Clustering and Variable Selection in the Presence of Mixed Variable Types and Missing Data

C. B. Storlie[†], **S. M. Myers**[‡], **S. K. Katusic**[†], **A. L. Weaver**[†], **R. Voigt**[§], **P. E. Croarkin**[†], **R. E. Stoeckel**[†], and **J. D. Port**[†]

[†]Mayo Clinic

[§]Texas Children's Hospital

[‡]Geisinger Autism & Developmental Medicine Institute

Abstract

We consider the problem of model-based clustering in the presence of many correlated, mixed continuous and discrete variables, some of which may have missing values. Discrete variables are treated with a latent continuous variable approach and the Dirichlet process is used to construct a mixture model with an unknown number of components. Variable selection is also performed to identify the variables that are most influential for determining cluster membership. The work is motivated by the need to cluster patients thought to potentially have autism spectrum disorder (ASD) on the basis of many cognitive and/or behavioral test scores. There are a modest number of patients (486) in the data set along with many (55) test score variables (many of which are discrete valued and/or missing). The goal of the work is to (i) cluster these patients into similar groups to help identify those with similar clinical presentation, and (ii) identify a sparse subset of tests that inform the clusters in order to eliminate unnecessary testing. The proposed approach compares very favorably to other methods via simulation of problems of this type. The results of the ASD analysis suggested three clusters to be most likely, while only four test scores had high (>0.5) posterior probability of being informative. This will result in much more efficient and informative testing. The need to cluster observations on the basis of many correlated, continuous/discrete variables with missing values, is a common problem in the health sciences as well as in many other disciplines.

Keywords

Model-Based Clustering; Dirichlet Process; Missing Data; Hierarchical Bayesian Modeling; Mixed Variable Types; Variable Selection

1 Introduction

Model-based clustering has become a very popular means for unsupervised learning[1–4]. This is due in part to the ability to use the model likelihood to inform, not only the cluster membership, but also the number of clusters M which has been a heavily researched

Corresponding Author: Curtis Storlie, storlie.curt@mayo.edu.

problem for many years. The most widely used model-based approach is the normal mixture model which is not suitable for mixed continuous/discrete variables. For example, this work is motivated by the need to cluster patients thought to potentially have autism spectrum disorder (ASD) on the basis of many correlated test scores. There are a modest number of patients (486) in the data set along with many (55) test score/self-report variables, many of which are discrete valued or have left or right boundaries. Figure 1 provides a look at the data across three of the variables; Beery_standard is discrete valued and ABC_irritability is continuous, but with significant mass at the left boundary of zero. The goals of this problem are to (i) cluster these patients into similar groups to help identify those with similar clinical presentation, and (ii) identify a sparse subset of tests that inform the clusters in an effort to eliminate redundant testing. This problem is also complicated by the fact that many patients in the data have missing test scores. The need to cluster incomplete observations on the basis of many correlated continuous/discrete variables is a common problem in the health sciences as well as in many other disciplines.

When clustering in high dimensions, it becomes critically important to use some form of dimension reduction or variable selection to achieve accurate cluster formation. A common approach to deal with this is a principal components or factor approach[5]. However, such a solution does not address goal (ii) above for the ASD clustering problem. The problem of variable selection in regression or conditional density estimation has been well studied from both the L_1 penalization[6–8] and Bayesian perspectives[9–11]. However, variable selection in clustering is more challenging than that in regression as there is no response to guide (supervise) the selection. Still, there have been several articles considering this topic; see Fop and Murphy[12] for a review. For example, Raftery and Dean[13] propose a partition of the variables into *informative* (dependent on cluster membership even after conditioning on all of the other variables) and *non-informative* (conditionally independent of cluster membership given the values of the other variables). They use BIC to accomplish variable selection with a greedy search which is implemented in the R package `clustvarsel`. Similar approaches are used by Maugis et al.[14] and Fop et al.[15]. An efficient algorithm for identifying the *optimal* set of informative variables is provided by Marbac and Sedki[16] and implemented in the R package `VarSelLCM`. Their approach also allows for mixed data types and missing data, however, it assumes both *local* and *global* independence (i.e., independence of variables within a cluster and unconditional independence of informative and non-informative variables, respectively). The popular LASSO or L1 type penalization has also been applied to shrink cluster means together for variable selection[17–19]. There have also been several approaches developed for sparse K-means and distance based clustering[20–22].

In the Bayesian literature Tadesse et al.[4] consider variable selection in the finite normal mixture model using reversible jump (RJ) Markov chain Monte Carlo (MCMC)[23]. Kim et al.[24] extend that work to the nonparametric Bayesian mixture model via the Dirichlet process model (DPM)[25–28]. The DPM has the advantage of allowing for a countably infinite number of possible components (thus making it nonparametric), while providing a posterior distribution for how many components have been *observed* in the data set at hand. Both Tadesse et al.[4] and Kim et al.[24] use a point mass prior to achieve sparse representation of the informative variables. However, for simplicity they assume all non-informative variables are (unconditionally) independent of the informative variables. This assumption is frequently violated in practice and it is particularly problematic in the case of the ASD analysis as it would force far too many variables to be included into the informative set as is demonstrated later in this paper.

There is not a generally accepted best practice to clustering with mixed discrete and continuous variables. Hunt and Jorgensen[29], Biernacki et al.[30], and Murray and Reiter[31] meld mixtures of *independent* multinomials for the categorical variables and mixtures of Gaussian for the continuous variables. However, it may not be desirable for the dependency between the discrete variables to be entirely represented by mixture components when clustering is the primary objective. As pointed out in Hennig and Liao[32], mixture models can approximate any distribution arbitrarily well so care must be taken to ensure the mixtures fall in line with the goals of clustering. When using mixtures of Gaussian combined with independent multinomials, a data set with many correlated discrete variables will tend to result in more clusters than a comparable dataset with mostly continuous variables. A discrete variable measure of some quantity instead of the continuous version could therefore result in very different clusters. Thus, a Gaussian latent variable approach[33–37] would seem more appropriate for treating discrete variables when clustering is the goal. An observed ordinal variable x_j, for example, is assumed to be the result of thresholding a latent Gaussian variable z_j. For binary variables, this reduces to the multivariate probit model[38,39]. There are also extensions of this approach to allow for unordered categorical variables.

In this paper, we propose a Bayesian nonparametric approach to perform simultaneous estimation of the number of clusters, cluster membership, and variable selection while explicitly accounting for discrete variables and partially observed data. The discrete variables as well as continuous variables with boundaries are treated with a Gaussian latent variable approach. The informative variable construct of Raftery and Dean[13] for normal mixtures is then adopted. However, in order to effectively handle the missing values and account for uncertainty in the variable selection and number of clusters, the proposed model is cast in a fully Bayesian framework via the Dirichlet process. This is then similar to the work of Kim et al.[24], however, they did not consider discrete variables or missing data. Further, a key result of this paper is a solution to allow for dependence between informative and non-informative variables in the nonparametric Bayesian mixture model. Thus, this work overcomes the assumption of (global) independence between informative and non-informative variables. Furthermore, by using the latent variable approach it also overcomes the (local) independence assumption among the informative/clustering variables often assumed when clustering data of mixed type[12].

The solution takes a particularly simple form and also provides an intuitive means with which to define the prior distribution in a manner that decreases prior sensitivity. The component parameters are marginalized out to facilitate more efficient MCMC sampling via a modified version of the split-merge algorithm of Jain and Neal[40]. Finally, missing data is then handled in a principled manner by treating missing values as unknown parameters in the Bayesian framework[41,42]. This approach implicitly assumes a missing at random (MAR) mechanism[43], which implies that the likelihood of a missing value *can* depend on the value of the unobserved variable(s), marginally, just not after conditioning on the observed variables.

The rest of the paper is laid out as follows. Section 2 describes the proposed nonparametric Bayesian approach to clustering observations of mixed discrete and continuous variables with variable selection. Section 3 evaluates the performance of this approach when compared to other methods on several simulation cases. The approach is then applied to the problem for which it was designed in Section 4 where a comprehensive analysis of the ASD problem is presented. Section 5 concludes the paper. This paper also has supplementary material which contains derivations, full exposition of the proposed MCMC algorithm, and MCMC trace plots.

2 Methodology

2.1 Dirichlet Process Mixture Models

As discussed above, the proposed model for clustering uses mixture distributions with a countably infinite number of components via the Dirichlet process prior[25,44,45]. Let $\boldsymbol{y} = (y_1, \ldots, y_p)$ be a p-variate random vector and let $\boldsymbol{y}_i$, $i = 1, \ldots, n$, denote the i^{th} observation of $\boldsymbol{y}$. It is assumed that $\boldsymbol{y}_i$ are independent random vectors coming from distribution $F(\theta_i)$. The model parameters θ_i are assumed to come from a mixing distribution G which has a Dirichlet process prior, i.e., the familiar model,

$$\boldsymbol{y}_i \mid \theta_i \sim F(\theta_i), \quad \theta_i \sim G, \quad G \sim \mathrm{DP}(G_0, \alpha), \quad (1)$$

where DP represents a Dirichlet Process distribution, G_0 is the base distribution and α is a precision parameter, determining the concentration of the prior for G about G_0 [44]. The prior distribution for θ_i in terms of successive conditional distributions is obtained by integrating over G, i.e.,

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{i'=1}^{i-1} \delta(\theta_{i'}) + \frac{\alpha}{i-1+\alpha} G_0, \quad (2)$$

where $\delta(\theta)$ is a point mass distribution at θ. The representation in (2) makes it clear that (1) can be viewed as a countably infinite mixture model. Alternatively, let $\Omega = [\omega_1, \omega_2, \ldots]$ denote the unique values of the θ_i and let ϕ_i be the index for the component to which observation i belongs, i.e., so that $\omega_{\phi_i} = \theta_i$. The following model[26] is equivalent to (2)

$$P(\phi_i = m \mid \phi_1, \ldots, \phi_{i-1}) = \begin{cases} 1 & \text{if } i = 1 \text{ and } m = 1. \\ \frac{n_{i,m}}{i - 1 + \alpha} & \text{if } \phi_{i'} = m \text{ for any } i' < i. \\ \frac{\alpha}{i - 1 + \alpha} & \text{if } m = \max(\phi_1, \ldots, \phi_{i-1}) + 1. \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

with $\boldsymbol{y}_i / \boldsymbol{\phi}_i, \Omega \sim F(\omega_{\phi_i})$, $\omega_m \sim G_0$ and $n_{i,m}$ is the number of $\boldsymbol{\phi}_{i'} = m$ for $i' < i$. Thus, a new observation i is allocated to an existing cluster with probability proportional to the cluster size or it is assigned to a new cluster with probability proportional to α. This is often called the Chinese restaurant representation of the Dirichlet process. It is common to assume that F is a normal distribution in which case $\omega_m = (\boldsymbol{\mu}_m, \Sigma_m)$ describes the mean and covariance of the m^{th} component. This results in a normal mixture model with a countably infinite number of components.

2.2 Discrete Variables and Boundaries/Censoring

Normal mixture models are not effective for clustering when some of the variables are too discretized as demonstrated in Section 3. This is also a problem when the data have left or right boundaries that can be achieved (e.g., several people score the minimum or maximum on a test). However, a Gaussian latent variable approach can be used to circumvent these issues. Suppose that variables y_j for $j \in \mathscr{D}$ are discrete, ordinal variables taking on possible values $\boldsymbol{d}_j = \{d_{j,1}, \ldots, d_{j,L_j}\}$ and that y_j for $j \in \mathscr{C} = \mathscr{D}^c$ are continuous variables with lower and upper limits of b_j and c_j, which could be infinite. Assume for some latent, p-variate, continuous random vector $\boldsymbol{z}$ that

$$y_j = \begin{cases} \sum_{l=1}^{L_j} d_{j,l} I_{\{a_{j,l-1} < z_j \le a_{j,l}\}} & \text{for } j \in \mathscr{D} \\ z_j I_{\{b_j \le z_j \le c_j\}} + b_j I_{\{z_j < b_j\}} + c_j I_{\{z_j > c_j\}} & \text{for } j \in \mathscr{C} \end{cases} \quad (4)$$

where I_A is the indicator function equal to 1 if A and 0 otherwise, $a_{j,0} = -\infty$, $a_{j,L_j} = \infty$, and $a_{j,l} = d_{j,l}$ for $l = 1, \ldots, L_j - 1$. That is, the discrete y_j are the result of thresholding the latent variable z_j on the respective cut-points. The continuous y_j variables are simply equal to the z_j unless the z_j cross the left or right boundary of what can be observed for y_j. That is, if there are finite limits for y_j, then y_j is assumed to be a left and/or right censored version of z_j, thus producing a positive mass at the boundary values of y_j.

A joint mixture model for mixed discrete and continuous variables is then,

$$z_i \mid \phi_i, \Omega \sim N(\boldsymbol{\mu}_{\phi_i}, \boldsymbol{\Sigma}_{\phi_i}), \quad (5)$$

with prior distributions for ω_m and $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]'$ as in (3).

Binary y_j such as gender can be accommodated by setting $\boldsymbol{d}_j = \{0, 1\}$. However, if there is only one cut-point then the model must be restricted for identifiability[39]; namely, if y_j is binary, then we must set $\Sigma_m(j, j) = 1$. The restriction that $\Sigma_m(j, j) = 1$ for binary y_j complicates posterior inference, however, this problem has been relatively well studied in the multinomial probit setting and various proposed solutions exist[46]. It is also straight-forward to use the latent Gaussian variable approach to allow for unordered categorical variables[47,46,48,49], however, inclusion of categorical variables also complicates notation and there are no such variables in the ASD data. For brevity, attention is restricted here to continuous and ordinal discrete variables.

2.3 Variable Selection

Variable selection in clustering problems is more challenging than in regression problems due to the lack of targeted information with which to guide the selection. Using model-based clustering allows a likelihood based approach to model selection, but exactly how the parameter space should be restricted when a variable is "out of the model" requires some care. Raftery and Dean[13] defined a variable y_j to be *non-informative* if conditional on the values of the other variables, it is independent of cluster membership. This implies that a non-informative y_j may still be quite dependent on cluster membership through its dependency with other variables. They assumed a Gaussian mixture distribution for the informative variables, with a conditional Gaussian distribution for the non-informative variables and used maximum likelihood to obtain the change in BIC between candidate models. Thus, they accomplished variable selection with a greedy search to minimize BIC. They further considered restricted covariance parameterizations to reduce the parameter dimensionality (e.g., diagonal, common volume, common shape, common orientation, etc.). We instead take a Bayesian approach to this problem via Stochastic Search Variable Selection (SSVS)[9,50] as this allows for straight-forward treatment of uncertainty in the selected variables and that due to missing values. Kim et al.[24] used such an approach with a DPM for infinite normal mixtures, however, due to the difficulty imposed they did not use the same definition as Raftery and Dean[13] for a non-informative variable. They defined a non-informative variable to be one that is (unconditionally) independent of cluster membership and *all* other variables. This is not reasonable in many cases, particularly in the ASD problem, and can result in negative consequences as seen in Section 3. Below, we layout a more flexible model specification akin to that taken in Raftery and Dean[13] to allow for (global) dependence between informative and non-informative variables in a DPM.

Let the informative variables be represented by the *model* $\boldsymbol{\gamma}$, a vector of binary values such that $\{y_j : \gamma_j = 1\}$ is the set of informative variables. A priori it is assumed that $\Pr(\gamma_j = 1) = \rho_j$. Without loss of generality assume that $\boldsymbol{y}$ has elements ordered such that $\boldsymbol{y} = [\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}]$, with $\boldsymbol{y}^{(1)} = \{y_j : \gamma_j = 1\}$ and $\boldsymbol{y}^{(2)} = \{y_j : \gamma_j = 0\}$, and similarly for $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$. The model in (5) becomes,

$$z_i \mid \boldsymbol{\gamma}, \phi_i, \Omega \sim N(\boldsymbol{\mu}_{\phi_i}, \boldsymbol{\Sigma}_{\phi_i}), \qquad (6)$$

with

$$\boldsymbol{\mu}_m = \begin{pmatrix} \boldsymbol{\mu}_{m1} \\ \boldsymbol{\mu}_{m2} \end{pmatrix}, \quad \boldsymbol{\Sigma}_m = \begin{pmatrix} \boldsymbol{\Sigma}_{m11} & \boldsymbol{\Sigma}_{m12} \\ \boldsymbol{\Sigma}_{m21} & \boldsymbol{\Sigma}_{m22} \end{pmatrix}. \qquad (7)$$

From standard multivariate normal theory, $[\boldsymbol{z}^{(2)} / \boldsymbol{z}^{(1)}, \boldsymbol{\phi} = m] \sim N(\boldsymbol{\mu}_{2/1}, \Sigma_{2/1})$ with $\boldsymbol{\mu}_{2 \mid 1} = \boldsymbol{\mu}_{m2} + \boldsymbol{\Sigma}_{m21}\boldsymbol{\Sigma}_{m11}^{-1}(z^{(1)} - \boldsymbol{\mu}_{m1})$ and $\boldsymbol{\Sigma}_{2 \mid 1} = \boldsymbol{\Sigma}_{m22} - \boldsymbol{\Sigma}_{m21}\boldsymbol{\Sigma}_{m11}^{-1}\boldsymbol{\Sigma}_{m12}$. Now in order for the non-informative variables to follow the definition of Raftery and Dean[13], the $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ must be parameterized so that $\boldsymbol{\mu}_{2/1}, \Sigma_{2/1}$ do not depend on m. In order to accomplish this, it is helpful to make use of the canonical parameterization of the Gaussian[51],

$$z \mid \boldsymbol{\gamma}, \Omega, \phi = m \sim \mathcal{N}_C(\boldsymbol{b}_m, \boldsymbol{Q}_m),$$

with precision $\boldsymbol{Q}_m = \boldsymbol{\Sigma}_m^{-1}$ and $\boldsymbol{b}_m = \boldsymbol{Q}_m\boldsymbol{\mu}_m$. Partition the canonical parameters as,

$$\boldsymbol{b}_m = \begin{pmatrix} \boldsymbol{b}_{m1} \\ \boldsymbol{b}_2 \end{pmatrix}, \quad \boldsymbol{Q}_m = \begin{pmatrix} \boldsymbol{Q}_{m11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix}. \qquad (8)$$

Result 1— *The parameterization in* (*8*) *results in* $(\boldsymbol{\mu}_{2/1}, \Sigma_{2/1})$ *that does not depend on m.*

Proof: The inverse of a partitioned matrix directly implies that $\boldsymbol{\Sigma}_{2 \mid 1} = \boldsymbol{Q}_{22}^{-1}$, which does not depend on m. It also implies that $-\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} = \boldsymbol{\Sigma}_{m21}\boldsymbol{\Sigma}_{m11}^{-1}$, and substituting $\Sigma_m\boldsymbol{b}_m$ for $\boldsymbol{\mu}_m$ in $\boldsymbol{\mu}_{2/1}$ gives $\boldsymbol{\mu}_{2 \mid 1} = \boldsymbol{Q}_{22}^{-1}(\boldsymbol{b}_2 - \boldsymbol{Q}_{21}z^{(1)})$, which also does not depend on m.

The $\boldsymbol{Q}_{21}$ does not depend on m which implies the same dependency structure across the mixture components. This is a necessary assumption in order for $\boldsymbol{z}^{(2)}$ to be non-informative variables, i.e., so that cluster membership conditional on $\boldsymbol{z}^{(1)}$ is independent of $\boldsymbol{z}^{(2)}$.

Now the problem reduces to defining a prior distribution for Ω, i.e., $\omega_m = \{\boldsymbol{b}_m, \boldsymbol{Q}_m\}$, $m = 1, 2, \ldots,$ conditional on the model $\boldsymbol{\gamma}$, that maintains the form of (8). Let $\omega_m^{(1)} = \{\boldsymbol{b}_{m1}, \boldsymbol{Q}_{m11}\}$ and $\omega_m^{(2)} = \omega^{(2)} = \{\boldsymbol{b}_2, \boldsymbol{Q}_{21}, \boldsymbol{Q}_{22}\}$. The prior distribution for Ω will be defined first unconditionally for $\omega^{(2)}$ and then for $\omega_m^{(1)}$, $m = 1, 2, \ldots,$ conditional on $\omega^{(2)}$. There are several considerations in defining these distributions: (i) the resulting $\boldsymbol{Q}_m$ must be positive definite, (ii) it is desirable for the marginal distribution of (μ_m, Σ_m) to remain unchanged for any model $\boldsymbol{\gamma}$ to limit the influence of the prior for ω_m on variable selection, and (iii) it is desirable for them to be conjugate to facilitate MCMC sampling[26,40].

Let Ψ be a $p \times p$ positive definite matrix, partitioned just as $\boldsymbol{Q}_m$, and for a given $\boldsymbol{\gamma}$ assume the following distribution for $\omega^{(2)}$,

$$\begin{aligned}&\boldsymbol{Q}_{22} \sim \mathscr{W}(\boldsymbol{\Psi}_{22\,|\,1}^{-1}, \eta), \quad \boldsymbol{b}_2 \mid \boldsymbol{Q}_{22} \sim \mathscr{N}(\boldsymbol{0}, \tfrac{1}{\lambda}\boldsymbol{Q}_{22}), \\ &\boldsymbol{Q}_{21} \mid \boldsymbol{Q}_{22} \sim \mathscr{M}\mathscr{N}(-\boldsymbol{Q}_{22}\boldsymbol{\Psi}_{21}\boldsymbol{\Psi}_{11}^{-1}, \boldsymbol{Q}_{22}, \boldsymbol{\Psi}_{11}^{-1}),\end{aligned} \qquad (9)$$

where $\mathscr{W}$ denotes the Wishart distribution, and $\mathfrak{M}\mathscr{N}$ denotes the matrix normal distribution.

The distribution of $\omega_m^{(1)}$, conditional on $\omega^{(2)}$ is defined implicitly below. A prior distribution is *not* placed on ($\boldsymbol{b}_{m1}$, $\boldsymbol{Q}_{m11}$), directly. It is helpful to reparameterize from ($\boldsymbol{b}_2$, $\boldsymbol{Q}_{22}$, $\boldsymbol{Q}_{21}$, $\boldsymbol{b}_{m1}$, $\boldsymbol{Q}_{m11}$) to ($\boldsymbol{b}_2$, $\boldsymbol{Q}_{22}$, $\boldsymbol{Q}_{21}$, $\boldsymbol{\mu}_{m1}$, Σ_{m11}). By doing this, independent priors can be placed on ($\boldsymbol{b}_2$, $\boldsymbol{Q}_{22}$, $\boldsymbol{Q}_{21}$) and ($\boldsymbol{\mu}_{m1}$, Σ_{m11}) and still maintain all of the desired properties as will be seen in Results 2 and 3.

The prior distribution of ($\boldsymbol{\mu}_{m1}$, Σ_{m11}) is

$$\boldsymbol{\Sigma}_{m11} \overset{iid}{\sim} \mathscr{W}^{-1}(\boldsymbol{\Psi}_{11}, \eta - p_2), \quad \boldsymbol{\mu}_{m1} \mid \boldsymbol{\Sigma}_{m11} \overset{ind}{\sim} \mathscr{N}(\boldsymbol{0}, \tfrac{1}{\lambda}\boldsymbol{\Sigma}_{m11}), \qquad (10)$$

where $\mathscr{W}^{-1}$ denotes the inverse-Wishart distribution and ($\boldsymbol{\mu}_{m1}$, Σ_{m11}) are independent of $\omega^{(2)}$. The resulting distribution of ($\boldsymbol{b}_{m1}$, $\boldsymbol{Q}_{m11}$) conditional on ($\boldsymbol{b}_2$, $\boldsymbol{Q}_{22}$, $\boldsymbol{Q}_{21}$) is not a common or named distribution, but it is well defined via the relations, $\boldsymbol{b}_{m1} = \boldsymbol{\Sigma}_{m11}^{-1}\boldsymbol{\mu}_{m1} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{b}_2$, and $\boldsymbol{Q}_{m11} = \boldsymbol{\Sigma}_{m11}^{-1} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$.

Result 2—*The prior distribution defined in* (9) *and* (10) *results in a marginal distribution for* ($\boldsymbol{\mu}_m$, Σ_m) *of* $\mathscr{N}\mathscr{I}\mathscr{W}$ ($\boldsymbol{0}$, λ, Ψ, η)*,i.e.,the same normal-inverse-Wishart regardless of* $\boldsymbol{\gamma}$.

<u>Proof:</u> It follows from Theorem 3 of Bodnar and Okhrin[52] that $\Sigma_m \sim \mathscr{I}\mathscr{W}$ (η, Ψ). It remains to show $\boldsymbol{\mu}_m / \Sigma_m \sim \mathscr{N}$ ($\boldsymbol{0}$, $(1/\lambda)$ Σ_m). However, according to (9) and (10) and the independence assumption,

$$\left.\begin{pmatrix}\boldsymbol{\mu}_{m1}\\ \boldsymbol{b}_2\end{pmatrix}\right| \boldsymbol{\Sigma}_m \sim \mathscr{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \frac{1}{\lambda}\begin{pmatrix}\boldsymbol{\Sigma}_{m11} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Q}_{22}\end{pmatrix}\right).$$

Also, $\boldsymbol{b}_m = \boldsymbol{Q}_m\boldsymbol{\mu}_m$ implies,

$$\begin{pmatrix} \boldsymbol{\mu}_{m1} \\ \boldsymbol{\mu}_{m2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} & \boldsymbol{Q}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_{m1} \\ \boldsymbol{b}_{2} \end{pmatrix}.$$

Using the relation $\boldsymbol{Ax} \sim \mathcal{N}(\boldsymbol{A\mu}, \boldsymbol{A}\Sigma\boldsymbol{A}')$ for $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ gives the desired result.

As mentioned above, the normal-inverse-Wishart distribution is conjugate for ω_m in the unrestricted (no variable selection) setting. It turns out that the distribution defined in (9) and (10) is conjugate for the parameterization in (8) as well, so that the component parameters can be integrated out of the likelihood. Let the (latent) observations be denoted as $\boldsymbol{Z} = [z_1', \ldots, z_n']'$, and the data likelihood as $f(\boldsymbol{Z} / \boldsymbol{\gamma}, \boldsymbol{\phi}, \Omega)$.

Result 3—*The marginal likelihood of* $\boldsymbol{Z}$ *is given by*

$$f(\boldsymbol{Z} \mid \gamma, \boldsymbol{\phi}) = \int f(\boldsymbol{Z} \mid \gamma, \boldsymbol{\phi}, \Omega) f(\Omega \mid \gamma) d\Omega$$

$$= \pi^{-\frac{np}{2}} \prod_{m=1}^{M} \left[\left(\frac{\lambda}{n_m+\lambda}\right)^{\frac{p_1}{2}} \frac{|\boldsymbol{\Psi}_{11}|^{\frac{\eta-p_2}{2}} \Gamma_{p_1}\left(\frac{n_m+\eta-p_2}{2}\right)}{|\boldsymbol{V}_{m11}|^{\frac{n_m+\eta-p_2}{2}} \Gamma_{p_1}\left(\frac{\eta-p_2}{2}\right)} \right] \left[\left(\frac{\lambda}{n+\lambda}\right)^{\frac{p_2}{2}} \frac{|\boldsymbol{\Psi}_{11}|^{\frac{p_2}{2}} |\boldsymbol{\Psi}_{2\mid 1}|^{\frac{\eta}{2}} \Gamma_{p_2}\left(\frac{n+\eta}{2}\right)}{|\boldsymbol{V}_{11}|^{\frac{p_2}{2}} |\boldsymbol{V}_{2\mid 1}|^{\frac{n+\eta}{2}} \Gamma_{p_2}\left(\frac{\eta}{2}\right)} \right],$$

where (i) $M = \max(\boldsymbol{\phi})$*, i.e., the number of observed components, (ii)* $p_1 = \Sigma \boldsymbol{\gamma}_j$ *is the number of informative variables, (iii)* $p_2 = p - p_1$*, (iv)* n_m *is the number of* $\boldsymbol{\phi}_i = m$*, (v)* $\Gamma_p(\cdot)$ *is the multivariate gamma function, and (vi)* $\boldsymbol{V}_{m11}$*,* $\boldsymbol{V}_{11}$*,* $\boldsymbol{V}_{2/1}$ *are defined as,*

$$\boldsymbol{V}_{m11} = \sum_{\phi_i = m} (z_i^{(1)} - \bar{z}_{m1})(z_i^{(1)} - \bar{z}_{m1})' + \frac{n_m \lambda}{n_m + \lambda} \bar{z}_{m1} \bar{z}_{m1}' + \boldsymbol{\Psi}_{11},$$

$$\boldsymbol{V}_{11} = \sum_{i=1}^{n} (z_i^{(1)} - \bar{z}_1)(z_i^{(1)} - \bar{z}_1)' + \frac{n\lambda}{n+\lambda} \bar{z}_1 \bar{z}_1' + \boldsymbol{\Psi}_{11},$$

$$\boldsymbol{V}_{2\mid 1} = \boldsymbol{V}_{22} - \boldsymbol{V}_{21} \boldsymbol{V}_{11}^{-1} \boldsymbol{V}_{21}',$$

with $\bar{z}_{m1} = \frac{1}{n_m} \sum_{\phi_i = m} z_i^{(1)}$, $\bar{z}_1 = \frac{1}{n} \sum_{i=1}^{n} z_i^{(1)}$, $\bar{z}_2 = \frac{1}{n} \sum_{i=1}^{n} z_i^{(2)}$,

$$\boldsymbol{V}_{22} = \sum_{i=1}^{n} (z_i^{(2)} - \bar{z}_2)(z_i^{(2)} - \bar{z}_2)' + \frac{n\lambda}{n+\lambda}\bar{z}_2\bar{z}_2' + \boldsymbol{\Psi}_{22}, \textit{ and} \qquad (11)$$
$$\boldsymbol{V}_{21} = \sum_{i=1}^{n} (z_i^{(2)} - \bar{z}_2)(z_i^{(1)} - \bar{z}_1)' + \frac{n\lambda}{n+\lambda}\bar{z}_2\bar{z}_1' + \boldsymbol{\Psi}_{21}.$$

The derivation of Result 3 is provided in Web Appendix B.

2.4 Hyper-Prior Distributions

Kim et al.[24] found there to be a lot of prior sensitivity due to the choice of prior for the component parameters. This is in part due to the separate prior specification for the parameters corresponding to informative and non-informative variables, respectively. The specification above treats all component parameters collectively, in a single prior, so that the choice will not be sensitive to the interplay between the priors chosen for informative and non-informative variables. A further stabilization can be obtained by rationale similar to that used in Raftery and Dean[13] for restricted forms of the covariance (such as equal shape, orientation, etc.). We do not enforce such restrictions exactly, but one might expect the components to have similar covariances or similar means for some of the components. Thus it makes sense to put hierarchical priors on λ, $\boldsymbol{\Psi}$, and η, to encourage such similarity if warranted by the data. A Gamma prior is also placed on the concentration parameter α, i.e.,

$$\begin{gathered}\lambda \sim \mathrm{Gamma}(A_\lambda, B_\lambda), \;\; \boldsymbol{\Psi} \sim \mathscr{W}(\boldsymbol{P}, N), \\ \eta - (p+1) \sim \mathrm{Gamma}(A_\eta, B_\eta), \;\; \alpha \sim \mathrm{Gamma}(A_\alpha, B_\alpha).\end{gathered} \qquad (12)$$

In the analyses below, relatively vague priors were used with $A_\lambda = B_\lambda = A_\eta = B_\eta = 2$. The prior for α was set to $A_\alpha = 2$, $B_\alpha = 2$, to encourage anywhere from 1 to 15 clusters from 100 observations. The results still have some sensitivity to the choice of $\boldsymbol{P}$. In addition, there are some drawbacks to Wishart priors which can be exaggerated when applied to variables of differing scale[53,54]. In order to alleviate these issues, we recommend first standardizing the columns of the data to mean zero and unit variance, then using $N = p + 2$, $\boldsymbol{P} = (1/N)\boldsymbol{I}$. Finally, the prior probability for variable inclusion was set to ρ_j =0.5 for all j. The data model in (4) and (6), the component prior distribution in (9) and (10), along with the hyper-priors in (12), completes the model specification.

2.5 MCMC Algorithm

Complete MCMC details are provided in the Web Appendix C. However, an overview is provided here to illustrate the main idea. The complete list of parameters to be sampled in the MCMC are $\Theta = \{\boldsymbol{\gamma}, \boldsymbol{\phi}, \lambda, \eta, \boldsymbol{\Psi}, \alpha, \tilde{\boldsymbol{Z}}\}$, where $\tilde{\boldsymbol{Z}}$ contains any latent element of $\boldsymbol{Z}$ (i.e., either corresponding to missing data, discrete variable, or boundary/censored observation). The only update that depends on the raw observed data $\boldsymbol{Y} = [\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n']'$ is the update of $\tilde{\boldsymbol{Z}}$. All other parameters, when conditioned on $\boldsymbol{Y}$ and $\boldsymbol{Z}$, only depend on $\boldsymbol{Z}$. The $\tilde{\boldsymbol{Z}}$ are block updated, each with a MH step, but with a proposal that looks almost conjugate, and is therefore accepted with high probability; the block size can be adjusted to trade-off between acceptance and speed (e.g., acceptance $\sim$40%). A similar strategy is taken with the $\boldsymbol{\Psi}$ update, i.e., a nearly conjugate update is proposed and accepted/rejected via an MH step. Because the component parameters are integrated out, the ϕ_i can be updated with simple Gibbs sampling[26], however, this approach has known mixing issues[40,55]. Thus, a modified split-merge algorithm[40] similar to that used in[24] was developed to sample from the posterior distribution of $\boldsymbol{\phi}$. The remaining parameters are updated in a hybrid Gibbs, Metropolis Hastings (MH) fashion. The $\boldsymbol{\gamma}$ vector is updated with MH by proposing an add, delete, or swap move[50]. The λ, η, α parameters have standard MH random walk updates on log-scale. The MCMC routine then consists of applying each of the above updates in turn to complete a single MCMC iteration, with the exception that the $\boldsymbol{\gamma}$ update be applied L_g times each iteration.

Two modifications were also made to the above strategy to improve mixing. The algorithm above would at times have trouble breaking away from local modes when proposing $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ updates separately. Thus, an additional joint update is proposed for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ each iteration which substantially improved the chance of a move each iteration. Also, as described in more detail in Web Appendix C, the traditional split merge algorithm proposes an update by first selecting two points, i and i', at random. If they are from the same cluster (according to the current $\boldsymbol{\phi}$) it then assigns them to separate clusters and assigns the remaining points from that cluster to each of the two new clusters at random. It then conducts several (L) restricted (to one of the two clusters) Gibbs sampling updates to the remaining ϕ_h from the original cluster. The resulting $\boldsymbol{\phi}^*$ then becomes the proposal for a split move. We found that the following adjustment resulted in better acceptance of split/merge moves. Instead of assigning the remaining points to the two clusters at random, simply assign them to the closest of the two observations i or i'. Then conduct L restricted Gibbs sample updates to produce the proposal. We found little performance gain beyond $L = 3$. Lastly, it would be possible to instead use a finite mixture approximation via the kernel stick breaking representation of a DPM[56,55]. However, this approach would be complicated by the dependency between $\boldsymbol{\gamma}$ and the structure and dimensionality of the component parameters. This issue is entirely avoided with the proposed approach as the component parameters are integrated out. The code to perform the MCMC for this model has been made available in a GitHub repository at https://github.com/cbstorlie/DPM-vs.git.

2.6 Inference for $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$

The estimated cluster membership $\hat{\boldsymbol{\phi}}$ for all of the methods was taken to be the respective mode of the estimated cluster membership probabilities. For the DPM methods, the cluster membership probability matrix P (which is an $n \times \infty$ matrix in principle) is not sampled in the MCMC, and is not identified due to many symmetric modes (thus their can be label switching in the posterior samples). However, the information theoretic approach of Stephens[57] (applied to the DPM in Fu et al.[58]) can be used to address this issue and relabel the posterior samples of $\boldsymbol{\phi}$ to provide an estimate of P. The resulting estimate $\hat{P}$ has i^{th} row, m^{th} column that can be thought of as the proportion of the relabeled posterior samples of ϕ_i that have the value m. While technically P is an $n \times \infty$ matrix, all columns after M^* have zero entries in $\hat{P}$, where M^* is the maximum number of clusters observed in the posterior. For the results below, the point estimate of $\hat{\boldsymbol{\gamma}}$ is determined by $\hat{\gamma}_j = 1$ if $\Pr(\gamma_j = 1) > \rho_j = 0.5$, and $\hat{\gamma}_j = 0$ otherwise.

3 Simulation Results

In this section the performance of the proposed approach for clustering is evaluated on two simulation cases similar in nature to the ASD clustering problem. Each of the cases is examined (i) without missing data or discrete variables/censoring, (ii) with missing data, but no discrete variables/censoring, (iii) with missing data and several discrete and/or censored variables.

The approaches to be compared are listed below.

DPM-vs – the proposed method.

DPM-cont – the proposed method without accounting for discrete variables/censoring (i.e., assuming all continuous variables).

DPM – the proposed method with variable selection turned off (i.e., a prior probability $\rho_j = 1$).

DPM-ind – the approach of Kim et al.[24] when all variables are continuous (i.e., assuming non-informative variables are independent of the rest), but modified to treat discrete variables/censoring and missing data when applicable just as the proposed approach.

Mclust-vs – the approach of Raftery and Dean[13] implemented with the `clustvarsel` package in R. When there are missing data, Random Forest Imputation[59] implemented with the `missForest` package in R is used prior to application of `clustvarsel`. However, the Mclust-vs approach does not treat discrete variables differently and thus treats all variables as continuous and uncensored.

VarSelLCM – the approach of Marbac and Sedki[16] implemented in the R package `VarSelLCM`. It allows for mixed data types and missing data, however, it assumes both *local* independence of variables within cluster and *global* independence between informative and non-informative variables.

Each simulation case is described below. Figure 2 provides a graphical depiction of the problem for the first eight variables from the first of the 100 realizations of Case 2(c). Case 1 simulations resulted in very similar data patterns as well.

Case 1(a) – $n = 150$, $p = 10$. The true model has $M = 3$ components with mixing proportions 0.5, 0.25, 0.25, respectively, and $\boldsymbol{y} \mid \boldsymbol{\phi}$ is a multivariate normal with no censoring nor missing data. Only two variables $\boldsymbol{y}^{(1)} = [y_1, y_2]'$ are informative, with means of $(2, 0)$, $(0, 2)$, $(-1.5, -1.5)$, unit variances, and correlations of 0.5, 0.5, −0.5 in each component, respectively. The non-informative variables $\boldsymbol{y}^{(2)} = [y_3, \ldots, y_{10}]'$ are generated as $iid\,\mathcal{N}(0, 1)$.

Case 1(b) – Same as the setup in 1(a) only the non-informative variables $\boldsymbol{y}^{(2)}$ are correlated with $\boldsymbol{y}^{(1)}$ through the relation $\boldsymbol{y}^{(2)} = \boldsymbol{B}\boldsymbol{y}^{(1)} + \boldsymbol{\varepsilon}$, where $\boldsymbol{B}$ is a 8×2 matrix whose elements are distributed as $iid\,\mathcal{N}(0, 0.3)$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, Q_{22}^{-1})$, with $\boldsymbol{Q}_{22} \sim \mathscr{W}(\boldsymbol{I}, 10)$.

Case 1(c) – Same as in 1(b), but variables y_1, y_6 are discretized to the closest integer, variables y_2, y_9 are left censored at −1.4 (~8% of the observations), and y_3, y_{10} are right censored at 1.4.

Case 1(d) – Same as 1(c), but the even numbered y_j have ~30% of the observations MAR.

Case 2(a) – $n = 300$, $p = 30$. The true model has $M = 3$ components with mixing proportions 0.5, 0.25, 0.25, respectively, $\boldsymbol{y}/\boldsymbol{\phi}$ is a multivariate normal with no censoring nor missing data. Only four variables (y_1, y_2, y_3, y_4) are informative, with means of $(0.6, 0, 1.2, 0)$, $(0, 1.5, -0.6, 1.9)$, $(-2, -2, 0, 0.6)$ and all variables with unit variance for each of the three components, respectively. All correlations among informative variables are equal to 0.5 in components 1 and 2, while component 3 has correlation matrix, $\boldsymbol{\Sigma}_{311}(i, j) =$

$0.5(-1)^{|i+j|} I_{\{i \neq j\}} + I_{\{i=j\}}$. The non-informative variables $\boldsymbol{y}^{(2)} = [y_5, \ldots, y_{30}]'$ are generated as *iid* $\mathcal{N}(0, 1)$.

Case 2(b) – Same as the setup in Case 2(a) only the non-informative variables $\boldsymbol{y}^{(2)}$ are correlated with $\boldsymbol{y}^{(1)}$ through the relation $\boldsymbol{y}^{(2)} = \boldsymbol{B}\boldsymbol{y}^{(1)} + \boldsymbol{\varepsilon}$, where $\boldsymbol{B}$ is a 26×4 matrix whose elements are distributed as *iid* $\mathcal{N}(0, 0.3)$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, Q_{22}^{-1})$, with $\boldsymbol{Q}_{22} \sim \mathcal{W}(\boldsymbol{I}, 30)$.

Case 2(c) – Same setup as in Case 2(b), but now variables y_1, y_6, y_{11} are discretized to the closest integer, variables y_2, y_9, y_{10}, y_{11} are left censored at −1.4 (~8% of the observations), and variables $y_3, y_{12}, y_{13}, y_{14}$ are right censored at 1.4.

Case 2(d) – Same as Case 2(c), but the even numbered y_j have ~30% MAR.

For each of the eight simulation cases, 100 data sets were randomly generated and each of the five methods above was fit to each data set. The methods are compared on the basis of the following statistics.

Acc – Accuracy calculated as the proportion of observations in the estimated clusters that are in the same group as they are in the true clusters, when put in the arrangement (relabeling) that best matches the true clusters.

FI – Fowlkes-Mallows index of $\hat{\boldsymbol{\phi}}$ relative to the true clusters.

ARI – Adjusted Rand index.

M – The number of estimated clusters. The estimated number of clusters for the Mclust-vs and VarSelLCM methods was chosen as the best of the possible $M = 1, \ldots 8$ cluster models via BIC. The number of clusters for the Bayesian methods is chosen as the posterior mode and is inherently allowed to be as large as *n*.

p_1 – The model size, $p_1 = \Sigma_j \hat{\gamma}_j$.

PVC – The proportion of variables correctly included/excluded from the model, $PVC = (1/p)\Sigma_j I_{\{\hat{\gamma}_j = \gamma_j\}}$.

CompT – The computation time in minutes (using 20,000 MCMC iterations for the Bayesian methods).

These measures are summarized in the columns of the tables below by their mean (and standard deviation) over the 100 data sets. It appeared that 20,000 iterations (10,000 burn in and 10,000 posterior samples) was sufficient for the Bayesian methods to summarize the posterior in the simulation cases via several trial runs, however not every simulation result was inspected for convergence.

The simulation results from Cases 1(a)–(d) are summarized in Table 1. The summary score for the *best* method for each summary is in bold along with that for any other method that was not statistically different from the *best* method on the basis of the 100 trials (via an uncorrected paired *t*-test with $\alpha = 0.05$). As would be expected, DPM-ind is one of the best methods on Case 1(a), however, it is not significantly better than DPM-vs or Mclust-vs on any of the metrics. VarSelLCM performs slightly worse than the top three methods in this case since the local independence assumption is being violated. All of the other methods solidly outperform DPM though, which had a difficult time finding more than a single cluster since it had to include all 10 variables. In Case 1(b) the assumptions of DPM-ind are now being violated and it is unable to perform adequate variable selection. It must include far too many non-informative variables due to the correlation within $\boldsymbol{y}^{(2)}$ and between $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$. The clustering performance suffers as a result and like DPM, it has difficulty finding more than a single cluster. VarSelLCM also struggles in this case for the same reason; the global independence assumption is being violated. Mclust-vs still performs well in this case, but DPM-vs (and DPM-cont) is significantly better on two of the metrics. In case 1(c) DPM-vs is now explicitly accounting for the discrete and left/right censored variables, while DPM-cont does not. When the discrete variables are incorrectly assumed to be continuous it tends to create separate clusters at some of the unique values of the discrete variables. This is because a very high likelihood can be obtained by normal distributions that are almost singular along the direction of the discrete variables. Thus, DPM-vs substantially outperforms DPM-cont and Mclust-vs, demonstrating the importance of explicitly treating the discrete nature of the data when clustering. Finally, Case 1(d) shows that the loss of 30% of the data for half of the variables (including an informative variable) does not degrade the performance of DPM-vs by much. In this case Mclust-vs uses Random Forest Imputation to first impute the data, then cluster. The imputation procedure does not explicitly take into account of the cluster structure of the data, rather it could mask this structure. This is another reason that the performance is worse than the proposed approach which incorporates the missingness directly into the clustering model. Mclust-vs and VarSelLCM both have *much* faster run-times than the Bayesian methods, however, when there are local or global correlations and discrete variables and/or missing data, they did not perform nearly as well as DPM-vs.

The simulation results from Cases 2(a)–(d) are summarized in Table 2. A similar story line carries over into Case 2 where there are now $p = 30$ (four informative) variables and $n = 300$ observations. Namely, DPM-vs is not significantly different from DPM-ind or Mclust-vs on any of the summary measures for Case 2(a), with the exception of computation time. DPM-vs is the best method on all summary statistics (except *CompT*) by a sizeable margin on the remaining cases. While Mclust is much faster than DPM-vs, the cases of the most interest in this paper are those with discrete variables, censoring and/or missing data (i.e., Cases 1(c), 1(d), 2(c), and 2(d)). In these cases, the additional computation time of DPM-vs might seem inconsequential relative to the enormous gain in accuracy. It is interesting that DPM-vs suffers far less from the missing values when moving from Case 2(c) to 2(d) than it did from Case 1(c) to 1(d). This is likely due to the fact that there are a larger number of observations to offset the additional complexity of a larger p. However, it is also likely that the additional (correlated) variables may help to reduce the posterior variance of the *imputed* values.

4 Application to Autism and Related Disorders

The cohort for this study consists of subjects falling in the criteria for “potential ASD” (PASD) on the basis of various combinations of developmental and psychiatric diagnoses obtained from comprehensive medical and educational records as described in Katusic et al.[60]. The population of individuals with PASD is important because this group represents the pool of patients with developmental/behavioral symptoms from which clinicians have to determine who has ASD and/or other disorders. Subjects 18 years of age or older were invited to participate in a face-to-face session to complete psychometrist-administered assessments of autism symptoms, cognition/intelligence, memory/learning, speech and language, adaptive functions, and maladaptive behavior. In addition, guardians were asked to complete several self-reported, validated questionnaires. The goal is to describe how the patients’ test scores separate them in terms of clinical presentation and which test scores are the most useful for this purpose. This falls in line with the new Research Domain Criteria (RDoC) philosophy that has gained traction in the field of mental health research. RDoC is a new research framework for studying mental disorders. It aims to integrate many levels of information (cognitive/self-report tests, imaging, genetics) to understand how all of these might be related to similar clinical presentations.

A total of 87 test scores measuring cognitive and/or behavioral characteristics were considered from a broad list of commonly used tests for assessing such disorders. A complete list of the individual tests considered is provided in Web Appendix A. Using expert judgment to include several commonly used aggregates in place of individual subtest scores, this list was reduced to 55 test score variables to be considered in the clustering procedure. Five of the 55 variables have fewer than 15 possible values and are treated here as discrete, ordinal variables. A majority (46) of the 55 variables also have a lower bound, which is attained by a significant portion of the individuals, and are treated as left censored. Five of the variables have an upper bound that is attained by many of the individuals and are thus treated as right censored. There are 486 observations (individuals) in the dataset, however, only 67 individuals have complete data, i.e., a complete case analysis would throw out 86% of the observations.

DPM-vs was applied to these data; four chains with random starting points were run in parallel for 85,000 iterations each, which took ~40 hours on a 2.2GHz processor. The first 10,000 iterations were discarded as burn-in. More iterations were used here than in the simulation cases due to the fact that this analysis is slightly more complicated (e.g., more variables and observations) and it only needed to be performed once. MCMC trace plots are provided in Web Appendix D. All chains converged to the same distribution (aside from relabeling) and were thus combined.Four of the tests (Beery standard, CompTsc_ol, WJ_Pass_Comprehen, and Adaptive Composite) had a high (>0.88) posterior probability of being informative (Table 3). There is also evidence that Ach_abc_Attention and Ach_abc_AnxDep are informative. The posterior samples were split on which of these two should be included in the model (they were only informative together for 0.1% of the MCMC samples). The next highest posterior inclusion probability for any of the remaining variables was 0.17 and the sum of the inclusion probabilities for all remaining variables was only 0.28. Thus, there is strong evidence to suggest that only five of the 55 variables are sufficient to inform the cluster membership.

A majority (54%) of the posterior samples identified three components/clusters, with 0.12 and 0.25 posterior probability of two and four clusters, respectively. The calculation of $\hat{\phi}$ also resulted in three components. Figure 3 displays the estimated cluster membership via pairwise scatterplots of the five most informative variables on a standardized scale. Ach_abc_Attention has also been multiplied by minus one so that higher values imply better functioning for *all* tests. The corresponding mean vectors of the three main components are also provided in Table 3. There are two groups that are very distinct (i.e., Clusters 1 and 2 are the “high” and “low” groups, respectively), but there is also a “middle” group (Cluster 3). Cluster 3 subjects generally have medium-to-high Adaptive_Composite, WJ_Pass_Comprehen, and Ach_abc_Attention scores, but low-to-medium Beery_standard and WJ_Pass_Comprehen.

Figure 4(a) provides a 3D scatter plot on the three most informative variables, highlighting separation between Cluster 1 and Clusters 2 and 3. However, Clusters 2 and 3 are not well differentiated in this plot. Figure 4(b) shows a 3D scatter plot on the variables CompTsc_ol, WJ_Pass_Comprehen, and Ach_Attention, illustrating differentiation between Clusters 2 and 3.

The goal of this work is not necessarily to identify clusters that align with clinical diagnosis of ASD, i.e., it is not a classification problem. The Research Domain Criteria (RDoC) philosophy is to get away from subjective based diagnosis of disease. The hope is that these clusters provide groups of similar patients that may have similar underlying physiological causes and can be treated similarly (whether the clinical diagnosis was ASD or not). That being said, the “high” cluster aligned with *no* clinical diagnosis of ASD for 92% of its subjects, while the “low” cluster aligned with positive clinical ASD diagnosis for 50% of its subjects. As these clusters result from a bottom-up, data-driven method, they may prove useful to determine imaging biomarkers that correspond better with cluster assignment than a more subjective diagnosis provided by a physician. This will be the subject of future work.

5 Conclusions & Further Work

In this paper we developed a general approach to clustering via a Dirichlet process model that explicitly allows for discrete and censored variables via a latent variable approach, and missing data. This approach overcomes the assumption of (global) independence between informative and non-informative variables and the assumption of (local) independence of variables within cluster often assumed when clustering data of mixed type. The MCMC computation proceeds via a split/merge algorithm by integrating out the component parameters. This approach was shown to perform markedly better than other approaches on several simulated test cases. The approach was developed for moderate p in the range of ~10–300. The computation is $\mathcal{O}(p^3)$, which makes it ill-suited for extremely large dimensions. However, it may be possible to use a graphical model[61,62] within the proposed framework to alleviate this burden for large p.

The approach was used to analyze test scores of individuals with potential ASD and identified three clusters. Further, it was determined that only five of the 55 variables were informative to assess the cluster membership of an observation. This could have a large impact for diagnosis of ASD as there are currently ~100 tests/subtest scores that could be used, and there is no universal standard. Further, the clustering results have served to generate hypotheses about what might show up in brain imaging to explain some of the differences between potential ASD patients. A follow-up study has been planned to investigate these possible connections.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by Public Health Service research grants MH093522 andAG034676 from the National Institutes of Health.

References

1. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association. 2002; 97:611–631.
2. Basu SanjibChib Siddhartha. Marginal likelihood and bayes factors for dirichlet process mixture models. Journal of the American Statistical Association. 2003; 98(461):224–235.
3. Quintana Fernando A, Iglesias Pilar L. Bayesian clustering and product partition models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2003; 65(2):557–574.
4. Tadesse Mahlet G, Sha NaijunVannucci Marina. Bayesian variable selection in clustering high-dimensional data. Journal of the American Statistical Association. 2005; 100(470):602–617.
5. Liu JS, Zhang JL, Palumbo MJ, Lawrence CE. Bayesian clustering with variable and transformation selections. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, editorsBayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting. Oxford University Press; USA: 2003. 249–275.
6. Tibshirani RJ. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B. 1996; 58:267–288.
7. Zou HuiHastie Trevor. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005; 67(2):301–320.
8. Lin Y, Zhang H. Component selection and smoothing in smoothing spline analysis of variance models. Annals of Statistics. 2006; 34:2272–2297.
9. George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993; 88:881–889.
10. Reich BJ, Storlie CB, Bondell HD. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. Technometrics. 2009; 51:110–120. [PubMed: 19789732]
11. Chung YeonseungDunson David B. Nonparametric bayes conditional distribution modeling with variable selection. Journal of the American Statistical Association. 2012
12. Fop MichaelMurphy Thomas Brendan. Variable selection methods for model-based clustering. 2017 arXiv preprint arXiv:1707.00306.
13. Raftery Adrian E, Dean Nema. Variable selection for model-based clustering. Journal of the American Statistical Association. 2006; 101(473):168–178.
14. Maugis CathyCeleux GillesMartin-Magniette Marie-Laure. Variable selection for clustering with gaussian mixture models. Biometrics. 2009; 65(3):701–709. [PubMed: 19210744]
15. Fop MichaelSmart KeithMurphy Thomas Brendan. Variable selection for latent class analysis with application to low back pain diagnosis. 2015 arXiv preprint arXiv:1512.03350.
16. Marbac MatthieuSedki Mohammed. Variable selection for model-based clustering using the integrated complete-data likelihood. Statistics and Computing. 2017; 27(4):1049–1063.
17. Pan WeiShen Xiaotong. Penalized model-based clustering with application to variable selection. Journal of Machine Learning Research. May.2007 8:1145–1164.
18. Wang SijianZhu Ji. Variable selection for model-based high-dimensional clustering and its application to microarray data. Biometrics. 2008; 64(2):440–448. [PubMed: 17970821]
19. Xie BenhuaiPan WeiShen Xiaotong. Variable selection in penalized model-based clustering via regularization on grouped parameters. Biometrics. 2008; 64(3):921–930. [PubMed: 18162109]
20. Friedman Jerome H, Meulman Jacqueline J. Clustering objects on subsets of attributes (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004; 66(4):815–849.
21. Hoff Peter D. Model-based subspace clustering. Bayesian Analysis. 2006; 1(2):321–344.

22. Witten Daniela M, Tibshirani Robert. A framework for feature selection in clustering. Journal of the American Statistical Association. 2012

23. Richardson SylviaGreen Peter J. On bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society: series B (statistical methodology). 1997; 59(4):731–792.

24. Kim SinaeTadesse Mahlet G, Vannucci Marina. Variable selection in clustering via dirichlet process mixture models. Biometrika. 2006; 93(4):877–893.

25. Ferguson Thomas S. A Bayesian analysis of some nonparametric problems. The Annals of Statistics. 1973; 1(2):209–230.

26. Neal Radford M. Markov chain sampling methods for dirichlet process mixture models. Journal of computational and graphical statistics. 2000; 9(2):249–265.

27. Teh Yee WhyeJordan Michael I, Beal Matthew J, Blei David M. Hierarchical dirichlet processes. Journal of the American Statistical Association. 2006; 101:1566–1581.

28. Hjort Nils LidHolmes ChrisMüller PeterWalker Stephen G. Bayesian Nonparametrics. Cambridge University Press; New York, NY: 2010.

29. Hunt LynetteJorgensen Murray. Mixture model clustering for mixed data with missing information. Computational Statistics & Data Analysis. 2003; 41(3):429–440.

30. Biernacki ChristopheDeregnaucourt ThibaultKubicki Vincent. Model-based clustering with mixed/missing data using the new software mixtcomp. CMStatistics 2015 (ERCIM 2015). 2015

31. Murray Jared S, Reiter Jerome P. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. 2016 arXiv preprint arXiv:1410.0438.

32. Hennig ChristianLiao Tim F. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2013; 62(3):309–369.

33. Muthen Bengt. Latent variable structural equation modeling with categorical data. Journal of Econometrics. 1983; 22(1):43–65.

34. Everitt Brian S. A finite mixture model for the clustering of mixed-mode data. Statistics & probability letters. 1988; 6(5):305–309.

35. Dunson David B. Bayesian latent variable models for clustered mixed outcomes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62(2):355–366.

36. Ranalli MoniaRocci Roberto. Mixture models for mixed-type data through a composite likelihood approach. Computational Statistics & Data Analysis. 2017; 110:87–102.

37. McParland DamienGormley Isobel Claire. Model based clustering for mixed data: clustmd. Advances in Data Analysis and Classification. 2016; 10(2):155–169.

38. Lesaffre EmmanuelMolenberghs Geert. Multivariate probit analysis: a neglected procedure in medical statistics. Statistics in Medicine. 1991; 10(9):1391–1403. [PubMed: 1925169]

39. Chib SiddharthaGreenberg Edward. Analysis of multivariate probit models. Biometrika. 1998; 85(2):347–361.

40. Jain SoniaNeal Radford M. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Journal of Computational and Graphical Statistics. 2004; 13(1):158–182.

41. Storlie Curtis B, Lane William A, Ryan Emily M, Gattiker James R, Higdon David M. Calibration of computational models with categorical parameters and correlated outputs via bayesian smoothing spline anova. Journal of the American Statistical Association. 2015; 110(509):68–82.

42. Storlie CB, Therneau TerryCarter RickeyChia NicholasBergquist JohnRomero-Brufau Santiago. Prediction and inference with missing data in patient alert systems. Journal of the American Statistical Association (in review). 2017. https://arxiv.org/pdf/1704.07904.pdf

43. Rubin Donald B. Inference and missing data. Biometrika. 1976; 63(3):581–592.

44. Escobar Michael D, West Mike. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association. 1995; 90(430):577–588.

45. MacEachern Steven N, Müller Peter. Estimating mixture of dirichlet process models. Journal of Computational and Graphical Statistics. 1998; 7(2):223–238.

46. Imai Kosukevan Dyk David A. A bayesian analysis of the multinomial probit model using marginal data augmentation. Journal of Econometrics. 2005; 124(2):311–334.

47. McCulloch Robert E, Polson Nicholas G, Rossi Peter E. A bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics. 2000; 99(1):173–193.

48. Zhang XiaoJohn Boscardin W, Belin Thomas R. Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. Computational statistics & data analysis. 2008; 52(7):3697–3708. [PubMed: 19396365]

49. Bhattacharya AnirbanDunson David B. Simplex factor models for multivariate unordered categorical data. Journal of the American Statistical Association. 2012; 107(497):362–377. [PubMed: 23908561]

50. George EI, McCulloch RE. Approaches for Bayesian variable selection. Statistica Sinica. 1997; 7:339–373.

51. Rue HavardHeld Leonhard. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC; Boca Raton, FL: 2005.

52. Bodnar TarasOkhrin Yarema. Properties of the singular, inverse and generalized inverse partitioned wishart distributions. Journal of Multivariate Analysis. 2008; 99(10):2389–2405.

53. Gelman Andrew, et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). Bayesian analysis. 2006; 1(3):515–534.

54. Huang AlanWand Matthew P. , et al. Simple marginally noninformative prior distributions for covariance matrices. Bayesian Analysis. 2013; 8(2):439–452.

55. Ishwaran HemantJames Lancelot F. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. 2001

56. Sethuraman J. A constructive definition of Dirichlet priors. Statistica Sinica. 1994; 4:639–650.

57. Stephens Matthew. Dealing with label switching in mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000; 62(4):795–809.

58. Fu Audrey QiuyanRussell StevenBray Sarah J, Tavaré Simon, et al. Bayesian clustering of replicated time-course gene expression data with weak signals. The Annals of Applied Statistics. 2013; 7(3):1334–1361.

59. Stekhoven Daniel J, Bühlmann Peter. Missforest: non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012; 28(1):112–118. [PubMed: 22039212]

60. Katusic Slavica K, Myers ScottColligan Robert C, Voigt RobertStoeckel Ruth E, Port John D, Croarkin Paul E, Weaver Amy. Developmental brain dysfunction-related disorders and potential autism spectrum disorder (pasd) among children and adolescents - population based 1976–2000 birth cohort. Lancet Neurology. 2016 (in review).

61. Giudici PaoloGreen PJ. Decomposable graphical gaussian model determination. Biometrika. 1999; 86(4):785–801.

62. Wong FrederickCarter Christopher K, Kohn Robert. Efficient estimation of covariance selection models. Biometrika. 2003; 90(4):809–830.

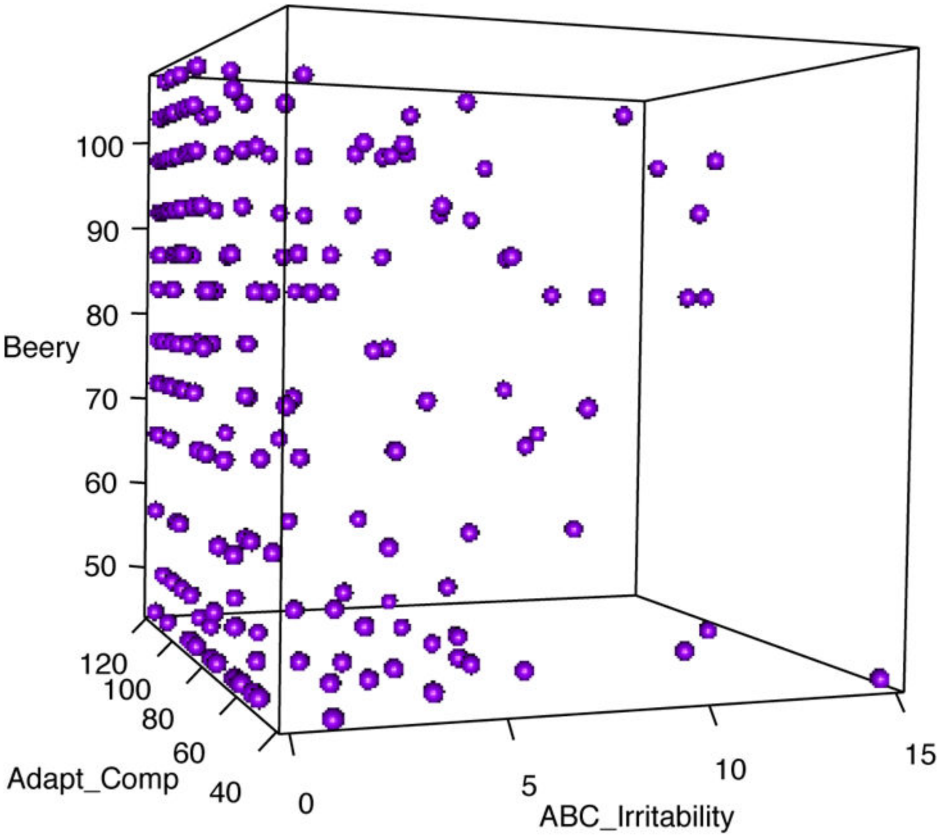

Figure 1.
3D scatter plot of three of the test score variables for potential ASD subjects.

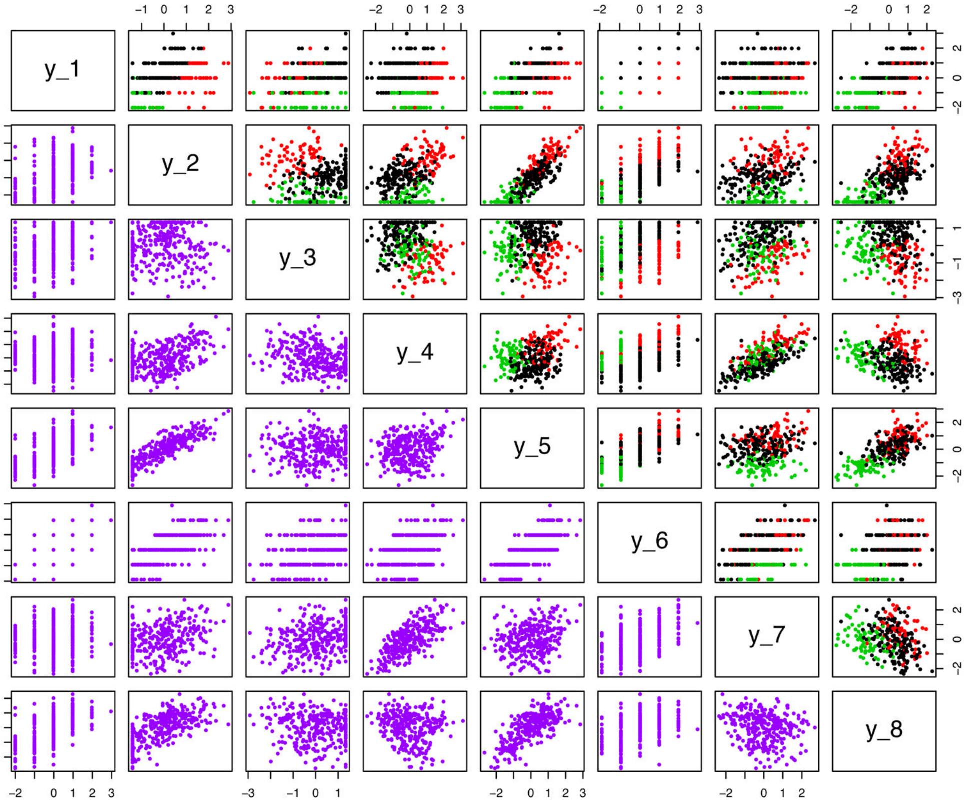

Figure 2.

Pairwise scatter plots of the first eight variables for simulation Case 2(c).

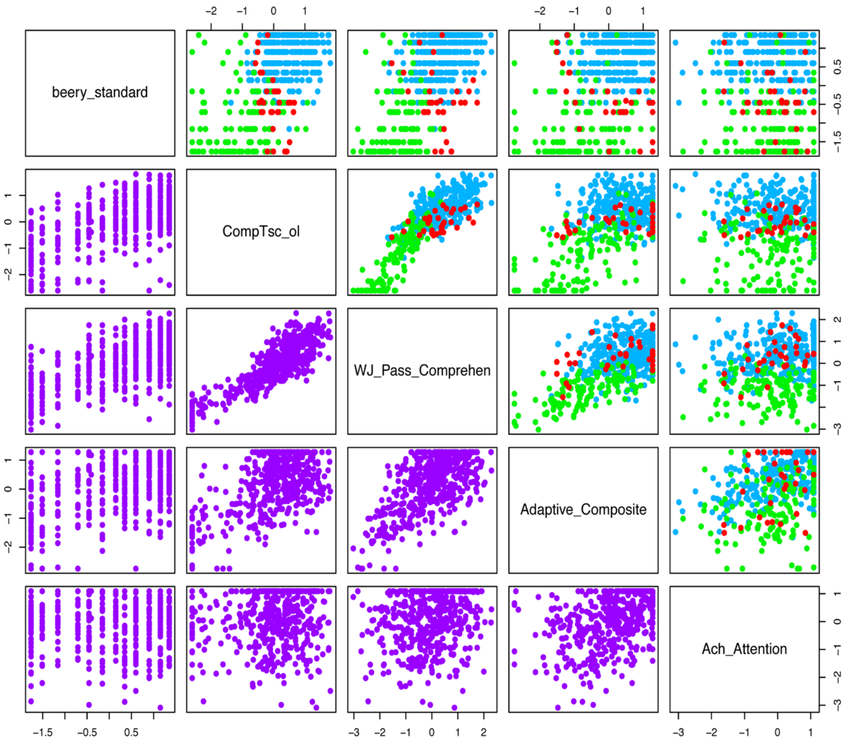

Figure 3.

Pairwise scatter plots of the standardized version of the five most informative variables in Table 3 with estimated cluster membership above the diagonal and the raw data below.

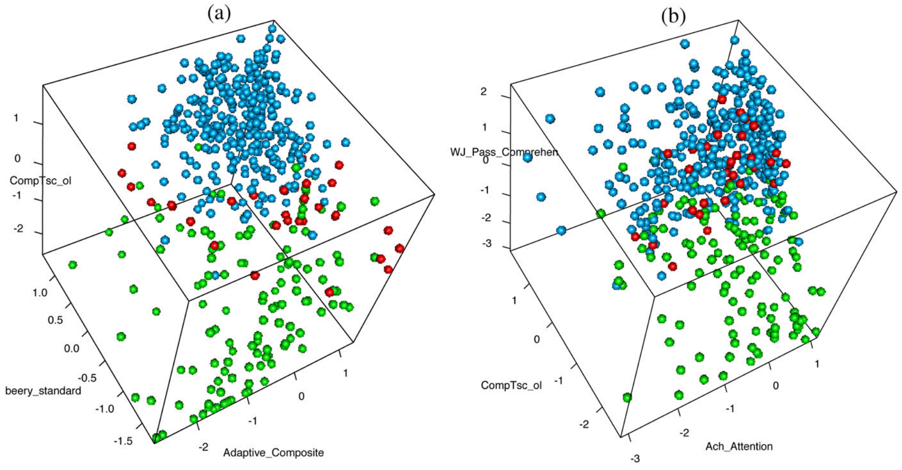

Figure 4.

Three dimensional scatter plots of the tests on standardized scale: (a) The most informative three variables with estimated cluster membership. (b) Observations plotted on the variables CompTsc_ol, WJ_Pass_Comprehen, and Ach_abc_Attention, to better illustrate the separation of clusters 2 and 3.

Table 1

Simulation Case 1 Results.

Method	*Acc*	*FI*	*ARI*	*M*	p_1	*PVC*	*CompT*
			Case 1(a)				
DPM-vs *	**0.91** (0.11)	**0.86** (0.07)	**0.78** (0.16)	**2.9** (0.4)	**2.0** (0.3)	**0.99** (0.04)	296 (18)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.1)	10.0 (0.0)	0.20 (0.00)	341 (28)
DPM-ind	**0.90** (0.13)	**0.86** (0.08)	**0.76** (0.20)	**3.0** (0.6)	**1.9** (0.4)	**0.99** (0.04)	295 (23)
Mclust-vs	**0.91** (0.10)	**0.86** (0.07)	**0.78** (0.15)	**3.0** (0.4)	**2.0** (0.2)	**0.99** (0.06)	**1** (0)
VarSelLCM	0.74 (0.21)	0.73 (0.10)	0.52 (0.29)	**2.9** (1.2)	3.5 (3.2)	0.84 (0.32)	6 (2)
			Case 1(b)				
DPM-vs *	**0.89** (0.14)	**0.85** (0.08)	**0.76** (0.20)	**3.0** (0.7)	1.9 (0.4)	**0.98** (0.05)	267 (19)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	10.0 (0.0)	0.20 (0.00)	292 (15)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	8.5 (0.8)	0.35 (0.08)	243 (18)
Mclust-vs	**0.87** (0.12)	0.83 (0.10)	**0.72** (0.19)	**2.8** (0.4)	**2.0** (0.1)	0.94 (0.12)	**1** (0)
VarSelLCM	0.52 (0.06)	0.57 (0.05)	0.40 (0.06)	6.7 (0.9)	9.1 (0.8)	0.29 (0.08)	11 (2)
			Case 1(c)				
DPM-vs	**0.85** (0.14)	**0.81** (0.08)	**0.68** (0.20)	**2.8** (0.6)	**1.9** (0.4)	**0.97** (0.07)	254 (18)
DPM-cont	0.62 (0.06)	0.52 (0.06)	0.31 (0.07)	4.9 (0.7)	**2.0** (0.5)	0.89 (0.11)	296 (25)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	10.0 (0.0)	0.20 (0.00)	285 (19)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	8.2 (1.0)	0.38 (0.10)	243 (24)
Mclust-vs	0.49 (0.10)	0.43 (0.09)	0.19 (0.12)	6.7 (1.5)	2.7 (0.8)	0.67 (0.14)	**1** (0)
VarSelLCM	0.48 (0.08)	0.51 (0.06)	0.34 (0.07)	6.8 (1.1)	8.4 (0.8)	0.36 (0.08)	10 (1)
			Case 1(d)				
DPM-vs	**0.77** (0.19)	**0.76** (0.10)	**0.57** (0.25)	**2.6** (0.8)	**1.9** (0.6)	**0.95** (0.09)	304 (21)
DPM-cont	0.62 (0.04)	0.52 (0.05)	0.31 (0.06)	4.8 (0.8)	**1.9** (0.5)	0.89 (0.11)	355 (29)
DPM	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.1)	10.0 (0.0)	0.20 (0.00)	343 (29)
DPM-ind	0.37 (0.02)	0.58 (0.00)	0.00 (0.00)	1.0 (0.0)	7.6 (1.3)	0.44 (0.13)	277 (40)

Method	*Acc*	*FI*	*ARI*	*M*	p_1	*PVC*	*CompT*
Mclust-vs	0.50 (0.10)	0.43 (0.08)	0.20 (0.11)	7.0 (1.1)	3.1 (0.9)	0.64 (0.14)	**1** (0)
VarSelLCM	0.50 (0.08)	0.51 (0.06)	0.34 (0.07)	6.5 (1.1)	8.3 (0.8)	0.36 (0.08)	12 (9)
True	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.0 (0.0)	2.0 (0.0)	1.00 (0.00)	–

[*] DPM-cont is identical to DPM-vs for cases 1(a) and 1(b) and is therefore not listed separately.

Table 2

Simulation Case 2 Results.

Method	***Acc***	***FI***	***ARI***	***M***	p_1	***PVC***	***CompT***
			Case 2(a)				
DPM-vs*	**0.89** (0.11)	**0.85** (0.09)	**0.74** (0.20)	**3.0** (0.6)	**3.8** (0.7)	**0.99** (0.02)	544 (45)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.4 (0.8)	30.0 (0.0)	0.13 (0.00)	1158 (120)
DPM-ind	**0.90** (0.10)	**0.85** (0.08)	**0.75** (0.17)	**3.1** (0.6)	**3.9** (0.4)	**1.00** (0.02)	543 (36)
Mclust-vs	**0.91** (0.12)	**0.88** (0.09)	**0.78** (0.23)	**2.9** (0.5)	**4.0** (0.8)	0.98 (0.07)	**8** (3)
VarSelLCM	0.68 (0.07)	0.68 (0.04)	0.53 (0.06)	4.7 (0.6)	**4.0** (0.0)	**1.00** (0.00)	25 (1)
			Case 2(b)				
DPM-vs*	**0.90** (0.10)	**0.85** (0.08)	**0.75** (0.17)	**3.1** (0.6)	**3.9** (0.5)	**0.99** (0.03)	557 (27)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	30.0 (0.0)	0.13 (0.00)	1009 (67)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.9 (0.2)	0.17 (0.01)	966 (71)
Mclust-vs	0.83 (0.14)	0.80 (0.12)	0.63 (0.24)	2.6 (0.5)	3.4 (0.8)	0.91 (0.10)	**8** (3)
VarSelLCM	0.43 (0.04)	0.46 (0.04)	0.26 (0.05)	7.3 (0.6)	27.3 (1.4)	0.22 (0.05)	53 (3)
			Case 2(c)				
DPM-vs	**0.93** (0.03)	**0.89** (0.05)	**0.82** (0.07)	**3.3** (0.6)	**4.0** (0.2)	**1.00** (0.02)	589 (47)
DPM-cont	0.62 (0.16)	0.57 (0.14)	0.34 (0.20)	4.4 (1.2)	3.4 (1.4)	0.87 (0.05)	597 (49)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	30.0 (0.0)	0.13 (0.00)	1078 (54)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.8 (0.5)	0.17 (0.02)	1060 (101)
Mclust-vs	0.45 (0.06)	0.40 (0.04)	0.13 (0.06)	6.4 (1.3)	2.8 (1.0)	0.81 (0.04)	**8** (24)
VarSelLCM	0.42 (0.05)	0.43 (0.05)	0.21 (0.06)	7.3 (0.8)	26.6 (1.5)	0.24 (0.05)	59 (3)
			Case 2(d)				
DPM-vs	**0.91** (0.08)	**0.86** (0.08)	**0.78** (0.14)	**3.2** (0.6)	**3.9** (0.4)	**0.99** (0.03)	577 (54)
DPM-cont	0.61 (0.16)	0.55 (0.14)	0.32 (0.19)	4.3 (1.1)	3.2 (1.2)	0.87 (0.05)	581 (41)
DPM	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.3 (0.5)	30.0 (0.0)	0.13 (0.00)	1028 (85)
DPM-ind	0.50 (0.03)	0.61 (0.01)	0.00 (0.00)	1.0 (0.0)	28.2 (1.1)	0.19 (0.04)	958 (81)

Method	*Acc*	*FI*	*ARI*	*M*	*p_1*	*PVC*	*CompT*
Mclust-vs	0.46 (0.06)	0.41 (0.04)	0.14 (0.05)	6.7 (1.4)	3.0 (1.2)	0.81 (0.04)	**6** (7)
VarSelLCM	0.43 (0.05)	0.43 (0.05)	0.20 (0.06)	7.0 (1.0)	26.0 (1.8)	0.26 (0.06)	73 (19)
True	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	3.0 (0.0)	4.0 (0.0)	1.00 (0.00)	–

[*] DPM-cont is identical to DPM-vs for cases 1(a) and 1(b) and is therefore not listed separately.

Table 3

Posterior inclusion probabilities and sample means for the six most informative tests.

Variable	**Pr(γ_j = 1)**	***Cluster Means***		
		1	**2**	**3**
Beery_standard	1.000	0.77	−1.02	−0.44
CompTsc_ol	1.000	0.46	−1.21	−0.01
WJ_Pass_Comprehen	0.944	0.38	−1.26	0.30
Adaptive_Composite	0.889	0.44	−0.68	0.16
ach_abc_Attention	0.460	−0.18	0.21	0.04
ach_abc_AnxDep	0.427	−0.04	0.06	−0.01