	$\eta = \infty$	0.949	0.951	0.947
$\eta = 2.5$	41.5	20.6	10.5	$\eta = 2.5$	0.952	0.951	0.944
$\eta = 2$	40.5	20.0	10.1	$\eta = 2$	0.956	0.951	0.950

Note: The first row contains results for the estimation of θ . The second row contains results for the estimation of β .

using the mean-squared error $MSE_{\theta} = \mathbb{E}[\|\widehat{\theta} - \theta\|_2]$ and $MSE_{\beta} = \mathbb{E}[\|\widehat{\beta} - \beta\|_2]$. For the simulations in Table 1, we estimate the mean-squared error by simulating $U = 1,000$ data sets in each scenario and average the L_2 norm of the error over all the data sets. The estimated MSEs are calculated as follows:

$$\widehat{MSE}_{\beta} = \frac{1}{U} \sum_{i=1}^U \sum_{k \neq 1} \|\widehat{\beta}_k^{(i)} - \beta_k\|_2, \quad \widehat{MSE}_{\theta} = \frac{1}{U} \sum_{i=1}^U \sum_k \|\widehat{\theta}_k^{(i)} - \theta_k\|_2, \quad (6.1)$$

where the superscript (i) denotes the estimated parameter of the i th generated data set. We also construct 95% confidence intervals for each parameter and report the coverage over all the $U = 1,000$ data sets. We check the convergence of Algorithms 1 and 3 by calculating the L_2 norm of the difference between the current and old parameter estimates and comparing against a set tolerance of $\varepsilon = 10^{-4}$. In the first column of Table 1, for a given η , we observe that the estimated MSE decreases at a roughly linear rate as we increase the sample size, which is expected in the parametric setting. This demonstrates that the estimation procedures for both the no missing data (Algorithm 1) and missing data cases (Algorithm 3) are consistent. The MSE for β is higher when there is missingness than when there is no missingness, which suggests that the estimating β becomes more difficult.

In the second column of Table 1, we generally observe nominal coverage as well. There is slight undercoverage of θ when $\eta = 2$. We attribute this to the fact that we do not have a closed-form expression for the MLE and there are many approximations in the algorithms. For example, Algorithm 3 is a stochastic EM algorithm because it relies on a prespecified number of imputations M . Increasing the number of imputations will decrease the Monte Carlo error at the cost of computational time. It is possible that the gradient of the log-likelihood surface can be fairly small in the neighborhood of the true MLE, leading to early stopping of the algorithms and returning a near-solution but not the true MLE. Moreover, when there is missing data, the effective sample size is lower than the number of observations n , which can be another reason for the slight undercoverage in finite samples. For most cases, the coverage is close to nominal, which suggests that the bootstrap procedure and the choice of initializing it at the original estimate returned by Algorithm 3 work in practice.

For an additional comparison, we conduct a second simulation using the same data-generating process. We first perform imputation using the *mice* R package (van Buuren & Groothuis-Oudshoorn, 2011),

Table 2. These results are the estimated MSEs after imputing $M = 20$ times using *mice* for $U = 200$ replicates

$MSE_{\theta}(\times 100)$				$MSE_{\beta}(\times 100)$			
	Sample size				Sample size		
	$n = 500$	$n = 1,000$	$n = 2,000$		$n = 500$	$n = 1,000$	$n = 2,000$
$\eta = 2.5$	0.183	0.096	0.045	$\eta = 2.5$	39.4	19.0	10.2
$\eta = 2$	0.194	0.105	0.045	$\eta = 2$	46.0	19.6	11.6

Note: The left table contains results for the estimation of θ . The right table contains results for the estimation of β .

Table 3. For the simulations in Tables 1 and 2, we report the average computation time in minutes and the standard deviation in parentheses across all the randomly generated data sets

Algorithms 3 and 4 (Point estimate and bootstrap CI)				<i>mice</i> and Algorithm 1 (Point estimate)			
Sample size				Sample size			
	$n = 500$	$n = 1,000$	$n = 2,000$		$n = 500$	$n = 1,000$	$n = 2,000$
$\eta = 2.5$	2.0 (2.1)	2.8 (0.5)	5.5 (1.1)	$\eta = 2.5$	8.1 (0.91)	14.2 (0.83)	28.2 (3.8)
$\eta = 2$	2.4 (1.9)	3.9 (3.5)	6.2 (1.9)	$\eta = 2$	8.1 (0.76)	14.5 (0.89)	27.4 (2.1)

Note: These were performed on a cluster using 48 cores. The first table contains the times for obtaining both a point estimate and a confidence interval using Algorithms 3 and 4. The second table contains the times for only obtaining a point estimate with *mice* and Algorithm 1.

generating $M = 20$ imputed datasets. We then fit our model using Algorithm 1 and compute the MSEs. The results are reported in Table 2.

All simulations in this section were run on a cluster with 48 cores and a parallelized implementation. Our first observation is that this two-step approach—imputing with *mice* followed by Algorithm 1—is substantially slower than Algorithms 3 and 4. As a result, we limit the *mice* experiment to 200 total repetitions for estimating the MSE and do not report coverage results since bootstrap-based interval estimation would be prohibitively time-consuming. For reference, obtaining *both* a point estimate and a bootstrapped confidence interval with our method requires approximately 3–6 times less computation time than obtaining *only* a point estimate after imputing with the *mice* method. This can be seen in Table 3.

Interestingly, we observe that the MSEs from the *mice* imputation and Algorithm 1 pipeline are comparable to those from our method. However, we believe this is largely due to the well-separated mixture components in the data-generating process. In general, we do not recommend using *mice* in this context, as its imputation model may not be compatible with the mixture model being fit, and may not lead to asymptotically valid estimates in more challenging settings. Model uncongeniality is a known problem in the imputation literature (Meng, 1994). In contrast, our method imputes from the same model that is estimated and fit, ensuring consistency.

Further simulations on the complete case data and the influence of the number of imputations are also included in Appendix E.2 of the Supplementary Material. When fitting the model on complete case data, we observe that the MSE does not decrease at a linear rate due to bias, and that the confidence intervals are not asymptotically valid. Moreover, the results appear fairly stable as we vary the number of imputations M across 10, 20, and 50, suggesting that ten imputations are already sufficient. To ensure better accuracy of our estimates while maintaining computational efficiency, we use $M = 20$ imputations in the real data analysis.

7. Application to the NACC data

As mentioned earlier, the data set that motivates our model is from the National Alzheimer’s Coordinating Center (NACC) Uniform Data Set, collected from the years 2005 to 2019. This is a longitudinal data set that comprises individuals of varying cognitive status: cognitively normal to mild cognitive

impairment (MCI) to dementia. Neuropsychological test scores are typically assessed annually with some baseline covariates collected upon entry to the study. However, there is also some missingness present in the outcome variables. The missingness can be due to various reasons. For example, some tests may no longer be administered after a given time because they have been replaced with another one. On the other hand, some individuals may have test scores missing because of a data recording error or they are too sick to take more tests.

7.1. Description of outcome variables and covariates

Our primary goal is to measure the cognitive ability of Alzheimer's disease patients while measuring its association with baseline covariates. Following similar analysis in Brenowitz et al. (2017), we focus on four cognitive domains: long-term memory (episodic memory), attention, language, and executive function. Specifically, we used the Logical Memory Story A immediate and delayed recall to assess memory, Digit Span Forward and Backward tests to evaluate attention, Animal listing test to measure language ability, and Trail Making Tests Parts A and B to measure executive function. Additionally, we used the mini-mental state examination (MMSE) to assess overall cognitive impairment, resulting in a total of eight outcome variables. The Trail A and B tests are assessed based on the time it takes to complete a given task, so a higher score is considered worse. On the other hand, for the other tests, higher scores are indicative of stronger cognitive ability.

We include four baseline covariates in this study: age, education, sex, and race. Age was kept on a yearly scale, and education was dichotomized based on whether an individual had obtained a college degree. Of particular interest was the association between education and cognitive decline. This analysis was motivated by the cognitive reserve hypothesis, which suggests that some mechanisms can provide individuals more resilience to cognitive decline. We would like to explore this hypothesis to see if individuals with higher levels of education are more resistant to cognitive impairment (Meng & D'Arcy, 2012; Thow et al., 2018).

For each individual, we have a Clinical Dementia Rating (CDR) score, which is assigned by a medical professional. The CDR score is based on a clinician's judgment on an individual's cognitive ability and is the standard procedure to determine whether someone has dementia or not. A CDR score 0 refers to cognitively normal, a score 0.5 refers to an MCI state, and a score of at least 1 refers to dementia. Note that the CDR score is all based on a clinician's judgment, which is very different from the neuropsychological tests (exam-based). This score takes values in the set $\{0, 0.5, 1, 2, 3\}$. Here, 0 indicates normal cognitive ability, 0.5 is MCI, 1 is mild dementia, 2 is moderate dementia, and 3 is severe dementia. Thus, we expect that a healthy group would contain individuals with lower CDR scores than an unhealthier group. We will use the CDR score to interpret some of our results, but we emphasize that we *do not* use this variable in the model.

Since there are eight outcome variables, there are up to $2^8 = 256$ missing data patterns. In this data set, we observe 93 missing data patterns with seven patterns comprising over 95% of the data. In particular, there are $n_{cc} = 25,041$ fully observed cases and $n = 41,181$ total observations, so that approximately only 60% of the individuals have completely observed outcomes.

In Figure 2, we plot the CDR distribution for the complete cases, the individuals missing at least one outcome variable, and the entire data set, respectively. If the data were MCAR, we would expect the CDR score distribution to be mostly constant across all missing patterns of the data. However, we observe that for the complete cases, the dementia group (CDR score of 1, 2, or 3) makes up less than 20%, but the dementia group is close to 40% of the data that has at least one missing outcome variable. This suggests that the dementia group is severely underrepresented when we only perform analysis on the complete cases. Thus, from the plots, we can visually postulate the MCAR assumption to be unreasonable.

We select the model using prior knowledge. Since the scientific community previously determined that five levels for the CDR score were appropriate, we also choose the same number. A similar number of groups will help with interpretability, and five is small enough that the number of parameters is manageable.

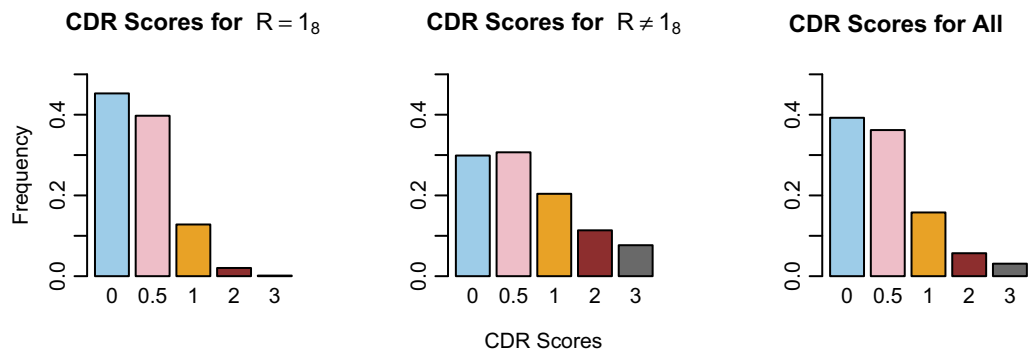


Figure 2. The CDR score distributions for the complete cases, the individuals missing at least one outcome variable, and the entire data set are provided in the left, middle, and right panels, respectively.

Table 4. These tables contain point estimates of the test score means for each latent class

		Outcome variables			
		MMSE (30)	LOGIMEM (25)	MEMUNITS (25)	DIGIF (12)
Class	1	28.9 (0.01)	13.4 (0.04)	12.0 (0.05)	8.8 (0.02)
	2	26.8 (0.03)	9.1 (0.05)	7.1 (0.06)	7.7 (0.02)
	3	25.3 (0.06)	7.6 (0.08)	5.5 (0.08)	7.1 (0.03)
	4	22.5 (0.06)	5.3 (0.05)	3.5 (0.05)	6.6 (0.03)
	5	14.4 (0.12)	3.1 (0.05)	1.8 (0.04)	5.3 (0.04)
		DIGIB (12)	ANIMALS (77)	TRAILA (150)	TRAILB (300)
Class	1	7.2 (0.05)	20.9 (0.05)	28.4 (0.09)	65.9 (0.2)
	2	5.7 (0.02)	15.9 (0.05)	44.2 (0.17)	117.9 (0.43)
	3	4.9 (0.03)	13.5 (0.07)	57.2 (0.39)	209.7 (0.84)
	4	4.2 (0.02)	11.0 (0.06)	66.7 (0.36)	298.5 (0.08)
	5	2.7 (0.03)	6.9 (0.07)	145.3 (0.20)	295.2 (0.09)

Note: In the title of each column, we report the maximum possible score for a given test in parentheses. In each cell, the standard errors are reported in parentheses.

7.2. Point estimates and confidence intervals

The data analysis was conducted on a 2022 Macbook Air with 8 cores, and it took about a day to fit the model and complete the bootstrap with 1,000 bootstrap samples. In Table 4, we report the point estimates of the test score means and the corresponding standard errors in parentheses. There are a few takeaways. As reported, Classes 1 and 5 can be interpreted as the most and least healthy individuals, respectively. Note that in 6 out of 8 tests (omitting TRAIL A and TRAIL B), a higher score corresponds to better performance. On the other hand, TRAIL A and TRAIL B are scored using time-to-completion in seconds, so a higher score corresponds to a worse performance. For every test aside from TRAIL B, we observe a clear monotonic behavior in the test score means from Class 1 to Class 5, which suggests that the model is reasonable and fairly interpretable. For TRAIL B, Classes 4 and 5 have very similar mean test scores that are close to the maximum possible test score of 300. Since a test score of 300 indicates that the individual timed out and did not complete the test, this suggests that individuals from Classes 4 and 5 have comparable performance on the TRAIL B test.

Table 5. This table contains point estimates for the coefficients for each of the covariates and each latent class

		Covariates			
		Age	Education	Race	Sex
Class	2	0.06 (0.001)	−0.62 (0.028)	−0.37 (0.026)	0.72 (0.037)
	3	0.07 (0.002)	−0.9 (0.039)	0.45 (0.034)	0.93 (0.043)
	4	0.06 (0.002)	−1.06 (0.035)	0.36 (0.032)	0.79 (0.04)
	5	0.05 (0.002)	−1.14 (0.0041)	0.41 (0.0035)	0.93 (0.0041)

Note: In each cell, the standard errors are reported in parentheses.

Additionally, the results also suggest the different tests may have value at distinguishing individuals from different latent cognitive classes. For example, TRAIL A and TRAIL B are routinely used to measure executive function. Classes 4 and 5 represent the most unhealthy individuals in the population, but individuals in both perform similarly on TRAIL B. On the other hand, the mean test scores in Classes 4 and 5 for TRAIL A are very different, which suggests that TRAIL A can be a good discriminator between very unhealthy and moderately unhealthy individuals. As TRAIL B is a more complex task than TRAIL A, this seems to agree with our intuition.

In Table 5, we report the point estimates of the coefficients in the logistic regressions. One initial observation is that the coefficients of Age and Education are all positive and negative at the $\alpha = 0.05$ significance level, respectively, across all classes. This implies that increasing age is associated with lower cognitive ability and higher education may provide some protection against dementia. The latter is known as the cognitive reserve hypothesis in the context of education (Meng & D'Arcy, 2012; Thow et al., 2018). We can compare the magnitudes of the coefficient of Education to the coefficient of Age. For example, in Class 2, we observe $0.62/0.06 \approx 10$, and in Class 5, we observe $1.14/0.05 \approx 23$. This suggests that having a college degree or higher education may be equivalent to being 10–20 years younger in terms of cognitive ability.

7.3. Latent classes and clustering

We report two clustering results using models fit on the complete cases (Figure 3) and the entire data (Figure 4). For each set of clustering results, we visualize the clusters by plotting the CDR score composition of each of the five clusters via barplots. Furthermore, we report the mean of the CDR scores for each group, which gives a rough estimate on the group-specific clinical cognitive ability. From the complete case results (CC) in Figure 3, we generally see that the first clusters are primarily composed of healthy individuals as individuals with CDR scores of 0 take up larger proportions. As we progress to the middle latent classes, healthier individuals take up fewer proportions and individuals with MCI (CDR score of 0.5) take up larger proportions. Then, in Group 5, the majority of the individuals have some form of dementia.

For the clustering that used the entire data with the MAR assumption (Figure 4), we observe a similar trend with the healthier individuals, the MCI individuals, and the dementia individuals being the dominant proportion in the earlier, middle, and later latent classes. However, the results from clustering the entire data set are more stark. In contrast with the complete clustering, Groups 4 and 5 contain a significantly higher number of individuals. Moreover, our model is able to detect a group of individuals with high dementia, as evidenced by Group 5's mean CDR score of 1.55. Unlike the complete case clustering, Group 5 has almost 75% of individuals with some form of dementia. Many of these individuals would be omitted from the analysis if the missing data were not accounted for. Thus, accounting for the missingness leads to a clustering result with a stronger correlation with clinical assessments (CDR score).

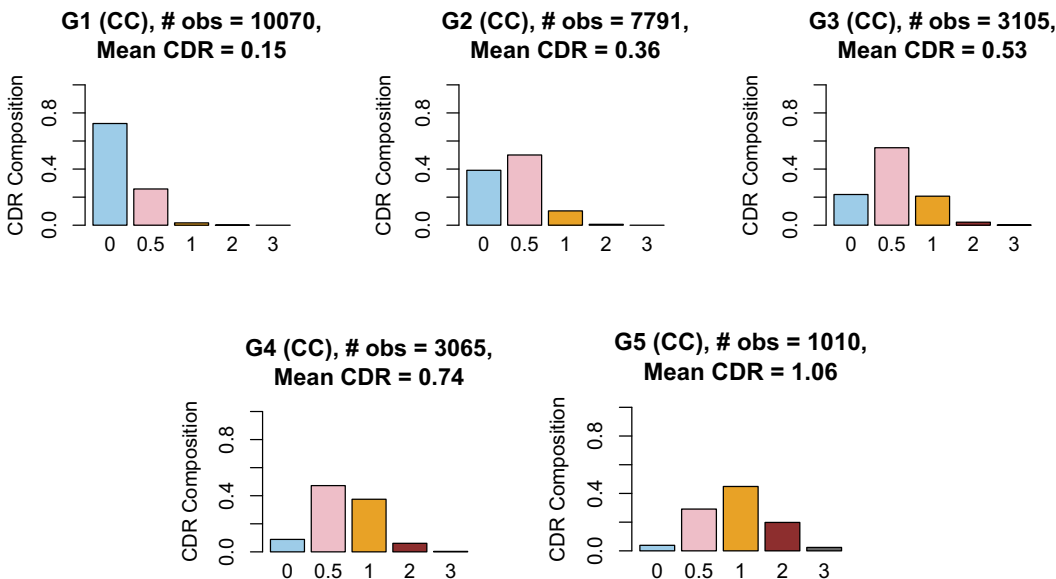


Figure 3. Clustering on complete data only.
Note: These barplots summarize the composition of each of the five latent groups. We order the groups from the most healthy to the least healthy. This is reflected in the mean CDR score of each group.

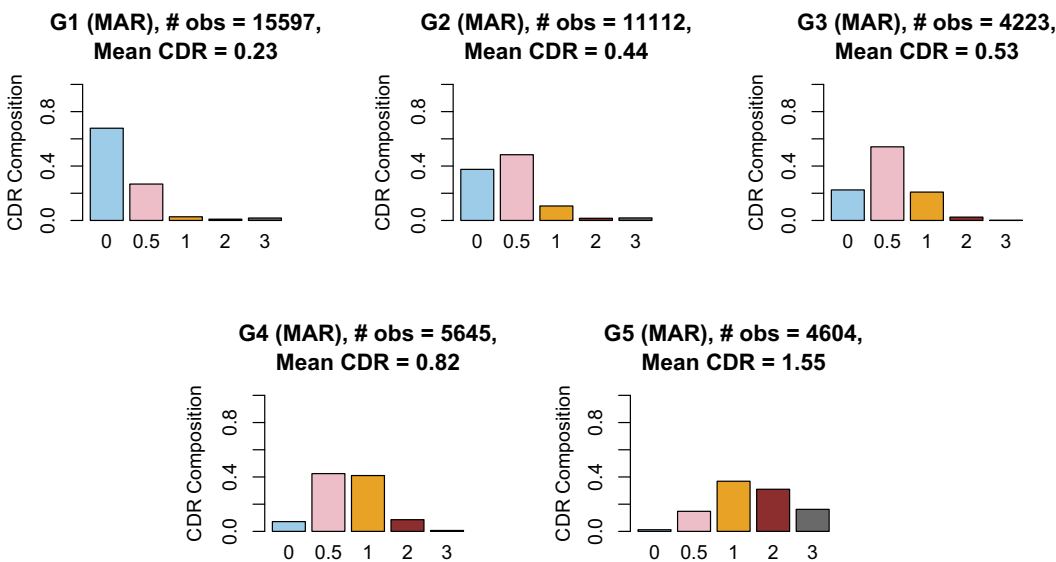


Figure 4. Clustering on the entire data with MAR assumption.
Note: These barplots summarize the composition of each of the five latent groups. We order the groups from the most healthy to the least healthy. This is reflected in the mean CDR score of each group.

8. Discussion

In this article, we have proposed a mixture of binomial experts model for modeling neuropsychological test score data. This model builds on classical ideas from the latent variable and mixture of experts literature. Through this model, we are able to construct a latent representation of an individual's cognitive ability using their test scores and relate it to baseline covariates. Because the Alzheimer's

disease data set is both enriched with multivariate information and plagued with missing data, we address both of these issues. We outline how to perform estimation and inference under the MAR assumption. Further comments on simulations are included in [Appendix E](#) of the Supplementary Material.

There are several avenues for extending this article. Our presented model has two key components: the weights $w_k(x; \beta_k)$ and the component probabilities $p_k(y; \theta_k)$. Due to the local independence assumption, the component probabilities $p_k(y; \theta_k)$ reduce to a product of marginal distributions, which is a binomial product in this article. One straightforward way to generalize this work is modifying these two working models to other parametric families. In our setting, the outcome data are discrete and bounded because we have neuropsychological test scores. However, often, outcome data can come in a variety of different forms such continuous, discrete, and mixed, as well as bounded and unbounded. For instance, in continuous and count data, respectively, the Gaussian and Poisson distributions may be applicable. Access to individual question level data may allow us to leverage ideas from IRT and the Rasch model (Rasch, 1960). Additionally, since there is longitudinal information in the data, there may be a way to incorporate time in the model, so that model-based clustering can be performed on entire trajectories rather than individuals at a given time point. There has been previous work on mixture of experts models applied to time series and longitudinal data (Huerta et al., 2003; Tang & Qu, 2016; Waterhouse et al., 1995), and it would be interesting to extend these frameworks to a model for multivariate longitudinal discrete data. Model selection is also another avenue for future exploration as we selected our model using prior knowledge. We included comments on AIC and BIC in [Appendix F](#) of the Supplementary Material. Also, since we have 33 ADRC centers collecting data, we may introduce a random effect model in the weight to handle the effect from different centers.

The MAR assumption is fundamental to our analysis of the NACC data as well as the inference procedure (Algorithm 3), but it is not the only option. MNAR assumptions encompass a rich class of potentially plausible assumptions. It would be interesting to explore MNAR assumptions in the context of latent variable modeling using methods such as pattern graphs (Chen, 2022a; Cheng et al., 2022; Suen & Chen, 2021). Additionally, perturbing the MAR as a form a sensitivity analysis remains largely an open question. We note that departures from MAR would likely result in a significant modification of the elegant estimation procedure outlined in the article because the ignorability condition would no longer hold. Missing not random assumptions often require stronger modeling assumptions because the relationship between **Y** and **R** needs to be modeled directly. We leave this for future work.

Supplementary material. The supplementary material for this article can be found at <https://doi.org/10.1017/psy.2025.10053>.

Funding statement. D.S. was supported by NSF DGE-2140004. Y.C. was supported by NSF DMS-195278, NSF DMS-2112907, NSF DMS-2141808, and NIH U24-AG072122.

Competing interests. The authors declare none.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Alashwal, H., El Halaby, M., Crouse, J., Abdalla, A., & Moustafa, A. (2019). The application of unsupervised clustering methods to Alzheimer's disease. *Frontiers in Computational Neuroscience*, 13, 31.
- Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A), 3099–3132. <https://doi.org/10.1214/09-AOS689>
- Anderson, M. J., de Valpine, P., Punnett, A., & Miller, A. E. (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution*, 9(6), 3276–3294. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.4948>
- Bishop, C. M., & Svenskn, M. (2002). Bayesian hierarchical mixtures of experts. In U. Kjærulff & C. Meek (Eds.), *Proceedings of the nineteenth conference on uncertainty in artificial intelligence* (pp. 57–64). Morgan Kaufmann Publishers Inc.
- Bock, H. H. (1994). Information and entropy in cluster analysis. In H. Bozdogan, S. L. Sclove, A. K. Gupta, D. Haughton, G. Kitagawa, T. Ozaki, & K. Tanabe (Eds.), *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach: Volume 2 multivariate statistical modeling* (pp. 115–147). Springer.

- Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). *Model-based clustering and classification for data science: With applications in R*. Cambridge University Press.
- Boyle, P. A., Yu, L., Wilson, R. S., Leurgans, S. E., Schneider, J. A., & Bennett, D. A. (2018). Person-specific contribution of neuropathologies to cognitive loss in old age. *Annals of Neurology*, 83(1), 74–83. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25123>
- Brenowitz, W. D., Hubbard, R. A., Keene, C. D., Hawes, S. E., Longstreth, W., Woltjer, R. L., & Kukull, W. A. (2017). Mixed neuropathologies and estimated rates of clinical progression in a large autopsy sample. *Alzheimer's & Dementia*, 13(6), 654–662.
- Brintz, B., Fuentes, C., & Madsen, L. (2018). An asymptotic approximation to the n-mixture model for the estimation of disease prevalence. *Biometrics*, 74(4), 1512–1518.
- Carreira-Perpinan, M. A., & Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1), 141–152.
- Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2, 73–82.
- Celeux, G., & Diebolt, J. (1987). *The EM and SEM algorithms for mixtures: Statistical and numerical aspects* (Tech. Rep. No. RR-0641). INRIA. <https://inria.hal.science/inria-00075912>
- Chen, T., Zhang, N. L., Liu, T., Poon, K. M., & Wang, Y. (2012). Model-based multidimensional clustering of categorical data. *Artificial Intelligence*, 176(1), 2246–2269. <https://www.sciencedirect.com/science/article/pii/S000437021100110X>
- Chen, Y.-C. (2022a). Pattern graphs: A graphical approach to nonmonotone missing data. *The Annals of Statistics*, 50(1), 129–146. <https://doi.org/10.1214/21-AOS2094>
- Chen, Y.-C. (2022b). Statistical Inference with Local Optima. *Journal of the American Statistical Association*, 118(543), 1940–1952. <https://doi.org/10.1080/01621459.2021.2023550>
- Cheng, G., Chen, Y.-C., Smith, M. A., & Zhao, Y.-Q. (2022). Handling nonmonotone missing data with available complete-case missing value assumption. Preprint, [arXiv:2207.02289](https://arxiv.org/abs/2207.02289).
- De Meyer, G., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., De Deyn, P. P., & Initiative, A. D. N. (2010). Diagnosis-independent Alzheimer disease biomarker signature in cognitively normal elderly people. *Archives of Neurology*, 67(8), 949–956.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22. <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x>
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.
- Escudero, J., Zajicek, J. P., & Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using bioprofile concepts and k-means. In *2011 annual international conference of the IEEE engineering in medicine and biology society* (pp. 6470–6473), Institute of Electrical and Electronics Engineers.
- Fruhwirth-Schnatter, S., & Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26, 78–89. <https://EconPapers.repec.org/RePEc:bes:jnlbes:v:26:y:2008:p:78-89>
- Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215–231. <http://www.jstor.org/stable/2334349>
- Gormley, I. C., & Frühwirth-Schnatter, S. (2018). Handbook of mixture analysis. Methodology. <https://api.semanticscholar.org/CorpusID:196182122>
- Gormley, I. C., & Murphy, T. B. (2008). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483), 1014–1027. <http://www.jstor.org/stable/27640140>
- Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4), 2082–2107. <https://doi.org/10.1214/19-AOS1878>
- Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2), 542–548. <http://www.jstor.org/stable/3215044>
- Haines, L. M. (2016). Maximum likelihood estimation for n-mixture models. *Biometrics*, 72(4), 1235–1245. <https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12521>
- Huang, G.-H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1), 5–32.
- Huerta, G., Jiang, W., & Tanner, M. A. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, 13(4), 1097–1118. <http://www.jstor.org/stable/24307162>
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79–87.
- Jordan, M., & Jacobs, R. (1993). Hierarchical mixtures of experts and the EM algorithm. In *Proceedings of 1993 international conference on neural networks (IJCNN-93-Nagoya, Japan)* (Vol. 2, pp. 1339–1344), Institute of Electrical and Electronics Engineers.
- Kang, H. (2013). The prevention and handling of missing data. *Korean Journal of Anesthesiology*, 64(5), 402–406. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/>

- Katzman, R., Aronson, M., Fuld, P., Kawas, C., Brown, T., Morgenstern, H., & Ooi, W. L. (1989). Development of dementing illnesses in an 80-year-old volunteer cohort. *Annals of Neurology*, 25(4), 317–324.
- Kéry, M., Royle, J. A., & Schmid, H. (2005). Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications*, 15(4), 1450–1461. <http://www.jstor.org/stable/4543451>
- Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3), 281–293.
- Kruskal, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18, 95–138.
- Lee, J. W., & Harel, O. (2022). Incomplete clustering analysis via multiple imputation. *Journal of Applied Statistics*, 50(9), 1962–1979.
- Levine, R. A., & Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10(3), 422–439. <http://www.jstor.org/stable/1391097>
- Linzer, D. A., & Lewis, J. B. (2011). polCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10), 1–29. <http://doi.org/10.18637/jss.v042.i10>
- Maslen, B., Popovic, G., Lim, M., Marzinelli, E., & Warton, D. (2023). How many sites? Methods to assist design decisions when collecting multivariate data in ecology. *Methods in Ecology and Evolution*, 14(6), 1564–1573. <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14094>
- Maydeu-Olivares, A., & Liu, Y. (2015). Item diagnostics in multivariate discrete data. *Psychological Methods*, 20(2), 276–292.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and stan*. (2nd ed.) CRC Press. <http://xcelab.net/rm/statistical-rethinking/>
- McParland, D., & Gormley, I. C. (2013). Clustering ordinal data via latent variable models. In B. Lausen, D. Van den Poel, & A. Ultsch (Eds.), *Algorithms from and for nature and life* (pp. 127–135). Springer International Publishing.
- Meng, X., & D'Arcy, C. (2012). Education and dementia in the context of the cognitive reserve hypothesis: A systematic review with meta-analyses and qualitative analyses. *PLoS One*, 7, 1–16.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4), 538–558.
- Mjørud, M., Røsvik, J., Rokstad, A. M., Kirkevold, M., & Engedal, K. (2014). Variables associated with change in quality of life among persons with dementia in nursing homes: A 10 months follow-up study. *PLoS One*, 9(12), e115248.
- O'Hagan, A., Murphy, T. B., Scrucca, L., & Gormley, I. C. (2019). Investigation of parameter uncertainty in clustering using a Gaussian mixture model via jackknife, bootstrap and weighted likelihood bootstrap. *Computational Statistics*, 34(4), 1779–1813. <https://doi.org/10.1007/s00180-019-00897-9>
- Ouyang, J., & Xu, G. (2022). Identifiability of latent class models with covariates. *Psychometrika*, 87(4), 1343–1360. <https://doi.org/10.1007/s11336-022-09852-y>
- Qiu, Y., Jacobs, D. M., Messer, K., Salmon, D. P., & Feldman, H. H. (2019). Cognitive heterogeneity in probable Alzheimer disease: Clinical and neuropathologic features. *Neurology*, 93(8), e778–e790.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Robins, J. M., & Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1), 113–124. <http://www.jstor.org/stable/2673565>
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282. <https://doi.org/10.1177/014662169001400305>
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Serafini, A., Murphy, T. B., & Scrucca, L. (2020). *Handling missing data in model-based clustering*. Preprint.
- Sportisse, A., Marbac, M., Biernacki, C., Boyer, C., Celeux, G., Josse, J., & Laporte, F. (2024). Model-based clustering with missing not at random data. *Statistics and Computing*, 34(4), Article 135, Springer. <https://doi.org/10.1007/s11222-024-10444-2>
- Stern, Y. (2012). Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurology*, 11(11), 1006–1012. [https://doi.org/10.1016/S1474-4422\(12\)70191-6](https://doi.org/10.1016/S1474-4422(12)70191-6)
- Suen, D., & Chen, Y.-C. (2021). *Multistage estimators for missing covariates and incomplete outcomes*. Preprint.
- Tang, X., & Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25(4), 1117–1137. <http://www.jstor.org/stable/44861912>
- Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1), 244–248. <https://doi.org/10.1214/aoms/117705155>
- Thomas, H., Lohaus, A., & Brainerd, C. J. (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development*, 58(9), 1–190. <http://www.jstor.org/stable/1166121>
- Thow, M. E., Summers, M. J., Saunders, N. L., Summers, J. J., Ritchie, K., & Vickers, J. C. (2018). Further education improves cognitive reserve and triggers improvement in selective cognitive functions in older adults: The Tasmanian healthy brain project. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10(1), 22–30.
- Tosto, G., Monsell, S. E., Hawes, S. E., Bruno, G., & Mayeux, R. (2016). Progression of extrapyramidal signs in Alzheimer's disease: Clinical and neuropathological correlates. *Journal of Alzheimer's Disease*, 49(4), 1085–1093.

- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://www.jstatsoft.org/index.php/jss/article/view/v045i03>
- van der Vaart, A. W. (1998). *Asymptotic statistics* (Vol. 3). Cambridge University Press.
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450–469. <http://www.jstor.org/stable/25792024>
- Wang, N., & Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85(4), 935–948. <http://www.jstor.org/stable/2337494>
- Waterhouse, S., MacKay, D., & Robinson, A. (1995). Bayesian methods for mixtures of experts. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8). MIT Press. https://proceedings.neurips.cc/paper_files/paper/1995/file/9da187a7a191431db943a9a5a6fec6f4-Paper.pdf
- White, A., & Murphy, T. B. (2014). BayesLCA: An R package for Bayesian latent class analysis. *Journal of Statistical Software*, 61(13), 1–28. <http://doi.org/10.18637/jss.v061.i13>
- Yesavage, J. A., Taylor, J. L., Friedman, L., Rosenberg, P. B., Lazzeroni, L. C., Leoutsakos, J. S., & Group, C. R. (2016). Principal components analysis of agitation outcomes in Alzheimer's disease. *Journal of Psychiatric Research*, 79, 4–7. <https://doi.org/10.1016/j.jpsychires.2016.04.004>
- Yuksel, S., Wilson, J., & Gader, P. (2012). Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23, 1177–1193.
- Zhang, M., Katzman, R., Salmon, D., Jin, H., Cai, G., Wang, Z., Qu, G., Grant, I., Yu, E., Levy, P., & Klauber, M.R. (1990). The prevalence of dementia and Alzheimer's disease in Shanghai, China: Impact of age, gender, and education. *Annals of Neurology*, 27(4), 428–437.
- Zhang, N. L. (2004). Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5, 697–723.