Identifying latent subgroups of children with developmental delay using Bayesian sequential updating and Dirichlet process mixture modelling.

Patricia Gilholm

School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology Brisbane, Australia.

Kerrie Mengersen

Australian Centre of Excellence in Mathematical and Statistical Frontiers, School of Mathematical Sciences, Science and Engineering Faculty, Queensland University of Technology Brisbane, Australia.

Corresponding Author: Helen Thompson

School of Mathematical Sciences
Queensland University of Technology
GPO Box 2434
Brisbane, Queensland, 4001
Australia

E-mail: helen.thompson@qut.edu.au

Summary. This research aimed to identify latent subgroups of children with developmental delay, by modelling and clustering developmental milestones. The main objectives were to (a) create a developmental profile for each child by modelling milestone achievements across multiple domains of development, and (b) cluster the profiles to identify groups of children who show similar deviations from typical development. The ensemble methodology used in this research incorporates Bayesian sequential updating for modelling the achievement of milestones, which allows for updated predictions of development to be made in real time, and Dirichlet process mixture modelling, a more flexible and adaptive clustering method. The data used were 348 binary developmental milestone measurements from birth to three years of age, from a small community sample of young children (N = 79). The model identified several groups of children with similar features, ranging from no delays in all functional domains, up to large delays in all domains. The performance of the Dirichlet process mixture model was validated with two small simulation studies.

Keywords: Bayesian sequential updating, developmental milestones, Dirichlet process mixture model, latent subgroups, unsupervised clustering

1. Introduction

The early identification of children who have a developmental disability or delay can sometimes be challenging, as developmental delays may occur gradually and only become more evident as a child grows older. As a consequence, children are often referred to intervention services when they are older than three years of age, which may not coincide with the timing of the delay (Bailey et al., 2004). An earlier diagnosis may lead to more prompt access to early intervention. Therefore, understanding the development of at-risk children prior to three years of age is necessary in order to facilitate diagnosis and access to early intervention (Sacrey et al., 2018). This research uses an ensemble method involving Bayesian sequential updating and Dirichlet process mixture modelling (DPMM) to identify the onset and trajectory of delays for children with, or at-risk of developmental disability and to identify latent subgroups of children who have a similar developmental profile from birth to three years of age.

A common approach that is used to identify at-risk children is monitoring and screening developmental milestones. Developmental milestones are behaviours that are displayed by children at certain times during their development from infancy through to school age, so monitoring milestones can provide a systematic approach in which to observe the progress of development over time (Johnson and Blasco, 1997; Petty, 2015). Developmental milestones are currently being used in research to classify children into subgroups that describe their trajectory of development, by using unsupervised clustering methods. Unsupervised clustering refers to a collection of statistical and machine learning methods that divide cohorts into subgroups based on the structure within the data, when there are no class labels available for classification (Marquand et al., 2016). Common unsupervised clustering methods include K-means (Hartigan and Wong, 1979) and finite mixture modelling (McLachlan and Peel, 2000), which is also known as latent class analysis or growth mixture modelling for longitudinal data (Collins and Lanza, 2010).

Unsupervised clustering methods have been applied in retrospective studies to identify subgroups of specific developmental disabilities including Attention-Deficit/Hyperactivity Disorder (Karalunas et al., 2014), Autism Spectrum Disorder(ASD)(Sacco et al., 2012; Wiggins et al., 2017) and Pervasive Developmental Disorders (Shen et al., 2007). Prospective designs have also been used to cluster at-risk infants (Jones et al., 2014). However, these studies often consider only a single developmental disorder, such as ASD (Kim et al., 2016; Landa et al., 2012; Ruzich et al., 2016), or focus on only one domain of development, such as language development (Pickles et al., 2014; Ukoumunne et al., 2012) or communication skills (Määttä et al., 2012). The method developed in this paper is more comprehensive than previous work as it incorporates milestones collected at 58 measurement occasions from birth to three years of age, and includes measurements from six domains of functioning. The sample is also more diverse, as the data are collected from a community sample of young children, including both typically developing children and children with a wide variety of developmental disorders and delays, such as cerebral palsy and ASD. Due to this diverse sample, the purpose of this research is not to identify specific disability groups, but to implement a more personalised approach by, firstly, learning and updating each child's developmental profile as the milestones are met over time and, secondly, identifying latent subgroups of children with similar

developmental profiles.

The proposed method has three main components: (a) model the probability of milestone achievement for the collective set of functional domains for each child using Bayesian sequential updating; (b) calculate the area between each child's probability sequence and a reference sequence; and (c) cluster the obtained areas into subgroups using DPMM. Through using this method, individual predictions of development can be made and updated for each functional domain and the obtained subgroups can be used to assist treatment planning by targeting the specific developmental delays that are characteristic of each subgroup.

Bayesian sequential updating provides a prediction of behaviour based on the information obtained at previous trials or measurement occasions. This is achieved by incorporating previous information into the prior, so that past behaviour has some influence on the posterior estimates (O'Flaherty and Komaki, 1992). This makes it an ideal method for analysing data that are collected over time, as the likelihood needs to only be calculated for the new data in order to update the model parameters (Oravecz et al., 2016). Bayesian sequential updating is commonly applied to clinical trials, including the continual reassessment method for Phase I clinical trials (Cai et al., 2014; Carlin et al., 2010) and Bayesian adaptive design for therapy development (Yin et al., 2012; Zhou et al., 2008). However, to the authors' knowledge, this approach has yet to be applied to modelling developmental milestones. As developmental milestones are sequential, with new behaviours expected to emerge at each month, the use of Bayesian sequential updating means that predictions of a child's development can be made and revised as the child develops, rather than waiting for retrospective reporting and analysis of milestone achievements.

The proposed method in this paper will summarise the sequence of posterior probabilities obtained from each child by calculating the area between the child's sequence and a reference sequence representing a theoretical child who has achieved all milestones. Inspired by the comparison of Kaplan-Meier survival curves, proposed by Chen et al. (2016), the rescaled area between the sequences provides a metric that indicates how dissimilar each child is from typical development. The rescaled areas range from 0 to 1, with values closer to 1 indicating larger differences between the sequences (Chen et al., 2016). The construction of the areas also aids clustering, as it significantly reduces the dimensionality of the data.

The clustering approach used in this paper is Dirichlet process mixture modelling. DPMM and its variants have been applied to numerous clustering problems in health, including stratification of children's health (Molitor et al., 2010), classification of Parkinson's disease (Shahbaba and Neal, 2009; White et al., 2012) and classification of fetal heart rates (Yu et al., 2017). The DPMM is a Bayesian nonparametric model that introduces uncertainty into the number of clusters through partitioning the data stochastically at each iteration of a Markov chain Monte Carlo (MCMC) sampler (Molitor et al., 2010). This approach has the distinct advantage over traditional clustering methods, such as finite mixture modelling and K-means, as it allows the number of clusters to be dictated by the data, meaning that the analyst does not need to specify the number of clusters *a priori* (Molitor et al., 2010). This flexibility is important for the current application, as the data are expected to grow as children respond to more milestones, or more children

join the program. By using a DPMM, the number of clusters can also grow or merge as new data are collected and included in the model.

This paper is structured as follows. In Section 2, the data on developmental milestones that are used in this paper are described. Section 3 outlines, in detail, the three steps for modelling and clustering the milestones. Section 4 presents the clustering results and describes the subgroups that emerged. Finally, a summary of the findings and future research are discussed in Section 5.

2. Data

The data for this study were provided by *The Developing Foundation* (The Developing Foundation Inc, 2018), a Brisbane-based Australian charity that supports families who are seeking treatment for a family member with a brain injury or developmental disability. The organisation collected data on developmental milestones using an online program *Developing Childhood* (Developing Childhood, 2011). The program allows parents and carers to assess and track their child's achievement of developmental milestones from birth to three years. There are 348 milestones in total, which are categorised into six functional domains: Vision, Auditory, Tactile, Movement, Speech and Hand function. Fifty-eight milestones are measured within each of these functional domains. The milestones are not measured uniformly across time; within each functional domain, there are three ordered milestones measured per month in the first 12 months, two ordered milestones measured per month between 13 to 18 months and one milestone measured per month from 19 to 25 months. The remaining three milestones are measured at 28, 31 and 34 months. Example milestones for each functional domain are shown in *Table 1.*

The original sample consisted of data from 118 infants whose parents or carers were voluntarily using the program. This sample consists of both typically developing children and children with a diverse range of developmental disabilities, including Autism Spectrum Disorders, Cerebral Palsy, Down Syndrome, and speech and hearing impairments, as well as more general developmental delays. Although the nature of each child's developmental status is confidential, it is assumed that this sample consists of a larger proportion of children with a developmental disorder or disability than in the general population, as the program was specifically designed for families who seek assistance from *The Developing Foundation.* In order to develop the method, only children with complete data sequences were included in the analysis. Extensions to accommodate missing data points are described in the Discussion. Of the original sample, complete data sequences were available for 79 children.

3. Model Specification

3.1. Bayesian sequential updating

In this study, a child's achievement of milestones over time was represented as a sequence of Bernoulli trials. The milestones were assumed to be independent, where milestone achievement was recorded as $y = 1$ and not achieving a milestone was recorded as $y = 0$. This was considered a reasonable assumption as milestone achievement is not necessarily

Table 1. Example 1, 12, 18 and 34 month milestones in each functional domain.

Functional Domain	1 month	12 months	18 months	34 months
Vision	Instantly blinks at bright light	Television or colourful moving objects capture attention	Visually aware of close and distant world	Recognises and points out tiny details in pictures
Auditory	Instantly startles to sudden loud noise	Listens to speech without distraction from other sounds	Follows simple two-step commands	Comprehends three key words in a sentence
Tactile	Negative response to pain, positive to comfort	Maintains balance with supported stepping	Begins to identify objects by touch alone	Aware of body size in relation to surroundings
Speech	Non-specific cry	Sound-making with intent	Social speech used for interacting	Regular use of speech to tell stories and experiences
Movement	Unrestricted range of movement in all limbs	Walks holding on to one hand	Attempts to run but without a lot of control	Can pedal a tricycle with good control
Hands	Hands mostly fisted or slightly open	Finger feeding with pincer grasp	Stacks 4-6 blocks	Can dress and undress completely

cumulative, in that some children can achieve later milestones without achieving earlier ones. Moreover, the dependency between milestones achieved within each child was modelled through the sequential updating of the prior.

Bayesian sequential updating is a recursive process that can be used for trials that are observed in a sequence, whereby the posterior distribution for the observation(s) in the first trial becomes the prior distribution for the observation(s) in the second trial. The sequential updating of the prior distribution for a series of Bernoulli trials is a simple procedure, as the posterior distribution for z successes out of N trials using a $Beta(\theta|a, b)$ prior has a posterior distribution of the form $Beta(\theta|z + a, N - z + b)$ (Kruschke, 2014). Therefore, for sequential data, the posterior distribution can be updated for each new observation by adding 1 to a for each subsequent success or 1 to b for each subsequent failure. A brief summary of the Bayesian beta-Bernoulli model, is provided in *Appendix 1.*

To perform the sequential updating, a $Beta(1, 1)$ prior was used for the first observation for all participants, as this is a uniform prior with equal probability of success or failure in achieving a milestone. The sequential updating procedure was then implemented for each individual child, resulting in a series of posterior means, which represent the probability of achieving each milestone based on the child's past milestone achievements. The posterior means for the observed milestone were then plotted across time for each functional domain, along with their 95% highest posterior density (HPD) intervals.

A selection of plots of the posterior means for six children in the Auditory functional domain is displayed in *Figure 1.* Here, it can be seen that there is variability across children in terms of the number of milestones recorded for each child, as well as the progress of development over time. For example, Child 1 responded to all of the milestones in the Auditory functional domain, has very high posterior means for most milestones and only starts to show a slight decline at around the 50th milestone. In contrast, the posterior means for Child 6 are much more variable, with a steeper decline beginning at the 20th milestone.

3.2. Area between posterior probability sequences

In order to compare and cluster the sequences of posterior means, the curve for each child's sequence of posterior means was compared to the curve for a theoretical typically developing child's sequence, by calculating the area between the curves. The theoretical "gold-standard" sequence of posterior means was created by performing Bayesian sequential updating on simulated data for a hypothetical child who had achieved all milestones.

As the sequences of posterior means are stepwise functions, the area between the sequences can be calculated as follows

$$area(F_1, F_2) = \frac{1}{w}\{|F_2(t_n) - F_1(t_n)|(w - t_n) + \sum_{i=1}^{n} |F_2(t_i) - F_1(t_i)|(t_{i+1} - t_i)\},$$

where F_1 is the step function of the child's posterior means, F_2 is the "gold-standard" development step function, t denotes the discrete milestone time points ranging from

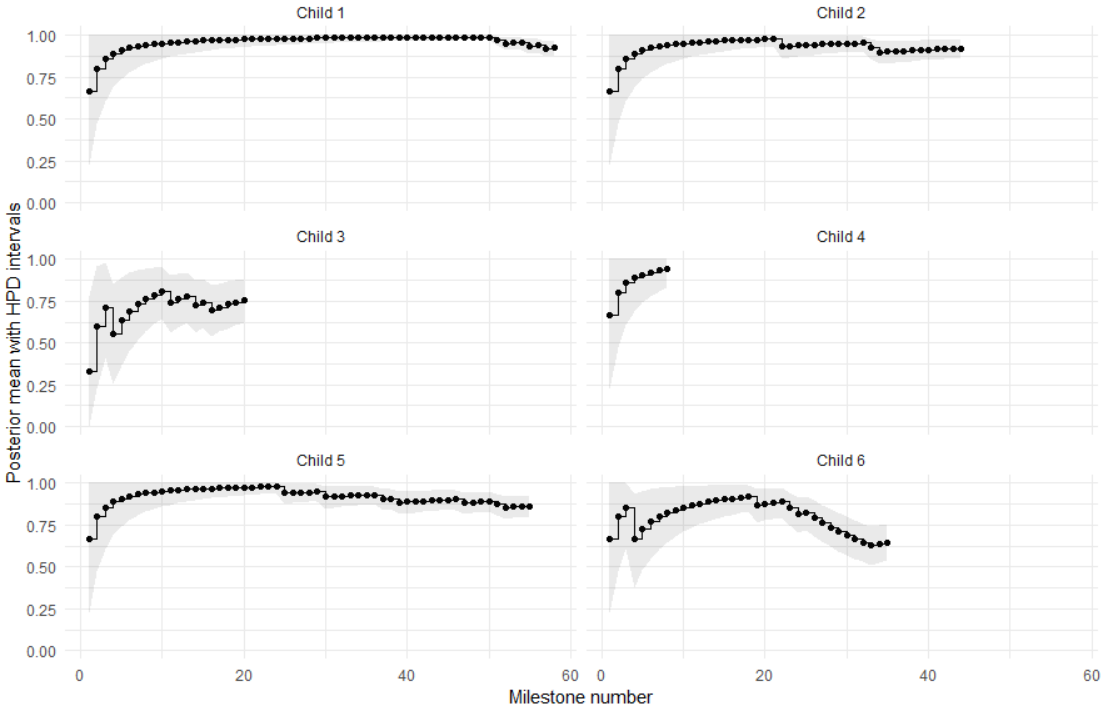

Fig. 1. Posterior means and 95% HPD intervals for the milestones in the Auditory functional domain for six children.

$t = 1 \leq i \leq n$ and w corresponds to the number of observed milestones (Chen et al., 2016). As the number of observed milestones varies across children, the areas are rescaled by the total number of milestones observed by each child, w. This results in a rescaled area between 0 and 1, where scores closer to 0 indicate children whose posterior means are more similar to the "gold-standard" posterior means. An example of the area that is calculated is displayed in *Figure 2*, where the shading represents the absolute area calculated, which is then rescaled by the number of observed milestones.

In order for the theoretical curve to remain the same for all children, the starting point for the theoretical sequence was set to be equal to the starting point of each child's sequence. Six areas were calculated for each child, one for each functional domain. The resulting areas were highly positively skewed with many scores close to 0. In order to assist clustering, the areas were transformed from $[0, 1]$ to $(-\infty, +\infty)$ using the logit transformation.

3.3. Dirichlet process mixture model

The Dirichlet process mixture model is a Bayesian nonparametric method for unsupervised clustering. A general description of the DPMM is available in *Appendix 2.* This paper used the *stick-breaking representation* method for drawing samples from a Dirichlet process, which was first established by Sethuraman (1994). In this representation,

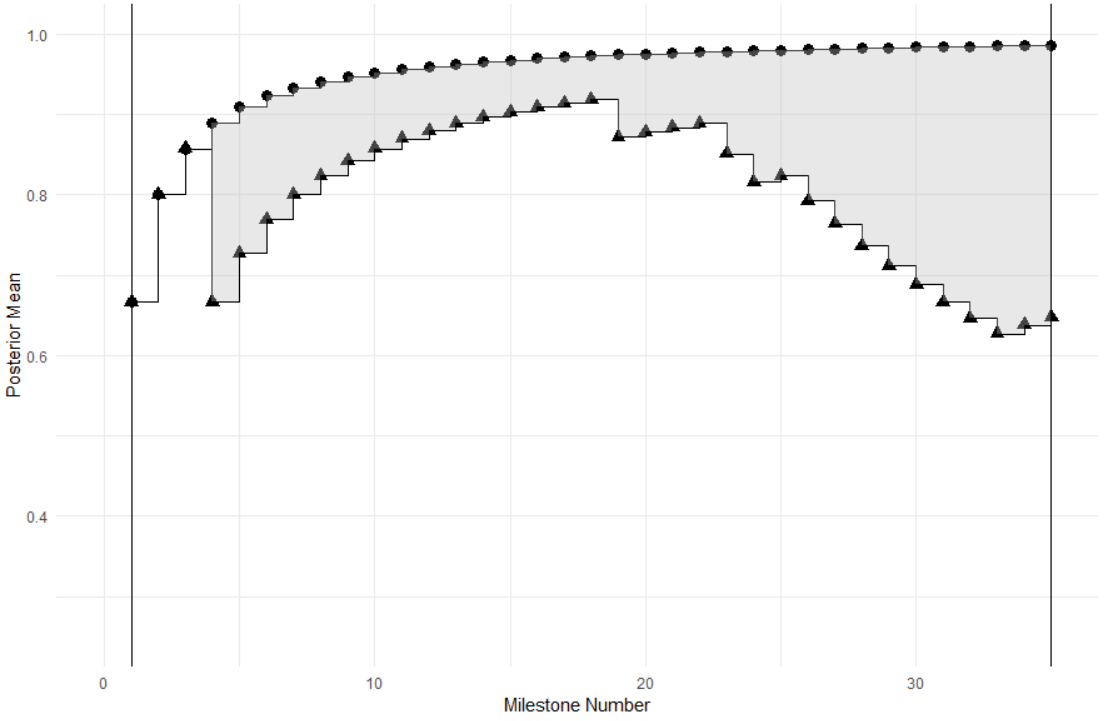

Fig. 2. Example of the area calculated between the theoretical "gold-standard" posterior means (circles) and the child's posterior means (triangles).

the mixing distribution G is represented by an infinite sum of weighted point masses:

$$G = \sum_{k=1}^{\infty} C_k \delta_{\theta_k},$$

where δ_{θ_k} represents a point mass of 1 located at θ_k which is sampled directly from the base distribution, G_0, i.e., $\theta_k \sim G_0$ (Walker, 2007). The weights C_k are generated sequentially through the stick-breaking process:

$$\begin{aligned} V_1, V_2, ... &\overset{iid}{\sim} Beta(1, \alpha) \\ C_1 &= V_1 \\ C_k &= V_k \prod_{j=1}^{k-1} (1 - V_j); \qquad k \geq 2. \end{aligned}$$

The stick-breaking analogy refers to the generation of the weights, where the stick starts with a length of one and the first weight is broken off from the stick at length C_1. The remaining stick has a length of $1 - C_1$ and C_2 is broken off from this length of stick (Teh, 2011). This process continues for each successive break, where the stick can theoretically be broken an infinite amount of times.

Posterior inference from a DPMM utilises Markov chain Monte Carlo(MCMC) posterior simulation (Müller and Quintana, 2004). A number of different methods have been established that use Gibbs sampling, including blocked sampling (Ishwaran and James, 2001), retrospective sampling (Papaspiliopoulos and Roberts, 2008) and slice sampling (Walker, 2007). This research implemented the slice sampling procedure, established by Walker (2007). An outline of the slice sampler is provided in *Appendix 3*.

Due to the nature of the stick-breaking construction of the Dirichlet process, there exists a size-biased ordering of the expected prior mixture probabilities, e.g., $E[C_k] > E[C_{k+1}]$ for all k (Hastie et al., 2015). Therefore, the Gibbs sampler needs to adequately mix over the cluster labels, otherwise clusters with lower labels will have an unfair advantage, as they are given higher prior probability (Porteous et al., 2012). In order to prevent the Gibbs sampler from getting stuck in local modes corresponding to one assignment of cluster labels, label-switching moves were implemented as outlined in Papaspiliopoulos and Roberts (2008).

In this paper, the Dirichlet process mixture model was implemented, as outlined above, to model a mixture of p-dimensional multivariate normal distributions, whereby, conditional on each cluster k, the likelihood for y_i is

$$p(y_i|z_i = k, \boldsymbol{\mu}_k, \Sigma_k) = MVN(\boldsymbol{\mu}_k, \Sigma_k)$$

with mean $\boldsymbol{\mu}_k = [\mu_{1k}, ..., \mu_{pk}]$ and variance-covariance matrix Σ_k. A joint prior distribution $p(\boldsymbol{\mu}_k, \Sigma_k) = p(\boldsymbol{\mu}_k|\Sigma_k)p(\Sigma_k)$ was used, as outlined in van Havre et al. (2016), where

$$\begin{aligned} p(\boldsymbol{\mu}_k|\Sigma_k) &= MVN(\boldsymbol{b}_0, \Sigma_k/N_0) \\ p(\Sigma_k) &= IW(c_0, C_0). \end{aligned}$$

The prior distribution for the concentration parameter, α, was $Gamma(\eta_1, \eta_2)$, as commonly used for DPMMs (Ishwaran and James, 2002).

As each iteration of the MCMC Gibbs sampler provides an estimate of the cluster labels, the posterior expected adjusted Rand index (PEAR) method, proposed by Fritsch et al. (2009), was used to obtain an overall estimate of the optimal number of clusters. As the number of clusters K varies between iterations, the proposed method uses the posterior similarity matrix $P(z_i = z_j|y)$, containing the pairwise probabilities that two observations belong to the same cluster (Fritsch et al., 2009). The best clustering is chosen as the one that maximises the posterior expected Rand index, $E(AR(z^*, z)|y)$, where z^* is any potentional clustering and z is the unknown true clustering (Fritsch et al., 2009).

4. Results

4.1. Grid experiment: Milestone data

A grid experiment was undertaken to test the performance of the DPMM for different hyperparameter specifications for the milestone data. The values of the hyperparameters varied as follows. The precision parameter N_0 was set to $1, 0.5, 0.2, 0.1, 0.05$, or 0.01. The degrees of freedom for the inverse Wishart, c_0, was either 6 or 7. The scale parameter

for the inverse Wishart, C_0, was selected as either the covariance matrix of the data Σ_y, $0.75\Sigma_y$ or $0.5\Sigma_y$, and finally, the prior for α was specified as either $Gamma(1,1)$ or $Gamma(2,2)$. The vector of prior means $\mathbf{b}_0$ did not vary between models and was fixed at $\mathbf{b}_0 = \bar{\mathbf{y}}$. This resulted in 72 combinations of hyperparameters. A subset of the combinations are displayed in *Table 2*. The full list of hyperparameter combinations, as well as the effective sample size and autocorrelation statistics for K and α, can be found in *Tables 1-3* of the supporting materials.

The prior values chosen for the multivariate normal parameters were considered relatively uninformative while remaining within a plausible range for each parameter (van Havre et al., 2016). Similar priors have been used in van Havre et al. (2016), Frühwirth-Schnatter (2006) and Fraley and Raftery (2007). The precision parameter N_0 is analogous to adding N_0 observations to each group in the data (Fraley and Raftery, 2007). This parameter greatly influences the dispersion of the group means and, therefore, the values for this hyperparameter varied from small to large in order to compare and select the hyperparameter that would provide optimal dispersion. The hyperparameters for the prior on α were selected based on recommendations in the literature. The $Gamma(1,1)$ prior was selected so that small values for α were more likely to be sampled, which results in the allocation of the data to fewer clusters (Gelman et al., 2013). The $Gamma(2,2)$ prior was selected as it encourages both small and large values of α to be sampled (Ishwaran and James, 2002).

The grid experiment was performed in two stages. In the first stage, each model ran for 100,000 iterations. Based on convergence statistics of the chains for the number of clusters K and α, 15 models that had the best effective sample size and autocorrelation for the different values of N_0 were selected. The hyperparameters for the 15 selected models are displayed in *Table 2*, along with the optimal number of clusters (based on the first chain) retrieved from each model. In the second stage, three chains for each of the 15 models were specified and run for 1,000,000 iterations each, to assess the long-term behaviour of the slice sampler. The three chains for each model were initialised using K-means, with the number of clusters specified as $K = 5$, $K = 10$ and $K = 15$, respectively. R statistical software (R Core Team, 2018) was used to conduct the slice sampling. The slice sampling code is publicly available on Github (White, 2015).

Due to the small sample size and "noisy"(i.e., individuals do not group easily into clusters) data, the precision hyperparameter, N_0, had the largest effect on the performance of the slice sampler. For smaller values of N_0, there was larger dispersion in the group means, so the sampler would only sample a small number of clusters. This resulted in very little movement in the chain for K, small effective sample sizes and high autocorrelation between iterations. For larger values of N_0, there was less dispersion, and more variation in the number of clusters sampled at each iteration, resulting in better convergence. However, this also resulted in a higher number of clusters to be sampled. The traceplots of α and K for the 15 selected models are accessible through *Section 4* of the supporting materials. The chains for each model were assessed for convergence using the Gelman-Rubin statistic (Gelman et al., 1992) and the optimal number of clusters for each chain were obtained using the PEAR post-processing method. The cluster allocations corresponding to the maximum PEAR value were assigned to each chain and then the consistency of allocations across the chains was assessed using alluvial plots.

Table 2. Hyperparameters and the number of obtained clusters for the 15 selected models.

Model	N_0	c_0	C_0	α	No. of clusters
1	0.01	6	Σ_y	Gamma(1,1)	3
2	0.01	7	$0.5\Sigma_y$	Gamma(2,2)	3
3	0.05	7	Σ_y	Gamma(1,1)	3
4	0.05	7	$0.5\Sigma_y$	Gamma(1,1)	12
5	0.05	6	Σ_y	Gamma(2,2)	3
6	0.05	7	$0.5\Sigma_y$	Gamma(2,2)	9
7	0.10	7	Σ_y	Gamma(1,1)	9
8	0.10	7	Σ_y	Gamma(2,2)	13
9	0.10	7	$0.75\Sigma_y$	Gamma(2,2)	14
10	0.20	7	$0.75\Sigma_y$	Gamma(1,1)	15
11	0.20	7	Σ_y	Gamma(2,2)	13
12	0.50	7	Σ_y	Gamma(1,1)	14
13	0.50	7	Σ_y	Gamma(2,2)	19
14	1.00	7	Σ_y	Gamma(1,1)	16
15	1.00	7	Σ_y	Gamma(2,2)	19

The convergence statistics and alluvial plots can be found in *Section 5* of the supporting materials.

4.2. Description of clusters

The procedure outlined above for assessing convergence and post-processing the cluster assignments was performed on the 15 selected models. Model convergence was achieved for 10 out of 15 models based on a Gelman-Rubin statistic of less than 1.1 for both K and α. Of the models that reached convergence, models 7, 8 and 9 were chosen for further investigation because they had the most consistent cluster allocations across the three chains. These three models had the same prior specification for N_0 ($N_0 = 0.10$), and the number of clusters were not too few or too many for the sample size and the study domain (range = 9 to 16 clusters). All three models were comparable, however, Model 7 was ultimately selected as the best model as it returned the smallest number of clusters across all three chains, so was deemed the most parsimonious. Chain 1 of Model 7 was selected as the final model as it did not produce clusters of size $N = 1$ compared to the other chains. This final model consisted of nine clusters and on further inspection the majority of these clusters were also represented in the clustering solutions for Models 8 and 9.

The descriptive statistics for each group of the chosen model can be found in *Table 3* and the profiles for each group can be found in *Figure 3*. The main characteristics of the nine groups are as follows. Group 1 is the largest group, consisting of children with relatively small delays for each functional domain. This group contains a number of outlying individuals, whose profiles do not fit with the characteristics of the other groups. Group 2 contains children who have large delays for all functional domains. Group 3 is characterised by larger delays for the auditory, tactile and vision domains, and close to typical development for movement. Group 4 only contains two children, who both have achieved all milestones in the auditory and hands domains and have slight

Table 3. Group size, mean area and (standard deviation) for each functional domain, per cluster.

Group	Group size	Auditory		Hands		Movement		Speech		Tactile		Vision	
1	22	0.027	(0.051)	0.009	(0.020)	0.013	(0.033)	0.032	(0.058)	0.008	(0.017)	<0.001	(0.001)
2	8	0.166	(0.124)	0.105	(0.087)	0.101	(0.085)	0.212	(0.185)	0.095	(0.108)	0.095	(0.075)
3	14	0.099	(0.078)	0.026	(0.045)	<0.001	(<0.001)	0.047	(0.103)	0.048	(0.054)	0.082	(0.099)
4	2	0.000	(0.000)	0.000	(0.000)	0.063	(0.087)	0.003	(0.004)	0.049	(0.054)	0.002	(0.002)
5	6	<0.001	(<0.001)	0.002	(0.003)	0.001	(0.001)	0.016	(0.018)	0.016	(0.026)	0.018	(0.014)
6	10	<0.001	(0.001)	<0.001	(0.001)	<0.001	(0.001)	<0.001	(<0.001)	<0.001	(<0.001)	<0.001	(0.002)
7	10	0.095	(0.072)	0.001	(0.001)	0.044	(0.055)	0.135	(0.140)	0.059	(0.057)	0.058	(0.057)
8	3	0.043	(0.033)	0.000	(0.000)	0.022	(0.005)	0.296	(0.161)	0.000	(0.000)	0.083	(0.056)
9	4	0.052	(0.067)	0.156	(0.206)	0.140	(0.100)	0.002	(0.004)	0.053	(0.102)	0.078	(0.142)

delays in the movement and tactile domains. Group 5 is characterised by children who have experienced minimal delays early on in development and they differ from Group 6 who have experienced no delays or very minimal delays later in development. Group 7 is characterised by some delays in all functional domains, except for hand function. Group 8 only consists of three individuals who have very large speech delays and finally Group 9 only consists of four children that have some delays across all domains, except for speech. Additional plots that display the cumulative sum of the achieved milestones for each group can be found in *Section 6* of the supporting materials.

Groups 4, 8 and 9 have the smallest cluster sizes and therefore could be considered outliers. Alternatively, these groups could represent emerging clusters that would have a larger representation if more data were collected. Similarly, the outlying observations in Group 1 may also split to form their own representative clusters when additional information is collected from new observations. The splitting and merging of clusters can be seen when observing the differences in cluster allocations across chains, even when the chains have converged and there is little difference in the maximum PEAR values. This uncertainty in the number of clusters is a key feature of DPMMs which can be influenced by prior specifications, particularly when being applied to noisy, multivariate data. Therefore, careful consideration of the resulting clusters needs to be made to determine if these clusters are feasible or meaningful to the study domain.

4.3. Sensitivity analysis: Simulated data

Two simulation studies were performed to illustrate how the proposed DPMM performs for (a) well separated, adjacent or overlapping clusters and (b) small, medium or large sample sizes. For each scenario, three bivariate clusters were simulated using the clusterlab package in R (John, 2018). For scenario 1, three small clusters, with 50 observations in each cluster, were simulated and compared under three conditions to assess the DPMM's performance when clusters are overlapping. A small sample size was chosen in order to make comparisons to the application data, which has a small sample size. In the first condition, the three clusters were visibly well-separated, in the second condition the three clusters were adjacent but not overlapping and in the third condition the three clusters were slightly overlapping. The simulated data used in scenario 1 are displayed in the first row of *Figure 4*. The same 15 models that were specified in the grid experiment were used in the simulation study, except for the c_0 hyperparameter, which was specified as 2 or 3, instead of 6 or 7, as fewer dimensions were simulated.

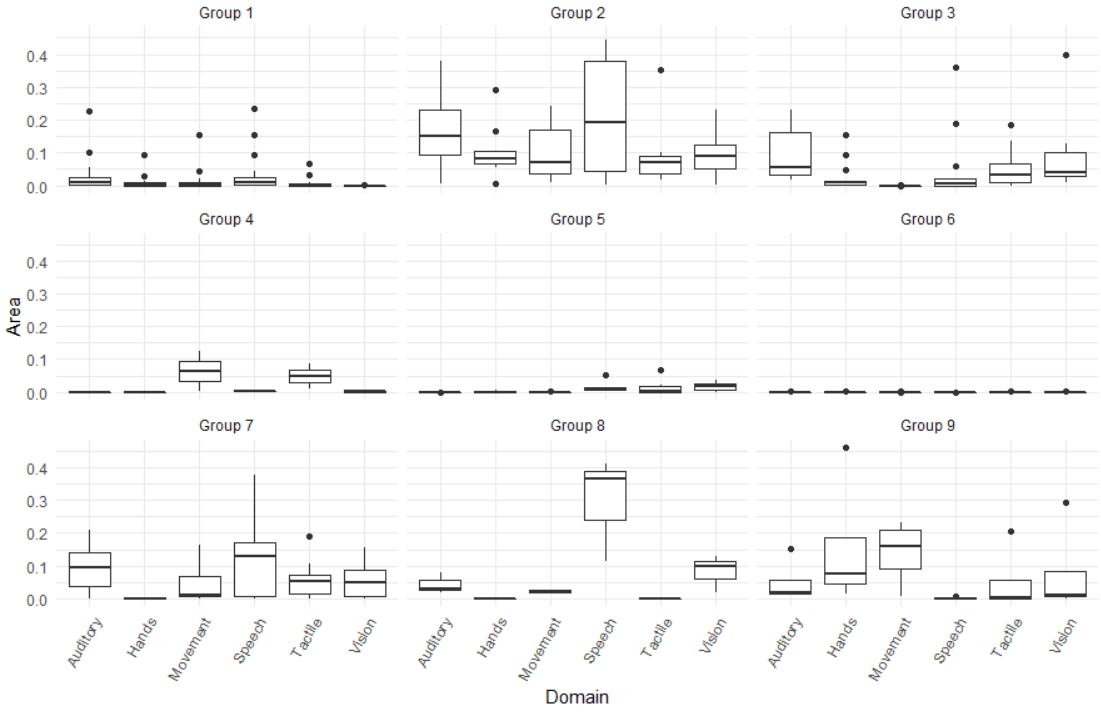

Fig. 3. Cluster profiles for the nine predicted groups from the selected model.

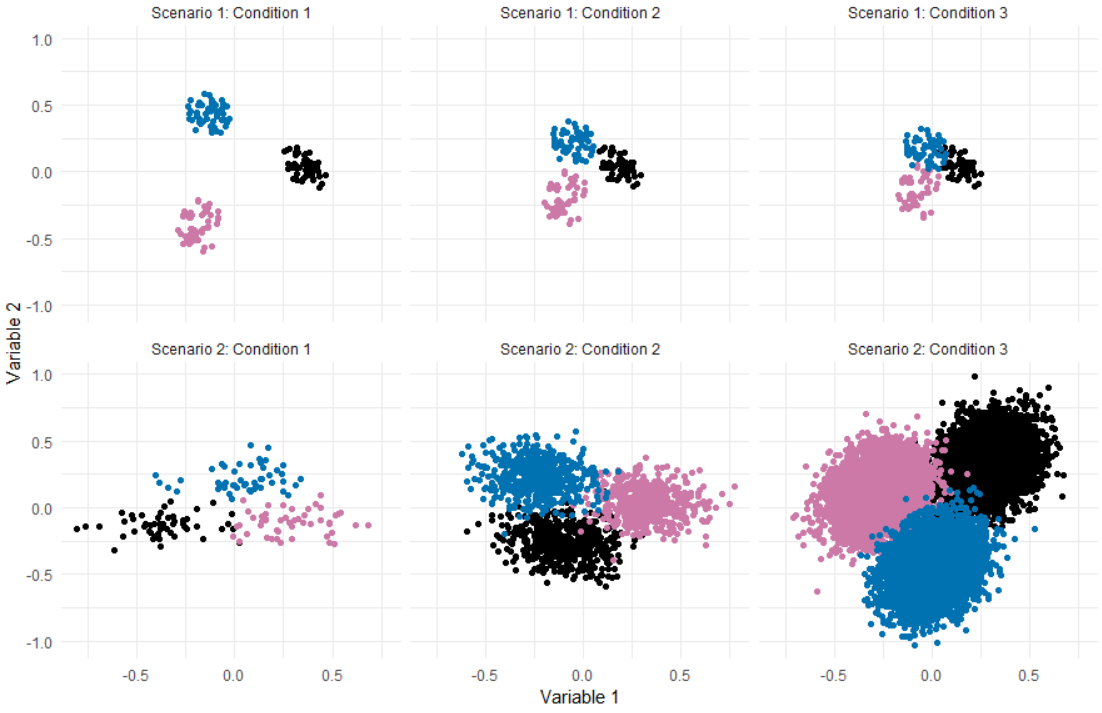

Fig. 4. Simulated data used for each condition in the sensitivity analysis. The first row contains the data used for scenario 1 and the second row contains the data used for scenario 2.

Three chains were specified for each model and each chain ran for 1,000,000 iterations, discarding the first 400,000 as burn-in. The Gelman-Rubin statistic for K and α, the average maximum-PEAR value and the average classification accuracy for each model are displayed in *Section 7* of the supporting materials.

For conditions 1 and 2, the DPMM returned exactly the same clusters that were simulated, regardless of which hyperparameters were used. This was not the case for condition 3 where the cluster compositions varied across the models. All the models correctly identified and classified one of the clusters, however the boundary for the other two clusters varied across the models, resulting in most models returning one larger cluster and one smaller cluster (average classification accuracy across all models = 80.43%). This demonstrates that, even for relatively simple cases, if the data are small, noisy and overlapping, clustering using a DPMM can result in different clustering solutions due to the uncertainty introduced into the clusters and the choice of hyperparameters.

The second simulation study was undertaken to identify if the difficulties associated with the clustering of overlapping, noisy data would remain if more observations were collected. For this scenario, three slightly overlapping bivariate clusters were sampled, which differed in terms of sample size for each condition. For condition 1, each cluster consisted of 50 observations, for condition 2, each cluster contained 500 observations and for condition 3 there were 5000 observations in each cluster. The sample sizes were selected such that condition 1 was the same size used in scenario 1 and conditions 2 and

3 increased the sample size by a magnitude of 10. The simulated data used for scenario 2 is displayed in the second row of *Figure 4*. For this simulation, the 15 models ran for only 40,000 iterations, with the first 20,000 discarded as burn-in. The number of iterations were less than those used in the previous applications as the average run time for the large sample size, using only 10,000 iterations, was 16hr:37min to run the MCMC sampler and 50hr:36min to run the PEAR post-processing of the cluster labels and used, on average, 23.9GB of RAM per model. All models were run on a HPC cluster, inclusive of Intel E5-2670, E5-2680v2, E5-2680v3 and 6140 CPU models. Due to these time and memory requirements, a different approach was used to assess convergence of the chains and a different method, based on hierarchical clustering, was used to post-process the chains. Details of these alternative methods can be found in *Section 8* of the supporting materials.

The average classification accuracy for each model within each condition is displayed in *Section 9* of the supporting materials. In summary, the average classification accuracy across all models for all scenarios was high (small = 94.88%, medium = 96.34%, large = 98.57%), despite the difficulties associated with processing the largest sample size. This high accuracy was due to all models retrieving three clusters as the optimal number of clusters, regardless of the prior specifications of the models. There were no major differences across models, indicating that the prior specification does not play a large role when there is only a slight amount of overlap in the clusters, particularly when the sample size increases. There is much more overlap and complexity in the data for the application, which explains why there is much more uncertainty in the number of clusters. These simulation results indicate that the clustering accuracy slightly improves with more observations, but at an infeasible computational cost. Although, if the sample size becomes too large, an alternative post-processing method can be implemented.

5. Discussion

This research used an ensemble method for modelling and clustering developmental milestones using Bayesian sequential updating and Dirichlet process mixture modelling. Using Bayesian sequential updating, the probability of achieving each milestone was modelled based on each child's own sequence of milestone measurements. This sequence of probabilities was summarised by calculating the area between each child's sequence and a reference sequence representing "gold-standard" development. The areas were then clustered using DPMM to identify subgroups of children who were experiencing similar delays in development.

This detailed method allows for personalised predictions of milestone achievements to be made, as the updated sequences are constructed only using the child's measurements. The model also introduces uncertainty into the predictions, as each parameter is modelled as a distribution of credible values. This means, in practice, that more detailed predictions can be communicated to parents regarding their child's likely trajectory of development and the certainty associated with each prediction can be conveyed. In addition, by using Bayesian sequential updating, the predictions can be easily updated with the collection of new data, so predictions can be made as the child develops rather than retrospectively. By clustering the probability sequences, children who are experiencing

similar delays are able to be identified, meaning that early interventions can be tailored to meet the needs of each group, allowing for more personalised treatment planning.

By using this approach, in the present application, 9 groups were identified that differed in terms of their level of deviation from typical development, across six functional domains. Although some of the cluster sizes were considered small, these groups represented children that did not have the same developmental pattern as the larger groups. Instead of being placed with the most likely group, which is typical for other unsupervised clustering methods (e.g., K-means), these children were placed in their own emerging cluster group. This is important for clinical practice, as these children can have treatments tailored to meet the unique characteristics of the emerging cluster, rather than have tailored treatments based on clusters that they are "most alike", which may not adequately address the needs of the child.

Despite the practical advantages of using this modelling approach, there are a number of methodological limitations that need to be taken into consideration. Firstly, the developmental milestones within each month were assumed to be sequential, based on information elicited from a domain expert. However, this may not necessarily be the case for every child. A more general model could use the binomial distribution to model the milestones for each month, but this approach was not considered here. Secondly, this model did not incorporate any covariates. Covariate information was not available for the sample that was used in this study, but there are a number of covariates that could have an influence on milestone achievement. Past studies have found significant environmental and prenatal predictors of developmental delay, including birth complications and maternal education (Sonnander and Claesson, 1999), poverty and caregiver cognitive impairment (Scarborough et al., 2009), and low birth weight (Schieve et al., 2016). The method developed in this paper could benefit from incorporating this type of covariate information into the model, to create more accurate predictions. Thirdly, the method developed in this paper is most effective with complete data sequences, as it can overestimate the degree of delay when there are missing data points. Imputation or functional data approaches could be explored to rectify this problem, but these approaches were outside the scope of this paper. Modelling the milestones using a functional data approach will be explored in future work.

Additional considerations need to be made when using DPMM. Despite its advantages over other clustering methods, several modelling decisions need to be made in order to obtain the most efficient results, including hyperparameter choice, method for sampling from the posterior and technique used for post-processing the MCMC chains. Each one of these aspects of modelling using a DPMM needs to be carefully considered, as different choices can lead to substantial differences in the clusters.

The importance of hyperparameter choice was highlighted in the grid experiment performed in Section 4, where changing the precision hyperparameter resulted in large differences in the obtained clusters. This hyperparameter significantly effects the dispersion of the cluster means. Selecting a value for this hyperparameter requires careful consideration as a value that is too small will result in over-dispersion, leading to poor convergence and a small number of clusters, whereas a value that is too large will result in too many clusters being sampled. Comparing different hyperparameter specifications is recommended, and the best model selected based on both an inspection of the conver-

gence statistics as well as judgement of the obtained clusters to see if they are sensible given the data and the application domain.

In this paper, the slice sampler was selected as the method for sampling from the posterior distribution of the Dirichlet process. There are, however, several alternative samplers that can be used, for example, the blocked sampler (Ishwaran and James, 2001) and the retrospective sampler (Papaspiliopoulos and Roberts, 2008). The slice sampler was used in this application as it adaptively selects the number of mixture components (Rodriguez et al., 2011) and easily updates them at each iteration (Papaspiliopoulos, 2008). Also, unlike the truncated methods, it targets the true posterior rather than an approximation (Favaro et al., 2013). However, this method does have some limitations. Due to the high correlation between each slice from the slice sampler and the mixture weights, the number of components sampled at each iteration can be large if the slice is small (Favaro et al., 2013; Kalli et al., 2011). This can result in slow mixing and high autocorrelations (Rodriguez et al., 2011), as was the case in this research. However, as these samplers are often developed and illustrated on simulated or low-dimensional datasets, it is likely that similar problems would be encountered using alternative samplers when applied to complex data, such as that used in the current application (Hastie et al., 2015).

The final aspect of DPMM that requires consideration is the choice of method for post-processing the MCMC chains to obtain the optimal number of clusters. The PEAR method was used in this paper, however, alternative methods have been proposed that also use the posterior simularity matrix, including Binder's loss function (Hurn et al., 2003) and hierarchical clustering (Medvedovic et al., 2004). The PEAR method was chosen as a previous study demonstrated that this method was better able to estimate clusters closer to the true clustering than the alternative methods, particularly for overlapping clusters (Fritsch et al., 2009). However, this method is computationally intensive, as shown in the sensitivity analysis in Section 4. Therefore, an alternative method for larger sample sizes was demonstrated, using a hierarchical clustering approach, which often produces comparable results to PEAR and is computationally fast (Fritsch et al., 2009). However, this approach can be considered *ad hoc*, has no consistent guidelines for cutting the dendrogram and tends to produce more clusters of size $N = 1$ (Fritsch et al., 2009).

Overall, the DPMM approach presented here allows for flexibility in modelling and does not require the specification of the number of clusters *a priori*. Additionally, the DPMM takes into account emerging clusters, which makes it ideal for the current application, as it is expected that the clusters will grow or merge as more data are collected. When combined with the Bayesian sequential updating, this model demonstrates a new approach to modelling developmental milestones, which can provide detailed information regarding a child's development, as well as assist in the formulation of personalised early interventions targeted for developmental delays that occur throughout the early, most critical, years of development.

6. Acknowledgements

Work by Patricia Gilholm was supported by an Australian Technology Network of Universities Industry Doctoral Training Centre scholarship, co-funded by QUT and the Developing Foundation. Furthermore, the data used in this research was generously provided by The Developing Foundation.

References

Bailey, D. B., Hebbeler, K., Scarborough, A., Spiker, D. and Mallik, S. (2004) First experiences with early intervention: a national perspective. *Pediatrics*, **113**, 887–896.

Cai, C., Yuan, Y. and Ji, Y. (2014) A bayesian dose finding design for oncology clinical trials of combinational biological agents. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 159–173.

Carlin, B. P., Berry, S. M., Lee, J. J. and Muller, P. (2010) *Bayesian adaptive methods for clinical trials*. CRC press.

Chen, D., Wang, H., Sheng, L., Hueman, M. T., Henson, D. E., Schwartz, A. M. and Patel, J. A. (2016) An algorithm for creating prognostic systems for cancer. *Journal of medical systems*, **40**, 160.

Collins, L. M. and Lanza, S. T. (2010) *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, vol. 718. John Wiley & Sons.

Developing Childhood (2011) Developing childhood. URL: `http://www.developingchildhood.com.au/home`.

Favaro, S., Teh, Y. W. et al. (2013) Mcmc for normalized random measure mixture models. *Statistical Science*, **28**, 335–359.

Ferguson, T. S. (1973) A bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Fraley, C. and Raftery, A. E. (2007) Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, **24**, 155–181.

Fritsch, A., Ickstadt, K. et al. (2009) Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis*, **4**, 367–391.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer Science & Business Media.

Gelman, A., Rubin, D. B. et al. (1992) Inference from iterative simulation using multiple sequences. *Statistical science*, **7**, 457–472.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian data analysis*. Chapman and Hall/CRC.

Ghosal, S. (2010) The dirichlet process, related priors and posterior asymptotics. *Bayesian nonparametrics*, **28**, 35.

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the dirichlet process. *Scandinavian journal of statistics*, **28**, 355–375.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 100–108.

Hastie, D. I., Liverani, S. and Richardson, S. (2015) Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, **25**, 1023–1037.

van Havre, Z., White, N., Rousseau, J. and Mengersen, K. (2016) Clustering action potential spikes: Insights on the use of overfitted finite mixture models and dirichlet process mixture models. *arXiv preprint arXiv:1602.01915*.

Hurn, M., Justel, A. and Robert, C. P. (2003) Estimating mixtures of regressions. *Journal of computational and graphical statistics*, **12**, 55–79.

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

— (2002) Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical statistics*, **11**, 508–532.

John, C. R. (2018) *clusterlab: Flexible Gaussian Cluster Simulator*. URL: `https://CRAN.R-project.org/package=clusterlab`. R package version 0.0.2.5.

Johnson, C. P. and Blasco, P. A. (1997) Infant growth and development. *Pediatr Rev*, **18**, 224–242.

Jones, E. J., Gliga, T., Bedford, R., Charman, T. and Johnson, M. H. (2014) Developmental pathways to autism: a review of prospective studies of infants at risk. *Neuroscience & Biobehavioral Reviews*, **39**, 1–33.

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

Karalunas, S. L., Fair, D., Musser, E. D., Aykes, K., Iyer, S. P. and Nigg, J. T. (2014) Subtyping attention-deficit/hyperactivity disorder using temperament dimensions: toward biologically based nosologic criteria. *JAMA psychiatry*, **71**, 1015–1024.

Kim, S. H., Macari, S., Koller, J. and Chawarska, K. (2016) Examining the phenotypic heterogeneity of early autism spectrum disorder: subtypes and short-term outcomes. *Journal of Child Psychology and Psychiatry*, **57**, 93–102.

Kruschke, J. (2014) *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Landa, R. J., Gross, A. L., Stuart, E. A. and Bauman, M. (2012) Latent class analysis of early developmental trajectory in baby siblings of children with autism. *Journal of Child Psychology and Psychiatry*, **53**, 986–996.

Määttä, S., Laakso, M.-L., Tolvanen, A., Ahonen, T. and Aro, T. (2012) Developmental trajectories of early communication skills. *Journal of Speech, Language, and Hearing Research*.

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J. and Beckmann, C. F. (2016) Beyond lumping and splitting: a review of computational approaches for stratifying psychiatric disorders. *Biological psychiatry: cognitive neuroscience and neuroimaging*, **1**, 433–447.

McLachlan, G. and Peel, D. (2000) Finite mixture models, willey series in probability and statistics.

Medvedovic, M., Yeung, K. Y. and Bumgarner, R. E. (2004) Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, **20**, 1222–1232.

Molitor, J., Papathomas, M., Jerrett, M. and Richardson, S. (2010) Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, **11**, 484–498.

Müller, P. and Quintana, F. A. (2004) Nonparametric bayesian data analysis. *Statistical science*, 95–110.

Neal, R. M. (2000) Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, **9**, 249–265.

— (2003) Slice sampling. *Annals of statistics*, 705–741.

O'Flaherty, B. and Komaki, J. L. (1992) Going beyond with bayesian updating. *Journal of Applied Behavior Analysis*, **25**, 585–597.

Oravecz, Z., Huentelman, M. and Vandekerckhove, J. (2016) Sequential bayesian updating for big data. *Big Data in Cognitive Science*, **13**.

Papaspiliopoulos, O. (2008) A note on posterior sampling from dirichlet mixture models. *manuscript, Department of Economics, Universitat Pompeu Fabra*.

Papaspiliopoulos, O. and Roberts, G. O. (2008) Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.

Petty, K. (2015) *Developmental milestones of young children*. Redleaf Press.

Pickles, A., Anderson, D. K. and Lord, C. (2014) Heterogeneity and plasticity in the development of language: A 17-year follow-up of children referred early for possible autism. *Journal of Child Psychology and Psychiatry*, **55**, 1354–1362.

Porteous, I., Ihler, A. T., Smyth, P. and Welling, M. (2012) Gibbs sampling for (coupled) infinite mixture models in the stick breaking representation. *arXiv preprint arXiv:1206.6845*.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `https://www.R-project.org/`.

Rodriguez, A., Lenkoski, A. and Dobra, A. (2011) Sparse covariance estimation in heterogeneous samples. *Electronic journal of statistics*, **5**, 981.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H. and Baron-Cohen, S. (2016) Subgrouping siblings of people with autism: Identifying the broader autism phenotype. *Autism Research*, **9**, 658–665.

Sacco, R., Lenti, C., Saccani, M., Curatolo, P., Manzi, B., Bravaccio, C. and Persico, A. M. (2012) Cluster analysis of autistic patients based on principal pathogenetic components. *Autism Research*, **5**, 137–147.

Sacrey, L.-A. R., Zwaigenbaum, L., Bryson, S., Brian, J., Smith, I. M., Roberts, W., Szatmari, P., Vaillancourt, T., Roncadin, C. and Garon, N. (2018) Parent and clinician agreement regarding early behavioral signs in 12-and 18-month-old infants at-risk of autism spectrum disorder. *Autism Research*, **11**, 539–547.

Scarborough, A. A., Lloyd, E. C. and Barth, R. P. (2009) Maltreated infants and toddlers: predictors of developmental delay. *Journal of Developmental & Behavioral Pediatrics*, **30**, 489–498.

Schieve, L. A., Tian, L. H., Rankin, K., Kogan, M. D., Yeargin-Allsopp, M., Visser, S. and Rosenberg, D. (2016) Population impact of preterm birth and low birth weight on developmental disabilities in us children. *Annals of epidemiology*, **26**, 267–274.

Sethuraman, J. (1994) A constructive definition of dirichlet priors. *Statistica sinica*, 639–650.

Shahbaba, B. and Neal, R. (2009) Nonlinear models using dirichlet process mixtures. *Journal of Machine Learning Research*, **10**, 1829–1850.

Shen, J. J., Lee, P. H., Holden, J. J. and Shatkay, H. (2007) Using cluster ensemble and validation to identify subtypes of pervasive developmental disorders. In *AMIA Annual Symposium Proceedings*, vol. 2007, 666. American Medical Informatics Association.

Sonnander, K. and Claesson, M. (1999) Predictors of developmental delay at 18 months and later school achievement problems. *Developmental Medicine and Child Neurology*, **41**, 195–202.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004) *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13. John Wiley & Sons.

Teh, Y. W. (2011) Dirichlet process. In *Encyclopedia of machine learning*, 280–287. Springer.

The Developing Foundation Inc (2018) The developing foundation. URL: `https://www.developingfoundation.org.au/`.

Ukoumunne, O., Wake, M., Carlin, J., Bavin, E., Lum, J., Skeat, J., Williams, J., Conway, L., Cini, E. and Reilly, S. (2012) Profiles of language development in pre-school children: a longitudinal latent class analysis of data from the early language in victoria study. *Child: care, health and development*, **38**, 341–349.

Walker, S. G. (2007) Sampling the dirichlet mixture model with slices. *Communications in StatisticsSimulation and Computation®*, **36**, 45–54.

White, N. (2015) Dirichlet process mixture model for spike sorting. URL: `https://github.com/nicolemwhite/spike_sorting_DPM`.

White, N., Johnson, H., Silburn, P. and Mengersen, K. (2012) Dirichlet process mixture models for unsupervised clustering of symptoms in parkinson's disease. *Journal of Applied Statistics*, **39**, 2363–2377.

Wiggins, L. D., Tian, L. H., Levy, S. E., Rice, C., Lee, L.-C., Schieve, L., Pandey, J., Daniels, J., Blaskey, L., Hepburn, S. et al. (2017) Homogeneous subgroups of young children with autism improve phenotypic characterization in the study to explore early development. *Journal of autism and developmental disorders*, **47**, 3634–3645.

Yin, G., Chen, N. and Jack Lee, J. (2012) Phase ii trial design with bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 219–235.

Yu, K., Quirk, J. G. and Djurić, P. M. (2017) Dynamic classification of fetal heart rates by hierarchical dirichlet process mixture models. *PloS one*, **12**, e0185417.

Zhou, X., Liu, S., Kim, E. S., Herbst, R. S. and Lee, J. J. (2008) Bayesian adaptive design for targeted therapy development in lung cancera step toward personalized medicine. *Clinical Trials*, **5**, 181–193.

7. Appendix

7.1. Appendix 1

Recall that for a single Bernoulli trial the likelihood of the data y given θ is

$$p(y|\theta) = \theta^y(1-\theta)^{1-y},$$

where $p(y|\theta) = \theta$ if $y = 1$ and $p(y|\theta) = (1-\theta)$ if $y = 0$. After observing y, the posterior probability for θ becomes

$$p(\theta|y) \propto \theta^y(1-\theta)^{1-y} \times p(\theta),$$

where $p(\theta)$ is the prior distribution for θ. The Beta distribution, which is conjugate to the Bernoulli distribution, can be used as a prior distribution for θ (Spiegelhalter et al., 2004). The density of the Beta distribution is given by

$$p(\theta|a,b) = \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b),$$

where a and b are hyperparameters and $B(a, b)$ is a normalising constant (Kruschke, 2014). When combining the Bernoulli likelihood function with the Beta prior distribution for a series of N independent trials with z successes, the posterior distribution for θ is again a Beta, given by

$$p(\theta|z, N) = \theta^{(z+a-1)}(1-\theta)^{(N-z+b-1)}/B(z+a, N-z+b).$$

The posterior mean for θ can be easily computed as $E(\theta|y) = \frac{z+a}{N+a+b}$ and the posterior variance can be computed as

$$var(\theta|y) = \frac{(a+z)(b+N-z)}{(a+b+N)^2(a+b+N+1)}.$$

7.2. Appendix 2

The Dirichlet process was first introduced by Ferguson (1973) and is defined as a probability distribution over random probability measures (Ghosal, 2010). The distribution of a Dirichlet process is (almost surely) discrete, in that a random sample drawn from a Dirichlet process has a nonzero probability that multiple draws will have identical values (Green and Richardson, 2001). It is this discreteness property which makes the Dirichlet process ideal for clustering, as it avoids the need to determine the number of clusters *a priori* (Neal, 2000). The basic Dirichlet process mixture model is formulated as follows:

$$\begin{aligned} y_i|\theta_i &\sim p(y_i|\theta_i) \\ \theta_i|G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

The Dirichlet process models the distribution from which data $y_1, ..., y_n$ are drawn as a mixture of distributions, $p(y_i|\theta_i)$, where each parameter θ_i is drawn from a mixing distribution G (Neal, 2000). This mixing distribution is given a Dirichlet process prior, with concentration parameter $\alpha > 0$ and base distribution G_0. The base distribution is the prior expectation of G, i.e., $E[G] = G_0$, and the concentration parameter acts as an inverse variance where larger values of α result in smaller variances, which creates more concentrated draws around the mean of the base distribution (Teh, 2011).

7.3. Appendix 3

Slice sampling is an efficient adaptation of Gibbs sampling which can adapt easily to non-standard distributions (Neal, 2003). The general premise is to introduce a latent variable u so that the joint density of y and u becomes

$$f_{C,\theta}(y, u) = \sum_k \mathbf{1}(u < C_k) N(y|\theta_k),$$

where u is uniformly distributed on the interval $(0, C_{k_i})$ (Walker, 2007). Given u, the number of components is now finite, consisting of a subset indexed by $A_u = \{k : C_k > u\}$ (Kalli et al., 2011). The complete data likelihood for a sample $i = 1, ..., n$ is given by

$$l_{C,\theta}(y_i, u_i, z_i = k_i) = \prod_{i=1}^{n} \mathbf{1}(u_i < C_{k_i}) N(y_i|\theta_{k_i}),$$

where z is a variable identifying which cluster the observation y_i belongs to, which has the following conditional density:

$$p(z_i = k|...) \propto \mathbf{1}(k \in A_{u_i})N(y_i|\theta_k).$$

The introduction of u means that only a finite set of stick weights, C_k, and corresponding parameters, θ_k need to be sampled at each iteration (Walker, 2007).