Chapter 8
Bayesian Spike Sorting: Parametric and Nonparametric Multivariate Gaussian Mixture Models

Nicole White, Zoé van Havre, Judith Rousseau, and Kerrie L. Mengersen

Abstract The analysis of action potentials is an important task in neuroscience research, which aims to characterise neural activity under different subject conditions. The classification of action potentials, or "spike sorting", can be formulated as an unsupervised clustering problem, and latent variable models such as mixture models are often used. In this chapter, we compare the performance of two mixture-based approaches when applied to spike sorting: the Overfitted Finite Mixture model (OFM) and the Dirichlet Process Mixture model (DPM). Both of these models can be used to cluster multivariate data when the number of clusters is unknown, however differences in model specification and assumptions may affect resulting statistical inference. Using real datasets obtained from extracellular recordings of the brain, model outputs are compared with respect to the number of identified clusters and classification uncertainty, with the intent of providing guidance on their application in practice.

Keywords Mixture model · Dirichlet process · Classification · Spike sorting

8.1 Introduction

Extracellular recordings are a form of electrophysiological data that allows real time monitoring of multiple neurons *in vivo*. Data collection focuses on the measurement of action potentials or "spikes", which characterise local neural activity at a

N. White (✉)
Institute for Health and Biomedical Innovation, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: nm.white@qut.edu.au

Z. van Havre · K. L. Mengersen
School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia

J. Rousseau
Department of Statistics, University of Oxford, Oxford, UK

K. L. Mengersen et al. (eds.), *Case Studies in Applied Bayesian Data Science*, Lecture Notes in Mathematics 2259, https://doi.org/10.1007/978-3-030-42553-1_8

given point in time. Analysis of these data aims to estimate both the number of active source neurons present and their relative frequency. Comparing the results of analysis across different subject conditions can therefore provide insight into changes in neural activity, for example, in different regions of the brain or in response to various stimuli.

The analysis of extracellular recordings consists of two main stages: (1) spike detection, and (2) the assignment of detected spikes to source neurons. This chapter focuses on the assignment stage, also known as *spike sorting* [1, 2]. A common assumption underpinning spike sorting methods is that different neurons generate action potentials with a characteristic, repeatable shape. Spike sorting can therefore be viewed as an unsupervised clustering problem where spikes with similar features are grouped together, for example, based on summary statistics [3, 4] or low-dimensional transformations of the data, such as wavelet transforms or principal components analysis [1, 5].

Mixture models offer a general solution for unsupervised clustering and are a popular tool for spike sorting, including cases where the number of source neurons (clusters) is unknown. Applications of mixture models to spike sorting have included finite mixtures of Gaussian [2, 6] and t-distributions [7], mixtures of factor analysers [8], Reversible Jump Markov chain Monte Carlo (RJMCMC) [9], and time-dependent mixtures to account for non-stationarity in waveforms [10, 11]. Nonparametric mixture models based on the Dirichlet Process (DP) have also been proposed [12, 13].

Different mixture-based approaches all aim to determine the optimal clustering of a dataset. However, differences in model specification can impact subsequent inferences, for example, the number of clusters identified and/or classification uncertainty for individual observations. This chapter aims to provide insight into this issue by comparing two mixture-based approaches to spike sorting. Both are formulated within the Bayesian framework and represent parametric and nonparametric approaches to mixture modelling. The first model is a finite mixture of multivariate Gaussian distributions, applying methodology proposed by [14]. This model initially overfits the number of clusters expected in the data. The prior distribution for the mixture model weights is then specified in a way that encourages excess clusters in the posterior distribution to have negligible weight [15]. The second model considers a nonparametric approach to mixture estimation which uses the DP as a prior over unknown mixture components. Clustering behaviour induced by properties of the DP is then used to estimate the most likely partition of the data.

Outcomes from each approach are compared with respect to the number of clusters identified, the predicted classification of individuals spikes, and the features of identified clusters.

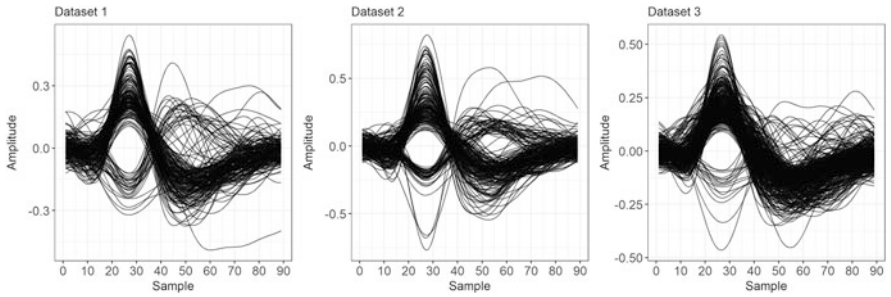

Fig. 8.1 Sampled spikes from three extracellular recordings. Each spike is represented by 89 samples, equivalent to 1 ms of recording. Datasets varied by sample size (L to R): $n = 192, 211, 348$

8.2 Data

Selected approaches were applied to data from three independent extracellular recordings of the brain (Fig. 8.1). Each spike was represented by a waveform consisting of 89 samples, corresponding to 1 millisecond of recording time. The number of detected spikes for analysis was equal to 192, 211 and 348 for Datasets 1, 2 and 3, respectively.

Dimension reduction was performed on sampled waveforms for each dataset in Fig. 8.1 using a robust version of Principal Components Analysis (PCA) [16]. This method was chosen to lessen the influence of outliers on the estimation of principal components. The first four principal components were used as inputs into each mixture model (Fig. 8.2), which explained 83% (Dataset 1), 91% (Dataset 2), and 85% (Dataset 3) of total variation in sampled waveforms.

8.3 Methodology

In this section, key features of each mixture modelling approach are outlined. Common to both approaches is the problem of classifying n spikes into K clusters, where K is *a priori* unknown. Individual spikes in each model are represented by a multivariate vector $\mathbf{y}_i = \{y_{i1}, \ldots, y_{ir}\}$, containing r measurements for spike i.

For the data described in Sect. 8.2, $\mathbf{y}_i$ is assumed to follow a Multivariate Gaussian distribution with mean $\boldsymbol{\mu}_k = [\mu_{1k}, \ldots, \mu_{rk}]$ and variance-covariance matrix Σ_k, $1 \leq k \leq K$. Conditional on assignment to cluster k, the likelihood for $\mathbf{y}_i$ is,

$$p\left(\mathbf{y}_i | z_i = k, \boldsymbol{\theta}_k\right) = N_r\left(\boldsymbol{\mu}_k, \Sigma_k\right), \tag{8.1}$$

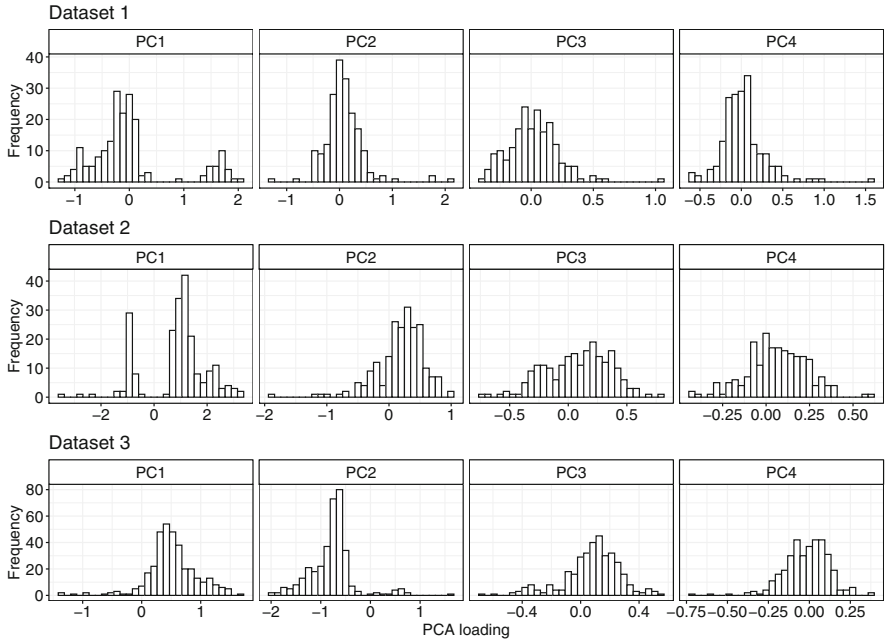

Fig. 8.2 Distribution of the first four principal components of each original dataset. Each row represents a dataset (Dataset 1, 2, 3) and each column represented a principal component (PC1, PC2, PC3, PC4)

with unknown parameters $\boldsymbol{\theta}_k = \left(\boldsymbol{\mu}_k, \Sigma_k\right)$. For each cluster, the joint prior distribution for θ_k takes the form:

$$p\left(\boldsymbol{\theta}_k\right) = p\left(\boldsymbol{\mu}_k | \Sigma_k\right) p\left(\Sigma_k\right) \tag{8.2}$$

with

$$\begin{aligned} p\left(\boldsymbol{\mu}_k | \Sigma_k\right) &= N_r\left(\mathbf{b}_0, \frac{\Sigma_k}{N_0}\right) \\ p\left(\Sigma_k\right) &= IW\left(c_0, \mathbf{C}_0\right). \end{aligned} \tag{8.3}$$

The assignment each spike to available clusters is inferred using a discrete latent variable z_i, where $z_i = k$ if spike i is assigned to cluster k. The inclusion of z_i is a form of data augmentation [17], and is required for sampling from the posterior distribution.

All models were estimated using Markov chain Monte Carlo (MCMC), with details provided Sects. 8.3.1 and 8.3.2. For analyses presented in Sect. 8.4, the

following values were chosen for the hyperparameters: $\mathbf{b}_0 = \overline{\mathbf{y}}$, $N_0 = 0.01$, $c_0 = 5$ and $\mathbf{C}_0 = 0.75\text{cov}(\mathbf{y})$. These values were chosen to reflect a plausible range of values for each parameter, whilst remaining relatively non-informative. Other hyperparameter choices for multivariate Gaussian mixture models are discussed in [18].

8.3.1 Overfitted Finite Mixture Model (OFM)

The first approach involves fitting a finite mixture model where the number of clusters is set to be greater than the number of clusters expected in the data. We refer to this approach as the Overfitted Finite Mixture model (OFM) [14]. Assuming $K^* > K$ clusters are fitted to the data, the likelihood of $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ under the OFM is,

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \sum_{k=1}^{K^*} \pi_k N_r\left(\boldsymbol{\mu}_k, \Sigma_k\right), \tag{8.4}$$

where $\pi_k = Pr\left(z_i = k\right)$, is the prior probability of a randomly selected observation being assigned to cluster k. Collectively, $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_{K^*}\}$ represent the mixture model weights and are subject to the constraint $\sum_{k=1}^{K^*} \pi_k = 1$.

Under the OFM, the prior distribution for z_i given $\boldsymbol{\pi}$ is Multinomial,

$$z_i|\boldsymbol{\pi} \sim MN\left(1; \pi_1, \ldots, \pi_{K^*}\right), \tag{8.5}$$

which allows $\mathbf{z} = \{z_1, \ldots, z_n\}$ to be sampled at each MCMC iteration via the posterior probabilities of cluster membership:

$$p(z_i = k|\mathbf{y}_i, \boldsymbol{\theta}) = \frac{\pi_k N_r\left(\boldsymbol{\mu}_k, \Sigma_k\right)}{\sum_{l=1}^{K} \pi_l N_r\left(\boldsymbol{\mu}_l, \Sigma_l\right)} \tag{8.6}$$

$$\propto \pi_k N_r\left(\boldsymbol{\mu}_k, \Sigma_k\right). \tag{8.7}$$

The defining feature of the OFM is the choice of prior distribution for the mixture model weights. As per the specification of a finite mixture model, weights are assumed to follow a Dirichlet distribution,

$$\left(\pi_1, \ldots, \pi_{K^*}\right) \sim D(\alpha_1, \ldots, \alpha_{K^*}), \tag{8.8}$$

which is characterised by the hyperparameters $\alpha_1, \ldots, \alpha_{K^*}$. In the absence of prior information, it is common to set these hyperparameters to a common value; i.e. $\alpha_1 = \cdots = \alpha_{K^*} = \gamma$. Building on results by [15], the OFM chooses an appropriate value for γ that results in weights for excess components $\{k = K + 1, \ldots, K^*\}$

being shrunk towards zero. When fitted to the observed data, the number of unique values of $\mathbf{z}$ is an estimate of the true number of clusters, K.

The proposed methodology was recently applied by [14] for the case of univariate Gaussian distributions. A key feature of the methodology was the use of prior tempering on the hyperparameter γ. Briefly, a ladder of T values $\{\gamma^{(1)}, \ldots, \gamma^{(T)}\}$ was created, where each element was chosen *a priori* to promote emptying behaviour, based on the results of [15]. The MCMC algorithm was implemented in parallel in combination with Gibbs sampling steps for the remaining model parameters. Code used to implement the MCMC algorithm for the OFM model presented in this chapter is available online (https://github.com/zoevanhavre/Zmix_devVersion2).

8.3.2 Dirichlet Process Mixture Model (DPM)

The second approach considers a nonparametric alternative to mixture modelling by using the Dirichlet Process (DP) as a prior over unknown mixture components. The DP is a stochastic process which is defined as a distribution over probability measures; i.e. a single draw from the DP is itself a distribution [19]. For a measureable space $\boldsymbol{\Theta}$, the data generating process for $\mathbf{y}_i$ under the DP is,

$$\begin{aligned} \mathbf{y}_i|\boldsymbol{\theta}_i &\sim \boldsymbol{\theta}_i \\ \boldsymbol{\theta}_i|G &\sim G \\ G &\sim DP\,(mG_0)\,. \end{aligned} \tag{8.9}$$

The random probability measure G follows a DP defined by a base distribution, G_0, and a concentration parameter $m > 0$. G_0 is interpreted as the mean of the DP, and is assigned as suitable distribution according to the form of $\boldsymbol{\theta}_i$.

Under the DP, draws for multiple $\boldsymbol{\theta}_i$ have a non-zero probability of taking the same value. This discreteness property induces clustering of the observed data, which can be seen in different formulations of the DP. Under the stick-breaking construction [20], G is replaced with an infinite weighted sum of point masses:

$$\begin{aligned} G &= \textstyle\sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \\ \pi_k &= v_k \textstyle\prod_{l<k} (1 - v_l) \\ v_k &\sim Beta\,(1, m) \\ \theta_k|G_0 &\sim G_0 \end{aligned} \tag{8.10}$$

where $G_0 = p\,(\boldsymbol{\theta}_k)$ and $\delta_{\boldsymbol{\theta}_k}$ denotes a Dirac mass at $\boldsymbol{\theta}_k$. The term 'stick-breaking' refers to the analogy that the weights $\pi_1, \pi_2, \ldots$ represent portions of a stick with total length equal to 1. Conditional on preceding clusters, each π_k is a randomly

drawn proportion of stick length remaining so that $\sum_{k=1}^{\infty} \pi_k = 1$. For this reason, the DPM is often referred to as an infinite mixture model [19].

An alternative construction of the DP is the Polya Urn scheme [21] or Chinese restaurant process. Under this construction, G in integrated out, resulting in the following prior predictive distribution distribution for $\boldsymbol{\theta}_i$,

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1}, \ldots, \boldsymbol{\theta}_1, m, G_0 \sim \frac{mG_0}{m+i-1} + \sum_{k=1}^{K-1} \frac{N_k \delta_{\boldsymbol{\theta}_k}}{m+i-1} \tag{8.11}$$

or, in terms of z_i,

$$p(z_i = k|z_1, \ldots, z_{i-1}, m) = \begin{cases} \frac{N_k}{i-1+m} & 1 \leq k \leq K \\ \frac{m}{i-1+m} & k = K+1. \end{cases} \tag{8.12}$$

where N_k is the number of observations already assigned to cluster k. The DPM therefore assumes that each observation has a probability of being assigned to an existing cluster $(1, \ldots, K)$, or representing a new cluster $(K+1)$.

The DPM includes a additional concentration parameter, m, which influences the level of clustering in the data. For example, under the stick-breaking construction in Eq. (8.10), m influences draws for the stick-breaking weights, $v_1, v_2, \ldots$ which, in turn, are used to compute the mixture weights. This parameter can be treated as an unknown parameter in the DPM; in this chapter, we assume $m \sim \Gamma(1, 1)$.

For results presented in Sect. 8.4, DPM models were estimated using slice sampling [22]. This algorithm is based on the stick-breaking construction (8.10) and involves a modified version of Eq. (8.6) to account for an unspecified number of clusters. Uniform auxiliary variables, $u_i \sim U(0, \pi_{z_i})$, based on current values for the mixture weights are introduced to sample each z_i. Additional clusters are proposed until the condition

$$\sum_{k=1}^{K^*} \pi_i > 1 - \min\{u_1, \ldots, u_n\} \tag{8.13}$$

is met, with K^* being the number of clusters sampled for the current MCMC iteration. R code to implement the DPM slice sampler is available online (https://github.com/nicolemwhite/spike_sorting_DPM).

8.3.3 *Comparing Spike Sorting Solutions*

For each dataset in Fig. 8.2, OFM and DPM model outputs were compared to determine the effects of model specification on the estimated number of clusters and classification outcomes.

Number of Clusters The number of non-empty clusters was recorded at the end of each MCMC iteration, as an estimate of the true number of clusters. The resulting distribution of K over all MCMC iteration provided an indication of the most likely number of clusters and associated uncertainty.

Optimal Classification Using MCMC samples for $\mathbf{z}$, pairwise posterior probabilities were calculated to infer the optimal partition of each dataset. For each pair of observations i and $i^{'}$, the posterior pairwise probability $Pr\left(z_i = z_{i'}|\mathbf{y}\right)$ was calculated as the proportion of MCMC iterations where i and i' were assigned to the same cluster, irrespective of the value of k. A benefit of using these probabilities is that it avoids the need to correct for label switching [23]. The resulting $n \times n$ matrix of probabilities was then used to determine the maximum a posteriori (MAP) estimate of $\mathbf{z}$ [24]. In this chapter, optimal partitions under each DPM were estimated using the Posterior Expected Rand (PEAR) index proposed by [25].

Modelling results were based on 20,000 MCMC iterations, following an initial burn-in phase of 20,000 iterations. OFM estimation assumed an initial estimate of 10 clusters and the proposed tempering algorithm was implemented using $\boldsymbol{\gamma} = 2^{\{-32,-16,-8,-4,-2,0,2,4\}}$. MCMC sampling was further initialised by applying the k-means clustering algorithm to each dataset with $k = 10$.

8.4 Results

Differences in the estimated number of clusters were observed between models, with DPM model outcomes subject to greater posterior uncertainty (Fig. 8.3). Across all datasets, fitted OFM models converged to 4 clusters and showed little to

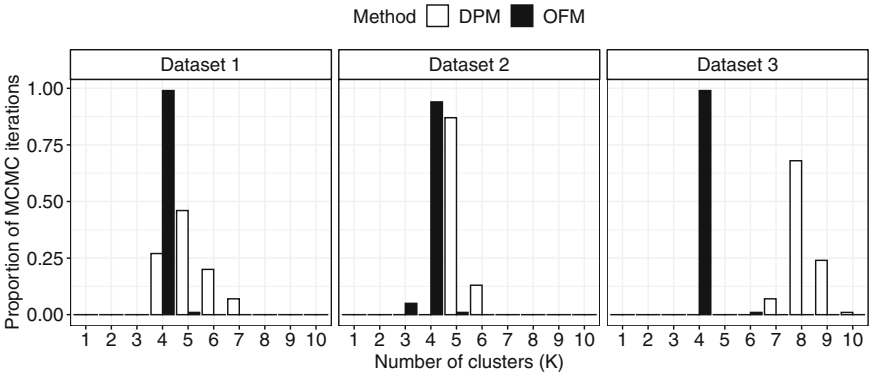

Fig. 8.3 Posterior distributions of the estimated number of clusters in Datasets 1, 2 and 3. Distributions were based on the MCMC output for the DPM (white) and OFM (black)

no support for other values of K. Uncertainty in the number of clusters among DPM models was greatest for Dataset 1, which inferred between 4 and 7 clusters with similar support across MCMC iterations. Discrepancies in the most likely number of clusters were largest for Dataset 3, with 63% of MCMC iterations proposing 8 clusters under the DPM model.

The visualisation of pairwise posterior probabilities suggested that the classification of spikes in Datasets 1 and 2 was robust to the choice of mixture model, despite evidence of differences in the true value of K (Fig. 8.4). Corresponding MAP estimates for $\mathbf{z}$ showed that the optimal clustering based on DPM models included an additional cluster, however in each case this cluster only contained a single observation (Table 8.1). Differences in pairwise posterior probabilities between models fitted to Dataset 3 were more pronounced, and were associated with a sparser clustering of spikes under the DPM model. However, additional clusters predicted by this model also had relatively low weights, representing between 0.3% and 4.3% of identified spikes.

The projection of optimal classifications onto the original data in Fig. 8.1 provided further insight into additional clusters generated under each DPM model (Fig. 8.5). For Datasets 1 and 2, the assignment of waveforms to Clusters 1, 2 and 3 was generally consistent under both approaches. Underlying spike shapes across these clusters were clearly defined, and were distinguished from one another based on minimum and maximum amplitudes. Defining features for Cluster 4 under each OFM model were less clear, and appeared to represent outlying observations; spike sorting solutions under corresponding DPM models instead attributed these observations to multiple clusters.

The assignment of outliers to singleton clusters was also observed for Dataset 3, however further inconsistencies between models indicated greater sensitivity in DPM parameter estimates. For example, spikes assigned to Cluster 3 of the OFM model varied substantially with respect to maximum amplitude. Results from the corresponding DPM model represented the same spikes by 2 smaller clusters with different maximum amplitudes.

8.5 Discussion

Using the example of spike sorting, this chapter has compared two popular approaches to mixture modelling, to assess the effect of model specification on statistical inference. Both methods represented the observed data as a mixture of multivatiate Gaussian distributions and assumed that true number of clusters was unknown *a priori*.

Differences in model specification affected the estimation of K, with fitted DPM models associated with greater numbers of clusters. This outcome can be attributed to the properties of the DP when used as a prior distribution over mixture components. Unlike the OFM which assumes an upper bound on K, the DP prior assumes that observations can either be assigned to an existing cluster or be

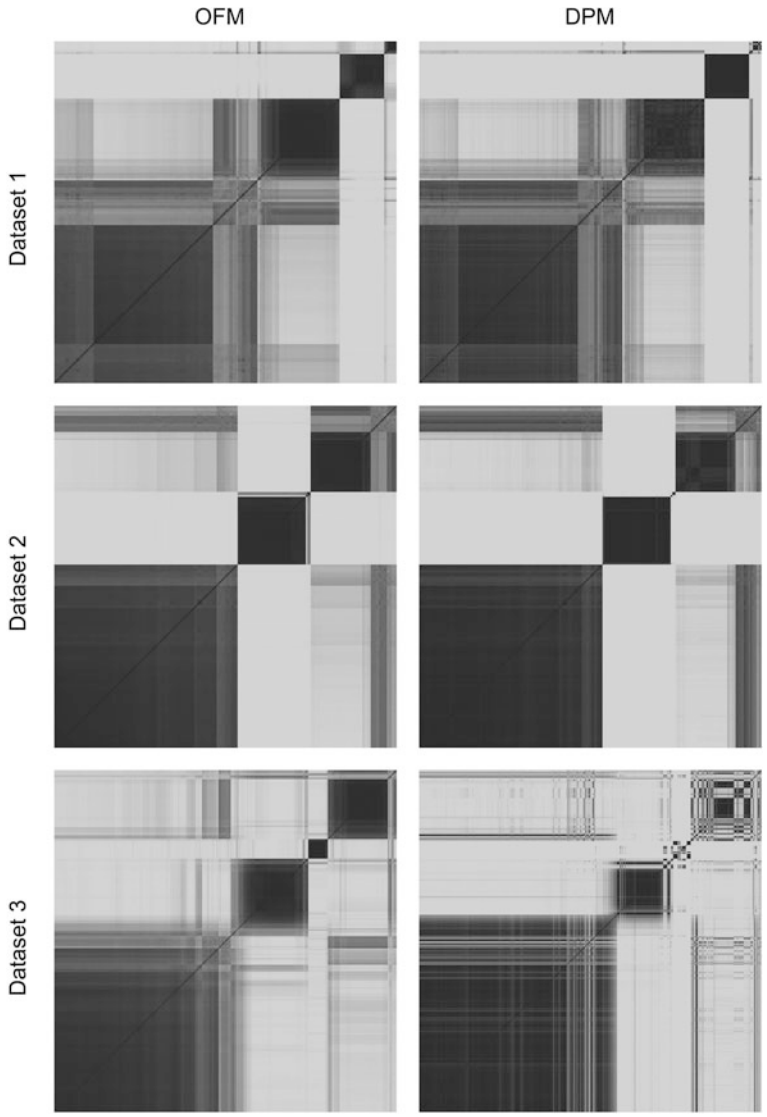

Fig. 8.4 Pairwise posterior similarity matrices for Datasets 1–3, Pairwise posterior similarity matrices for Datasets 1–3, based on MCMC output from the OFM (left column) and DPM (right column). Pairwise posterior probabilities range from 0 (light grey) to 1 (black)

Table 8.1 Frequencies of cluster membership, as determined by the optimal partition under each OFM and DPM model

	Dataset 1 ($n = 192$)			Dataset 2 ($n = 211$)			Dataset 3 ($n = 349$)		
Model	Cluster	Count	%	Cluster	Count	%	Cluster	Count	%
OFM	1	110	57	1	125	59	1	176	50
	2	48	25	2	44	21	2	80	23
	3	25	13	3	40	19	3	72	21
	4	9	5	4	2	1	4	20	6
DPM	1	108	56	1	127	60	1	200	57
	2	50	26	2	42	20	2	51	15
	3	25	13	3	39	18	3	38	11
	4	5	3	4	2	1	4	31	9
	5	4	2	5	1	1	5	15	4
	6	–	–	6	–	–	6	10	3
	7	–	–	7	–	–	7	2	<1
	8	–	–	8	–	–	8	1	<1

Inferred clusters under both models are labelled in decreasing order by frequency

associated with the generation of a new cluster. For results presented in Sect. 8.4, this behaviour led to the generation of additional clusters, however in most cases, these represented a single observation. In contrast, OFM models promoted a parsimonious approach to clustering, whereby outlying observations were allocated to the same cluster. This outcome can be attributed to the prior distribution specified for the unknown mixture weights, as it strongly discourages the posterior from assigning weight to clusters with limited support from the observed data. When applied to spike sorting, small clusters inferred under either approach should therefore be interpreted with care, as these are likely to represent noise as opposed to distinct source neurons.

Optimal spike sorting solutions proposed by OFM and DPM models were similar among clusters with larger weights, and performed well in capturing different waveform shapes. However, greater classification uncertainty under the DPM model reflected potential sensitivity in parameter estimation of multivariate Gaussian distributions. Whilst not considered in this chapter, the use of alternative distributions such as the multivariate-t distribution may help to address this sensitivity. Future studies in this area should therefore consider the effects of model misspecification on the performance of different mixture-based approaches.

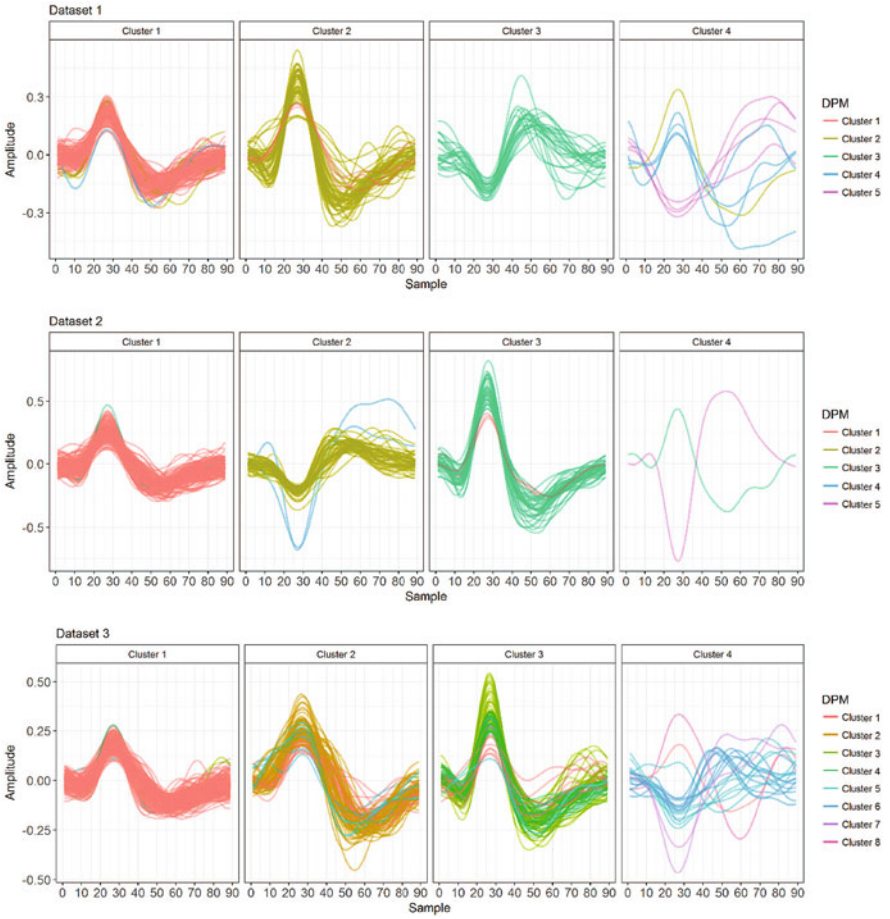

Fig. 8.5 Optimal classifications for Datasets 1, 2 and 3 based on MAP estimates produced by OFM and DPM model. For each dataset, spikes are clustered according to the OFM model. Within each OFM cluster, individual spikes are coloured based on their corresponding classifcation under the DPM model

References

1. M.S. Lewicki, A review of methods for spike sorting: the detection and classification of neural action potentials. Network **9**, R53–R78 (1998)
2. M. Sahani, Latent Variable Models for Neural Data Analysis. PhD Thesis, California Institute of Technology, Pasadena (1999)
3. M. Delescluse, C. Pouzat, Efficient spike-sorting of multi-state neurons using inter-spike intervals information. J. Neurosci. Methods **150**, 16–29 (2006)
4. C. Pouzat, M. Delescluse, P. Viot, J. Diebolt, Improved spike-sorting by modeling firing statistics and burst-dependent spike amplitude attenuation: a Markov chain Monte Carlo approach. J. Neurophysiol. **91**, 2910–2928 (2004)

5. J.C. Letelier, P.P. Weber, Spike sorting based on discrete wavelet transform coefficients. J. Neurosci. Methods **101**, 93–106 (2000)
6. E. Wood, M. Fellows, J.R. Donoghue, M.J. Black, Automatic spike sorting for neural decoding, in *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2 (2004), pp. 4009–4012
7. S. Shoham, M.R. Fellows, R. Normann, Robust, automatic spike sorting using mixtures of multivariate t-distributions. J. Neurosci. Methods **127**, 111–122 (2003)
8. D. Görür, C.E. Rasmussen, A.S. Tolias, F. Sinz, N.K. Logothetis, Modelling spikes with mixtures of factor analysers, in *Joint Pattern Recognition Symposium* (Springer, Berlin, Heidelberg, 2004), pp. 391–398
9. D.P. Nguyen, L.M. Frank, E.N. Brown, An application of reversible-jump Markov chain Monte Carlo to spike classification of multi-unit extracellular recordings. Network **14**, 61–82 (2003)
10. A. Bar-Hillel, A. Spiro, E. Stark, Spike sorting: Bayesian clustering of non-stationary data. J. Neurosci. Methods **157**, 303–316 (2006)
11. A. Calabrese, L. Paninski, Kalman filter mixture model for spike sorting of non-stationary data. J. Neurosci. Methods **196**, 159–169 (2011)
12. F. Wood, M.J. Black, A non-parametric Bayesian approach to spike sorting. J. Neurosci. Methods. **173**, 1–12 (2008)
13. J. Gasthaus, F. Wood, D. Gorur, Y.W. Teh, Dependent Dirichlet process spike sorting, in *Advances in Neural Information Processing Systems* (2009), pp. 497–504
14. Z. van Havre, N. White, J. Rousseau, K. Mengersen, Overfitting Bayesian mixture models with an unknown number of components. PLoS One. **10**, e0131739 (2015)
15. J. Rousseau, K. Mengersen, Asymptotic behaviour of the posterior distribution in overfitted mixture models. J. R. Stat. Soc. (Ser. B) **73** 689–710 (2011)
16. M. Hubert, P.J. Rousseeuw, B.K. Vanden, ROBPCA: a new approach to robust principal component analysis. Technometrics **47**, 64–79 (2005)
17. M.A. Tanner, W.H. Wong, The calculation of posterior distributions by data augmentation. J. Am. Stat. Assoc. **82**, 528–540 (1987)
18. S. Frühwirth-Schnatter, *Finite Mixture and Markov Switching Models* (Springer, New York, 2006)
19. Y.W. Teh, Dirichlet process, in *Encyclopedia of Machine Learning*, ed. by C. Sammut, G.I. Webb (Springer, Boston, 2011)
20. J. Sethuramna, A constructive definition of Dirichlet priors. Stat. Sin. **4**, 639–650 (1994)
21. D. Blackwell, J.B. MacQueen, Ferguson distributions via polya urn schemes. Ann. Stat. **1**, 353–355 (1973)
22. S. Walker, Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. **36**, 45–54 (2007)
23. M. Stephens, Dealing with label switching in mixture models. J. R. Stat. Soc. (Ser. B) **62**, 795–809 (2000)
24. M. Medvedovic, S. Sivaganesan, Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics **18**, 1194–1206 (2002)
25. A. Fritsch, K. Ickstadt, Improved criteria for clustering based on the posterior similarity matrix. Bayesian Anal. **4**, 367–392 (2009)