Published in final edited form as:

Comput Stat Data Anal. 2012 January 1; 56(1): 114–125. doi:10.1016/j.csda.2011.06.014.

Comparison of Methods for Identifying Phenotype Subgroups Using Categorical Features Data With Application to Autism Spectrum Disorder

Mulugeta Gebregziabher^{a,b}, Matthew S Shotwell, Jane M Charles, and Joyce S Nicholas
Department of Medicine, Division of Biostatistics and Epidemiology, 135 Cannon St., Charleston
Suite 303, SC 29425, Fax: 843-876-1126

Abstract

We evaluate the performance of the Dirichlet process mixture (DPM) and the latent class model (LCM) in identifying autism phenotype subgroups based on categorical autism spectrum disorder (ASD) diagnostic features from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision. A simulation study is designed to mimic the diagnostic features in the ASD dataset in order to evaluate the LCM and DPM methods in this context. Likelihood based information criteria and DPM partitioning are used to identify the best fitting models. The Rand statistic is used to compare the performance of the methods in recovering simulated phenotype subgroups. Our results indicate excellent recovery of the simulated subgroup structure for both methods. The LCM performs slightly better than DPM when the correct number of latent subgroups is selected *a priori*. The DPM method utilizes a maximum a posteriori (MAP) criterion to estimate the number of classes, and yielded results in fair agreement with the LCM method. Comparison of model fit indices in identifying the best fitting LCM showed that adjusted Bayesian information criteria (ABIC) picks the correct number of classes over 90% of the time. Thus, when diagnostic features are categorical and there is some prior information regarding the number of latent classes, LCM in conjunction with ABIC is preferred.

Keywords

autism; Dirichlet process mixture; latent class model

© 2011 Elsevier B.V. All rights reserved.

Correspondence to: Mulugeta Gebregziabher.

^agebregz@musc.edu

^bMedical University of South Carolina,

The manuscript represents the views of the authors and not those of the funding agencies.

All authors had access to the data and contributed to the manuscript.

Disclosure: None of the authors disclosed any financial or other conflicts of interest.

Supplementary Materials

Additional tables and figures are in the appendix.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. INTRODUCTION

Autism Spectrum Disorders (ASD) are a group of developmental disabilities characterized by impairments in socialization and communication, as well as by unusual behaviors or interests [1]. There have been some efforts to understand and characterize subgroups within the ASD spectrum [2, 3, 4, 5] in order to assist in earlier identification of cases and to tailor intervention efforts. Since 2000, the South Carolina Autism and Developmental Disabilities Monitoring Network (SC ADDM) has been conducting ongoing, multiple-source, records-based surveillance of ASD [6] within the framework of the Centers for Disease Control and Prevention's (CDC) ADDM [7]. On children with a previous diagnosis of ASD and on those at risk of having ASD, information on early developmental concerns, regression or plateaus in development, problem behaviors, presence of associated behaviors, and previous diagnoses were collected [6]. The collected data represent a heterogeneous collection of autism diagnostic features (behaviors). The goal of this study is to identify phenotype subgroups within the ASD spectrum. The word phenotype is used in the sense of behavioral phenotype [8].

Methods such as hierarchical clustering which are based on dissimilarity of individual characteristics have been proposed to identify subgroups or clusters of observations using characteristics that subjects share in observed data. Latent class models (LCM), Dirichlet process mixtures (DPM) and other mixture models are also used for similar purposes. When the measured characteristics are in continuous or ordinal scale, hierarchical cluster analysis methods [9] as well as mixture model approaches [10, 11, 12] are used. Similarly, when the measured individual characteristics are categorical, LCM [13] or DPM [14] are often used to identify subgroups/clusters [4, 15, 16, 17, 18]. Hierarchical clustering with flexible-beta algorithm has also been adopted for categorical variables [19, 20]. When the variables for clustering are a mixture of continuous and categorical, methods such as converting variables to homogeneous types, analyzing variables separately and finding a weighted average of standardized dissimilarity measures [21] have been suggested. While LCM makes the assumption that the number of clusters is known/fixed *a priori*, DPM does not require such an assumption. Traditionally, model fit indices such as Akaike information Criteria (AIC) or Bayesian information criteria (BIC) have been used to identify LCM. However, there is no consensus on what model fit index identifies the best fitting LCM. The key contribution of this study is that it jointly uses the partitioning from DPM and information criteria to identify phenotype subgroups from categorical features data. In addition, it is not clear whether DPM or LCM has superior performance than the other when the observed variables are categorical; especially, when the categorical variables are characterized by "absent" values (atypical values that do not occur in the majority of the study population) as in our data example.

The objectives of the study four three fold. First, to evaluate and compare LCM and DPM in identifying phenotype subgroups/clusters using simulated data that mimics characteristic variables from the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition Text Revision (DSM-IV-TR). Second, to compare the performance of fit indices such as Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and adjusted BIC (ABIC) in picking the best fitting LCM. Third, to use DPM partitioning to choose optimal number of latent classes for LCM. Fourth, to use data from 806 confirmed and suspected cases of ASD and their diagnostic and associated features to identify phenotype subgroups using both LCM and DPM.

2. DATA EXAMPLE

The ASD data example used in this manuscript consists of de-identified information from the SC ADDM surveillance project on a total of 806 children (413 ASD confirmed and 393 ASD suspected cases) aged 8 years in the surveillance years 2000, 2002, and 2004 (birth years 1992, 1994, and 1996, respectively). A more detailed description of the SC ADDM study can be found in [6]. In brief, the ADDM Network uses a standardized case definition based on the presence of a particular number and quality of behaviors (social, communication, and unusual behaviors) from the DSM-IV-TR for autistic disorder, Asperger disorder, or pervasive developmental disorder not otherwise specified (PDD-NOS). For example, autistic disorder, the most severe end of the spectrum, requires a total of 6 or more of the DSM-IV-TR criteria including at least two social, 1 communication, 1 unusual behavior (see Table 1). Additional associated behaviors are not used in the confirmation of case status but are identified in order to further characterize and describe the children [6, 7]. The DSM-IV-TR criteria and associated features are given in Table 1. Among this sample of children, 598(74%) were males and 208(26%) were females. The racial composition was 324(40%) white, 366(45.5%) black, 12(1.5%) other and 104(12.9%) unknown racial/ethnic groups.

LCM and DPM analyses were applied to the DSM-IV-TR criteria and associated features to identify homogenous subgroups of ASD. Associations between the emergent subgroups and demographic characteristics were assessed via multinomial logistic regression. Odds ratios (ORs) and confidence intervals (CIs) were calculated to evaluate the associations.

3. METHODS

Among the several methods that have been proposed to identify clusters of observations based on characteristics that subjects shared in observed data, we present LCM and DPM. These methods have been developed as methods for discovering subpopulations with similar response profiles and as a flexible way of modeling the correlation of categorical multivariate data [13, 14, 17, 18].

3.1. Latent Class Models (LCM)

Let y denote the entire collection of observed diagnostic features data, where $y_i = \{y_{i1}, \dots, y_{iJ}\}$ is the complete features for subject i for $i = 1, 2, \dots, N$. Let η_k denote the probability of membership in latent class k and let z_i denote the latent class to which subject i belongs, where z_i may take an integer value in $\{1, \dots, K\}$. We assume that a latent class exists in which, if it were observed, it would explain the relationship among the features (y_{ij}). The variable y_{ij} may take an integer value in $\{0, \dots, C_j - 1\}$ where $C_j \geq 2$. Thus, C_j represents the number of possible categories for y_{ij} . Without loss of generality, we assume $C_j = 2$ for all j . Given the latent class, y_{ij} is modeled as a Bernoulli random variable with probability

$$\pi_{jk} = \frac{\exp(\beta_{jk})}{1 + \exp(\beta_{jk})},$$

where β_{jk} is an unknown latent class-specific parameter whose collection is denoted $\beta = \{\beta_{11}, \dots, \beta_{jk}\}$. Due to local independence assumption [13], the features in y_i are independent within each latent class. Thus, the joint probability $\phi(y_i) = \Pr(y_{i1}, \dots, y_{iJ})$ can be expressed as,

$$\varphi(y_i) = \sum_{k=1}^K Pr(k) Pr(y_{i1}, \dots, y_{ij} | k) = \sum_{k=1}^K \eta_k \prod_{j=1}^J (\pi_{jk})^{y_{ij}} (1 - \pi_{jk})^{1-y_{ij}}. \quad (1)$$

The latent class model may be fitted by maximizing the likelihood function

$$L(\eta_k, \pi_{jk}) = \sum_{i=1}^N \log[\varphi(y_i)]$$

with respect to η_k and π_{jk} using the EM algorithm [22]. The EM algorithm involves iterating between posterior probabilities of latent class membership

$$P(z_i = k | y_i) = \frac{\eta_k}{\varphi(y_i)} \prod_{j=1}^J (\pi_{jk})^{y_{ij}} (1 - \pi_{jk})^{1-y_{ij}}.$$

One challenge with LCM is model selection among models with different numbers of latent classes [23]. Currently, there is no consensus in the literature regarding a single statistical index that identifies the most appropriate number of classes in a given population [24]. Thus, we build a series of models with a successive number of classes. Selection of the optimal model utilizes measures of goodness of fit (see section 3.4). Another challenge with fitting latent class models is their vulnerability to converging on local rather than global solutions. Moreover, identifiability and estimability are additional challenges. These challenges may be minimized using multiple random sets of starting values and multiple final stage optimization to avoid local maxima and to ensure that all values converge on identical solutions. If most of the observed likelihood ratio statistics converge to the same value, this can be considered as sufficient evidence for model identifiability [13].

LCM was implemented in SAS via the procedure LCA Version 1.2.4 [25]. A maximum of 500 iterations were considered with 100 different random starting values. Convergence was evaluated using maximum absolute deviation with a criterion of less than 0.000001. In order to handle sparseness issues in the estimation of parameters, optimal Bayesian stabilizing priors are used [13]. The data derived priors applied to both π_{jk} and η_k help to avoid parameter estimates on the boundary values of zero and one. These priors for latent class membership (η_k) as well as item response probabilities (π_{jk}) are decided based on how strongly we want it to influence the parameter estimates of the latent class model. Collins and Lanza (2010) recommend a prior strength of 1 which essentially adds a small amount of pseudo-cases to each class, in order to improve estimation by biasing estimates away from zero. The prior for π_{jk} also acts like adding a small amount of pseudo-cases to each response category for each class, in order to improve estimation by biasing it away from zero or one solutions for some π_{jk} 's. These are similar to Bayes constant in the latent class clustering functionality in LatentGOLD [26].

3.2. Dirichlet Process Mixtures (DPM)

Continuing with the established notation, the collection of latent class membership variables $z = \{z_1, \dots, z_N\}$ represents the partition of subjects into clusters, where the term “cluster” is used synonymously with “latent class”. Hence, z is termed the cluster partition variable. In

hierarchical notation, the augmented Dirichlet process mixture of Bernoulli models is given by,

$$\begin{aligned} y_{ij}|z_i=k &\sim \text{Bernoulli}(\pi_{kj}) \\ \pi_{kj} &\sim \text{Beta}(a, b) \\ P(z) &\propto \alpha^K \prod_{k=1}^K \Gamma(n_k) \end{aligned}$$

where π_{jk} is the probability that a subject in cluster k will exhibit diagnostic feature j . The probability mass function $P(z)$ assigns prior probability over all possible cluster partitions, where α is a fixed tuning parameter, $\Gamma(\cdot)$ is the gamma function, and n_k is the number of subjects assigned to the k^{th} cluster. The associated likelihood and posterior density/mass functions are given by,

$$\begin{aligned} P(y|z, \pi) &= \prod_{k=1}^K \prod_{j=1}^m (\pi_{kj})^{n_{kj}} (1 - \pi_{kj})^{n_k - n_{kj}} \\ P(\pi, z|y) &\propto \alpha^K \prod_{k=1}^K \Gamma(n_k) \prod_{j=1}^m (\pi_{kj})^{a_{kj}-1} (1 - \pi_{kj})^{b_{kj}-1} \\ P(z|y) &\propto \alpha^K \prod_{k=1}^K \Gamma(n_k) \prod_{j=1}^m \frac{\Gamma(a_{kj})\Gamma(b_{kj})}{\Gamma(a_{kj}+b_{kj})} \end{aligned}$$

respectively, where y and π are the collections of diagnostic feature indicators $\{y_{ij}\}$ and their probabilities $\{\pi_{jk}\}$, and n_{kj} is the number of subjects in cluster k that exhibit feature j . The posterior statistics a_{kj} and b_{kj} are given by $a_{kj} = a + n_{kj}$, and $b_{kj} = b + n_k - n_{kj}$.

In order to identify phenotype clusters in these data, the cluster partition variable z was estimated. The estimate for z was taken to be the cluster partition that maximized the marginal posterior mass function $P(z|y)$, or the maximum *a posteriori* (MAP) estimate. The Polya urn gibbs sampler [14, 27] was used to draw samples from the marginal posterior distribution over z which was computed with a uniform (0, 1) prior ($a = b = 1$). The sample value with greatest posterior mass over 50000 iterations was taken as the MAP estimate. The DPM analysis was performed in R Version 2.11.0 with the package `profdpm`.

3.3. Relationship between Class Membership and Covariates

After identifying the phenotype subgroups through the latent class model, association between class membership and covariates can be assessed via multinomial logistic regression [28]. Suppose $x_i = (x_{i1}, \dots, x_{ip})'$ be a vector of covariates for the i th subject. The probability of belonging to the k th latent class ($k = 1, \dots, K$) can be expressed as a function of x_i by,

$$P(z_i = k | x_i) = \frac{\exp(x_i \beta_k)}{1 + \sum_{k=1}^{K-1} \exp(x_i \beta_k)},$$

where K is the reference group and $\beta_k = (\beta_{1k}, \dots, \beta_{pk})'$ is a $p \times 1$ vector of log odds ratio parameters. The log-odds ratio parameters could be estimated jointly with the latent class model parameters or separately after modal assignment of latent classes. Since there is conditional independence between y_i and x_i within a latent class, these log odds ratios can be used to measure the association between phenotype subgroups (latent class) and observed covariates that were not used in the latent class model. In the AS-D data example, this will

help to study the association of autism subgroups with demographic variables such as race and gender.

3.4. Model Assessment and Comparison

Discovering phenotype subgroups using LCM requires proper model selection with the optimal number of subgroups [23, 29]. Several methods have been suggested. For example, statistical indices such as AIC [30], BIC [31], consistent AIC (CAIC) [32] and adjusted BIC (ABIC) [33] and an entropy measure [34]. These statistics account for both model fit (deviance) and model complexity by penalizing models with a larger number of parameters. The criteria are compared across competing models and the lowest value on each criterion indicates the best fitting model. AIC is often used in LCM applications [35]. However, Woodruffe [36] showed that AIC is not theoretically consistent and consequently may not select the correct model when sample size is large. On the other hand, BIC has been shown to have a consistent property [37] that will lead to a correct choice of model for large sample size and is considered to be one of the most reliable indicators in determining the number of latent classes [24]. Nylund et al (2007) made comparisons among several methods of deciding number of latent classes via simulation study. They concluded that BIC performed the best of the ICs and the bootstrap likelihood ratio test proved to be a very consistent indicator of classes across all of the models they considered. Similarly, CAIC tends to favor a model with fewer parameters since it penalizes over-parameterization more severely than BIC or AIC. Another approach based on the posterior class membership probabilities is entropy (R^2). It evaluates how well each of the classes is separated and represented by the data with values ranging from 0 to 1. Values approaching or exceeding 0.80 are preferred. We will make comparisons among these methods in terms of their performance in picking the correct number of classes in LCM via simulation studies.

After partitioning the subjects into clusters via DPM and estimating latent classes via LCM, we will use the method of Rand [38] to compute an index of agreement between the estimated and true cluster partitions by LCM and DPM. The Rand index is the proportion of subject pairs that are consistently clustered in the true (simulated) and estimated cluster partitions and can take values in the interval [0, 1], where one indicates perfect agreement. A pair of subjects that is clustered consistently either occur in the same cluster or in separate clusters, in *both* partitions.

4. SIMULATION STUDY

4.1. Data Generation and Simulation Parameters

Simulated data sets that mimic the ASD data example were used to evaluate LCM and DPM. Preliminary results (from latent class models of $k = 1$ to $k = 5$ classes), indicated the presence of four or five phenotype subgroups in our ASD data example. Latent Class Analysis (LCA) was used to estimate the latent class membership probabilities η_k and latent class-specific probabilities, π_{jk} for each diagnostic feature ($j = 1, \dots, J$ and $k = 4, 5$). The estimated probabilities were used to simulate 100 new datasets for each latent class number ($K = 4, K = 5$). Each new dataset consisted of $N = 806$ subjects whose latent class was drawn according to the estimated latent class probabilities η_k and diagnostic features y_{ij} was drawn according to the Bernoulli probability π_{jk} for $j = 1, \dots, 24$. Hence, the simulated, or 'true' latent class of each subject was available for evaluating the correctness of the estimated latent classes/cluster partition. Information criteria based on CAIC, BIC and ABIC will be used to select the best fitting LCM, while MAP will be used to estimate the classes for DPM. The method of Rand will be used to compute an index of agreement between the estimated and true cluster partitions.

5. RESULTS

5.1. Simulation Results

The Dirichlet process mixture and latent class model methods were used to estimate the cluster partitions. DPM partitions were obtained for each of the 200 ($K = 4$ and $K = 5$) simulated datasets and for each of three values of the precision parameter $\alpha \in \{1, 1/50, 1/500\}$. Smaller values of the precision parameter place additional penalty on partitions with many clusters. Table 2 gives the distribution of estimated number of clusters for each combination of α and K values. With a precision index of $1/50$, the four class model was identified correctly in 97% of the data sets, while the five classes model was correctly identified in 86% of the datasets. When the precision index was reduced to $1/500$, the four classes model was identified correctly in 99% of the datasets while the five classes model was identified in 80% of the datasets.

Figure 1 depicts the statistical indices based on likelihood information criteria and deviance for the LCM estimates for the partitions of the simulated datasets. The AIC, BIC, ABIC and deviance divided by two ($D/2$) are reported. The error bars represent intertertile range of the distribution of these indices across the simulated data sets. The minimum values for the BIC and ABIC correctly identified the true simulated latent classes for both the 4 and 5 class populations. Additionally, the distribution of the estimated number of classes by goodness of fit statistics for the four and five class simulated datasets are depicted in Figure 2. ABIC identified the correct model over 90% of the time for both the 4 and 5 classes models. BIC identified over 90% for the 4 classes model but under 50% for the 5 classes model while AIC was correct in only 50% or less for both models.

Figure 3 illustrates the computed Rand indices between the estimated and true (simulated) cluster partitions for the LCM and DPM analyses of the four and five classes simulated datasets. The LCM estimates performed best in both the four and five classes simulated datasets. However, empirical confidence intervals for the difference in Rand index between the DPM and best performing LCM estimates indicated that the difference was not significant.

5.2. ASD Data Results

Selection of the best fitting model was based on goodness of fit statistics and on whether the model reflected conceptually meaningful and coherent subgroups that account for the heterogeneity in the study population. Model fit indices for LCM under different number of class assumptions are reported in Table 5 (See Appendix). After comparing models with one to seven classes, the best fitting LCM had 4 classes based on BIC and 5 classes based on a combination of BIC and ABIC. Given that the simulation study favored ABIC as the preferred criteria and the fact that DPM also preferred the 5 classes model, we selected the 5 classes model. The minimum entropy R^2 was 0.84 for this model indicating a very good fit (results not tabulated). Each of the LCM models were estimated under 100 different random starting values and they were sufficiently well identified with 100%, 91% and 85% of the likelihood ratio statistics converging to the same value for $K=4$, $K=5$ and $K=3$ respectively.

The best fitting five classes or clusters model is described by the DSM-IV-TR criteria and associated features in Table 3, which shows the estimated probabilities of item response given latent class, $\hat{\pi}_{jk}$. An item was *highly likely* to characterize the cluster if the probability of item response was 0.8 or higher. It was considered as *likely* if $\hat{\pi}_{jk}$ was in $[0.6, 0.8)$. It was considered *chance*, *unlikely* or *very unlikely* if $\hat{\pi}_{jk}$ was in $[0.4, 0.6)$, $[0.2, 0.4)$ and less than 0.2 respectively.

Using the criteria highly likely (≥ 0.8) and likely $[0.6-0.8]$, Cluster 1 (35% of sample) is the *least affected*, with no diagnostic criteria or associated features occurring at a probability ≥ 0.8 , and only the DSM-IV-TR diagnostic communication criteria 2a and 2b in the lower probability range $[0.6-0.8]$. Cluster 3 (6.8% of sample) is *mildly affected*, with some associated features but no diagnostic criteria at a probability ≥ 0.8 . The high probability associated features are abnormality in mood or affect, and hyperactivity/attention problems. DSM-IV-TR diagnostic social interaction (1d) and unusual behavior criteria (3b) occur in the lower probability range $[0.6-0.8]$ along with the associated features aggression and argumentative/oppositional/defiant behaviors. Cluster 4 (*moderately affected*, 21% of sample) is the first cluster with a DSM-IV-TR diagnostic criterion appearing in the probability range ≥ 0.8 . This is the communication criteria 2a, along with the associated features aggression, argumentative/oppositional/defiant behaviors, and hyperactivity/attention problems. Lower probability $[0.6-0.8]$ characteristics are the DSM-IV-TR social interaction (1d) and communication criteria (2b) and the associated features abnormality in mood or affect, delayed motor movement, and temper tantrums. Cluster 2 (*highly affected*, 21% of sample) has an increasing number of DSM-IV-TR diagnostic criteria in the ≥ 0.8 probability range (social interaction 1a, 1d; communication 2a), but picks up criteria from all categories in the $[0.6-0.8]$ probability range (social interaction 1b; communication 2b, 2c; unusual behaviors 3b, 3c, 3d) as well as the associated feature hyperactivity/attention problems. Cluster 5 (16.2% of sample) is the *very highly affected* group, with probability ≥ 0.8 for DSM-IV-TR criteria from all categories (social interaction 1a, 1b, 1d; communication 2a, 2b; unusual behaviors 3b, 3c), as well as the associated features delayed motor movement, hyperactivity/attention problems, and odd response to sensory stimuli. Additionally, all characteristics from Table 1 except for three (self-injurious behavior, lack of fear or excessive fear, and staring spells or seizure-like activity) fall in the lower probability range $[0.6-0.8]$. Additional illustration and confirmation of the level of expression of the DSM criteria and associated behaviors in the 5 classes is given by the heat maps in Figure 4. The clusters from the (*least affected* to the (*highly affected* are labeled from A to E respectively and the number of people assigned to the cluster is proportional to the size of the heat map box.

Examining estimated latent class membership by identified case status (confirmed ASD versus suspected ASD) indicated that the percentage of confirmed cases increases with increasing symptom severity (14.8% in the *least affected*, 30.9% in the *mildly affected*, 41.1% in the *moderately affected*, 91.7% in the *highly affected* and 100% in the *very highly affected* groups). While intellectual disability (ID) was not used as a clustering factor, it is interesting to note the proportions of children with intellectual disability in each cluster. Of the 806 children in the sample, 735 had information on cognitive ability and they were distributed as follows: 47.7% in the *least affected*, 17.0% in the *mildly affected*, 44.2% in the *moderately affected*, 60.6% in the *highly affected* and 57.0% in the *very highly affected*. The corresponding estimates for DPM are provided in the heat map for DPM (see Figure A.5 and additional online tables).

Maximum likelihood estimates of probability of class membership by number of latent classes and their corresponding standard errors are reported in Table 6 (See Appendix). The classification probabilities from DPM were also very similar (see supplementary online files).

Table 4 shows odds ratios (ORs) and accompanying confidence intervals (CIs) for the associations between the emergent subgroups and demographic characteristics from a multinomial logistic regression analysis with cluster 1 (*least affected*) as the reference group. These were obtained using modal assignment of latent classes. The results indicate that there were highly significant positive association between male gender and these phenotypes. For

instance, males were at an almost two fold risk of being in the very highly, highly or mildly affected cluster than females. On the other hand, black race was negatively associated with these phenotypes. Blacks had 80% lower odds to be in the very highly affected subgroup (class 5) than whites. Similarly, blacks were 70%, 80% and 60% less likely to be in the highly (class 2), mildly (class 3) and moderately affected (class 4) groups respectively. However, being in other racial groups (Asians, American Indians, etc) was not associated with these phenotype subgroups.

6. DISCUSSION

In this study we investigated the performance of the Dirichlet process mixture (DPM) and latent class models (LCM) in identifying ASD phenotype subgroups based on DSM-IV-TR criteria and ASD associated features. Our results indicate excellent recovery of the true subgroup structure in the simulated data for both methods. The LCM method performs slightly (though not significantly) better than the DPM method when the correct number of latent subgroups is selected *a priori*. DPM which utilizes a maximum a posteriori (MAP) criterion to estimate the number of classes yielded results that mostly agree with LCM. The simulation results indicate that, adjusted BIC identified the correct latent class model over 90% of the time, while the maximum a posteriori (MAP) estimate gave promising results for the DPM method.

Most cluster analyses of similar datasets in the past have identified a 2 grouping cluster, with one group consisting of lower-functioning, more symptomatic children and a second group of higher-functioning children [3, 2]. But, the analysis of our ASD data resulted in five phenotype subgroups, ranging from *least affected* to *very highly affected* individuals. The *least affected* cluster contained no DSM-IV-TR diagnostic criteria nor associated features at a probability of 80% or more. Diagnostic features appeared across the clusters with increasing probability, until ultimately the *very highly affected* cluster contained DSM-IV-TR diagnostic criteria from each of the social interaction, communication, and unusual behaviors subgroups, as well as multiple associated features with probability of 80% or more.

The key clinical message that stands out from this analysis is that the five phenotype subgroups identified by LCM range from least affected with no diagnostic criteria and no associated features to the very highly affected with diagnostic features from the three core areas that define an ASD: language, social interaction, and repetitive, ritualistic behaviors. The very highly affected subgroup had the greatest number of associated behaviors as well. This spectrum-like model follows the proposed ASD spectrum definition for the DSM-V to be published sometime in the next few years in which the specific diagnoses of “autistic disorder”, “PDD-NOS”, and “Asperger disorder” may no longer be used. This spectrum model of increasing diagnostic criteria and associated features gives the clinician a sense of the child’s prognosis.

Latent class models (LCM) have been applied in identification of sub-populations or phenotype subgroups in several studies. For example, in Foroud et al [15], 13 tests for alcoholic diagnosis were conducted and 830 individuals were surveyed and a 4-latent-class phenotype was chosen. In Pickles et al. [4], using data from 2373 patients on 15 diagnostic tests for autism (each with three categories), 3- and 4-latent-class model were chosen. Similarly, Sullivan et al. [16] had 2914 subjects (twins) whom were assessed by 14 major depression symptoms (all binary) and a 7 latent classes model was selected. Satten et al. [35] also used LCM to select latent classes of the 6 and 12 short loci. LCM is also used to identify obsessive-compulsive disorder subgroups based on observed individual level comorbidity data [39]. Some of these studies used the likelihood ratio [4, 15, 16] and others

used A-IC [35] to select their respective best fitting latent class model. Our results reveal that adjusted BIC is a reasonable fit index to identify the correct model with the appropriate number of classes. This is consistent with results in Lyund et al (2007). In a study comparing the bootstrap likelihood ratio test (BLRT) with some of the commonly used approaches, Lyund et al (2007) concluded that BIC performed the best of the ICs and the bootstrap likelihood ratio test (BLRT) proved to be a very consistent indicator of classes. Looking at their results for the 15 item simple and 10 item complex models, for samples sizes bigger than 500 (similar to our scenario of 24 items and $n=806$), both BIC and adjusted BIC identified the true number of classes more than 99% of the time while BLRT identified about 95%. Moreover, our results are further confirmed by DPM, which does not require specifying *K a priori* since it results in partitions that make up the possible classes.

Other approaches include agglomerative hierarchical cluster analysis, mixture models and latent class mixed models. Such methods have been applied to gene expression data for the detection of subgroups of samples with similar expression profiles across genes or for the detection of subgroups of genes with similar expression profiles across samples. One special strategy of hierarchical cluster analysis which exhibits a good recovery for categorical data which are characterized by many “absent” variables [19] where “absent” refers to behaviors that are atypical and do not occur in the majority of the study subjects is the flexible-beta clustering (FBC). Association between clusters is done using flexible-beta method [20]. On the other hand, association between observations is done using commonly used measures that accept asymmetric binary variables such as Jaccard and Kulczynski coefficients [19, 41]. Within each method, the number of clusters is determined by finding the breaking point below which further branching is ignored [19]. However, this method is not easy to implement and does not have clear model assessment criteria to select the best fit.

Moreover, in agglomerative hierarchical cluster analysis each observation starts as its own cluster and at successive steps of the algorithm, clusters are merged until only a single cluster remains. Although the approach is appealing for visualizing the data, it can be unstable for classification because once an observation is assigned to a cluster it cannot be removed. To deal with this, methods to reallocate already assigned observations have been developed. An alternative approach is to use model-based methods that split the dataset. Such methods can be more stable when the object is to find a few important clusters [21]. There are also other approaches that can be used for similar purposes [11, 10, 43, 42]. A head to head comparison of all these methods could be a topic of further research.

Several measures for partition comparison have been proposed, including the Rand [38], adjusted Rand [44], Fowlkes and Mallows [45], and Wallace [46] indices. Each of these statistics enumerate some proportion of observation pairs that satisfy some condition, such as being clustered together or separately in the two partitions under scrutiny. The adjusted Rand further corrects for the expected number of pairs that meet a criterion. Hence, the adjusted Rand is suitable when the magnitude of the index is the object of inference. When the Rand index is used to compare several estimated partitions against a known partition, then the adjustment is not necessary, and the interpretation is simpler. That is why the (unadjusted) Rand index was selected for partition comparison in this study.

From both the simulation study and the data example, we conclude that jointly using LCM with ABIC could be beneficial in modeling individual differences in which subjects are assumed to belong to one of a finite number of subgroups. It is also important that involving experts with experience and familiarity with prior literature in the subject matter, as we did, play a role in deciding the number of classes.

Acknowledgments

This work is partially supported by funds from MUSC Provost Office. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the CDC.

This study was partially supported by funds from MUSC Provost Office and by grant IIR-06-219 (PI: Egede and Mauldin) funded by the VHA Health Services Research and Development (HSR&D) program.

The SC ADDM project is funded by the Centers for Disease Control and Prevention (CDC). We thank the SC ADDM for allowing us to use the data example. The funding agencies did not participate in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

We thank the Penn State University Methodology Center for allowing us to use their SAS Procedure LCA.

References

1. Myhr G. Autism and other pervasive developmental disorders: exploring the dimensional view. *Canadian Journal of Psychiatry*. 1998; 43:589–595.
2. Belinger LJ, Smith TH. A review of subtyping in autism and proposed dimensional classification mode. *Journal of Autism and Developmental Disorders*. 2001; 31:411–422. [PubMed: 11569587]
3. Stevens MC, Fein DA, Dunn M, Allen D, Waterhouse LH, Feinstein C, Rapin I. Subgroups of children with autism by cluster analysis: a longitudinal examination. *Journal of the American Academy of Child and Adolescent Psychiatry*. 2000; 39:1–6. [PubMed: 10638056]
4. Pickles A, Bolton P, Macdonald H, Bailey A, Le Couteur A, Sim CH, Rutter M. Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Amer J Hum Genet*. 1995; 57:717–726. [PubMed: 7668301]
5. King, LB. Unpublished Dissertation. Medical University of South Carolina; 2009. Autism Spectrum Disorders in South Carolina: Healthcare Utilization, Phenotypes, and Environmental Exposures.
6. Nicholas JS, Charles JM, Carpenter LA, King LB, Jenner W, Spratt EG. Prevalence and characteristics of children with autism spectrum disorders. *Annals of Epidemiology*. 2008; 18:130–136. [PubMed: 18083540]
7. Center for Disease Control and Prevention. Prevalence of autism spectrum disorders - autism and developmental disabilities monitoring network, 14 sites, United States, 2002. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2002 Principal Investigators. *Morbidity and Mortality Weekly Report*. 2007; 56:12–28.
8. O'Brien, G. Behavioural Phenotypes in Clinical Practice. Mac Keith Press; London, UK: 2002.
9. Manly, BFJ. Multivariate Statistical Methods: A primer. 3. Sage Publications; 1990. p. 125-136.
10. McLachlan, GJ.; Peel, D. Finite Mixture Models. John Wiley & Sons, Inc; New York, NY: 2000.
11. McLachlan, GJ.; Basford, KE. Mixture Models: Inference and Applications to Clustering. Marcel Dekker; New York, NY: 1988.
12. Everitt BS. A Monte Carlo investigation of the likelihood ratio test for number of components in a mixture of normal distributions. *Multivariate Behav Res*. 1981; 16:171–180.
13. Collins, LM.; Lanza, ST. Latent class and Latent Transition Analysis. John Wiley & Sons Inc; Hoboken, NJ: 2010.
14. MacEachern SN, Muller P. Estimating Mixture of Dirichlet Process Models. *J Comp and Graph Stat*. 1998; 7:223–238.
15. Foroud T, Bucholz KK, Edenberg HJ, Goate A, Neuman RJ, Porjesz B, Koller DL, Rice J, Reich T, Bierut LJ, Cloninger CR, Nurnberger Jr, Li TK, Conneally PM, Tischfield JA, Crowe R, Hesselbrock V, Schuckit M, Begleiter H. Linkage of an alcoholism-related severity phenotype to chromosome 16. *Alcoholism: Clin Exp Res*. 1998; 22:2035–2042.
16. Sullivan PF, Prescott CA, Kendler KS. The subtypes of major depression in a twin registry. *J Affect Disord*. 2002; 68:273–284. [PubMed: 12063155]
17. Navarro DJ, Griffiths TL, Stetvers M, Lee MD. Modeling individual differences using Dirichlet Process. *J math Psychology*. 2006; 50:101–122.

18. Dunson DB. Bayesian dynamic modeling of latent trait distributions. *Biostatistics*. 2006; 7:551–568. [PubMed: 16488893]
19. Khattree, R.; Naik, DN. *Multivariate Data Reduction and Discrimination with SAS software*. Cary NC: SAS Institute Inc; 2000.
20. Milligan, GW. College of Administrative Science Working Paper Series. Columbus: Ohio State University; 1987. A Study of the Beta-Flexible Clustering Method; p. 8761
21. Kaufman, L.; Rousseeuw, PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc; New York, NY: 1990.
22. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*. 1974; 61:215–231.
23. Clogg, CC. Latent class models. In: Arminger, G.; Clogg, CC.; Sobel, ME., editors. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Plenum Press; New York: 1995. p. 311–359.
24. Nylund KL, Asparouhov T, Muthen BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling*. 2007; 14:535–69.
25. Lanza ST, Collins LM, Lemmon DR, Schafer JL. PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*. 2007; 14:671–694. [PubMed: 19953201]
26. Vermunt, JK.; Magidson, J. *Latent GOLD 4.0 User's Guide*. Statistical Innovations Inc; Belmont, Massachusetts: 2005.
27. Blackwell D, MacQueen JB. Ferguson distribution via poly urn schemes. *Annals of statistics*. 1973; 1:353–355.
28. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*. 1997; 92:1375–86.
29. Collins LM, Fidler PL, Wugalter SE, Long JD. Goodness-of-fit testing for latent class models. *Multivariate Behav Res*. 1993; 28:375–389.
30. Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, BN.; Csaki, F., editors. *Second International Symposium on Information Theory*. Akademiai Kiado; Budapest: 1973. p. 267–281.
31. Schwarz G. Estimating the dimension of a model. *Ann Statist*. 1978; 6:461–464.
32. Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytic extensions. *Psychometrika*. 1987; 52:345–370.
33. Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*. 1987; 52:333–343.
34. Ramaswamy V, DeSarbo W, Reibstein D, Robinson W. An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Sci*. 1993; 12:103–124.
35. Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population in case-control studies of genetic association using a novel latent-class model. *Amer J Hum Genet*. 2001; 68:466–477. [PubMed: 11170894]
36. Woodruffe M. On model selection and the arcsine laws. *Ann Statist*. 1982; 10:1182–1194.
37. Houghton D. On the choice of a model to fit data from an exponential family. *Ann Statist*. 1988; 16:342–355.
38. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*. 1971; 66:846–850.
39. Nestadt G, Addington A, Samuels J, Liang KY, Bienvenu OJ, Riddle M, Grados M, Hoehn-Saric R, Cullen B. The identification of OCD-related subgroups based on comorbidity. *Biological Psychiatry*. 2003; 53:914–920. [PubMed: 12742679]
40. Stephens M. Dealing with label switching in mixture models. *J Roy Statist Soc Ser B*. 2000; 62:795–809.
41. Gower JC. A Comparison of Some Methods of Cluster Analysis. *Biometrics*. 1967; 23:623–637. [PubMed: 6080200]
42. Siegmund KD, Laird PW, Laird-Offringa IA. A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics*. 2004; 20:1896–904. [PubMed: 15044245]

43. McCulloch CE, Lin H, Slate EH, Turnbull BW. Discovering subpopulation structure with latent class mixed models. *Stat Med*. 2002; 21:417–29. [PubMed: 11813228]
44. Hubert L, Arabie P. Comparing partitions. *Journal of Classification*. 1985; 2:193–218.
45. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*. 1983; 78:553–569.
46. Wallace DL. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*. 1983; 78:569–576.

Appendix A. Additional Figures

Figure A.5. Heat Map for the five class DPM. The level of dark color indicates the severity of the classes where (*least affected* is less dark and (*highly affected* is the darkest. The number of people assigned to the cluster is proportional to the size of the heat map box

Appendix B. Additional Tables

Table B.5

Fit Indices for LC Models with different number of classes for the Autism data example (AIC= Akaike Information Criteria, BIC=Bayesian Information Criteria, ABIC=Adjusted BIC).

Model	Fit Statistics for LCM		
Classes (K)	AIC	BIC	ABIC
1	13709	13822	13746
2	11451	11779	11623
3	11106	11453	11218
4	10823	11287	10973
5	10762	11344	10950
6	10692	11390	10971
7	10656	11473	10990

Table B.6
Maximum Likelihood Estimates of Probability of Class Membership and standard error (SE) for Autism data example.

Model	Pr(class membership) estimates with (SE)						
Classes (K)	1	2	3	4	5	6	7
2	0.57(0.021)	0.43(0.021)	NA	NA	NA	NA	NA
3	0.38(0.023)	0.33(0.023)	0.29(0.028)	NA	NA	NA	NA
4	0.35(0.020)	0.17(0.016)	0.27(0.022)	0.22(0.022)	NA	NA	NA
5	0.35(0.020)	0.21(0.021)	0.08(0.016)	0.21(0.021)	0.17(0.016)	NA	NA
6	0.35(0.020)	0.12(0.022)	0.07(0.014)	0.19(0.020)	0.16(0.016)	0.11(0.024)	NA
7	0.35(0.020)	0.12(0.020)	0.07(0.014)	0.19(0.020)	0.11(0.016)	0.11(0.021)	0.06(0.016)

Highlights for COMSTA 5035

- evaluated Dirichlet Process Mixture (DPM) and Latent Class Model (LCM)
- Applied DPM and LCM to autism-spectrum-disorder (ASD) and simulated datasets
- LCM performed slightly better than DPM in recovering simulated subgroups.
- Jointly used partitioning (DPM) and information criteria to identify subgroups.

Figure 1. Information criteria and deviance for the latent class analysis of the four and five class simulated datasets. Y-axis is value of the IC and x-axis is the number of estimated latent classes.

Figure 2.

Distribution of the estimated number of classes by Information criteria for the four and five class simulated datasets. The X axis shows the counts of the estimated number of clusters as a function of the true simulated number of clusters (Y-axis) for different types of information criteria (legend).

Figure 3.

Box and whisker plots of the Rand indices between the estimated and true (simulated) cluster partitions, computed for each LCM and DPM analysis across the four and five class simulated datasets. Higher values indicate better classification.

Figure 4.

Heat Map for the five class LCM. The level of dark color indicates the severity of the classes where (*least affected* is less dark and (*highly affected* is the darkest. The number of people assigned to the cluster is proportional to the size of the heat map box

Table 1

DSM-IV-TR and associated behaviors identified in SC ADDM.

DSM-IV-TR items	Corresponding Behaviors
Social Interaction	
DSM1a	Deficits in non-verbal behavior
DSM1b	Poor peer relationships
DSM1c	Failure to share interests with others
DSM1d	Poor social or emotional reciprocity
Communication	
DSM2a	Delayed spoken language
DSM2b	Conversational deficits
DSM2c	Repetitive language
DSM2d	Deficits in imaginative play/imitation
Unusual behavior	
DSM3a	Restricted Interests
DSM3b	Routines and rituals
DSM3c	Stereotyped mannerisms
DSM3d	Preoccupation with parts of objects
Associated Behaviors	
hyper	Hyperactivity, attention problems
DMM	Delayed Motor Movement
Mood	Abnormality in mood or affect
EDS	Abnormality in Eating or Drinking or Sleeping
tantrum	Temper tantrums
argue	Argumentative, oppositional, defiant
aggress	Aggression
cog	Abnormality in cognitive skills
SS	Odd response to Sensory Stimuli
SIB	Self-Injurious Behavior
fear	Lack of fear or excessive fear
seizure	Staring spells or seizure-like activity

Table 2

Distribution of the estimated number of classes for each combination of α and K values. K is the number of clusters discovered by DPM as a function of the tuning parameter α .

α	K	Estimated Classes ($\hat{K}$)							
		4	5	6	7	8	9		
1	4	41	37	19	2	1	0		
1	5	1	29	34	21	11	4		
1/50	4	97	3	0	0	0	0		
1/50	5	7	86	7	0	0	0		
1/500	4	99	1	0	0	0	0		
1/500	5	20	80	0	0	0	0		

Table 3

Item response probabilities or $Pr(y = 1|k) = \pi_{jk}$ estimates for Autism data example (standard errors are in Appendix table).

DSM-IV-TR Items	Latent Class				
	1	2	3	4	5
DSM1a	0.50	0.81	0.37	0.54	0.95
DSM1b	0.15	0.60	0.45	0.35	0.89
DSM1c	0.27	0.47	0.17	0.25	0.69
DSM1d	0.31	0.82	0.72	0.60	0.95
DSM2a	0.73	0.98	0.16	0.88	0.92
DSM2b	0.75	0.74	0.46	0.79	0.90
DSM2c	0.30	0.68	0.25	0.49	0.70
DSM2d	0.12	0.56	0.11	0.23	0.74
DSM3a	0.05	0.51	0.32	0.14	0.74
DSM3b	0.06	0.74	0.73	0.38	0.97
DSM3c	0.06	0.72	0.19	0.39	0.94
DSM3d	0.03	0.64	0.01	0.11	0.77
Associated Behaviors					
EDS	0.10	0.44	0.26	0.55	0.76
mood	0.21	0.29	0.82	0.69	0.78
cog	0.51	0.35	0.48	0.59	0.63
aggress	0.08	0.30	0.64	0.82	0.71
argue	0.19	0.21	0.75	0.83	0.72
DMM	0.38	0.46	0.25	0.70	0.81
hyper	0.59	0.71	0.88	0.88	0.96
fear	0.04	0.27	0.15	0.15	0.54
SS	0.02	0.41	0.04	0.16	0.86
SIB	0.06	0.22	0.10	0.32	0.58
seizure	0.12	0.16	0.12	0.28	0.39
tantrum	0.12	0.32	0.37	0.69	0.76

Table 4

Odds Ratio and 95% CI for Phenotypes/clusters by gender (1=Male, 0=Female) and race (1=White, 2=Black, 3=Other).

Phenotype	LCM		
	Male	Black	Other
1	1.0(NA)	1.0(NA)	1.0(NA)
2	1.9(1.2,3.2)	0.3(0.2,0.5)	3.6(0.4,33.8)
3	1.8(0.8,3.9)	0.2(0.1,0.4)	2.6(0.2,43.4)
4	1.0(0.7,1.6)	0.4(0.3,0.7)	2.7(0.3,26.9)
5	2.1(1.2,3.6)	0.2(0.1,0.3)	2.7(0.3,26.7)
Phenotype	DPM		
	Male	Black	Other
1	1.0(NA)	1.0(NA)	1.0(NA)
2	2.0(1.2,3.5)	0.2(0.1,0.4)	0.9(0.2,4.5)
3	2.1(0.9,4.6)	0.2(0.1,0.5)	1.0(0.1,9.7)
4	1.3(0.8,2.0)	0.4(0.2,0.6)	0.8(0.2,3.9)
5	2.4(1.4,4.3)	0.2(0.1,0.3)	0.8(0.1,5.1)