Dirichlet Process mixture models for unsupervised clustering of symptoms in Parkinsons disease

Nicole White[a**]**Corresponding author. Email: nm.white@qut.edu.au Helen Johnson [a] Peter Silburn [b,c] and Kerrie Mengersen[a]

[a]*Mathematical Sciences, Queensland University of Technology, Brisbane, Australia*; [b]*St. Andrew's War Memorial Hospital and Medical Institute, Brisbane, Australia*; [c]*University of Queensland Centre for Clinical Research, Brisbane, Australia*

Abstract

In this paper, the goal of identifying disease subgroups based on differences in observed symptom profile is considered. Commonly referred to as phenotype identification, solutions to this task often involve the application of unsupervised clustering techniques. In this paper, we investigate the application of a Dirichlet Process mixture (DPM) model for this task. This model is defined by the placement of the Dirichlet Process (DP) on the unknown components of a mixture model, allowing for the expression of uncertainty about the partitioning of observed data into homogeneous subgroups. To exemplify this approach, an application to phenotype identification in Parkinson's disease (PD) is considered, with symptom profiles collected using the Unified Parkinson's Disease Rating Scale (UPDRS).

Clustering, Dirichlet Process mixture, Parkinson's disease, UPDRS.

1 Introduction

The search for patient subgroups that potentially explain observed heterogeneities in complex disease has been a long pursued goal in medical research. Statistical solutions for this task include various methods of unsupervised clustering, whereby patients are partitioned into less heterogeneous subgroups based on similarities in responses to discriminating variables, with no *a priori* classification rule. Common examples of unsupervised clustering tools include $k-$means [20], nearest neighbour [62] and hierarchical clustering methods [27], in addition to finite mixture models [39, 16]. A comprehensive review of these methods and others is provided by [24].

An area of research that has benefited greatly from unsupervised clustering methods is subgroup identification with respect to differences in observed symptom profile. Often referred to as syndrome or phenotype identification, examples of this type of study can be found in migraine [43, 7], Alzhiemer's disease [58, 9], Respiratory illness [49], and Schizophrenia [32, 29].

In this study, the identification of potential phenotypes is explored for Parkinson's disease (PD). This commonly diagnosed, neurodegenerative disorder affects an estimated 1-2% of the over 60 population [52], and is principally characterised by motor impairment, in the form of tremor, rigidity, postural instability and bradykinesia or involuntary movement. Like other complex diseases, a hallmark of PD is marked heterogeneity in symptom expression [46, 13], including differences in symptom severity and rates of progression.

In light of this heterogeneity, a number of previous studies have sought to identify clinical subtypes for PD via unsupervised clustering methods. Recent examples include [45, 53, 55] and [33]. The most recent review of findings from cluster analyses for symptoms associated with PD was given by [54]. In [61], a finite mixture model approach was considered with an application to subgroup identification in PD, based on responses to select symptoms.

In this paper, we consider a Dirichlet Process Mixture (DPM) model [1] for

the unsupervised clustering of symptoms in PD. The DPM model involves the placement of a Dirichlet Process (DP) [12] prior over the unknown components of a mixture model. The clustering behaviour induced by this prior results in a model that captures uncertainty in partitions of the observed data. To date, DPM models have found applications in various areas of medical research. However, to our knowledge, the DPM model is yet to be applied to symptom-based subgroup identification and thus forms the focus of this paper.

This paper is organised as follows. Section 2 outlines the DPM model, and details the features of the DP that make it suitable for unsupervised clustering problems. Similar to the model presented in [61], that considers the same application in Parkinson's disease, the model represents a latent class analysis of multiple, categorical variables, this time in the form of Multinomial distributed symptoms, each represented by three response levels. While the intention of this paper is not to provide a lengthy discussion of theory underlying DPM modelling, key details are highlighted where appropriate and references the relevant literature are made. Section 3 presents the application of the DPM model to subgroup identification in PD. The focus of modelling is placed on clustering in terms of differences in progression of three key, motor-specific symptoms - Tremor, Rigidity and Akinesia - in addition to Activities of Daily Living (ADL). In line with the majority of studies into PD subgroups, information on these symptoms is collected using the Unified Parkinson's Disease Rating Scale (UPDRS). Considered as the gold standard in the assessment of symptoms and overall severity of PD, an excellent discussion of the UPDRS is provided by [44]. A summary of findings and directions for future research is presented in Section 4.

2 Methodology

2.1 Model specification

This paper considers the problem of clustering responses to multiple symptoms for a sample of n patients/subjects. The observed data is denoted by the matrix $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with rows corresponding to subjects and columns corresponding to symptoms. Throughout this paper, the number of included symptoms is equal to J.

For a single subject i and in the absence of clustering, each element of the vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ is assumed to have been generated from an underlying probability distribution

$$y_{ij} \sim p\left(y_{ij}|\theta_{ij}\right) \qquad j = 1, \ldots, n \tag{1}$$

defined by parameter/s θ_{ij}. For the application at hand, each symptom is observed as a categorical variable with three possible outcomes. Hence,

$$p\left(y_{ij}|\theta_{ij}\right) = \prod_{r=1}^{3} \theta_{ijr}^{\mathbb{I}_{y_{ij}=r}}. \tag{2}$$

For J symptoms, the likelihood of $\mathbf{y}_i$ is given by,

$$p\left(\mathbf{y}_i|\theta_i\right) = \prod_{j=1}^{J} p\left(y_{ij}|\theta_{ij}\right) \tag{3}$$

and, for all n subjects, $p\left(\mathbf{y}|\theta\right) = \prod_{i=1}^{n} p\left(\mathbf{y}_i|\theta_i\right)$.

Under the assumption of clustering, observations collected from multiple subjects are described by the same parameters defining the likelihood in Equation (3). Thus, for subjects belonging to the same cluster, indexed by k, the parameters defining their contribution to the likelihood are replaced by $\theta_k = (\theta_{k1}, \ldots, \theta_{kJ})$. It follows then, for a general K clusters, the likelihood for

the entire sample is

$$p\left(\mathbf{y}|\theta\right) = \prod_{k=1}^{K} \prod_{\substack{i: \\ z_i=k}}^{n} \prod_{j=1}^{J} p\left(y_{ij}|\theta_{kj}\right). \tag{4}$$

In Equation (4), the latent variable $z_i \in (1, \ldots, K)$ is introduced to represent the cluster membership for subject i and to augment the parameters θ_{kj} such that $\theta_{ij} = \theta_{kj}$ if $z_i = k$.

Inferences about this latent variable, and the parameters that subsequently define the partition of subjects into homogeneous clusters, has led to the proposition of numerous clustering methods in the literature. Whilst we do not seek to present a comprehensive review here, common examples are distance-based methods such as k-means clustering [20], hierarchical clustering [27] and finite mixture models [39, 16]. In this paper, DPM model is considered for the clustering task and is the focus for the remainder of this Section.

A DPM model is characterised by its use of a DP prior on θ_i. Originally defined by [12], a DP is a distribution over random probability measures defined by a base distribution G_0 and concentration parameter α. For a random probability measure G distributed according to a DP, we write $G \sim DP\left(\alpha, G_0\right)$.

A defining property of a DP that makes it suitable for use in clustering is as follows. For any finite, disjoint partition of a measurable space Θ, $\{A_1, \ldots, A_K\}$, the vector $(G(A_1), \ldots, G(A_K))$ follows a finite dimensional Dirichlet distribution. In other words, the DP is used in clustering as a distribution over disjoint partitions.

The clustering behaviour induced by a DP can also be understood in terms of its so-called discreteness property. For the problem under study, the corresponding DPM model (excluding the latent variable) can be written as follows, in terms

of $\mathbf{y}_i$

$$\begin{aligned} \mathbf{y}_i|\theta_i &\sim p(\mathbf{y}_i|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0). \end{aligned} \tag{5}$$

Applying the property that draws from a DP are discrete with probability 1 [3], there exists a nonzero probability that multiple draws assume the same value for θ, thus inducing clustering behaviour. This discreteness can also be understood via the definition of the DP from [48], whereby a draw from the DP can be represented by an infinite sum of point masses on θ_k, $k = 1, 2, \ldots$. For a discussion on other representations of the DP, see [35].

Comparing this approach to others in the literature, the types of inferences drawn from a DPM model are comparable to those produced by a finite mixture model. However, for the DPM model, the number of unique values for θ_i and, in turn, the number of clusters present is not limited to K. Instead, inferences are based on averaging over all proposed partitions of the data. For the interested reader, further discussions on this comparison are given by [19] and [31].

To complete the DPM model specification for the application at hand, the base distribution G_0 and concentration parameter α require definition. The base distribution incorporates prior information about θ into the model, and can be viewed as the average distribution of the DP; ie. $E(G) = G_0$. In this paper, a conjugate prior in the form of a Dirichlet distribution is assumed, The subscript k has been dropped to indicate prior exchangeability amongst latent components.

$$\begin{aligned} G_0 &= p(\theta_1, \ldots, \theta_J) \\ &= \prod_{j=1}^{J} D\left(\frac{1}{R}, \ldots, \frac{1}{R}\right). \end{aligned} \tag{6}$$

The concentration parameter α controls the prior level of clustering induced by the DP, with no clustering produced as $\alpha \to 0$ [59]. This can either be assigned a fixed value or inferred as part of the model. This paper assumes $\alpha \sim G(a, b)$, a prior first proposed by [11]. Alternatives to this prior are highlighted in the Discussion.

2.2 Model inference

Numerous procedures suited for the estimation of DPM models exist in the literature, with the vast majority involving the use of Markov chain Monte Carlo (MCMC). Amongst the earliest of these methods to appear were the Gibbs samplers devised by [10] and [11], based on the original DP definition by [12] and assuming conjugate priors. However, the use of these approaches is no longer widespread, due to their focus on sampling θ_i, as opposed to the latent variable z_i which is more computationally feasible. Gibbs samplers that conditionally sample each z_i, commonly referred to as the collapsed Gibbs sampler, appear in [36] and [37]. For nonconjugate priors, a Gibbs sampling approach was developed by [60]. For further discussion of these algorithms, the reader is directed to [41].

Other MCMC based algorithms for DPM model estimation include the split-merge algorithm, first developed for conjugate priors by [25] and later for nonconjugate priors by [26]. [57] proposed an algorithm based on Slice sampling [42], recently extended by [28]. An example of an non-MCMC based algorithm is the Variational Bayes' approached proposed by [5].

In this paper, we adopt an algorithm known as the blocked Gibbs sampler [21, 22]. Briefly, the algorithm is based on the stick-breaking construction of the DP [48], whereby draws from the DP are expressed as an infinite sum of point masses on θ, namely,

$$\begin{aligned} G &= \sum_{k=1}^{\infty} \eta_k \delta(\theta_k) \\ \theta_k &\sim G_0. \end{aligned} \tag{7}$$

The stick-breaking weights, η_k, are constructed sequentially from a series of random $Beta\,(1, \alpha)$ variates. The algorithm makes the simplifying assumption that the stick-breaking process can be "blocked" or truncated at $k << \infty$, leading to a Gibbs sampler similar to those developed for finite mixture models (for example, see [38] and [16]). While the use of truncation in DPM model estimation is reserved for the Discussion, care must be taken in the choice of truncation level, L, to ensure it well exceeds the expected number of components.

The blocked Gibbs sampler consists of iterating between sampling from the full conditional for $\mathbf{z} = (z_1, \ldots, z_n)$, θ, η and α. For current estimates of η and θ, the latent variable for subject i is updated from

$$z_i|\eta, \theta, \alpha \sim Mult\left(\frac{\eta_1 \prod_{j=1}^{J} p(y_{ij}|\theta_{j1})}{C}, \ldots, \frac{\eta_L \prod_{j=1}^{J} p(y_{ij}|\theta_{jL})}{C}\right) \tag{8}$$

with $C = \sum_{k=1}^{L} \eta_k \prod_{j=1}^{J} p(y_{ij}|\theta_{jk})$, representing the normalising constant. Given this updated allocation of subjects, the component proportions are updated by an $L-$dimensional Dirichlet distribution:

$$\eta_1, \ldots, \eta_L|\alpha, \mathbf{z} \sim D\left(\frac{\alpha}{L} + N_1, \ldots, \frac{\alpha}{L} + N_L\right), \tag{9}$$

known as the L degree weak approximation to the DP [23]. Here, $N_k = \sum_{i=1}^{n} \{z_i = k\}$ An alternative update for η based on the stick-breaking representation, in the form of Beta distributions, is also possible [21].

Finally, for each occupied cluster, defined by the unique values for $\mathbf{z}$, the component parameters θ are updated by the Dirichlet distribution,

$$(\theta_{jk1}, \ldots, \theta_{jkR}) \sim D\left(\frac{1}{R} + \sum_{i=1}^{n} \mathbb{I}\left\{z_i = k \cap y_{ij} = 1\right\}, \ldots, \frac{1}{R} + \sum_{i=1}^{n} \mathbb{I}\left\{z_i = k \cap y_{ij} = R\right\}\right). \tag{10}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

2.3 Clustering and parameter inference

There exist two main questions are of interest when fitting a DPM model. The first is to estimate the probabilistic clustering of subjects, modelled by the latent variable $\mathbf{z}$, given all clusterings proposed by the Gibbs sampler. The second is to make inferences on the component weights and component specific parameters, as these provide valuable information about the prevalence of each cluster and their dominant characteristics with respect to the Multinomial items included in the model.

Inference on model parameters augmented by the latent variable, and the latent variable itself, are complicated by label switching. There are a number of solutions to the label switching problem, including the imposition of prior constraints [15] and relabelling algorithms [50, 6, 38]. In this paper, a similarity matrix based approach is used for clustering inference. For each realisation of $\mathbf{z}$, denoted by $\mathbf{z}^{(d)}$, the similarity matrix for the proposed classification, $S^{(d)}$, is of size $n \times n$ and consists of elements $S_{ii'}^{(d)}$ satisfying the relation

$$S_{ii'}^{(d)} = \begin{cases} 1 & \text{if } z_i^{(d)} = z_{i'}^{(d)} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

The posterior expected clustering is found by computing the element-wise average of $S^{(d)}$ over all iterations, producing the average similarity matrix $\overline{S}$, with elements $\overline{S}_{i,i'}$ representing the posterior pairwise probability $Pr(z_i = z_{i'})$. From Equation (11), it can be seen that the construction of each $S^{(d)}$ and ultimately $\overline{S}$ is invariant to permutations of labelling on $\mathbf{z}$, with the focus placed on similarities in classification over all subjects.

Given both $S^{(d)}$ and $\overline{S}$, there exist a number of criteria for inferring the most likely clustering $\mathbf{z}^*$, including Binder's loss function [2], least squares [8] and hierarchical agglomerative clustering [40]. These approaches have in common the goal of maximising the association between the posterior expected classification and classifications proposed by the Gibbs sampler, to produce the single

likely clustering $\mathbf{z}^*$. In this paper, the criterion of [14] is employed, in light of its tendency to produce non-singleton clusters over other methods, attributed to its shrinkage property. This criterion involves calculation of the Posterior Expected Adjusted Rand (PEAR) index for each realisation of $\mathbf{z}$. For a single proposed clustering, the PEAR index is,

$$PEAR(d) = \frac{\sum_{i<i'} \mathbb{I}\left\{z_i^{(d)} = z_{i'}^{(d)}\right\} \overline{S}_{ii'} - \sum_{i<i'} \mathbb{I}\left\{z_i^{(d)} = z_{i'}^{(d)}\right\} \sum_{i<i'} \overline{S}_{iji'} / \binom{n}{2}}{0.5(\sum_{i<i'} \mathbb{I}\left\{z_i^{(d)} = z_{i'}^{(d)}\right\} + \sum_{i<i'} \overline{S}_{ii'}) - \sum_{i<i'} \mathbb{I}\left\{z_i^{(d)} = z_{i'}^{(d)}\right\} \sum_{i<i'} \overline{S}_{ii'} / \binom{n}{2}}. \tag{12}$$

Given the symmetry of the similarity matrix, only the lower diagonal is considered. The Adjusted Rand index is featured in this criterion as it rescales the Rand index so it does not exceed 1, allowing for improved interpretation and comparison of different models over its unadjusted form. In practice, Equation (12) is computed for each sampled classification, such that $\mathbf{z}^*$ is taken as the sampled classification equaling the largest PEAR value. In a similar spirit to other aforementioned methods, this PEAR value indicates maximal associations between the chosen and posterior expected clustering.

Given $\mathbf{z}^*$, θ and η are simulated from their respective full conditional distributions to provide parameter inference. Alternatively, for parameters with a conjugate prior distribution, an approximation to the posterior can be obtained by computing the analytic expectation and variance given $\mathbf{z}^*$. For example, the posterior expectation of η given $\mathbf{z}^*$ can be estimated by the theoretical expectation of the Dirichlet distribution, in light of Equation (9). This expectation takes the form,

$$\eta^* = \left(\frac{\alpha/L + N_1(\mathbf{z}^*)}{\sum_{l=1}^{L} \alpha/L + N_l(\mathbf{z}^*)}, \ldots, \frac{\alpha/L + N_L(\mathbf{z}^*)}{\sum_{l=1}^{L} \alpha/L + N_l(\mathbf{z}^*)} \right) \tag{13}$$

where $N_l(\mathbf{z}^*) = \sum_{i=1}^{n} \mathbb{I}\{z_i^* = l\}$. The posterior variance is then given by,

$$Var(\eta|\mathbf{z}^*) = \frac{\eta^*(1-\eta^*)}{1 + \sum_{l=1}^{L} \alpha/L + N_l(\mathbf{z}^*)}. \tag{14}$$

Similar expressions are easily obtained for both θ^* and α^* (see for example [17], Appendix A).

3 Results

The proposed methodology was applied to the problem of symptom based subgroup identification in PD. As mentioned in the Introduction, a hallmark of PD is marked heterogeneity in observed symptom profiles between subjects, not only with respect to presence but also the severity of different symptoms. To this end, the application of clustering methods to data collected on PD symptoms may aid the identification of clinically meaningful subtypes that attempt to explain this heterogeneity.

The focus in this paper is on the overall severity of key motor symptoms associated with PD, in addition to the impact of PD on Activities of Daily Living (ADL). Data used for this analysis were collected as part of the Queensland Parkinson's Project [51], involving the recruitment of patients diagnosed with PD from five specialist neurological clinics across Brisbane, Australia. Information on patients was collected using the Unified Parkinson's Disease Rating Scale (UPDRS), in particular UPDRS Sections II and III. Variables included in the model were in the form of sums over individual UPDRS items, specific to a given symptom. These variables were ADL (items 5–17), Tremor (items 20 and 21), Rigidity (item 22) and Akinesia (items 23–26).

In preliminary analyses, it was thought that clusters formed with these variables were strongly influenced by duration of PD diagnosis, given the progressive nature of PD, with higher scores generally indicative of a more advanced case of the disease. This was flagged as a potential confounder in the identification of clusters. Therefore, to adjust for possible effects attributed to duration of diagnosis, variables were divided by each subject's duration of diagnosis, such that each variable was now interpreted as a patient's average score per year of diagnosis. Patients with either missing information on duration of diagnosis or a

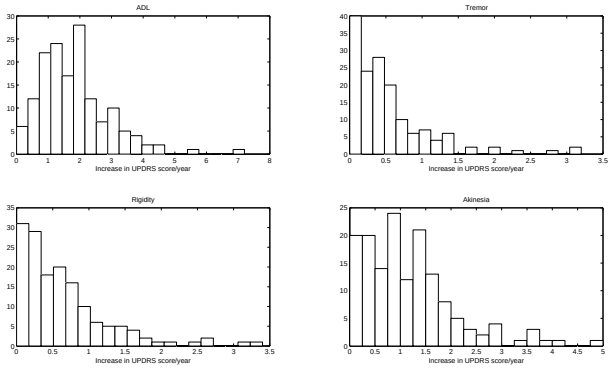

Figure 1: Frequencies histograms for each included symptom in DPM analysis. Variables are interpreted as the increase in relevant UPDRS score per year of PD diagnosis.

Table 1: Cut offs for recoding of UPDRS-based variables into Multinomial categories, to be used in the clustering analysis.

Symptom	Minimum	33^{rd} percentile	66^{th} percentile	Maximum
ADL	0	1.23	2.05	7.2
Tremor	0	0.20	0.55	3.2
Rigidity	0	0.29	0.74	3.4
Akinesia	0	0.72	1.37	5.0

duration of diagnosis less than five years were further excluded from the analysis. Application of these exclusions resulted in a sample size of $n = 153$ patients, with an average duration of PD diagnosis of 7.84 years and corresponding standard deviation of 6.23 years. Figure 1 provides observed data histograms for each included variable.

As a final step, each variable included in the analysis was recoded into one of three categories, determined by its empirical quantiles. For all variables, Level 1 corresponded to the lowest level of severity and Level 3 the highest severity level. Table 1 summarises the scores for each UPDRS-based variable, representing the cut-offs for the creation of the three Multinomial categories after accounting for duration of diagnosis. Of these variables, ADL had the highest rate of increase per year of diagnosis.

The truncation level for the construction of the stick breaking weights was set

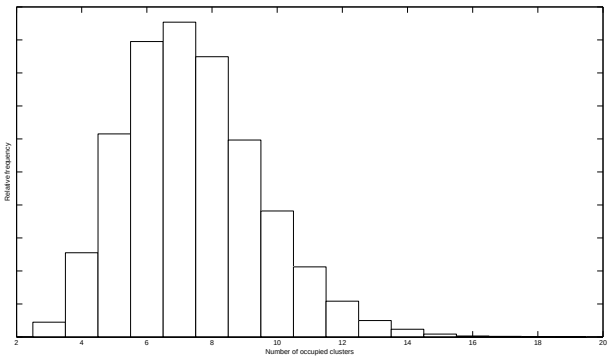

Figure 2: Relative frequencies of occupied clusters over the course of the blocked Gibbs sampler, given a truncation level of 100 and a $G(2, 2)$ prior on α.

to $L = 100$. Uncertainty in α was modelled by a $G(2, 2)$ prior. The blocked Gibbs sampler described in the previous Section was run on the recoded data for 100,000 iterations, discarding the first 50,000 iterations as the burnin period. To explore variation in the level of clustering during the course of the sampler, the number of occupied clusters was recorded at the end of each iteration. This quantity of interest is summarised in terms of relative frequencies in Figure 2. Overall, the sampler inferred a median of seven clusters, with a 95% credible interval of four to twelve clusters and 80% credible interval of five to ten clusters. The posterior distribution of α is displayed in Figure 3. In summary, the posterior mean for α was 1.333, with a posterior variance equal to 0.376.

To infer the most likely clustering of subjects based on the MCMC output, the classification similarity matrix was constructed by sampling the proposed classification at every 100^{th} iteration post burnin. Using the PEAR criterion, the most likely classification consisted of seven mixture components. The posterior means and standard deviations for each parameter, using the method described in Section 2.3, are given in Table 2.

Describing each cluster in turn, Cluster 1 was characterised by high probabilities of high progression over all symptoms. Cluster 2 was characterised by high probabilities of moderate progression in ADL and Akinesia. Cluster 3 repre-

Table 2: Parameter estimates for the component weights (η) and component specific parameters (θ) for the DPM analysis based on a $G(2, 2)$ prior for α. Estimates are summarised in terms of posterior expectation and standard deviation. Asterisks denote singleton clusters

ADL							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7*
1	0.09	0.16	0.85	0.45	0.62	0.11	0.17
	(0.04)	(0.05)	(0.06)	(0.10)	(0.17)	(0.16)	(0.21)
2	0.20	0.60	0.11	0.32	0.19	0.78	0.66
	(0.06)	(0.07)	(0.05)	(0.09)	(0.14)	(0.21)	(0.27)
3	0.71	0.24	0.04	0.23	0.19	0.11	0.17
	(0.06)	(0.06)	(0.03)	(0.08)	(0.14)	(0.16)	(0.21)
Tremor							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7*
1	0.09	0.43	0.69	0.02	0.05	0.11	0.17
	(0.04)	(0.07)	(0.08)	(0.02)	(0.07)	(0.16)	(0.21)
2	0.26	0.24	0.30	0.88	0.62	0.11	0.66
	(0.06)	(0.06)	(0.08)	(0.06)	(0.17)	(0.16)	(0.27)
3	0.65	0.33	0.01	0.10	0.33	0.78	0.17
	(0.07)	(0.07)	(0.02)	(0.06)	(0.16)	(0.21)	(0.21)
Rigidity							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7*
1	0.01	0.52	0.59	0.06	0.90	0.44	0.66
	(0.01)	(0.07)	(0.08)	(0.05)	(0.10)	(0.25)	(0.27)
2	0.13	0.30	0.33	0.93	0.05	0.12	0.17
	(0.05)	(0.06)	(0.08)	(0.05)	(0.07)	(0.16)	(0.21)
3	0.86	0.18	0.08	0.01	0.05	0.44	0.17
	(0.05)	(0.05)	(0.04)	(0.02)	(0.07)	(0.25)	(0.21)
Akinesia							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7*
1	0.07	0.01	0.95	0.49	0.90	0.78	0.17
	(0.04)	(0.01)	(0.04)	(0.10)	(0.10)	(0.20)	(0.21)
2	0.05	0.88	0.04	0.23	0.05	0.11	0.17
	(0.03)	(0.04)	(0.03)	(0.08)	(0.07)	(0.15)	(0.21)
3	0.88	0.11	0.01	0.28	0.05	0.11	0.66
	(0.05)	(0.04)	(0.02)	(0.09)	(0.07)	(0.15)	(0.27)
η	0.30	0.30	0.20	0.14	0.04	0.01	0.01
	(0.03)	(0.03)	(0.03)	(0.03)	(0.01)	(0.01)	(0.01)

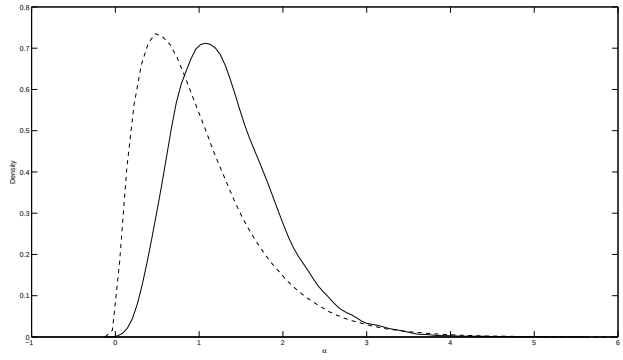

Figure 3: Posterior distribution of the DP concentration parameter α (solid line), based on a $G(2, 2)$ prior (dashed line).

sented the reverse trend to Cluster 3, with strong evidence of low progression over all four symptoms. Representing the last of the dominant clusters, Cluster 4 appeared to be described by high probabilities of moderate progression for Tremor and Rigidity.

The remaining three clusters, (5,6 and 7) had relatively smaller weights, all below 0.1, with Cluster 7 containing a single patient. Given these low weights, it was thought that Clusters 5, 6 and 7 represented multivariate outliers in the observed data. Alternatively, they may have represented so-called "emerging" clusters; ie. clusters that were not supported by the observed data but may be more prominent with the collection of information of additional patients. Finally, the presence of these clusters could have resulted from the prior specified for α, given its direct role in clustering inference. This prompted a sensitivity analysis on α, presented in the next Section.

It is of interest to note similarities and differences between the inferred partition and results from previous studies of PD. In a systematic review of previous studies by [55], a number of studies cited identified "slow progression" and "high progression" subgroups with respect to all included symptoms. This finding was similar to the identification of Clusters 1 and 3 in this analysis. In [45] and [33], tremor-dominant and rigidity dominant subtypes were identified. This appeared

to be in contrast to the combined Tremor/Rigidity cluster found here (Cluster 4), however, in the former studies, the focus was on overall severity as opposed to progression.

3.1 Sensitivity analysis

To investigate the role of α in clustering inference, a sensitivity analysis was conducted for different choices of the hyperparameters a and b, involved in the prior distribution for α. The results that follow are based on two priors, namely a $G(4, 2)$ and $G(2, 4)$ prior, each representing different prior expectations for α. Based on the result by [1][1], the specification of these priors resulted in a prior expected number of clusters equal to 8.70 and 2.86 clusters, for priors $G(4, 2)$ and $G(2, 4)$ respectively.

The blocked Gibbs sampler was again run for 100,000 iterations with the first 50,000 iterations discarded. Figures 4(a) and 4(b) display the posterior distributions for α and the number of occupied clusters, for each prior tested.

In Figure 4(a), differences in the posterior distributions for α were observed, most notably with respect to spread. Differences in posterior means were also noted, with $E(\alpha|\mathbf{z}, \mathbf{y}) = (1.99, 0.82)$ for $G(4, 2)$ and $G(2, 4)$ respectively. The comparison of posterior distributions in Figure 4(b) showed evidence of a higher median number of clusters for the $G(4, 2)$, as expected given its corresponding prior expectation. Conversely, the equivalent posterior distribution for the $G(2, 4)$ prior exhibited a slightly narrow range of occupied clusters.

Parameter inference, for each choice of prior, is summarised in Tables 3 and 4, again in terms of posterior means and standard deviations.

Of immediate note in the comparison of these two clustering solutions were differences in the number of inferred clusters, at most seven for the $G(4, 2)$ prior. However, in comparing results for the first four clusters under each prior, it was seen that these clusters were very similar with respect to dominant features. Cluster descriptions with respect to these features were also aligned with those

[1] $E(K) \approx \alpha \log\left(1 + \frac{n}{\alpha}\right)$

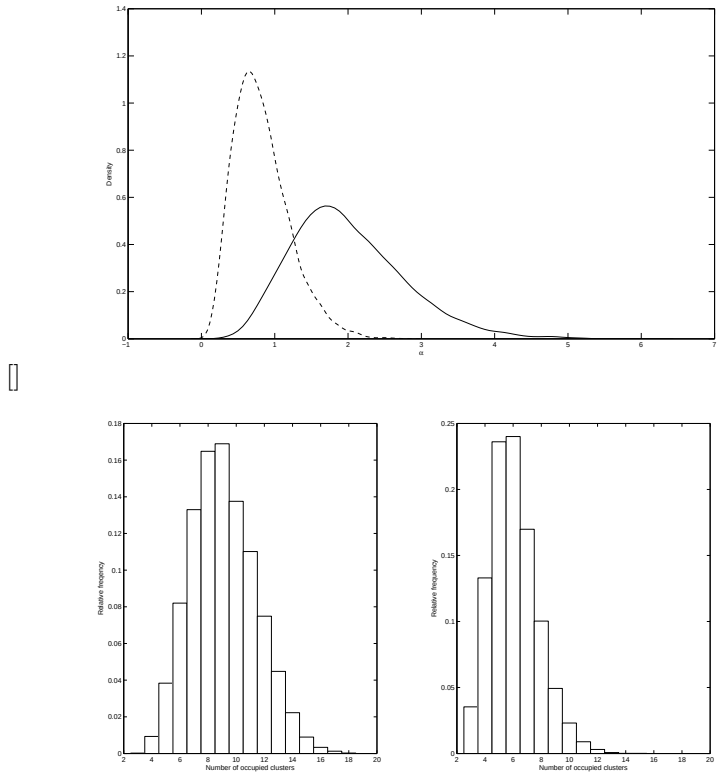

Figure 4: Posterior distributions for fig:ch3aalphasens α for each prior tested. Solid line = $G(4,2)$, Dashed line = $G(2,4)$; fig:ch3aKsens the number of occupied clusters (left to right): $G(4,2)$ and $G(2,4)$.

Table 3: Parameter estimates for revised DPM analysis, based on $\alpha \sim G(4, 2)$, summarised in terms of posterior expectation and standard deviation.

ADL							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
1	0.09	0.77	0.22	0.63	0.03	0.05	0.08
	(0.04)	(0.07)	(0.07)	(0.11)	(0.05)	(0.07)	(0.12)
2	0.20	0.14	0.74	0.24	0.03	0.90	0.58
	(0.06)	(0.05)	(0.07)	(0.10)	(0.05)	(0.10)	(0.22)
3	0.71	0.09	0.04	0.13	0.94	0.05	0.34
	(0.06)	(0.04)	(0.03)	(0.08)	(0.07)	(0.07)	(0.21)
Tremor							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
1	0.09	0.56	0.59	0.02	0.14	0.05	0.08
	(0.04)	(0.08)	(0.08)	(0.03)	(0.10)	(0.07)	(0.12)
2	0.26	0.32	0.22	0.96	0.43	0.62	0.08
	(0.06)	(0.07)	(0.07)	(0.04)	(0.15)	(0.17)	(0.12)
3	0.65	0.12	0.19	0.02	0.43	0.33	0.84
	(0.07)	(0.05)	(0.07)	(0.03)	(0.15)	(0.17)	(0.17)
Rigidity							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
1	0.05	0.69	0.53	0.02	0.63	0.05	0.08
	(0.03)	(0.07)	(0.08)	(0.03)	(0.14)	(0.07)	(0.12)
2	0.15	0.27	0.22	0.91	0.33	0.90	0.58
	(0.05)	(0.07)	(0.07)	(0.06)	(0.14)	(0.10)	(0.22)
3	0.80	0.04	0.25	0.07	0.04	0.05	0.34
	(0.05)	(0.03)	(0.07)	(0.06)	(0.05)	(0.07)	(0.21)
Akinesia							
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
1	0.07	0.90	0.01	0.74	0.03	0.05	0.33
	(0.03)	(0.05)	(0.02)	(0.10)	(0.05)	(0.07)	(0.21)
2	0.05	0.01	0.92	0.24	0.94	0.33	0.58
	(0.03)	(0.01)	(0.05)	(0.10)	(0.07)	(0.17)	(0.22)
3	0.88	0.09	0.07	0.02	0.03	0.62	0.09
	(0.04)	(0.04)	(0.04)	(0.03)	(0.05)	(0.17)	(0.12)
η	0.32	0.24	0.21	0.11	0.06	0.04	0.02
	(0.04)	(0.03)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)

Table 4: Parameter estimates for revised DPM analysis, based on $\alpha \sim G(2, 4)$, summarised in terms of posterior expectation and standard deviation.

ADL				
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.11	0.05	0.89	0.39
	(0.04)	(0.03)	(0.05)	(0.10)
2	0.61	0.22	0.10	0.37
	(0.07)	(0.06)	(0.05)	(0.10)
3	0.28	0.73	0.01	0.24
	(0.06)	(0.06)	(0.01)	(0.09)
Tremor				
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.47	0.07	0.47	0.06
	(0.07)	(0.04)	(0.08)	(0.05)
2	0.11	0.31	0.47	0.88
	(0.04)	(0.07)	(0.08)	(0.07)
3	0.42	0.62	0.06	0.06
	(0.07)	(0.07)	(0.04)	(0.05)
Rigidity				
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.53	0.05	0.52	0.15
	(0.07)	(0.03)	(0.08)	(0.07)
2	0.30	0.14	0.42	0.65
	(0.06)	(0.05)	(0.08)	(0.10)
3	0.17	0.81	0.06	0.20
	(0.05)	(0.06)	(0.04)	(0.08)
Akinesia				
Level	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	0.13	0.09	0.93	0.15
	(0.05)	(0.04)	(0.04)	(0.07)
2	0.80	0.03	0.01	0.47
	(0.06)	(0.02)	(0.01)	(0.10)
3	0.07	0.88	0.06	0.38
	(0.04)	(0.05)	(0.04)	(0.10)
η	0.31	0.29	0.26	0.14
	(0.04)	(0.04)	(0.03)	(0.03)

obtained from the original analysis, involving a $G(2, 2)$ prior on α. That said, the comparison of component weights under different priors revealed some differences between Clusters 1, 2 and 3. For example, the moderate ADL/Akinesia cluster was ranked third with respect to weight under the $G(4, 2)$ prior, compared to first for $G(2, 4)$ and second under the $G(2, 2)$ prior. These discrepancies could be attributed to the relative closeness of inferred weights for these three clusters. In greater detail, in transitioning from models with more to fewer clusters, the reclassification of some patients into more dominant clusters resulted in changes to the posterior means for the corresponding weights. Overall, the choice of prior on α was seen to impact on the number of clusters inferred; however, in terms of identifying dominant clusters and their features, resulting inferences were largely robust to the choice of prior.

4 Discussion

In review, this paper has investigated the problem of clustering multivariate, categorical data with an application to the identification of symptom subgroups in PD. It represents a novel application of the DPM model, an unsupervised clustering method that allows for uncertainty about the partition of patients into an unknown number of subgroups.

The results presented in Section 3 beared similarities to other studies that have considered clustering techniques for the identification of PD subtypes. The identification of Clusters 1 and 3 were in line with previous results [45, 33]. In these studies, the presence of these clusters was linked to age on PD onset, with younger age of onset patients typically characterised under Cluster 1. Conversely, the identification of Clusters 2 and 4 appear to differ from the results of previous studies and are thus worthy of further investigation.

The analysis presented in this paper brought to attention the role of the concentration parameter involved in the DP prior, in particular how the choice of prior on α affected clustering inference. This aspect of DPM modelling has been

discussed by many authors, among those [22] and [59] where larger values of α tended to produce more smoothing and larger numbers of occupied clusters. In this study, a conjugate Gamma prior on α was chosen, and the choice of hyperparameters affected the number of inferred clusters. However, parameter inference on dominant clusters was not largely affected by this choice. That said, there exist other, more flexible, prior distributions for the concentration parameter. In [34], the DP is replaced by a prior obtained by a generalised Gamma process. Alternatively, in [56], an enriched conjugate prior for the DP is developed. Although not considered in this paper, these distributions provide competitive alternatives to the use of a Gamma prior and could be considered in future studies of this nature.

Out of the numerous algorithms available for DPM model inference, the blocked Gibbs sampler was chosen in light its relative computational speed. However, this choice of this algorithm required the specification of a truncation level, which is not required for other sampling approaches. Examples of these alternative approaches include those based on the Pòlya Urn representation [4] and the normalised, completely randomised measures representation first proposed by [12]. As a reviewer rightly pointed out, a problem with the use of truncation on the stick-breaking process may ignore the influence of outlying observations or heavy tail behaviour. In the present study, the blocked Gibbs sampler was rerun with multiple different choices for L, with resulting inferences not affected by this choice. Nevertheless, the choice of sampler for DPM model estimation should be carefully considered as part of any analysis.

A broader issue faced in the use of unsupervised clustering methods, including the DPM model, is that of variable selection. It is widely acknowledged in the literature that the choice of variables included in a clustering analysis may affect the identification of different clusters and, in many cases, some variables may play little to no role in differentiating between clusters. For the featured application, while variables were chosen for clinical reasons, it would be of great interest to compare results with models including variable selection, in this case, the se-

lection of symptoms that best discriminate between patients. Furthermore, in other applications, variable selection may represent a significant challenge. A recent innovation in clustering problems is the incorporation of Stochastic Search Variable Selection (SSVS) [18], whereby variables are randomly included or excluded during the course of MCMC and their role in discriminating between clusters assessed. For DPM models, this extension was considered by [30] and also presents opportunities for future work.

Acknowledgements

The authors wish to thank George Mellick and Nadeeka Dissanayaka for their assistance in providing data for this study. We also kindly thank two anonymous reviewers for their thoughtful comments and feedback on revisions of this manuscript.

References

[1] C.E. Antoniak, *Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems*, Annals of Statistics 2 (1974), pp. 1152–1174.

[2] D.A. Binder, *Bayesian Cluster Analysis*, Biometrika 65 (1978), pp. 31–38.

[3] D. Blackwell, *Discreteness of Ferguson selections*, The Annals of Statistics. 1 (1973), pp. 356–358.

[4] D. Blackwell and J.B. MacQueen, *Ferguson distributions via Pólya urn schemes*, The Annals of Statistics 1 (1973), pp. 353–355.

[5] D.M. Blei and M.I. Jordan, *Variational inference for Dirichlet process mixtures*, Bayesian Analysis 1 (2006), pp. 121–144

[6] G. Celeux, M. Hurn and C.P. Robert, *Computational and Inferential Diffi-*

culties with Mixture Posterior Distributions, Journal of the American Statistical Association 95 (2000), pp. 957–970.

[7] C. Chen, J.M. Keith, D.R. Nyholt, N.G. Martin and K.L. Mengersen, *Bayesian latent trait modeling of migraine symptom data*, Human genetics 126 (2009), pp. 277–288.

[8] Dahl, DB. Model-based clustering for expression data via a Dirichlet process mixture model. In. *Bayesian inference for gene expression and proteomics* (Eds. Do, K.A and Müller,P. and Vannucci, M.). Cambridge University Press. Cambridge, 2006. pp. 201–218.

[9] J.E. Davidson, M.C. Irizarry, B.C. Bray, S. Wetten, N. Galwey, R. Gibson, M. Borrie,R. Delisle, H.H. Feldman, G.Y. Hsiung, L. Fornazzari, S. Gauthier, D. Guzman, I. Loy-English, R. Keren, A. Kertesz, P. St. George-Hyslop, J. Wherrett and A.U. Monsch, *An exploration of cognitive subgroups in Alzheimer's disease*, Journal of the International Neuropsychological Society 16 (2010), pp. 233–243.

[10] M.D. Escobar, *Estimating normal means with a Dirichlet process prior*, Journal of the American Statistical Association 89 (1994), pp. 268–277.

[11] M.D. Escobar and M. West, *Bayesian Density Estimation and Inference Using Mixtures*, Journal of the American Statistical Association 90 (1995), pp. 577–588.

[12] T.S. Ferguson, *A Bayesian analysis of some nonparametric problems*, The annals of statistics 1 (1973), pp. 209–230.

[13] T. Foltynie, C. Brayne and R.A. Barker, *The heterogeneity of idiopathic Parkinson's disease*, Journal of neurology 249 (2002), pp. 138–145.

[14] A. Fritsch and K. Ickstadt, *Improved criteria for clustering based on the posterior similarity matrix*, Bayesian Analysis 4 (2009), pp. 367–392.

[15] S. Frühwirth-Schnatter, *Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models*, Journal of the American Statistical Association 96 (2001), pp. 194–209.

[16] S. Frühwirth-Schnatter, *Finite mixture and Markov switching models*, Springer Verlag, 2006.

[17] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, *Bayesian data analysis*, 2nd edition, Chapman and Hall/CRC, 2004.

[18] E.I. George and R.E. McCulloch, *Variable selection via Gibbs sampling*, Journal of the American Statistical Association 88 (1993), pp. 881–889.

[19] P.J. Green and S. Richardson, *Modelling heterogeneity with and without the Dirichlet process*, Scandinavian journal of statistics 2 (2001), pp 355–375.

[20] J.A. Hartigan and M.A. Wong, *A k-means clustering algorithm*, Journal of the Royal Statistical Society Series C 28 (1979), pp. 100–108.

[21] H. Ishwaran and L.F. James, *Gibbs sampling methods for stick-breaking priors*, Journal of the American Statistical Association 96 (2001), pp. 161–173.

[22] H. Ishwaran and L.F. James, *Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information*, Journal of Computational and Graphical Statistics 11 (2002), pp. 508–532.

[23] H. Ishwaran and M. Zarepour, *Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models*, Biometrika 87 (2000), pp. 371–390.

[24] A.K. Jain, M.N. Murty and P.J. Flynn, *Data clustering: a review*, ACM computing surveys (CSUR) 31 (1999), pp. 264–323.

[25] S. Jain and R.M. Neal, *A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model*, Journal of Computational and Graphical Statistics 13 (2004), pp. 158–182.

[26] S. Jain and R.M. Neal, *Splitting and merging components of a nonconjugate Dirichlet process mixture model*, Bayesian Analysis 2 (2007), pp. 445–472.

[27] S.C. Johnson, *Hierarchical clustering schemes*, Psychometrika 32 (1967), pp. 241–254.

[28] M. Kalli, J.E. Griffin and S.G. Walker, *Slice sampling mixture models*, Statistics and computing 21 (2011), pp. 93–105.

[29] K.S. Kendler, L.M. Karkowski and D. Walsh, *The structure of psychosis: latent class analysis of probands from the Roscommon Family Study*, Archives of General Psychiatry 55 (1998), pp. 492–509.

[30] S. Kim, M.G. Tadesse and M. Vannucci, *Variable selection in clustering via Dirichlet process mixture models*, Biometrika 93 (2006), pp. 877–893.

[31] J.W. Lau and P.J. Green, *Bayesian model-based clustering procedures*, Journal of Computational and Graphical Statistics 16 (2007), pp. 526–558.

[32] S. Leask, J. Vermunt, D. Done, T. Crow, M. Blows and M. Boks, *Beyond symptom dimensions: Schizophrenia risk factors for patient groups derived by latent class analysis*, Schizophrenia research 115 (2009), pp. 346–350.

[33] S. Lewis, T. Foltynie, A. Blackwell, T. Robbins, A. Owen and R. Barker, *Heterogeneity of Parkinson's disease in the early clinical stages using a data-driven approach*, Journal of Neurology, Neurosurgery and Psychiatry 76 (2005), pp. 343–348.

[34] A. Lijoi, R.H. Mena and I. Prünster, *Controlling the reinforcement in Bayesian non-parametric mixture models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2007), pp. 715–740.

[35] A. Lijoi and I. Prünster, *Models beyond the Dirichlet Process.* In Bayesian Nonparametrics (eds. Hjort, N., Holmes, C., Müller, P. and Walker, S.) Cambridge University Press, 2010. pp. 80–136.

[36] S.N. MacEachern, *Estimating normal means with a conjugate style Dirichlet process prior*, Communications in Statistics-Simulation and Computation 23 (1994), pp. 727–741.

[37] S.N. MacEachern and P. Müller, *Estimating mixture of Dirichlet process models*, Journal of Computational and Graphical Statistics 7 (1998), pp. 223–238.

[38] J.M. Marin, K.L. Mengersen and C.P. Robert, *Bayesian Modelling and Inference on Mixtures of Distributions*, Handbook of Statistics 25 (2005), pp. 459–507.

[39] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.

[40] M. Medvedovic and S. Sivaganesan, *Bayesian Infinite Mixture Model Based Clustering of Gene Expression Profiles*, Bioinformatics 18 (2002), pp. 1194–1206.

[41] R.M. Neal, *Markov Chain sampling methods for Dirichlet Process mixture models*, Journal of computational and graphical statistics 9 (2000), pp. 249–265.

[42] R.M. Neal, *Slice sampling*, Annals of Statistics 31 (2003), pp. 705–741.

[43] D.R. Nyholt, N.G. Gillespie, A.C. Heath, K.R. Merikangas, D.L. Duffy and N.G. Martin, *Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities*, Genetic epidemiology 26 (2004), pp. 231–244.

[44] *Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations*, Movement Disorders 18 (2003), pp. 738–750.

[45] J. Reijnders, U. Ehrt, R. Lousberg, D. Aarsland and A. Leentjens, *The association between motor subtypes and psychopathology in Parkinson's disease*, Parkinsonism and Related Disorders 15 (2009), pp. 379–382.

[46] M.C. Schiess, H. Zheng, V.M. Soukup, J.G. Bonnen and H.J. Nauta, *Parkinson's disease subtypes: clinical classification and ventricular cerebrospinal fluid analysis*, Parkinsonism and Related Disorders 6 (2000), pp. 69–76.

[47] G. Schwartz, *Estimating the dimension of a model*, Annals of Statistics 6 (1978), pp. 461–464.

[48] J. Sethuraman, *A constructive definition of Dirichlet priors*, Statistica Sinica 4 (1994), pp. 639–650.

[49] B. Spycher, M. Silverman, A. Brooke, C. Minder and C. Kuehni, *Distinguishing phenotypes of childhood wheeze and cough using latent class analysis*, European Respiratory Journal 31 (2008), pp. 974–981.

[50] M. Stephens, *Dealing with label switching in mixture models*, Journal of the Royal Statistical Society: Series B 62 (2000), pp. 795–809.

[51] G.T. Sutherland, G.M. Halliday, P.A. Silburn, F.L. Mastaglia, D.B Rowe, R.S. Boyle, J.D. O'Sullivan, T. Ly, S.D. Wilton and G.D. Mellick, *Do polymorphisms in the familial Parkinsonism genes contribute to risk for sporadic Parkinson's disease?*, Movement disorders 24 (2009), pp. 833–838.

[52] C.M. Tanner and S.M. Goldman *Epidemiology of Parkinson's disease*, Neurologic Clinics 14 (1996), pp. 317–335.

[53] S.M. van Rooden, M. Visser, D. Verbaan, J. Marinus and J.J. van Hilten, *Motor patterns in Parkinson's disease: A data-driven approach*, Movement disorders 24 (2009), pp. 1042–1047.

[54] S.M. van Rooden, W.J. Heiser, J.N Kok, D. Verbaan, J.J van Hilten and J. Marinus, *The identification of Parkinson's disease subtypes using cluster analysis: a systematic review*, Movement Disorders 25 (2010), pp. 969–978.

[55] S.M. van Rooden, F. Colas, P. Martínez-Martín, M. Visser, D. Verbaan, J.

Marinus, R.K. Chaudhuri, J.N. Kok and J.J. van Hilten, *Clinical subtypes of Parkinson's disease*, Movement Disorders 29 (2011), pp. 51–58.

[56] S. Wade, S. Mongelluzzo and S. Petrone, *An enriched conjugate prior for Bayesian nonparametric inference*, Bayesian Analysis 6 (2011), pp. 359–386.

[57] S.G. Walker, *Sampling the Dirichlet mixture model with slices*, Communications in Statistics-Simulation and Computation 36 (2007), pp. 45–54.

[58] C.D. Walsh, *Latent Class Analysis Identification of Syndromes in Alzheimer's Disease: A Bayesian Approach*, Metodoloski Zvezki: Advances in Methodology and Statistics 3 (2006), pp. 147–162.

[59] M. West, *Hyperparameter Estimation in Dirichlet Process Mixture Models*, ISDS Discussion Paper 92-A03, Duke University, 1992.

[60] M. West, P. Müller and MD. Escobar, *Hierarchical Priors and Mixture Models, with Application in Regression and Density Estimation*, In: Aspects of Uncertainty: A Tribute to D. V. Lindley (eds. AFM. Smith and P. Freeman), Wiley. 1994. pp. 363–386.

[61] N. White, H. Johnson, P. Silburn, G. Mellick, N. Dissanayaka and K. Mengersen, *Probabilistic subgroup identification using Bayesian finite mixture modelling: A case study in Parkinson's disease phenotype identification*, preprint(2010), to appear in Statistical Methods in Medical Research, doi: 10.1177/0962280210391012.

[62] M.A. Wong and T. Lane, *A kth nearest neighbour clustering procedure*, Journal of the Royal Statistical Society: Series B 45 (1983), pp. 362–368.